18th European Young Statisticians Meeting

26 - 30 August 2013 Osijek, Croatia

Proceedings

Organized by

Department of Mathematics
J.J. Strossmayer University of Osijek, Croatia

and

Bernoulli Society for Mathematical
Statistics and Probability

Bernoulli Society
for Mathematical Statistics
and Probability

18th European Young Statisticians Meeting

26-30 August 2013

Department of Mathematics

J.J. Strossmayer University of Osijek, Croatia

Proceedings

Edited by

Nenad Šuvak, Department of Mathematics, J.J. Strossmayer University of Osijek

Bernoulli Society
for Mathematical Statistics
and Probability

Publisher

J.J. Strossmayer University of Osijek, Department of Mathematics

Editor

Nenad Šuvak, Department of Mathematics, J.J. Strossmayer University of Osijek

Technical Editors

Danijel Grahovac, Department of Mathematics, J.J. Strossmayer University of Osijek

Nenad Šuvak, Department of Mathematics, J.J. Strossmayer University of Osijek

CIP record is available in the catalog of the City and University Library in Osijek under number 131021055.

ISBN 978-953-6931-70-5

Preface

European Young Statisticians Meetings are organized every two years under the auspices of the European Regional Committee of the Bernoulli Society for Mathematical Statistics and Probability. The aim is to provide a scientific forum for the next generation of European researchers in probability theory and statistics. It represents an excellent opportunity to promote new collaborations and international cooperation. Participants are less than 30 years old or have 2 to 8 years of research experience, and are invited on the basis of their scientific achievements, in a uniformly distributed way in Europe (at most 2 participants per country).

The 18th European Young Statisticians Meeting (18th EYSM) was held at the Department of Mathematics, J.J. Strossmayer University of Osijek, Croatia, $26-30$ August 2013.

The conference was attended by 45 participants from 25 European countries. The 45 talks were organized in 14 sessions covering 11 different topics (there were no parallel sessions):

1. Statistical inference – estimation
2. Statistical inference – testing procedures
3. Theory of continuous time stochastic processes
4. Inequalities and stochastic ordering
5. Diagnostics and decision theory
6. Optimal design
7. Statistical applications – biology and medicine
8. Statistical applications – economics and insurance

9 Statistical applications – image analysis

10. Statistical applications – engineering, industry and seismology
11. Other topics in statistics and probability.

Five eminent scientist gave keynote lectures:

1. Bojan Basrak, Department of Mathematics, University of Zagreb, Croatia - *On dependent regularly varying observations*
2. Nikolai N. Leonenko, School of Mathematics, Cardiff University, United Kingdom - *Multifractal products of geometric stationary processes*

3. Jürgen Pilz, Alpen-Adria Universität Klagenfurt, Austria - *Some advances in Bayesian spatial prediction and sampling design*

4. Johan Segers, Institut de statistique, biostatistique et sciences actuarielles, Université catholique de Louvain, Belgium - *Semiparametric Gaussian copula models: Geometry and efficient rank-based estimation*

5. Michael Sørensen, Department of Mathematical Sciences, University of Copenhagen, Denmark - *Statistics for stochastic differential equations - two approaches*.

We would like to express our gratitude to the members of the International Organizing Committee for selecting the high-level young scientists for attending this conference, as well as to the reviewers of the papers published in the conference proceedings. We would like to thank to our colleagues from the Department of Mathematics, J.J. Strossmayer University of Osijek, for all their efforts and help. Furthermore, we would like to thank all the sponsors that helped in organizing the 18th EYSM, particularly to the Bernoulli Society for Mathematical Statistics and Probability and to the Department of Mathematics, J.J. Strossmayer University of Osijek. Last, but not least, we thank all the participants for providing an excellent scientific program and a lot of fun during the social events.

It is our pleasure to announce that the 19th European Young Statisticians Meeting will take place in Prague, the capital of Czech Republic. We wish them all the luck!

March, 2014

Nenad Šuvak

on behalf of the Local Organizing Committee

18th European Young Statisticians Meeting

Organizer

Department of Mathematics, J.J. Strossmayer University of Osijek

Auspices

Bernoulli Society for Mathematical Statistics and Probability

J.J. Strossmayer University of Osijek

International Organizing Committee

Jürgen Pilz, Alpen-Adria Universitaet Klagenfurt, Austria

Gregor Kastner, Vienna University of Economics and Business, Austria

Johan Segers, Université catholique de Louvain, Belgium

Metodi Nikolov, University of Sofia, Bulgaria

Marek Dvořák, Charles University in Prague, Czech Republic

Mirta Benšić, J.J. Strossmayer University of Osijek, Croatia

Michael Sørensen, University of Copenhagen, Denmark

Paavo Salminen, Abo Akademi University, Finland

Olivier Bouaziz, Universite Paris Descartes – Paris 5, France

Andrea Krajina, Georgia Augusta University Göttingen, Germany

Ioannis Ntzoufras, Athens University of Economics and Business, Greece

Gyorgy Terdik, University of Debrecen, Hungary

Emanuele Taufer, University of Trento, Italy

Francesca Ieva, Politecnico di Milano, Italy

Harry van Zanten, University of Amsterdam, The Netherlands

Grzegorz Wyłupek, University of Wrocław, Poland

Paulo C. Rodrigues, Nova University of Lisbon, Portugal

Eugenia Panaitescu, ”Carol Davila” Univ. of Medicine and Pharmacy, Romania

Alexey Muravlev, Steklov Mathematical Institute, Russia

Vladimír Lacko, Comenius University, Slovakia

Aleš Toman, University of Ljubljana, Slovenia

Alba Maria Franco Pereira, University of Vigo, Spain

Silvelyn Zwanzig, Uppsala University, Sweden

Johanna Ziegel, University of Bern, Switzerland

Deniz Inan, Marmara University, Turkey

Ludmila Sakhno, Taras Shevchenko National University of Kyiv, Ukraine

Steven Gilmour, University of Southampton, United Kingdom

Nikolai N. Leonenko, Cardiff University, United Kingdom

Local Organizing Committee

Mirta Benšić, J.J. Strossmayer University of Osijek

Danijel Grahovac, J.J. Strossmayer University of Osijek

Petra Posedel, Zagreb School of Economics and Management

Nenad Šuvak, J.J. Strossmayer University of Osijek

Conference structure: keynote lectures, invited lectures

Conference language: English

Contents

18th
European Young
Statisticians Meeting

Abstracts of keynote lectures

Multifractal products of geometric stationary processes

Nikolai N. Leonenko[*]

School of Mathematics, Cardiff University, United Kingdom

Abstract

This is joint work with D. Denisov (Cardiff University).

Multifractal and monofractal models have been used in many applications in hydrodynamic turbulence, finance, genomics, computer network traffic, etc. (see, for example, [7]). There are many ways to construct random multifractal models ranging from simple binomial cascades to measures generated by branching processes and the compound Poisson process ([2] - [7]).

Anh, Leonenko and Shieh ([1]-[3]) and Leonenko and Shieh [8] considered multifractal products of stochastic processes as defined in [9], but they provide a new interpretation of the conditions on the characteristics of geometric stationary processes in terms of the moment generating functions.

We investigate the properties of multifractal products of geometric Gaussian processes with possible long-range dependence and geometric Ornsteinf-Uhlenbeck processes driven by Lévy motion and their finite and infinite superpositions. We present the general conditions for the $\mathcal{L}_q$ convergence of cumulative processes to the limiting processes and investigate their q-th order moments and Rényi functions, which are nonlinear, hence displaying the multifractality of the processes as constructed. We also establish the corresponding scenarios for the limiting processes, such as log-normal, log-gamma, log-tempered stable or log-normal tempered stable scenarios.

Bibliography

[1] Anh, V. V., Leonenko, N. N. and Shieh, N.-R. (2008). Multifractality of products of geometric Ornstein-Uhlenbeck-type processes. *Adv. in Appl. Probab*. **40** 1129–1156.

[2] Anh, V. V., Leonenko, N. N. and Shieh, N.-R. (2009). Multifractal scaling of products of birth-death processes. *Bernoulli* **15** 508–531.

[3] Anh, V. V., Leonenko, N. N., Shieh, N.-R. and Taufer, E. (2010). Simulation of multifractal products of Ornstein-Uhlenbeck type processes. *Nonlinearity* **23** 823–843.

[4] Bacry, E. and Muzy, J.F. (2003). Log-infinitely divisible multifractal processes. *Comm. Math. Phys.* 236 (2003), 449–475.

[5] Barndorf-Nilsen, O.E. and Shmigel, Yu (2004). Spatio-temporal modeling based on Lévy processes, and its applications to turbulence. (Russian) *Uspekhi Mat. Nauk* 59, 63–90; translation in *Russian Math. Surveys* 59, 65–90.

[6] Denisov, D. and Leonenko, N. (2011). Multifractality of products of geometric stationary processes. *Submitted*, published in arxiv.org/abs/1110.2428.

[7] Doukhan, P., Oppenheim, G. and Taqqu, M.S.(2003). *Theory and Applications of Long-range Dependence*. Birkhäuser Boston.

[8] Leonenko, N.N and Shieh N.-R. (2013). Rényi function for multifractal random fields. *Fractals*, in press.

[9] Mannersalo, P., Norris, I. and Riedi, R. (2002). Multifractal products of stochastic processes: construction and some basic properties. *Adv. Appl. Prob.*, 34, 888–903.

[*] e-mail: LeonenkoN@Cardiff.ac.uk

Statistics for stochastic differential equations — two approaches

Michael Sørensen[*]

Department of Mathematical Sciences, University of Copenhagen, Denmark

Abstract

For discrete-time observations of the solution to a stochastic differential equation, there is usually no explicit expression for the likelihood function, which is a product of transition densities. Therefore, the likelihood function must be approximated. A brief review will be given of a broad spectrum of approximation methods. Two approaches will be presented in detail. Martingale estimating functions are a simple way of approximating likelihood inference that provides estimators which are easy to calculate. These estimators are generally consistent, and if the estimating function is chosen optimally, they are efficient in a high frequency asymptotic scenario, where the sampling frequency goes to infinity. At low sampling frequencies, efficient estimators can be obtained by more accurate approximations to likelihood inference based on simulation methods, including both the stochastic EM-algorithm and Bayesian approaches like the Gibbs sampler. These methods are much more computer intensive. Simulation of diffusion bridges plays a central role. Therefore this highly non-trivial problem has been investigated actively over the last 10 years. A simple method for diffusion bridge simulation will be presented and applied to likelihood inference for stochastic differential equations.

[*] e-mail: michael@math.ku.dk

Some advances in Bayesian spatial prediction and sampling design

Jürgen Pilz[*]

Alpen-Adria Universitaet Klagenfurt, Austria

Abstract

In my talk, I will report on recent work with my colleagues G. Spoeck and H. Kazianka in the area of Bayesian spatial prediction and design [1]-[4].

The Bayesian approach not only offers more flexibility in modeling but also allows us to deal with uncertain distribution parameters, and it leads to more realistic estimates for the predicted variances. We report on some experiences gained with our approach during a European project on ”Automatic mapping of radioactivity in case of emergency”.

We then go on and apply copula methodology to Bayesian spatial modeling and derive predictive distributions. Moreover, I report on recent results on finding objective priors for the crucial nugget and range parameters of the widely used Matern-family of covariance functions.

Further on, I briefly consider the challenges in stepping from the purely spatial setting to spatio-temporal modeling and prediction.

Finally, I will consider the problem of choosing an ”optimal” spatial design, i.e. finding an optimal spatial configuration of the observation sites minimizing the total mean squared error of prediction over an area of interest. Using Bessel-sine/cosine- expansions for random fields we arrive at a design problem which is equivalent to finding optimal Bayes designs for linear regression models with uncorrelated errors, for which powerful methods and algorithms from convex optimization theory are available. I will also indicate problems and challenges with optimal Bayesian design when dealing with more complex design criteria such as minimizing the averaged expected lengths of predictive intervals over the area of interest.

Bibliography

[1] H. Kazianka and J. Pilz (2011). Bayesian spatial modeling and interpolation using copulas. *Computers & Geosciences.* 37(3): 310-319.

[2] H. Kazianka and J. Pilz (2012). Objective Bayesian analysis of spatial data taking account of nugget and range parameters. *The Canadian Journal of Statistics.* 40(2): 304-327.

[3] J. Pilz, H. Kazianka and G. Spoeck (2012). Some advances in Bayesian spatial prediction and sampling design. *Spatial Statistics.* 1: 65-81.

[4] G. Spoeck and J. Pilz (2013). Spatial sampling design based on spectral approximations of the error process. In: *Spatio-temporal design: Advances in Efficient Data Acquisition* (W.G. Mueller and J. Mateu, Eds.), Wiley, New York, 72-102

[*] e-mail: Juergen.Pilz@uni-klu.ac.at

Semiparametric Gaussian copula models: Geometry and efficient rank-based estimation

Johan Segers[*][1], **Ramon van den Akker**[2] **and Bas Werker**[2]

[1] *Université catholique de Louvain, Belgium*
[2] *Tilburg University, The Netherlands*

Abstract

For multivariate Gaussian copula models with unknown margins and general correlation structures, a simple, rank-based and semiparametrically efficient estimator is proposed. An algebraic representation of relevant subspaces of the tangent space is constructed that allows to easily study questions of adaptivity with respect to the unknown marginal distributions and of efficiency of the pseudo-likelihood estimator and the normal-scores rank correlation coefficient. Some well-known examples are treated explicitly: circular correlation matrices, factor models, and Toeplitz matrices, special cases being exchangeable structures, moving average models and autoregressive models. For constructed examples, the asymptotic relative efficiency of the pseudo-likelihood estimator can be as low as 20 percent. For finite samples, these findings are confirmed by Monte Carlo simulations.

On dependent regularly varying observations

Bojan Basrak[§]

Department of Mathematics, University of Zagreb, Croatia

Abstract

It is well known that the extremal behavior of stationary sequences can be nicely captured using the language of point processes. We explain how this theory extends from iid to dependent sequences as long as this dependence disappears in time. The theory turns out to be especially elegant when applied to stationary regularly varying sequences, which we discuss in detail.

In particular, the dependence structure of extremes for such sequences can be described using the concept of the tail process. By application of the point processes theory, this leads to various asymptotic results for extremes and sums of such sequences, including some nonstandard functional limit theorems.

[*] e-mail: johan.segers@uclouvain.be
[§] e-mail: bbasrak@math.hr

18th
European Young
Statisticians Meeting

Papers

Robust multivariate process control of multi-way data with root cause analysis

Peter Scheibelhofer[*][1,2] **Günter Hayderer**[2] **and Ernst Stadlober**[1]

[1] *Graz University of Technology, Austria*
[2] *ams AG, Unterpremstätten, Austria*

Abstract

The evaluation of the manufacturing process conditions is a crucial challenge in modern semiconductor fabrication. With growing complexity large numbers of process variables are recorded during equipment operations of each process step. To monitor these processes, traditional fault detection and classification methods were implemented, but they are mostly univariate. Multivariate techniques such as Principal Component Analysis and Hotelling's T^2 are capable of advanced process control but they are mainly based on statistically calculated indicators such as means or standard deviations of one wafer over its process time. Thereby, information of the time variation of the variables is omitted. In this work, we present a generalized methodology for multivariate process control that considers the whole recorded information of a wafer by using multi-way principal component analysis (MPCA). The use of Hotelling's T^2 statistics makes outcomes easy to monitor as it can be summarized into one control chart. By grouping similar variables into reasonable functional groups and by applying decomposition methods for the T^2 signal, a root cause analysis is possible. Furthermore, special attention is paid on the robustness of the MPCA and T^2 procedure as an analysis independent of frequently observed outliers is crucial. In a case study of production data from the Austrian semiconductor manufacturer ams AG an observed production machine error can be detected and its root cause can be tracked down successfully.

Keywords: fault detection, multivariate process control, multi-way principal component analysis, robust statistics
AMS subject classifications: 62P30

1 Introduction

Modern semiconductor fabrication consists of a series of highly complex manufacturing steps resulting in a final product with well-defined electrical properties. To achieve this goal, each step of the batch process has to be monitored adequately. During the processing of one batch (wafer) at a given process stage data information for every observed status variable is typically recorded by sensors with a fixed frequency, e.g. one data point per second. Thus, the total recorded data information of a wafer over its process time can be arranged in a multi-way array with dimension $I \times J \times K$ which holds the information of I wafers on J variables at K observed time points. A suitable method for handling such multi-way data is multi-way principal component analysis (MPCA, see [11]). The result of an MPCA decomposition of a multi-way array is a series of principal components consisting of score vectors of dimension $1 \times I$, loading matrices of dimension $J \times K$ and an $I \times J \times K$ dimensional error array. See figure 1 for an illustration.

[*] Corresponding author, e-mail: peter.scheibelhofer@ams.com

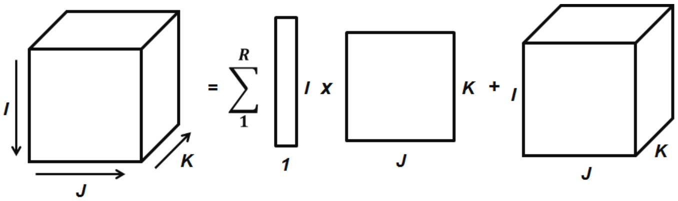


Figure 1: Arrangement and decomposition of a three-way array as a result of MPCA.

As for ordinary principal component analysis (PCA), every wafer gets a unique score value for each principal component based on its respective variation over the process time. Therefore the monitoring of these scores is of interest for process control (see [6]). In order to ensure a reasonable root cause analysis, in a first step the observed variables are arranged in functional variable groups according to their physical relationship. This grouping is achieved with the help of process experts. Then, in a second step an MPCA analysis is performed for each functional group seperately and all of the resulting scores are monitored by implementing Hotelling's T^2 statistics (see [3]) for phase 1 and 2 observations (see [5]). For suspect wafers this approach allows a meaningful application of the Mason-Young-Tracy (MYT) decomposition of the T^2 signal (see [5]) for out-of-control wafers. Thereby, a given problem can be tracked down to a single functional group or a relationship between groups.

2 Robustness of the approach

In order to get robust MPCA results with score vectors and loading matrices not influenced by outlying observations several approaches are possible. Engelen and Hubert (see [1]) proposed an approach for robustly exploring a multi-way array using a parallel factor analysis (PARAFAC) model. Their method is based on the ROBPCA algorithm for robust principal component analysis by Hubert et al. (see [4]). Another possibility to decompose a three-way array is to use the approach by Nomikos and MacGregor (see [6]) based on unfolding the given three-way array to a large matrix of dimension $I \times JK$ and then performing ordinary PCA via the nonlinear iterative partial least squares (NIPALS) algorithm. Based on the ideas of Engelen and Hubert, a robustification of the Nomikos and MacGregor approach can be achieved by also using ROBPCA as a starting point instead of ordinary PCA to calculate scores and loadings. This way, one is able to avoid problems of PARAFAC models with degenerate solutions where the algorithm has difficulties in correctly fitting a model (see e.g. [9], section 5.4).
Furthermore, the well-known Hotelling's T^2 statistics, which is applied to the score vectors, can be robustified by using robust estimates of the mean and variance-covariance structure of the given data (see [10]). Here, we use the minimum covariance determinant (MCD) method by Rousseeuw and van Driessen (see [8]).

3 Case Study

We studied an error of the magnetic field in a plasma etch tool used during wafer processing at Austrian semiconductor manufacturer ams AG. The failure caused a severe decrease in the etch rate of the equipment (see [2]). Classical univariate analysis did not show any severe out-of-control alarms. In the affected month June 2011 the proposed approach was applied to data from about 900 wafers. All computation was done using `R` (see [7]). About 430 wafers were used as historical phase 1 data set to characterize the in-control situation by using the robust MCD estimators. The resulting T^2 control chart of all observed wafers clearly shows significant out-of-control signals for wafers affected by the error (around wafer 500) as visualized in figure 2.

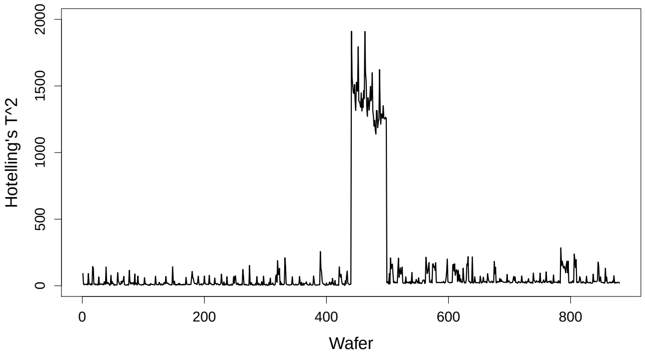


Figure 2: Values of Hotelling's T^2 statistics for about 1000 analyzed wafers.

The respective MYT decomposition of the T^2 signal correctly tracks down the root cause of the problem mainly to the RF functional variable group of the etch tool. The root cause was also confirmed by process engineers. One possible resulting error profile from the T^2 signal decomposition is shown in figure 3. Only wafers affected by the magnetic field error show this particular error fingerprint.

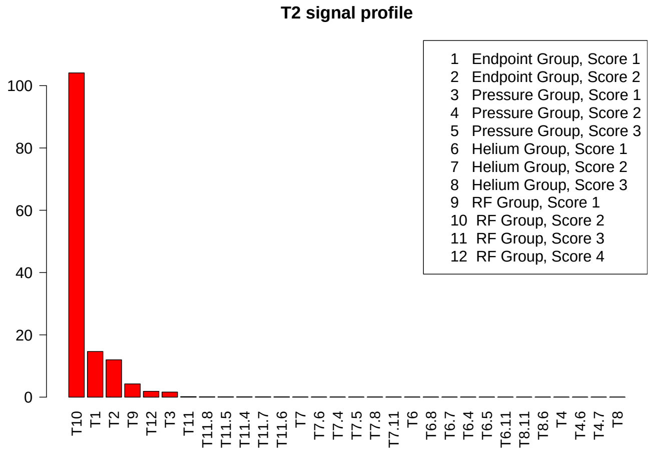


Figure 3: MYT decomposition profile of Hotelling's T^2 signal exemplarily for one wafer affected by the magnetic field error.

Acknowledgements: This work was supported by ams AG. The provision of the data set as well as the company's support and supervision is gratefully acknowledged.

Bibliography

[1] Engelen, S. and Hubert, M. (2011). Detecting Outlying Samples in a PARAFAC Model. *Analytica Chimica Acta* 705, 155–165.

[2] Hayderer, G. (2012). Multivariate Fault Detection and Classification of Magnetic Field Breakdown on a Plasma Etch Tool. *Proceedings of the European Advanced Process Control and Manufacturing (APCM) Conference, April 16–19, 2012, Grenoble, France*.

[3] Hotelling, H. (1947). Multivariate Quality Control Illustrated by Air Testing of Sample Bombsights. *Techniques of Statistical Analysis*, C. Eisenhart, H. Hastay and W.A. Wallis, eds. McGraw-Hill, New York, 111–184.

[4] Hubert, M., Rousseeuw, P., Vanden Branden, K. (2005). ROBPCA: A New Approach To Robust Principal Component Analysis. *Technometrics* 47, 64–79.

[5] Mason, R.L., Young, J.C. (2001). *Multivariate Statistical Process Control with Industrial Applications*. ASA-SIAM series on statistical and applied probability. ASA-SIAM, Philadelphia, PA, USA.

[6] Nomikos, P. and MacGregor, J.F. (1994). Monitoring Batch Processes Using Multiway Principal Component Analysis. *AlChE Journal* 40(8), 1361–1375.

[7] R Core Team (2012). R: A Language and Environment for Statistical Computing [Computer software manual]. Vienna, Austria. Available form http://www.R-project.org/ (ISBN 3-900051-07-0)

[8] Rousseeuw, P.J. and Van Driessen, K. (1999). A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics* 41, 212–223.

[9] Smilde, A., Bro, R. and Geladi, P. (2004). *Multi-way Analysis with Applications in the Chemical Sciences*, John Wiley and Sons, West Sussex, England.

[10] Vargas, J.A. (2003). Robust Estimation in Multivariate Control Charts for Individual Observations. *Journal of Quality Technology* 35, 367–376.

[11] Wold, S., Geladi, P., Esbensen, K. and Ohman, J. (1987). Multi-Way Principal Components and PLS Analysis. *Journal of Chemometrics* 1, 41–56.

Alternative based thresholding for pre-surgical fMRI

Joke Durnez *[1], **Beatrijs Moerkerke**[1], **Andreas Bartsch**[2] **and Thomas E. Nichols**[3]

[1] *Department of Data Analysis, Ghent University, Belgium*
[2] *Department of Neuroradiology, University of Heidelberg, Germany*
[3] *Department of Statistics & Warwick Manufacturing Group, University of Warwick, United Kingdom*

Abstract

Functional Magnetic Reasonance Imaging (fMRI) plays an important role in pre-surgical planning for patients with resectable brain lesions such as tumors. With appropriately designed tasks, the results of fMRI studies can guide resection, thereby preserving vital brain tissue.

The mass univariate approach to fMRI data analysis consists of performing a statistical test in each voxel, which is used to classify voxels either as active or inactive, i.e. related, or not, to the task of interest. In cognitive neuroscience, the focus is on controlling the rate of false positives while accounting for the severe multiple testing problem of searching the brain for activations. However, stringent control of false positives is accompanied by a risk of false negatives which can be detrimental, particularly in clinical settings where false negatives may lead to surgical resection of vital brain tissue. Consequently, for clinical applications we argue for a testing procedure with a stronger focus on preventing false negatives.

We present a thresholding procedure that incorporates information on false positives and false negatives. We combine 2 measures of significance for each voxel: a classical p-value which reflects evidence against the null hypothesis of no activation and an alternative p-value which reflects evidence against activation of a prespecified size. This results in a layered statistical map for the brain. One layer marks voxels exhibiting strong evidence against the traditional null hypothesis, while a second layer marks voxels where activation cannot be confidently excluded. The third layer marks voxels where the presence of activation can be rejected.

Keywords: fMRI, power, false negative errors, multiple testing, pre-surgical fMRI
AMS subject classifications: 62P07

1 Introduction

A common treatment for patients suffering from a brain tumor is surgical resection of the tumor. In order to minimize the risk of resecting brain tissue involved in essential brain functions, such as speech or language comprehension, these patients often undergo pre-surgical functional Magnetic Resonance Imaging (fMRI). This is a technique that shows subject-specific neural activity changes in the brain. The resulting fMRI data can assist the surgeon in performing the tumor resection while preserving the brain tissue involved in important cognitive and sensorimotor functions, and can even be used to predict the outcome of post-operative cognitive functioning [4].

For an fMRI data analysis, the brain is divided in more than 100,000 voxels. The mass univariate approach to fMRI data analysis consists of performing a statistical test in each voxel. In cognitive neuroscience, this technique is used to link neurological and neuropsychological functions with their respective location in the brain, supporting different theories of brain function. To be confident that a brain area is associated with a task it is essential to account for the multiple testing problem. However, multiple testing corrections result in a more stringent control of the null hypothesis of no activation, and consequently, the probability of a false

*Corresponding author, e-mail: Joke.Durnez@UGent.be

negative increases [5]. However, the scientific discipline generally deems stringent control of false positives necessary, accepting the concomitant sacrifices in sensitivity.
In a clinical setting such as pre-surgical fMRI however, a loss in power means true activation is not discovered, and this might result in the resection of vital brain tissue. Inversely, false positives have a less negative impact on the surgical result [3]. The goal of classical hypothesis testing is to prevent the null hypothesis from being rejected, by only considering voxels as being active when enough evidence against the null of no activation is found. This asymmetrical way of penalising errors in statistical inference is undesirable in this context [4], and instead the focus should be on protecting the alternative hypothesis: one only wants to exclude activation when enough evidence against activation is found. We therefore present a new hypothesis thresholding procedure that incorporates both information on false positives and false negatives and thus is ideally suited for pre-surgical fMRI.

2 Methods

2.1 Measures of evidence against the null and alternative

At each voxel i, $i = 1, \ldots, I$, we assume that a linear model is fit and produces $\widehat{\Delta}_i$, an unbiased estimate of the BOLD effect of interest Δ_i, and an estimate of the standard deviation of $\widehat{\Delta}_i$, its “standard error” SE($\widehat{\Delta}_i$). We henceforth suppress the voxel subscript unless needed for clarity.

The null and the alternative hypothesis The null hypothesis $H_0 : \Delta = 0$ states that the true effect magnitude is zero, and an underlying difference between conditions Δ is equal to 0. Classical statistical inference involves computing a test statistic, converted to a p-value, that measures the evidence against this null hypothesis. The decision procedure to reject H_0 is calibrated to maintain the Type I error at α. However, failing to reject H_0 does not allow one to conclude that H_0 is true.
Our procedure considers an “alternative hypothesis” p-value, p_1, that measures the evidence against $H_a : \Delta = \Delta_1$, the non-zero effect magnitude expected under activation. fMRI-studies are often preceeded with power analyses for sample size calculations which also require the specification of Δ_1. In literature, different approaches to choose a meaningful Δ_1 have been presented [1, 7]. Alternatively, in pre-surgical fMRI, one can estimate Δ_1 based on data in previous patients.

Measures of significance At a given voxel we have a test statistic T with observed value t, We assume that T has a known distribution under H_0 (e.g. Student’s t with given degrees-of-freedom, or Gaussian), so that we can compute the classical p-value

$$p_0 = P(T \geq t|H_0). \tag{1}$$

That is, p_0 quantifies the evidence against the null hypothesis H_0 of no task-related activation.
In a symmetrical fashion, the alternative p-value is defined as in Moerkerke et al. [6]:

$$p_1 = P(T \leq t|H_a). \tag{2}$$

Correspondingly, p_1 measures the evidence against H_a, and corresponds to the classical p-value for testing a “null” H_a versus “alternative” H_0. In generally, as the evidence in favor of H_a grows, p_0 becomes smaller and p_1 becomes larger.
In order to compute p_1 we need the distribution of T under H_a, which requires specification of Δ_1. However, we don’t expect a single magnitude of true activation, but a distribution of different true values [1]. Therefore, in a Bayesian spirit, we specify a distribution of likely values of Δ_1 instead of fixed value:

$$\Delta_1 \sim \mathcal{N}\left(\mu, \tau^2\right) \tag{3}$$

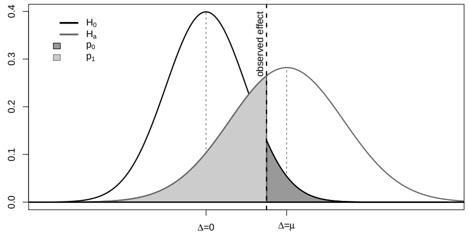


Figure 1: The distributions of an effect under H_0 and H_a are displayed for an observed effect of $t = 1.5$, SE($\widehat{\Delta}$) $= 1$, $\Delta_1 = 2$ and $\tau = 1$. Note that H_a has a wider distribution than H_0 due to the uncertainty on Δ_1.

where μ is the expected magnitude of effect under true activation while acknowledging variation among voxels, specifically Gaussian variation with standard deviation τ.
Assuming that T also follows a Gaussian distribution, it has the following distribution under H_a at voxel i:

$$T_i \sim \mathcal{N}\left(\frac{\mu}{\text{SE}(\widehat{\Delta}_i)}, \frac{\text{SE}(\widehat{\Delta}_i)^2 + \tau^2}{\text{SE}(\widehat{\Delta}_i)^2},\right) \tag{4}$$

where voxel subscripts are used to emphasize that the values of μ and τ are *fixed* for the entire brain, and based on prior knowledge or other experiments, while SE($\widehat{\Delta}_i$) is from each individual voxel. With this distribution we can compute p_1 at each voxel. An illustration of both measures of significance can be seen in Figure 1. As the alternative distribution depends the voxel-specific standard error, the distance between the null and alternative distributions will be voxel-specific. In particular a large standard error results in a large overlap between H_0 and H_a, while small standard errors lead to a large distance and little overlap between H_0 and H_a.

2.2 Combining measures of significance

In classical null hypothesis significance testing, a threshold α on p_0 can be translated into a threshold t_α for the test statistic t. In parallel, a threshold β on p_1 can be translated into a test statistic threshold t_β. While t_α is determined by α (and degrees-of-freedom if not using a Gaussian), t_β further depends on β, μ, τ *and* SE($\widehat{\Delta}_i$). Thus t_β varies over the brain depending on the (estimated) voxelspecific standard error.

3 Results

We consider data from a patient suffering from a left prefrontal brain tumor. The study design was a box-car design, where the patient was asked to alternate between recitation of tongue-twisters and quiescence. For the application to mass univariate linear modeling, the data were analyzed with FEAT in FSL 4.1 [8].
We derive the expected effect magnitude for Δ_1 and the variability of that effect τ from 5 patients who underwent the same fMRI paradigm. We threshold the image of each individual using an FDR-control at 0.05 and look at the average percent BOLD change units in each individual. We specify the expected effect magnitude for Δ_1 of $\mu = 0.73$ percent BOLD change units, and variability of that effect as $\tau = \sqrt{\widehat{\tau}^2} = 0.21$ percent BOLD change. These results are consistent with others in the literature (see e.g. Desmond and Glover, Figure 7A [1])
Results are shown in Figure 2 with thresholds $\alpha = 0.001$ and $\beta = 0.20$. In other words, we specified a p_0 threshold for declaring an activation when there is none at 1-in-1000; and we set the p_1 threshold

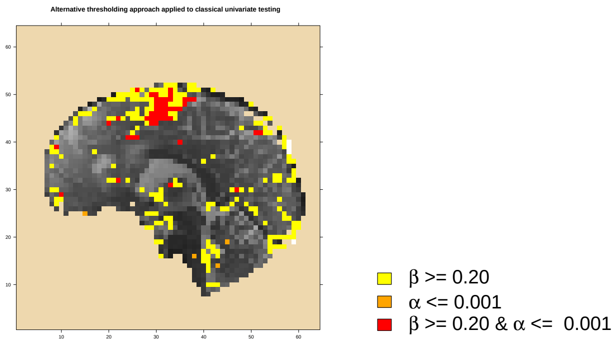

Alternative thresholding approach applied to classical univariate testing
60
50
40
30
20
10
10
20
30
40
50
60
🟨 β >= 0.20
🟧 α <= 0.001
🟥 β >= 0.20 & α <= 0.001

Figure 2: Sagittal slice of "layered" activation inference overlaying grayscale T2* reference image, threshold values of $\alpha = 0.001$ and $\beta = 0.20$. Red areas show areas of high confidence of activation (H_0 rejected, H_a not rejected), while yellow areas show areas where activation cannot be ruled out (neither H_0 nor H_a rejected); uncolored areas have high confidence of no activation (H_0 not rejected, H_a rejected), while the few orange voxels indicate voxels with significant but surprisingly small BOLD response magnitude (H_0 and H_a rejected).

for declaring the absence of activation when in fact the specified activation magnitude is present at 1-in-5. The red and the (scant) orange voxels show where H_0 can be confidently rejected, and, if presurgical planning was done only on the basis of classical null hypothesis testing, all other tissue would be regarded as "safe". Considering information on the alternative, we have the red voxels where, specifically, H_0 can be rejected and H_a cannot be rejected; i.e. the red voxels are incompatible with the null and compatible with the alternative, and thus are strong evidence for the effect. The yellow areas are areas are where neither H_0 *nor* H_a can be rejected; here the data is compatible with both the null and alternative, and suggest a lack of confidence in ruling out activation. Finally, for voxels with no coloration, the H_0 cannot be rejected but H_a can; the data are compatible with the null and incompatible with the alternative, and thus have good evidence for a lack of activation and suggest that these brain regions can be safely resected. This shows the key strength of the procedure: Among voxels traditionally classified as "nonactive", i.e. those with insufficiently small p_0's, it distinguishes between voxels where there is compelling evidence for non-activation (not colored) and those voxels where we cannot rule out the possibility of activation (yellow).
The orange voxels represent voxels for which the observed effect size is between the null hypothesis of no activation and the expected effect size. In these voxels, both the null and the alternative hypothesis are rejected which corresponds to very low residual noise in the GLM.

4 Discussion

Statistical thresholding in the context of multiple tests is generally driven by the need to limit false positives. These stringent testing procedures in fMRI research leads to an abundance of false negatives [5] and are therefore less useful in the context of pre-surgical fMRI where a false negative can have dire consequences. While many attempts have been made to propose more liberal testing criteria for example by controlling the FDR instead of the FWER [2], the focus is still on protecting the type I error rate. The unilateral focus on preventing false positives leads to a bias towards large obvious effects and against complex cognitive and affective effects [5]. We therefore propose a measure that quantifies the evidence against the alternative hypothesis as introduced in [6]. We use this quantity p_1 in addition with the classical p_0-value in a procedure that results in a thresholding procedure with multiple layers of significance. One layer consists of voxels exhibiting strong evidence of activation (red, in Fig. 2), while a another layer shows voxels with ambiguous

evidence (yellow and orange), and a final layer then consists of voxels for which the presence of activation can be confidently rejected (an absense of overlaid statistic values). Thereby we offer a more symmetrical interest towards both false positives and false negatives.
This procedure has been developed in light of pre-surgical fMRI, as false negatives can have harmful consequences for the patient. However the lack of power is omnipresent in fMRI-analyses [5] and therefore this procedure is also very useful in all branches of cognitive neuroscience. In this procedure, control of false positives remains possible but our procedure also takes into account information on the false negative rate. We do not assert that our method alleaviates all concerns with multiplicity, and one possible direction of future work is a multiplicity correction that adjusts both null and alternative hypothesis inferences for the number of tests.

Bibliography

[1] Desmond, J.E. and Glover, G.H. (2002). Estimating sample size in functional MRI (fMRI) neuroimaging studies: statistical power analyses. *Journal of neuroscience methods* 118(2), 115–28.

[2] Genovese, C. R., Lazar, N. and Nichols, T.E. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage* 15(4), 870–8.

[3] Gorgolewski, K.J., Storkey, A. J., Bastin, M.E. , Pernet, C.R. (2012). Adaptive thresholding for reliable topological inference in single subject fMRI analysis. *Frontiers in Human Neuroscience* 6(245), 1–14.

[4] Haller, S. and Bartsch, A.J. (2009). Pitfalls in fMRI. *European radiology* 19(11), 2689–706.

[5] Lieberman, M.D. and Cunningham, W. (2009). Type I and Type II error concerns in fMRI research: re-balancing the scale. *Social cognitive and affective neuroscience* 4(4), 423–8.

[6] Moerkerke, B., Goetghebeur, E., De Riek, J. and Roldan-Ruiz, I. (2006). Significance and impotence: towards a balanced view of the null and the alternative hypotheses in marker selection for plant breeding. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 169(1), 61–79.

[7] Mumford, J.A. and Nichols, T.E. (2008). Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation. *NeuroImage* 39(1), 261–8.

[8] Smith, S.M. , Jenkinson, M, Woolrich, M.W., Beckmann, C.F., Behrens, T.E.J., Johansen-Berg, H., Bannister, P.R., De Luca, M., Drobnjak, I, Flitney, D.E., Niazy, R.K., Saunders, J., Vickers,J., Zhang, Y., De Stefano, N., Brady, J.M. and Matthews, P.M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage* 23 (S1), 208–19

Adaptive Bayesian estimation in Gaussian sequence space models

Jan Johannes[1], **Rudolf Schenk**[*1] **and Anna Simoni**[2]

[1] *Université catholique de Louvain, Voie du Roman Pays 20, B-1348 Louvain-la-Neuve*
[2] *Université de Cergy-Pontoise, 33 boulevard du Port, F - 95011 Cergy-Pontoise*

Abstract We consider the nonparametric Bayesian estimation in a Gaussian sequence space model. The procedure is studied from a frequentist point of view, that is, we are interested in an optimal concentration rate of the posterior distribution shrinking to the distribution that generates the data. In a first step, we derive lower and upper bounds for the posterior concentration rates over a family of Gaussian prior distributions indexed by a tuning parameter. This result establishes posterior consistency, however the concentration rate depends on the parameter of interest and a tuning parameter. Under a suitable choice of the tuning parameter we derive a concentration rate uniformely over a class of parameters and show that this rate coincides with the minimax rate. As the choice of the tuning parameter depends on the considered class, we introduce in a second step a hierarchical prior and show that the resulting posterior concentration rate coincides in a direct sequence space model with the minimax rate and prove, furthermore, that the fully data-driven Bayes estimate is minimax-optimal.

Keywords: Bayesian methods
AMS subject classifications: 62C10

1 Introduction

Let ℓ_2 be the Hilbert space of square summable real valued sequences endowed with the usual inner product $\langle\cdot,\cdot\rangle_{\ell_2}$ and associated norm $\|\cdot\|_{\ell_2}$. In a Gaussian sequence space model we want to recover $\theta = (\theta_j)_{j\geq 1} \in \ell_2$ from a version that is blurred by Gaussian white noise. We adopt a Bayesian approach, where the conditional distribution of the observations given the parameter is Gaussian:

$$\mathrm{Y}_j \,|\, \boldsymbol{\vartheta}_j = \theta_j \sim \mathcal{N}\big(\lambda_j\theta_j, \epsilon\big), \quad \text{independent}, \quad j \in \mathbb{N}, \tag{1}$$

with sequence $(\lambda_j)_{j\geq 1}$, λ for short, and noise level $\epsilon > 0$. The sequence space model is called indirect if the sequence λ tends to zero. The particular case of a constant sequence λ is also called direct sequence space model. We will introduce a Gaussian prior distribution $P_{\boldsymbol{\vartheta}}$ of $\boldsymbol{\vartheta}$ having a well-known Gaussian posterior distribution $P_{\boldsymbol{\vartheta}\,|\,\mathrm{Y}}$. The objective is to derive its posterior concentration rate which are based on tail bounds for noncentral χ^2 distributions established in Birgé [2001]. To be more precise, we are seeking for a rate R_ϵ which is up to a constant a lower and an upper bound of the concentration rate of the posterior distribution $P_{\boldsymbol{\vartheta}\,|\,\mathrm{Y}}$, i.e.,

$$\lim_{\epsilon\to 0} \mathbb{E}_{\theta_o} P_{\boldsymbol{\vartheta}\,|\,\mathrm{Y}}(\underline{C}R_\epsilon \leq \|\boldsymbol{\vartheta} - \theta_o\|^2_{\ell_2} \leq \overline{C}R_\epsilon) = 1.$$

For a more detailed discussion see Barron et al. [1999], Castillo [2008] or Goshal et al. [2000]. This result establishes posterior consistency, however, the concentration rate depends on the parameter of interest and the choice of a tuning parameter. Under a suitable choice of the tuning parameter we derive a concentration rate uniformly over a class of parameters and show that this rate coincides with the minimax rate derived by Johannes and Schwarz [2013]. As the choice of the tuning parameter depends on the considered class, we introduce a hierarchical prior and show that the resulting posterior concentration rate coincides in a direct sequence space model with the minimax rate and prove, furthermore, that the fully data-driven Bayes estimate is minimax-optimal. The proofs are given in Johannes and Schenk [2013].

[*] Corresponding author, e-mail: rudolf.schenk@uclouvain.be

2 Basic model assumptions

We assume a Gaussian prior distribution for the parameter $\boldsymbol{\vartheta} = (\boldsymbol{\vartheta}_j)_{j\geq 1}$, that is $\{\boldsymbol{\vartheta}_j\}_{j\geq 1}$ are independent, normally distributed with prior means $(\theta_j^\times)_{j\geq 1}$ and prior variances $(\varsigma_j)_{j\geq 1}$:

$$\boldsymbol{\vartheta}_j \sim \mathcal{N}(\theta_j^\times, \varsigma_j), \quad \text{independent}, \quad j \in \mathbb{N}. \tag{2}$$

Standard calculus shows that the posterior distribution of $\boldsymbol{\vartheta}$ given $\mathrm{Y} = (\mathrm{Y}_j)_{j\geq 1}$ is Gaussian, that is $\{\boldsymbol{\vartheta}_j\}_{j\geq 1}$ are conditionally independent , normally distributed random variables given Y with posterior mean $\theta_j^{\mathrm{Y}} := \mathbb{E}[\boldsymbol{\vartheta}_j \,|\, \mathrm{Y}] = \frac{\varsigma_j^{-1}\theta_j^\times + \lambda_j \epsilon^{-1}\mathrm{Y}_j}{\lambda_j^2\epsilon^{-1} + \varsigma_j^{-1}}$ and posterior variance $\sigma_j^2 := \mathrm{Var}(\boldsymbol{\vartheta}_j \,|\, \mathrm{Y}) = (\lambda_j^2\epsilon^{-1} + \varsigma_j^{-1})^{-1}$, for all $j \in \mathbb{N}$. Moreover, a common Bayes estimate of the unknown parameter θ is the posterior mean $\mathbb{E}[\boldsymbol{\vartheta} \,|\, \mathrm{Y}]$. Taking this as a starting point, we construct a sequence of prior distributions: To be more precise, let us denote by δ_x the Dirac measure in the point x. Given $m \in \mathbb{N}$, we consider the independent random variables $\{\boldsymbol{\vartheta}_j^m\}_{j\geq 1}$ and their marginal distributions

$$\boldsymbol{\vartheta}_j^m \sim \mathcal{N}(\theta_j^\times, \varsigma_j),\ 1 \leq j \leq m \text{ and } \boldsymbol{\vartheta}_j^m \sim \delta_{\theta_j^\times},\ m < j, \text{ independent } j \in \mathbb{N} \tag{3}$$

resulting in the degenerate prior distribution $P_{\boldsymbol{\vartheta}^m}$.Consequently, $\{\boldsymbol{\vartheta}_j^m\}_{j\geq 1}$ are conditionally independent given Y and their posterior distribution is Gaussian with mean θ_j^{Y} and variance σ_j^2 for $1 \leq j \leq m$ while being degenerate on $\theta_j^\times$ for $j > m$. Hence, the Bayes estimate $\widehat{\theta}^m := \mathbb{E}[\boldsymbol{\vartheta}^m \,|\, \mathrm{Y}]$ is given for $j \geq 1$ by $\widehat{\theta}_j^m := \theta_j^{\mathrm{Y}}\mathbf{1}\{j \leq m\} + \theta_j^\times \mathbf{1}\{j > m\}$. The dimension parameter m plays the role of a tuning parameter. From a Bayesian point of view it is a hyperparameter and we will introduce now a prior distribution on the same which leads to a hierarchical prior distribution. In the following we consider a random parameter M taking its values in $\{1, \dots, G_\epsilon\}$ for some $G_\epsilon \in \mathbb{N}$ and prior distribution P_{M}. Both G_ϵ and P_{M} will be specified below. Now given M we consider the random variables $\{\mathrm{Y}_j\}_{j\geq 1}$ and $\left\{\boldsymbol{\vartheta}_j^{\mathrm{M}}\right\}_{j\geq 1}$ and their distributions are determined by

$$\mathrm{Y}_j = \lambda_j\, \boldsymbol{\vartheta}^{\mathrm{M}} + \sqrt{\epsilon}\zeta_j \quad \text{ and } \quad \boldsymbol{\vartheta}_j^{\mathrm{M}} = \theta_j^\times + \sqrt{\varsigma_j}\eta_j \mathbf{1}\{1 \leq j \leq \mathrm{M}\}$$

where $\{\zeta_j, \eta_j\}_{j\geq 1}$ are iid. standard normally distributed and independent of M. The Bayes estimate $\widehat{\theta} := \mathbb{E}[\boldsymbol{\vartheta}^{\mathrm{M}} \,|\, \mathrm{Y}]$ satisfies $\widehat{\theta}_j = \theta_j^\times$ for $j > G_\epsilon$ and for all $1 \leq j \leq G_\epsilon$

$$\widehat{\theta}_j = \theta_j^\times\, P(1 \leq \mathrm{M} \leq j-1 |\, \mathrm{Y}) + \theta_j^{\mathrm{Y}}\, P(j \leq \mathrm{M} \leq G_\epsilon |\, \mathrm{Y}).$$

3 Theoretical results

A major step towards establishing a concentration rate of the posterior distribution consists a finite sample bound for a fixed $m \in \mathbb{N}$. We express these bounds in terms of

$$\mathfrak{b}_m := \sum_{j>m}(\theta_{oj} - \theta_j^\times)^2, \quad \mathfrak{v}_m := \sum_{j=1}^m \sigma_j^2 = \sum_{j=1}^m \frac{1}{\lambda_j^2\epsilon^{-1} + \varsigma_j^{-1}};$$

$$\mathfrak{t}_m := \max_{1\leq j\leq m} \sigma_j^2 \quad \text{and} \quad \mathfrak{r}_m := \sum_{j=1}^m (\mathbb{E}_{\theta_o}[\theta_j^{\mathrm{Y}}] - \theta_{oj})^2 = \sum_{j=1}^m \frac{\varsigma_j^{-2}(\theta_j^\times - \theta_{oj})^2}{(\lambda_j^2\epsilon^{-1} + \varsigma_j^{-1})^2}.$$

Proposition 3.1. *For all $m \in \mathbb{N}$, for all $\epsilon > 0$ and for all $0 < c < 1/8$ we have*

$$\mathbb{E}_{\theta_o} P_{\boldsymbol{\vartheta}^m \,|\, \mathrm{Y}}(\|\boldsymbol{\vartheta}^m - \theta_o\|_{\ell_2}^2 > \mathfrak{b}_m + 3\mathfrak{v}_m + (3/2)\, m\, \mathfrak{t}_m + 4\mathfrak{r}_m) \leq 2\exp(-m/36);$$
$$\mathbb{E}_{\theta_o} P_{\boldsymbol{\vartheta}^m \,|\, \mathrm{Y}}(\|\boldsymbol{\vartheta}^m - \theta_o\|_{\ell_2}^2 < \mathfrak{b}_m + \mathfrak{v}_m - 4\, c\, (m\, \mathfrak{t}_m + \mathfrak{r}_m)) \leq 2\exp(-c^2 m).$$

The proof of the last result makes use of tail bounds for sums of independent squared Gaussian random variables. The next assertion presents a version which is due to Birgé [2001].

Lemma 3.1. *Let $\{X_j\}_{j \geq 1}$ be independent and normally distributed r.v. with mean $\alpha_j \in \mathbb{R}$ and standard deviation $\beta_j \geq 0$, $j \in \mathbb{N}$. For $m \in \mathbb{N}$ set $S_m := \sum_{j=1}^m X_j^2$ and consider $v_m \geq \sum_{j=1}^m \beta_j^2$, $t_m \geq \max_{1 \leq j \leq m} \beta_j^2$ and $r_m \geq \sum_{j=1}^m \alpha_j^2$. Then for all $c \geq 0$ we have*

$$\sup_{m \geq 1} e^{\frac{c(c \wedge 1)(v_m + 2r_m)}{4t_m}} P\big(S_m - \mathbb{E}S_m \leq -c(v_m + 2r_m)\big) \leq 1;$$
$$\sup_{m \geq 1} e^{\frac{c(c \wedge 1)(v_m + 2r_m)}{4t_m}} P\big(S_m - \mathbb{E}S_m \geq \frac{3c}{2}(v_m + 2r_m)\big) \leq 1.$$

The desired convergence of all the aforementioned sequences to zero necessitates to consider appropriate subsequences in dependence of ϵ, notably $(\mathfrak{v}_{m_\epsilon})_{m_\epsilon \geq 1}$, $(\mathfrak{t}_{m_\epsilon})_{m_\epsilon \geq 1}$ and $(\mathfrak{r}_{m_\epsilon})_{m_\epsilon \geq 1}$. To be more precise, we demande that the subsequences satisfy the following assumption.

Assumption A.1. *There exist constants $0 < \epsilon_o := \epsilon_o(\theta_o, \theta^\times, \varsigma) < 1$ and $0 < K := K(\theta_o, \theta^\times, \varsigma) < 1$ such that the prior distribution satisfies the condition $\sup_{0 < \epsilon < \epsilon_o} (\mathfrak{r}_{m_\epsilon} \vee \mathfrak{t}_{m_\epsilon})/(\mathfrak{b}_{m_\epsilon} \vee \mathfrak{v}_{m_\epsilon}) \leq K$.*

Corollary 3.1. *Under Assumption A.1 we have for all $0 < \epsilon < \epsilon_o$ and $0 < c < 1/(8K)$*

$$\mathbb{E}_{\theta_o} P_{\boldsymbol{\vartheta}^{m_\epsilon} \,|\, \mathrm{Y}}(\|\boldsymbol{\vartheta}^{m_\epsilon} - \theta_o\|_{\ell_2}^2 > (4 + (11/2)K)[\mathfrak{b}_{m_\epsilon} \vee \mathfrak{v}_{m_\epsilon}]) \leq 2\exp(-\frac{m_\epsilon}{36}); \tag{4}$$
$$\mathbb{E}_{\theta_o} P_{\boldsymbol{\vartheta}^{m_\epsilon} \,|\, \mathrm{Y}}(\|\boldsymbol{\vartheta}^{m_\epsilon} - \theta_o\|_{\ell_2}^2 < (1 - 8\, c\, K)[\mathfrak{b}_{m_\epsilon} \vee \mathfrak{v}_{m_\epsilon}]) \leq 2\exp(-c^2 m_\epsilon). \tag{5}$$

Thereby, assuming that $m_\epsilon := m(\epsilon)$ is chosen such that $\mathfrak{t}_{m_\epsilon} = o(\mathfrak{v}_{m_\epsilon})$ as $\epsilon \to 0$ implies the convergence to zero of the posterior probability. Furthermore, if we assume in addition that $\mathfrak{v}_{m_\epsilon} = o(1)$ and $m_\epsilon \to \infty$ as $\epsilon \to 0$ then we obtain by the dominated convergence theorem that also $\mathfrak{b}_{m_\epsilon} = o(1)$. Hence, $(\mathfrak{b}_{m_\epsilon} \vee \mathfrak{v}_{m_\epsilon})_{m_\epsilon \geq 1}$ converges to zero and is indeed a posterior concentration rate.

Theorem 3.1 (Posterior consistency)**.** *Under Assumption A.1 if $m_\epsilon \to \infty$ and $\mathfrak{v}_{m_\epsilon} = o(1)$ as $\epsilon \to 0$, then*

$$\lim_{\epsilon \to 0} \mathbb{E}_{\theta_o} P_{\boldsymbol{\vartheta}^{m_\epsilon} \,|\, \mathrm{Y}}((1 - 8cK)[\mathfrak{b}_{m_\epsilon} \vee \mathfrak{v}_{m_\epsilon}] \leq \|\boldsymbol{\vartheta}^{m_\epsilon} - \theta_o\|_{\ell_2}^2 \leq (4 + 11K/2)[\mathfrak{b}_{m_\epsilon} \vee \mathfrak{v}_{m_\epsilon}]) = 1.$$

Proposition 3.2 (Bayes estimate consistency)**.** *Let the assumptions of Theorem 3.1 be satisfied. Consider the Bayes estimate $\widehat{\theta}^{m_\epsilon} := \mathbb{E}[\boldsymbol{\vartheta}^{m_\epsilon} \,|\, \mathrm{Y}]$ then*

$$\mathbb{E}_{\theta_o} \|\widehat{\theta}^{m_\epsilon} - \theta_o\|_{\ell_2}^2 \leq (3 + K)[\mathfrak{b}_{m_\epsilon} \vee \mathfrak{v}_{m_\epsilon}]$$

and consequently $\mathbb{E}_{\theta_o} \|\widehat{\theta}^{m_\epsilon} - \theta_o\|_{\ell_2}^2 = o(1)$ as $\epsilon \to 0$.

The last assertion shows that $(\mathfrak{b}_{m_\epsilon} \vee \mathfrak{v}_{m_\epsilon})_{m_\epsilon \geq 1}$ is up to a constant a lower and upper bound of the concentration rate. The result, however, is obtained under Assumption A.1 which depends on the particular choice of the prior distribution. We suppose that the prior distribution, and more precisely, the prior variances are chosen such that the following assumption holds.

Assumption A.2. *Define $\Lambda_j := \lambda_j^{-2}$, $j \geq 1$, $\Lambda_{(m)} := \max_{1 \leq j \leq m} \Lambda_j$ and $\overline{\Lambda}_m := m^{-1} \sum_{j=1}^m \Lambda_j$, $m \geq 1$. There exists a constant d such that $\varsigma_j \geq d\Lambda_j$ for all $j \geq 1$.*

Corollary 3.2. *Under Assumption A.2, let $m_\epsilon = m(\epsilon)$ be chosen such that $m_\epsilon \to \infty$ and $\epsilon m_\epsilon \overline{\Lambda}_{m_\epsilon} = o(1)$ as $\epsilon \to 0$, and suppose in addition*

$$\limsup_{\epsilon \to 0} \Lambda_{(m_\epsilon)} \big\{ \mathfrak{b}_{m_\epsilon} (\epsilon\, m_\epsilon)^{-1} \vee \overline{\Lambda}_{m_\epsilon} \big\}^{-1} < \infty \tag{6}$$

then there exist a constant K such that

$$\lim_{\epsilon \to 0} \mathbb{E}_{\theta_o} P_{\boldsymbol{\vartheta}^{m_\epsilon} \,|\, \mathrm{Y}}(K^{-1}[\mathfrak{b}_{m_\epsilon} \vee \epsilon m_\epsilon \overline{\Lambda}_{m_\epsilon}] \leq \|\boldsymbol{\vartheta}^{m_\epsilon} - \theta_o\|_{\ell_2}^2 \leq K[\mathfrak{b}_{m_\epsilon} \vee \epsilon m_\epsilon \overline{\Lambda}_{m_\epsilon}]) = 1.$$

Under the conditions of the last assertion, the sequence $(\mathfrak{b}_{m_\epsilon} \vee \epsilon m_\epsilon \overline{\Lambda}_{m_\epsilon})_{m_\epsilon \geq 1}$ provides up to constants a lower and upper bound for the concentration rate. The result implies consistency but it does not answer the question of optimality in a satisfactory way. Observe that the rate depends on the parameter of interest θ and we could optimize the rate for each θ separately, but we are rather interested in a uniform rate over a class of parameters. Given a strictly positive sequence $\mathfrak{a} = \big(\mathfrak{a}_j\big)_{j \geq 1}$ consider for $\theta \in \ell_2$ its weighted norm $\|\theta\|_{\mathfrak{a}}^2 := \sum_{j \geq 1} \mathfrak{a}_j \theta_j^2$. We define $\ell_2^{\mathfrak{a}}$ as the completion of ℓ_2 with respect to $\|\cdot\|_{\mathfrak{a}}$. We assume in the following that the parameter θ_o belongs to the ellipsoid $\Theta_{\mathfrak{a}}^r := \big\{\theta \in \ell_2^{\mathfrak{a}} : \|\theta - \theta^\times\|_{\mathfrak{a}}^2 \leq r\big\}$. Define for all $\epsilon > 0$

$$m_\epsilon^\star := m_\epsilon^\star(\mathfrak{a}, \lambda) := \arg\min_{m \geq 1} (\mathfrak{a}_{m+1}^{-1} \vee \epsilon\, m\, \overline{\Lambda}_m) \quad \text{and}$$

$$\mathcal{R}_\epsilon^\star := \mathcal{R}_\epsilon^\star\big[\mathfrak{a}, \lambda\big] := (\mathfrak{a}_{m_\epsilon^\star+1}^{-1} \vee \epsilon\, m_\epsilon^\star\, \overline{\Lambda}_{m_\epsilon^\star}).$$

Theorem 3.2 (Optimal posterior concentration rate)**.** *Under Assumption A.2, suppose in addition that $m_\epsilon^\star$ satisfies (6) then there exists a constant $K := K(\Theta_{\mathfrak{a}}^r, \lambda)$ such that*

$$\lim_{\epsilon \to 0} \inf_{\theta_o \in \Theta_{\mathfrak{a}}^r} \mathbb{E}_{\theta_o} P_{\boldsymbol{\vartheta}^{m_\epsilon^\star} \,|\, \mathrm{Y}}(\|\boldsymbol{\vartheta}^{m_\epsilon^\star} - \theta_o\|_{\ell_2}^2 \leq K \mathcal{R}_\epsilon^\star) = 1$$

moreover, if $\Psi_\epsilon / \mathcal{R}_\epsilon^\star = o(1)$ as $\epsilon \to 0$ then

$$\lim_{\epsilon \to 0} \sup_{\theta_o \in \Theta_{\mathfrak{a}}^r} \mathbb{E}_{\theta_o} P_{\boldsymbol{\vartheta}^{m_\epsilon^\star} \,|\, \mathrm{Y}}(\|\boldsymbol{\vartheta}^{m_\epsilon^\star} - \theta_o\|_{\ell_2}^2 \leq \Psi_\epsilon) = 0.$$

It is interesting to note that the rate $\mathcal{R}_\epsilon^\star = \mathcal{R}_\epsilon^\star\big[\Theta_{\mathfrak{a}}^r, \lambda\big]$ is optimal in a minimax sense. To be more precise, given an estimator $\widehat{\theta}$ of θ let $\sup_{\theta \in \Theta_{\mathfrak{a}}^r} \mathbb{E}_\theta \|\widehat{\theta} - \theta\|^2$ denote the maximal mean integrated squared error (MISE) over the class $\Theta_{\mathfrak{a}}^r$. It has been shown in Johannes and Schwarz [2013] that $\mathcal{R}_\epsilon^\star$ provides up to a constant a lower bound for the maximal MISE over the class $\Theta_{\mathfrak{a}}^r$ and that there exists an estimator attaining this rate. The next assertion establishes the minimax optimality of the Bayes estimate.

Proposition 3.3 (Minimax-optimal Bayes estimate)**.** *Let the assumptions of Theorem 4.1 be satisfied and $\widehat{\theta}^{m_\epsilon^\star} := \mathbb{E}[\boldsymbol{\vartheta}^{m_\epsilon^\star} \,|\, \mathrm{Y}]$ then there exists a constant $K := K(\Theta_{\mathfrak{a}}^r, \lambda)$ such that*

$$\sup_{\theta_o \in \Theta_{\mathfrak{a}}^r} \mathbb{E}_{\theta_o} \|\widehat{\theta}^{m_\epsilon^\star} - \theta_o\|_{\ell_2}^2 \leq K \mathcal{R}_\epsilon^\star.$$

4 Adaptivity in the direct sequence space model

We will derive a concentration rate given the aforementioned hierarchical prior distribution in a direct sequence space model, that is $\lambda_j = 1$, $j \geq 1$. For this purpose set $G_\epsilon := \lfloor \epsilon^{-1} \rfloor$ and

$$p_{\mathrm{M}}(m) = \frac{\exp(\frac{-m}{\epsilon}) \prod_{j=1}^m (1 + \varsigma_j \epsilon^{-1})^{1/2}}{\sum_{m'=1}^{G_\epsilon} \exp(\frac{-m'}{\epsilon}) \prod_{j=1}^{m'} (1 + \varsigma_j \epsilon^{-1})^{1/2}} \quad \text{for } 1 \leq m \leq G_\epsilon\,. \tag{7}$$

Theorem 4.1 (Optimal posterior concentration rate)**.** *Under Assumption A.2 suppose in addition that* $m^\star_\epsilon$ *satisfies* (6) *then there exists a constant* $K := K(\Theta^r_{\mathfrak{a}}, \lambda)$ *such that*

$$\lim_{\epsilon \to 0} \inf_{\theta_o \in \Theta^r_{\mathfrak{a}}} \mathbb{E}_{\theta_o} P_{\boldsymbol{\vartheta}^{\mathrm{M}} \,|\, \mathrm{Y}}(\|\boldsymbol{\vartheta}^{\mathrm{M}} - \theta_o\|^2_{\ell_2} \leq K \mathcal{R}^\star_\epsilon) = 1.$$

moreover, if $\Psi_\epsilon / \mathcal{R}^\star_\epsilon = o(1)$ *as* $\epsilon \to 0$ *then*

$$\lim_{\epsilon \to 0} \sup_{\theta_o \in \Theta^r_{\mathfrak{a}}} \mathbb{E}_{\theta_o} P_{\boldsymbol{\vartheta}^{\mathrm{M}} \,|\, \mathrm{Y}}(\|\boldsymbol{\vartheta}^{\mathrm{M}} - \theta_o\|^2_{\ell_2} \leq \Psi_\epsilon) = 0.$$

We shall emphasize that the concentration rate derived from the hierarchical prior coincides with the minimax optimal rate $\mathcal{R}^\star_\epsilon = \mathcal{R}^\star_\epsilon\big[\Theta^r_{\mathfrak{a}}, \lambda\big]$ of the maximal MISE over the class $\Theta^r_{\mathfrak{a}}$. In particular this prior does not involve any knowledge of the class $\Theta^r_{\mathfrak{a}}$, therefore, the corresponding Bayes estimate is fully-data driven. The next assertion establishes its minimax-optimality.

Proposition 4.1 (Minimax-optimal Bayes estimate)**.** *Under the assumptions of Theorem 4.1. Consider the Bayes estimate* $\widehat{\theta} := \mathbb{E}[\boldsymbol{\vartheta}^{\mathrm{M}} \,|\, \mathrm{Y}]$ *then there exists a constant* $K := K(\Theta^r_{\mathfrak{a}}, \lambda)$ *such that* $\sup_{\theta_o \in \Theta^r_{\mathfrak{a}}} \mathbb{E}_{\theta_o} \|\widehat{\theta} - \theta_o\|^2_{\ell_2} \leq K \mathcal{R}^\star_\epsilon$ *for all* $\epsilon > 0$.

Conclusions and perspectives. In this paper we have presented a hierarchical prior leading to a fully-data Bayes estimate that is minimax-optimal in a direct sequence space model. Obviously, the concentration rate based on a hierarchical prior in an indirect sequence space model possibly with additional noise in the eigenvalues is only one amongst the many interesting questions for further research and we are currently exploring this topic.

Acknowledgements: This work was supported by the IAP research network no. P7/06 of the Belgian Government (Belgian Science Policy) and by the “Fonds Spéciaux de Recherche” from the Université catholique de Louvain.

Bibliography

[1] A. Barron, M. J. Schervish, and L. Wasserman (1999), The consistency of posterior distributions in nonparametric problems. Annals of Statistics 27, 2, 536–561.

[2] L. Birgé (2001), An alternative point of view on Lepski’s method, *IMS Lecture Notes*, State of the art in probability and statistics, Leiden 1999, 36, 113–133.

[3] I. Castillo (2008), Lower bounds for posterior rates with Gaussian process priors. Electronic Journal of Statistics, 2, 1281-1299.

[4] S. Ghosal, J. K. Ghosh and A. W. van der Vaart (2000), Convergence rates of posterior distributions. Annals of Statistics 28, 2 , 500–531.

[5] J. Johannes and R. Schenk (2013), Adaptive Bayesian estimation in Gaussian sequence space models, Discussion paper at Université catholique de Louvain.

[6] J. Johannes and M. Schwarz (2013), Adaptive Gaussian inverse regression with partially unknown operator, To appear in Communications in Statistics - Theory and Methods.

Skewed sub-Gaussian multivariate distribution

Teodosi Geninski *[1], Ivan Mitov[2] and Zari Rachev[3]

[1] *Faculty of Mathematics and Informatics, Sofia University, Bulgaria*
[2] *FinAnalytica Inc.*
[3] *Stony Brook University*

Abstract

Normal variance mixture models are used as an extension of the Gaussian framework to allow heavier tails and add flexibility to the Wiener processes' time concept. The Sub-Gaussian model is a typical representative of this class. It is a parametric sub-class of the multivariate α-stable distribution which is an elliptical, infinitely divisible and has a tractable representation of its characteristic function. It possesses heavy tails but it is also a symmetric distribution.

To overcome the latter drawback a ρ-weighted, univariate, α-stable skewness component is introduced. The domain of ρ and its connection to the skewness and the dependence structure are explored as well as some of the border cases. By varying ρ from 0 to 1 the distribution transforms from a regular Sub-Gaussian to multivariate α-stable with independent and not necessary symmetric components.

Keywords: Variance mixture, Multivariate stable models, Sub-Gaussian model, Asymmetric distributions
AMS subject classifications: 60E07, 62P05, 62E17

1 Introduction

A particular parametric subclass of the multivariate α-stable distributions is the class of $\alpha-$stable sub-Gaussian distributions. In this report we introduce a distribution based on the multivariate $\alpha-$stable sub-Gaussian distribution. All the marginal distributions within our model are $\alpha-$stable however not symmetric since different skewness parameters are allowed. This is extremely important extension because there is a significant empirical evidence that many real world observable variables, e.g. the financial asset returns, are not symmetric ([2], [5]). In the next two sections we give the definitions of the $\alpha-$stable distributions and the multivariate $\alpha-$stable sub-Gaussian distributions. We provide without proofs some important properties which are used later in the paper. Section 3 defines our multivariate distribution and investigates its key properties and in Section 4 we discuss the model estimation methods and scenarios generation. In Section 5 we use the skewed sub-Gaussian distribution to model the dependence between US stock index and large cap US stock. The last section summarizes the results and concludes the findings.

2 $\alpha-$stable Distributions

2.1 Univariate and multivariate $\alpha-$stable distributions

The class of $\alpha-$stable distributions arises from the generalization of the central limit theorem. The stable distributions are the only possible weak limits of properly normalized sums of independent identically distributed (i.i.d.) random variables. The normal distribution is a special case. They possess *domains of*

[*] Corresponding author, e-mail: teodosi.g@gmail.com

attraction; that is, a sum of i.i.d. random variables has properties close to the properties of the limit distribution and we can adopt the limit distribution as an approximate model. The domains of attraction property is very desirable and it is not possessed by any other distribution for the summation scheme. The most natural definition of a stable random vector is the following.

Definition 2.1. *A random vector $\mathbf{X} = (X_1, \ldots, X_d)^T$ is said to be $\alpha-$stable random vector in $\mathbf{R}^d$, $\alpha \in (0, 2]$, if for any positive numbers A and B there is a vector $\mathbf{d} \in \mathbf{R}^d$ such that*

$$A\mathbf{X}^{(1)} + B\mathbf{X}^{(2)} \overset{d}{=} (A^\alpha + B^\alpha)^{1/\alpha}\mathbf{X} + \mathbf{d} \tag{1}$$

where $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ are independent copies of $\mathbf{X}$. The random vector $\mathbf{X}$ is called strictly stable if $\mathbf{d} = 0$, and is said to be symmetric stable if $\mathbf{P}(\mathbf{X} \in U) = \mathbf{P}(-\mathbf{X} \in U)$ for any Borel set $U \in \mathbf{R}^d$.

This definition extends to n i.i.d. copies of $\mathbf{X}$ for each $n \in \mathbb{N}$ which justifies the term 'stable' because the sum of i.i.d. random variables has the same distribution as X up to a scale and shift parameter.
Another equivalent way to define $\alpha-$stable random vector is through its characteristic function.

Definition 2.2. *A random vector $\mathbf{X} = (X_1, \ldots, X_d)^T$ is said to be $\alpha-$stable random vector in $\mathbf{R}^d$, $\alpha \in (0, 2]$, if there is a finite measure Γ on the unit sphere $\mathcal{S}^d$ and a vector $\mu \in \mathbf{R}^d$ such that the characteristic function $\Phi_\alpha(\mathbf{u}) := E(e^{i\mathbf{u}^T\mathbf{X}})$ has the following form:*

(a) For $\alpha \neq 1$,

$$\Phi_\alpha(\mathbf{u}) = \exp\left\{-\int_{\mathcal{S}_d} |(\mathbf{u}^T\mathbf{s})^\alpha| \left(1 - i\,\mathrm{sign}(\mathbf{u}^T\mathbf{s}) \tan\frac{\pi\alpha}{2}\right)\Gamma(d\mathbf{s}) + i\mathbf{u}^T\boldsymbol{\mu}\right\};$$

(b) For $\alpha = 1$,

$$\Phi_\alpha(\mathbf{u}) = \exp\left\{-\int_{\mathcal{S}_d} |\mathbf{u}^T\mathbf{s}| \left(1 + i\frac{2}{\pi}\mathrm{sign}(\mathbf{u}^T\mathbf{s}) \ln(\mathbf{u}^T\mathbf{s})\right)\Gamma(d\mathbf{s}) + i\mathbf{u}^T\boldsymbol{\mu}\right\}.$$

The pair $(\Gamma, \boldsymbol{\mu})$ is unique and is called spectral decomposition. The measure Γ is called spectral measure of the stable random vector. The distribution of $\mathbf{X}$ is denoted by $S_\alpha(\Gamma, \boldsymbol{\mu})$.

In the symmetric case equations (a) and (b) become the following one

(a') For $0 < \alpha \neq 2$,

$$\Phi_\alpha(\mathbf{u}) = \exp\left\{-\int_{\mathcal{S}_d} |(\mathbf{u}^T\mathbf{s})^\alpha|\Gamma(d\mathbf{s}) + i\mathbf{u}^T\boldsymbol{\mu}\right\}.$$

The symmetric $\alpha-$stable distributions are usually denoted by $S\alpha S(\Gamma, \boldsymbol{\mu})$.
Next, we give three important properties of the one-dimensional stable distributions which are used further in the paper.

Property 1. *If $X_i \sim S_\alpha(\sigma_i, \beta_i, \mu_i)$, $i = 1, \ldots, n$ are i.i.d. rv's then*

$$S = \sum_{i=1}^{n} X_i \sim S_\alpha(\sigma, \beta, \mu)$$

where

$$\mu = \sum_{i=1}^{n} \mu_i, \quad \sigma = (\sigma_1^\alpha + \ldots + \sigma_n^\alpha)^{1/\alpha}, \quad \beta = \frac{\beta_1\sigma_1^\alpha + \ldots \beta_n\sigma_n^\alpha}{\sigma_1^\alpha + \ldots + \sigma_n^\alpha}.$$

Property 2. *If $X \sim S_\alpha(\sigma, \beta, \mu)$ then $sX \sim S_\alpha(|s|\sigma, sign(s)\beta, s\mu)$ and $m + X \sim S_\alpha(\sigma, \beta, m + \mu)$.*

Property 3. *Let $Z \sim S_{\alpha'}(\sigma, 0, 0)$ and let $0 < \alpha < \alpha'$. Let Y be an $\alpha/\alpha'-$stable random variable, totally skewed to the right*

$$Y \sim S_{\alpha/\alpha'}\left(\left(\cos\frac{\pi\alpha}{2\alpha'}\right)^{\alpha'/\alpha}, 1, 0\right)$$

and assume that Z and Y are independent. Then

$$X = Y^{1/\alpha'} Z \sim S_\alpha(\sigma, 0, 0).$$

This property implies that if Z is a zero mean Gaussian random variable and if Y is a positive $\alpha/2-$stable random variable independent of X, then

$$X = Y^{1/2} Z$$

is symmetric $\alpha-$stable. This property implies that every symmetric $\alpha-$stable random variable is conditionally Gaussian.

2.2 $\alpha-$stable sub-Gaussian distributions

An important subset of the $\alpha-$stable distributions is the class of sub-Gaussian distributions. They are a special case of symmetric α-stable distributions, but their spectral measure is always discrete and this allows us to have a tractable expression for the characteristic function. Property 3 plays a crucial role in the investigation of the sub-Gaussian distributions.

Definition 2.3. *Let $\mathbf{Z} \sim N(0, I_d)$ be a standard normal random vector in $\mathbf{R}^d$. Let A be an $d \times d$ matrix, $\mu \in \mathbf{R}^d$, and $Y \sim S_{\alpha/2}\left(\left(\cos\frac{\pi\alpha}{4}\right)^{2/\alpha}, 1, 0\right)$. Then the random vector $\mathbf{X}$ defined by*

$$\mathbf{X} = \boldsymbol{\mu} + \sqrt{Y} A\mathbf{Z} \tag{2}$$

is called $\alpha-$stable sub-Gaussian random vector.

The matrix $\Sigma = AA^T$ is called dispersion matrix of the sub-Gaussian distribution. Equation (2) is equivalent to

$$\mathbf{X} = \boldsymbol{\mu} + \sqrt{Y}\mathbf{U} \tag{3}$$

where $\mathbf{U} \sim N(0, \Sigma)$. The sub-Gaussian random vector $\mathbf{X}$ inherits its dependence from the underlying normal random vector $\mathbf{U}$. Further in the paper we denote this class of distributions by $S_\alpha^{SG}(\Sigma, \boldsymbol{\mu})$. It is a special case of the so called normal mean-variance mixtures.

Property 4. *Every sub-Gaussian random vector $\mathbf{X}$ defined as in Definition 2.3 has stable marginals with parameters $(\alpha, 0, \sigma_i/\sqrt{2}, \mu_i)$, i.e. $X_i \sim S_\alpha(\sigma_i/\sqrt{2}, 0, \mu_i)$, for $i = 1, \ldots, d$, where σ_i^2 are the diagonal elements of the dispersion matrix Σ.*

Proof. The property is a direct consequence from Property 3 for $\alpha' = 2$. Note that by the definition of the stable distribution we have $S_2(\sigma, 0, 0)$ is Gaussian distribution with standard deviation $\sigma\sqrt{2}$. Applying Property 2 for the constant term $\boldsymbol{\mu}$ concludes the proof. □

Definition 2.4. *The random vector $\mathbf{X}$ is said to have a (multivariate) normal mean-variance mixture distribution if*

$$\mathbf{X} \overset{d}{=} \boldsymbol{\mu} + Y\boldsymbol{\gamma} + \sqrt{\mathbf{Y}} A\mathbf{Z}$$

where

(i) $\mathbf{Z} \sim N(0, I_k)$;

(ii) $Y \geq 0$ *is a non-negative, scalar-valued rv which is independent of* $\mathbf{Z}$*;*

(iii) $A \in \mathbf{R}^{d \times k}$ *is a matrix; and*

(iv) $\boldsymbol{\mu}$ *and* $\boldsymbol{\gamma}$ *are parameter vectors in* $\mathbf{R}^d$.

In this case we have that

$$\mathbf{X}|Y = y \sim N(\boldsymbol{\mu} + y\boldsymbol{\gamma}, y\Sigma)$$

where $\Sigma = AA^T$ and it is clear why such distributions are known as mean-variance mixtures of normals. The characteristic function of $\mathbf{X}$ is given by

$$\Phi_{\mathbf{X}}(\mathbf{u}) = e^{i\mathbf{u}^T\boldsymbol{\mu}}\hat{H}\left(\frac{1}{2}\mathbf{u}^T\Sigma\mathbf{u} - i\mathbf{u}^T\boldsymbol{\gamma}\right), \tag{4}$$

where $\hat{H}(s) = \mathbf{E}e^{-sY}$ is the Laplace-Stieltjes transform of the mixing rv Y, which is also called subordinator.

In our particular case $Y \sim S_{\alpha/2}\left(\left(\cos\frac{\pi\alpha}{4}\right)^{2/\alpha}, 1, 0\right)$, and $\hat{H}(s) = e^{-s^{\alpha/2}}$. In this way using (4) with the particular form of $\hat{H}$ we obtain the characteristic function of the $\alpha-$stable sub-Gaussian distribution formulated in the following proposition.

Proposition 2.1. *The characteristic function of the* $\alpha-$*stable sub-Gaussian random vector* $\mathbf{X} \sim S_\alpha^{SG}(\Sigma, \boldsymbol{\mu})$*, defined by (2), has the form*

$$\Phi_\alpha^{SG}(\mathbf{u}) = e^{i\mathbf{u}^T\boldsymbol{\mu}}e^{-\left(\frac{1}{2}\mathbf{u}^T\Sigma\mathbf{u}\right)^{\alpha/2}}. \tag{5}$$

For α-stable sub-Gaussian random vectors, we do not need the spectral measure Γ in the characteristic function. This fact simplifies the fit and the simulation of such distributions. The $\alpha-$stable sub-Gaussian distributions are a special subclass of the multivariate symmetric stable distributions, and therefore they are elliptical distributions. It is well known that the elliptical distributions do not allow for modeling different lower and upper tail dependence. Therefore, in the next section we define a modification of the classical $\alpha-$stable sub-Gaussian distribution allowing for asymmetry.

3 Skewed sub-Gaussian Distributions

Definition 3.1. *Let* $\mathbf{Z} \sim N(0, I_d)$ *be a standard normal random vector in* $\mathbf{R}^d$*. Let* A *be a* $d \times d$ *matrix,* $\boldsymbol{\mu} \in \mathbf{R}^d$*, and* $Y \sim S_{\alpha/2}\left(\left(\cos\frac{\pi\alpha}{4}\right)^{2/\alpha}, 1, 0\right)$ *and let* $\rho \in (0, 1)$*. For each* $i = 1, \ldots, d$ *define the random variable* $W_i \sim S_\alpha(\rho^{1/\alpha}, \rho^{-1}\beta_i, 0)$ *independent of* Y *and* $W_j, j \neq i$*, where* $\beta_i = \beta_i(\rho) \in (-\rho, \rho)$*. Then the random vector* $\mathbf{X}$ *defined by*

$$\mathbf{X} = \boldsymbol{\mu} + \mathbf{W} + (1-\rho)^{1/\alpha}\sqrt{Y}A\mathbf{Z}, \tag{6}$$

where $\mathbf{W} = \frac{1}{\sqrt{2}}(W_1\sigma_1, \ldots, W_d\sigma_d)^T$ *and* σ_i^2 *is the* i*-th diagonal element of* $\Sigma = AA^T$*, is called skewed sub-Gaussian random vector with parameters* $(\Sigma, \boldsymbol{\beta}, \boldsymbol{\mu}, \rho)$ *with skewness parameter* $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^T$.

The family of the skewed sub-Gaussian distributions will be denoted by $S_\alpha^{SSG}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\mu}, \rho)$, i.e. we write $\mathbf{X} \sim S_\alpha^{SSG}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\mu}, \rho)$ for the random vector $\mathbf{X}$ defined by (6).

Property 5. *Every slewed sub-Gaussian random vector* $\mathbf{X}$ *defined as in Definition 3.1 has asymmetric stable marginals with parameters* $(\alpha, \beta_i, \sigma_i/\sqrt{2}, \mu_i)$*, i.e.* $X_i \sim S_\alpha(\sigma_i/\sqrt{2}, \beta_i, \mu_i)$*, for* $i = 1, \ldots, d$.

Proof. From (6) for each $i = 1, \ldots, d$ f we have

$$\mathbf{X} = \boldsymbol{\mu} + \mathbf{W} + (1-\rho)^{1/\alpha} \sqrt{Y} A\mathbf{Z} = \boldsymbol{\mu} + \mathbf{W} + (1-\rho)^{1/\alpha} \mathbf{C}, \tag{7}$$

where $\mathbf{C}$ is sub-Gaussian vector. By applying Property 4 for the marginals of $\mathbf{C}$ and by the independence of $\mathbf{W}$, Y and $\mathbf{Z}$ we obtain

$$\begin{aligned} X_i = \mu_i + \frac{1}{\sqrt{2}} W_i \sigma_i + (1-\rho)^{1/\alpha} C_i \sim \mu_i + \frac{\sigma_i}{\sqrt{2}} S_\alpha \left(\rho^{1/\alpha}, \rho^{-1}\beta_i, 0 \right) + \\ + (1-\rho)^{1/\alpha} S_\alpha \left(\frac{\sigma_i}{\sqrt{2}}, 0, 0 \right) \end{aligned} \tag{8}$$

Now from Property 1 and Property 2 we have

$$\begin{aligned} X_i \sim \mu_i + S_\alpha \left(\frac{\sigma_i}{\sqrt{2}} \rho^{1/\alpha}, \rho^{-1}\beta_i, 0 \right) + S_\alpha \left(\frac{\sigma_i}{\sqrt{2}} (1-\rho)^{1/\alpha}, 0, 0 \right) \\ \sim S_\alpha \left(\frac{\sigma_i}{\sqrt{2}}, \beta_i, \mu_i \right) \end{aligned} \tag{9}$$

□

The skewed sub-Gaussian multivariate distribution depends on a new scalar parameters $\rho \in (0,1)$ and $\beta_i \in (-\rho, \rho)$. This restriction shows that the skewness of all the marginals is controlled by a single parameter $\rho \in (0,1)$. The distribution is not anymore a member of the elliptical class. It is characterized by the following theorem.

Theorem 3.1. *Let* $\mathbf{X} \sim S_\alpha^{SSG}(\Sigma, \boldsymbol{\beta}, \boldsymbol{\mu}, \rho)$ *be a skewed sub-Gaussian random vector. Then the characteristic function of* $\mathbf{X}$ *is given by*

$$\begin{aligned} \Phi_\alpha^{SSG}(\mathbf{u}) = \quad & \exp \Big\{ i\mathbf{u}^T \boldsymbol{\mu} - \tfrac{1}{2}(1-\rho) \left(\mathbf{u}^T \Sigma \mathbf{u} \right)^{\alpha/2} - \\ & -\rho\, 2^{-\alpha/2} \textstyle\sum_{j=1}^d \left(|u_j|^\alpha \sigma_j^\alpha \left(1 - i\frac{\beta_j}{\rho} \mathrm{sign}(u_j) \mathrm{C}(u_j, \alpha) \right) \right) \Big\}, \end{aligned} \tag{10}$$

where $\mathrm{C}(u,\alpha) = \tan\frac{\pi\alpha}{2}$ *for* $\alpha \neq 1$, *and* $\mathrm{C}(u,1) = -\frac{2}{\pi}\ln|u|$. *Moreover,*

(1) $\mathbf{X} \overset{d}{\to} \mathbf{X}^0 \sim S_\alpha^{SG}(\Sigma, \boldsymbol{\mu})$, *as* $\rho \to 0$;

(2) $\mathbf{X} \overset{d}{\to} \mathbf{X}^1$, *which is a multivariate* $\alpha-$*stable vector with independent components as* $\rho \to 1$.

Proof. We calculate the respective limits in the characteristic function (10) when $\rho \to 0$ and $\rho \to 1$. When $\rho \to 0$ we use the fact that $\beta_i/\rho < 1$ (since β_i depends on ρ) in order to obtain $\Phi_\alpha^{SSG}(\mathbf{u}) \to \Phi_\alpha^{SG}(\mathbf{u})$, where $\Phi_\alpha^{SG}(\mathbf{u})$ is defined by (5) in Proposition 2.1.
When $\rho \to 1$ we have

$$\Phi_\alpha^{SSG}(\mathbf{u}) \to \exp\left\{ i\mathbf{u}^T \mu - 2^{-\alpha/2} \sum_{j=1}^d \left(|u_j|^\alpha \sigma_j^\alpha \left(1 - i\frac{\beta_j}{\rho} \mathrm{sign}(u_j) \mathrm{C}(u_j, \alpha) \right) \right) \right\},$$

which is the characteristic function of $\mathbf{W}$, i.e. an $\alpha-$stable vector of independent components. This completes the proof.

□

	α (tail)	β (skew)	σ (scale)	μ (location)	S&P 500	Microsoft
S&P 500	1.65	-0.30	0.0057	0.0056	1	0.655
Microsoft	1.65	0.28	0.0081	0.0011	0.655	1

Table 1: S&P 500 and Microsoft estimations based on 3 years of daily data

4 Simulation

Both sub-Gaussian and the proposed skewed sub-Gaussian distributions posses convenient stochastic representations. Thus the task of sampling from those distributions is reduced to sampling from multivariate Gaussian distribution and sampling from one-dimensional α-stable distribution.

The problem of simulating from a multivariate Gaussian distribution is well studied and it can be solved by applying singular value decomposition to the covariance and for example Box-Muller method for sampling from a standard normal distribution.

In order to simulate from an α-stable distribution $S_\alpha(\sigma, \beta, \mu)$ we can rely on the following algorithm:

- Generate U from $U(-\frac{\pi}{2}, \frac{\pi}{2})$ (uniformly distributed in the interval $(-\frac{\pi}{2}, \frac{\pi}{2})$) and E from $Exp(1)$ (exponentially distributed with mean 1).
- For $\alpha \neq 1$ compute

$$X = \mu + \sigma \left(1 + \beta^2 \tan^2 \frac{\pi\alpha}{2}\right)^{\frac{1}{2\alpha}} \cdot \frac{\sin\left(\alpha(U+B)\right)}{(cos(U))^{\frac{1}{\alpha}}} \cdot \left(\frac{\cos\left(U - \alpha\left(U+B\right)\right)}{E}\right)^{\frac{1-\alpha}{\alpha}}, \tag{11}$$

  where

$$B = \frac{1}{\alpha} \arctan\left(\beta \tan \frac{\pi\alpha}{2}\right)$$

- For $\alpha = 1$ compute

$$X = \mu + \sigma \frac{2}{\pi} \left(\beta \log \sigma + \left(\frac{\pi}{2} + \beta U\right) \tan U - \beta \log\left(\frac{\frac{\pi}{2} E \cos U}{\frac{\pi}{2} + \beta U}\right)\right). \tag{12}$$

Rigorous proof and deviation of the above algorithm based on the Chambers-Mallows-Stuck method can be found in [1], [8] and [9].

5 Example

In this section we provide some empirical results and produce simulations for the series of Standard & Poors 500 (SPX) index and Microsoft (MSFT). The multivariate distribution is fitted on the log-return series r_t which is obtained from the price series p_t using the following transformation

$$r_t = \log\left(\frac{r_t}{r_{t-1}}\right).$$

We use 750 daily log-returns (which is approximately 3 years of data) up to 27 Aug 2013.

Utilizing MLE approach (see [6]) we estimate the asymmetric stable parameters for the two series. Numbers are available in Table 1.

We also estimate the correlation matrix of the two series and use it to approximate the dispersion matrix of the sub-Gaussian component.

Comparison between historically observed log-returns and simulations using different multivariate models — multivariate Gaussian, sub-Gaussian and skewed sub-Gaussian with $\rho = 0.3$ is provided in Figure 2.

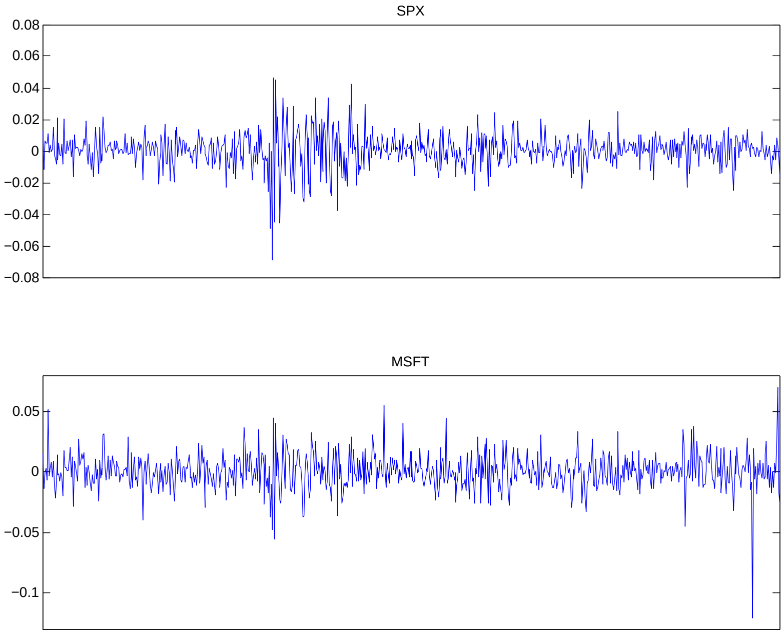


Figure 1: 750 S&P 500 and MSFT daily log-returns up to 27 Aug 2013

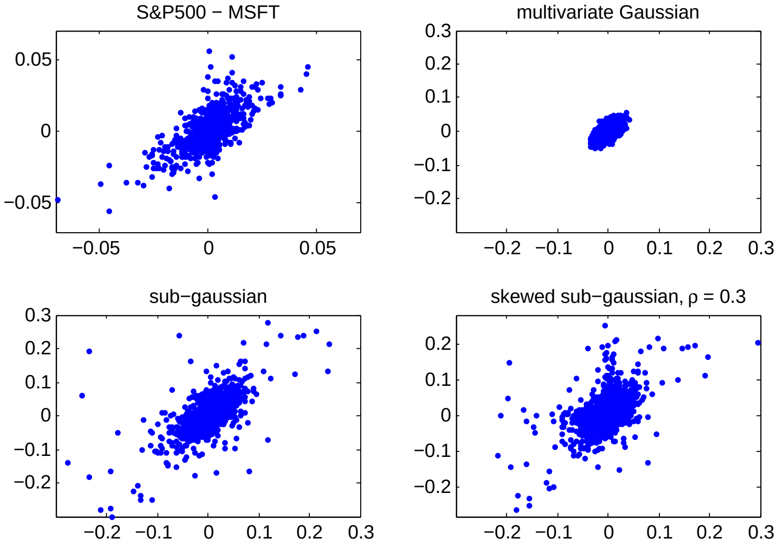


Figure 2: Historical and simulated returns of S&P 500 vs MSFT

6 Summary

We use the asymmetric stable distributions and their appealing properties to extend and add skew to the sub-Gaussian multivariate distribution. The derived distribution depends on a skew-weight parameter $\rho \in$

$(0, 1)$. We obtain the limiting distributions in case ρ tends to 0 (the regular sub-Gaussian distribution) and 1 (multivariate stable distribution with independent components).
Simulation techniques for the asymmetric stable distributions and hence for the sub-Gaussian and skewed sub-Gaussian distributions are outlined. We also present some empirical results using the proposed distribution although its estimation is still an open question.

Acknowledgements: This work was supported by the European Social Fund through the Human Resource Development Operational Programme under contract BG051PO001-3.3.06-0052 (2012/2014). The research is partially supported by appropriated state fund for research allocated to Sofia University (contract 111/2013), Bulgaria.

Bibliography

[1] Chambers, J.M., Mallows, C.L., Stuck B.W. (1976). *A method for simulating stable random variables*, Journal of the American Statistical Association 71, 340–344.

[2] Fama, E. (1963). *Mandelbrot and the Stable Paretian hypothesis*, Journal of Business 36, 420–429.

[3] Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, John Wiley and Sons, New Jersey.

[4] Feller, W. and Orey, S. (1961). *A Renewal Theorem*, Indiana Univ. Math. J. 10, 619–624.

[5] Mandelbrot, B. (1963). *The variation of certain speculative prices*, Journal of Business 26, 394–419.

[6] Rachev S. and Mittnik S., (2000), *Stable Paretian Models in Finance*, Wiley

[7] Samorodnitsky G., Taqqu M. S., (1994), *Stable Non-Gaussian Random Processes: Stochastic Models With Infinite Variance*, Chapman & Hall.

[8] Weron R., (1996), *On the Chambers-Mallows-Stuck method for simulating skewed stable random variables*, Statistics & Probability Letters 28.

[9] Weron R., (1996), *Correction to: On the Chambers-Mallows-Stuck method for simulating skewed stable random variables*.

Minimum description length principle and distribution complexity of spherical distributions

Bono Nonchev[*]

Faculty of Mathematics and Informatics, University of Sofia, Bulgaria

Abstract

The application of the MDL principle to discern from which distribution a sample originates is discussed with the focus is on the general class of spherical distributions. Their trivial generalization the elliptical distributions are widely used in financial theory and have properties that enable us to calculate a closed form solution of their distribution complexity.

The MDL principle and its codelength/model interpretation is discussed first, as well as its application in model selection. Then the NML model is introduced as a suitable choice and its equivalent formulation as the model complexity is explored. After that the distribution complexity is presented as a solution of the problem of infinite model complexity, with the rest of the paper exposing the main result - the calculation of the distribution complexity for spherical distributions.

The analytical formulas for the distribution complexity are explicitly shown in three cases - the Gaussian distribution, the Student-T distribution and the Laplace distribution. Thoughts on their interpretation of the change of complexity with the size of the sample are presented with a somewhat surprising characterization of the NML model for the spherical distributions that has potential impact on robust estimation.

Keywords: MDL, Model Selection, Complexity, Distribution Selection, Spherical distributions, Student-T distribution, Laplace distribution
AMS subject classifications: 94A17, 62B10, 62F03

1 Introduction

The problem discussed in this paper is problem of determining the distribution of a sample using the Minimum Description Length principle (MDL). The choice of distributions is the general class of spherical distributions.

This problem of model selection is one of the classical problems in statistics. Using a naïve approach to selection using the Neyman-Pearson lemma runs into problems as soon as the simple exact two distribution test is extended to a continuum of hypotheses.

Naturally a more sophisticated Bayesian approach like that of [1] yields more convincing results by casting the problem into a Bayesian framework, however there is something deeply unsatisfying in assigning subjective prior probabilities. Using Jeffreys' objective priors instead turns out to be very closely related to the MDL principle.

The Minimum Description Length principle (MDL) in its most basic form states that the more the data generated by a process can be compressed, the more we know about it. This simple idea has some very interesting applications and is presented briefly in section 2.

The purpose of this paper is to present the surprising result that in a very important sense all spherical distributions have identical descriptive power.

[*] e-mail: bono.nonchev@gmail.com

2 Minimum Description Length Principle

A classic example problem used in the inspirational paper of Kolmogorov [2] is that if you are charged with the task of transmitting three sequences of a million symbols, each 0 or 1, like the following

- 01010101010101010101010101010101010...
- 11011001111111011111101100111111111...
- 10101010001110100011101000111010111110...

you can certainly do better than transmitting the whole sequence bit by bit, if you exploit the regularities in the data.

In each of those cases knowledge of the patterns in the data would allows more effective transmission, which is why the MDL principle equates knowledge with compression. The first sequence is just 01 repeated, so sending this instruction instead is quite a lot faster. The second has about 9 ones for each zero, so long strings of ones can be encoded with shorter codes than strings of zeroes, and transmitted by shorter codes than the trivial. For the third not much can be done, as it is generated random and independent with equal probabilities of 0 and 1.

Regrettably allowing the use of any code renders the problem of finding the shortest codes uncomputable. The main insight of Rissanen is for the MDL principle to restrict the set of codes to those corresponding to probability distributions. In addition the distributions are only used as a description method and are not assumed to actually generate the modeled process.

Suppose there is a random variable X with distribution f. There is an optimal code called Shannon-Fano code for X that encodes an obvservation x with a codeword with length

$$L(x) = -\ln f(x) + \Delta \tag{1}$$

where Δ is a constant dependent only on the desired precision of x, so is usually skipped.

Most research in the area is focused on extending this and finding suitable coding schemes when there are many possible distributions for X. In the literature a set of distributions with some defining characteristic (e.g. the set of all normal distributions) is called a model. For each distribution there is an optimal code, namely the Shannon-Fano code. A single distribution that approximates all distributions in a model “well” is called an universal model and the main line of research on the MDL principle is the discovery and application of those models.

More general overview of the MDL principle can be found in [3] and one focused on statistical modelling in [4].

In this paper the Normalized Maximum Likelihood model is used, first introduced in [5] and subsequently thoroughly explored for various problems. It is defined as follows: suppose a sample is to be modeled using a parametric family with parameter θ and have its MLE $\hat{\theta}(x)$. A natural idea is to use code with length

$$L_{NML}(x) = -\ln f(x|\hat{\theta}(x)) + \ln \int f\left(y|\hat{\mu}(y), \hat{\sigma}(y)\right) dy = -\ln f(x|\hat{\theta}(x)) + COMP_n\left(f\right)$$

The last term is called the complexity of the model and the NML model is defined only when $COMP_n\left(f\right) < \infty$.

This is the basis for the stochastic complexity (SC) criterion for model selection: having a finite number of competing models, encode the sample using the NML distribution for each model and choose the one having the smallest codelength $L_{NML}^{\mathcal{M}}(x)$. The chosen model is the best description of the data at hand.

The main result in this paper is the computation of the model complexity for the spherical distributions and the fact that they are all in some sense equivalent.

3 Scale-location families

In this section some basic definitions and previous results are provided as discussed in [6].
As it turns out for scale-location families the model complexity is infinite. There are several ways to deal with the infinities, most notable of which are the *renormalization* by complexity conditional on the data space as presented in [7] and the usage of complexity conditional on the parameter space as in [8]. The first approach allows comparison between different distributions, but the renormalization step is not really needed.
A scale-location family is a family of distributions having p.d.f. $f(\mathbf{x}^n|\mu,\sigma)$ for which a function $g(\mathbf{y}^n)$ exists satisfying

$$f(\mathbf{x}^n|\mu,\sigma) = \sigma^{-n} g\left(\frac{\mathbf{x}^n-\mu}{\sigma}\right)$$

The model complexity for many families, including the above, turns out to be infinite, so a natural choice is to use the complexity conditional on $\mathbf{x}^n$. This has been studied in [6] and the following important decomposition applies (*Theorem 1*, pp. 109):

$$COMP_n(\mathcal{M}|\{-R\le\hat{\mu}\le R, D\le\hat{\sigma}\}) = \ln\int_{\mathbf{x}^n\in\mathcal{A}} f(\mathbf{x}^n|\hat{\mu},\hat{\sigma})\,d\mathbf{x}^n =$$

$$= \ln 2RD^{-1} + \ln DC_n(\mathcal{M})$$

The last term is called the distribution complexity, because it does not depend on the restriction of $\mathbf{x}^n$, freeing the model comparison procedure of the arbitrary bounds R and D. It is defined as

$$DC_n(\mathcal{M}) = \mathbb{E}_{\mathbf{Y}^n}\left[\delta\left(\hat{\mu}(\mathbf{Y}^n)\left(1-\hat{\sigma}(\mathbf{Y}^n)\right)\right)\right] = \int \delta\left(\mu(Y^n)\right)\delta\left(1-\sigma(Y^n)\right) g(y^n)dy^n$$

where $\mathbf{Y}^n \sim g(\mathbf{y}^n)$ and δ is the Dirac delta function.

4 Elliptical distributions

Let $f(x)$ be an arbitrary univariate distribution satisfying $f(x) = ch(x)$ for an even functon h and a normalizing constant c. The multivariate spherical generalization of g is defined as

$$f(x^n|\mu,\sigma) = c\sigma^{-n}h\left(\sigma^{-2}\left(x^n-\mu\right)^T\left(x^n-\mu\right)\right)$$

The defining feature of a spherical family is that $\overline{x} = \frac{1}{n}\sum x_i$ and $s^2 = \frac{1}{n}\sum(x_i-\overline{x})^2$ are sufficient statistics for μ and σ. If in addition h is decreasing and differentiable and w_0 is the smallest non-negative solution of $-\frac{2}{n}w\frac{\partial h}{\partial w}(w) = h(w)$, then the MLEs for (μ,σ) are $\hat{\mu} = \overline{x}$ and $\hat{\sigma}^2 = \frac{n}{w_0}s^2$. Thus the spherical family can be re-parameterized to have $\hat{\sigma}^2 = s^2$ by multiplying σ by $\sqrt{\frac{n}{w_0}}$.
Note: Other parameterizations require different conditions on the model complexity in order to use the same region of $\mathbf{x}^n$ between models, which cancels their effect on the distribution complexity.
Using this parameterization allows direct application of the results for scale-location family from section 3. Combined with the application of the properties of the δ-function and the fact that on the peak of the distribution's p.d.f. its value is equal to $c\cdot h(n)$ (hence is independent of the sample), an analytic formula for model complexity can be obtained as follows

$$DC_n(\mathcal{M}) = \qquad 2n^2\int_{2n-2q-p^2>0} g(y^{n-2}, y_{n-1}(\cdot), y_n(\cdot))\left(2n-2q-p^2\right)^{-\frac{1}{2}} dy^{n-2} =$$

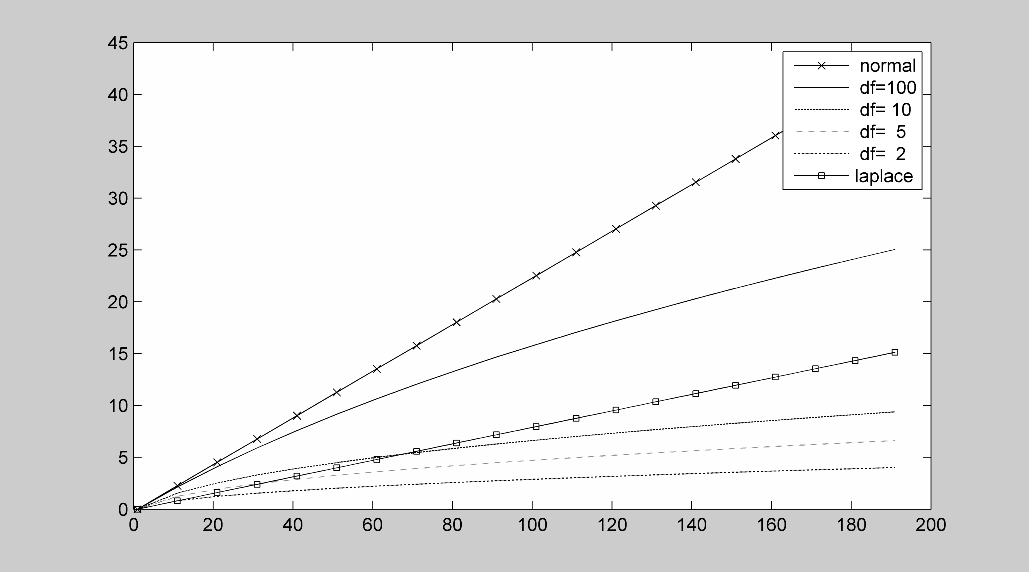


Figure 1: The distribution complexity of various distributions vs sample size.

$$= \frac{2n^{\frac{n}{2}}\pi^{\frac{n-1}{2}}}{\Gamma\left(\frac{n-1}{2}\right)}[c \cdot h(n)]$$

So for any spherical distribution satisfying the above relatively weak conditions allow an analytic formula for the model complexity. After substituting the likelihood evaluated at the MLE is $f(x|\hat{\theta}(x)) = s^{-n}c \cdot h\left(n\right)$, the codelength used in the SC becomes

$$L_{NML}(x) = n \ln s - \ln\left[c \cdot h\left(n\right)\right] + \ln \frac{2n^{\frac{n}{2}}\pi^{\frac{n-1}{2}}}{\Gamma\left(\frac{n-1}{2}\right)} + \ln\left[c \cdot h(n)\right] = n \ln s + \ln \frac{2n^{\frac{n}{2}}\pi^{\frac{n-1}{2}}}{\Gamma\left(\frac{n-1}{2}\right)}$$

which does not depend on the actual spherical distribution, so regrettably it cannot be used to distinguish between the spherical distributions.

5 Examples: Normal, Student-T and Laplace

In addition to the classical result for the complexity of the normal distribution, two more distributions' compexities are plotted on figure 1:

$$DC_n\left(\mathcal{M}\right) = \begin{cases} \frac{2\left(\frac{n}{2}\right)^{\frac{n}{2}} e^{-\frac{n}{2}}}{\sqrt{\pi}\Gamma\left(\frac{n-1}{2}\right)} & \text{for the Gaussian distribution} \\ \frac{2n^{\frac{n}{2}}\Gamma\left(\frac{n+\nu}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{n-1}{2}\right)\nu^{\frac{n}{2}}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{n}{\nu}\right)^{-\frac{n+\nu}{2}} & \text{for the Student-T distribution} \\ \frac{n^n\Gamma\left(\frac{n}{2}\right)}{2\sqrt{\pi}\Gamma\left(\frac{n-1}{2}\right)\Gamma(n)}e^{-n} & \text{for the Laplace distribution} \end{cases}$$

6 Conclusions

The fact that the spherical distributions have complexity that offsets exactly their log-likelihood is surprising and it shows that the fitting method is actually the one responsible for the model complexity in the NML model, not the distribution itself. This means that for robust estimators, where a computing method for the parameters is usually described instead of a distribution, the model complexity as defined by Shtarkov in [5] can be extended to be used for any method of fitting and is a sensible way to measure its model complexity. A classical result is that for the Student-T distribution there is no MLE for μ, σ and the degrees of freedom ν simultaneously. This is caused by the fact that the derivative of the likelihood is positive for all ν, however it can be interpreted through the fact that the model complexity increases with ν and the log-likelihood's bias toward choosing the more complex model. Accounting for the model complexity in the SC criterion will also fail to choose a model, but for a different reason - because of the indifference to the actual distribution, as long as it is spherical.

Acknowledgements: I want to the thank my PhD advisor Prof. Plamen Mateev for his continuing support and inspiration. This work was supported by the European Social Fund through the Human Resource Development Operational Programme under contract BG051PO001-3.3.06-0052 (2012/2014). The research is partially supported by appropriated state fund for research allocated to Sofia University (contract 111/2013), Bulgaria.

Bibliography

[1] Kass, R. and Raftery, A. (1995) Bayes factors. *Journal of the American Statistical Association*. 90, 773-–795.

[2] Kolmogorov, A. N. (1963) On Tables of Random Numbers. *Sankhya: The Indian Journal of Statistics, Series A*. 25, 369-–376.

[3] Grünwald, P. (2007) *The Minimum Description Length Principle*, The MIT Press, Cambridge MA.

[4] Rissanen, J. (2007) *Information and Complexity in Statistical Modeling (Information Science and Statistics)*, Springer, 2007.

[5] Shtarkov, Y. (1987) Universal Sequential Coding of Single Messages. *Problems of Information Transmission*. 23, 175-–186.

[6] Nonchev, B. (2013) Minimum Description Length Principle in Discriminating Marginal Distributions. *Pliska Stud. Math. Bulgar.*. 22, 101-–114.

[7] Rissanen, J. (2000) MDL Denoising. *IEEE Transactions on Information Theory*. 46, 2537-–2543.

[8] Stine, R. and Foster, D. (2001) The Competitive Complexity Ratio. *Proceedings of the 2001 Conference on Information Sciences and Systems*. WP8 1–6.

[9] Grünwald, P., Myung, J. I., and Pitt, M. (2005) *Advances in Minimum Description Length: Theory and Applications*, The MIT Press, Cambridge MA.

[10] Qian, G. and Künsch, H. (1996) On Model Selection in Robust Linear Regression. *Research Report NO. 80*, ETH Zürich.

On the tail index inference based on the scaling function method

Danijel Grahovac*[1], Mofei Jia[2], Nikolai Leonenko[3] and Emanuele Taufer[2]

[1] *Department of Mathematics, Josip Juraj Strossmayer University, Osijek, Croatia*
[2] *Department of Economics and Management, University of Trento, Italy*
[3] *Cardiff School of Mathematics, Cardiff University, UK*

Abstract

In [3], a new method has been presented for making inference about the tail of samples coming from unknown heavy-tailed distribution. Method is based on asymptotic properties of the empirical structure function, a variant of statistic that resembles usual sample moments. Using this approach one can successfully inspect the nature of the tail of the underlying distribution, as well as provide estimated values on the unknown tail index. Here we briefly describe the method and test its performance on some simulated and real world data by comparing it with the well known Hill estimator.

Keywords: heavy-tailed distributions, tail index, empirical structure function, scaling functions, Hill estimator.
AMS subject classifications: 62F10, 62F12, 62E20.

1 Introduction

Heavy-tailed distributions are of considerable importance in modeling a wide range of phenomena in finance, geology, hydrology, physics, queuing theory and telecommunication. Since the work of Mandelbrot [4], where stable distributions with index less than 2 have been advocated for describing fluctuations of cotton prices, there has been an exhausting research concerning the use of heavy-tailed distribution in the context of finance.
We define that the distribution of some random variable X is heavy-tailed with index $\alpha > 0$ if it has a regularly varying tail with index $-\alpha$, i.e.

$$P(|X| > x) = \frac{L(x)}{x^{\alpha}}, \quad |x| \to \infty,$$

where $L(t), t > 0$, is a slowly varying function, i.e., $L(tx)/L(x) \to 1$ as $|x| \to \infty$, for every $t > 0$. In particular, this implies that $E|X|^q < \infty$ for $q < \alpha$ and $E|X|^q = \infty$ for $q > \alpha$, which can be used as the alternative definition. We are interested in the estimation of the unknown tail index α, measuring the ”thickness” of the tails, based on the finite data sample with no additional assumptions on the distribution of the data.
There exists a range of estimators for this particular problem. The most well known estimators are the Pickand’s, Hill and moment estimator by Dekkers, Einmahl and de Haan. A nice survey of these estimators and their properties can be found in [2] and [1]. Tail index estimators are usually based on upper order statistics and their asymptotic properties. As an alternative, [5] proposed an estimator based on the asymptotics of the partial sum. In this paper we present a novel approach given in [3]. We evaluate the performance of this estimator on some real world examples and compare it with probably the most popular one, the Hill estimator.

[*] Corresponding author, e-mail: dgrahova@mathos.hr

2 Estimation method

The estimator presented in [3] is based on asymptotic properties of the empirical structure function (also called partition function), a kind of statistic that resembles usual sample moments. More precisely, given a sample $X_1, \dots, X_n$ coming from a strictly stationary stochastic process $\{X_t, t \in \mathbb{Z}_+\}$ (discrete time) or $\{X_t, t \in \mathbb{R}_+\}$ (continuous time) which has a heavy-tailed marginal distribution with unknown tail index α, define

$$S_q(n,t) = \frac{1}{\lfloor n/t \rfloor} \sum_{i=1}^{\lfloor n/t \rfloor} \left| \sum_{j=1}^{\lfloor t \rfloor} X_{t(i-1)+j} \right|^q, \tag{1}$$

where $q > 0$ and $1 \le t \le n$. In words, we partition the data into consecutive blocks of length $\lfloor t \rfloor$, then sum each block and take the power q of the absolute value of the sum. Finally, we average over all $\lfloor n/t \rfloor$ blocks. Notice that for $t = 1$ one gets the usual empirical q-th absolute moment.

Asymptotic properties of $S_q(n,t)$ have been considered before in the context of multifractality detection (see [3] and the references therein). Instead of keeping t fixed, we take it to be of the form $t = n^s$ for some $s \in (0,1)$, which allows the blocks to grow as the sample size increases. It is clear that then $S_q(n,n^s)$ will diverge since $s > 0$. The quantity of interest is the rate of divergence of this statistic, i.e. we consider the limiting behavior of $\ln S_q(n,n^s)/\ln n$. This has been established in [3] under the assumptions of strict stationarity of the sequence $X_t, t \in \mathbb{Z}_+$ and mild dependence condition in the form of the strong mixing property with an exponentially decaying rate (for details see [3]). It is also assumed that the expectation is zero in case when it is finite. The proof of the theorem can be found in [3].

Theorem 2.1. *Suppose $X_t, t \in \mathbb{Z}_+$ is a strictly stationary sequence that has a strong mixing property with an exponentially decaying rate and suppose that $X_t, t \in \mathbb{Z}_+$ has a heavy-tailed marginal distribution with tail index $\alpha > 0$. Suppose also that $EX_i = 0$ when $\alpha > 1$. Then for $q > 0$ and every $s \in (0,1)$*

$$\frac{\ln S_q(n,n^s)}{\ln n} \xrightarrow{P} R_\alpha(q,s) := \begin{cases} \frac{sq}{\alpha}, & \text{if } q \le \alpha \text{ and } \alpha \le 2, \\ s + \frac{q}{\alpha} - 1, & \text{if } q > \alpha \text{ and } \alpha \le 2, \\ \frac{sq}{2}, & \text{if } q \le \alpha \text{ and } \alpha > 2, \\ \max\left\{s + \frac{q}{\alpha} - 1, \frac{sq}{2}\right\}, & \text{if } q > \alpha \text{ and } \alpha > 2, \end{cases} \tag{2}$$

as $n \to \infty$, where $\xrightarrow{P}$ stands for convergence in probability.

It is clear that the limit considered in the preceding theorem heavily depends on the tail index α, which makes it possible to make inference about the unknown tail index. First notice that if for some non-negative sequence $\{Z_n\}$ of random variables $\ln Z_n/\ln n \xrightarrow{P} a \in \mathbb{R}$, then for some function M such that $\ln M(n)/\ln n \to 0$, $Z_n/n^a M(n) \xrightarrow{d} Z$ as $n \to \infty$, where Z is a random variable not identically equal to zero (possible degenerate). So, it follows from Theorem 2.1 that $\varepsilon_n := \frac{S_q(n,n^s)}{n^{R_\alpha(q,s)}M(n)} \xrightarrow{d} \varepsilon$, where ε is a random variable not identically equal to zero. By simply rewriting this, one arrives at the

$$\frac{\ln S_q(n,n^s)}{\ln n} = R_\alpha(q,s) + \frac{\ln M(n)}{\ln n} + \frac{\ln \varepsilon_n}{\ln n}. \tag{3}$$

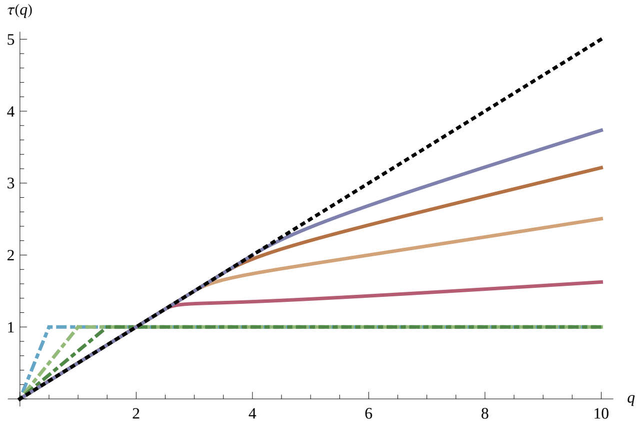
This equation can be seen as the regression model. The term $\ln \varepsilon_n/\ln n$ can be considered as an error term in the regression of $\ln S_q(n,n^s)/\ln n$ on q and s. One should count on the intercept in the model, in order to compensate for the $\ln M(n)/\ln n$ term. The possible nonzero mean of an error can be subtracted and considered as a part of the intercept.

The basic idea of the approach presented in [3] is to estimate the tail index α by the means of Equation (3). To avoid bivariate regression, one can assume the limit is linear in s, i.e. $R_\alpha(q,s) = \tau(q)s + c(q)$. This holds exactly except in the case $q > \alpha > 2$. By theoretically regressing $\ln S_q(n, n^s)/\ln n$ on s, for a range of values $s \in (0, 1)$, one gets the expression for $\tau(q)$ (notice that this is obvious in case $\alpha \leq 2$):

$$\tau(q) = \begin{cases} \frac{q}{\alpha}, & \text{if } 0 < q \leq \alpha \;\&\; \alpha \leq 2, \\ 1, & \text{if } q > \alpha \;\&\; \alpha \leq 2, \\ \frac{q}{2}, & \text{if } 0 < q \leq \alpha \;\&\; \alpha > 2, \\ \frac{q}{2} + \frac{2(\alpha-q)^2(2\alpha+4q-3\alpha q)}{\alpha^3(2-q)^2}, & \text{if } q > \alpha \;\&\; \alpha > 2. \end{cases} \tag{4}$$

$\tau(q)$ is refereed to as the scaling function. When α is large, i.e., $\alpha \to \infty$, it follows from (4) that $\tau(q) = q/2$. This corresponds to data coming from a distribution with all moments finite, e.g., an independent normally distributed sample. This line will be referred to as the baseline. Theoretical plots of scaling functions for a range of α values are shown in Fig. 1. It is clear that the shape of the scaling function is heavily influenced by the value of tail index α.

Figure 1: Plots of scaling function $\tau(q)$ against the moment q



The baseline is shown by a dashed line. The case $\alpha \leq 2$ ($\alpha = 0.5, 1.0, 1.5$) and $\alpha > 2$ ($\alpha = 2.5, 3.0, 3.5, 4.0$) are shown by dot-dashed and solid lines, respectively.

Having a finite data sample, one can estimate $\tau(q)$ in a single point q as the slope in the simple linear regression model by regressing $\ln S_q(n, n^s)/\ln n$ on s, for a range of values of $s \in (0, 1)$. More precisely, fix $q > 0$ and for $s_i \in (0, 1),\ i = 1, \ldots, m$ calculate $S_i = \ln S_q(n, n^{s_i})/\ln n,\ i = 1, \ldots, m$ based on the data sample. Now, estimate the value of the scaling function at the point q as

$$\left(\hat{\tau}(q), \hat{b}\right) = \underset{(a,b)\in\mathbb{R}^2}{\arg\min} \sum_{i=1}^{m} \left(S_i - as_i - b\right)^2. \tag{5}$$

Repeating this for a range of q makes it possible to give a plot of empirical scaling function $\hat{\tau}$. By comparing empirical scaling function with Fig. 1, one can make inference about the nature of the tails of the underlying distribution. Moreover, by minimizing the difference between the theoretical scaling function (4) and the empirical one $\hat{\tau}(q)$ for some range of $q \in (0, q_{max})$ one can find the estimate for α. More precisely, for points $q_i \in (0, q_{max}),\ i = 1, \ldots, n$, estimate $\tau_i = \hat{\tau}(q_i)$ by the means of Equation (5). Estimator is defined as

$$\hat{\alpha} = \underset{\alpha\in(0,\infty)}{\arg\min} \sum_{i=1}^{m} \sum_{j=1}^{k} (\tau_i - \tau(q_i))^2. \tag{6}$$

Method is divided in two cases, $\alpha \leq 2$ and $\alpha > 2$, in order to simplify the estimation procedure. Cases can be distinguished graphically by plotting the empirical scaling function.

3 Examples and comparison with Hill estimator

In this section we test the performance of estimator (6) on some known data sets and compare it with the Hill estimator. Let $X_{(1)} \geq X_{(2)} \geq \cdots \geq X_{(n)}$ denote the order statistics of the sample $X_1, X_2, \cdots X_n$, and k_n be a sequence of positive integers satisfying $1 \leq k_n < n$, $\lim_{n\to\infty} k_n = \infty$, and $\lim_{n\to\infty} (k_n/n) = 0$. The Hill estimator based on k_n upper order statistics is

$$\hat{\alpha}_{k_n} = \left(\frac{1}{k_n} \sum_{i=1}^{k_n} \log \frac{X_{(i)}}{X_{(k_n+1)}} \right)^{-1}. \tag{7}$$

Hill estimator is known to be weakly consistent as well as strongly consistent and asymptotically normal under certain conditions. For details see [2]. However, performance of the Hill estimator is heavily influenced by the choice of k_n. There is no generally accepted method on how to choose k_n, and it is usually recommended to plot the values for a range of k_n values and to look for the part of the graph where the value stabilizes. The resulting plot is usually called Hill plot.

3.1 Example 1 - non-constant slowly varying function in the tail

Hill estimator is known to behave poorly if the slowly varying function in the tail is far away from constant. We compare this behavior with the performance of the estimator (6). Consider two distribution F_1, F_2 defined by their survival functions

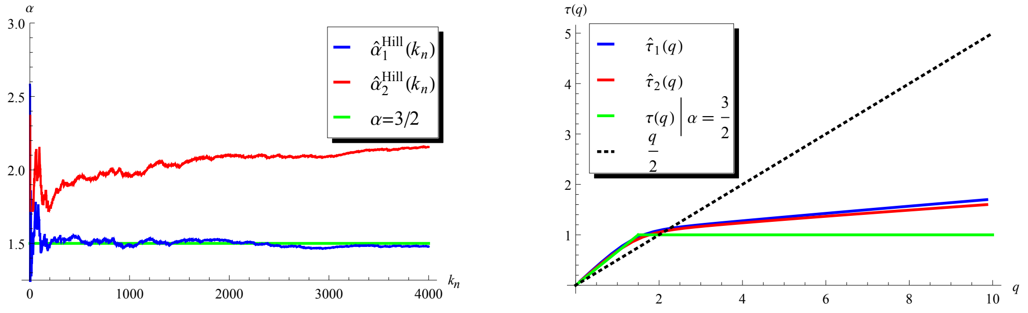
$$\overline{F}_1(x) = 1 - F_1(x) = \frac{1}{x^{\frac{3}{2}}}, \quad x \geq 1, \tag{8}$$

$$\overline{F}_2(x) = 1 - F_2(x) = \frac{e^{\frac{3}{2}}}{x^{\frac{3}{2}} \ln x}, \quad x \geq e. \tag{9}$$

Both distributions are heavy-tailed with tail index equal to $3/2$. We generate samples from these two distributions with 5000 observations. Corresponding Hill plots are shown in Figure 2(a). For F_2, one could wrongly conclude that the value of the tail index is around 2. The Hill's method is highly sensitive to the presence of non-constant slowly varying function in the tail. This is sometimes called Hill horror plot (see [2]). Figure 2(b) shows empirical scaling functions for the same samples together with the theoretical one and the baseline. One can see that scaling functions almost coincide with the theoretical one. Calculating estimates using (6) yields values $\hat{\alpha}_1 = 1.441$ and $\hat{\alpha}_2 = 1.5141$. It seems that non-constant slowly varying function affects the estimation but the effect is not so dramatical as for the Hill estimator. Most important part of the scaling functions for the inference about the tail is before the breakpoint and the breakpoint itself. For example, one can try estimating α only based on the values of $\hat{\tau}(q)$ for q less than a breakpoint observed graphically by fitting simple linear regression through origin. Theoretically, slope of the regression line should be $1/\alpha$. For example above, using $q \in (0, 1.5)$ one gets estimates for α: 1.454 for F_1 and 1.527 for F_2.

3.2 Example 2 - non heavy-tailed distribution

For the next example we compare the behavior of two estimators when the underlying distribution is not heavy-tailed. For this purpose, sample of 2000 observations was generated from standard logistic distribution



(a) Hill plot

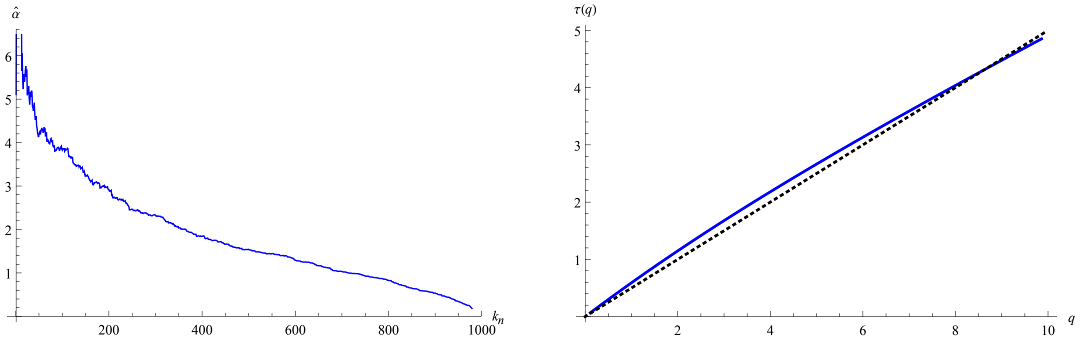
(b) Scaling functions

Figure 2: Example 1

given by probability density function

$$f(x) = \frac{e^{-x}}{(1+e^{-x})^2},\ x \in \mathbb{R}.$$

Figure 3(a) shows the Hill plot. It is impossible to draw any conclusion by only analyzing the Hill plot. This is why it is always necessary to use some other techniques for detecting heavy tails in data samples. On the other hand, estimated scaling function provides self contained characterization of the tail. From Figure 3(b) one can surely doubt the existence of heavy-tails since the empirical scaling function almost coincides with the baseline $q/2$.



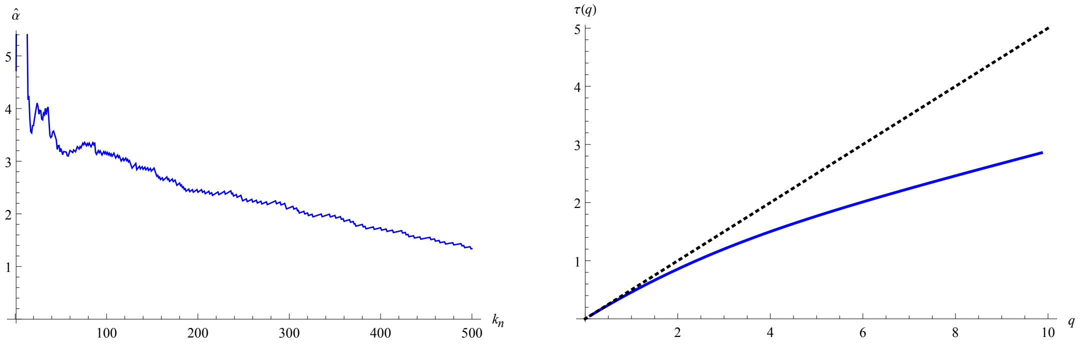
(a) Hill plot

(b) Scaling function

Figure 3: Scaling function

3.3 Example 3 - EUR/USD exchange rates

In this example we analyze daily closing rates of euro against U.S. dollar during the period $2007-2012$. The data consists of differences of rates and has 1868 observations. Hill plot is shown in Figure 4(a) and corresponding scaling function in the Figure 4(b). Hill plot fails to stabilize, but one could say this happens for k_n around 100 yielding, for example, value $\hat{\alpha} = 3.133$ for $k_n = 100$. Scaling function evidently points that the variance is finite since the break occurs after $q = 2$ and the plot coincides with the baseline before the break. Estimation for the case $\alpha > 2$ yields the value 3.112, consistent with the Hill estimator.



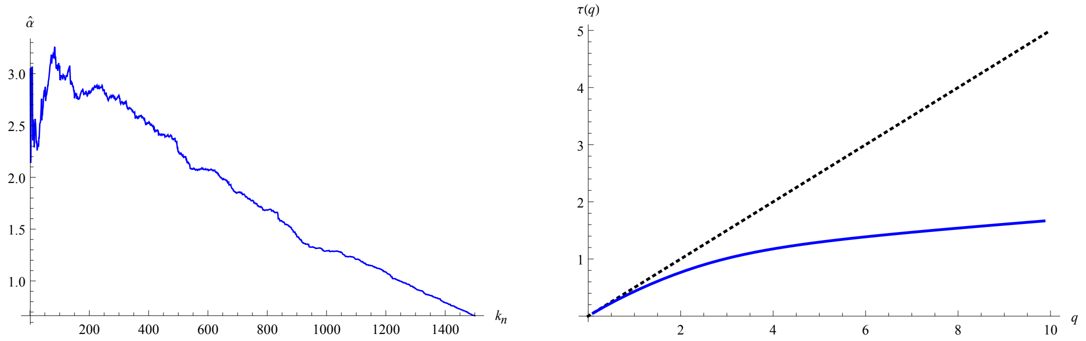
(a) Hill plot

(b) Scaling functions

Figure 4: Scaling function

3.4 Example 4 - daily log-returns of DAX

Next example again involves financial data. We use daily log-returns of the German stock index DAX (September 20, 1988 - August 24, 1995), similar to Figure 6.4.12 in [2]. Hill plot in Figure 5(a) is made by using absolute value of the data. Following [2], one can conclude that the plot stabilizes around 2.8 for $100 \leq k_n \leq 300$. However, plot fails to stabilize for larger k_n, similar as in the Example 1. Data has been centered for the estimation of the scaling function on Figure 5(b). Plot shows that α could be somewhere between 2 and 2.5. Calculating the estimate (6) yields the value 2.465. Thus, there is a significant discrepancy between two estimates. Considering the inconclusiveness of the Hill plot, one could give preference to the estimate (6).



(a) Hill plot

(b) Scaling function

Figure 5: Scaling function

Bibliography

[1] L. De Haan and A. Ferreira. *Extreme Value Theory: An Introduction*. Springer, 2006.

[2] Paul Embrechts, Claudia Klüppelberg, and Thomas Mikosch. *Modelling extremal events: for insurance and finance*, Volume 33. Springer Verlag, 1997.

[3] Danijel Grahovac, Mofei Jia, Nikolai N Leonenko, and Emanuele Taufer. Asymptotic properties of the partition function and applications in tail index inference of heavy-tailed data. *arXiv preprint arXiv:1310.0333*, 2013.

[4] B. Mandelbrot. The variation of certain speculative prices. *Journal of Business*, 36(4):394–419, 1963.

[5] M.M. Meerschaert and H.P. Scheffler. A simple robust estimation method for the thickness of heavy tails. *Journal of Statistical Planning and Inference*, 71(1):19–34, 1998.

Accelerated failure time model for repairable systems

Petr Novák[*]

Charles University in Prague, Faculty of Mathematics and Physics, Department of Probability and Mathematical Statistics

Abstract

When studying the service record of a device which is a subject to degradation, we want to estimate the time-to-failure distribution for maintenance optimization. The dependency of the failure time distribution on applicable regression variables can be described with a suitable model. For instance, we may use the number of repairs and maintenance actions or their cost as time-varying covariates. For this situation, the Cox proportional hazards model has been suggested, with the repairs and maintenance actions influencing the hazard function multiplicatively. Alternatively, we can use the Accelerated failure time model, where the covariates cause the internal time of the device to flow faster or slower. In this work we describe such models and demonstrate their application on real data.

Keywords: Reliability analysis, Repair models, Regression, Accelerated Failure Time model.
AMS subject classifications: 62N02.

1 Introduction

We study data describing a service record of one or more devices which degrade over time. In case of a failure, it is necessary to perform a repair. Preventive maintenance is performed to avoid breakdowns, and to optimize the maintenance costs, it is desirable to estimate the distribution of the time to failure with the help of available information. Regression models used in survival analysis can be adjusted to accommodate recurring repairs and maintenance actions. The Cox proportional hazards model for repairable systems was described by Percy and Kobbacy [6], with covariates multiplicatively influencing a parametric baseline hazard. In this work, we show a similar approach with the Accelerated failure time model (AFT) with time-varying covariates (Lin and Ying [3]), where the covariates and regression parameters influence multiplicatively the flow of the internal time of the device. Further, we show methods of estimating the cumulative baseline hazard nonparametrically if we have data on more devices, which allows us to estimate the regression parameters without assumptions on the shape of the baseline. Finally, we show the application of all described methods on real data from oil industry.

2 Modeling the lifetime of a repairable system

Suppose we observe n independent devices. Let $T_{i1}, ..., T_{in_i}$ be random variables representing the ordered times of actions (repair or maintenance) performed on the i-th device. Denote $\Delta_{i1}, ..., \Delta_{in_i}$ the indicators whether in j-th time on the i-th device a repair ($\Delta_{ij} = 1$) or a maintenance ($\Delta_{ij} = 0$) was performed and let $\boldsymbol{X}_i(t) = (X_{i1}(t), ..., X_{ip}(t))^T$ be explanatory variables, possibly time-varying.

[*] e-mail: novakp@karlin.mff.cuni.cz

We work with counting processes denoting the number of repairs and maintenance actions on the i-th device up to time t:

$$N_{i\bullet}(t) = \sum_{j=1}^{n_i} I(T_{ij} \le t, \Delta_{ij} = 1), \qquad M_{i\bullet}(t) = \sum_{j=1}^{n_i} I(T_{ij} \le t, \Delta_{ij} = 0).$$

Denote the hazard function for the i-th device

$$\lambda_i(t) = \lim_{h \to 0} P(N_{i\bullet}(t+h) - N_{i\bullet}(t) \ge 1 | \mathcal{H}(t))/h$$

where $\mathcal{H}(t)$ is the history of events up to time t. Further denote the cumulative hazard functions $\Lambda_i(t) = \int_0^t \lambda_i(s)ds$ and $S_i(t) = \exp(-\Lambda_i(t))$ corresponding survival functions of the time to failure distributions. We assume that a repair returns the device to working state and that it affects the hazard function. We parametrize the hazard function and estimate the parameters using the maximum likelihood method. The likelihood can be written as

$$L = \prod_{i=1}^{n} \left(\prod_{j=1}^{n_i} \lambda_i(T_{ij}^-)^{\Delta_{ij}} \cdot S_i(T_{in_i}) \right).$$

We need the left limit, as λ_i change at the times of the events. The log-likelihood has then the form

$$l = \sum_{ij} \Delta_{ij} \log \lambda_i(T_{ij}^-) - \sum_i \int_0^{T_{in_i}} \lambda_i(t)dt. \tag{1}$$

With help of the counting processes of the failures $N_{ij}(t) = \Delta_{ij} I(T_{ij} \le t)$ and the at-risk indicators $Y_{ij}(t) = I(T_{i,j-1} < t \le T_{ij})$, we may write the log-likelihood as

$$l = \sum_{ij} \int_0^{\infty} \Big(\log \lambda_i(t^-) dN_{ij}(t) - Y_{ij}(t)\lambda_i(t^-)dt \Big).$$

3 The Accelerated failure time model

We assume that the covariates cause the internal time of the device to flow faster or slower (Accelerated Failure Time model, AFT). We use the time transformation (Lin and Ying [3])

$$t \to \int_0^t e^{\boldsymbol{X}_i^T(s)\boldsymbol{\beta}} ds =: h_i(t, \boldsymbol{\beta}),$$

Denote λ_0 the baseline hazard function. The AFT model works with the hazard function for the i-th device in the form

$$\lambda_i(t) = \lambda_0(h_i(t, \boldsymbol{\beta}))e^{\boldsymbol{X}_i^T(t)\boldsymbol{\beta}}.$$

If the baseline hazard function is constant (corresponding with the exponential distribution), the AFT model coincides with the Cox proportional hazards $\lambda_i(t) = \lambda_0(t)e^{\boldsymbol{X}_i^T(t)\boldsymbol{\beta}}$, where the covariates affect the hazard function multiplicatively.

3.1 Parametric inference

We can assume that each repair or maintenance action has an influence on the flow of the time and set the number of actions ($N_{i\bullet}(t)$ and $M_{i\bullet}(t)$) as explanatory variables. Furthermore, we can add the cost or type of the last action as a covariate (Percy and Alkali [4]). If the covariate values change only in the times of observed events and the baseline hazard $\lambda_0(t)$ is parametric, it is possible to insert the hazard function into the log-likelihood 1 and maximize. The significance of the parameters can be then assessed by a likelihood ratio test, with $2\left(l(\hat{\boldsymbol{\beta}}) - l(\beta_{10}, \hat{\boldsymbol{\beta}}_{(2,\dots p)})\right) \sim \chi^2_1$ for testing $\beta_1 = \beta_{10}$ etc.

3.2 Semiparametric inference

If we have data on more than one device, it is possible to estimate the baseline nonparametrically. This may be desirable, since we do not need to pose any assumptions on the form of the baseline and focus solely on the regression parameters.

For each device we have the time transformation $t \to h_i(t, \boldsymbol{\beta})$. We work with time-transformed processes

$$N^*_{ij}(t,\boldsymbol{\beta}) = \Delta_{ij} I(h_i(T_{ij},\boldsymbol{\beta}) \le t), \qquad M^*_{ij}(t,\boldsymbol{\beta}) = (1-\Delta_{ij}) I(h_i(T_{ij},\boldsymbol{\beta}) \le t),$$

$$Y^*_{ij}(t,\boldsymbol{\beta}) = I(h_i(T_{i,j-1},\boldsymbol{\beta}) < t \le h_i(T_{ij},\boldsymbol{\beta})), \qquad \boldsymbol{X}^*_i(t,\boldsymbol{\beta}) = \boldsymbol{X}_i(h_i^{-1}(t,\boldsymbol{\beta})).$$

The score obtained by taking the derivative of the log-likelihood with respect to $\boldsymbol{\beta}$ has the form

$$U(\boldsymbol{\beta}) = \sum_{ij} \int_0^\infty W_i(t^-,\boldsymbol{\beta}) \left(dN^*_{ij}(t,\boldsymbol{\beta}) - Y^*_{ij}(t,\boldsymbol{\beta}) d\Lambda_0(t) \right),$$

where $W_i(t,\boldsymbol{\beta}) = \frac{\lambda_0'(t)}{\lambda_0(t)} \int_0^{h_i^{-1}(t,\boldsymbol{\beta})} \boldsymbol{X}_i^T(s) e^{\boldsymbol{X}_i^T(s)\boldsymbol{\beta}} ds + \boldsymbol{X}^*_i(t,\boldsymbol{\beta})$. This form is relatively complicated, with terms λ_0' and λ_0 not easy to estimate. However, it can be replaced by the approximate score by inserting $W_i^0(t,\boldsymbol{\beta}) = \boldsymbol{X}^*_i(t,\boldsymbol{\beta})$ instead of $W_i(t,\boldsymbol{\beta})$ (Lin and Ying [3]). We can then insert the Nelson-Aalen estimate of the cumulative baseline hazard function

$$\hat{\Lambda}_0(t,\boldsymbol{\beta}) = \int_0^t \frac{dN^*_{\bullet\bullet}(s,\boldsymbol{\beta})}{\sum_{ij} Y^*_{ij}(s,\boldsymbol{\beta})}$$

and get the score in form

$$U(\boldsymbol{\beta}) = \sum_{ij} \int_0^\infty \left(\mathbf{X}^*_i(t^-,\boldsymbol{\beta}) - \frac{\sum_{kl} \boldsymbol{X}^*_k(t^-,\boldsymbol{\beta}) Y^*_{kl}(t,\boldsymbol{\beta})}{\sum_{kl} Y^*_{kl}(t,\boldsymbol{\beta})} \right) dN^*_{ij}(t,\boldsymbol{\beta}).$$

Because the score is not continuous in $\boldsymbol{\beta}$, we obtain the parameter estimates by minimizing $\|U(\boldsymbol{\beta})\|$.

The variance of $\hat{\boldsymbol{\beta}}$ depends on the unknown $\lambda_0'(t)$ and $\lambda_0(t)$ and cannot be estimated easily. A resampling technique has been developed to avoid this problem (Lin et.al [2]) but is not described here due to pressure of space. The significance of the parameters can be tested with the score statistics

$$U(\beta_{10}, \hat{\boldsymbol{\beta}}_{(2,\dots p)})^T \hat{I}^{-1}(\beta_{10}, \hat{\boldsymbol{\beta}}_{(2,\dots p)}) U(\beta_{10}, \hat{\boldsymbol{\beta}}_{(2,\dots p)})$$

etc., where $\hat{I}$ is the estimate of the observed information matrix,

$$\hat{I}(\boldsymbol{\beta}) = \sum_{ij} \int_0^\infty \left(\mathbf{X}^*_i(t^-,\boldsymbol{\beta}) - \frac{\sum_{kl} \boldsymbol{X}^*_k(t^-,\boldsymbol{\beta}) Y^*_{kl}(t,\boldsymbol{\beta})}{\sum_{kl} Y^*_{kl}(t,\boldsymbol{\beta})} \right)^{\otimes 2} dN^*_{ij}(t,\boldsymbol{\beta}).$$

4 Modeling the lifetime of oil pumps

We explore data on the service of oil pumps during several years (Kobbacy *et al.* [1] and Percy and Alkali [5]). For one device we have detailed data on $n_1 = 65$ times of repairs and maintenance actions and the cost of each action in man-hours. This data has been studied by Percy and Alkali [4] using the Cox model. We model the lifetime using the AFT model as shown above with various parametrized baseline hazard functions and covariates. For four other pumps we have only the times of actions at disposal, $(n_2, \ldots, n_4) = (51, 90, 30, 30)$. We use both the semiparametric methods and parametrized baseline hazards to estimate the regression parameters utilizing data of all the five pumps.

4.1 Parametric modeling of one pump service

As covariates, we use the number of repairs $N_{i\bullet}(t)$ and maintenances $M_{i\bullet}(t)$, the indicator whether the last action was a repair $N_{i\Delta}(t) = \sum_{j=1}^{n} \Delta_{ij} I(T_{ij} \leq t < T_{i,j+1})$ and the cost of the actions, with parameters $\beta = (\sigma, \rho, \tau, b)^T$. Using methods from above we estimate the parameters in the AFT model. We try to maximize the likelihood for exponential, Weibull $\lambda_0(t) = a\lambda^a t^{a-1}$, gamma $f(t) \propto t^{a-1}e^{-\lambda t}$ and truncated Gumbel $\lambda_0(t) = \lambda a^t$ baseline distributions. We perform the likelihood ratio test to determine whether each covariate can be replaced by zero.

λ_0	log - lik	$e^{\hat{\sigma}}$	$e^{\hat{\rho}}$	$e^{\hat{\tau}}$	$e^{\hat{b}}$	$\hat{\lambda}$	$\hat{a}$
Exp.	-213.3	0.910	1.443	1.448	1.0054	0.0011	–
Significance		< 0.001	< 0.001	0.052	0.086		
Weibull	-212.6	0.913	1.315	1.243	1.0054	0.0009	1.514
Significance		< 0.001	< 0.001	0.140	0.068		
Gamma	-213.2	0.901	1.501	1.506	1.0052	0.0007	0.722
Significance		< 0.001	< 0.001	0.034	0.095		
Gumbel	-210.2	0.870	1.373	1.204	1.0043	0.0004	1.001
Significance		< 0.001	< 0.001	0.017	0.047		

Table 1: The log-likelihood, parameter estimates and the p-values of the tests of nullity from parametric models of the lifetime of one oil pump.

Comparing the likelihood values in Table 1 we find that it is highest with the truncated Gumbel distribution. Further we see that the more each action did cost, the more it accelerated the internal time, because $e^{\hat{b}} > 1$. Each man-hour of the action means an acceleration of time by about 0.5%. This covariate is, however, significant on $\alpha = 0.05$ only in the last case. The cumulative number of repairs has a positive influence ($e^{\hat{\sigma}} < 1$), whereas the number of maintenance actions has interestingly a negative influence ($e^{\hat{\rho}} > 1$). This could be due to repairs often taking much more man-hours than maintenances, resulting in negative influence of both. The number of repairs and maintenances are both significant. If the last action was a repair, the time flows about $20 - 50\%$ faster compared to when it was a maintenance, but the influence is significant only with Gamma and Gumbel baselines.

4.2 Semiparametric modeling of the lifetime of five pumps

For five devices we have only the times of repairs and maintenances available. The data on the cost of the actions was not available for all pumps, therefore we estimate only the regression parameters σ, ρ and τ. We tried both the parametric model with the same baseline distributions as above and the semiparametric model. In the parametric cases we maximize the log-likelihood whereas in the semiparametric approach we insert the estimate of the cumulative baseline hazard into the score function and minimize $\|U(\boldsymbol{\beta})\|$. We test the significance of each covariate with the likelihood ratio test for the parametric cases and with the score test for the semiparametric model. The likelihood in the semiparametric methods depends on the unknown baseline hazard and therefore is not available for a direct comparison.

λ_0	log - lik	$e^{\hat{\sigma}}$	signif.	$e^{\hat{\rho}}$	signif.	$e^{\hat{\tau}}$	signif.	$\hat{\lambda}$	$\hat{a}$
Exp.	-875.1	1.012	< 0.001	0.988	0.055	1.655	< 0.001	0.012	–
Weibull	-875.1	1.012	< 0.001	0.986	0.019	1.653	< 0.001	0.012	1.021
Gamma	-874.8	1.012	< 0.001	0.993	0.271	1.673	< 0.001	0.011	0.759
Gumbel	-871.3	1.033	< 0.001	1.025	0.033	1.544	< 0.001	0.010	0.999
nonparam.	–	1.072	0.007	1.016	0.015	1.064	0.030	–	–

Table 2: The log-likelihood, parameter estimates and the p-values of the tests of nullity from modeling the lifetime of five pumps.

In Table 2 we see that in all cases the internal time accelerates with each repair ($e^{\hat{\sigma}} > 1$). Among the parametric models, the Gumbel distribution has the highest likelihood. In that case and also in the semiparametric model, the maintenance actions have also a negative influence, whereas in the other cases it is positive. The time flows faster if the last action was a repair, but the extent of the temporary acceleration is much larger in the parametric models ($54 - 67\%$) than in the semiparametric (6.4%). This difference requires further study, it might be explained to some extent as a compensation for different tail behavior of the parametric and semiparametric estimates of the baseline hazard. For $\alpha = 0.05$, all covariates have a significant influence, except for the number of maintenances in the parametric case with exponential and Gamma distributions.

5 Conclusion

We explored methods for modeling the influence of maintenance and repairs on the lifetime of the observed device with the help of the Accelerated failure time model. The model has a straightforward interpretation, stating that the covariates accelerate or decelerate the flow of the internal time of the device and therefore cause it to age faster or slower. For a parametric baseline hazard function, the service record of one device is enough to obtain the estimates of the regression parameters. If we have data on more devices at disposal, it is possible to estimate the cumulative baseline hazard function nonparametrically. Further research could concern developing goodness-of-fit tests or testing whether a nonparametric estimate may be replaced by a suitable parametrized baseline hazard.

Acknowledgements: This work was supported by grants SVV 261315/2013 and GAUK 11122/2013.

Bibliography

[1] Kobbacy K.A.H., Fawzi B.B., Percy D.F. and Ascher H.E. (1997). A full history proportional hazards model for preventive maintenance scheduling. *Quality and Reliability Engineering Intl.* 13, 187–198.

[2] Lin D.Y., Wei L.J. and Ying Z. (1998) Accelerated failure time models for counting processes. *Biometrika* 85(3), 605–618.

[3] Lin D.Y. and Ying Z. (1995). Semiparametric inference for the accelerated life model with time-dependent covariates. *Journal od Statistical Planning and Inference* 44, 47–63.

[4] Percy D.F. and Alkali B.M. (2005) Generalized proportional intensities models for repairable systems. *IMA Journal of Management Mathematics* 17, 171–185.

[5] Percy D.F. and Alkali B.M. (2007) Scheduling preventive maintenance for oil pumps using generalized proportional intensities models. *International Transactions of Operational Research* 14, 547–563.

[6] Percy D.F. and Kobbacy K.A.H. (1998) Using proportional-intensities models to schedule preventive-maintenance intervals *IMA Journal of Mathematics Applied in Business & Industry*, 9, 289–302.

Score test statistic for change-point detection in AR time series with dependent errors

Katarína Starinská[*]

Department of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University, Prague

Abstract

Detecting changes in the parameter values of any model is of great importance for many sectors. With a model up to date we are able to give better predictions. Finding a change-point can help us to understand the influence of some events on observed data. The efficient score test statistic was introduced in [5] for detecting changes in the parameters of autoregressive(AR) time series with independent identically distributed(i.i.d) errors. This test allows us to detect a change in all the parameters at once or in every parameter separately. We study the behavior of this statistic when the assumption of i.i.d. white noise is violated and replaced with the assumption of having martingale difference sequence. We present the simulation study which shows us the asymptotic behavior and the power of this test statistic.

Keywords: Change-point detection, Invariance principle, Autoregressive time series.
AMS subject classifications: primary 62F05, secondary 60F17, 62M10.

1 Introduction

One of the first tests for a parameter change at an unknown time was suggested in the article [8] in 1954. Since then, many methods were developed and studied. For a review see [1]. The changes in the parameters of the autoregressive (AR) time series were studied e.g. in [3], [5] or [6]. Our interest is in the efficient score statistic, which was introduced in [5] under the assumption of independent identically distributed (i.i.d.) error sequence. There are processes that can be expressed as AR processes, for example general integer autoregressive process (GINAR) or random coefficient autoregressive process (RCA), in which the assumption of i.i.d. errors does not hold. We are studying how the efficient score statistic works under weaker assumption than an i.i.d. white noise sequence.

2 Score test statistic

In this section we describe the construction of score test statistic and shortly explain the test for detecting changes in one parameter or in all the parameters at once.
Let the sequence $\{Y_i\}$ satisfy the autoregressive model of order p AR(p)

$$\begin{aligned} Y_i - \mu &= \sum_{j=1}^{p} \phi_j (Y_{i-j} - \mu) + \varepsilon_i, \quad 1 \le i \le k \\ Y_i - \mu^* &= \sum_{j=1}^{p} \phi_j^* (Y_{i-j} - \mu^*) + \varepsilon_i^*, \quad k+1 \le i \le n. \end{aligned} \tag{1}$$

[*] e-mail: starinskak@gmail.com

where $\{\varepsilon_i\}$ resp. $\{\varepsilon_i^*\}$ is a noise sequence with zero mean and $\mathbb{E}[\varepsilon_i^2] = \sigma^2$ resp. $\mathbb{E}[\varepsilon_i^{*2}] = \sigma^{*2}$,

$$\xi = (\mu, \phi_1, \ldots, \phi_p, \sigma^2)'$$

is the vector of parameters before change and

$$\xi^* = (\mu^*, \phi_1^*, \ldots, \phi_p^*, \sigma^{*2})'$$

is the vector of parameters after change-point k.
We test the null hypothesis that there is no change in the parameter values against the alternative hypothesis, that there exist time k such that at least one of the observed parameters changed:

$$\begin{aligned} H &: k = n, \\ A &: k < n. \end{aligned}$$

We derive the efficient score statistic under the null hypothesis and under the assumption of normally distributed error sequence $\{\varepsilon_i\}$. Therefore we know the analytical expression of the conditional density $f(Y, \xi) = f(Y_i | Y_{i-1}, \ldots, Y_{i-p})$ of Y_i under $Y_{i-1}, \ldots, Y_{i-p}$ and the (conditional) logarithmic likelihood function of $Y_{-p+1}, \ldots, Y_0, Y_1, \ldots, Y_k$

$$\ell_k(\xi) = -\frac{k}{2} \log(2\pi\sigma^2) - \frac{1}{2} \sum_{i=1}^{k} \left[Y_i - \mu - \sum_{j=1}^{p} \alpha_j (Y_{i-j} - \mu) \right]^2 .$$

The efficient score vector is the vector of partial derivations of the logarithmic likelihood function with respect to unknown parameters $\nabla_\xi \ell_k(\xi) = \left(\frac{\partial \ell_k(\xi)}{\partial \mu}, \frac{\partial \ell_k(\xi)}{\partial \phi_i}, \ldots, \frac{\partial \ell_k(\xi)}{\partial \phi_p}, \frac{\partial \ell_k(\xi)}{\partial \sigma^2} \right)'$. To normalize the scores, we need the information matrix which in this case has the block-diagonal form

$$I(\xi) = \left(-\mathbb{E} \left[\frac{\partial^2 \ln f(Y, \xi)}{\partial \xi_i \partial \xi_j} \right] \right)_{i,j=1}^{p+1} = \begin{pmatrix} \frac{1}{\sigma^2}(1 - \sum_{j=1}^{p} \phi_j)^2 & 0 & 0 \\ 0 & \frac{1}{\sigma^2}\Gamma & 0 \\ 0 & 0 & \frac{1}{2\sigma^2} \end{pmatrix}, \tag{2}$$

where Γ is a covariance matrix of the vector $(Y_1, \ldots, Y_p)'$. The special form of the information matrix allows us to test the change in all the parameters at once or in a smaller group of parameters assuming that all the other parameters are not changed.
Denote $\hat{\xi}_n$ the estimate of ξ based on the whole observed sequence of length n. Then, the efficient score test statistic is

$$\hat{B}(u) = n^{-1/2} I^{-1/2}(\hat{\xi}_n) \nabla_\xi \ell_{[nu]}(\hat{\xi}_n), \quad 0 \le u \le 1, \tag{3}$$

where $[x]$ is the integer part of x.
Although, we derived the statistic $\hat{B}(u)$ under the assumption of Gaussian white noise, it is not a necessary condition. When we replace the normality of errors by appropriate conditions on the moments of errors, the likelihood function will become quasi-likelihood and the test will still be valid.
Let us formulate the following assumptions:

(A1) $\{\varepsilon_i\}$ is a sequence of i.i.d. random variables with zero mean and variance σ^2.

(A2) $\{\varepsilon_i\}$ is a m.d.s., where $\mathbb{E}[\varepsilon_i^2 | \mathcal{F}_{i-1}] = \sigma^2$, such that $\mathcal{F}_{i-1} = \sigma(\varepsilon_s, s \le i-1)$.

In [5] the following theorem is proven:

Theorem 2.1. *Let $\{Y_i\}$ be a sequence satisfying model AR(p), where $\{\varepsilon_i\}$ satisfies (A1) and $\mathbb{E}|\varepsilon_i|^\kappa < \infty$ for some $\kappa > 4$. Furthermore, assume that characteristic polynomial $\phi(z) = 1 - \sum_{j=1}^{p} \phi_j z^j$ has roots outside the unit circle. Then there exists a $(p+2)$-dimensional Gaussian process $\boldsymbol{B}(u)$ with independent Brownian bridge components $B^{(j)}(u)$, $j = 1, 2, \ldots, p+2$, such that*

$$\max_{1\le j\le p+2} \sup_{0\le u\le 1} |\hat{B}^{(j)}(u) - B^{(j)}(u)| = o_p(1).$$

The assumption (A1) can be replaced by (A2) and similar theorem can be proved by following the steps of the proof in [5] and using corresponding limit theorems for martingale difference sequences.

Proposition 2.1. *Let $\{Y_i\}$ be a AR(p) process with a stationary, ergodic m.d.s. $\{\varepsilon_i\}$ that satisfies (A2) and $\mathbb{E}|\varepsilon_i|^\kappa < \infty$ for some $\kappa > 4$. Moreover, let the characteristic polynomial $\phi(z) = 1 - \sum_{j=1}^{p} \phi_j z^j$ have roots outside the unit circle. Then there exists a $(p+2)$-dimensional Gaussian process $\boldsymbol{B}(u)$ with independent Brownian bridge components $B^{(j)}(u)$, $j = 1, 2, \ldots, p+2$, such that*

$$\max_{1\le j\le p+2} \sup_{0\le u\le 1} |\hat{B}^{(j)}(u) - B^{(j)}(u)| = o_p(1).$$

We say that there is a change in the parameter ξ_j for some $j = 1, \ldots, p+2$ if

$$\sup_{0\le u\le 1} |\hat{B}^{(j)}(u)| \ge C(\alpha),$$

where $C(\alpha)$ is a critical value corresponding to a significance level α gained from the properties of supremum of Brownian bridge

$$\mathrm{P}\left(\sup_{0\le t\le 1} |B(t)| \ge x\right) = 2\sum_{k=1}^{\infty} (-1)^{k+1} \exp\{-2k^2x^2\}.$$

Critical values can be found in some statistical tables, for example in [7].
We reject the null hypothesis (there is a change in one or more parameters) on a significance level α if the following inequality holds

$$\max_{1\le j\le p+2} \sup_{0\le u\le 1} |\hat{B}^{(j)}(u)| \ge C(\alpha^*),$$

where $\alpha^* = 1 - (1-\alpha)^{1/(p+2)}$, because we are testing $(p+2)$ parameters for a possible change.

3 Simulation Study

Firstly, we simulate the AR sequences under the null hypothesis and we use the score test statistic to detect the change in one of the parameters. We count how many times the test statistic indicate a change-point, even if there is no change in the parameter values for both cases (A1) and (A2) and compare the results.
Then, we look at the power of this test when the value of one of the parameters change and again compare the results for AR time series with (A1) and (A2) assumption.
The significance level is always set to be $\alpha = 0.05$.

We generate 1000 sequences of the AR(1) process with i.i.d. errors and 1000 realizations of AR(1) process with m.d.s. errors. We set the parameters for both processes as follows $\xi = (\mu, \phi_1, \sigma^2)' = (5, 0.25, 5)'$ and the length of the generated sequences is $n = 300$ (the process of length 400 is generated and the first 100 observations are discarded to gain the stationary process). Then we test this sequences for changes in any of the parameters. In Table 1 we can see the relative number of rejections of the null hypothesis even if it holds.

Significance level	AR(1) under (A1)	AR(1) under (A2)
0.05	0.046	0.052

Table 1: The relative number of falsely rejected null hypothesis for AR(1) process on a significance level 0.05.

Next, we would like to know the power of our test. We simulated one change in the parameter values of AR(1) sequence at fixed time and we watched how many times the change was detected. Again, we studied both cases (A1) and (A2). The results are in Table 2 and Table 3. First column is the time when the parameter changed (change-point). The other columns show the relative number of detected changes when the parameters changed from ξ to ξ^* at time k. There are two types of changes, the first one is relatively small compared to the values of ξ and the other one is relatively big. Table 2 corresponds to the assumption (A1) and Table 3 corresponds to (A2).

change-point	$\xi = (5, 1/4, 5)'$	$\xi = (5, 1/4, 5)'$
k	$\xi^* = (6, 1/3, 16/3)'$	$\xi^* = (8, 1/2, 6)'$
75	0.082	0.788
150	0.212	0.999
225	0.125	0.985

Table 2: The relative number of detected changes in AR(1) process under the assumption (A1).

change-point	$\xi = (5, 1/4, 5)'$	$\xi = (5, 1/4, 5)'$
k	$\xi^* = (6, 1/3, 16/3)'$	$\xi^* = (8, 1/2, 6)'$
75	0.070	0.643
150	0.093	0.993
225	0.073	0.960

Table 3: The relative number of detected changes in AR(1) process under the assumption (A2).

Detection of small changes in the parameters of AR(1) process is much more difficult for processes with m.d.s. white noise. Further, we see that the score test is strongest when the change appears in the middle of observed sequences. As expected, moving the change-point to the beginning or end of the sequences caused lowering the ability to detect a change.

4 Conclusion

Presented generalization of the efficient score statistic gives the opportunity to detect change-points in more general models than common AR(p) model. As the simulation study shows, the number of rejected null hypothesis, if there is no change, is approximately the same as the significance level α for both cases. From the second part of the simulation study it seems, that replacing (A1) by (A2) lowered the power of the score test.

Acknowledgements: This work was supported by grants SVV-2013-267-315 and GAUK 586721.

Bibliography

[1] Csörgö, M. and Horváth, L. (1997). *Limit Theorems in Change-Point Analysis*, Chichester: Wiley.

[2] Davidson, J. (1994). *Stochastic Limit Theory: Advanced Texts in Econometrics*, Oxford University Press, USA.

[3] Davis, R.A., Huang, D. and Yao, Y.-C. (1995). Testing for a Change in the Parameter Value and Order of an Autoregressive Model. *Ann. Statist.*, 23, 282-304.,

[4] Eberlein, E. (1986). On Strong Invariance Principles Under Dependence Assumptions. *Ann. Probab.*, 14, 260-270.

[5] Gombay, E. (2008). Change Detection in Autoregressive Time Series. *J. Multivar. Anal.*, 99, 451-464.

[6] Hušková, M., Prášková, Z. and Steinebach, J. (2007). On the Detection of Changes in Autoregressive Time Series I & II. *J. Stat. Plann. Inference*, 137.

[7] Owen, D.B. (1966). *Handbook of Statistical Tables, Russian title: Sbornik statisticzeskich tablic*, Moscow.

[8] Page, E.S. (1954). Continuous Inspection Schemes. *Biometrika*, 41, 100-105.

Some properties of a class of continuous time moving average processes

Andreas Basse-O'Connor[*]

Department of Mathematics, Aarhus University, Denmark

Abstract

A continuous time moving average is a process $X = \{X_t : t \in \mathbb{R}_+\}$ of the form

$$X_t = \int_{-\infty}^{t} \phi(t-s)\, dZ_s \tag{1}$$

where $\phi : \mathbb{R}_+ \to \mathbb{R}$ is a deterministic function and $Z = \{Z_t : t \in \mathbb{R}\}$ is a process with stationary and independent increments (a so-called Lévy process). In the case where the kernel function ϕ is the gamma density, i.e. $\phi(t) = e^{-\lambda t} t^{\gamma-1}$ where $\gamma, \lambda > 0$, we derive necessary and sufficient conditions on γ, λ for X to have sample paths of finite variation, or more generally, to be a semimartingale.

Keywords: Finite variation, Semimartingales, Moving averages, Gamma density
AMS subject classifications: 60G48, 60H05, 60G51, 60G17

1 Introduction

In discrete time, moving average processes play an important role in time series analysis. A moving average is a process of the form $X_n = \sum_{k=-\infty}^{n} \phi_{n-k} Z_k$ where $\{\phi_k\}_{k\in\mathbb{Z}_+}$ is a deterministic sequence of real numbers and $\{Z_k\}_{k\in\mathbb{Z}}$ is a sequence of independent and identically distributed random variables. Their continuous time analogue (called continuous time moving averages) are processes $X = \{X_t : t \in \mathbb{R}_+\}$ of the form

$$X_t = \int_{-\infty}^{t} \phi(t-s)\, dZ_s \tag{2}$$

where $\phi : \mathbb{R}_+ \to \mathbb{R}$ is a deterministic function and $Z = \{Z_t : t \in \mathbb{R}\}$ is a process with stationary and independent increments (i.e. a Lévy process). As usual we will assume that Z has right-continuous sample paths and $Z_0 = 0$. Suppose moreover that Z has no Gaussian component. The process Z is completely determined (in law) by its shift parameter $b \in \mathbb{R}$ and its Lévy measure ν from the Lévy–Khintchine representation

$$\mathbb{E}[e^{i\theta Z_1}] = \exp\Big(ibt + \int_{\mathbb{R}} \big(e^{i\theta x} - 1 - i\theta [\![x]\!]\big)\, \nu(dx)\Big), \qquad \theta \in \mathbb{R}, \tag{3}$$

where ν is a Borel measure on $\mathbb{R}$ satisfying $\int_{\mathbb{R}} (|x|^2 \wedge 1)\, \nu(dx) < \infty$ and $[\![x]\!] := x/(\max\{1, |x|\})$, $x \in \mathbb{R}$, is a truncation function. We will assume that Z is non-deterministic in the sense that $\nu(\mathbb{R}) > 0$. All stochastic integrals are defined as in Rajput and Rosiński [5, page 460].

[*] e-mail: basse@imf.au.dk

A function $f\colon \mathbb{R} \to \mathbb{R}$ is said to be of finite variation if $\mathrm{V}(f;[a,b]) < \infty$ for all finite $a < b$ where

$$\mathrm{V}(f;[a,b]) = \sup_{\substack{n\in\mathbb{N} \\ a=t_0<\cdots<t_n=b}} \sum_{i=1}^{n} |f(t_i) - f(t_{i-1})|. \tag{4}$$

A key example of a function of finite variation is an absolutely continuous function, that is, where f satisfies

$$f(t) = f(u) + \int_u^t f'(s)\,ds, \qquad \text{for all } u,t \in \mathbb{R},\ u \le t \tag{5}$$

for some locally integrable function $f'\colon \mathbb{R} \to \mathbb{R}$. The following result, which is a special case of [3, Theorems 3.1 and 3.3], will be used to characterize the finite variation property of a class of moving averages described below.

Theorem 1.1 (Basse-O'Connor and Rosiński)**.** *Consider a continuous time moving average X given by* (2)*. If X has sample paths of finite variation a.s. and $\int_{[-1,1]} |x|\,\nu(dx) = \infty$, then ϕ is absolutely continuous with a derivative ϕ' satisfying*

$$\int_{|x|\le 1} \int_0^\infty \big(|\phi'(s)x| \wedge |\phi'(s)x|^2\big)\,ds\,\nu(dx) < \infty. \tag{6}$$

Conversely, if ν is concentrated on $[-1,1]$ and ϕ is absolutely continuous with a derivative ϕ' satisfying (6)*, then X has right-continuous sample paths of finite variation a.s.*

Besides the finite variation property we will study the more general semimartingale property. A stochastic process $Y = \{Y_t : t \in \mathbb{R}_+\}$ is called a semimartingale with respect to a filtration $\mathbb{F} = (\mathcal{F}_t)_{t\ge 0}$ if it has a decomposition of the form

$$Y_t = Y_0 + M_t + A_t, \tag{7}$$

where M is a right-continuous local martingale, A is a right-continuous adapted process with sample paths of finite variation and Y_0 is $\mathcal{F}_0$-measurable. We refer to [4] for more information on semimartingales, here we will just note that one can define a Lebesgue–Stieltjes integral (resp. an Itô integral) with respect to a stochastic process if and only if it has sample paths of finite variation (resp. is semimartingales). Both properties play a fundamental role in stochastic analysis and its applications.

In this note we will focus on moving averages X given by (2) where the kernel function ϕ is the density of a gamma distribution (up to a multiplicative constant), that is, for $\lambda, \gamma > 0$

$$\phi(t) = \phi_{\lambda,\gamma}(t) = e^{-\lambda t} t^{\gamma-1}, \qquad t > 0. \tag{8}$$

The gamma kernel (8) is used to model turbulence using *ambit processes*; see [1] and the reference therein. In Theorem 2.2 below we characterize the set of λ, γ and ν for which X has sample paths of finite variation, or more generally, is a semimartingale with respect to the filtration $\mathbb{F}^Z = (\mathcal{F}_t^Z)_{t\ge 0}$ given by

$$\mathcal{F}_t^Z = \sigma(Z_s : s \in (-\infty, t]) \vee \mathcal{N}, \tag{9}$$

where $\mathcal{N}$ is the set of all null sets.

2 Main results

Throughout this section $X = \{X_t : t \ge 0\}$ is a continuous time moving average process given by

$$X_t = \int_{-\infty}^{t} \phi_{\lambda,\gamma}(t-s)\,dZ_s \tag{10}$$

where Z is a Lévy process without Gaussian component given by (3) with Lévy measure ν and $\phi_{\lambda,\gamma}$ is the gamma density given by (8). Our main results, Theorems 2.1 and 2.2, are stated below and are followed by three examples. The proofs are postponed to the next section.

Theorem 2.1. *Process X given by (10) is well-defined, i.e. the stochastic integrals exists, if and only if the following two conditions are satisfied:*

(i) $\int_{|x|\geq 1} \log(|x|)\,\nu(dx) < \infty$,

(ii) One of the following (a)–(c) are satisfied:

(a) $\gamma > \frac{1}{2}$,

(b) $\gamma = \frac{1}{2}$ *and* $\int_{|x|\leq 1} |x|^2 |\log(|x|)|\,\nu(dx) < \infty$,

(c) $\gamma \in (0, \frac{1}{2})$ *and* $\int_{|x|\leq 1} |x|^{1/(1-\gamma)}\,\nu(dx) < \infty$.

The next result gives an explicit characterization of when X has sample paths of finite variation or is a semimartingale. In the case $\gamma = 1$, X is a Lévy driven Ornstein–Uhlenbeck process and hence a semimartingale with respect to $\mathbb{F}^Z$. Moreover, X has sample paths of finite variation if and only if Z has sample paths of finite variation (i.e. $\int_{[-1,1]} |x|\,\nu(dx) < \infty$). Thus in the following result we will assume that $\gamma \neq 1$.

Theorem 2.2. *Let X be a stochastic process given by (10) with $\gamma \neq 1$. Then the following three conditions are equivalent:*

(I) X is a semimartingale with respect to $\mathbb{F}^Z$,

(II) X has right-continuous sample paths of finite variation with probability one,

(III) One of the following (A)–(C) are satisfied:

(A) $\gamma > \frac{3}{2}$,

(B) $\gamma = \frac{3}{2}$ *and* $\int_{|x|\leq 1} |x|^2 |\log(|x|)|\,\nu(dx) < \infty$,

(C) $\gamma \in (1, \frac{3}{2})$ *and* $\int_{|x|\leq 1} |x|^{1/(2-\gamma)}\,\nu(dx) < \infty$.

Theorem 2.2 shows that when $\gamma \neq 1$ then X is a semimartingale if and only if it is of finite variation. Below we use Theorems 2.1 and 2.2 to study three examples.

Example 2.1. *Suppose that Z is a symmetric α-stable Lévy process with $\alpha \in (0, 2)$, i.e. $\nu(dx) = c|x|^{-1-\alpha}\,dx$ where $c > 0$. Then X is well-defined if and only if $\gamma > (\alpha - 1)/\alpha$, and it is a semimartingale with respect to $\mathbb{F}^Z$ if and only if $\gamma > (2\alpha - 1)/\alpha$ or $\gamma = 1$.*

Example 2.1 may be viewed as a natural generalization of the case where Z is a Gaussian Lévy process (i.e. a Brownian motion). In this case Z is a 2-stable Lévy process and X is a semimartingale with respect to $\mathbb{F}^Z$ if and only if $\gamma = 1$ or $\gamma > \frac{3}{2}$; see [1].

Example 2.2. *Suppose that Z is a normal inverse Gaussian Lévy process. In this case $\nu(dx) = f(x)\,dx$ where $f(x) \sim c|x|^{-2}$ as $x \to 0$ and f decays exponential fast as $|x| \to \infty$. The process X is well-defined for all $\gamma, \lambda > 0$, and it is a semimartingale with respect to $\mathbb{F}^Z$ if and only if $\gamma \geq 1$.*

Example 2.3. *Suppose that Z is an inverse Gaussian Lévy process, i.e.*

$$\nu(dx) = ce^{-rx}x^{-3/2}\mathbf{1}_{\{x>0\}}\,dx$$

where $c, r > 0$. For all $\gamma, \lambda > 0$, X is well-defined and is a semimartingale with respect to $\mathbb{F}^Z$.

3 Proofs

For notation simplicity we will suppress λ and γ in $\phi_{\lambda,\gamma}$, that is, $\phi(t) = e^{-\lambda t} t^{\gamma-1}$ for $t \geq 0$.

Proof of Theorem 2.1. Recall that $[\![x]\!] := x/(\max\{1, |x|\})$. The stochastic integral (10) is well-defined if and only if the following two conditions are satisfied (cf. [5, Theorem 2.7])

$$\int_{\mathbb{R}} \int_0^\infty \big(|\phi(s)x|^2 \wedge 1\big)\, ds\, \nu(dx) < \infty, \tag{11}$$

$$\int_0^\infty \Big| b\phi(s) + \int_{\mathbb{R}} \big([\![x\phi(s)]\!] - \phi(s)[\![x]\!]\big)\, \nu(dx) \Big|\, ds < \infty. \tag{12}$$

Note that $\phi(s)/s^{\gamma-1} \to 1$ as $s \to 0$ and there exist two constants $c_1, c_2 > 0$ such that $c_1 e^{-2\lambda s} \leq \phi(s) \leq c_2 e^{-(\lambda/2)s}$ for all $s \geq 1$. Hence the inner integral in (11) is bounded from above and below by constants times

$$\underbrace{\int_0^1 \big(|s^{\gamma-1}x|^2 \wedge 1\big)\, ds}_{I_1(x)} + \underbrace{\int_1^\infty \big(|e^{-as}x|^2 \wedge 1\big)\, ds}_{I_2(x)} \tag{13}$$

where $a = 2\lambda$ in the lower bound and $a = \lambda/2$ in the upper bound. Calculating the integral $I_2(x)$ yields that $I_2(x)$ is bounded from above and below by constants times

$$\log(|x|)\mathbf{1}_{\{|x|>e\}} + |x|^2\mathbf{1}_{\{|x|\leq e\}}. \tag{14}$$

A similar calculation shows that $I_1(x)$ is bounded from above and below by constants times

$$\begin{cases} |x|^2\mathbf{1}_{\{|x|\leq 1\}} + \mathbf{1}_{\{|x|>1\}} & \gamma > \frac{1}{2} \\ |x|^2|\log(|x|)|\mathbf{1}_{\{|x|\leq 1/2\}} + \mathbf{1}_{\{|x|>1/2\}} & \gamma = \frac{1}{2} \\ |x|^{1/(1-\gamma)}\mathbf{1}_{\{|x|\leq 1\}} + \mathbf{1}_{\{|x|>1\}} & \gamma \in (0, \frac{1}{2}). \end{cases} \tag{15}$$

This shows that ϕ satisfies (11) if and only if (i)–(ii) of Theorem 2.1 are satisfied. Furthermore, if (i)–(ii) are satisfied then a similar calculation shows that (12) is satisfied. This completes the proof. □

To prove Theorem 2.2 we will need the following remark:

Remark 3.1. *For $\gamma > 1$ the function ϕ is absolutely continuous with a derivative $\phi'(t) = e^{-\lambda t}((\gamma - 1)t^{\gamma-2} - \lambda t^{\gamma-1})$. We have that $\phi'(t)/t^{\gamma-2} \to (\gamma - 1)$ as $t \to 0$ and $c_1 e^{-2\lambda t} \leq |\phi'(t)| \leq c_2 e^{-(\lambda/2)t}$ for all $t \geq 1$, where $c_1, c_2 > 0$ are two constants. These estimates show that the inner integral in (6) is bounded from above and below by constants times*

$$\int_0^1 \big(|s^{\gamma-2}x| \wedge |s^{\gamma-2}x|^2\big)\, ds + \int_1^\infty \big(|e^{-as}x| \wedge |e^{-as}x|^2\big)\, ds \tag{16}$$

where $a = 2\lambda$ in the lower bound and $a = \lambda/2$ in the upper bound. Thus by a similar calculation as used in the proof of Theorem 2.1, it follows that ϕ satisfied (6) if and only if (III) of Theorem 2.2 holds.

Proof of Theorem 2.2. If $\gamma \in (0, 1)$ then $\lim_{t\downarrow 0} \phi(t) = \infty$ and by [7, page 86], X has unbounded sample paths on each compact interval with strictly positive probability, which exclude that X satisfies (I) or (II). Thus we may and do assume that $\gamma > 1$. The implication (II) → (I) is true in general, and the implication (II) → (I) follows by [2, Theorem 4.1]. That is, we need to show the equivalence between (II) and (III) under the assumption $\gamma > 1$.
(II) → (III): Suppose that X has right-continuous sample paths of finite variation almost surely. If $\int_{[-1,1]} |x|\, \nu(dx) < \infty$ then (III) follows from the inequality $\int_{[-1,1]} |x|\, \nu(dx) \leq \int_{[-1,1]} |x|^{1/(2-\gamma)}\, \nu(dx)$. Therefore we may and

do assume that $\int_{[-1,1]} |x| \, \nu(dx) = \infty$. By Theorem 1.1, ϕ is absolutely continuous with a derivative ϕ' satisfying (6), which implies that (III) is satisfied cf. Remark 3.1.
(III) → (II): Suppose that (III) holds. According to the Lévy–Itô decomposition (see [8, Theorem 19.2]) there exists a decomposition $Z = Z^{\mathrm{s}} + Z^{\mathrm{l}}$ where Z^{s} and Z^{l} are two independent Lévy processes with Lévy measures $\nu^{\mathrm{s}} = \nu_{|[-1,1]}$ and $\nu^{\mathrm{l}} = \nu_{|[-1,1]^c}$, respectively. Decompose X as $X_t = X_t^{\mathrm{s}} + X_t^{\mathrm{l}}$ where

$$X_t^{\mathrm{s}} = \int_{-\infty}^{t} \phi(t-u) \, dZ_u^{\mathrm{s}} \qquad \text{and} \qquad X_t^{\mathrm{l}} = \int_{-\infty}^{t} \phi(t-u) \, dZ_u^{\mathrm{l}}. \tag{17}$$

The two integrals exist thanks to Theorem 2.1. By Remark 3.1, ϕ is absolutely continuous with a derivative ϕ' satisfying (6) and since ν^{s} is concentrated on $[-1,1]$, X^{s} has right-continuous sample paths of finite variation cf. Theorem 1.1. To show that X^{l} has a.a. sample paths of finite variation let $B = \{B_t : t \in \mathbb{R}\}$ be a process with $B_0 = 0$ satisfying $B_t - B_u = \mathrm{V}(Z^{\mathrm{l}}, (u,t])$ for all $u \le t$. Then B is an increasing Lévy process with Lévy measure ν_B given by $\nu_B(A) = \int_{\mathbb{R}} \mathbf{1}_A(|x|) \, \nu^{\mathrm{l}}(dx)$ for all Borel measurable sets $A \subseteq \mathbb{R}$ cf. [8, Theorem 21.9]. We extend ϕ to $\mathbb{R}$ by setting $\phi(t) = 0$ for $t < 0$. From the fact that ϕ is increasing on $(-\infty, \lambda/(\gamma-1)]$ and decreasing on $[\lambda/(\gamma-1), \infty)$ it follows that there exists a constant $c > 0$, depending only on t, γ and λ, such that

$$\mathrm{V}(\phi(\cdot - u); [0,t]) = \mathrm{V}(\phi; [-u, t-u]) \le c e^{(\lambda/2)u} \qquad \text{for all } u \in \mathbb{R}. \tag{18}$$

For all $0 = t_0 < \cdots < t_n = t$ we have

$$\sum_{i=1}^{n} |X_{t_i}^{\mathrm{l}} - X_{t_{i-1}}^{\mathrm{l}}| = \sum_{i=1}^{n} \Big| \int_{-\infty}^{t} \big(\phi(t_i - u) - \phi(t_{i-1} - u)\big) \, dZ_u^{\mathrm{l}} \Big| \tag{19}$$

$$\le \int_{-\infty}^{t} \mathrm{V}(\phi(\cdot - u); [0,t]) \, dB_u \le c \int_{-\infty}^{t} e^{(\lambda/2)u} \, dB_u \tag{20}$$

where the latter integral is finite a.s. due to the fact that $\int_{|x|\ge 1} \log(|x|) \, \nu_B(dx) < \infty$, see e.g. [6, Theorem 55 (i)]. Thus X^{l} has a.a. sample paths of finite variation. By Lebesgue's dominated convergence theorem it follows that X^{l} has continuous sample paths a.s. This completes the proof. □

Bibliography

[1] Barndorff-Nielsen, O. E. (2012). Notes on the gamma kernel. *Thiele Research Reports;* No. 03. Available at http://math.au.dk/publs?publid=946.

[2] Basse-O'Connor, A. and Rosiński, J. (2012). Structure of infinitely divisible semimartingales. arXiv:1209.1644v2 [math.PR].

[3] Basse-O'Connor, A. and Rosiński, J. (2013). Characterization of the finite variation property for a class of stationary increment infinitely divisible processes. *Stochastic Process. Appl.* 123(6), 1871–1890.

[4] Jacod, J. and Shiryaev, A. (2003). *Limit Theorems for Stochastic Processes*. Springer-Verlag, Berlin.

[5] Rajput, B. S. and Rosiński, J. (1989). Spectral representations of infinitely divisible processes. *Probab. Theory Related Fields* 82(3), 451–487.

[6] Rocha-Arteaga, A. and Sato, K. (2003). *Topics in Infinitely Divisible Distributions and Lévy Processes*. In: Aportaciones Matemáticas: Investigación 17, Sociedad Matemática Mexicana.

[7] Rosiński, J. (1989). On path properties of certain infinitely divisible processes. *Stochastic Process. Appl.* 33(1).

[8] Sato, K. (1999). *Lévy Processes and Infinitely Divisible Distributions*. Cambridge University Press.

Intervention in Ornstein-Uhlenbeck SDEs

Alexander Sokol[*]

University of Copenhagen, Denmark

Abstract

We introduce a notion of intervention for stochastic differential equations and a corresponding causal interpretation. For the case of the Ornstein-Uhlenbeck SDE, we show that the SDE resulting from a simple type of intervention again is an Ornstein-Uhlenbeck SDE. We discuss criteria for the existence of a stationary distribution for the solution to the intervened SDE. We illustrate the effect of interventions by calculating the mean and variance in the stationary distribution of an intervened process in a particularly simple case.

Keywords: Causality, Intervention, SDE, Ornstein-Uhlenbeck process, Stationary distribution.
AMS subject classifications: 60G15.

1 Introduction

Causal inference for continuous-time processes is a field in ongoing development. Similar to causal inference for graphical models, see [8], one of the primary objectives for causal inference for continuous-time processes is to estimate the effect of an intervention given assumptions on the distribution and causal structure of the observed continuous-time process.

Several flavours of causal inference are available for continuous-time processes, see for example [3], [4] and [9]. In this paper, we outline a notion of intervention for stochastic differential equations and a corresponding causal interpretation, we calculate the solution to an intervened Ornstein-Uhlenbeck SDE, and we calculate analytical expressions for the mean and variance of the stationary distribution of the resulting process for particular examples of interventions.

2 Causal interpretation of stochastic differential equations

Consider a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, P)$ satisfying the usual conditions, see [10] for the definition of this and other notions related to continuous-time stochastic processes. Let Z be a d-dimensional semimartingale and assume that $a : \mathbb{R}^p \to \mathbb{M}(p,d)$ is a Lipschitz mapping, where $\mathbb{M}(p,d)$ denotes the space of real $p \times d$ matrices. Consider the stochastic differential equation (SDE)

$$X_t^i = x_0^i + \sum_{j=1}^{d} \int_0^t a_{ij}(X_{s-})\,\mathrm{d}Z_s^j, \qquad i \leq p. \tag{1}$$

By the Lipschitz property of a, it holds by Theorem V.7 of [10] that there exists a pathwisely unique solution to (1). The following definition yields a causal interpretation of (1) based on simple substitution and inspired by ideas outlined in Section 4.1 of [1].

[*] e-mail: alexander@math.ku.dk

Definition 2.1. *Consider some $m \leq p$ and $c \in \mathbb{R}$. The $(p-1)$-dimensional intervened SDE arising from the intervention $X^m := c$ is defined to be*

$$U_t^i = x_0^i + \sum_{j=1}^{d} \int_0^t b_{ij}(U_{s-})\,\mathrm{d}Z_s^j \text{ for } i \leq p \text{ with } i \neq m, \tag{2}$$

where $b_{ij}(y_1, \ldots, y_{m-1}, y_{m+1}, \ldots, y_p) = a_{ij}(y_1, \ldots, c, \ldots, y_p)$, and the c is on the m'th coordinate. Letting U be the unique solution to the SDE and defining $Y = (U^1, \ldots, U^{m-1}, c, U^{m+1}, \ldots, U^p)$, we refer to Y as the intervened process and write $(X|X^m := c)$ for Y.

By Theorem V.16 and Theorem V.5 of [10], the solutions to both (1) and (2) may be approximated by the Euler schemes for their respective SDEs. Making these approximations and applying Pearl's notion of intervention in an appropriate sense, see [8], we may interpret Definition 2.1 as intervening in the system (1) under the assumption that the driving semimartingales $Z^1, \ldots, Z^d$ are noise processes unaffected by interventions, while the processes $X^1, \ldots, X^p$ are affected by interventions. Note that the operation of making an intervention takes a p-dimensional SDE as its input and yields a $(p-1)$-dimensional SDE as its output, and this operation is crucially dependent on the coefficients in the SDE: These coefficients in a sense corresponds to the directed acyclic graphs of [8]. A major benefit of causality in systems such as (1) as compared to the theory of [8] is the ability to capture feedback systems and interventions in such feedback systems.

As the solutions to (1) and (2) are defined on the same probability space, we may even consider the process $Y - X$, where $Y = (X|X^m := c)$, allowing us to calculate for example the variance of the effect of the intervention. As Y and X are never observed simultaneously in practice, however, we will concentrate on analyzing the differences between the laws of Y and X separately.

Recent developments related to causality for continuous-time processes have been focused on weak conditional local independence (WCLI), see for example [4]. A link between our notion of intervention and WCLI is the following: If X^i is equal to the intervened process $(X^i|X^m := c)$ for some c, then X^i is WCLI of X^m.

3 Intervention in Ornstein-Uhlenbeck SDEs

Recall that for an $\mathcal{F}_0$ measurable variable X_0 and for $A \in \mathbb{R}^p$, $B \in \mathbb{M}(p,p)$ and $\sigma \in \mathbb{M}(p,d)$, the Ornstein-Uhlenbeck SDE with initial value X_0, mean reversion level A, mean reversion speed B, diffusion matrix σ and d-dimensional driving noise is

$$X_t = X_0 + \int_0^t B(X_s - A)\,\mathrm{d}s + \sigma W_t, \tag{3}$$

where W is a d-dimensional $(\mathcal{F}_t)$ Brownian motion, see Section II.72 of [11]. The unique solution to this equation is $X_t = \exp(tB)(X_0 - \int_0^t \exp(-sB)BA\,\mathrm{d}s + \int_0^t \exp(-sB)\sigma\,\mathrm{d}W_s)$ where the matrix exponential is defined by $\exp(A) = \sum_{n=0}^{\infty} A^n/n!$. This is a Gaussian homogeneous Markov process with continuous sample paths. The following lemma shows that making an intervention in an Ornstein-Uhlenbeck SDE yields an SDE whose nontrivial coordinates solve another Ornstein-Uhlenbeck SDE.

Lemma 3.1. *Consider the Ornstein-Uhlenbeck SDE (3) with initial value x_0. Fix $m \leq p$ and $c \in \mathbb{R}$, and let X be the unique solution to (3). Let $Y = (X|X^m := c)$ and let Y^{-m} be the $p-1$ dimensional process obtained by removing the m'th coordinate from Y. Let $\tilde{B}$ be the submatrix of B obtained by removing the m'th row and column of B, and assume that $\tilde{B}$ is invertible. Then Y^{-m} solves*

$$Y_t^{-m} = y_0 + \int_0^t \tilde{B}(Y_s^{-m} - \tilde{A})\,\mathrm{d}s + \tilde{\sigma} W_t, \tag{4}$$

where y_0 is obtained by removing the m'th coordinate from x_0, $\tilde{\sigma}$ is obtained by removing the m'th row of σ and $\tilde{A} = \alpha - \tilde{B}^{-1}\beta$, where α and β are obtained by removing the m'th coordinate from A and from the vector whose i'th component is $b_{im}(c - a_m)$, respectively, where b_{im} is the entry corresponding to the i'th row and the m'th column of B, and a_m is the m'th element of A.

Proof. By Definition 2.1, $Y_t^i = y_0 + \int_0^t b_{im}(c - a_m) + \sum_{j \neq m} b_{ij}(Y_s^j - a_j)\,\mathrm{d}s + \sum_{j=1}^p \sigma_{ij} W_t^j$ for $i \neq m$. Note that for any vector y, the system of equations in $\tilde{a}$

$$b_{im}(c - a_m) + \sum_{j \neq m} b_{ij}(y_j - a_j) = \sum_{j \neq m} b_{ij}(y_j - \tilde{a}_j) \text{ for } i \neq m, \tag{5}$$

is equivalent to the system of equations

$$\sum_{j \neq m} b_{ij}\tilde{a}_j = \left(\sum_{j \neq m} b_{ij} a_j \right) - b_{im}(c - a_m) \text{ for } i \neq m, \tag{6}$$

which, since we have assumed $\tilde{B}$ to be invertible, has the unique solution $\tilde{A} = \tilde{B}^{-1}(\tilde{B}\alpha - \beta) = \alpha - \tilde{B}^{-1}\beta$. For $i \neq m$, we then obtain $Y_t^i = y_0 + \int_0^t \sum_{j \neq m} b_{ij}(Y_s^j - \tilde{a}_j)\,\mathrm{d}s + \sum_{j=1}^p \sigma_{ij} W_t^j$, proving the result. □

Recall that a principal submatrix of a matrix is a submatrix with the same rows and columns removed. In words, Lemma 3.1 states that if a particular principal submatrix $\tilde{B}$ of the mean reversion speed is invertible, then making the intervention $X^m := c$ in an Ornstein-Uhlenbeck SDE results in a new Ornstein-Uhlenbeck SDE with mean reversion speed $\tilde{B}$ and modified mean reversion level involving the inverse of $\tilde{B}$. Now assume that an Ornstein-Uhlenbeck SDE is given such that the solution has a stationary initial distribution. A natural question to ask is what interventions will yield intervened processes where stationary initial distributions also exist. In the following, we consider this question.

Recall that a square matrix is called stable if its eigenvalues have negative real parts and semistable if its eigenvalues have nonpositive real parts, see [2]. Theorem 4.1 of [12] yields necessary and sufficient criteria for the existence of a stationary probability measure for the solution of (3). One criterion is expressed in terms of the controllability subspace of of the matrix pair (B, σ), which is the span of the columns in the matrices $\sigma, B\sigma, \ldots, B^{p-1}\sigma$. In the case where σ has full column span, meaning that the columns of σ span all of $\mathbb{R}^p$, the controllability subspace is all of $\mathbb{R}^p$, and Theorem 4.1 of [12] shows that the existence of a stationary probability measure is equivalent to B being stable. The case where σ is not required to have full column span is more involved.

In the following, we will restrict our attention to Ornstein-Uhlenbeck processes with σ having full column span. By Theorem 4.1 of [12], it then holds that there exists a stationary distribution if and only if B is stable. Furthermore, applying Theorem 2.4 and Theorem 2.12 of [7], it holds in the affirmative case that the stationary distribution is the normal distribution with mean μ and variance Γ solving $B\mu = BA$ and $\sigma\sigma^t + B\Gamma + \Gamma B^t = 0$. Note that as B is stable, zero is not an eigenvalue of B, thus B is invertible and $\mu = A$. Also, stability of B yields that $\Gamma = \int_0^\infty e^{sB}\sigma\sigma^t e^{sB^t}\,\mathrm{d}s$. For the $(p-1)$-dimensional Ornstein-Uhlenbeck process resulting from an intervention according to Lemma 3.1, the diffusion matrix $\tilde{\sigma}$ is obtained by removing the m'th row of σ. As the columns of σ span $\mathbb{R}^p$, the columns of $\tilde{\sigma}$ span $\mathbb{R}^{p-1}$. Therefore, it also holds for the intervened process that there exists a stationary distribution if and only if the mean reversion speed is stable. We conclude that for diffusion matrices with full column span, the existence of stationary distributions for both the original and the intervened SDE is determined solely by stability of the mean reversion speed matrix B and corresponding principal submatrices.

Consider a stable matrix B. It then holds that if all principal submatrices of B are stable, all interventions will preseve stability of the system. We are thus lead to the question of when a principal submatrix of a matrix is stable. In general, stability or semistability does not lead to stability or semistability of principal submatrices. There are, however, classes of matrices satisfying that all principal submatrices are stable. For example, by the inclusion principle for symmetric matrices, see Theorem 4.3.15 of [6], it follows that a principal submatrix of any symmetric stable matrix again is stable. In general, though, it is difficult to ensure that all principal submatrices are stable. However, there are criteria ensuring that all principal submatices are semistable. For example, Lemma 2.4 of [5] shows that if B is stable and sign symmetric, then all principal submatrices of B are semistable. Here, sign symmetry is a somewhat involved matrix criterion, it does however hold that any stable symmtric matrix also is sign symmetric. Furthermore, by Theorem 1 of [2], either of the following three properties are also sufficient for having all principal submatrices being semistable: that $A - D$ is stable for all nonnegative diagonal D, that DA is stable for all positive diagonal D, or that there is positive diagonal D such that $AD + DA^t$ is negative definite.

4 An example of a particular intervention

Consider now a three-dimensional Ornstein-Uhlenbeck process X with σ being the identity matrix of order three and upper diagonal mean reversion speed matrix B, and assume that the diagonal elements of B all are negative. As the diagonal elements of B in this case also are the eigenvalues, B and all of its principal submatrices are then stable. The interpretation of having B upper diagonal is that the levels of both X^1, X^2 and X^3 directly influence the average change in X^1, while only the levels of X^2 and X^3 directly influence the average change in X^2 and only X^3 directly influences the average change in X^3.

We will investigate the details of what happens to the system when making the interventions $X^2 := c$ or $X^3 := c$. To this end, we calculate the stationary mean and variance, that is, the mean and variance in the stationary distribution, for each of the intervened processes. Consider first the case of the intervention $X^2 := c$. Let μ and Γ denote the mean and variance in the stationary distribution after intervention. Applying Lemma 3.1, the result of making this intervention is an Ornstein-Uhlenbeck process with mean reversion speed and mean reversion level

$$\left[\begin{array}{cc} b_{11} & b_{13} \\ 0 & b_{33} \end{array}\right] \quad \text{and} \quad \left[\begin{array}{c} a_1 \\ a_3 \end{array}\right] - \left[\begin{array}{cc} b_{11} & b_{13} \\ 0 & b_{33} \end{array}\right]^{-1} \left[\begin{array}{c} b_{12}(c-a_2) \\ 0 \end{array}\right]. \tag{7}$$

By explicit calculations, we obtain

$$\mu = \left[\begin{array}{c} a_1 - \frac{b_{12}}{b_{11}}(c-a_2) \\ a_3 \end{array}\right] \text{ and } \Gamma = \left[\begin{array}{cc} -\frac{1}{2b_{11}} - \frac{b_{13}^2}{2b_{11}b_{33}(b_{11}+b_{33})} & \frac{b_{13}}{2b_{33}(b_{11}+b_{33})} \\ \frac{b_{13}}{2b_{33}(b_{11}+b_{33})} & -\frac{1}{2b_{33}} \end{array}\right]. \tag{8}$$

Next, considering the intervention $X^3 := c$, we let ν and Σ denote the mean and variance in the stationary distribution. By Lemma 3.1, the result of making this intervention is an Ornstein-Uhlenbeck process with mean reversion speed and mean reversion level

$$\left[\begin{array}{cc} b_{11} & b_{12} \\ 0 & b_{22} \end{array}\right] \quad \text{and} \quad \left[\begin{array}{c} a_1 \\ a_2 \end{array}\right] - \left[\begin{array}{cc} b_{11} & b_{12} \\ 0 & b_{22} \end{array}\right]^{-1} \left[\begin{array}{c} b_{13}(c-a_3) \\ b_{23}(c-a_3) \end{array}\right], \tag{9}$$

yielding by calculations similar to the previous case that

$$\nu = \left[\begin{array}{c} a_1 - \left(\frac{b_{13}}{b_{11}} - \frac{b_{12}b_{23}}{b_{11}b_{22}}\right)(c-a_3) \\ a_2 - \frac{b_{23}}{b_{22}}(c-a_3) \end{array}\right] \text{ and } \Sigma = \left[\begin{array}{cc} -\frac{1}{2b_{11}} - \frac{b_{12}^2}{2b_{11}b_{22}(b_{11}+b_{22})} & \frac{b_{12}}{2b_{22}(b_{11}+b_{22})} \\ \frac{b_{12}}{2b_{22}(b_{11}+b_{22})} & -\frac{1}{2b_{22}} \end{array}\right]. \tag{10}$$

We now interpret these results. In the original system, all of X^1, X^2 and X^3 negatively influenced themselves, and in addition to this, X^2 influenced X^1 and X^3 influenced X^1 both directly and through its influence on X^2. Based on this, we would expect that making the intervention $X^2 := c$, the stationary mean of X^3 would not be changed, while the stationary mean of X^1 would change, depending on the level of influence b_{12} of X^2 on X^1. This is what we see in (8). When making the intervention $X^3 := c$, however, we obtain a change in the stationary mean of X^1 based both on the direct influence of X^3 on X^1, depending on b_{13}, but also on the indirect influence of X^3 on X^1 through X^2, depending also on b_{23} and b_{12}. Furthermore, the stationary mean of X^2 also changes. This is what we see in (10).

As for the stationary variance, the changes resulting from interventions are in both cases of the same type, both independent of c. This implies that while we in most cases will be able to obtain any stationary mean for, say, X^1, by picking c suitably, the stationary variance is influenced only by the parts of the system for which the interventions are made. Furthermore, by considering explicit formulas for the stationary variance in the original system, it may be seen that for example positive covariances may turn negative and vice versa when making interventions.

Acknowledgements: The development of the notion of intervention for SDEs is joint work with my thesis advisor, Niels Richard Hansen, whom I also thank for valuable discussions and advice.

Bibliography

[1] Aalen, Odd O., Røysland, Kjetil; Gran, Jon Michael. Causality, mediation and time: A dynamic viewpoint. *J. R. Statist. Soc. A* 175 (2012), no. 4, 831–861.

[2] Cross, G. W. Three types of matrix stability. *Linear Algebra and Appl.* 20 (1978), no. 3, 253–263.

[3] Florens, Jean-Pierre; Fougere, Denis. Noncausality in continuous time. *Econometrica* 64 (1996), no. 5, 1195–1212.

[4] Gégout-Petit, Anne; Commenges, Daniel. A general definition of influence between stochastic processes. *Lifetime Data Anal.* 16 (2010), no. 1, 33–44.

[5] Hershkowitz, Daniel; Keller, Nathan. Positivity of principal minors, sign symmetry and stability. *Linear Algebra Appl.* 364 (2003), 105–124.

[6] Horn, Roger A.; Johnson, Charles R. Matrix analysis. *Cambridge University Press, Cambridge*, 1985.

[7] M. Jacobsen: A Brief Account of the Theory of Homogeneous Gaussian Diffusions in Finite Dimensions, in ”Frontiers in Pure and Applied Probability 1”, p. 86-94, 1993.

[8] Pearl, Judea. Causality. Models, reasoning, and inference. Second edition. *Cambridge University Press, Cambridge*, 2009.

[9] Petrović, Ljiljana; Dimitrijević, Sladjana. Invariance of statistical causality under convergence. *Statist. Probab. Lett.* 81 (2011), no. 9, 1445–1448.

[10] Protter, Philip E. Stochastic integration and differential equations. Second edition. Version 2.1. Corrected third printing. *Springer-Verlag, Berlin*, 2005.

[11] Rogers, L. C. G.; Williams, David. Diffusions, Markov processes, and martingales. Vol. 1. Foundations. *Cambridge University Press*, Cambridge, 2000.

[12] Zakai, Moshe; Snyders, Jakov. Stationary probability measures for linear differential equations driven by white noise. *J. Differential Equations* 8 (1970), 27–33.

Bayesian multiscale analysis of images

Leena Pasanen[*][1] **and Lasse Holmström**[1]

[1] *Department of Mathematical Sciences, University of Oulu, Finland*

Abstract

Two novel multiscale methods for digital images are proposed. The first method detects differences between two images obtained from the same object at two different instants of time. It detects both small scale, sharp changes and large scale, average changes. The second method extracts features that differ in intensity from their surroundings and produces a multiresolution analysis of an image as a sum of scale-dependent components.

As images are usually noisy, Bayesian inference is used to separate real differences and features from artefacts caused by random noise. The use of the Bayesian paradigm allows the use of various noise types, incorporation of expert knowledge about the images at hand and facilitates analysis of non-linear transformation of images.

The methods are instants of SiZer (Significant zero crossings of derivatives) methodology that was originally considered for one-dimensional nonparametric probability density estimation and curve fitting [1, 2]. The new methods, iBSiZer (Bayesian SiZer for images) and MRBSiZer (Multiresolution Bayesian SiZer), were originally proposed in [7] and [8], respectively.

Keywords: Bayesian methods, Scale space, Image analysis, SiZer
AMS subject classifications: 62M40

1 Introduction

When two images have been obtained from the same object at two different times, the changed areas can in principle be detected from the pixelwise difference image. However, as images usually contain noise, statistical methods are needed to separate the real changes from noise artefacts. In particular, large areas where the pixel intensity has changed only slightly may easily be masked by noise. If such areas are smoothed, the pixelwise noise variance can be reduced making the underlying signal easier to detect. Special challenges arise when one wants to detect changes in multispectral images, such as in the case of the Landsat ETM+ satellite. To facilitate change detection, the multidimensional data are often first transformed to make it one dimensional. The noise in the transformed image may then have a complicated structure making statistical inference challenging.

We propose a statistical method for the analysis of the transformed images that also allows detection of changes in many spatial scales. The changes are detected using Bayesian inference that facilitates the incorporation of expert information about the images at hand. Analysis in multiple spatial scales is achieved by employing many different smoothing levels.

The methods have their origin in SiZer technology first introduced in [1, 2] for the purposes of one-dimensional nonparametric probability density estimation and curve fitting. Since then it has been developed into various directions including Bayesian approaches, analysis of two-dimensional densities and images; see the review articles [5] and [6] and the references therein. The change detection method discussed here is called Bayesian SiZer for images, or iBSiZer [7].

[*] Corresponding author, e-mail: leena.pasanen@oulu.fi

Another closely related problem is the detection of image features that differ in intensity values from their surroundings. To solve this problem, we decompose the image into additive components that corresponds to features in different spatial scales. The resulting method is referred to as Multiresolution Bayesian SiZer, or MRBSiZer [8].

The Bayesian framework is presented in Section 2, the iBSiZer and MRBSiZer methods are outlined in Section 3, and example images are analyzed in Section 4.

2 Bayesian framework

A grayscale digital image can be considered as an $M \times N$ array of real numbers x_{ij}, but in mathematical derivations we treat it is as a vector $\mathbf{x} = [x_1, \ldots, x_n] \in \mathbb{R}^n, \;\; n = NM.$ Landsat ETM+ satellite images consist of eight image bands, each representing a different wavelength of light. We vectorize each $M \times N$ band image and then combine the bands corresponding to the two instants of time considered into one $16n \times 1$ vector $\mathbf{x} = [\mathbf{x}_{11}^T, \ldots, \mathbf{x}_{18}^T, \mathbf{x}_{21}^T, \ldots, \mathbf{x}_{28}^T]^T$, where $\mathbf{x}_{ij}$ is the band j at time i.

An observed image $\mathbf{y}$, satellite or otherwise, is modeled as

$$\mathbf{y} = \mathbf{x} + \boldsymbol{\varepsilon}, \tag{1}$$

where, $\mathbf{x}$ is the true image and $\boldsymbol{\varepsilon}$ is the corrupting noise. The posterior distribution of $\mathbf{x}$ given $\mathbf{y}$ is then

$$p(\mathbf{x}|\mathbf{y}) = \tfrac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})} \propto p(\mathbf{y}|\mathbf{x})p(\mathbf{x}), \tag{2}$$

where $p(\mathbf{y}|\mathbf{x})$ is the likelihood of $\mathbf{y}$ given $\mathbf{x}$ and $p(\mathbf{x})$ is the prior distribution of $\mathbf{x}$. The noise is assumed to have a Gaussian distribution $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$ and hence $p(\mathbf{y}|\mathbf{x})$ is a Gaussian density.

As the prior distribution of $\mathbf{x}$, we use a Gaussian smoothing prior that penalizes for image roughness as measured by the second differences of neighboring pixel intensities [7]. For Landsat ETM+ images, the prior models the prior temporal dependence in the images corresponding to the same band as well as the smoothness of each band image $\mathbf{x}_{ij}$. This prior model for Landsat ETM+ images is discussed in more detail in [9].

By substituting the Gaussian likelihood and the Gaussian smoothing prior one obtains a multivariate Gaussian posterior [7]. If the values of the parameters in prior or likelihood are unknown, one can use the empirical Bayes approach that estimates them from the data or alternatively use the fully Bayesian approach that treats them as random variables (see [7] and [9]).

3 Scale space analysis

IBSiZer detects credible changes between two images in many scales. The scales are defined by applying various degrees of smoothing to the images. However, instead of using the smooths of the observed images directly, change detection is based on the posterior distribution of the smooths of the true underlying difference image.

We use the roughness penalty smoother

$$\mathbf{S}_\lambda = (\mathbf{I} + \lambda \mathbf{Q})^{-1},$$

where $\mathbf{Q} = \mathbf{C}^T\mathbf{C}$, and $\mathbf{C}$ is the matrix that defines the second differences of neighboring pixels and $\mathbf{I}$ is the identity matrix. The parameter λ controls the smoothness of the smooth $\mathbf{S}_\lambda\mathbf{x}$ and we require that as $\lambda \to \infty$, $\mathbf{S}_\lambda\mathbf{x}$ approaches the mean of $\mathbf{x}$; see [7] for details.

Change detection is performed in three steps. The first step is to obtain the posterior distribution of the true underlying difference image. When changes are detected from Landsat ETM+ satellite images, the image dimensionality is first reduced using a transformation t so that instead of having a vector of length $16n$ to analyze, a vector of length n is analyzed. Examples of such a transformation are for example the temporal difference of vegetation indexes or the temporal difference of a certain image band. Different transformations give different information about the change and therefore, the transformation t needs to be chosen carefully to match the change type of interest. The likelihood function $p(t(\mathbf{y})|t(\mathbf{x}))$ can then be complex, whereas the likelihood function $p(\mathbf{y}|\mathbf{x})$ is simply a multivariate Gaussian. It is also easier to formulate one's prior knowledge about the images in terms of $\mathbf{x}$ rather than in terms of $t(\mathbf{x})$. Especially, the dependency between bands on different time points can be taken into account in the prior of $\mathbf{x}$. Therefore, instead of $p(t(\mathbf{x})|t(\mathbf{y}))$, we consider $p(t(\mathbf{x})|\mathbf{y})$. For digital grayscale images, the transformation t can be simply the difference of images $\mathbf{x}_1 - \mathbf{x}_2$. The posterior distribution $p(t(\mathbf{x})|\mathbf{y})$ can be analyzed by first drawing a sample from the posterior distribution (2) and then transforming each sampled image by t.

The second step is to obtain the posterior distribution of the smooths $p(\mathbf{S}_\lambda t(\mathbf{x})|\mathbf{y})$. In practice, this posterior is approximated by smoothing the images sampled in the first step. The third step is to use the smoothed sample images to detect the pixels that have high enough posterior probability to differ credibly from zero. The inference here is simultaneous over all pixels of the image and we use the "simultaneous credible intervals" (CI) method that was first proposed for one dimensional data in [3] and then extended for digital images in [8].

MRBSiZer considers a single image and aims to detect areas that differ in intensity from their neighborhood. This is accomplished by employing differences of image smooths. In the following, $\mathbf{x}$ can represent a grayscale image, a difference image or a transformed image.

Let $0 = \lambda_1 < \lambda_2 < \cdots < \lambda_{L-1} < \lambda_L = \infty$ be a set of smoothing levels. Then

$$\mathbf{x} = \sum_{i=1}^{L-1} (\mathbf{S}_{\lambda_i} - \mathbf{S}_{\lambda_{i+1}})\mathbf{x} + \mathbf{S}_{\lambda_L}\mathbf{x} \equiv \sum_{i=1}^{L-1} \mathbf{z}_i + \mathbf{z}_L, \tag{3}$$

where $\mathbf{z}_i = (\mathbf{S}_{\lambda_i} - \mathbf{S}_{\lambda_{i+1}})\mathbf{x}$ for $i = 1, \ldots, L-1$, and $\mathbf{z}_L = \mathbf{S}_\infty \mathbf{x}$. Here $\mathbf{z}_i$ for $i = 1, \ldots, L-1$ is the difference between two consecutive smooths and is referred to as the ith "detail" of the decomposition (3). It can be interpreted as the detail lost when smoothing is increased from λ_i to λ_{i+1}.

As in iBSiZer, the first step is to obtain the posterior distribution of $\mathbf{x}$ given the noisy data $\mathbf{y}$. The second step is to obtain the posterior distributions of the differences of smooths $\mathbf{z}_i = (\mathbf{S}_{\lambda_i} - \mathbf{S}_{\lambda_{i+1}})\mathbf{x}$ using the sample generated in the first step. Finally, the credibly nonzero image parts of each $\mathbf{z}_i$ are detected as in iBSiZer.

4 Experiments

To illustrate the idea of iBSiZer, we will first detect changes in a pair of images based on a real Landsat ETM+ satellite image and manually constructed changes, using the difference of their NDVI (Normalized difference vegetation index) images. For a pair of satellite images represented by $\mathbf{v} = [\mathbf{v}_{11}^T, \ldots, \mathbf{v}_{28}^T]^T$ the NDVI difference is computed as

$$\mathbf{N}_\mathbf{v} \equiv \tfrac{\mathbf{v}_{24} - \mathbf{v}_{23}}{\mathbf{v}_{24} + \mathbf{v}_{23}} - \tfrac{\mathbf{v}_{14} - \mathbf{v}_{13}}{\mathbf{v}_{14} + \mathbf{v}_{13}}.$$

The true and the noisy NDVI difference images are denoted by $\mathbf{N}_\mathbf{x}$ and $\mathbf{N}_\mathbf{y}$, respectively. The test image pair is constructed from a 176 × 165 subimage of a full Landsat ETM+ satellite image [7]. The images were first smoothed to obtain $\mathbf{x}_{13}$ and $\mathbf{x}_{14}$. The images $\mathbf{x}_{23}$ and $\mathbf{x}_{24}$ were then obtained by making manually changes to $\mathbf{x}_{13}$ and $\mathbf{x}_{14}$. Finally, the "observed images" $\mathbf{y}_{13}, \mathbf{y}_{14}, \mathbf{y}_{23}$, and $\mathbf{y}_{24}$ were constructed by adding Gaussian iid noise with variance 16 to each band. The noisy band images, the corresponding NDVI images, the NDVI-difference of the true and the noisy images are presented in Figure 1.

Assuming the model (1), we use a Gaussian smoothing prior for $\mathbf{x}$ where, in addition to spatial dependence, also temporal dependence is modeled; see [9] for more details. The resulting posterior mean, and the credibility maps with smoothing levels $[0,\ 1,\ 100]$ are displayed in Figure 1. The two smallest scale maps detect the small changed areas with high absolute intensity. The large positive area in the lower right corner is detected in the larger scale. Note that none of the maps alone would reveal all the interesting features. If simple thresholding would be applied to the noisy NDVI-difference image, the true changes would be masked by noise [9].

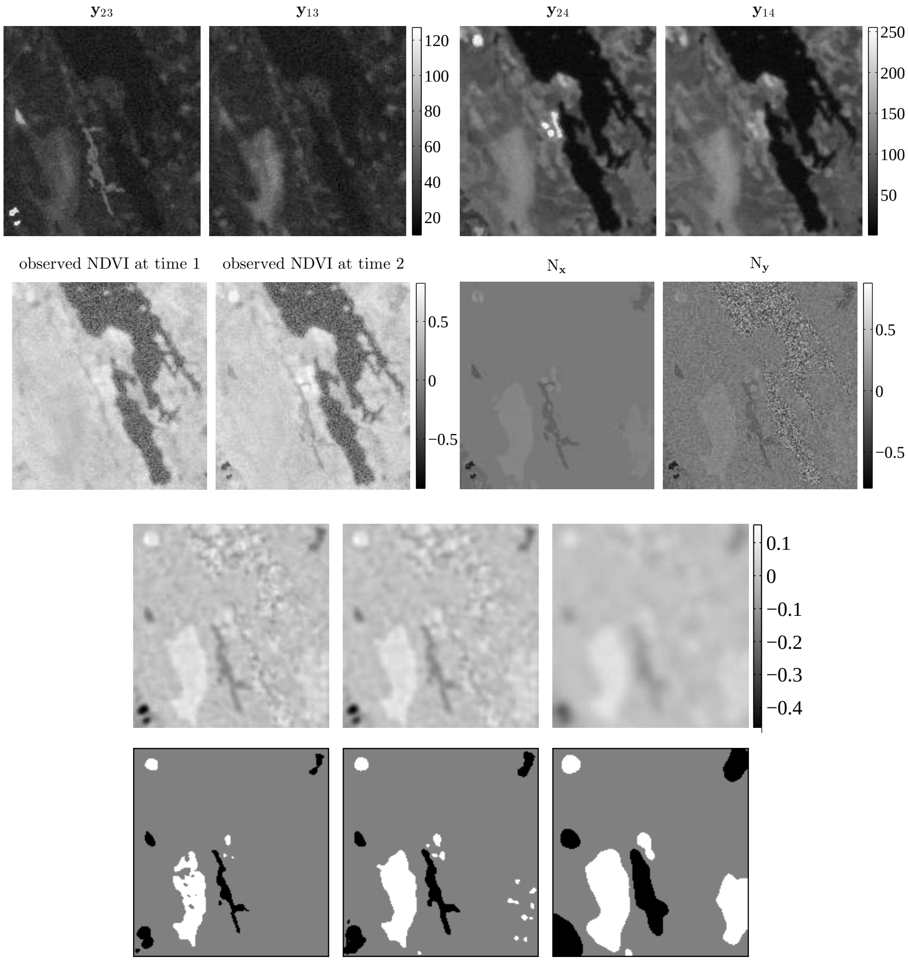


Figure 1: A partly artificially constructed pair of satellite images. 1st row: Noisy band images. 2nd row: Noisy NDVI-images, the true and the observed difference of NDVI-images. 3th row: Posterior means of $\mathbf{S}_\lambda \mathbf{N}_\mathbf{x}$, $\lambda = 0, 1, 100$. 4th row: Corresponding iBSiZer-maps.

Next, MRBSiZer is used to analyze predicted global climate change between 1980-2000 (present) and 2080-2100 (future). We had available a posterior sample of climate change fields from a hierarchical Bayesian model that combined predictions of several atmosphere-ocean general circulation models. For the details of the statistical model employed, see [4]. The mean of the sample is presented in the first row of Figure 2. The rest of the rows present the posterior means of the multiresolution details $\mathbf{z}_i$ (on the left) and the corresponding MRBSiZer maps (on the right). The smoother used in the scale space analysis operates on a sphere and is defined in [8]. The four rows represent a multiresolution decomposition of predicted global temperature rise into an overall rising mean (bottom panel), a north-south gradient that accounts for the bigger temperature increase in the northern hemisphere (second panel from the bottom) and a pattern of more complex small scale change that concentrates in the northernmost latitudes (the two uppermost panels).

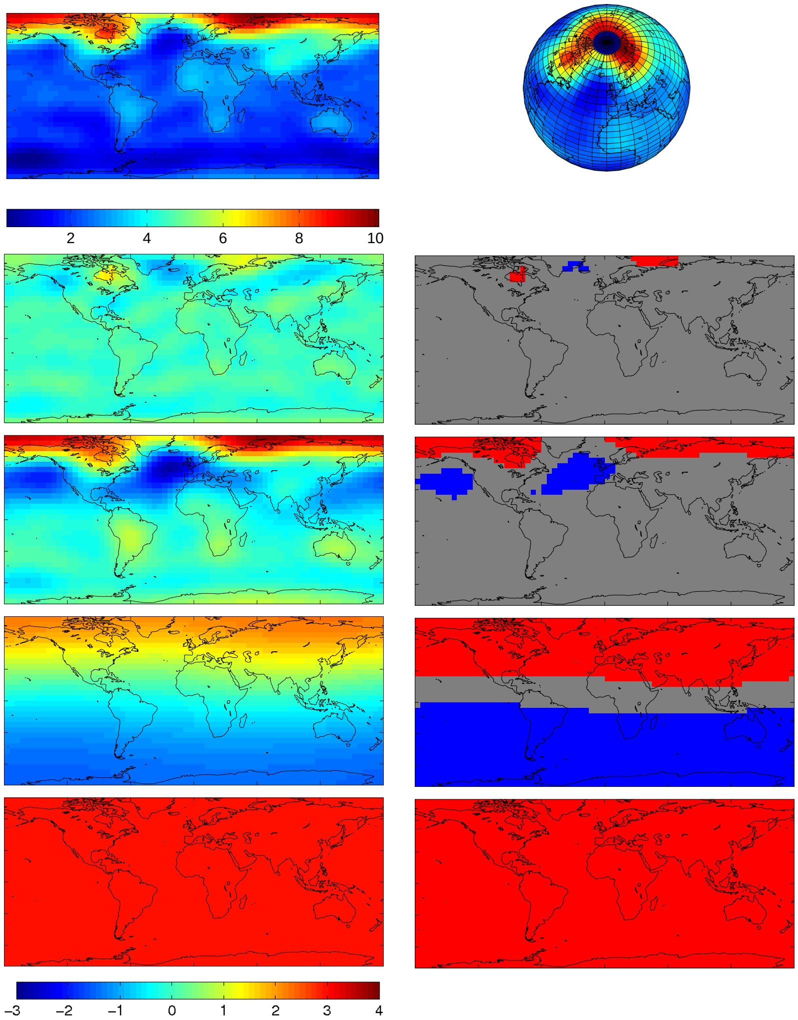


Figure 2: Scale space multiresolution analysis of predicted global climate chance. 1st row: Mean global temperature change field. Rows 2-4: MRBSiZer analysis of temperature change. The left hand column presents the multiresolution detail posterior means and the right hand column shows the corresponding credibility maps. Blue and red corresponds to negative (colder) and positive (warmer) changes respectively.

Bibliography

[1] Chaudhuri, P. and Marron, J. S. (1999). SiZer for Exploration of Structures in Curves. *Journal of the American Statistical Association* 94, 807–823.

[2] Chaudhuri, P. and Marron, J. S. (2000). Scale Space View of Curve Estimation. *The Annals of Statistics* 28, 408–428.

[3] Erästö, P. and Holmström, L. (2005). Bayesian Multiscale Smoothing for Making Inferences About Features in Scatter Plots. *Journal of Computational and Graphical Statistics*, 14, 569–589.

[4] Furrer,R., Sain, S. R., Nychka, D. W., and Meehl, G. A. (2007). Multivariate Bayesian Analysis of Atmosphere-Ocean General Circulation Models. *Environ. Ecol. Stat* 14,249–266.

[5] Holmström, L. (2010). Scale Space Methods.*Wiley Interdisciplinary Reviews: Computational Statistics* 2, 150–159. Available on-line at http://dx.doi.org/10.1002/wics.79.

[6] Holmström, L (2010). BSiZer. *Wiley Interdisciplinary Reviews: Computational Statistics* 2, 526–534. Available on-line at http://dx.doi.org/10.1002/wics.115.

[7] Holmström, L. and Pasanen, L. (2012). Bayesian Scale Space Analysis of Differences in Images. *Technometrics* 54, 16–29.

[8] Holmström, L., Pasanen, L., Furrer, R. and Sain, S. R. (2011). Scale Space Multiresolution Analysis of Random Signals. *Computational Statistics & Data Analysis* 55, 2840–2855.

[9] Pasanen, L. and Holmström, L. (2012) Bayesian Scale Space Analysis of Temporal Changes in Landsat ETM+ Satellite Images. Submitted for publication. Available on-line at http://cc.oulu.fi/~lpasanen/iBSiZer/ChangeDetection/ChangeDetection.pdf.

Multiresolution methods for ranking

Eric Sibony[*]

Institut Mines-Telecom, LTCI, Telecom Paristech / CNRS, France

Abstract

We use a recently introduced framework for multiresolution analysis on the symmetric group to predict rankings. Viewing preferences as sets of permutations, ranking prediction implies to handle probability distributions on the symmetric group, which usually leads to intractable storage or computations. We define a new smoothing technique based on wavelet decomposition that allows to obtain sparse representations for a large class of probability distributions. We show that in many practical cases, our method performs efficiently, in terms of storage and from a computational cost perspective as well.

Keywords: Ranking, Multiresolution analysis, Wavelets, Statistical estimation, Pairwise preferences.
AMS subject classifications: 62G05, 43A65.

1 Introduction

Ranking problems have been the subject of much attention these last few years in the machine-learning literature. The need to deal with orderings of items has indeed found more and more applications in modern technological devices such as recommendation systems or search engines, where they model for example preferences on movies or products, relevance of websites to a query, closeness of a relationship in a social network or the order of the words in automatic translation. They all fit in the following general model: the n items are labeled $1, 2, ..., n$, and each ordering is seen as a permutation σ over the set $\{1, ..., n\}$, where $\sigma(i)$ is the rank of the object i. Learning and prediction of rankings is then made through probabilistic modeling and statistical estimation on the set $\mathfrak{S}_n$ of all the permutations of the n items, the symmetric group. The latter's cardinal, equal to $n!$, rapidly exploding when n increases, any naive approach leads to intractable computations and even the simple storage of a probability vector becomes unfeasible as soon as $n \geq 15$, while the considered applications usually require to deal with $n \geq 10^5$ or more.

Hence, efficient inference methods necessarily require smoothing techniques so as to produce estimators with good statistical properties and compact representation both at the same time. Inspired by the remarkable achievements of traditional wavelet methods in signal and image processing, we develop the framework for multiresolution analysis on the symmetric group recently introduced in [3] to perform sparse estimation of probability distribution on $\mathfrak{S}_n$.

2 Mathematical setting and problem statement

We consider the problem of recovering a probability distribution p on $\mathfrak{S}_n$ from a set of observations. Let $L(\mathfrak{S}_n) = \{f : \mathfrak{S}_n \to \mathbb{R}\}$, $p \in L(\mathfrak{S}_n)$ is such that for all $\pi \in \mathfrak{S}_n$, $p(\pi) \geq 0$, and $\sum_{\pi \in \mathfrak{S}_n} p(\pi) = 1$. It typically models the variability of (a community of) customers complete preferences on the n items: if σ is a random permutation of law p, then the probability that a customer ranks the n items according to

[*] e-mail: eric.sibony@telecom-paristech.fr

a given ranking $\pi \in \mathfrak{S}_n$ is $\mathbb{P}[\sigma = \pi] = p(\pi)$. More generally, for any subset $S \subset \mathfrak{S}_n$, $\mathbb{P}[\sigma \in S] = \sum_{\pi \in S} p(\pi) = \langle p, \mathbf{1}_S \rangle$, where $\mathbf{1}_S$ denotes the indicative function of S and $\langle ., . \rangle$ is the usual inner product on $L(\mathfrak{S}_n)$. The observations take the form of subsets of $\mathfrak{S}_n$, modeling the preferences expressed by the customers: for instance, if a customer prefers the item i to the item j, the corresponding observation is the set of all orderings that rank i before j, *i.e.* $\{\pi \in \mathfrak{S}_n \mid \sigma(i) < \sigma(j)\}$, denoted by abuse $\{i \prec j\}$. Any type of observation fits in this model (see [6]).

In most cases of interest, each customer only expresses his preferences on a small subset of items, leading to very restricted observations and thus very limited information. The problem presents therefore three main challenges, in providing an estimator that has good statistical performance even with very limited information, compact representation and effective computation.

In this paper, we restrict ourselves to the case where the observations are pairwise comparisons, *i.e.* of the form $\{i \prec j\}$ with $i \neq j$. Thus we assume that we are given T pairwise comparisons $i_1 \prec j_1$, ..., $i_T \prec j_T$ drawn IID from an observation process defined as follows:

- each pair $\{i_t, j_t\}$ is drawn IID from a probability distribution μ on the pairs among $\{1, ..., n\}$ that is assumed to be known;
- knowing that $\{i_t, j_t\} = \{i, j\}$, the comparison is $i \prec j$ with probability $\langle p, \mathbf{1}_{\{i \prec j\}} \rangle =: p_{i \prec j}$ and $i \succ j$ with probability $p_{i \succ j} = 1 - p_{i \prec j}$.

In this setting, the empirical estimator is defined as the average of all the observations normalized to be a probability distribution:

$$\widehat{p} = \frac{1}{T} \sum_{t=1}^{T} \frac{2}{n!} \mathbf{1}_{\{i_t \prec j_t\}}. \tag{1}$$

Let M be the linear operator of $L(\mathfrak{S}_n)$ defined by

$$Mf = \frac{2}{n!} \sum_{1 \leq i < j \leq n} \mu(\{i, j\}) \left[\langle f, \mathbf{1}_{\{i \prec j\}} \rangle \mathbf{1}_{\{i \prec j\}} + \langle f, \mathbf{1}_{\{i \succ j\}} \rangle \mathbf{1}_{\{i \succ j\}} \right]. \tag{2}$$

Then the empirical estimator is a statistical approximation of Mp with $\mathbb{E}\left[\|\widehat{p} - Mp\|_2^2\right] \leq 2/(T.n!)$ where $\|.\|_2$ denotes the usual l^2 norm on $L(\mathfrak{S}_n)$ (the proof is straightforward), and the problem can therefore be seen as an inverse problem, where the goal is to recover p from a noisy version of Mp. Since for any function $f \in L(\mathfrak{S}_n)$, denoting $\overline{f} = \frac{1}{n!} \sum_{\pi \in \mathfrak{S}_n} f(\pi)$,

$$Mf = \overline{f}\, \mathbf{1}_{\mathfrak{S}_n} + \frac{2}{n!} \sum_{1 \leq i < j \leq n} \mu(\{i, j\}) \langle f - \overline{f}, \mathbf{1}_{\{i \prec j\}} \rangle \left[\mathbf{1}_{\{i \prec j\}} - \mathbf{1}_{\{i \succ j\}} \right],$$

M has rank at most $\binom{n}{2} + 1$. Hence, recovering p comes down to approximately solve a linear system of $n(n-1)/2 + 1$ equations with $n!$ unknowns. The solution is far from being unique and the problem requires structural assumptions on p to be resolved. For instance, an assumption of sparsity is made in [2], but empirical evidence show that probability distributions are rather diffuse on $\mathfrak{S}_n$. However, we claim that multiresolution analysis allows to build wavelet bases in which a large class of probability distributions have a sparse decomposition, and therefore that p can be recovered by the following optimization problem :

$$\min_{q \in L(\mathfrak{S}_n)} \|q\|_{\psi,0} \qquad \text{subject to} \qquad \|Mq - \widehat{p}\|_2^2 \leq \frac{2}{Tn!}, \tag{3}$$

where $\|q\|_{\psi,0}$ is the number of coefficients in the decomposition of q in a wavelet basis defined in the sequel. The solution to this problem provides furthermore an estimator with compact representation.

3 Multiresolution and wavelet analysis on $\mathfrak{S}_n$

The space of real valued functions on $\mathfrak{S}_n$, $L(\mathfrak{S}_n)$, is equipped with the canonical Dirac basis $\{\delta_\sigma\}_{\sigma\in\mathfrak{S}_n}$, where $\delta_\sigma(\pi) = 1$ if $\pi = \sigma$ and 0 otherwise. For a function $f \in L(\mathfrak{S}_n)$, the decomposition

$$f = \sum_{\sigma\in\mathfrak{S}_n} f(\sigma)\delta_\sigma \tag{4}$$

shall be referred to as the "spatial" decomposition.

3.1 Fourier analysis on $\mathfrak{S}_n$

Equipped with the composition operator $\circ$, $\mathfrak{S}_n$ is a non-commutative group. The translation operator on $L(\mathfrak{S}_n)$ related to a permutation $\tau \in \mathfrak{S}_n$ is defined by $T_\tau f : \pi \mapsto f(\tau^{-1} \circ \pi)$. It would be expected that a Fourier basis of $L(\mathfrak{S}_n)$ would be an orthonormal basis in which all the operators T_τ are diagonal. Unfortunately, since $\mathfrak{S}_n$ is not commutative, the translation operators do not commute and such a basis cannot exist. However, for any $(\tau, \sigma) \in \mathfrak{S}_n^2$, we have $T_\tau \circ T_\sigma = T_{\tau\circ\sigma}$, which means that the mapping $\tau \in \mathfrak{S}_n \mapsto T_\tau$ is a representation of the group $\mathfrak{S}_n$ (it is actually the left regular representation of $\mathfrak{S}_n$). Therefore classic results in group representation theory guarantee the existence of an orthonormal basis in which all the translation operators are block diagonal, with the same blocks. One may refer to [1] for further details. Let us denote by $\rho_\lambda(\sigma)$ the block indexed by λ of the matrix coefficient of T_σ, $\sigma \in \mathfrak{S}_n$, in this basis. The Fourier (matrix) coefficients of any $f \in L(\mathfrak{S}_n)$ are defined by: $\forall\sigma \in \mathfrak{S}_n$,

$$\widehat{f}(\lambda) = \sum_{\sigma\in\mathfrak{S}_n} f(\sigma)\rho_\lambda(\sigma) \in \mathcal{M}_{d_\lambda}(\mathbb{R}),$$

where d_λ is the size of the block indexed by λ. By virtue of the inversion formula, the function f can then be expanded as :

$$f = \sum_\lambda \frac{d_\lambda}{n!} \left\langle \widehat{f}(\lambda), \rho_\lambda(.) \right\rangle_{HS}, \tag{5}$$

denoting $\langle ., . \rangle_{HS}$ the scalar product on matrices. This expansion shall be referred to as the "spectral" decomposition.

Decompositions (4) and (5) describe local properties either only in space, or else in frequency solely. Hence, there is no reason that they would allow sparse representations of functions with spatially varying degrees of smoothness . This motivates the need for building a basis, which would be localized both in space and in frequency, such as that proposed in [3]. We now briefly recall the principles underlying its construction (using slightly different notations).

3.2 The multiscale structure of $\mathfrak{S}_n$

As a first go, a *multiscale structure* on $\mathfrak{S}_n$ is defined. For $1 \leq k \leq n$, set $\mathcal{A}_k = \{i = (i_1, ..., i_k) \in \{1, ..., n\}^k \mid p \neq q \Rightarrow i_p \neq i_q\}$ and $\mathcal{A}_0 = \{0\}$ by convention. Define the sets $A_0^0 = \mathfrak{S}_n$, $A_j^1 = \{\sigma \in \mathfrak{S}_n \mid \sigma^{-1}(1) = j\}$ for $j \in \{1, \ldots, n\}$, and more generally, for $i \in \mathcal{A}_k$ and $k \in \{1, ..., n-1\}$, consider:

$$A_i^k = \{\sigma \in \mathfrak{S}_n \mid \sigma^{-1}(1) = i_1, \ldots, \sigma^{-1}(k) = i_k\}. \tag{6}$$

Observe that, for all $k \in \{0, \ldots, n-1\}$, $|A_i^k| = (n-k)!$ (in particular for $k = n-1$, A_i^k is a singleton) and that $\left(\{A_i^k\}_{i\in\mathcal{A}_k}\right)_{0\leqslant k\leqslant n-1}$ is a sequence of nested partitions of $\mathfrak{S}_n$:

$$\begin{aligned} \mathfrak{S}_n &= \textstyle\bigcup_{i\in\mathcal{A}_k} A_i^k \quad \text{for all } k \in \{0, ..., n-1\}, \\ A_i^k &= \textstyle\bigcup_{j\notin i} A_{(i,j)}^{k+1} \quad \text{for all } k \in \{0, ..., n-1\} \text{ and } i \in \mathcal{A}_k, \end{aligned}$$

where $j \notin i$ abusively means that $j \in \{1, ..., n\} \setminus \{i_1, ..., i_k\}$.
Beyond the multiscale structure thus defined for $\mathfrak{S}_n$, notice that the A_i^k's interact well with the group structure of $\mathfrak{S}_n$. Indeed, identifying the isomorphic groups $\mathfrak{S}_{n-k}$ and $\{\sigma \in \mathfrak{S}_n \mid \sigma(1) = 1, ..., \sigma(k) = k\}$, one may write $A_i^k = \{\sigma' \circ \pi_i \mid \sigma' \in \mathfrak{S}_{n-k}\}$, where π_i is any permutation in $\mathfrak{S}_n$ such that $\pi_i^{-1}(1) = i_1, \ \ldots, \ \pi_i^{-1}(k) = i_k$. This means that A_i^k is a right coset of $\mathfrak{S}_{n-k}$ in $\mathfrak{S}_n$, which is denoted by $A_i^k = \mathfrak{S}_{n-k}\pi_i$. Therefore, the sequence of nested partitions is directly related to the embedding of subgroups $\mathfrak{S}_1 \subset \cdots \subset \mathfrak{S}_{n-1} \subset \mathfrak{S}_n$. This key point allow the basis defined below to enjoy good localization properties in frequency.

3.3 Multiresolution analysis on $\mathfrak{S}_n$

The definition of a multiresolution analysis on the symmetric group given in [3] is based on the multiscale tree-structure of $\mathfrak{S}_n$ described above. It involves the projectors $P_i : f \in L(\mathfrak{S}_n) \mapsto f\mathbf{1}_{A_i^k}$, $i \in A_i^k$ and $k \in \{0, \ldots, n-1\}$.

Definition 3.1. *A sequence of subspaces $V^0 \subseteq V^1 \subseteq ... \subseteq V^{n-1} = L(\mathfrak{S}_n)$ forms a coset based multiresolution analysis (CMRA) for $\mathfrak{S}_n$ if the following properties are satisfied.*

A. For any $f \in V^k$ and $\tau \in \mathfrak{S}_n$, $T_\tau f \in V^k$.

B. If $f \in V^k$, then $P_i f \in V^{k+1}$ for any $i \in \mathcal{A}_{k+1}$.

C. If $g \in V^{k+1}$, then for any $i \in \mathcal{A}_{k+1}$ there exist $f \in V^k$ such that $P_i f = g$.

We refer to [3] for a detailed description of a general method to construct a CMRA for $\mathfrak{S}_n$, starting from any given subspace V^0. In this paper, focus is on the simple, but sufficiently rich, case where $V^0 = \{f \in L(\mathfrak{S}_n) : \quad f \text{ constant on } \mathfrak{S}_n\}$. It yields the subspaces :

$$V^k = \{f \text{ constant on each } A_i^k, \ \ i \in \mathcal{A}_k\}, \ \ k \in \{1, \ \ldots, \ n-1\}.$$

3.4 Wavelets on $\mathfrak{S}_n$

Starting from a multiresolution analysis $\left(V^k\right)_{1 \leq k < n}$, the general construction scheme for wavelet bases was formalized in [4]. The principle is to define W^{k+1} as the orthogonal of V^k in V^{k+1} and to consider the decomposition:

$$V^N = V^0 \bigoplus \left[\bigoplus_{k=1}^{n-1} W^k\right].$$

Then one has to define orthonormal bases for V^0 and the W^k, that interact well with the multiresolution structure. For the considered case of piecewise constant functions, the associated wavelet bases are similar to Haar bases, and have simple expressions, such as:

$$\phi = \frac{1}{\sqrt{n!}}\mathbf{1}_{\mathfrak{S}_n} \text{ and } \psi_{i,m}^k = \frac{1}{\sqrt{(n-k)!}} \frac{1}{\sqrt{m(m+1)}} \left[\sum_{t=1}^{m} \mathbf{1}_{A_{i,j_t}^k} - m\mathbf{1}_{A_{i,j_{m+1}}^k}\right], \tag{7}$$

for $1 \leqslant k \leqslant n-1$, $i \in \mathcal{A}_{k-1}$ and $1 \leqslant m \leqslant n-k$, with $\{j_1 \leqslant ... \leqslant j_{n-k+1}\} = \{1, ..., n\} \setminus \{i_1, ..., i_{k-1}\}$.
Any function $f \in L(\mathfrak{S}_n)$ thus admits the expansion:

$$f = \langle f, \phi \rangle \phi + \sum_{k=1}^{n-1} \sum_{i \in \mathcal{A}_{k-1}} \sum_{m=1}^{n-k} \left\langle f, \psi_{i,m}^k \right\rangle \psi_{i,m}^k, \tag{8}$$

referred to as the "spatial-frequency" decomposition.

4 Sparse estimation in the wavelet basis

For a function $f \in L(\mathfrak{S}_n)$, let Ψf denote its wavelet transform in the basis 7, *i.e.* the collection of its wavelet coefficients: $\Psi f = \{\langle f, \phi\rangle\} \cup \{\langle f, \psi^k_{i,m}\rangle \mid 1 \leqslant k \leqslant n-1, i \in \mathcal{A}_{k-1}, 1 \leqslant m \leqslant n-k\}$. The initial inverse problem can now be written as:

$$\min_{q \in L(\mathfrak{S}_n)} \|\Psi q\|_0 \qquad \text{subject to} \qquad \|Mq - \widehat{p}\|_2^2 \leq \frac{2}{Tn!}. \tag{9}$$

This type of problem has been the subject of much attention these last few years, and numerous methods have been proposed to solve it with their theoretical guarantees, one the most famous being the l^1 Lagrangian pursuit (see [5] for the details). Let Θ denote the set of indexes of the wavelet basis, $\Theta = \{0\} \cup \{(k, i, m) \mid 1 \leqslant k \leqslant n-1, i \in \mathcal{A}_{k-1}, 1 \leqslant m \leqslant n-k\}$, where 0 is the index of ϕ. The principle is to compute the estimator defined by:

$$\tilde{p} = \sum_{\theta \in \Theta} c_\theta \frac{\psi_\theta}{\|M\psi_\theta\|_2} \qquad \text{with} \qquad c = \underset{c \in \mathbb{R}^{n!}}{\operatorname{argmin}}\ \frac{1}{2}\left\|\widehat{p} - \sum_{\theta \in \Theta} c_\theta \frac{M\psi_\theta}{\|M\psi_\theta\|_2}\right\|_2^2 + T\|c\|_1, \tag{10}$$

where T is a Lagrangian multiplier to be adjusted. Let $p = \sum_{\theta \in \Theta^*} \langle p, \psi_\theta\rangle \psi_\theta$ be the decomposition of p in the wavelet basis 7, where Θ^* is the support of Ψp. The theory of sparse approximation ensures that under sufficient "incoherence" conditions of $\{M\psi_\theta / \|M\psi_\theta\|\}_{\theta \in \Theta^*}$, the estimator $\tilde{p}$ is well-defined and converges toward the solution of problem 9, *i.e.* p.

Acknowledgements: This work is supported by the Labex LMH grant n°ANR-11-IDEX-0003-02.

Bibliography

[1] Diaconis, P. Group representations in probability and statistics. (1988). *Institute of Mathematical Statistics Lecture Notes - Monograph Series.*

[2] Jagabathula, S and Shah, D. (2011). Inferring Rankings Using Constrained Sensing. *IEEE Transactions on Information Theory* 57(11):7288–7306.

[3] Kondor, R and Dempsey, W. (2012). Multiresolution analysis on the symmetric group. *Neural Information Processing Systems* 25.

[4] Stéphane Mallat. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *Pattern Analysis and Machine Intelligence, IEEE* II(7).

[5] Stéphane Mallat. (2008).*A Wavelet Tour of Signal Processing, The Sparse Way*, Academic Press.

[6] Mingxuan Sun, M., Lebanon, G. and Kidwell, P. (2012). Estimating probabilities in recommendation systems. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 61(3):471–492.

Constructing hierarchical copulas using the Kendall distribution function

Eike Christian Brechmann[*]

Center for Mathematical Sciences, Technische Universität München.

Abstract

While there is substantial need for dependence models in higher dimensions, most existing models are rather restrictive and barely balance parsimony and flexibility. The class of hierarchical Kendall copulas is proposed as a new approach to these problems. By aggregating dependence information of non-overlapping groups of variables in different hierarchical levels using the Kendall distribution function, hierarchical Kendall copulas provide a new and attractive option to model dependence patterns between large numbers of variables.

Keywords: multivariate copula, hierarchical copula, Kendall distribution function
AMS subject classifications: 62H20

1 Introduction

The statistical modeling of dependence has made significant progress in the last years. This is mainly due to the appealing copula approach. According to the famous theorem by Sklar [9], every n-dimensional distribution function can be expressed in terms of its univariate marginal distribution functions and an n-dimensional copula, which is a multivariate distribution function on $[0, 1]^n$ with uniform marginal distribution functions. Many of the standard, and also of the newly proposed, copula models however turn out to be rather restrictive in higher dimensions and barely balance parsimony and flexibility. Grouping variables, for instance by industry sectors or nationality, is therefore a common procedure to approach such problems. Examples of such copula models are the grouped Student's t copula by [5], elliptical copulas with clustered correlation matrix and hierarchical Archimedean copulas (see, e.g., [7]).

To overcome limitations of these models, we introduce the new class of hierarchical Kendall copulas as a flexible, but yet parsimonious dependence model (see [2]). It is built up by copulas for non-overlapping groups (clusters) of variables in different hierarchical levels. Dependence information of the clusters is aggregated using the Kendall distribution function, which is the multivariate analog of the probability integral transform for univariate random variables. The model does not restrict the choice of copulas and their parameters, so that hierarchical Kendall copulas provide a new and attractive option to model dependence patterns between large numbers of variables.

This paper presents main results of [2, 3] and shows how hierarchical Kendall copulas can be used to conduct systemic risk stress testing exercises—an important issue in the finance and insurance sectors today.

[*] e-mail: brechmann@ma.tum.de

2 Hierarchical Kendall copulas

The hierarchical construction, which we will investigate here, is based on the notion of the Kendall distribution function (see, e.g., [1]). Let $\boldsymbol{U} := (U_1, ..., U_n)' \sim C$, where C is an n-dimensional copula, then the *Kendall distribution function* K is defined as

$$K(t) := P(C(\boldsymbol{U}) \leq t), \quad t \in [0, 1].$$

It holds that $t \leq K(t) \leq 1$, for $t \in [0, 1]$, as well as $K(0-) = 0$. We assume here that copulas are absolutely continuous and possess continuous Kendall distribution functions.

The Kendall distribution function is the multivariate analog of the probability integral transform. More precisely, it is the univariate distribution function of the random variable $Z := C(\boldsymbol{U})$, so that $K(Z) \sim U(0, 1)$. An alternative interpretation is that it describes the distribution of the *level sets* of a copula, which are given by

$$L(z) = \{\boldsymbol{u} \in [0, 1]^d : C(\boldsymbol{u}) = z\}, \quad z \in (0, 1).$$

Generally, Kendall distribution functions are not available in closed form. A notable exception are Archimedean copulas (see [1]).

The idea of our hierarchical dependence model, which we call “hierarchical Kendall copula”, is to aggregate dependence information of non-overlapping groups of variables (clusters) using the Kendall distribution function and thus mimic the classical copula approach for univariate margins. Now, let $U_1, ..., U_n \sim U(0, 1)$ and let $C_0, C_1, ..., C_d$ be copulas of dimensions $d, n_1, ..., n_d$, respectively, where $n_i \geq 1,\ i = 1, ..., d$, and $n = \sum_{i=1}^{d} n_i$. Further, let $K_1, ..., K_d$ denote the Kendall distribution functions corresponding to $C_1, ..., C_d$. We define $m_i = \sum_{j=1}^{i} n_j,\ i = 1, ..., d$, and $m_0 = 0$ as well as $\boldsymbol{U}_i := (U_{m_{i-1}+1}, ..., U_{m_i})'$ and $V_i := K_i(C_i(\boldsymbol{U}_i))$ for $i = 1, ..., d$. Under the assumptions that

- $\mathcal{A}_1$: $\boldsymbol{U}_1, ..., \boldsymbol{U}_d$ are mutually independent conditionally on $(V_1, ..., V_d)'$, and
- $\mathcal{A}_2$: the conditional distribution of $\boldsymbol{U}_i|(V_1, ..., V_d)'$ is the same as the conditional distribution of $\boldsymbol{U}_i|V_i$ for all $i = 1, ..., d$,

the random vector $(U_1, ..., U_n)'$ is said to be distributed according to the *hierarchical Kendall copula* $C_{\mathcal{K}}$ with *nesting copula* C_0 and *cluster copulas* $C_1, ..., C_d$ if

A. $\boldsymbol{U}_i \sim C_i\ \forall i \in \{1, ..., d\}$,

B. $(V_1, ..., V_d)' \sim C_0$.

In contrast to other hierarchical dependence models, this approach allows to combine copulas from different classes to account for complex dependence patterns. The two-level construction is illustrated in Figure 1. It can also easily be extended to an arbitrary number of levels (see [2]).

The intuition behind the two assumptions $\mathcal{A}_1$ and $\mathcal{A}_2$ is that, given the information of the nesting variables $V_1, ..., V_d$, the clusters are independent of each other and also of other nesting variables, since the dependence among the clusters is explained through the “representatives” $V_1, ..., V_d$. In other words, $V_1, ..., V_d$ can be interpreted as unobserved factors, whose joint behavior determines the dependence of the different clusters. In finance, such factors may be, e.g., industry sectors.

We now characterize the hierarchical Kendall copula in terms of its density. Let $\boldsymbol{U} = (U_1, ..., U_n)'$ be distributed according to a hierarchical Kendall copula $C_{\mathcal{K}}$ with cluster copulas $C_1, ..., C_d$ and nesting copula C_0. Corresponding densities are denoted by $c_0, c_1, ..., c_d$, respectively. According to [2], it holds that

$$c_{\mathcal{K}}(\boldsymbol{u}) = c_0(K_1(C_1(\boldsymbol{u}_1)), ..., K_d(C_d(\boldsymbol{u}_d))) \prod_{i=1}^{d} c_i(\boldsymbol{u}_i),$$

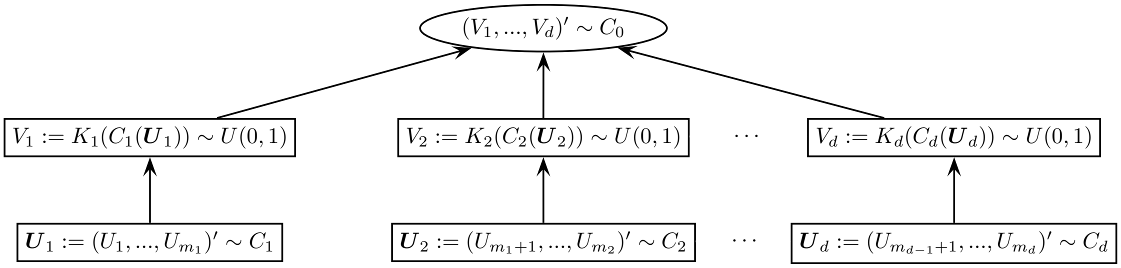

$(V_1, ..., V_d)' \sim C_0$
$V_1 := K_1(C_1(\boldsymbol{U}_1)) \sim U(0,1)$
$V_2 := K_2(C_2(\boldsymbol{U}_2)) \sim U(0,1)$
$\cdots$
$V_d := K_d(C_d(\boldsymbol{U}_d)) \sim U(0,1)$
$\boldsymbol{U}_1 := (U_1, ..., U_{m_1})' \sim C_1$
$\boldsymbol{U}_2 := (U_{m_1+1}, ..., U_{m_2})' \sim C_2$
$\cdots$
$\boldsymbol{U}_d := (U_{m_{d-1}+1}, ..., U_{m_d})' \sim C_d$

Figure 1: Illustration of the construction of hierarchical Kendall copulas.

where $\boldsymbol{u} = (u_1, ..., u_n)'$ and $\boldsymbol{u}_i = (u_{m_{i-1}+1}, ..., u_{m_i})',\ i = 1, ..., d.$
The availability of the density expression then renders feasible maximum likelihood estimation of dependence parameters. Furthermore, it can be shown that the two important special cases of independence as well as of comonotonicity are hierarchical Kendall copulas, while, in general, dependence between clusters ranges between these cases and can also be negative.
Sampling from a given hierarchical Kendall copula is however rather challenging. In general, a sample from a hierarchical Kendall copula can be obtained using the following top-down procedure (see Figure 1).

A. Obtain a sample $(v_1, ..., v_d)'$ from C_0.

B. Set $z_i := K_i^{-1}(v_i)$ for all $i = 1, ..., d$.

C. Obtain a sample $\boldsymbol{u}_i$ from $\boldsymbol{U}_i|(C_i(\boldsymbol{U}_i) = z_i)$ for $i = 1, ..., d$.

D. Return $\boldsymbol{u} := (u_1, ..., u_n)'$.

The crucial step is the third one. It requires sampling from the distribution of $\boldsymbol{U}_i|(C_i(\boldsymbol{U}_i) = z_i)$, that is, sampling from a multivariate distribution given the level set $L(z_i)$ at level $z_i \in (0,1)$. In general, no closed-form solutions are known for this problem and approximate procedures such as rejection sampling have to be used. However, for Archimedean, Plackett and Archimax copulas, closed-form methods are derived in [2, 3]. In particular, let $\boldsymbol{U} := (U_1, ..., U_n)' \sim C$, where C is an n-dimensional Archimedean copula with generator φ (see [8]). It holds then, for all $j = 1, ..., n-1$, that

$$F_{U_j|U_1,...,U_{j-1},C(\boldsymbol{U})}(u|u_1, ..., u_{j-1}, z) = \left(1 - \frac{\varphi(u)}{\varphi(z) - \sum_{1\leq i<j} \varphi(u_i)}\right)^{n-j}, \qquad (1)$$

$u \in (C^{-1}_{u_1,...,u_{j-1}}(z), 1)$, where $C^{-1}_{u_1,...,u_{j-1}}(\cdot)$ is the inverse of

$$C_{u_1,...,u_{j-1}}(\cdot) := C(u_1, ..., u_{j-1}, \cdot, 1, ..., 1).$$

This expression can be used for closed-form conditional inverse sampling.

3 Systemic risk stress testing

In the aftermath of the financial crisis of 2007-2009, the discussion about systemic risk is central in order to prevent similar crises in the future, and therefore regulators seek to identify systemically important institutions (see [6]). Systemic importance is closely linked to contagion effects among financial institutions and hence their interconnectedness. Statistically, this interconnectedness can be expressed and characterized using suitable dependence models. For this purpose, hierarchical dependence models are particularly useful,

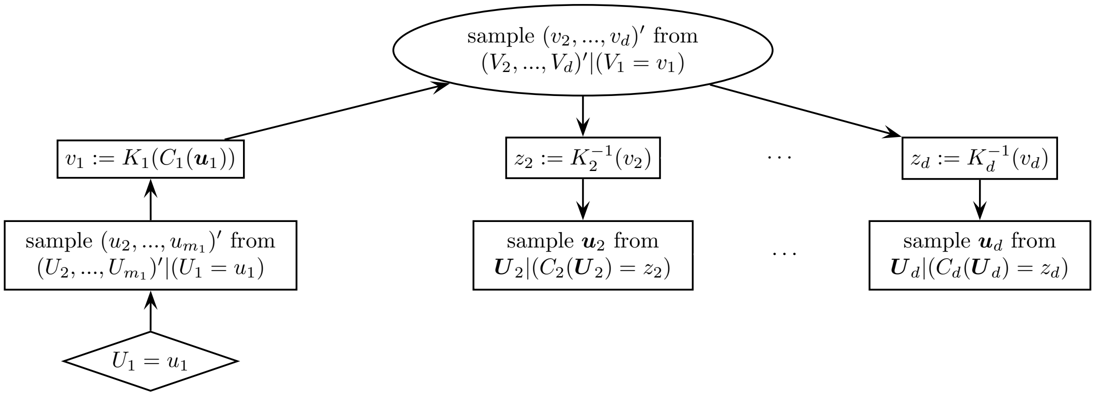

sample $(v_2, ..., v_d)'$ from $(V_2, ..., V_d)'|(V_1 = v_1)$
$v_1 := K_1(C_1(\boldsymbol{u}_1))$
$z_2 := K_2^{-1}(v_2)$
$\cdots$
$z_d := K_d^{-1}(v_d)$
sample $(u_2, ..., u_{m_1})'$ from $(U_2, ..., U_{m_1})'|(U_1 = u_1)$
sample $\boldsymbol{u}_2$ from $\boldsymbol{U}_2|(C_2(\boldsymbol{U}_2) = z_2)$
$\cdots$
sample $\boldsymbol{u}_d$ from $\boldsymbol{U}_d|(C_d(\boldsymbol{U}_d) = z_d)$
$U_1 = u_1$

Figure 2: Illustration of conditional sampling from hierarchical Kendall copulas.

since different institutions are typically clustered by region or industry sector (banks, insurers, hedge funds, etc.).

Having identified a network of financial institutions, it can be used for systemic risk assessment and classification. An important tool for this purpose is stress testing, that is, the analysis of the stability of the financial system under shocks such as the failure of an institution. The potential impact of such critical events decisively determines the systemic relevance of an institution.

Now, let $\boldsymbol{X} := (X_1, ..., X_n)'$ be a random vector of risk quantities. Then we are interested in the case $\boldsymbol{X}_{-k}|(X_k = x_k)$, $k \in \{1, ..., n\}$, where $\boldsymbol{X}_{-k}$ denotes the random vector $\boldsymbol{X}$ with the kth component removed and the event $\{X_k = x_k\}$ corresponds to a stress situation. For instance, if X_k is the company value, then a stress situation occurs when x_k is very small (near-failure of company k). The conditional distribution of $\boldsymbol{X}_{-k}|(X_k = x_k)$ given the specific underlying dependence model is however typically not known in closed form. We will therefore derive a conditional simulation algorithm for hierarchical Kendall copulas, which can be used for scenario analyses.

We set $U_j := F_{X_j}(X_j)$ for all $j = 1, ..., n$, and assume that $\boldsymbol{U} := (U_1, ..., U_n)' \sim C_{\mathcal{K}}$ for a hierarchical Kendall copula $C_{\mathcal{K}}$ with cluster copulas $C_1, ..., C_d$ and nesting copula C_0. Without loss of generality, let $k = 1$ and define $u_1 := F_{X_1}(x_1)$. This means that the stress situation occurs in the first cluster. The sampling strategy is then as follows:

A. Obtain a sample $(u_2, ..., u_{m_1})'$ from $(U_2, ..., U_{m_1})'|(U_1 = u_1)$.

B. Set $v_1 := K_1(C_1(\boldsymbol{u}_1))$.

C. Obtain a sample $(v_2, ..., v_d)'$ from $(V_2, ..., V_d)'|(V_1 = v_1)$.

D. Set $z_i := K_i^{-1}(v_i)$ for all $i = 2, ..., d$.

E. Obtain a sample $\boldsymbol{u}_i$ from $\boldsymbol{U}_i|(C_i(\boldsymbol{U}_i) = z_i)$ for $i = 2, ..., d$.

F. Return $(u_2, ..., u_n)'$.

Samples x_j, $j = 2, ..., d$, from $\boldsymbol{X}_{-1}|(X_1 = x_1)$ are then given by $x_j := F_{X_j}^{-1}(u_j)$. This conditional sampling procedure for hierarchical Kendall copulas is illustrated in Figure 2. Since the fifth step is as in the standard sampling approach described above, conditional sampling of hierarchical Kendall copulas boils down to conditional sampling of the cluster and nesting copulas. While conditional sampling of elliptical distributions (and copulas) is straightforward and well-known, appropriate strategies for Archimedean and vine copulas are derived in [4]. In particular, the procedure for Archimedean copulas exploits (1).

The described conditional sampling approach for hierarchical Kendall copulas then allows to evaluate the impact of stress events to certain financial institutions to assess their systemic relevance. A case study using Archimedean, elliptical and vine copulas can be found in [4].

Acknowledgements: The author gratefully acknowledges the support of the TUM Graduate School's International School of Applied Mathematics as well as of Allianz Deutschland AG.

Bibliography

[1] Barbe, P., Genest, C., Ghoudi, K., and Rémillard, B. (1996). On Kendall's process. *Journal of Multivariate Analysis* 58 (2), 197–229.

[2] Brechmann, E. C. (2014). Hierarchical Kendall copulas: Properties and inference. Canadian Journal of Statistics, 42 (1), 78–108.

[3] Brechmann, E. C. (2013). Sampling from Hierarchical Kendall copulas. Journal de la Société Française de Statistique, 154 (1), 192-209.

[4] Brechmann, E. C., Hendrich, K., and Czado, C. (2013). Conditional copula simulation for systemic risk stress testing. Preprint.

[5] Daul, S., De Giorgi, E., Lindskog, F., and McNeil, A. (2003). The grouped t-copula with an application to credit risk. *Risk* 16 (11), 73–76.

[6] Financial Stability Board, International Monetary Fund, and Bank for International Settlements (2009). Report to the G-20 finance ministers and central bank governors: Guidance to assess the systemic importance of financial institutions, markets and instruments: Initial considerations.

[7] Hofert, M. (2010). Construction and sampling of nested Archimedean copulas. In Jaworski, P., Durante, F., Härdle, W., and Rychlik, T. (Eds.), *Copula Theory and Its Applications*, Springer, Berlin.

[8] McNeil, A. and Nešlehová, J. (2009). Multivariate Archimedean copulas, d-monotone functions and ℓ_1-norm symmetric distributions. *Annals of Statistics* 37 (5B), 3059–3097.

[9] Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de L'Université de Paris* 8, 229–231.

Discriminating between long-range dependence and non-stationarity

Philip Preuß[*][1] **and Mathias Vetter**[1]

[1] *Ruhr-Universität Bochum, Fakultät für Mathematik*

Abstract

This paper is devoted to the discrimination between a stationary long-range dependent model and a non stationary process. We develop a nonparametric test for stationarity in the framework of locally stationary long memory processes which is based on a Kolmogorov-Smirnov type distance between the time varying spectral density and its best approximation through a stationary spectral density. We show that the test statistic converges to the same Gaussian limit as in the short memory case if the (possibly time varying) long memory parameter is smaller than $1/4$ and justify why the limiting distribution is different if the long memory parameter exceeds this boundary. Concerning the latter case the novel FARI(∞) bootstrap is introduced which provides a bootstrap-based test for stationarity that only requires the long memory parameter to be smaller than $1/2$ which is the usual restriction in the framework of long-range dependent time series. Note that the present paper is a very condensed version of [13] to which we refer for all technical details and simulation results.

Keywords: testing stationarity, locally stationary process, long memory, integrated periodogram, spectral density.
AMS subject classifications: 62M10, 62M15, 62G10.

1 Introduction

For many decades, the assumption of second order stationarity has been the dominating paradigm in time series analysis. It allows for a straightforward implementation of estimation or forecasting techniques, and therefore a vast amount of literature exists in this framework; see for example [2] for an comprehensive overview. It is, however, well-known that many processes in the reality change their dependency structure over time, which yields that the assumption of stationarity becomes problematic in many applications. Therefore several approaches exist to model time-varying dependencies and one proposal which become particularly popular throughout the last decade is that of a local stationarity. These kinds of stochastic processes were introduced by [3] in the short memory context and extended to the long range dependent case by [1], [9] and [14].

There exist a vast amount of articles in which tests for stationarity are derived in the framework of local stationary [see for example [5], [6], [10], [11], [12] or [16]], but in all these articles long-range dependence is excluded, i.e. these methods cannot be employed for discriminating between long memory and non-stationarity. Such a discrimination is, however, of great importance, since many effects in the reality can be both explained by using either a complicated stationary long memory process or a simple non-stationary short memory model; see for example [8]. The aim of this paper is to fill the just described gap, i.e. to derive a test for stationarity which works also in the presence of long memory.

In order to achieve this goal, we will construct an estimator for a Kolmogorov-Smirnov-type distance between the time varying spectral density and its best approximation through a stationary spectral density, and discuss its asymptotic Gaussianity if the long memory parameter is smaller than $1/4$. If this boundary is

[*] Corresponding author, e-mail: philip.preuss@ruhr-uni-bochum.de

exceeded, the asymptotics become different. In order to obtain the asymptotic quantiles of our test statistic, we will then propose a bootstrap procedure which basically works by transforming the data to something which is 'close' to short memory and then applying the AR(∞) bootstrap of [7] to the transformed dataset. Since this corresponds to the case, where an FARIMA($p, d, 0$) model is fitted with growing order p, we call this new procedure the FARI(∞) bootstrap. It turns out that for consistency of this method, we only require that $1/2$ is an upper bound of the long memory parameter.

This paper is organized is follows: In Section 2 we introduce locally stationary long memory processes and our measure of stationarity. We then construct an estimator for this quantitiy in Section 3 and present our new bootstrap prceodure in Section 4. As mentioned in the abstract we refer to [13] for all technical details and a comprehensive simulation sutdy.

2 The framework

2.1 Locally stationary long memory processes

Consider the triangular array $X_{t,T}$, $t = 1, ..., T$, which follows an $MA(\infty)$ representation, i.e.

$$X_{t,T} = \sum_{l=0}^{\infty} \psi_{l,t,T} Z_{t-l}, \quad t = 1, ..., T,$$

with

$$\sup_{t,T} \sum_{l=0}^{\infty} \psi_{l,t,T}^2 < \infty,$$

and independent, standard normal distributed innovations Z_t. For the time varying coefficents $\psi_{l,t,T}$ we assume that there exist twice continuously differentiable functions $\psi_l : (0, 1] \to \mathbb{R}$ such that

$$\sup_{t=1,..,T} |\psi_{l,t,T} - \psi_l(t/T)| \le \frac{C}{T} \left(\frac{\log(l)}{l^{1-D}} 1_{\{l \neq 0\}} + 1_{\{l=0\}} \right), \quad \forall l \in \mathbb{N},$$

holds for some $0 < D < 1/2$ and a constant $C \in \mathbb{R}$. Concerning the sequence of approximating functions $(\psi_l(u))_{l \in \mathbb{N}}$ we furthermore assume that the conditions from Assumption 1 in [13] are fulfilled. This ensures that the time-varying spectral density

$$f(u, \lambda) = \frac{1}{2\pi} | \sum_{l=0}^{\infty} \psi_l(u) \exp(-i\lambda l)|^2$$

behave like a constant times $1/\lambda^{2d(u)}$ as $\lambda \to 0$, where $d(u) : (0, 1] \to (0, D)$ is a twice continuously differentiable function, which is called 'time-varying long memory parameter' and describes how heavy current observations are influenced by data observed a long time ago. While a value close to $1/2$ indicates a very strong memory, the closer it is to zero the weaker the dependency becomes.

2.2 Measure of stationarity

In order to obtain a measure of stationarity, we follow [4] and consider a Kolmogorov-Smirnov type distance between the time varying spectral density $f(u, \lambda)$ and its approximation through the stationary spectral density $\int_0^1 f(u, \lambda) du$, namely

$$E = \sup_{(v,\omega) \in [0,1]^2} |E(v, \omega)|,$$

where

$$E(v,\omega) := \frac{1}{2\pi}\Big(\int_0^v \int_0^{\pi\omega} f(u,\lambda)d\lambda du - v\int_0^{\pi\omega}\int_0^1 f(u,\lambda)dud\lambda\Big), \quad (v,\omega)\in[0,1]^2.$$

Note that if $f(u,\lambda)$ does not depend on u (which is the case if the underlying process is stationary), then the expression E is equal to zero while being strictly positive otherwise. So in order to obtain a test for the null hypothesis that $f(u,\lambda)$ does not depend on the rescaled time u, it is natural to estimate E and to reject the null hypothesis if the estimator becomes 'big'.

3 The estimator

For obtaining an estimator for E we require an estimator for the local spectral density $f(u,\lambda)$, which is given by the so called local periodogram. This quantity is obtained by choosing an $N = o(T)$ and calculating

$$I_N(u,\lambda) = \frac{1}{2\pi}|\sum_{p=0}^{N-1} X_{\lfloor uT\rfloor - N/2+p+1,T}\exp(-i\lambda p)|^2,$$

where we set $X_{t,T} = 0$ for $t \notin \{1,...,T\}$. This is the usual periodogram but only using N data around the time point $\lfloor uT\rfloor$. One can show that, if $N\to\infty$, then, as for the classical periodogram, the local periodogram is an asymptotically unbiased (but not consistent) estimator for the time-varying spectral density. We now divide the T data into M intervals with length N each (i.e. it is $T = NM$). An estimator for the quantity E is then given by

$$\hat{E} = \sup_{v,\omega\in[0,1]} |\hat{E}(v,\omega)|,$$

where

$$\hat{E}_T(v,\omega) := \frac{1}{T}\sum_{j=1}^{\lfloor vM\rfloor}\sum_{k=1}^{\lfloor \omega\frac{N}{2}\rfloor} I_N(u_j,\lambda_k) - \frac{\lfloor vM\rfloor}{M}\frac{1}{T}\sum_{j=1}^{M}\sum_{k=1}^{\lfloor \omega\frac{N}{2}\rfloor} I_N(u_j,\lambda_k).$$

Here u_j denote the rescaled midpoints of the M intervals and $\lambda_k = 2\pi k/N$ correspond to the usual Fourier frequencies. Theorem 2 in [13] now states that if $D < 1/4$ and

$$N\to\infty, \quad N/T\to 0, \quad T^{1/2}\log(N)/N^{1-2D}\to 0 \tag{1}$$

is satisfied, then a normalized version of $\hat{E}_T(v,\omega)$ converges to some Gaussian process, which covariance structure depends in a complicated way on the spectral density $f(u,\lambda)$. So the asymptotic distribution of $\hat{E}$ is unknown in general and resampling methods are required to obtain critical values. Note further, that it is essential that N/T and $\sqrt{T}/N^{1-2D}$ both tend to zero, so asymptotic Gaussianity is no longer obtained for $D \geq 1/4$ (which is in line with the findings of [15] in the stationary case). However, we will provide a bootstrap procedure in the next section which approximates the critical values in this situation as well.

4 Bootstrap

In order to obtain critical values we proceed as follows.

1) Choose $p = p(T) \in \mathbb{N}$ and calculate $\hat{\theta}_{T,p} = (\hat{\underline{d}}, \hat{\sigma}_p^2, \hat{a}_{1,p}, \ldots, \hat{a}_{p,p})$ as the minimizer of

$$\frac{1}{T}\sum_{k=1}^{T/2}\Big(\log f_{\theta_p}(\lambda_{k,T}) + \frac{I_T(\lambda_{k,T})}{f_{\theta_p}(\lambda_{k,T})}\Big)$$

where $\lambda_{k,T} = 2\pi k/T$ for $k = 1, \ldots, T/2$, $I_T(\lambda) = \frac{1}{2\pi T}|\sum_{t=1}^{T} X_{t,T}\exp(-i\lambda t)|^2$ is the usual periodogram for stationary processes and

$$f_{\theta_p}(\lambda) = \frac{|1-\exp(-i\lambda)|^{-2\underline{d}}}{2\pi} \times \frac{\sigma_p^2}{|1-\sum_{j=1}^{p} a_{j,p}\exp(-i\lambda j)|^2}$$

is the spectral density of a stationary FARIMA$(p, \underline{d}, 0)$ model which we want to fit. Note that the estimator $\hat{\theta}_{T,p}$ is the classical Whittle estimator of a stationary process; see [17].

2) Calculate $Y_{t,T} = (1-B)^{\hat{\underline{d}}} X_{t,T}$ for $t = 1, \ldots, T$ and simulate a pseudo-series $Y_{1,T}^*, \ldots, Y_{T,T}^*$ according to

$$Y_{t,T}^* = Y_{t,T};\ t = 1, \ldots, p, \quad Y_{t,T}^* = \sum_{j=1}^{p} \hat{a}_{j,p} Y_{t-j,T}^* + \hat{\sigma}_p Z_t^*;\ p < t \leq T,$$

where the Z_t^* are independent standard normal distributed random variables.

3) Create the pseudo-series $X_{1,T}^*, \ldots, X_{T,T}^*$ by calculating $X_{i,T}^* = (1-B)^{-\hat{\underline{d}}} Y_{i,T}^*$ and compute $\hat{E}_T^*(v,\omega)$ in the same way as $\hat{E}_T(v,\omega)$ but with the original observations $X_{1,T}, \ldots, X_{T,T}$ replaced by the bootstrap replicates $X_{1,T}^*, \ldots, X_{T,T}^*$.

By repeating the above steps $B \in \mathbb{N}$ times, one obtaines an estimator for the $1-\alpha$ quantile of $\hat{E}$ under the null hypothesis. The null hypothesis is then rejected if $\hat{E}$ exceeds this estimated qantile. If we let $p = p(T)$ grow to infinity as the sample size T increases, then, under suitable conditions, a level α test is obtained, as it is shown in Section 4 of [13]. Note D only has to be smaller than $1/2$ which is the usual restriction in the framework of long-range dependence.

Acknowledgements: This work has been supported in part by the Collaborative Research Center "Statistical modeling of nonlinear dynamic processes" (SFB 823, Teilprojekt C1) of the German Research Foundation (DFG). It was written while the first author was a Visiting Fellow at the London School of Economics whose hospitality is gratefully acknowledged.

Bibliography

[1] Beran, J. (2009). On parameter estimation for locally stationary long-memory processes. *Journal of Statistical Planning and Inference*, 139, 900–915.

[2] Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods.*, Springer Verlag, New York.

[3] Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *Annals of Statistics*, 25(1), 1–37.

[4] Dahlhaus, R. (2009). Local inference for locally stationary time series based on the empirical spectral measure. *Journal of Econometrics*, 151, 101–112.

[5] Dette, H. and Preuß, P. and Vetter, M. (2011). A measure of stationarity in locally stationary processes with applications to testing. *Journal of the American Statistical Association*, 106(495), 1113-1124.

[6] Dwivedi, J. and Subba Rao, S. (2010). A test for second order stationarity based on the discrete Fourier transform. *Journal of Time Series Analysis*, 32, 68-91.

[7] Kreiss, J.-P. (1988). Asymptotic statistical inference for a class of stochastic processes. *Habilitationsschrift, Fachbereich Mathematik, Universität Hamburg*.

[8] Mikosch, T. and Starica, C. (2004). Non-stationarities in financial time series, the long range dependence and the IGARCH effects. *The Review of Economics and Statistics*, 86, 378–390.

[9] Palma, W. and Olea, R. (2010). An efficient estimator for locally stationary Gaussian Long-Memory processes. *Annals of Statistics*, 38(5), 2958–2997.

[10] Paparoditis, E. (2009). Testing temporal constancy of the spectral structure of a time series. *Bernoulli*, 15, 1190–1221.

[11] Paparoditis, E. (2010). Validating stationarity assumptions in time series analysis by rolling local periodograms. *Journal of the American Statistical Association*, 105, 839-851.

[12] Preuß, P. and Vetter, M. and Dette, H. (2012). A test for stationarity based on empirical processes. *to appear in Bernoulli*.

[13] Preuß, P. and Vetter, M. (2012). Discriminating between long-range dependence and non-stationarity. *Technical Report*.

[14] Roueff, F. and von Sachs, R. (2011). Locally stationary long memory estimation. *Stochastic Processes and their Applications*, 121, 813–844.

[15] Terrin, N. and Taqqu, M. S. (1990). A Noncentral Limit Theorem for Quadratic Forms of Gaussian Stationary Sequences. *Journal of Theoretical Probability*, 3(3), 449–475.

[16] von Sachs, R. and Neumann, M. H. (2000). A wavelet-based test for stationarity. *Journal of Time Series Analysis*, 21, 597-613.

[17] Whittle, P. (1951). Hypothesis Testing in Time Series Analysis. *HUppsala: Almqvist and Wiksell*.

Hidden Markov models in modelling time series of earthquakes

Katerina Orfanogiannaki[*][1] **and Dimitris Karlis**[2]

[1] *Institute of Geodynamics, National Observatory of Athens, Greece*
[2] *Department of Statistics, Athens University of Economics and Business, Greece*

Abstract

Discrete valued Hidden Markov Models (HMMs) are used to model time series of event counts in several scientific fields. The model has two parts: the observed sequence of event counts and an unobserved (hidden) sequence of states that consist a Markov chain. Each state is characterized by a specific distribution and the progress of the hidden process from state to state is controlled by a transition probability matrix. In this work we aim to present an application of HMMs to a bivariate discrete valued time series occurring in seismology by extending the existing univariate models. In particular, on 26 December 2004 and 28 March 2005 occurred two of the largest earthquakes of the last 40 years between the Indo-Australian and the southeastern Eurasian plates with moment magnitudes $Mw = 9.1$ and $Mw = 8.6$ respectively. An interesting question is to examine whether the events can be correlated. To do so we examine the time series containing the daily number of events in the region of each mainshock. Our aim is firstly to identify the dynamics for each series separately by fitting univariate Poisson HMMs and secondly to account for any correlation between the two series.

While models for univariate discrete valued time series are well known we extend the HMMs to the bivariate case by assuming appropriate bivariate discrete distributions for each state. We examine properties of the model and propose inference. Maximum likelihood estimators of the models' parameters are derived using an EM algorithm.

Keywords: Hidden Markov Models; Poisson; Bivariate Poisson; earthquake counts
AMS subject classifications: 62M10

1 Introduction

HMMs are well known models with many applications incluing seismology (see [6], [3] and [8]). They allow for overdispersion (variance larger than the mean), autocorrelation and zero inflation in the data. Characteristics that are often present in many real data. While the literature on HMMs now flourishes, there is a lack of such models for multivariate count data, i.e. when many count random variables are observed together in different time points. Such models may arise in several disciplines and they also constitute a class of time series models for multivariate counts which is less developed and has itself particular interest.

We extend HMMs to the bivariate case by assuming both the standard bivariate Poisson distribution and a bivariate discrete distribution with Poisson marginals based on the Frank Copula for each state. We use the proposed models to jointly model the daily frequency of earthquakes in 2 adjacent areas in the Sumatra rupture zone. These areas were activated with a time difference approximately 3 months. The occurrence of an earthquake is a result of strain accumulation in the area, also known in the literature as ”tectonic loading”. When the amount of strain build up exceeds the ability of a fault to prevent slip, energy is released with the earthquake occurrence. Stress changes generated by large earthquakes influence the timing and locations of subsequent earthquakes. Usually the activity migrates from one area two the other and this appears as

[*] Corresponding author, e-mail: korfanogiannaki@gmail.com

negative correlation between the two time series. Our purpose is to use statistical models and particularly HMMs to examine the presence of correlation in the time series of the two strong earthquakes that occurred in the region of Sumatra in December 2004 and March 2005. Due to the nature of the data models that allow for both positive and negative correlation are required. The inclusion of copula in the model allows for more flexible dependence structure. The selection of the particular copula family is based on the fact that it allows not only for positive but also negative correlation. Of course, any other copula can be used. The models' parameters are estimated using an EM algorithm.

2 Proposed models

2.1 Hidden Markov models: the general context

HMMs are discrete time stochastic processes that consist of two parts. The first part is an unobserved finite state Markov chain $\{C_i : i \in N\}$ on m states. The second part is a non-negative integer valued sequence of random variables $\{Y_i : i \in N\}$ such that, conditionally on C_i are mutually independent. Each state is associated with a probability distribution function f from the same parametric family Λ. When C_i is known Y_i takes the value y_i with probability $f(y_i \mid c_i)$. If in the univariate case we assume that each observation is generated from a Poisson distribution then f takes the form $f(y_i|c_i = j) = \frac{\exp^{-\lambda_j} \lambda_j^{y_i}}{y_i!}$ where $\lambda_j \geq 0$ and $y_i = 0, 1, \ldots,$ for all $i = 1, \ldots, n$ and $j = 1, \ldots, m.$
The parameters of the model are the transition probabilities of the Markov chain and the parameters of the probability distributions that are associated with the states. The transition probabilities γ_{lj} are defined as: $\gamma_{lj} = P(C_i = j \mid C_{i-1} = l)$. This is the probability that given the hidden process was in state l at the previous time point, it will be in state j at the current. If we denote with Ψ the parameter vector to be estimated then the likelihood of an HMM is:

$$L(\Psi \mid y_1, \ldots, y_n) = \sum_{c_1=1}^{m} \cdots \sum_{c_n=1}^{m} P(C_1 = c_1) f(y_1 \mid c_1) \prod_{i=2}^{n} \gamma_{c_{i-1}c_i} f(y_i \mid c_i)$$

where n is the length of the observation sequence. For the likelihood of a HMM to be calculated, the backward, $\beta_j(i)$, and forward, $\alpha_j(i)$, probabilities were introduced by [1]: $\beta_j(i) = P(y_{i+1}, \ldots, y_n|C_i = j)$, $\alpha_j(i) = P(y_1, \ldots, y_i, C_i = j)$. The likelihood can then be calculated in terms of the forward probabilities as: $L = \sum\limits_{l=1}^{m} \alpha_l(n)$.
We will formulate 2 different models. The Bivariate Poisson Hidden Markov model (BPHMM) and the a Hidden Markov model based on a Frank copula (HMMC). What differentiates the models is the selection of the distribution family. In the BPHMMs each state corresponds to a standard bivariate Poisson distribution while in the HMMCs each state corresponds to bivariate distribution with Poisson marginals defined via a Frank copula.

Bivariate Poisson Model

In the BPHMM the distribution associated with each state is a bivariate Poisson distribution. Assume that X_i, are independent Poisson distributions with parameters λ_i, respectively where $i = 0, 1, 2$. The random variables Y_1, Y_2 defined as: $Y_1 = X_1 + X_0$ and $Y_2 = X_2 + X_0$ follow the bivariate Poisson distribution with parameters $(\lambda_0, \lambda_1, \lambda_2)$. In BPHMM each state is associated with a different bivariate Poisson distribution. Consider, that $\boldsymbol{\lambda^j} = (\lambda_0^j, \lambda_1^j, \lambda_2^j)'$ is the vector of parameters of the bivariate Poisson distribution that corresponds to state j and $\boldsymbol{y}_i = (y_{1i}, y_{2i})'$ for $i = 1, \ldots, n$ is the vector of observed data that corresponds to the i-th observation. The joint probability distribution of Y_1, Y_2 that corresponds to state j is given by the formula:

$$f(y_{1i}, y_{2i}) = e^{-(\lambda_1^j+\lambda_2^j+\lambda_0^j)} \frac{\lambda_1^{j\,y_{1i}}}{y_{1i}!} \frac{\lambda_2^{j\,y_{2i}}}{y_{2i}!} \sum_{r=0}^{min(y_{1i},y_{2i})} \left(\begin{array}{c} y_{1i} \\ r \end{array} \right) \left(\begin{array}{c} y_{2i} \\ r \end{array} \right) r! \left(\frac{\lambda_0^j}{\lambda_1^j \lambda_2^j} \right)^r ,$$

$y_1, y_2 = 0, \dots,$ and $\lambda_0^j, \lambda_1^j, \lambda_2^j \geq 0$, while $Cov(Y_{1t}, Y_{2t}) = \lambda_{0j} \geq 0$ is the covariance between the random variables Y_1 and Y_2 and since it is always equal to a non negative number, only positive correlation is allowed. When the covariance is equal to zero independence is implied.

Hidden Markov Models with Copula

Copulas are bivariate (multivariate) distributions with uniform marginals. Copulas are currently fashionable models to describe dependence. For a more formal definition see [7]. In the discrete case in order to derive the joint probability mass function (pmf) we need to take differences i.e.: for the bivariate case with marginals $F(x)$ and $G(y)$ the joint pmf is given by the formula:

$$f(y_{1i}, y_{2i}) = C(F(y_{1i}), G(y_{2i})) - C(F(y_{1i} - 1), G(y_{2i})) -$$
$$-C(F(y_{1i}), G(y_{2i} - 1)) + C(F(y_{1i} - 1), G(y_{2i} - 1))$$

where $F(\cdot)$ and $G(\cdot)$ are the marginal cdfs. This can be generalized to larger dimensions, but as dimensions increase excessive summation is needed. This creates a challenge on selecting copulas that can be efficient for the calculations. So, far we have worked with a bivariate Frank copula given by $C(u, v) = -\frac{1}{\tau} \log\{1 + \frac{(\exp^{-\tau u} - 1)(\exp^{-\tau v} - 1)}{(\exp^{-\tau} - 1)}\}$, where τ is the copula parameter representing the dependence implied. Frank copula allows for both negative and positive correlation. In addition the parameter of the Frank copula is unbounded and can take any real value. Finally, dependence in the Frank copula is symmetric in both tails. Of course any other copula can be used. In our case we have selected Poisson maginals.

3 Parameter estimation

Due to the underlying structure of HMMs that allow for a missing data representation of the model an EM algorithm ([2]) is adopted for Maximum Likelihood estimation of the parameters of interest both in the univariate and in the bivariate case. We define the indicator random variables $u_j(i)$ and $v_{jk}(i)$ respectively, where $u_j(i) = 1$, if $C_i = j$ and 0 otherwise and $v_{jk}(i) = 1$, if $C_{i-1} = j$ and $C_i = k$. The complete data log-likelihood, in terms of the indicator random variables is given by:

$$\log \pi_{c_1} + \sum_{i=2}^{n} \sum_{j=1}^{m} \sum_{k=1}^{m} v_{kj}(i) \log \gamma_{kj} + \underbrace{\sum_{i=2}^{n} \sum_{j=1}^{m} u_j(i) \log f(y_i \mid \lambda_j)}_{\text{weighted likelihood}}$$

At the E-step of the EM algorithm we estimate u and v through their conditional expectations: $\hat{u}_j(i) = P(C_i = j \mid y_1, \dots, y_n)$ and $\hat{v}_{jk}(i) = P(C_i = k, C_{i-1} = j \mid y_1, \dots, y_i)$. The backward and forward probabilities are used to calculate u and v and the likelihood of the model. The reader can find full details about how the backward and forward probabilities are computed in [6].

At the M-step we maximize the complete data log-likelihood. In the univariate case (for details see [6]) and in the case of bivariate Poisson (for details see [4]) the parameters can be estimated by closed form equations. In the model with copula there are no close equations so numerical maximization technics are used (we implemented them in R)

4 Application

4.1 Data

Two large earthquakes occurred between the Indo-Australian and the southeastern Eurasian plates on 26 December 2004 and 28 March 2005 with moment magnitudes $Mw = 9.1$ and $Mw = 8.6$ respectively. Both earthquakes were shallow and were followed by many aftershocks. The spatial distribution of aftershocks gives an approximation of the fault zone of each earthquake. In the case of the 2004 earthquake the rapture started from the South and propagated further North. While in the case of the 2005 earthquake the rupture followed the opposite direction. The regions that correspond to the rupture zones of the two earthquakes are denoted by N (North) and S (South), respectively. The fact that the two rupture zones do not overlap indicate that both earthquakes are mainshocks with their one aftershocks each. Earthquakes with moment magnitudes $mb \geq 4.2$ were selected from the USGS and ISC earthquake data files for the region defined by the rectangle with coordinates $-1.00N - 15.00N$ and $91.00E - 100.00E$. Each time series consists of earthquake counts, in daily time units, for the two regions N and S. The mean number (variances) of earthquakes per day is 0.14 (0.43) in the North region and 0.10 (0.22) for the South. The variance in the North region is almost 3 times the mean while in the South it is double than the mean. There is evidence of overdispersion, in both time series. 248 earthquakes have occurred in the North region and 185 in the South, in the time period examined. The maximum number of earthquakes observed in the North region is 12, while in the South is 14. The correlation between the two regions is 0.03. Hidden Markov models are adequate models to describe for both overdispersion and serial correlation that appear in the univariate series of earthquake counts. The extended HMMs to the bivariate case allow us to jointly model the frequency of earthquakes in the 2 regions and estimate the correlation between them.

	Independent PHMMs		BPHMMs		HMMC	
m	Loglik	BIC	Loglik	BIC	Loglik	BIC
2	-1291.183	2627.409	-1291.085	2642.226	-1254.663	2569.383
3	-1250.091	2590.268	-1248.038	2608.683	-1235.939	2584.486
4	-1239.323	2628.789	-1237.453	2655.076	-1225.423	2631.018

Table 1: Comparison of the fitted models on the basis of BIC. Key: m is the number of states

4.2 Results

We applied PHMMs to the 2 univariate time series of earthquake counts for different number of states. We also modeled jointly the two time series considering both BPHMMs and HMMC. The comparison between the different models was based on the values of the Bayesian information criterion (BIC) defined as: $BIC(m) = -2l(k) + \ln(n)d_f$, where $l(k)$ is the maximized log-likelihood for the model with m states, d_f is the number of free parameters of that model and n is the size of the sample. The values of the maximized log-likelihood and of BIC for the different models with different number of states are summarized in Table 1. The model that best describes the data is the HMMC with 2 states. The parameter estimates for the HMMC model with 2 states are shown in Table 2. Each column of the table corresponds to a different state. The first 2 rows correspond to the seismicity rates for the 2 regions while the last row corresponds to the parameter of the Frank copula.

The estimated transition probability matrix determined for the HMMC model with 2 states is

$$\hat{\Gamma} = \begin{bmatrix} 0.99 & 0.01 \\ 0.61 & 0.39 \end{bmatrix}.$$

State 1 corresponds to a state of seismic quiescence. State 2 is a very active state for the North region. When the two regions are in state 1 they remain in that state with probability 0.99. This state tries to describe the excessive number of zeros present in the two time series. The use of copula based multivariate discrete distributions allows for more flexible dependence structure.

	1st state	2st state
North	0.076	4.020
South	0.094	0.077
copula parameter	0.825	2.844

Table 2: Parameter estimates for the HMMC model with 2 states.

Acknowledgements: This research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program ”Education and Lifelong Learning” of the National Strategic Reference Framework (NSRF) - Research Funding Program: Heracleitus II. Investing in knowledge society through the European Social Fund.

Bibliography

[1] Baum, L.E., Petrie, T., Soules, G. and Weiss, N. (1970). A maximization technique in the Statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics* 41, 164–171.

[2] Dempster, A.P., Laird, N.M. and Rubin, D.B (1997). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B* 39, 1–38.

[3] Ebel, J.E., Chambers, D.W., Kafka, A.L. and Baglivo, J. A. (2007). Non-Poissonian Earthquake Clustering and the Hidden Markov Model as Bases for Earthquake Forecasting in California. *Geophys. Res. Lett.* 78, 57–65.

[4] Karlis, D and Ntzoufras, J. (2003). Analysis of Sports Data Using Bivariate Poisson Models. *The Statistician* 52, 381–393.

[5] Lay, T., Kanamori, H., Ammon, C.J., Nettles, M., Ward, S.N., Aster, R.C., Beck, S.L., Bilek, S.L., Brudzinski, M.R., Butler, R., DeShon, H.R., Ekstrom, G., Satake, K., Sipkin, S. (2005). The great Sumatra-Andaman earthquake of 26 December 2004.*Science* 308, 1127-1133

[6] MacDonald, I.L. and Zucchini, W. (1997). *Hidden Markov models and other models for discrete-valued time series*, Chapman and Hall, London.

[7] Nelsen, R. (2006). *An Introduction to copulas, 2nd Edition*, Springer.

[8] Orfanogiannaki, K., Karlis, D. and Papadopoulos, G.A. (2010). Identifying seismicity patterns using Poisson Hidden Markov Models. *Pure and Applied Geophysics* 167, 919–931.

Approximating the posterior distribution of mixture weights with application to transcript expression estimation

Panagiotis Papastamoulis[*1] **and Magnus Rattray**[1]

[1] *Faculty of Life Sciences, University of Manchester*

Abstract

This study focuses on approximating the posterior distribution of mixture weights ($\boldsymbol{\theta}$) given some data ($\boldsymbol{x}$) using Variational Bayes (VB) methods [1]. Standard VB implementation [4] for this problem approximates the joint posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{x})$ of parameters and latent variables ($\boldsymbol{z}$). It is demonstrated via simulation that this approach leads to variance underestimation. For this reason a new variational scheme is developed by integrating out the latent variables and targeting the marginal posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{x})$. The new approximation belongs to the richer family of Generalized Dirichlet distributions [8], while a stochastic approximation algorithm [6] performs the optimization in the corresponding spaces arising from two different parameterizations. Moreover, it is proven that the new solution leads to a better marginal log-likelihood bound compared to the former.

The method is applied to transcript expression estimation using high throughput sequencing of RNA (RNA-seq) technology. Mixture models are a natural way to deal with such problems, and Gibbs sampling has already been applied [3]. The application of Variational methods to these datasets is novel and leads to encouraging results. Finally, the variational solution is exploited in order to improve Markov Chain Monte Carlo (MCMC) sampling with the Delayed Rejection algorithm [7].

Keywords: Kullback-Leibler divergence, marginal likelihood bound, BitSeq, RNA-seq.
AMS subject classifications: 62F15, 81T80, 92B15.

1 Introduction

Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ denote n independent observations identically distributed according to a mixture of $K > 1$ known distributions, $f_1, \ldots, f_K$, that is,

$$x_i \sim \sum_{k=1}^{K} \theta_k f_k(x_i), \quad i = 1, \ldots, n. \tag{1}$$

Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{K-1}) \in \Theta_K := \{\theta_k > 0, k = 1, \ldots, K-1 : \sum_{k=1}^{K-1} \theta_k < 1\}$ denote the unknown weights with $\theta_K := 1 - \sum_{k=1}^{K-1} \theta_k$. Moreover, let $\boldsymbol{z}_i := (z_{i1}, \ldots, z_{iK})$ be the latent vector which assigns the i-th observation to one of the components, that is, $x_i | z_{ik} = 1 \sim f_k(x_i)$, with $\boldsymbol{z}_i|\boldsymbol{\theta} \sim \mathcal{M}(1; \theta_1, \ldots \theta_K)$, independently for $i = 1, \ldots, n$, where $\mathcal{M}$ denotes the multinomial distribution.

Under a Bayesian setup, let $p(\boldsymbol{\theta})$ be the prior distribution, which in our context is a Dirichlet $\mathcal{D}(\alpha_1, \ldots, \alpha_K)$. The marginal likelihood, defined as $m(\boldsymbol{x}) := \int_{\Theta_K} p(\boldsymbol{\theta}|\boldsymbol{x})p(\boldsymbol{\theta})d\boldsymbol{\theta}$, is an important quantity to estimate because it allows for model selection. MCMC estimation of $m(\boldsymbol{x})$ is possible but not straightforward (see for example [2], p.139). VB methods [1] provide an attractive alternative based on an approximating distribution, while the model selection problem is dealt by providing a lower bound to the marginal likelihood.

[*] Corresponding author, e-mail: panagiotis.papastamoulis@manchester.ac.uk

Standard implementation of VB methodology [4] approximates jointly $\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{x}$, rather than the actual (non-augmented) posterior. In this paper we show that it is better to approximate $p(\boldsymbol{\theta}|\boldsymbol{x})$. The proposed methodology exploits the solution of standard VB by performing an optimization into a class of distributions that share the same mean with the initial solution. Two different parameterizations are taken into account: the first one forces the approximating distribution to remain inside the Dirichlet family, while the second one relaxes this assumption by using the Generalized Dirichlet family.

The rest of the paper is organized as follows. Standard VB implementation is briefly described in Section 2. In Section 2.1 a better bound is constructed and the optimization problem is stated in its general form. Moreover, two different parameterizations for the optimization problem are given. The methodology is illustrated in a simulation study and a real RNA-seq dataset in Sections 3.1 and 3.2. Finally, Section 3.3 uses the VB approximations in the Delayed Rejection MCMC algorithm.

2 Variational Approximation

VB methods aim at finding a lower-bound (L) of $\log m(\boldsymbol{x})$, by performing a free-form minimization of the Kullback-Leibler divergence $\mathrm{KL(q||p)}$ between an approximating distribution q and the target distribution p. Hence, we may write

$$\log m(\boldsymbol{x}) = L + \mathrm{KL(q||p)}. \tag{2}$$

According to the standard VB methodology [4], the joint posterior $p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{x})$ is approximated by another distribution $q(\boldsymbol{\theta}, \boldsymbol{z})$. In order to make the problem tractable this minimization is done considering the family of distributions

$$\mathcal{G} = \{g(\boldsymbol{\theta}, \boldsymbol{z}) = g(\boldsymbol{\theta})g(\boldsymbol{z}) : g(\boldsymbol{z}) = \prod_{i=1}^{n}\prod_{k=1}^{K} \phi_{ik}^{z_{ik}}\}, \tag{3}$$

where $\boldsymbol{\phi} := \{\phi_{ik} : i = 1, \ldots, n, k = 1, \ldots, K\}$ are the variational parameters. It turns out that the approximate distribution for $\boldsymbol{\theta}$ is

$$q(\boldsymbol{\theta}) = \mathcal{D}(\gamma_k; k = 1, \ldots, K), \tag{4}$$

$\gamma_k := \alpha_k + \sum_{i=1}^{n} \phi_{ik}$, and the optimization with respect to $\boldsymbol{\phi}$ is done using a steepest descent algorithm.

Figure 1 displays the estimates of $\theta_k|\boldsymbol{x}$ based on the simulation studies in Section 3. The black lines are considered as the ground "truth" and they are estimated by a long MCMC run, while the dashed ones are the estimates corresponding to the standard VB method. It is obvious that this approach exhibits good performance in terms of posterior means, but it leads to variance underestimation.

2.1 Bounding the non-augmented posterior

The distribution in (4) is optimal in terms of minimizing the KL divergence between the *joint* posterior $p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{x})$ and the distributions considered in (3). However, this does not mean that it is the "best" Dirichlet approximation of the marginal posterior $p(\boldsymbol{\theta}|\boldsymbol{x})$. This is proven in the following proposition.

Proposition 2.1. *Let $\mathcal{F}$ denote any subset/family of distributions with $q(\boldsymbol{\theta}) \in \mathcal{F}$. Then,*

$$\min_{f\in\mathcal{F}} \mathrm{KL}(f(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\boldsymbol{x})) \leqslant \mathrm{KL}(q(\boldsymbol{\theta}, \boldsymbol{z})||p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{x})), \tag{5}$$

and the equality holds if and only if $q(\boldsymbol{\theta}, \boldsymbol{z}) = p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{x})$, $\forall \boldsymbol{\theta}, \boldsymbol{z}$.

Proof. By the log-sum inequality, $\forall \boldsymbol{\theta} \in \Theta$

$$q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\boldsymbol{x})} = \left(\sum_{\boldsymbol{z}} q(\boldsymbol{\theta}, \boldsymbol{z})\right) \log \frac{\sum_{\boldsymbol{z}} q(\boldsymbol{\theta}, \boldsymbol{z})}{\sum_{\boldsymbol{z}} p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{x})} \leqslant \sum_{\boldsymbol{z}} q(\boldsymbol{\theta}, \boldsymbol{z}) \log \frac{q(\boldsymbol{\theta}, \boldsymbol{z})}{p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{x})} \Rightarrow$$

$$\int q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\boldsymbol{x})} d\boldsymbol{\theta} \quad \leqslant \quad \int \sum_{\boldsymbol{z}} q(\boldsymbol{\theta}, \boldsymbol{z}) \log \frac{q(\boldsymbol{\theta}, \boldsymbol{z})}{p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{x})} d\boldsymbol{\theta} \Leftrightarrow$$

$$\mathrm{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|x)) \leqslant \mathrm{KL}(q(\boldsymbol{\theta}, \boldsymbol{z})||p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{x})).$$

Hence (5) stems by the assumption that $q(\boldsymbol{\theta}) \in \mathcal{F}$. □

Now, for a given family $\mathcal{F}$, let $\boldsymbol{\delta} \in \Delta_{\mathcal{F}}$ denoting the corresponding (possibly high-dimensional) parameter space and $f \in \mathcal{F}$. Equation (2) implies that the lower bound (L) of the log-marginal likelihood corresponding to $f(\cdot; \boldsymbol{\delta})$ can be expressed as

$$L(\boldsymbol{\delta}) \quad = \quad \int_{\Theta_K} \{\log p(\boldsymbol{x}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) - \log f(\boldsymbol{\theta}; \boldsymbol{\delta})\} f(\boldsymbol{\theta}; \boldsymbol{\delta}) d\boldsymbol{\theta}. \tag{6}$$

We have to stress the fact that (6) cannot be computed directly even for fixed $\boldsymbol{\delta}$. However, (6) can be approximated via simulation, since it is expressed as the mean value of the random variable $g(\boldsymbol{\theta}) := \log p(\boldsymbol{x}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) - \log f(\boldsymbol{\theta}; \boldsymbol{\delta})$, $\boldsymbol{\theta} \sim f(\cdot; \boldsymbol{\delta})$. So, our objective function is written as:

$$\max_{\boldsymbol{\delta} \in \Delta_{\mathcal{F}}} L(\boldsymbol{\delta}) = \max_{\boldsymbol{\delta} \in \Delta_{\mathcal{F}}} \mathbb{E}_{\boldsymbol{\delta}} g(\boldsymbol{\theta}), \quad \boldsymbol{\theta} \sim f(\cdot; \boldsymbol{\delta}) \in \mathcal{F}. \tag{7}$$

Having in mind that the best variational approximation targeting the joint posterior is the Dirichlet distribution in (4), an obvious choice for $\mathcal{F}$ would be (a subset of) the Dirichlet family of distributions. However, it will prove useful to take into account an even broader family as well, that is, the Generalized Dirichlet family of distributions [8]. The VB solution (4) can be expressed as a Generalized Dirichlet distribution: $q(\boldsymbol{\theta}) = \mathcal{D}(\gamma_1, \ldots, \gamma_K) \equiv \mathcal{GD}(\gamma_1, \ldots, \gamma_{K-1}; \gamma_1^+, \ldots, \gamma_{K-1}^+)$, $\gamma_\ell^+ := \sum_{j=\ell+1}^K \gamma_\ell$, $\ell = 1, \ldots, K-1$. Next we define two specific sets $\mathcal{F}_{\mathcal{D}}$ and $\mathcal{F}_{\mathcal{GD}}$, with $\mathcal{F}_{\mathcal{D}} \subset \mathcal{F}_{\mathcal{GD}}$, in order to make the optimization tractable in (7). Our guide is to keep the same mean as the original VB distribution (4). These two sets are the following

$$\mathcal{F}_{\mathcal{D}} := \{\mathcal{D}(e^{\delta}\gamma_1, \ldots, e^{\delta}\gamma_K) : \delta \in \mathbb{R}\}, \tag{8}$$

$$\begin{aligned} \mathcal{F}_{\mathcal{GD}} := \{\mathcal{GD}(e^{\delta_1}\gamma_1, \ldots, e^{\delta_{K-1}}\gamma_{K-1}; e^{\delta_1}\gamma_1^+, \ldots, e^{\delta_{K-1}}\gamma_{K-1}^+) : \\ \delta_k \in \mathbb{R}, k = 1, \ldots, K-1\}. \end{aligned} \tag{9}$$

Note that the number of parameters in (8) and (9) equals to one and K, respectively. Moreover, for all $f \in \mathcal{F}_{\mathcal{GD}}$ it holds that $\mathbb{E}\theta_k = \gamma_k / \sum_{j=1}^K \gamma_j$, $\forall k = 1, \ldots, K$, while the same remains true for $f \in F_{\mathcal{D}}$ as well, since $\mathcal{F}_{\mathcal{D}} \subset \mathcal{F}_{\mathcal{GD}}$. Consequently, both families (8) and (9) contain distributions having the same means as the distribution $q(\boldsymbol{\theta})$ in (4). In order to maximize (7) under parameterizations (8) or (9) a stochastic approximation algorithm [6] was implemented.

3 Applications

3.1 Simulated Data

Let e_j, $j = 1, \ldots, 4$ denote given sequences with replacement of the letters “A”, “T”, “C” and “G”, having lengths equal to $1000, 5, 5, 1000$, respectively. Moreover, consider three discrete sets I_1, I_2, I_3, arising by joining different combinations of e_j, $j = 1, \ldots, 4$ one after the other. In particular, $I_1 = \{e_1, e_2\}$, $I_2 = \{e_2, e_4\}$ and $I_3 = \{e_2, e_3, e_4\}$. Let now $x_i \sim \sum_{k=1}^K \theta_k \mathcal{U}_{I_k}$, $K = 3$, $i = 1, \ldots, 2000$ be randomly sampled short sequences of 50 consecutive letters from a mixture of uniform distributions defined in I_k, $k = 1, 2, 3$. The true values of the weights used for the simulation is $\boldsymbol{\theta} = (2/9, 2/9, 5/9)$.

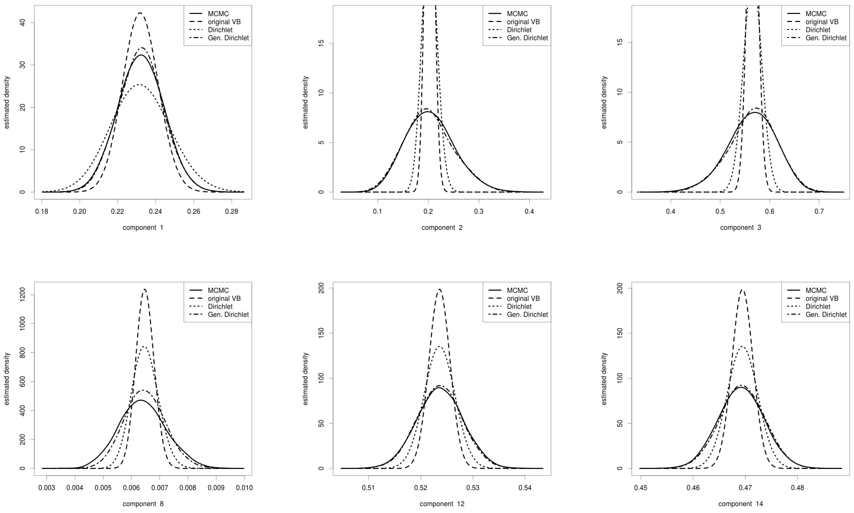


Figure 1: Density estimates of $\theta_k|\boldsymbol{x}$. Up: Simulated data $k = 1, 2, 3$. Down: RNA-seq data $k = 8, 12, 14$.

After imposing a $\mathcal{D}(1, 1, 1)$ prior on $\boldsymbol{\theta}$, we applied the three VB algorithms. The estimates of the lower bound of $\log m(\boldsymbol{x})$ are shown in Table 1. The first row of Figure 1 displays the estimated posterior marginal densities. Compared to a long MCMC run estimate, we conclude that the Dirichlet modification is better than standard VB, however the problem of variance underestimation is still apparent. Finally, the Generalized Dirichlet distribution is quite close to the one estimated by MCMC.

In order to make a connection with the next section, under a biological framework the 2000 observations would be called short reads, while the terms exons and transcripts refer to the sets e_j and I_k, respectively. Transcriptome is the set of all available transcripts and it is considered known. RNA-seq is a technology aiming to identify and quantify mRNA transcripts in a biological sample of short reads from the transcriptome. Some of the transcripts share much of their sequence (exons), hence the origin of a sampled read is unknown. In statistical terms, this problem reduces to estimate the weights of a mixture model, see [3] for details.

3.2 RNA-seq Data

A sample of human brain tissue reads was downloaded from NCBI (accession number GSM343511). This is part of a much bigger study (see [5]), but for our illustration we used as reference the gene ENSG00000102078. The resulting sample consists of $n = 61875$ reads and the number of components (transcripts) is equal to $K = 14$. Finally, we used the methodology described in [3] in order to compute the likelihood of the reads to the transcripts.

The estimates of the lower bound of $\log m(\boldsymbol{x})$ are shown in Table 1. Once again the marginal densities arising from the optimization in $\mathcal{F}_{\mathcal{GD}}$ are quite close to the ones obtained by a long MCMC run, as displayed in the second row of Figure 1 (only the 3 more highly expressed transcripts are shown).

3.3 Comparing the Approximations and Improving MCMC

The distributions arising from the VB methods can be used in order to obtain a MCMC sample from the posterior, via the two-stage Delayed Rejection MCMC technique [7]. At the 1st stage a value is generated

Dataset	standard VB	Dirichlet	Gen.Dirichlet
Section 3.1	-15341.24 (15.72%)	-15340.26 (31.00%)	-15339.38 (96.46%)
Section 3.2	-1367478.31 (2.26%)	-1367475.71 (17.6%)	-1367474.57 (39.64%)

Table 1: Marginal log-likelihood bounds according to the three VB methods. The percentages correspond to the 1st stage acceptance rate discussed in Section 3.3.

by an approximating distribution, independently from the current state of the chain. This is accepted with the usual Metropolis-Hastings acceptance ratio. If it is rejected, a random walk proposes a second candidate state, which is based on the previous state of the chain (details not shown here). Clearly, the 1st stage acceptance rate should be larger as the Variational approximation gets "better", while this can serve as a measure of efficiency of the approximation. Moreover, a high 1st stage acceptance rate improves the mixing of the MCMC sample, as the 1st stage draws are uncorrelated. A long MCMC run resulted to the 1st stage acceptance rates shown in Table 1, highlighting the improved performance of the proposed method over the standard VB.

Acknowledgements: The first author wishes to thank Associate Professor Ioannis Ntzoufras for recommending him to participate in the 18th EYSM. This work was supported by a BBSRC award ref. BB/J009415/1.

Bibliography

[1] Bishop, C. (2006). *Pattern Recognition and Machine Learning*, Springer, New York.

[2] Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*, Springer, New York.

[3] Glaus, P., Honkela, A. and Rattray, M. (2012). Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics* 28, 1721–1728.

[4] Hensman, J., Rattray, M. and Lawrence, N.D. (2012). Fast Variational Inference in the Conjugate Exponential Family. *Advances in Neural Information Processing Systems*, arXiv:1206.5162v2.

[5] Pan, Q., Shai, O., Lee, L.J., Frey, B.J. et al. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics* 40:1413–1415.

[6] Spall, J.C. (1992). Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control* 37, 332–341.

[7] Tierney, L. and Mira, A. (1999). Some adaptive Monte Carlo methods for Bayesian inference. *Statistics in Medicine* 18, 2507–2515.

[8] Wong, T.T. (1998). Generalized Dirichlet distribution in Bayesian analysis. *Applied Mathematics and Computation* 97, 165–181.

Models for dependent paired comparison data

Manuela Cattelan[*][1] **and Cristiano Varin**[2]

[1] *Department of Statistical Sciences, University of Padova*
[2] *D.A.I.S., Università Ca' Foscari Venezia*

Abstract

Paired comparison data are often analyzed employing regression models in which the probability that an item wins the comparison with another item is a function of the difference between the "worth" of the two items. Traditional models generally assume that all comparisons are independent. This assumption is often unrealistic, since it is difficult to believe that, for example, the results of two matches with a player in common are independent. Here, two different approaches that account for dependence in the data are illustrated. The first one is a random-effects model designed in way to produce a scheme of cross-correlations between observations with common items. The second approach is a marginal model specified only through means and covariances reflecting comparisons dependencies. Both approaches pose inferential difficulties either because the likelihood is computationally complex or because the joint distribution of the data is unavailable. These difficulties are overcome by means of different forms of composite likelihood inference.

Keywords: Bradley-Terry model, optimal estimating equations, paired comparisons, pairwise likelihood, Thurstone model.
AMS subject classifications: 62F10, 62J12.

1 Introduction

Paired comparison data derive from the comparison of objects or items in couples. This type of data can be encountered in many areas, including marketing and consumer behavior data, sport data, psychometric experiments and many more. In some instances there is a person that performs the paired comparisons, as in psychometric experiments, but there may also be a direct comparison between items, as in sport data.

Let Y_{ij} denote the result of the comparison between item i and item j. Then $Y_{ij} = 1$ if i wins the comparison against j, $Y_{ij} = 0$ otherwise. Let μ_i denote the "worth" or ability of item i, $i = 1, \ldots, n$, then traditional models for the analysis of paired comparison data assume that the probability that i wins against j is

$$\pi_{ij} = \mathrm{E}(Y_{ij}) = F(\mu_i - \mu_j), \tag{1}$$

where $F(\cdot)$ is the cumulative distribution function of a zero-symmetric random variable. The classical models employed for the analysis of paired comparison data are the Thurstone model [9], which assumes that F in formula (1) is the cumulative distribution function of a standard normal random variable, and the Bradley-Terry model [1] that specifies a logistic distribution function F. If there are explanatory variables, [8] suggests to set

$$\mu_i = x_i^T \beta,$$

where β is a d-dimensional vector of regression parameters.

Traditional models for the analysis of paired comparison data assume that all comparisons are independent. However, this assumption appears unrealistic since it implies that, for example, the results of two matches

[*] Corresponding author, e-mail: manuela.cattelan@unipd.it

involving a common player are independent. We illustrate two possible approaches to extend traditional models in order to account for dependence in the data. The first extension specifies a conditional model in which a multivariate distribution of all observations is described. The second approach accounts for dependence in the data without specifying a whole multivariate distribution, but only the first two moments.

2 Conditional models

It is possible to model the dependence among the observations through the inclusion of item-specific random effects [5]. In this case the worth of item i is described as

$$\mu_i = x_i^T \beta + U_i,$$

where U_i, $i = 1, \ldots, n$, are zero-mean independent random effects with density function $f(\cdot; \sigma^2)$ that depends on the parameter σ^2. The random effects allow to account also for the imperfect representation of the worth by the linear predictor.

The binary observations may be represented as censored continuous latent variables such that $Y_{ij} = 0$ iff $Z_{ij} < 0$ where

$$Z_{ij} = (x_i - x_j)^T \beta + U_i - U_j + \epsilon_{ij},$$

where ϵ_{ij} are independent zero-mean continuous random variables. Computational complexity is reduced if we assume that the random effects are normally distributed with mean zero and variance σ^2 and that the comparison-specific errors are normally distributed with mean 0 and variance 1, and they are independent of the random effects. The variance of the errors is set to 1 for identification purposes. Then $Z_{ij} \sim N((x_i - x_j)^T \beta; 1 + 2\sigma^2)$ and the correlation between two latent variables is

$$\text{corr}(Z_{ij}, Z_{kl}) = \begin{cases} \sigma^2/(1+2\sigma^2), & \text{if } i = k \text{ or } j = l, \\ -\sigma^2/(1+2\sigma^2), & \text{if } i = l \text{ or } j = k, \\ 0, & \text{if } i \neq j \neq k \neq l. \end{cases} \tag{2}$$

Hence, if two paired comparisons have an item in common, they are correlated, otherwise they are independent. The above specification implies that the correlation $\sigma^2/(1+2\sigma^2)$ lies in the interval $(0, 0.5)$.

Unfortunately, the likelihood function associated with the random effects model requires the approximation of an integral of dimension equal to the number of comparisons

$$L(\theta; y) = \int_{R^n} \left\{ \prod_{i=1}^{n-1} \prod_{j=(i+1)}^{n} \mathrm{P}(Y_{ij} = y_{ij} | U_i = u_i, U_j = u_j; \theta) \right\} \left\{ \prod_{i=1}^{n} \phi(u_i; \theta) \mathrm{d}u_i \right\},$$

where $\theta = (\beta, \sigma^2)$, $y = (y_{12}, \ldots y_{n-1\,n})$, and $\phi(\cdot)$ is the density function of a standard normal random variable.

Considering the latent variable specification, the likelihood function can be written as an integral with dimension equal to the number of paired comparisons

$$L(\theta; y) = \int_{A_{12}} \cdots \int_{A_{n-1\,n}} \phi_N(v; \Sigma) \mathrm{d}v,$$

where

$$A_{ij} = \begin{cases} (-\infty, -(x_i - x_j)^T \beta / \sqrt{1+2\sigma^2}), & \text{if } y_{ij} = 0, \\ (-(x_i - x_j)^T \beta / \sqrt{1+2\sigma^2}, +\infty), & \text{if } y_{ij} = 1, \end{cases}$$

$\phi_N(\cdot; \Sigma)$ denotes the density function of an N-dimensional normal random variable with correlation matrix Σ, N denotes the total number of paired comparisons and the elements of Σ are as shown in (2). This specification requires the approximation of a normal multivariate integral of dimension equal to the number of paired comparisons observed. Since the dimension of the integral can be very high, [2] suggests to employ pairwise likelihood to make inference in this model.

2.1 Pairwise likelihood estimation

Pairwise likelihood is an instance of composite likelihoods [10] that consists of the product of all marginal bivariate probabilities. In the paired comparisons context, the pairwise likelihood is the product of all bivariate probabilities of all couples of comparisons

$$PL(\theta; y) = \prod_{(ij)\neq(kl)} \mathrm{P}(Y_{ij} = y_{ij}, Y_{kl} = y_{kl}; \theta). \tag{3}$$

Under regularity conditions, the maximum pairwise likelihood estimator is asymptotically normally distributed with mean θ and covariance matrix $H(\theta)^{-1}J(\theta)H(\theta)^{-1}$, where $J(\theta) = \mathrm{var}(\nabla pl(\theta; Y)$, $H(\theta) = E(-\nabla^2 pl(\theta; Y))$ and $pl(\theta; Y) = \log PL(\theta; Y)$.

The use of pairwise likelihood noticeably reduces the computational complexity. For example, if we consider the censored latent random variable specification, then the bivariate probability that i loses the comparisons both against j and k is

$$\mathrm{P}(Y_{ij} = 0, Y_{ik} = 0) = \mathrm{P}(Z_{ij} < 0, Z_{ik} < 0) =$$
$$= \Phi_2\left(-\frac{(x_i - x_j)^T\beta}{\sqrt{1+2\sigma^2}}, -\frac{(x_i - x_k)^T\beta}{\sqrt{1+2\sigma^2}}; \frac{\sigma^2}{1+2\sigma^2}\right),$$

where $\Phi_2(\cdot, \cdot; \rho)$ denotes the cumulative distribution function of a bivariate normal random variable with standard marginals and correlation ρ. The pairwise likelihood (3) requires the approximation of at most bivariate normal integrals. Simulation studies presented in [2] show that pairwise likelihood estimators perform well with a modest loss of efficiency.

3 Marginal models

In some instances, one may be unwilling to specify the whole distribution of the data. In these cases, it is still possible to extend the traditional models to take into account the dependence in the data, but specifying only the first two moments of the distribution.

The maximum likelihood estimates of the regression parameters in both the Bradley-Terry and the Thurstone models are computed by solving the equations

$$DV^{-1}(y - \pi) = 0,$$

where D denotes the Jacobian of $\pi = (\pi_{12}, \dots, \pi_{n-1\,n})$ with respect to the components of β, and V is the covariance matrix computed under the assumption of independence, hence it is a diagonal matrix with entries $\pi_{ij}(1 - \pi_{ij})$. Under the independence assumption, $\hat{\beta}$ has an asymptotically normal distribution with mean β and variance $(D^TV^{-1}D)^{-1}$. Dependence can be accounted for by substituting the variance matrix computed under the independence assumption V with a non-diagonal covariance matrix W in which not all $\mathrm{cov}(Y_{ij}, Y_{kl})$ are zeros.

The classical measure of dependence in binary data is the cross-ratio

$$\psi_{ij,kl} = \frac{\mathrm{P}(Y_{ij} = 1, Y_{kl} = 1)\mathrm{P}(Y_{ij} = 0, Y_{kl} = 0)}{\mathrm{P}(Y_{ij} = 1, Y_{kl} = 0)\mathrm{P}(Y_{ij} = 0, Y_{kl} = 1)}.$$

It is reasonable to assume that only comparisons with an item in common are dependent, so $\psi_{ij,kl} = 1$ if $i \neq j \neq k \neq l$. Moreover, paired comparison models must assure the symmetry condition $\mathrm{P}(Y_{ij} = 1) =$

P$(Y_{ji} = 0)$, hence it follows that $\psi_{ij,ik} = 1/\psi_{ij,ki}$. Thereafter, we assume a common cross-ratio ψ for all couples

$$\psi_{ij,kl} = \begin{cases} \psi, & \text{if } i = k \text{ or } j = l, \\ 1/\psi, & \text{if } i = l \text{ or } j = k, \\ 0 & \text{otherwise.} \end{cases}$$

Hence, following [4], the bivariate probability of observing a win for i against both j and k is

$$\text{pr}(Y_{ij} = 1, Y_{ik} = 1) = \begin{cases} \pi_{ij}\pi_{ik}, & \text{if } \psi = 1, \\ \dfrac{1 + (\pi_{ij} + \pi_{ik})(\psi - 1) - G(\pi_{ij}, \pi_{ik}, \psi)}{2(\psi - 1)}, & \text{if } \psi \neq 1, \end{cases} \tag{4}$$

where $G(\pi_{ij}, \pi_{ik}, \psi) = \sqrt{\{1 + (\pi_{ij} + \pi_{ik})(\psi - 1)\}^2 + 4\psi(1 - \psi)\pi_{ij}\pi_{ik}}$. The probabilities of the other three possible combinations of results can be computed from equation (4) and marginal univariate probabilities.

The elements of the covariance matrix W can be computed as cov$(Y_{ij}, Y_{ik}) =$ P$(Y_{ij} = 1, Y_{ik} = 1) -$ P$(Y_{ij} = 1)$P$(Y_{ik} = 1)$. Since they depend also on the regression parameters β, it is not possible to employ standard generalized estimating equations [7]. For this reason, we resort to the hybrid pairwise likelihood method to estimate this marginal model extension.

3.1 Hybrid pairwise likelihood

Hybrid pairwise likelihood [6] suggests to iterate between solving optimal estimating equations for estimation of the regression parameters and maximizing the pairwise likelihood equation for estimation of the dependence parameter. Given the dependence parameter ψ, the regression parameters estimates $\hat{\beta}_{\text{dep}}(\psi)$ are computed by solving the equations

$$DW^{-1}(y - \pi) = 0,$$

while, for a fixed β, the estimate $\hat{\psi}(\beta)$ is obtained maximizing the pairwise likelihood

$$PL(\theta; y) = \prod_{(ij),(kl)} \text{P}(Y_{ij} = y_{ij}, Y_{ik} = y_{kl}; \theta).$$

The procedure iterates between the solution of optimal estimating equations for β given $\hat{\psi}(\hat{\beta}_{\text{dep}})$ and maximum pairwise likelihood estimation of ψ given $\hat{\beta}_{\text{dep}}(\hat{\psi})$. At convergence, the estimates of the regression coefficients are asymptotically normally distributed with mean β and covariance $(D^T W^{-1} D)^{-1}$. This procedure requires only the specification of the first two moments of the distribution both for estimation of the regression parameters and computation of their standard errors [3].

4 Conclusions

Inference in traditional models for the analysis of paired comparison data is performed by assuming independence among all observations. Often, this assumption is unrealistic, therefore different extensions that account for dependence among observations are proposed. The first extension specifies a conditional model in which dependence is introduced by item-specific random effects. The conditional models extension describes the whole distribution of all observations and inferential difficulties are overcome by means of pairwise likelihood methods.

The second approach proposed specifies a marginal model, which requires assumptions only about the first two moments of the distribution of the data. In this case, the recourse to the hybrid pairwise likelihood method is suggested.

Pairwise likelihood provides a straightforward solution to the problem of estimating complex models. This characteristic and the nice theoretical properties possessed by pairwise likelihood, have promoted the use of this estimating method in many contexts [10]. In particular, we employ pairwise likelihood in the conditional model to overcome the problem of computation or approximation of a high dimensional integral. In the marginal model proposed, pairwise likelihood is used in combination with optimal estimating equations in order to estimate the dependence parameter, which cannot be recovered from estimating functions that do not depend on the regression parameters.

Acknowledgements: The first Author acknowledges the financial support of the *Finanziamento Giovani Studiosi* for the project "Inferential Issues in Regression Models for Dependent Categorical and Discrete Data" carried out at the Department of Statistical Sciences, University of Padova.

Bibliography

[1] Bradley, R. A. and Terry, M. E. (1952). Rank Analysis of Incomplete Block Designs. I. The Method of Paired Comparisons. *Biometrika* 39, 324–345.

[2] Cattelan, M. (2009). Correlation Models for Paired Comparison Data. *PhD Thesis*, Department of Statistical Sciences, University of Padova.

[3] Cattelan, M. and Varin, C. (2013). Hybrid Pairwise Likelihood Analysis of Animal Behavior Experiments. *Submitted*.

[4] Dale, J. R. (1986). Global Cross-Ratio Models for Bivariate, Discrete, Ordered Responses. *Biometrics* 42, 909–917.

[5] Firth, D. (2005). Bradley-Terry Models in R. *Journal of Statistical Software* 12, 1–12.

[6] Kuk, A. Y. C. (2007). A Hybrid Pairwise Likelihood Method. *Biometrika* 94, 939–952.

[7] Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika* 73, 13–22.

[8] Springall, A. (1973). Response Surface Fitting Using a Generalisation of the Bradley-Terry Paired Comparison Model. *Journal of the Royal Statistical Society Series C - Applied Statistics* 22, 59–68.

[9] Thurstone, L. L. (1927). A Law of Comparative Judgement. *Psychological Review* 34, 368–389.

[10] Varin, C., Reid, N. and Firth, D. (2011). An Overview of Composite Likelihood Methods. *Statistica Sinica* 21, 5–42.

Confidence sets from empirical Bayes procedures with conditionally Gaussian priors on Sobolev balls

Botond Szabó[*][1]

[1] *Eindhoven University of Technology*

Abstract Consider the problem of constructing Bayesian based confidence sets that are adaptive in L^2-loss over a continuous scale of Sobolev classes in the Gaussian White noise model. We show that both the hierarchical Bayes and marginal likelihood empirical Bayes approaches lead to credible sets with asymptotic coverage zero for certain oddly behaving functions. Then we give a new empirical Bayes method based on the results of [7], which solves this problem and provides uniform and adaptive confidence sets over a whole collection of Sobolev classes.

Keywords: Nonparametric Bayes, adaptation, credible sets, coverage, Gaussian processes.
AMS subject classifications: Primary 62G15; secondary 62G20.

Introduction

Adaptive techniques for nonparametric estimation have been widely studied in the literature and many rate-adaptive results have been provided for a variety of statistical problems. However, an adaptive estimator without any knowledge of its uncertainty is rather uninformative, since one knows that the estimator is optimally close to the true function, but has no information about the actual distance. The uncertainty of an estimator can be characterized by a confidence set. For confidence sets there are two antagonistic features of interest, the size of the confidence sets and the coverage probability, in the sense that one can be achieved at the expense of the other. The aim is to construct minimal size confidence sets such that the coverage probability achieves a certain level.

The construction of adaptive confidence sets across a range of nested sub-models is even more involved. On the one hand we require that the size of the confidence sets are optimal on every single sub-model. On the other hand the coverage of the confidence sets is uniformly high over the collection of sub-models. It was pointed out by Low [3] that the preceding two competing features can not hold simultaneously in general. Introducing additional conditions, for instance shape restrictions [2] or the "self-similarity" assumption [6], can solve the problem and the construction of adaptive confidence sets with uniformly good coverage is possible.

In the Bayesian framework credible sets can be constructed to quantify the uncertainty in the posterior distribution. Due to the heavy Bayesian computational machinery (MCMC methods, ABC techniques, etc), in some cases the construction of credible sets can be easier than the construction of confidence sets from frequentist estimators. The frequentist coverage of credible sets describes to what extent credible sets can be viewed as frequentist confidence sets. From the celebrated Bernstein von Mises theorem follows that in parametric models under some regularity conditions the Bayesian credible sets are asymptotically equivalent to the frequentist confidence sets. However, in nonparametric problems the preceding equivalence does not hold in general. In our work we are interested in the frequentist properties of nonparametric credible sets based on adaptive Bayesian techniques.

[*] b.szabo@tue.nl

We consider the Gaussian sequence model which is often used as a platform to investigate more difficult nonparametric statistical problems. In the analysis we work with the L_2-loss function. Furthermore, we assume that the true sequence is contained in a Sobolev ball with regularity $\beta \in [D, 2D]$, for some given positive constant D. In the preceding model there exist adaptive confidence sets with uniformly good coverage properties [7]. We investigate the asymptotic behaviour of the marginal likelihood empirical Bayes method and show that there are certain oddly behaving truths for which the credible sets have asymptotically zero coverage. Good coverage properties of the credible sets rely on the correct bias-variance trade-off. However, in the case of the preceding adaptive Bayesian procedure for certain unregular functions the bias can dominate the variance which leads to a coverage probability zero. Keeping this in mind we construct a new empirical Bayes method based on risk estimation which provides adaptive confidence sets with good coverage properties.

The main message of the paper is that one has to choose the Bayesian procedure adequately to its purpose. For instance if one evaluates the performance of the posterior mean with the mean integrated squared error then it could happen that the likelihood based procedures attain sub-optimal behaviour. The reason behind it is that the mean integrated squared error is connected to the L_2-loss function, while the likelihood based methods are related to the Kullback-Leibler divergence. In the present paper we consider the problem of constructing credible sets with optimal size and good coverage. This boils down to finding a hyperparameter α which balances out the bias and the variance terms. However, the marginal likelihood empirical Bayes method selects that hyperparameter α which minimizes the Kullback-Leibler divergence between the marginal Bayesian likelihood function and the truth.

In Section 4 we make these loose statements more precise. First we introduce the Gaussian sequence model in more details. The negative result on the coverage of marginal likelihood empirical Bayes credible sets is given in Section 4. The new empirical Bayes method is referred to Section 4. The whole Section 2 is based on the paper [9]. We conclude the paper paper with a short simulation study in Section 4.

Main results

In the paper we work with the Gaussian white noise model and consider the sequence formulation

$$X_i = \theta_{0,i} + \frac{1}{\sqrt{n}} Z_i, \quad \text{for all } i = 1, 2, ...$$

where $X = (X_1, X_2, ...)$ is the observed infinite sequence, Z_i are independent standard normal distributed random variables and $\theta_0 = (\theta_{0,1}, \theta_{0,2}, ..)$ is the unknown infinite dimensional parameter of interest. Assume that θ_0 belongs to the Sobolev ball $S^\beta(M) = \{\theta : \sum \theta_i^2 i^{1+2\beta} \leq M\}$, where β is the regularity parameter and M is the squared radius of the Sobolev ball. The minimax rate of convergence over $S^\beta(M)$ is constant times $n^{-\beta/(1+2\beta)}$.

In the Bayesian framework to make inference about the unknown sequence θ_0 as a first step we endow it with a prior distribution. In our analysis we chose the infinite dimensional Gaussian distribution

$$\Pi_\alpha = \prod_{i=1}^{\infty} N(0, i^{-1-2\alpha}),$$

where the parameter $\alpha > 0$ denotes the regularity level of the prior distribution. The optimal choice of the hyperparameter $\alpha = \beta$ leads to posterior contraction rates $n^{-\beta/(1+2\beta)}$, while for other choices we get sub-optimal contraction rate [5],[1]. Since the smoothness parameter β of the true function θ_0 is usually not available one has to use a data driven method to choose α, as in [4]. Throughout the paper we assume the a priori knowledge that the parameter β lies in the interval $[D, 2D]$, with some given positive parameter D.

Marginal likelihood empirical Bayes method

In this section we state that the marginal likelihood empirical Bayes (MLEB) credible sets have asymptotic coverage zero for certain irregular sequences. The marginal log-likelihood function for α (relative to an infinite product of $N(0,1/n)$-distributions) is equal to

$$\ell_n(\alpha) = -\frac{1}{2}\sum_{i=1}^{\infty}\Big(\log\Big(1+\frac{n}{i^{1+2\alpha}}\Big) - \frac{n^2}{i^{1+2\alpha}+n}X_i^2\Big).$$

Let's denote by $\hat{\alpha}_n$ the maximizer of the likelihood function on the interval $[D,2D]$. The MLEB posterior is given by substituting $\hat{\alpha}_n$ for α in the posterior distribution:

$$\Pi_{\hat{\alpha}_n}(A|X) = \Pi_\alpha(A|X)\Big|_{\alpha=\hat{\alpha}_n}$$

for measurable subsets $A \subset \ell^2$.

The posterior distribution for fixed hyperparameter $\alpha > 0$ is conditionally Gaussian, hence a natural choice for the $1-\gamma$-credible set is the ℓ_2-ball centered at the posterior mean with radius $r_{n,\gamma}(\alpha)$:

$$\Pi_\alpha(\theta :\, \|\theta - \hat{\theta}_{n,\alpha}\| \leq r_{n,\gamma}(\alpha)|X) = 1-\gamma.$$

Substituting $\hat{\alpha}_n$ for the hyperparameter α and (possibly) blowing up the ball by a constant multiplier L we get the MLEB credible set

$$\hat{C}_n^E(L) = \{\theta :\, \|\theta - \hat{\theta}_{n,\hat{\alpha}_n}\| \leq L r_{n,\gamma}(\hat{\alpha}_n)\}. \tag{1}$$

We are interested in the frequentist coverage of the MLEB credible sets $P_{\theta_0}(\theta_0 \in \hat{C}_n^E(L))$. As we have already mentioned in the introduction, Robins and Van der Vaart [7] showed that there exist adaptive confidence sets with good uniform coverage properties for $\beta \in [D,2D]$. Unfortunately the MLEB credible sets do not attain this good coverage property. For any given parameter β there exists a sequence $\theta_0 \in S^\beta(M)$ for which the credible set has zero coverage asymptotically.

Theorem 0.1 (Theorem 3.1 in [8]). *For an arbitrary sequence of positive integers n_j satisfying $n_1 \geq 2$, $n_j \geq n_{j-1}^4$ and positive parameter K lets define the sequence $\theta_0 = (\theta_{0,1}, \theta_{0,2}, ...)$ as*

$$\theta_{0,i}^2 = \begin{cases} K n_j^{-1}, & \textit{if}\, n_j^{1/(1+2\beta)} \leq i < 2n_j^{1/(1+2\beta)}\, \textit{for any}\, j=1,2,..., \\ 0, & \textit{else}. \end{cases} \tag{2}$$

The constant K can be chosen such that for every $L>0$ the coverage of the credible set $\hat{C}_n^E(L)$ defined in (1) *tends to zero along the sub-sequence n_j.*

A detailed description of the intuition behind the bad coverage property of the credible sets for the preceding "inconvenient" truth can be found in [8]. We just mention in addition that even the prior knowledge on the smoothness ($\beta \in [D,2D]$) can not fix the poor performance of the method. A technical explanation relies on the incorrect bias-variance trade-off, caused by the different behaviour of the KL divergence and the L_2-loss function.

Risk-based empirical Bayes method

To correct the sub-optimal behaviour of the preceding adaptive Bayesian techniques we provide a new empirical Bayes method which gives adaptive credible sets with optimal coverage. The approach relies on minimizing the mean squared error of the posterior mean instead of maximizing the likelihood function. The estimator applied in the new empirical Bayes method is based on [7].

First we give an estimator for the squared bias $B_n^2(\alpha) = \sum i^{2+4\alpha}\theta_{0,i}^2/(i^{1+2\alpha}+n)^2$ with fixed hyperparameter α:

$$\hat{B}_{n,k_n}^2(\alpha) = \sum_{i=1}^{k_n}(X_i - \hat{\theta}_{n,i}(\alpha))^2 - \frac{i^{2+4\alpha}}{n(i^{1+2\alpha}+n)^2} = \sum_{i=1}^{k_n}\frac{i^{2+4\alpha}}{(i^{1+2\alpha}+n)^2}(X_i^2 - \frac{1}{n}),$$

where the sequence $k_n = n^{1/(1/2+2D)}$. Then the estimator $\hat{\alpha}_n$ is defined as

$$\hat{\alpha}_n = \inf\{\alpha \geq D : \hat{B}_{n,k_n}(\alpha) \geq n^{-2\alpha/(1+2\alpha)}\} \wedge (2D - C_0/\log n),$$

for some large enough constant C_0 specified later. Plugging in the estimator $\hat{\alpha}_n$ into the posterior distribution we get the risk-bases empirical Bayes (REB) posterior

$$\Pi_{\hat{\alpha}_n}(A|X) = \Pi_\alpha(A|X)\Big|_{\alpha=\hat{\alpha}_n}$$

for measurable subsets $A \subset \ell^2$. Similarly to the MLEB method we define the REB credible sets as

$$\hat{C}_n^R(L) := \{\theta : \|\theta - \hat{\theta}_{n,\hat{\alpha}_n}\| < Lr_{n,\gamma}(\hat{\alpha}_n)\}, \tag{3}$$

where L is a scaling parameter. The so constructed REB credible sets have in the minimax sense optimal size across the collection of Sobolev balls $S^\beta(M)$ with $\beta \in [D, 2D]$ and have uniformly good coverage properties over the largest Sobolev ball $S^D(M)$.

Theorem 0.2 (Theorem of 2.3 [9]). *For arbitrary positive parameters D, M and γ there exist constants C_0 and L such that the REB credible sets defined in (3) have at least $1-\gamma$ coverage uniformly*

$$\inf_{\theta_0 \in S^D(M)} P_{\theta_0}\big(\theta_0 \in \hat{C}_n^R(L)\big) \geq 1 - \gamma,$$

and the radius of the credible sets are rate optimal in a minimax sense for all $\beta \in [D, 2D]$

$$\inf_{\theta_0 \in S^\beta(M)} P_{\theta_0}\big(r_{n,\gamma}(\hat{\alpha}_n) \lesssim n^{-\beta/(1+2\beta)}\big) \to 1.$$

Numerical investigation

We give a short simulation study to illustrate that the new empirical Bayes method provides better confidence sets than the MLEB method. Consider the continuous representation of the Gaussian white noise model

$$X_t = \int_0^t f_0(s)ds + 1/\sqrt{n}W_t, \quad t \in [0,1],$$

where the function $f_0 = \sum \theta_{0,i}\sqrt{2}\sin(i\pi t)$ and W_t is the Brownian motion. The Fourier coefficients θ_0 are taken to be $\theta_{0,i} = \cos(i)i^{-1.4}$ for $i = 10, ..., 15$; $\theta_{0,i} = 4\cos(i)i^{-1.4}$ for $i = 150, ..., 200$; $\theta_{0,i} = i^{-1.4}$, for $i = 4^{4^j}, ..., 2 * 4^{4^j}$, $j = 2, ...$, and $\theta_{0,i} = 0$ else. Furthermore we choose a priori the regularity interval $[D, 2D]$ to be $[0.6, 1.2]$.

In the top line of Figure 1 we plot the MLEB credible sets given in (1), while in the bottom line stands the REB credible sets with sample sizes $n = 10^3, 5 * 10^4$ and $5 * 10^6$. We note that we did not blow up the credible sets by a factor $L > 1$ in any of the two cases. The true function is plotted by black, the posterior mean by blue and the gray area is the collection of the 95% closest draws to the posterior mean from the posterior distribution. One can see that the MLEB credible sets have poor coverage property for $n = 10^3$ and $n = 5 * 10^4$. The posterior mean is far away from the truth and in the meanwhile the credible sets

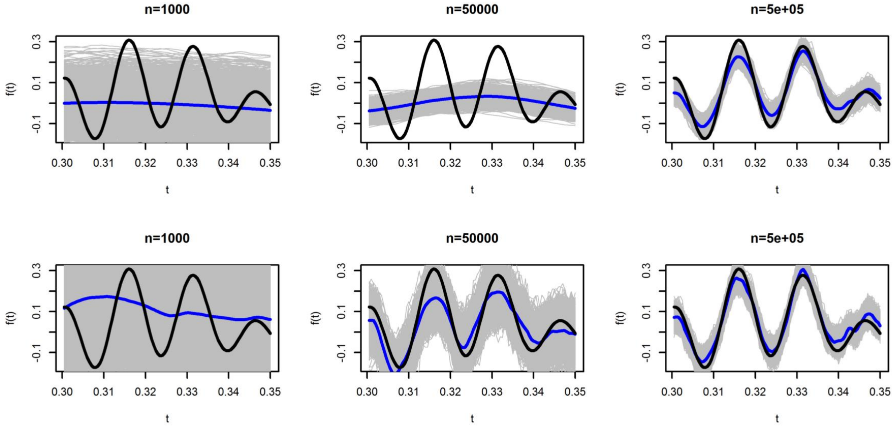


Figure 1: MLEB and the REB credible sets

are too narrow. Finally we note that the REB credible sets contain the truth in all cases, correcting the over-confidence of the MLEB method.

Acknowledgements: We would like to thank the referee for his/her useful comments in substantially improving the presentation of this article.

Bibliography

[1] Castillo, I. Lower bounds for posterior rates with Gaussian process priors. *Electron. J. Stat.* 2 (2008), 1281–1299.

[2] Hengartner, N. W., and Stark, P. B. Finite-sample confidence envelopes for shape-restricted densities. *Ann. Statist.* 23, 2 (1995), 525–550.

[3] Low, M. G. (1997). On nonparametric confidence intervals.*Ann. Statist.* 25(6), 2547–2554.

[4] Knapik, B. T., Szabo, B. T., van der Vaart, A. W., and van Zanten, J. H. Bayes procedures for adaptive inference in inverse problems for the white noise model. *available on http://arxiv.org/abs/1209.3628.*

[5] Knapik, B. T., van der Vaart, A. W., and van Zanten, J. H. Bayesian inverse problems with gaussian priors. *Ann. Statist.* 39, 5 (2011), 2626–2657.

[6] Picard, D., and Tribouley, K. Adaptive confidence interval for pointwise curve estimation. *Ann. Statist.* 28, 1 (2000), 298–335.

[7] Robins, James and van der Vaart, Aad W. (2006). Adaptive nonparametric confidence sets. *Ann. Statist.* 34, 229–253.

[8] Szabo, B. T., Vaart, A. W., and Zanten, J. H. (2013). Frequentist coverage of adaptive nonparametric Bayesian credible sets. *available on http://arxiv.org/abs/1310.4489.*

[9] Szabo, B. T., Vaart, A. W., and Zanten, J. H.(2013). Honest Bayesian confidence sets for the L2-norm. preprint. *available on http://arxiv.org/abs/1311.7474.*

Model selection approach for genome wide association studies in admixed populations

Piotr Szulc[*]

Department of Mathematics and Computer Science
Wroclaw University of Technology, Poland

Abstract

The main purpose of genome wide association studies (GWAS) is the identification of genes responsible for quantitative traits (Quantitative Trait Loci, QTL) or disease causing genes in human populations. Localization of genes in such outbred populations is relatively difficult. We present this problem and two criteria, mBIC and mBIC2, which were successfully used in GWAS.
However, it turns out that we can find much more influential genes if we perform GWAS in *admixed population*, obtained as a result of interbreeding between previously separated ancestral populations. In that case, apart from genotypes, we have information about the origin of genome's fragments. We introduce modification of mBIC and mBIC2, which can use this additional information. Finally, we present results of simulations which confirm that we are able to identify more influential genes in those populations.

Keywords: linear regression, model selection criteria, GWAS, admixed population
AMS subject classifications: 62J05, 92D20

1 Introduction

In the recent time we can observe a very popular problem in regression, so called *sparsity*, which denotes a situation when we have a very large number of variables, often bigger than the sample size, but the number of significant predictors is small. This situation takes place e.g. in genetics, in the problem of localizing genes responsible for quantitative traits (Quantitative Trait Loci, QTL). We have to deal with a large number of markers (fragments of DNA which can occur in different variants for different individuals and their genotype can be identified). Usually, we observe two versions (alleles) of markers, let's say A and B, but when we consider diploid organisms in which chromosomes appear in pairs, a genotype of j-th marker for i-th individual can be coded in the following way:

$$x_{ij} = \begin{cases} -1 & \text{if } AA \\ 0 & \text{if } AB \\ 1 & \text{if } BB \end{cases} \qquad (1)$$

When we have values of a trait of interest, we can use the linear regression and find markers which are related to the trait. If these markers are correlated with genes influencing the trait (what usually means that they occur near these genes), we can localize them.
In practice, the hardest is to localize QTL or disease causing genes in human populations, so called Genome Wide Association Studies (GWAS), because correlations between genotypes of QTL and neighbouring markers can be very small. Therefore, we need a large number of densely spaced markers. Relatively

[*] e-mail: piotr.a.szulc@pwr.wroc.pl

easier is to localize genes in experimental populations (plants and animals), in which we interbreed highly related individuals. Thanks to that, correlations between markers are notably higher. Much more about localizing genes in experimental populations can be found in [9] or [8] and information about GWAS is provided in [3].
The problem of a low correlation in genome wide association studies can be substantially decreased if one performs GWAS in admixed populations. These populations originate from a recent interbreeding between two previously isolated populations, let's say P_1 and P_2. In this case we have additional information on the ancestry states of a given region of a genome, which for diploid organisms we can present in the following way:

$$z_{ij} = \begin{cases} -1 & \text{if } P_1P_1 \\ 0 & \text{if } P_1P_2 \\ 1 & \text{if } P_2P_2. \end{cases} \tag{2}$$

We can assume that each marker comes either from population P_1 or P_2. Correlation between ancestry states is much higher than between genotypes of markers, unfortunately correlation between a genotype and an ancestry state is much lower. However, we will show that if we combine both pieces of information, we can find more significant genes.

2 Modified versions of Bayesian Information Criterion

We aim to choose the best linear regression model in the situation when the number of explanatory variables p is very large, bigger than the samples size n. Statistics knows a lot of criteria which can be used to find a suitable model, unfortunately classic versions like Akaike Information Criterion (AIC, [1]) or Bayesian Information Criterion (BIC, [13]) are inappropriate in our situation. This is due to the fact that they are derived based on the assumption that n goes to infinity, while p remains constant. This is obviously not a good assumption when p is larger than n and it was shown that classical criteria overestimate the number of significant regressors k[11].
To construct a suitable criterion, we need some *a priori* knowledge about the number of nonzero coefficients k because when p is larger then n, the least squares estimators for regression coefficients are not unique and regression models are not identifiable. In applications in genetics we usually assume the sparsity, i.e. that k/n is very small. This assumption was used to construct modifications of BIC: mBIC [5], mBIC2 ([11], [12]) and EBIC [7]. Further we will deal with two of them, mBIC and mBIC2, and show how to use them in the problem of admixed populations.

2.1 mBIC and mBIC2

We are interested in the following linear regression model:

$$y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i,\ i = 1, ..., n, \tag{3}$$

where y_i is a trait value for i-th individual, p is the number of available markers, x_{ij} is a genotype of j-th marker for i-th individual and ϵ_i's are independent variables with the normal distribution $\mathcal{N}(0, \sigma^2)$. Let s denote a subset of $\{1, ..., p_n\}$ (a model) of a size $v(s)$. If we want to find nonzero coefficients, we can use one of the most popular criterion, Bayesian Information Criterion, which suggests choosing the model minimalizing the formula

$$BIC(s) = n \ln(RSS(s)) + v(s) \ln n, \tag{4}$$

where $RSS(s)$ means residual sum of squares for the model s.
BIC was derived in a Bayesian context and it is used to approximate the logarithm of the posterior probability

of the given model. This probability is proportional to the product of the integrated likelihood of the data given the model and the prior probability of the model. BIC neglects this second factor and as a result it assigns the same prior probability to all models. It has undesirable consequences in the context of localizing genes because if we consider nonzero coefficients, the prior distribution on their number k is binomial $\mathcal{B}(p, 1/2)$. It leads to the overestimation of the number of predictors because this distribution is concentrated almost entirely on $[p/2 - 2\sqrt{2}, p/2 + 2\sqrt{2}]$.
Two modification of BIC dealing with that problem were proposed:

$$mBIC(s) = n\ln(RSS(s)) + v(s)\ln n + 2v(s)\ln\left(\frac{p}{c} - 1\right), \tag{5}$$

$$mBIC2(s) = n\ln(RSS(s)) + v(s)\ln n + 2v(s)\ln\frac{p}{c} - 2\ln(v(s)!), \tag{6}$$

where $c = Ek$, i.e. the expected value of k. The first modification is a result of the assumption that k has the binomial distribution but with mean equals c/p. The additional penalty in mBIC is closely related to the Bonferroni correction for multiple testing ([6], [11]), which is substantially worse than the Benjamini-Hochberg procedure [4]. The second modification is a product of exploiting good properties of B-H procedure.
There are quite a few papers which show good properties of those criteria, both theoretical and practical ([2], [6], [11], [14], [15]). If c is unknown, one can use $c = 4$ to control Family Wise Error Rate at the level of 10% (mBIC) and False Discovery Rate at the same level (mBIC2).

2.2 Admixed populations

Now we want to apply mBIC and mBIC2 in admixed populations. In this case our target is to fit the regression model

$$y_i = \sum_{j=1}^{p_1}\beta_j x_{ij} + \sum_{k=1}^{p_2}\gamma_k z_{ik} + \epsilon_i,\ i = 1, ..., n, \tag{7}$$

so we treat ancestry states as additional variables. Adequate criteria change to the following forms:

$$\begin{aligned} mBIC(s) = n\ln(RSS(s)) + (v_1(s) + v_2(s))\ln n + 2v_1(s)\ln\left(\frac{p_1}{c_1} - 1\right) + \\ + 2v_2(s)\ln\left(\frac{p_2}{c_2} - 1\right), \end{aligned} \tag{8}$$

$$\begin{aligned} mBIC2(s) = n\ln(RSS(s)) + (v_1(s) + v_2(s))\ln n + 2v_1(s)\ln\frac{p_1}{c_1} - 2\ln(v_1(s)!) + \\ + 2v_2(s)\ln\frac{p_2}{c_2} - 2\ln(v_2(s)!), \end{aligned} \tag{9}$$

where $v_1(s)$ means the number of markers in the model, $v_2(s)$ is the number of ancestry states in the model and p_1 denotes the number of all markers. The problem is with p_2 because due to high correlation it is not a good idea to treat p_2 as the total number of ancestry states. Instead, we can calculate so called the effective number of tests p_{eff} for ancestry variables and p_2 will be that value. The effective number of tests is the number of independent single tests which we can perform to control FWER at the level α. According to [10], we can write

$$\begin{aligned} \alpha = P_{H_0}\left(max_{j\in\{1,\ldots,p\}} LRT_j > c\right) \approx 1 - \exp(-2[1 - \Phi(\sqrt{c})]) - \\ - 0.02ptL\sqrt{c}\phi(\sqrt{c})\nu\left(\sqrt{0.02tLc}\right), \end{aligned} \tag{10}$$

where LRT is the likelihood ratio test, Φ is the normal distribution, ϕ is the normal density and $\nu(x) \approx \exp(-0.583x)$. Possibility of using this formula results from appropriate correlation between ancestry states, equal to $\exp(-tL)$, where L is the distance between markers (in Morgans) and t denotes an admixing time for a given individual. If we do not know admixing times and distances between markers are different, we suggest replacing tL with $-\ln r$, where $\ln r$ is the average of the logarithms of the correlations between neighbouring ancestry variables.
On the other hand, if we perform p_{eff} independent tests, we can write

$$\alpha = P_{H_0}\left(\max_{i\in\{1,\ldots,p_{eff}\}} LRT_j > c\right) \approx 1 - \left[1 - 2\left(1 - \Phi(\sqrt{(c)})\right)\right]^{p^{eff}}. \tag{11}$$

If we compare 10 and 11, we get

$$p_2 = p_{eff} = \frac{\ln(1-\alpha)}{\ln\left(2\Phi(\sqrt{c}) - 1\right)}. \tag{12}$$

3 Simulations

We performed simulations on the data which were very close to real data. We had 482906 markers (22 chromosomes) and 482906 ancestry states, all for 1000 unrelated individuals. We chose 20 causal markers which were differed in the linkage disequilibrium (LD; this characteristic says how much a marker is correlated with neighbours) and the ancestry frequency (AF; it says how much an ancestry state is correlated with the corresponding marker). We simulated values of a trait according to the model

$$y_i = \sum_{j=1}^{20} 0.5x_{ij} + \epsilon_i,\ i = 1, ..., 1000, \tag{13}$$

where $\epsilon_i \sim \mathcal{N}(0,1)$. Then we removed these causal markers from the design matrix and looked for the best linear model, using the strategy described in [12]. We considered the identified marker a true discovery if the correlation between this marker and the corresponding causal marker was greater than 0.5. Results are shown in Table 1. In the first design we used only genotypes of markers and in the second design we added ancestry states. The power is calculated as the number of true discoveries divided by 20 and averaged over 100 replicates.

	First design		Second design	
	mBIC	mBIC2	mBIC	mBIC2
Power	0.427	0.534	0.670	0.723
FDR	0.023	0.109	0.040	0.082

Table 1: Power and FDR

Our simulations show that the power of GWAS in admixed population can be increased if we add information about ancestry state to the regression model (and FDR stays at the low level). We were able to identify genes which have low LD (so it was impossible to find them using only genotypes) but high AF. What is more, because of these new genes, our model was better so other genes were found more often. FDR for mBIC2 is higher than for mBIC, what was expected (the penalty in mBIC2 is smaller).

Acknowledgements: This research was supported by NCN grant 2012/05/N/ST1/03136. I am working through that problem in cooperation with M. Bogdan, H. Tang, F. Frommlet and I want to thank them.

Bibliography

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

[2] Baierl, A., Bogdan, M., Frommlet, F. and Futschik, F. (2006). On locating multiple interacting quantitative trait loci in intercross designs. *Genetics* 173, 1693–1703.

[3] Balding, D.J. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics* 7, 781–791.

[4] Bogdan, M., Chakrabarti, A., Ghosh, J. and Frommlet, F. (2011). Asymptotic Bayesoptimality under sparsity of some multiple testing procedures. *Annals of Statistics* 39, 1551–1579.

[5] Bogdan, M., Ghosh, J. and Doerge, R.W. (2004). Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci. *Genetics* 167, 989–999.

[6] Bogdan, M., Ghosh, J. and Zak-Szatkowska, M. (2008). Selecting explanatory variables with the modified version of Bayesian Information Criterion, *Qualita and Reliability Engineering International* 24, 627–641.

[7] Chen, J. and Chen, Z. (2008). Extended Bayesian information criterion for model selection with large model space, *Biometrika* 94, 759–771.

[8] Doerge, R.W. (2002). Mapping and analysis of quantitative trait loci in experimental populations. *Nature Reviews Genetics* 3, 43–52.

[9] Doerge, R.W., Zeng, Z-B. and Weir, B.S. (1997). Statistical issues in the search for genes affecting quantitative traits in experimental populations. *Statistical Science* 12, 195–219.

[10] Dupuis, J. and Siegmund, D.O. (1999). Statistical methods for mapping quantitative trait loci from a dense set of markers. *Genetics* 151, 373–386.

[11] Frommlet, F., Bogdan, M., Murawska, M. and Chakrabarti, A. (2011), Asymptotic Bayes optimality under sparsity for general priors under the alternative (technical report available at arxiv.org/abs/1005.4753v2).

[12] Frommlet, Ruhaltinger, F., Twarog, P. and Bogdan, M. (2012). A model selection approach to genome wide association studies. *Computational Statistics and Data Analysis* 56, 1038–1051.

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

[14] Szulc, P. and Bogdan, M. (2012). Localizing influential genes with modified versions of Bayesian Information Criterion. *Mathematica Applicanda* 40, 3–14.

[15] Szulc, P. (2012). Weak conistency of modified versions of Bayesian Information Criterion in a sparse linear regression. *Probability and Mathematical Statistics* 32, 47–55.

Some remarks on normal conditionals and normal projections

Barry C. Arnold[1] **and B.G. Manjunath**[*2]

[1] *University of California Riverside, USA*
[2] *CEAUL and DEIO, FCUL, University of Lisbon, Portugal*

Abstract

It is always possible to construct a d-dimensional non-normal distribution having any finite number of normal projections and all $(d-1)$ dimensional marginals normal. Also, there can exist d-dimensional non-normal distribution with all conditional distributions being normal. In the present note we introduce two new characterizations of the classical d-dimensional normal distribution. (1) Having normal conditionals and a finite number of normal projections uniquely characterizes the classical d-dimensional normal distribution. (2) Having normal conditionals and each of $(d-1)$ coordinate random variables having a one dimensional normal distribution is sufficient to ensure that the d-dimensional distribution has to be classical normal.

Keywords: linear transformation, normal conditionals, normal marginals, non-normal distributions
AMS subject classifications: 62E10 and 62E15

1 Introduction

Classical distribution theory in higher dimensions is largely focused on the multivariate normal distribution. For the multivariate normal density it is well known that every marginal distribution, every conditional distribution and all linear transformations are also normal. Besides, it is also obvious that these properties chosen individually are not sufficient conditions to characterize the multivariate normal density. There are many multivariate non-normal distributions which share some of these features with the classical normal distribution. So, it is interesting to find combinations of these characteristics which will be sufficient to characterize the classical normal density. In the present article we review some available results in this direction and contribute two new characterization results involving normal conditional distributions. It is to be expected that similar results can be obtained when dealing with distributions with conditionals in arbitrary exponential families. As an illustration, two results will be presented dealing with distributions with exponential conditionals.
In the following section we present two examples where a non-normal bivariate density shares common features with the classical bivariate normal distribution. Stoyanov [5] is a useful source of other examples of multivariate non-normal distributions having classical normal properties.

1.1 Counterexample: Non-normal bivariate distribution with marginals, sum and difference which are normal

Let (X_1, X_2) be a bivariate random variable with density

$$f_\epsilon(x_1,x_2)=\frac{1}{2\pi}e^{-\frac{1}{2}(x_1^2+x_2^2)}\left\{1+\epsilon(x_1^3x_2-x_2^3x_1)e^{-\frac{1}{2}(x_1^2+x_2^2)}\right\} \tag{1}$$

[*] Corresponding author, e-mail: bgmanjunath@gmail.com

where

$$|\epsilon| \leq \frac{e^2}{4}.$$

Now, consider the following linear combination of the coordinates of (X_1, X_2), $X = X_1 \sin\theta + X_2 \cos\theta$. The density of X is

$$g_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \left\{ 1 - \frac{\sqrt{2}\ \epsilon\ sin(4\theta)}{32} (4x^4 - 12x^2 + 3) e^{-\frac{1}{2}x^2} \right\}. \tag{2}$$

Clearly, X is non-normal. It may be noted here that X_1, X_2, $X_1 + X_2$ and $X_1 - X_2$ are normal, but this does not imply that (X_1, X_2) is bivariate normal. Hence, marginal normality and finite number of normal projections are not sufficient conditions to characterize the multivariate normal distribution.
We refer to Hamedani and Tata [3] for a closely related bivariate normal characterization, that is, bivariate normality is completely determined by a countable dense number of one-dimensional normal projections. Another important reference is Manjunath and Parthasarathy [4] dealing with a generalization of the characterization in [3].

1.2 Counterexample: Non-normal bivariate density for which one set of conditionals and one marginal are normal

Let (X_1, X_2) be a bivariate random variable with density

$$f(x_1, x_2) \propto (1 + x_2^2)^{\frac{1}{2}} \exp\left\{ -\frac{1}{2} \left[x_1^2 x_2^2 + x_1^2 + x_2^2 \right] \right\}. \tag{3}$$

In this example, $X_2 \sim N(0, 1)$ and $X_1 | X_2 = x_2 \sim N\left(0, \frac{1}{1+x_2^2}\right)$ for all x_2. But note that the conditional distribution of X_2 given X_1 is not normal. This confirms the fact that marginal normality and one family of conditionals being normal are not sufficient to guarantee joint normality.
Bhattacharyya [2] observed that, even if one assumes that both families of conditional densities (of X_1 given X_2, and of X_2 given X_1), bivariate normality is not guaranteed. See Bhattacharyya [2] and Arnold et al. [1] for discussion of several sufficient conditions to characterize the classical bivariate and multivariate normal distribution within the class of distributions with normal conditionals.

2 Normal Conditionals

2.1 Bivariate normal conditionals

As in Arnold et al. [1], assume that a joint density $f(x, y)$ has all conditionals in the univariate normal family. Then writing the joint density as a product of a marginal and a conditional density in both possible ways, we have

$$\frac{f_1(x)}{\sigma_2(x)} \exp\left[-\frac{1}{2} \left(\frac{y - \mu_2(x)}{\sigma_2(x)} \right)^2 \right] = \frac{f_2(y)}{\sigma_1(y)} \exp\left[-\frac{1}{2} \left(\frac{x - \mu_1(y)}{\sigma_1(y)} \right)^2 \right], \tag{4}$$

where $f_1(x) > 0$ and $f_2(y) > 0$ are marginal densities and $\mu_2(x)$, $\sigma_2(x)$, $\sigma_1(y)$ and $\mu_1(y)$ are functions of marginal variables. By solving above equation, the joint density of $f(x, y)$ can be expressed as:

$$f(x, y) = \exp\left\{ (1, x, x^2) \begin{pmatrix} m_{00} & m_{01} & m_{02} \\ m_{10} & m_{11} & m_{12} \\ m_{20} & m_{21} & m_{22} \end{pmatrix} \begin{pmatrix} 1 \\ y \\ y^2 \end{pmatrix} \right\} \tag{5}$$

where the constants $\{m_{ij} : i, j = 0, 1, 2\}$ are chosen to ensure nonnegativity of $f(x, y)$ and its marginals and the integrability of those marginals. To guarantee integrability the coefficients must satisfy one of the two following sets of conditions: (1) $m_{22} = m_{12} = m_{21} = 0$, $m_{20} < 0, m_{02} < 0$ and $m_{11}^2 < 4m_{02}m_{20}$. (2) $m_{22} < 0$, $4m_{22}m_{02} > m_{12}^2$ and $4m_{20}m_{22} > m_{21}^2$. Condition (1) yields the classical bivariate normal density.

2.2 Multivariate extension

Let $\boldsymbol{X}$ be a d-dimensional random variable which has normal conditionals. Its joint density can be written in the form

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \exp\left\{\sum_{\boldsymbol{i}\in T_d} m_{i_1\ldots i_d} x_1^{i_1}\ldots x_d^{i_d}\right\}, \tag{6}$$

where T_d is the set of all vectors of 0's, 1's and 2's of dimension d and $\boldsymbol{i} = (i_1, .., i_d)^T$ each $i_j \in \{0, 1, 2\}$, $j = 1, 2, ..., d$. There are certain constraints on the $m_{i_1\ldots i_d}$'s in order to guarantee integrability, just as there were in dimension 2.

3 New characterizations of the multivariate normal

In this section we present the two new characterizations of d-dimensional classical normal density within the class of d-dimensional distributions with normal conditionals. We begin with a useful Lemma.

Lemma 3.1. *Let $(X_1, X_2, ...X_m, X_{m+1}, ..., X_d)$ have a normal conditionals distribution (of the form (6)). The following are equivalent:*

(1) $(X_1, X_2, ..., X_m)$ has a normal conditionals distribution.

(2) $(X_1, X_2, ...X_m)$ and $(X_{m+1}, ..., X_d)$ are independent and $(X_{m+1}, ..., X_d)$ also has a normal conditionals distribution.

Theorem 3.1. *Let $\boldsymbol{X} = (X_1, ..., X_d)^T$ have a normal conditionals density of the form (6). If each of the coordinate random variables $X_1, X_2, .., X_{d-1}$ has a one dimensional normal distribution then $\boldsymbol{X}$ has a d-dimensional classical normal distribution.*

An alternative necessary and sufficient condition for a d-dimensional density with normal conditionals to have a classical normal distribution is to have all d of its $(d - 1)$ dimensional marginals of the classical normal form.

Theorem 3.2. *Let $\boldsymbol{X} = (X_1, ..., X_d)^T$ be a d-dimensional random vector which satisfies the following conditions: (1) it has a normal conditionals density of the form (6); (2) for some vector $\boldsymbol{a} = (a_1, a_2, ..., a_d)$ with at least $(d - 1)$ of its coordinates being nonzero, $\sum_{i=1}^d a_iX_i$ has a normal distribution. Then, $\boldsymbol{X}$ has a classical d-dimensional normal density.*

4 Exponential conditionals

In this section we present a new characterization of a d-dimensional exponential density.
Let $\boldsymbol{X}$ be a d-dimensional random variable having exponential conditionals, then the joint density can be written in the form

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \exp\left\{ - \sum_{\boldsymbol{i}\in T_d} \lambda_{i_1 \ldots i_d} \left(x_1^{i_1} \ldots x_d^{i_d} \right) \right\}, \tag{7}$$

where T_d is the set of all vectors of 0's and 1's of dimension d and $\boldsymbol{i} = (i_1, .., i_d)^T$ each $i_j \in \{0, 1\}$, $j = 1, 2, ..., d$ and the parameters $\lambda_{i_1 \ldots i_d}$ are nonnegative.

Lemma 4.1. *If $(X_1, ..., X_d)$ has an exponential conditionals distribution of the form (7) and X_1 is exponential then X_1 and $(X_2, ..., X_d)$ are independent and $(X_2, ..., X_d)$ has a $(d - 1)$ dimensional exponential conditionals distribution*

Theorem 4.1. *Let $\boldsymbol{X} = (X_1, ..., X_d)$ have an exponential conditionals distribution of the form (7). If each of the coordinate random variables $X_1, X_2, .., X_{d-1}$ has a one dimensional exponential distribution then $\boldsymbol{X}$ has a d-dimensional distribution with independent exponential marginals.*

Acknowledgements: Research enabled through a grant from the Foundation for Science and Technology (FCT), project SFRH/BPD/72184/2010 and FCT /PTDC/MAT/101736/2008 "EXTREMA: EXTREMES IN TODAY'S WORLD". We would like to thank Bernoulli Society for the financial aid to attain the 18th EYSM.

Bibliography

[1] Barry C. Arnold, Enrique Castillo and Jose M. Sarabia, (1999). *Conditional Specification of Statistical Models*, Springer, New York.

[2] Bhattacharyya, A. (1943). On some sets of sufficient conditions leading to the normal bivariate distribution. *Sankhyá*, Vol. 6 Part 4, 399–406.

[3] Hamedani, G.G. and Tata, M.N. (1975). On the determination of the bivariate normal distribution from distributions of linear combinations of the variables. *The American Mathematical Monthly* 82, 913–915.

[4] Manjunath, B.G. and Parthasarathy, K.R. (2012). A note on Gaussian distributions in R^n. *Proc. Indian Acad. Sci. (Math. Sci.)* 122 (4), 635–644.

[5] Stoyanov, J. (1997). *Counterexamples in Probability (2nd ed.)*, Wiley, New York.

Competing risks analysis in Nephrology research: An example in peritoneal dialysis

Laetitia Teixeira[*1] **and Denisa Mendonca**[2]

[1] *Doctoral Program in Applied Mathematics - Faculty of Sciences and Institute of Biomedical Sciences Abel Salazar, University of Porto*
[2] *Institute of Biomedical Sciences Abel Salazar, University of Porto*

Abstract

In clinical and epidemiological research, increasing importance has been given to the competing risk approach and this methodology has been referred as the rule rather than the exception in follow-up studies [1]. It is an extension of classical survival analysis.

In the presence of competing risks, two types of analysis can be performed: modelling the cause-specific hazard and modelling the hazard of the subdistribution [5, 8]. The context of the research question is the main determinant for the choice of an appropriate statistical model. When the hazard of the subdistribution is analysed, the goal is to compare the probability of the event of interest and therefore can be translated into actual numbers of patients with this event. Comparing the cause-specific hazards gives an insight into the biological process [7, 8, 9].

In peritoneal dialysis programs, several endpoints can be observed: death, transfer to haemodialysis and renal transplantation. In our study, we were interested in modelling the time from the entrance in the peritoneal dialysis program until the occurrence of the event of interest, death, in the presence of competing risks (transfer to haemodialysis and renal transplantation). Regression models based on cause-specific hazard and hazard of the subdistribution were performed, considering time-independent covariates (gender, automatic peritoneal dialysis, first renal replacement therapy, reason for peritoneal dialysis), time-varying covariates (age and diabetes) and time-dependent covariates (peritonitis) and the estimates obtained by such models were examined and discussed.

Keywords: Competing risks, cause-specific hazard, hazard of the subdistribution, peritoneal dialysis.
AMS subject classifications: 62P10

1 Introduction

In survival analysis, the problem of competing risks occurs when a patient may experience only one event of a set of K possible events. Several definitions of competing risks can be considered. Gooley et al. [4] defined this concept as the situation where one type of event either precludes the occurrence of another event or fundamentally alters the probability of occurrence of this other event. When we are interested in evaluating a peritoneal dialysis program, we are in the presence of competing risk problem because patients can be experienced several events, such as death, transfer to haemodialysis or renal transplantation.

In the presence of competing risks, two main approaches associated with a two different hazard can be considered: testing the ‘pure’ effect by ignoring the competing risks (cause-specific hazard) and including the competing risks (cumulative incidence function or hazard of the subdistribution). Cause-specific hazard and cumulative incidence function are the most important quantities in competing risks problem. While the first quantity provides information about instantaneous failure rate from a particular event cumulative incidence

[*] Corresponding author, e-mail: laetitiateixeir@gmail.com

curve estimates the chance of ultimately experiencing that event [2]. These two approaches give different information about the effect of a covariate and depending on the clinical or medical research question, we may want to compare the cause-specific hazard or the cumulative incidence functions [9].
In this work, the objective is to perform and to discuss regression models based on cause-specific hazard and hazard of the subdistribution, considering time-independent (gender, Automatic Peritoneal Dialysis (APD), first renal replacement therapy, reason for peritoneal dialysis (PD), time-varying covariates (age and diabetes) and internal time-dependent covariates (peritonitis). The regression models were applied in all consecutive incident end-stage renal disease patients starting peritoneal dialysis between October 1985 and February 2013 in Peritoneal Dialysis Unit, Nephrology Department, CHP – Santo António Hospital, Porto, Portugal (n=444). Patient outcome was defined as the earliest event among: death, transfer to haemodialysis or renal transplantation. In the present study, the interest is the analysis of patient survival and, in this case, the event of interest was death and the competing risks were transfer to haemodialysis and renal transplantation. Patients without any of these outcomes were censored at the date of their last recorded visit or at the end of the study period (February 2013).

2 Regression modelling in competing risks setting

Two types of regression models were considered in this work, based on two types of hazard: cause-specific hazard (CSH) and cumulative incidence function (or hazard of the subdistribution (SH)). The CSH function is the principal estimable quantity in competing risks setting and is defined by the instantaneous risk of dying from a particular cause k given that the subjects is still at risk (i.e. the subjects is still alive) at time t:

$$h_k(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t, C = k \mid T \geq t)}{\Delta t}$$

The unit of h_k is number of events per person-time unit and there are as many CSH functions as there are types of events [10]. As the CSH definition conditions on $T \geq t$, the presence of other events affects the 'risk set' [10], that is, all patients with any event before time t are removed from the risk set at that time point [7].
Cumulative incidence function describes the actual risk of experiencing the event of interest k until time t [7] it is defined as the cumulative probability of event k having occurred in the presence of other competing events, i.e., is the probability that an event of type i occurs at or before time t:

$$F_k(t) = P(\text{failure time } T \leq t, cause = k) = \int_0^t S(u) h_k(u) du$$

Cumulative incidence function of event k is defined as a function of both the probability of not having failed from some other event first, $S(u)$, up t time t and the CSH for the event of interest at that time. Therefore, cumulative incidence estimator for an event of type k depends not only on the number of individuals who have experienced this type of event but also on the number of individuals who have not experienced any other type of event [9]. A graphical display of the estimated cumulative incidence functions for all competing events is a key summary of the competing risks process (similar to the Kaplan-Meier curve in survival analysis) [7]. The hazard associated with the cumulative incidence function is the SH and can be interpreted as the instantaneous risk of dying from a particular cause k given that no other events has occurred thus far (i.e. the subject is still alive at time t) k [9] and it is defined as:

$$\gamma_k(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t, C = k \mid \{T \geq t \text{ or } (T \leq t \text{ and } C \neq k)\})}{\Delta t}$$

The condition in the curled brackets expresses the fact that the event of interest did not happen until t, but it is possible that the observation for a subject has stopped because a competing event was observed [8]. To analyse CSH, usual techniques for the time to event analysis can be employed, such as Cox proportional hazards model:

$$h_k(t,x) = h_{k0}e^{\sum_{i=1}^{K}\beta_i x_i}$$

If there are r events at the time points $t_1 < t_2 < \ldots < t_{r-1} < t_r$ and R_j is the risk set at time t_j then the partial likelihood to be maximized is:

$$L(\beta) = \prod_{j=1}^{r}\left(\frac{exp(\beta x_j)}{\sum_{i\in R_j} exp(\beta x_i)}\right)$$

and $R_j = \{i : t_i \geq t_j\}$ represent the risk set. The quantity $exp(\beta)$ is called the CSH ratio (CSHR) and represents the increase of the CSH due to one unit increase of the covariate x. The cause k-specific hazard gives the rate of event k per time unit for individual who are still alive [1]. Regression modelling for estimating the association between the cumulative incidence function and covariates is complicated because these models have complex non-linear functional forms for the effect of the covariates on the cumulative incidence function [9]. The model proposed by Fine and Gray [3] is a direct approach assuming a proportional hazard form for the 'hazard rate' of the subdistribution function [10] and it is based on:

$$\gamma_k(t,x) = \gamma_{k0}(t)e^{\beta x}$$

The partial likelihood was constructed as:

$$\tilde{L}(\beta) = \prod_{j=1}^{r}\left(\frac{exp(\beta x_j)}{\sum_{i\in \tilde{R}_j} \omega_{ji} exp(\beta x_i)}\right)$$

The differences in the partial likelihood defined for the hazard of the subdistribution comparing with the likelihood defined for the CSH are the inclusion of weights at the denominator and the risk set is defined differently. The observation for which the competing risk event is observed is in the risk set at all times ($\tilde{R} = \{i : t_i \geq t_j$ or $t_i \leq t_j$ and the subject had a competing risk event$\}$). The interpretation for $exp(\beta)$ in this framework is similar: it represents the increase of the SH due to one unit increase of x [8].

In competing risks setting, the covariates may affect the CSH and SH differently [1]. The results obtained in the SH model are influenced by the way the competing risks were distributed. If patients with a characteristic were more likely to have a competing risk, the event of interest could not be observed and therefore the effect of this covariate would be diminished [9]. In this work, we consider two types of covariates: time-independent covariates and time-dependent covariates. A time-independent covariate is a variable whose value for a patient remains constant over time (e.g. gender, APD, first renal replacement therapy, reason for peritoneal dialysis). A time-dependent covariate is defined as any covariate whose value for a given subject may differ over time. Time-dependent covariates can be divided into several types and we consider only two different types of time-dependent covariates: defined and internal time-dependent covariates. Most defined time-dependent covariates, also called time-varying covariates, are of the form of the product of a time-independent variable multiplied by some function of time (e.g. diabetes and age). Internal time-dependent covariates are covariates whose values may change over time and the reason for a change depends on "internal" characteristics or behaviour specific to the individual (e.g. peritonitis is a covariate which start with value=0 and may increase to 1, if the patient experiences a peritonitis episode) [6].

3 Application

Survival analysis methods taking competing risks into account were performed for analysing patient survival. First, estimates of cumulative incidence function were calculated. Regression models taking competing risks

into account (Cox CSH model and Fine and Gray model based on SH model) were carried out to analyse the effect of covariates in the patient survival. Variable associated with the event of interest (death) at the 10% significance level on the basis of univariate models were introduced in the multivariable models. Diabetes and age were tested as time-varying covariates and peritonitis as internal time-dependent covariate. In the case of time-varying covariate, the covariate defined by the interaction between covariate and time (function identity) was only included if statistically significant.
All analyses were performed with R software using the packages *coxph* and *cmprsk* and significance level for multivariable models was set at 0.05.

3.1 Results

The sample comprises 444 patients, 59.7% women (n=265) and the mean age was 48.1 years. Transfer to haemodialysis was the main reason for PD discontinuation, followed by renal transplantation (n=119, 26.8%) and death (n=101, 22.7%). At the end of the study period, 15.5% of the patients were still on PD. 59.2% were PD first (i.e. the first renal replacement therapy was PD), 23.4% had diabetes and 60.5% had started PD by option. Finally, 51.1% of patients have experienced at least one peritonitis episode.
Analyzing the cumulative incidence estimates for the event of interest, the probabilities of death by 1, 3 and 5 years after starting PD were 0.065, 0.17 and 0.23, respectively.
In the SH model, significant risk factors for death are: (1) age as time-independent covariate (SHR=1.05, 95% CI 1.03-1.06); (2) first treatment (haemodialysis, considering peritoneal dialysis as reference category - SHR=1.66, 95% CI 1.11-2.49); (3) reason for peritoneal dialysis (SHR=0.55, 95% CI 0.37-0.81); (4) diabetes as time-independent covariate (SHR=2.04, 95% CI 1.36-3.06).
According to the results obtained in the univariable CSH model, the significant risk factors for death are: (1) APD (CSHR=0.63, 95% CI 0.52-0.78); (2) first treatment (haemodialysis, considering peritoneal dialysis as reference category - CSHR=1.28, 95% CI 1.02-1.62); (3) reason for peritoneal dialysis (CSHR=0.79, 95% CI 0.64-0.97); (4) diabetes as time-independent covariate (CSHR=1.27, 95% CI 1.01-1.59); (5) peritonitis as time-dependent covariate (CSHR=1.56, 95% CI 1.01-2.44).
In multivariable SH model, the final model (adjusted for gender) includes as significant risk factors for death: (1) age as time-independent covariate (SHR=1.05, 95% CI=1.03-1.06); (2) diabetes as time-independent covariate (SHR=2.04, 95% CI=1.34-3.11); (3) reason for peritoneal dialysis (SHR=0.62, 95% CI=0.42-0.93). Risk of death increased with age, and was also higher for diabetic patients and for patients included in the peritoneal program because of an access failure.
The final multivariable CSH model (adjusted for gender) found diabetes (time-independent covariate - CSHR=1.30, 95% CI 1.03-1.66) and reason for peritoneal dialysis (CSHR=0.77, 95% CI 0.62-0.96) as significant risk factor for death and, additionally, DPA (CSHR=0.64, 95% CI 0.52-0.79).

4 Conclusion

In this work, we present two different approaches in the analysis of time-to-event data in the presence of competing risks and each method has different advantages. When we estimate the cumulative incidence function or its hazard (SH), the main advantages are: it is a direct approach; it compares the observed probabilities of events or the observed rates of events; it does not assume independence between the types of events. In the case of the comparison of the CSH, the advantages are: it gives insight into the biological mechanism; it is invariant to the size of the competing risks [9].
Different results were obtained according the type of hazard considered and the decision about the choice of the model depends on the research question. CSH may be more relevant when the disease aetiology is of interest, since it quantifies the event rate among the ones at risk of developing the event of interest. Cumulative incidences are easier to interpret and are more relevant for the purpose of prediction [1].

Bibliography

[1] Andersen, P.K., Geskus, R.B., de Witte, T. and Putter, H. (2012). Competing risks in epidemiology: possibilities and pitfalls. *International Journal of Epidemiology.* 41(3), 861–870.

[2] Bajorunaite, R. and Klein, J.P. (2008). Comparison of failure probabilities in the presence of competing risks. *J Stat Comput Sim.* 78(10), 951–966.

[3] Fine, J.P. and Gray, R.J. (1999). A proportional hazards model for the subdistribution of a competing risk. *J Am Stat Assoc.* 94(446), 496–509.

[4] Gooley, T.A., Leisenring, W., Crowley, J. and Storer, B.E. (1999). Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. *Stat Med.* 18(6), 695–706.

[5] Klein, J.P. (2006). Modelling competing risks in cancer studies. *Stat Med.* 25(6), 1015–1034.

[6] Kleinbaum, D.G. and Klein, M. (2005). *Survival Analysis: A Self-Learning Text*, Springer Verlag, New York.

[7] Koller, M.T., Raatz, H., Steyerberg, E.W. and Wolbers, M. (2012). Competing risks and the clinical community: irrelevance or ignorance? *Stat Med.* 31(11), 1089–1097.

[8] Pintilie, M. (2007). Analysing and interpreting competing risk data. *Statist. Med.* 26(6), 1360–1367.

[9] Pintilie, M. (2006). *Competing risks. A practical perspective.*, John Wiley and Sons, New Jersey.

[10] Varadhan, R., Weiss, C.O., Segal, J.B., Wu, A.W., Scharfstein, D. and Boyd, C. (2010). Evaluating Health Outcomes in the Presence of Competing Risks A Review of Statistical Methods and Clinical Applications. *Med Care.* 48(6), S96–S105.

Mixture Lorenz curves. Three new models

Drăgulin Mircea[*][1] **and Gheorghe Carmen Adriana**[2]

[1]*Faculty of Mathematics and Computer Science, University of Bucharest*
[2]*National Institute of Economic Research, Romanian Science Academy*

Abstract

The Lorenz curve is one of the most investigated and also significant tool in the study of distribution and inequality of income. The main difficulty in finding a good analytical form is the lack of fitting on some intervals, especially in the tail of the function. Mixture parametric approach may overdue these problematical issues by introducing better constraints.
In this paper, three new mixture Lorenz Curves are generated from initial Lorenz Curve families. In order to analyze the inequality in the income distribution, for the third proposed curve the Gini indexes are obtained.

Keywords: parametric Lorenz curve, Gini index.
AMS subject classifications: 60E15, 91B82

1 Introduction. Lorenz Curve

In 1905 was published in ”Publications of the American Statistical Association” a short paper called ”Methods of measuring the concentration of wealth”. The article proposed a simple method, named later Lorenz Curve, for the view of distribution of income or wealth according to the inequality or concentration of the income gained. Max Otto Lorenz completed his doctoral study at the University of Wisconsin – Madison without any reference to this paper, his only publication in a scientific journal.
The term ”Lorenz Curve” appears in the first statistical methods book from America. It was written by King (1912)[7] primarily for the use of sociologists, political or administration economists. After 1970 the papers of Atkinson (1970)[1] and Gastwirth (1971)[2] the interest on LC distribution increased.
Let L be the class of all non-negative random variable with finite mean and let X from L, with the probability distribution function $f(x)$. Then the distribution function $F(x) = \int_0^x f(y)dy$ will be the percent of units with the income less or equal to x. The values of $F(x)$ are between 0 and 1. We assume there exists the mean of income, and has the form $\mu = \int_0^\infty f(x)dx$. Then the first order moment of X will be $F_1(x) = \mu^{-1}\int_0^x yf(y)dy$ and it represents the share of total income earned by a person with the income less or equal to x. The graphic representation on the unit square that has $F(x)$ on the abscissa and $F_1(x)$ on the ordinate represents the Lorenz Curve, where x takes values from 0 to ∞.
Definition 1.1. - Gastwirth (1971)[2] - Let $X \in L$ with the density function $F(\cdot)$ and its inverse $F^{-1}(y) = inf\{x : F(x) \geq y\}$. The Lorenz curve $L(\cdot)$ is defined by

$$L(p) = \mu^{-1}\int_0^p F^{-1}(y)dy; 0 \leq p \leq 1. \tag{1}$$

In fact, the Lorenz curve is the correlation between the percentage of population and the percentage of income that they earn.
Kakwani (1980)[6] showed that the necessary properties of Lorenz Curve existence are:

[*] Corresponding author, e-mail: mircea.mate@yahoo.com

A. $L(p) = 0$, if $p = 0$;

B. $L(p) = 1$, if $p = 1$;

C. $L'(0^+) \geq 0$, for any $0 \leq p \leq 1$;

D. $L''(p) \geq 0$, for any $0 \leq p \leq 1$.

2 Parametric families of Lorenz Curves

Kakwani and Podder (1973[4], 1976[4]) proposed the first models to estimate parametric Lorenz curves. In 1973 they introduced the curve $L(p) = p^{\gamma} e^{-\eta(1-p)}$, for $0 < p < 1$ and $\eta > 0; ; 1 < \gamma < 2$. Using the coordinate system proposed by Gini in 1932, of the form $\eta = \frac{u+v}{\sqrt{2}}$ and $\pi = \frac{u-v}{\sqrt{2}}$, where $0 < u < 1$, Kakwani and Podder gave another definition of the curve $v = L(u)$, characterized by $\eta = a\pi^{\alpha}(\sqrt{2} - \pi)^{\beta}$, with $a \geq 0$, $0 \leq \alpha \leq 1$ and $0 < \beta \leq 1$.
Further, we propose three new models of parametric Lorenz Curves:

$$L_1(p; \theta, \nu) = \frac{1}{ln(p)} \frac{p^{\nu} - p^{\theta}}{\nu - \theta}, \nu > \theta \tag{2}$$

$$L_2(p; \theta, k, \nu) = \frac{1}{ln(p)} \frac{p^{\nu} - p^{\theta}}{\nu - \theta}[1 - (1-p)^k], \nu > \theta; \theta \geq 0; ; 0 < k \leq 1 \tag{3}$$

$$L_3(p; \theta, \nu) = pe^{-\theta(2-p)} \frac{e^{\nu} - e^{\theta}}{\nu - \theta}, \nu > \theta; \theta \geq 0 \tag{4}$$

Particular cases:
After applying the limit on ν, with $\nu \to \theta$ we get

$$\lim_{\nu \to \theta} L_1(p; \theta, \nu) = p^{\theta} \tag{5}$$

$$\lim_{\nu \to \theta} L_2(p; \theta, \nu) = p^{\theta}[1 - (1-p)^k] \tag{6}$$

$$\lim_{\nu \to \theta} L_3(p; \theta, \nu) = pe^{-\theta(1-p)} \tag{7}$$

Theorem 2.1. *Assume that $L_1(p; \theta, \nu)$ is defined and continuous on $[0, 1]$, with the second derivative $L"(\cdot)$. The function $L_1(\cdot)$ is a Lorenz Curve if and only if $L_1(0; \theta, \nu) = 0, L'_1(0^+; \theta, \nu) \geq 0, L_1(1; \theta, \nu) = 1, L''_1(p; \theta, \nu) \geq 0, p \in (0, 1)$.*

Proof:

$$\begin{matrix} lim \\ p \to 0 \\ p > 0 \end{matrix} \quad L_1(p; \theta, \nu) = \frac{1}{\nu - \theta} \begin{matrix} lim \\ p \to 0 \\ p > 0 \end{matrix} \quad (\nu p^{\nu} - \theta p^{\theta}) = 0$$

$$\begin{matrix} lim \\ p \to 1 \\ p < 1 \end{matrix} \quad L_1(p; \theta, \nu) = \frac{1}{\nu - \theta} \begin{matrix} lim \\ p \to 1 \\ p < 1 \end{matrix} \quad (\nu p^{\nu} - \theta p^{\theta}) = 1$$

$$L'_1(0^+) = \frac{1}{\nu - \theta} \begin{matrix} lim \\ p \to 0 \\ p > 0 \end{matrix} \quad \frac{\nu(\nu-1)p^{\nu-1} lnp + p^{\nu-1} + (\theta - 1)p^{\theta-1}}{ln(p)} = \frac{\nu(\nu-1)^2}{2(\nu - \theta)}$$

$$\begin{matrix} lim \\ p\to 0 \\ p>0 \end{matrix} \quad \frac{p^{-1}}{(1-\nu)p^{-\nu}} = 0.$$

The most known way to measure inequality using Lorenz curves is the Gini index. It was introduced in the article ”Variability and Mutability” (1912) by Corrado Gini (1884-1965), an famous Italian sociologist and demographer.

The Gini index is the ratio that has at the numerator (A) the area between egalitarian line and the Lorenz Curve and at the denominator (A+B) all the area under the first bisector: $G = \frac{A}{A+B}$. A small value of the Gini index indicates a smooth distribution of the income. In practice we won't get the value 0 that corresponds to equal incomes and 1 which means totally inequality among the income units.

A well known definition of the Gini index using Lorenz curve is $G = 1 - 2\int_0^1 L(p)dp$. Using this formula we obtained the Gini indexes of the new parametric Lorenz Curves. We state here the index for L_3:

$$G_3 = 1 - 2\int_0^1 L_3(p)dp.$$

$$G_3 = 1 - 2\frac{1-e^{\nu-\theta}}{\theta^2(\theta-\nu)}\left(\theta - 1 + e^{-\theta}\right)$$

3 New Mixture Lorenz Curves

Mixture Lorenz curves are an important way to get a better data fit by constructing more complex models that combines a parametric Lorenz curve with known distribution function. The mixture method was introduced by Sarabia (2005)[8].

Let $L_1(p;\theta,\nu)$ be a parametric Lorenz curve, with parameter vectors θ,ν. For example, θ can be a scalar parameter that represents an factor of the homogeneity of the population.

Let $\pi(\theta;\alpha,\lambda)$ an absolute continuous probability density function, where α,λ are real parameters.

Theorem 2: The expression $\tilde{L}_1(p;\nu;\alpha,\lambda) = \int_\Theta L_1(p;\theta,\nu)\pi(\theta;\alpha,\lambda)d\theta$ defines a Lorenz curve, where $\pi(\theta;\alpha,\lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)}(\theta-1)^{\alpha-1}e^{-\lambda(\theta-1)}I(\theta>1)$, for any $\alpha,\lambda>0$, and $\theta>1$.

Proof:

$$\tilde{L}_1(p;\nu;\alpha,\lambda) = \int_1^\infty L_1(p;\theta,\nu)\pi(\theta;\alpha,\lambda)d\theta =$$

$$= \frac{\lambda^\alpha}{ln(p)\cdot\Gamma(\alpha)}\int_1^\infty \frac{p^\nu - p^\theta}{\nu-\theta}(\theta-1)^{\alpha-1}e^{-\lambda(\theta-1)}d\theta$$

After a few steps and applying the change of variable $t = \theta - 1$ we get:

$$\tilde{L}_1(p;\nu;\alpha,\lambda) = \frac{\lambda^\alpha p(1-\nu)^{\alpha-2}e^{-\lambda+1}}{ln(p)}\left[pE_n(-\nu\lambda - \nu lnp) - E_n(-\lambda\nu)\right],$$

Where E_n is the exponential integral defined by $E_n = -\int_{-x}^\infty \frac{e^{-t}dt}{t}$.

The necessary conditions for the existence of Lorenz Curve $\tilde{L}_1$ will be:

$$\tilde{L}_1(0;\nu;\alpha,\lambda) = 0; \widetilde{L'}_1(0^+;\nu;\alpha,\lambda) \geq 0$$

$$\tilde{L}_1(1;\nu;\alpha,\lambda) = 1; \widetilde{L''}_1(1;\nu;\alpha,\lambda) \geq 0, p \in (0,1).$$

These can be proved as the proof of theorem 1.

Theorem 3: The expression $\tilde{L}_2(p;\alpha,\lambda,k,\nu) = \int_0^\infty L_2(p;\theta,k,\nu)\pi(\theta;\alpha,\lambda)d\theta$ defines a lorenz Curve, where $\pi(\theta;\alpha,\lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\lambda\theta}I(\theta>0)$, for any $\alpha,\lambda>0$.

Theorem 4: The expression $\tilde{L}_3(p;\alpha,\lambda,\nu) = \int_0^\infty L_3(p;\theta,\nu)\pi(\theta;;\alpha,\lambda)d\theta$ defines a lorenz Curve, where $\pi(\theta;\alpha,\lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\lambda\theta}I(\theta>0)$, for any $\alpha,\lambda>0$.
Next we want to find the Gini index for the new Lorenz curve $\tilde{L}_3$. For this we will use Theorem 3 from Sarabia (2005)[8] that gives the following property:

$$G(\tilde{L}_3) = E_\pi[G(L_3)],$$

where $E_\pi$is the mathematical expectation of $G(L_3)$ as defined in section 2 with respect to the probability density function $\pi(\theta)$. We get

$$G_{\tilde{L}_3}(\lambda,\nu,\alpha) = \int_0^\infty \left[1-2\frac{1-e^{\nu-\theta}}{\theta^2(\theta-\nu)}\left(\theta-1+e^{-\theta}\right)\right]\frac{\lambda^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\lambda\theta}d\theta$$

The final form of the Gini index is given by,

$$G_{\tilde{L}_3}(\lambda,\nu,\alpha) = 1-2\lambda^\alpha e^{-\nu\lambda}(-\nu)^{\alpha-3}\frac{\Gamma(\alpha-2)}{\Gamma(\alpha)}\left[\Gamma(3-\alpha,-\nu\lambda)-\Gamma(3-\alpha,\nu(1+\lambda))\right]$$

4 Conclusions

The importance of Lorenz Curves in economic and statistical analysis in the inequality of income and wealth motivates the desire to find new parametric families of Lorenz curves. Multitude of parametric models proposed in the literature is not an inconvenience, but an additional reason given by the mismatch of the empirical curves in totally on the data set of income. We conclude that mixture parametric approach gives a better fit by introducing tighter constraints. In this paper we defined three mixture Lorenz Curves which are generated from new initial Lorenz Curve families. In order to analyse the inequality in the income distribution, for the third proposed curve the Gini indexes are obtained. As a further research, by using appropriate statistical tools there can be made comparisons between the new mixture curves and classical ones proposed by Sarabia[8]. The parameter estimates of the models can be obtained by using non-linear last squares.

Acknowledgements: We would like to thank to our supervisor of this project, professor Vasile Preda for the valuable guidance and advice.

Bibliography

[1] Atkinson, A.B. (1970). On the measurement of inequality, *J. Economic Theory* 2, 244-263

[2] Gastwirth, J. L., (1971). A General Definition of the Lorenz Curve, *Econometrica*, 39, 1037-1039.

[3] Kleiber, C. (2007). The Lorenz curve in economics and econometrics, *WWZ Working paper*.

[4] Kakwani, N. C.; Podder, N. (1973). On the estimation of Lorenz curves from grouped observations, *International Economic Review*, 14, 278-292.

[5] Kakwani, N. C.; Podder, N. (1976). Efficient estimation of the Lorenz curve and associated inequality measures from grouped observations, *Econometrica*, 44, 137-148.

[6] Kakwani, N. C. (1980). On a class of poverty measures, *Ecomometrica*, 48, 437-446

[7] King, W. I. (1912). The elements of statistical method, *The Macmillan Co.*, 89, 156-158.

[8] Sarabia, J.M.; Castillo, E.; Pascual, M; Sarabia, M. (2005). Mixture Lorenz curves. *Economics Letters,* Elsevier, vol. 89(1).

[9] Wang, Z. X.; Ng, Y-K.; Smyth, R. (2007). Revisiting the ordered family of Lorenz curves. *Discussion paper 19/07* Department of Economics Monash University.

On oracle inequality for exponential weighting of ordered smoothers

Chernousova, E.[1], **Golubev, Yu.**[2] **and Krymova, E.**[*3]

[13] *Moscow Institute of Physics and Technology (State University),*
[23] *Institute for Information Transmission Problems ,*
[2] *CNRS, Université de Provence,*
[3] *DATADVANCE*

Abstract

This paper deals with recovering an unknown vector from noisy data with the help of special family of linear estimates, namely, a family of ordered smoothers. The estimators withing this family are aggregated using the exponential weighting method. Our goal is to derive oracle inequalities controlling the risk of the aggregated estimate. Based on probabilistic properties of the unbiased risk estimate, we show that for the exponential weighting we can get better remainder terms than that one in Kneip's oracle inequality [9].

Keywords: ordered smoother, exponential weighting, unbiased risk estimation, oracle inequality
AMS subject classifications: 62G05

1 Introduction and main results

In this paper we focus on a simple sequence space model

$$Y_i = \mu_i + \sigma\xi_i, \quad i = 1, 2, \dots, n, \tag{1}$$

where ξ_i is a standard white Gaussian noise. For the sake of simplicity it is assumed that the noise level $\sigma > 0$ is known.
The goal is to estimate an unknown vector $\mu \in \mathbb{R}^n$ based on the data $Y = (Y_1, \dots, Y_n)^\top$. In this paper, μ is recovered with the help of linear estimates

$$\hat{\mu}_i^h(Y) = h_i Y_i,\ h \in \mathcal{H}, \tag{2}$$

where $\mathcal{H}$ is a finite set of vectors in $\mathbb{R}^n$ which will be described later on.
In what follows, the risk of an estimate $\hat{\mu}(Y) = (\hat{\mu}_1(Y), \dots, \hat{\mu}_n(Y))^\top$ is measured by

$$R(\hat{\mu}, \mu) = \mathbf{E}_\mu \|\hat{\mu}(Y) - \mu\|^2,$$

where $\mathbf{E}_\mu$ is the expectation with respect to the measure $\mathbf{P}_\mu$ generated by the observations from (1) and $\|\cdot\|$, $\langle\cdot,\cdot\rangle$ stand for the norm and the inner product in $\mathbb{R}^n$

$$\|x\|^2 = \sum_{i=1}^n x_i^2, \quad \langle x, y\rangle = \sum_{i=1}^n x_i y_i.$$

It is seen easily that the mean square risk of $\hat{\mu}^h(Y)$ is computed as follows

$$R(\hat{\mu}^h, \mu) = \|(1-h)\mu\|^2 + \sigma^2\|h\|^2.$$

[*] Corresponding author, e-mail: ekkrym@gmail.com

This risk depends on $h \in \mathcal{H}$ and one can minimize it choosing properly $h \in \mathcal{H}$. Often the minimal risk

$$r^{\mathcal{H}}(\mu) = \min_{h \in \mathcal{H}} R(\hat{\mu}^h, \mu)$$

is called the oracle risk.
Obviously, one cannot make use of the oracle estimate

$$\mu^*(Y) = h^* \cdot Y, \quad h^* = \arg\min_{h \in \mathcal{H}} R(\hat{\mu}^h, \mu)$$

because it depends on the underlying vector. However, one could try to construct an estimator $\tilde{\mu}^{\mathcal{H}}(Y)$ based on the family of linear estimates $\hat{\mu}^h(Y),\ h \in \mathcal{H}$, with the risk mimicking the oracle risk. This idea means that the risk of $\tilde{\mu}^{\mathcal{H}}(Y)$ should be bounded by the so-called oracle inequality

$$R(\tilde{\mu}^{\mathcal{H}}, \mu) \le r^{\mathcal{H}}(\mu) + \tilde{\Delta}^{\mathcal{H}}(\mu),$$

which holds uniformly in $\mu \in \mathbb{R}^n$. In general, such an estimator doesn't exist, but for certain statistical models it is possible to construct an estimator $\tilde{\mu}^{\mathcal{H}}(Y)$ (see, e.g., Theorem 1.1 below) such that:

- $\tilde{\Delta}^{\mathcal{H}}(\mu) \le \tilde{C} r^{\mathcal{H}}(\mu)$ for all $\mu \in \mathbb{R}^n$, where $\tilde{C} > 1$ is a constant.
- $\tilde{\Delta}^{\mathcal{H}}(\mu) \ll r^{\mathcal{H}}(\mu)$ for all $\mu :\ r^{\mathcal{H}}(\mu) \gg \sigma^2$.

It is well-known that one can find such an estimator provided that $\mathcal{H}$ is not very rich (see, e.g., [2]). In particular this can be done for the so-called ordered smoothers [9]. This is why this paper deals with the set of ordered multipliers $\mathcal{H}$ defined as follows:

- $h_i \in [0, 1],\ i = 1, \dots, n$ for all $h \in \mathcal{H}$,
- $h_{i+1} \le h_i,\ i = 1, \dots, n$ for all $h \in \mathcal{H}$,
- if for some integer k and some $h, g \in \mathcal{H}$, $h_k < g_k$, then $h_i \le g_i$ for all $i = 1, \dots, n$.

The last condition means that vectors in $\mathcal{H}$ may be naturally ordered, since for any $h, g \in \mathcal{H}$ there are only two possibilities $h_i \le g_i$ or $h_i \ge g_i$ for all $i = 1, \dots, n$. Notice that ordered smoothers are common in statistics (see, e.g., [9]). For example, smoothing splines, spectral regularization methods (see [15], [6]).
Nowadays, there are a lot of approaches aimed to construct an estimate mimicking the oracle risk. At the best of our knowledge, the principal idea in obtaining such estimates goes back to [1] and [11] and related to the method of the unbiased risk estimation [14]. The literature on this approach is so vast that it would be impractical to cite it here. We mention solely the following result by Kneip [9] since it plays an important role in our presentation. Denote by

$$\bar{r}(Y, \hat{\mu}^h) \stackrel{\text{def}}{=} \|Y - \hat{\mu}^h(Y)\|^2 + 2\sigma^2 \sum_{i=1}^{n} h_i - \sigma^2 n, \tag{3}$$

the unbiased risk estimate of $\hat{\mu}^h(Y)$.

Theorem 1.1. *Let*

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \bar{r}(Y, \hat{\mu}^h).$$

Then uniformly in $\mu \in \mathbb{R}^n$,

$$\mathbf{E}_\mu \|\hat{\mu}^{\hat{h}} - \mu\|^2 \le r^{\mathcal{H}}(\mu) + K\sigma^2 \sqrt{1 + \frac{r^{\mathcal{H}}(\mu)}{\sigma^2}}, \tag{4}$$

where K is a universal constant.

Another way to construct a good estimator based on the family $\hat{\mu}^h,\ h \in \mathcal{H}$ is to aggregate the estimates within this family using a held-out sample (see [12],[3]).
To overcome the well-know drawbacks of sample splitting one would like to construct estimators using the same observations and performing the aggregation. This can be done, for instance, with the help of the exponential weighting. The motivation of this method is related to the problem of functional aggregation, see [13]. It has been shown that this method yields rather good oracle inequalities for certain statistical models [10], [5], [13].
The exponential weighting estimate is defined as follows:

$$\bar{\mu}(Y) = \sum_{h\in\mathcal{H}} w^h(Y)\hat{\mu}^h(Y), \tag{5}$$

where

$$w^h(Y) = \pi^h \exp\biggl[-\frac{\bar{r}(Y,\hat{\mu}^h)}{2\beta\sigma^2}\biggr] \bigg/ \sum_{g\in\mathcal{H}} \pi^g \exp\biggl[-\frac{\bar{r}(Y,\hat{\mu}^g)}{2\beta\sigma^2}\biggr],\ \beta > 0,$$

and $\bar{r}(Y,\hat{\mu}^h)$ is the unbiased risk estimate of $\hat{\mu}^h(Y)$ defined by (3).
The main goal of this paper is to show that for the exponential weighting we can get oracle inequalities with smaller remainder terms than that one in Theorem 1.1, Equation (4).
In order to cover $\mathcal{H}$ with low and very hight cardinalities, we make use of the special prior weights defined as follows:

$$\pi^h \stackrel{\text{def}}{=} 1 - \exp\biggl\{-\frac{\|h^+\|_1 - \|h\|_1}{\beta}\biggr\}.$$

Here

$$h^+ = \min\{g \in \mathcal{H} : g > h\}$$

$\pi^{h_{\max}} = 1$, where $h^{\max}$ is the maximal multiplier in $\mathcal{H}$, and $\|\cdot\|_1$ stands for the l_1-norm in $\mathbb{R}^n$, i.e.,

$$\|h\|_1 = \sum_{i=1}^n |h_i|.$$

Along with these weights we will need also the following condition:

Condition 1. *There exists a constant $K_\circ \in (0,\infty)$ such that*

$$\|h\|^2 - \|g\|^2 \ge K_\circ\bigl(\|h\|_1 - \|g\|_1\bigr) \tag{6}$$

for all $h \ge g$ from $\mathcal{H}$.

The next theorem, yielding an upper bound for the mean square risk of $\bar{\mu}(Y)$ defined by (3.1), is the main result of this paper.

Theorem 1.2. *Assume that $\mathcal{H}$ is a set of ordered multipliers, $\beta \ge 4$, and Conditions 1 holds. Then, uniformly in $\mu \in \mathbb{R}^n$,*

$$\mathbf{E}_\mu\|\bar{\mu} - \mu\|^2 \le r^{\mathcal{H}}(\mu) + 2\beta\sigma^2 \log\biggl[C\biggl(1 + \frac{r^{\mathcal{H}}(\mu)}{\sigma^2}\biggr)\biggr]. \tag{7}$$

Here and in what follows $C = C(K_\circ, \beta)$ denotes strictly positive and bounded constants depending on $K_\circ, \beta$.

We begin a short discussion concerning this theorem. We finish this section with some remarks regarding this theorem.

Remark 1. The condition $\beta \geq 4$ may be improved when the ordered multipliers $h \in \mathcal{H}$ take only two values 0 and 1. In this case it is sufficient to assume that $\beta \geq 2$ (see [8]).

Remark 2. Usually Condition 1 may be checked rather easily. For instance, for smoothing splines and equidistant design, the set of ordered multipliers is given by

$$\mathcal{H} = \left\{ h : \ h_k = \frac{1}{1+\alpha\lambda_k}, \ \alpha \in \mathbb{R}^+ \right\}.$$

with $\lambda_k \asymp (\pi k)^{2m}$ for large k (see [4] for details). Heuristically, for small α and large n we have

$$\|h^\alpha\|^2 \approx \pi^{-1}\alpha^{-1/(2m)} \int_0^\infty \frac{1}{[1+x^{2m}]^2}\, dx$$

and

$$\|h^\alpha\|_1 \approx \pi^{-1}\alpha^{-1/(2m)} \int_0^\infty \frac{1}{1+x^{2m}}\, dx.$$

With these equations Condition 1 becomes obvious. A rigorous proof of (6) is based on a non-asymptotic version of these arguments. It is technical but unfortunately cumbersome and therefore, in order not to overload the paper, we omit it.

Remark 3. In contrast to Proposition 2 in [5], the remainder term in (7) does not depend neither on the cardinality of $\mathcal{H}$ nor n. It has the same structure as Kneip's oracle inequality in Theorem 1.1.

Remark 4. Comparing (7) with (4), we see that when $r^{\mathcal{H}}(\mu)/\sigma^2 \approx 1$, then the remainder terms in (4) and (7) have the same order, namely, σ^2. However, when $r^{\mathcal{H}}(\mu)/\sigma^2 \gg 1$, we get

$$2\beta\sigma^2 \log\left[C\left(1 + \frac{r^{\mathcal{H}}(\mu)}{\sigma^2}\right)\right] \ll K\sigma^2\sqrt{1+\frac{r^{\mathcal{H}}(\mu)}{\sigma^2}},$$

thus showing that the upper bound for the remainder term in the oracle inequality related to the exponential weighting is better than the one in Theorem 1.1.

Acknowledgements: This research was supported by Laboratory of Structural Methods of Predictive Modeling and Optimization, Moscow Institute Physics and Technology, Russian Federation government grant, ag. 11.G34.31.0073.

Bibliography

[1] AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle *Proc. 2nd Intern. Symp. Inf. Theory*. 267–281.

[2] BIRGÉ, L. and MASSART, P. (2007). Minimal penalties for Gaussian model selection, *Probab. Theory Relat. Fields* **138** 33–73.

[3] CATONI, O. (2004). *Statistical learning theory and stochastic optimization.* Lectures Notes in Math. **1851** Springer-Verlag, Berlin.

[4] CHERNOUSOVA E., GOLUBEV YU., KRYMOVA E. Ordered smoothers with exponential weighting // *arXiv:1211.4207 [Probab. Theory Relat. Fields].*

[5] DALAYAN, A. and SALMON J. (2011). Sharp oracle inequalities for aggregation of affine estimators. *arXiv:1104.3969v2 [math.ST].*

[6] ENGL, H.W., HANKE, M., and NEUBAUER, A. (1996). *Regularization of Inverse Problems. Mathematics and its Applications, 375.* Kluwer Academic Publishers Group. Dordrecht.

[7] GOLUBEV, YU. (2010). On universal oracle inequalities related to high dimensional linear models. *Ann. Statist.* **38** No. 5 2751-2780.

[8] GOLUBEV, YU. (2012). Exponential weighting and oracle inequalities for projection methods. *Problems of Information Transmission* No. 3 *arXiv:1206.4285*

[9] KNEIP, A. (1994). Ordered linear smoothers. *Annals of Stat.* **22** 835–866.

[10] LEUNG, G. and BARRON, A. (2006). Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory* **52** no. 8 3396-3410.

[11] MALLOWS, C. L. (1973). Some comments on C_p *Technometrics* **15** 661–675.

[12] NEMIROVSKI, A. (2000). *Topics in non-parametric statistics.* Lectures Notes in Math. **1738** Springer-Verlag, Berlin.

[13] RIGOLET, PH. and TSYBAKOV, A. (2011). Sparse estimation by exponential weighting. *arXiv:1108.5116v1 [math.ST].*

[14] STEIN, C. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. of Statist.*, **9** 1135-1151.

[15] WAHBA, G. (1990). *Spline Models for Observational Data.* SIAM, Philadelphia.

A stochastic optimization method for constructing optimal block designs with linear constraints

Alena Bachratá *[1] **and Radoslav Harman**[1]

[1] *Department of Applied Mathematics and Statistics, Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava, Slovakia*

Abstract

We propose a stochastic optimization method for constructing efficient designs of experiments under a broad class of linear constraints on the design weights. The linear constraints can represent restrictions on various types of "limits" associated with the experiment. To illustrate the method, we computed D- and A-optimal designs for estimating a set of treatment contrasts in the case of block size-two experiments with upper limits on the number of replications of each non-control treatment.

Keywords: stochastic optimization, design of experiments, linear constraints, block designs.
AMS subject classifications: 62K05, 62K10.

1 Introduction

Consider an experiment consisting of b trials with real-valued observations $Y_1, ..., Y_b$ satisfying the linear regression model

$$Y_i = \mathbf{f}'(x_i)\tau + \varepsilon_i,\ i = 1, \ldots, b. \tag{1}$$

In this model, $\mathbf{f}(x_1), ..., \mathbf{f}(x_b) \in \mathbb{R}^v$ are known regression vectors, $\tau \in \mathbb{R}^v$ is a vector of unknown parameters, and $\varepsilon_1, ..., \varepsilon_b$ are independent identically distributed random errors with variance $\sigma^2 < \infty$. The design points $x_1, ..., x_b$ are selected from a finite design space $\mathfrak{X}$.
Let the function $\xi : \mathfrak{X} \to \{0, 1, 2 \ldots\}$ be an exact experimental design, that is, $\xi(x)$ is the number of trials to be performed in the design point $x \in \mathfrak{X}$. The moment matrix for the design ξ is given by

$$\mathbf{M}(\xi) = \sum_{i=1}^{b} \xi(x_i)\mathbf{f}(x_i)\mathbf{f}'(x_i).$$

The $v \times v$ positive semidefinite matrix $\mathbf{M}(\xi)$ captures the amount of the information about the vector τ of all unknown parameters. Therefore, the aim of optimal design of experiment is to select the design ξ such that some aspect of the "size" of $\mathbf{M}(\xi)$ is maximized, depending on the aim of the experiment.
In many applications, the aim is to estimate a linear parameter system $\mathbf{A}'\tau$, where $\mathbf{A}$ is a $v \times s$ matrix of a full rank $s \leq v$. It is well known that an unbiased linear estimator of $\mathbf{A}'\tau$ exists if and only if

$$\mathcal{M}(\mathbf{A}) \subseteq \mathcal{M}(\mathbf{M}(\xi)), \tag{2}$$

where $\mathcal{M}$ denotes the linear space generated by the columns of a matrix. We will call (2) estimability conditions. If the estimability conditions are satisfied, then the information about the linear parameter subsystem $\mathbf{A}'\tau$ gained from the experiment can be represented by the information matrix

$$(\mathbf{A}'\mathbf{M}^{-}(\xi)\mathbf{A})^{-1}, \tag{3}$$

[*] Corresponding author, e-mail: Alena.Bachrata@fmph.uniba.sk

which is thoroughly justified in [8]. Note that if the estimability conditions are satisfied, then the information matrix (3) does not depend on the choice of the generalized inverse $\mathbf{M}^-(\xi)$.
To select the best experimental design ξ, we can use the following real-valued measures of the information matrix, known as the criteria of $D_{\mathbf{A}}$-optimality, and $A_{\mathbf{A}}$-optimality, respectively (see [8], cf. [1]):

$$\Phi_D(\mathbf{M}(\xi)) = \begin{cases} \det^{-1/s}(\mathbf{A}'\mathbf{M}^-(\xi)\mathbf{A}) & \text{if } \mathcal{M}(\mathbf{A}) \subseteq \mathcal{M}(\mathbf{M}(\xi)), \\ 0 & \text{else,} \end{cases}$$

$$\Phi_A(\mathbf{M}(\xi)) = \begin{cases} \left(\frac{1}{s}\,\mathrm{trace}(\mathbf{A}'\mathbf{M}^-(\xi)\mathbf{A})\right)^{-1} & \text{if } \mathcal{M}(\mathbf{A}) \subseteq \mathcal{M}(\mathbf{M}(\xi)), \\ 0 & \text{else.} \end{cases}$$

The design ξ^* is called Φ-optimal (that is $D_{\mathbf{A}}$-optimal or $A_{\mathbf{A}}$-optimal, depending on the chosen criterion), if

$$\xi^* \in \mathrm{argmax}_{\xi \in \Xi} \Phi(\mathbf{M}(\xi)),$$

where Ξ is the set of all designs satisfying required constraints. Note that if the errors are normally distributed, then the $D_{\mathbf{A}}$-optimal design minimizes the volume of the s-dimensional confidence ellipsoid for the linear parameter system $\mathbf{A}'\tau$, and the $A_{\mathbf{A}}$-optimal design minimizes the sum of variances of the s components of $\mathbf{A}'\tau$, see, e.g., [7].
In this paper, we will consider the class of linear constraints of the form

$$\sum_{x \in \mathfrak{X}} c_j(x)\xi(x) \leq \gamma_j; \quad j = 1, \ldots, K. \tag{4}$$

We will assume that for any design point $x \in \mathfrak{X}$ we have $c_j(x) \geq 0$ for all $j \in \{1, ..., K\}$, and $c_j(x) > 0$ for at least one $j \in \{1, ..., K\}$. We will also assume that $\gamma_j > 0$ for all $j \in \{1, ..., K\}$. Note that the assumptions imply that the set Ξ of designs satisfying restrictions (4) is non-empty and bounded.
The constraints (4) can be used, for instance, to set an upper limit on the total number of all trials, upper limits on the numbers of trials in individual design points, or an upper limit on the total cost of the experiment, provided that each trial is associated with a cost depending on the design point.

2 The stochastic optimization method

In general, computing a Φ-optimal design of experiments is a difficult problem of discrete optimization, see, e.g., [5] for a brief recent review of possible computational approaches. For the purpose of computing Φ-optimal designs under the constraints (4), we propose the following stochastic optimization method.
The computation begins with a permissible design $\zeta \in \Xi$. In the first phase, which we call saturation, the method adds trials to ζ by the greedy method, until it achieves a design ξ that is "saturated", i.e., addition of any trial to ξ would entail violation of some of the constraints (4). In the second phase, which we call sub-saturation, the method removes a random number of trials from the design ξ to obtain a new "sub-saturated" design ζ. The phases of saturation and sub-saturation are alternately repeated N times. At the end, the saturated design with the best value of the optimality criterion Φ is chosen to be the output of the method. The principle is illustrated in Figure 1.

3 Optimal block size-two designs for estimating a set of treatment contrasts

Consider the block experiment with b blocks of size two and treatments labeled $1, 2, ..., v$. Let the observations $Y_1, ..., Y_b$ correspond to the differences of the two responses within the same blocks. More formally,

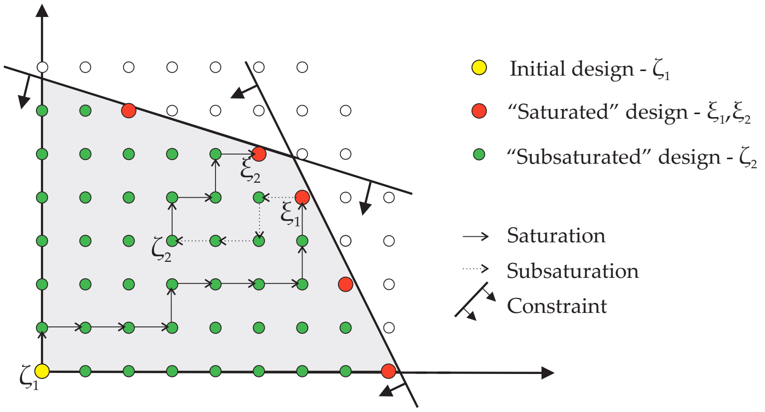

Initial design - ζ_1
“Saturated” design - ξ_1,ξ_2
“Subsaturated” design - ζ_2
Saturation
Subsaturation
Constraint
ζ_1
ζ_2
ξ_1
ξ_2

Figure 1: Illustration of the stochastic optimization algorithm for constructing constrained optimal designs.

we will assume that the design space is

$$\mathfrak{X} = \{(t_1, t_2) : 1 \leq t_1 < t_2 \leq v\},$$

and for each $(t_1, t_2) \in \mathfrak{X}$ the regressor $\mathbf{f}(t_1, t_2) \in \mathbb{R}^v$ is defined by $\mathbf{f}_{t_1}(t_1, t_2) = 1$, $\mathbf{f}_{t_2}(t_1, t_2) = -1$ and $\mathbf{f}_t(t_1, t_2) = 0$ for all $t \in \{1, ..., v\} \setminus \{t_1, t_2\}$. Then it is straightforward to show that the moment matrix of a design ξ is

$$\mathbf{M}(\xi) = \begin{pmatrix} r_1 & -a_{12} & -a_{13} & \dots & -a_{1v} \\ -a_{12} & r_2 & -a_{23} & \dots & -a_{2v} \\ -a_{13} & -a_{23} & r_3 & \dots & -a_{3v} \\ \dots & \dots & \dots & \dots & \dots \\ -a_{1v} & -a_{2v} & -a_{3v} & \dots & r_v \end{pmatrix}. \tag{5}$$

In the previous expression $a_{t_1 t_2} = \xi(t_1, t_2)$ for all $1 \leq t_1 < t_2 \leq v$, and

$$r_t = \sum_{(t_1,t_2) \in \mathfrak{X}_t} \xi(t_1, t_2), \; t = 1, ..., v,$$

where $\mathfrak{X}_t$ denotes the set of all design points (t_1, t_2) such that either $t_1 = t$ or $t_2 = t$. Therefore, the moment matrix (5) is equal to the information matrix of a block design with b blocks of size two, $r_1, ..., r_v$ replications of each treatment and $a_{t_1 t_2}$ occurrences of the treatments t_1 and t_2 in the same block (see, e.g., [2] and [3]).

In addition to the standard constraint $\sum_{(t_1,t_2) \in \mathfrak{X}} \xi(t_1, t_2) = b$ on the size of the experiment, we will assume that $\sum_{(t_1,t_2) \in \mathfrak{X}_t} \xi(t_1, t_2) \leq r$ for some given $r \in \mathbb{N}$ and for all $t \in \{2, ..., v\}$. That is, we assume that 1 is a “control” treatment, and each of the “non-control” treatments $2, ..., v$ can be replicated at most r times. Note that these constraints can be written in the form (4). Our aim will be to find the design that is optimal for estimating the set of $s = v - 1$ contrasts $\tau_1 - \tau_2, \dots, \tau_1 - \tau_v$. Formally, we will choose $\mathbf{A} = (\mathbf{1}_{v-1}, -\mathbf{I}_{v-1})'$, where $\mathbf{1}_{v-1} = (1, ..., 1)' \in \mathbb{R}^{v-1}$ and $\mathbf{I}_{v-1}$ is the identity matrix of type $(v-1) \times (v-1)$.

Every block design can be represented by a ”concurrence” graph with vertices corresponding to treatments and edges corresponding to the blocks. That is, if ξ is a design, then the set of the vertices of its concurrence graph is $\{1, ..., v\}$ and each couple t_1, t_2 of vertices is connected by $a_{t_1 t_2} = \xi(t_1, t_2)$ edges (cf. [4], [6], [2] and [3]). It can be shown that the problem of $D_{\mathbf{A}}$-optimality for the block size-two model is equivalent to the problem of constructing a graph with v vertices and b possibly multiple undirected edges, maximizing the number of its spanning trees (see [4] and [6]). Similarly, for our choice of the matrix $\mathbf{A}$, the problem of

$A_{\mathbf{A}}$-optimality is equivalent to the problem of constructing the graph with v vertices and b possibly multiple undirected edges, minimizing average electrical resistance between node 1 and all other nodes, if we assume that each edge has the electrical resistance of 1 ohm (cf. [2] and [3]).

Figure 2 shows the concurrence graphs of the resulting $D_{\mathbf{A}}$-optimal and $A_{\mathbf{A}}$-optimal designs of experiments for $v = 6$ treatments, $b = 5, \dots, 10$ blocks and the upper limit $r = 2$ on the number of replications of each non-control treatment. The designs were computed by the method explained in Section 2 with $N = 500$ iterations. The values of the criteria of $D_{\mathbf{A}}$-optimality and $A_{\mathbf{A}}$-optimality for all involved optimal designs are given in Table 1.

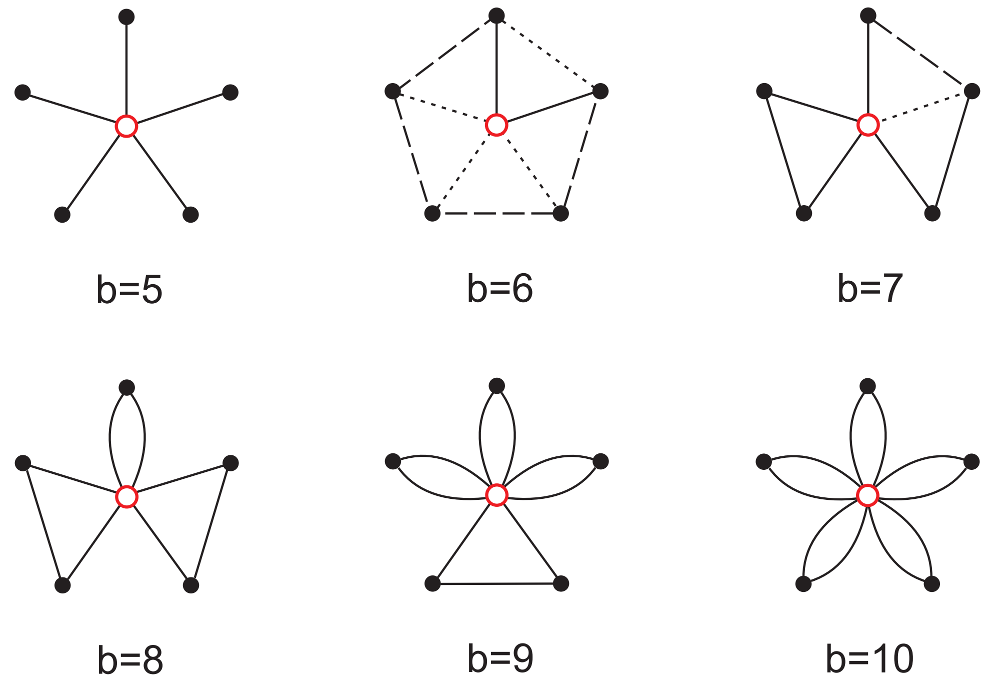


Figure 2: The concurrence graphs of the $D_{\mathbf{A}}$-optimal and the $A_{\mathbf{A}}$-optimal designs for the models with $v = 6$ treatments and $b = 5, ..., 10$ blocks of size two. The empty circles denotes the control treatments and the solid discs denote the non-control treatments. The edges represent the pairing of the treatments into blocks. The upper limit on the number of replications of each non-control treatment is $r = 2$. Note the for $b = 5, 8, 9, 10$ the $D_{\mathbf{A}}$-optimal and the $A_{\mathbf{A}}$-optimal designs coincide. For $b = 6, 7$ the edges of the concurrence graphs of the $D_{\mathbf{A}}$-optimal designs are denoted by the solid and dashed lines, and the edges of the concurrence graphs of the $A_{\mathbf{A}}$-optimal designs are denoted by the solid and dotted lines. Note that for $b = 6, 7$ the $A_{\mathbf{A}}$-optimal designs perform more replications of the control treatment than the $D_{\mathbf{A}}$-optimal designs.

	5	6D	6A	7D	7A	8	9	10
Φ_D	1.0000	1.4311	1.2457	1.6438	1.5518	1.7826	1.8882	2.0000
Φ_A	1.0000	0.8572	1.1538	1.3044	1.3636	1.5789	1.7647	2.0000

Table 1: The values of the criterion of $D_{\mathbf{A}}$-optimality (denoted by Φ_D) and the values of the criterion of $A_{\mathbf{A}}$-optimality (denoted by Φ_A) for optimal designs illustrated in Figure 2, see the text for details. The labels $5, 8, 9$ and 10 denote the designs for $b = 5, 8, 9, 10$ that are optimal for both criteria. The labels 6D and 7D denote the $D_{\mathbf{A}}$-optimal designs for $b = 6$ and $b = 7$, respectively. Similarly, 6A and 7A denote the $A_{\mathbf{A}}$-optimal designs for $b = 6$ and $b = 7$, respectively.

Acknowledgements: The research was supported by the Comenius University Grant UK/654/2013 and the Slovak VEGA Grant 1/0163/13.

Bibliography

[1] Atkinson, A. C. and Donev, A. N. and Tobias, R. D. (2007). *Optimum Experimental Designs, with SAS*, Oxford University Press.

[2] Bailey, R. A. (2007). Designs for two-colour microarray experiments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 56(4), 365–394.

[3] Bailey, R. A. and Cameron, P. J. (2009). Combinatorics of optimal designs. *Surveys in Combinatorics - London Mathematical Society Lecture Note Series* 365, 19-73.

[4] Cheng, C.-S. (1981). Maximizing the total number of spanning trees in a graph: Two related problems in graph theory and optimum design theory. *Journal of Combinatorial Theory* 31(2), 240–248.

[5] Harman, R. and Filová, L. (2013). Computing efficient exact designs of experiments using integer quadratic programming. *Computational Statistics and Data Analysis*, http://dx.doi.org/10.1016/j.csda.2013.02.021.

[6] Gaffke, N. (1982). D-optimal block designs with at most six varieties. *Journal of Statistical Planning and Inference* 6(2), 183–200.

[7] Pázman, A. (1986). *Foundations of Optimum Experimental Design*, Reidel.

[8] Pukelsheim, F. (2006). *Optimal Design of Experiments*, SIAM.

Stochastic interest rates in life insurance mathematics

Gábor Szűcs[*]

Comenius University, Faculty of Mathematics, Physics and Informatics, Department of Applied Mathematics and Statistics

Abstract

Basic life insurance mathematics applies some simplifications, e.g., the assumption of constant interest rates during the period of insurance. Insurance companies in Slovakia usually follow these assumptions and they calculate the premium using the technical interest rate. According to Decree of the National Bank of Slovakia the maximum technical rate of interest shall be 2.5% p. a. From a practical point of view, insurance corporations invest collected premiums on behalf of policyholders in different types of assets (e.g., bonds, shares, deposits). However, their yields have stochastic character, because the situation in the financial and capital markets is continually changing. For insurance companies it is important to know what kind of risks and losses they will face, if premium is computed using technical interest rate, while return on investments is not guaranteed. The aim of this paper is to present an alternative method for pricing the present value of potential future insurance losses. We assume that the potential losses are derived from the stochastic behavior of interest rates and market yields.

Keywords: technical rate of interest, actuarial present value, $ARIMA$ time series
AMS subject classifications: 91B30, 91G30

1 Introduction

Life insurance business is an important part of the insurance sector and the national economy. The primary function of life insurance is to provide protection, certainty as well as additional savings accumulation. We can say that a life insurance company is, strictly speaking, a ”set” of assets and liabilities. On the one side, premium paid by policyholders is subsequently allocated to various assets (e. g., bonds, shares, options, etc.) and, on the other side, the insurance companies have an obligation to provide insurance benefits upon occurrence of an insurance event.

The main aim of the actuarial mathematics is to develop appropriate models which can be applied to calibration of insurance products. For life insurance corporations one of the most important challenges is to correctly and accurately appreciate future liabilities (obligations). In this paper, we deal primarily with the examination of models by which the life insurance company is able to estimate the present value of their future expenses. We focus on the modeling of market interest rates, and thereby on estimation of potential future value of investments of a life insurance company.

The scheme of the paper is as follows: Section 2 describes the basic model of life insurance and methods of calculating the net single premium in case of different insurance types. Section 3 contains some important terms and definitions from the theory of stochastic processes, which are used to model stochastic interest rates. In Section 4, two different approaches of calculating the actuarial present value are compared.

[*] e-mail: szucs@fmph.uniba.sk

2 Basic model of life insurance

The basic model of life insurance applies some special notations and assumptions (see [1] or [5]), we will introduce only the most relevant ones. Let us consider a person of age x (a person aged x years, also called a *life aged* x). The probability that this person dies within the next year is denoted by the symbol q_x. The probability of complementary event, i. e., that the person of age x will survive to age $(x+1)$, is defined by the formula $p_x = 1 - q_x$. In actuarial life tables, there are generally given these one-year probabilities of death q_x, $\forall x \in \{0, 1, 2, \dots\}$. More generally, ${}_kp_x$ denotes the probability that the person of age x will survive at least k years and is defined by

$$ {}_kp_x = p_x \; p_{x+1} \; \dots \; p_{x+k-1} = \prod_{h=0}^{k-1} p_{x+h}, \quad k = 1, 2, 3, \dots $$

Similarly, ${}_kq_x$ is the probability that person dies within the coming k years. It can be expressed in the form

$$ {}_kq_x = {}_0p_x \; q_x + {}_1p_x \; q_{x+1} \; \dots \; {}_{k-1}p_x \; q_{x+k-1} = \sum_{h=0}^{k-1} {}_hp_x \; q_{x+h}, \quad \text{for } k = 1, 2, 3, \dots, $$

where ${}_0p_x \equiv 1$ and ${}_1p_x \triangleq p_x$.

2.1 Elementary insurance types

Life insurance is a contract between an insurance policy holder (insured) and an insurer, where the insured pays a *premium* and the insurer promises to pay a designated sum of money, the *sum insured*. The time and amount of sum insured may be random variables because of the stochastic character of the insured's future lifetime. One of the most important tasks of actuarial mathematics is to calculate the expected present value (EPV) of this payment. According to the principle of equivalence, the expected present value of the sum insured is equal to the net single premium. The EPV is in basic model calculated by discounting future cash payments by the technical interest rate, which is usually an effective annual rate of interest. Let us denote by ι the *technical interest rate* and by $\nu = (1+\iota)^{-1}$ the *discount factor*. The *force of interest*, denoted by δ, characterizes continuous compounding. The formula to convert between ι and δ is $\delta = \log(1+\iota)$.
More generally, for all $t \in \Upsilon$ we get following analogues

$$ (1+\iota)^t = \mathrm{e}^{\delta t}, \qquad \nu^t = \mathrm{e}^{-\delta t}, \tag{1} $$

where Υ is a given set.
A *pure endowment of duration* n *years* provides for payment of 1 unit at the end of the n-th year only if the insured survives until the age $(x+n)$. The net single premium is given by

$$ A_{x:\overset{\;1}{\overline{n|}}} = \nu^n \; {}_np_x = \mathrm{e}^{-\delta n} \; {}_np_x. \tag{2} $$

An n*-year term insurance* pays 1 unit at the end of the year of policyholder's death if he or she dies within the n-year period. The formula for the net single premium is

$$ A^1_{x:\overline{n|}} = \sum_{k=0}^{n-1} \nu^{k+1} \; {}_kp_x \; q_{x+k} = \sum_{k=0}^{n-1} \mathrm{e}^{-\delta(k+1)} \; {}_kp_x \; q_{x+k}. \tag{3} $$

An *endowment of duration* n *years* pays 1 unit either at end of the year of death of the insured or at the end of the n-th year if the insured survives until that time. The net single premium is denoted by $A_{x:\overline{n|}}$ and given by

$$ \begin{aligned} A_{x:\overline{n|}} &= \nu \; q_x + \nu^2 \; p_x \; q_{x+1} + \dots + \nu^n \; {}_{n-1}p_x \; q_{x+n-1} + \nu^n \; {}_np_x, \\ A_{x:\overline{n|}} &= A^1_{x:\overline{n|}} + A_{x:\overset{\;1}{\overline{n|}}}. \end{aligned} \tag{4} $$

3 Stochastic processes and time series

In this section we will introduce a stochastic model of force of interest applicable to pricing insurance products and to estimate present value of future payments. Let us consider the force of interest $\delta(t)$ which changes in time and has stochastic character. This function $\delta(t)$ and the stochastic actuarial present value have been studied in several papers (e. g., [4]). We was dealing with a methodology based on $ARIMA$ time series which can be used to model the stochastic interest rates.

Definition 1. (see [2]) Let us denote by $\mathbb{Z}$ the set of all integers. A discrete-time stochastic process (time series) $\mathbb{Y} = \{Y(t), t \in \mathbb{Z}\}$ is called *white noise* if $\mathbb{Y}$ is a sequence of uncorrelated random variables with mean 0 and variance σ^2, where $\sigma > 0$.

The *autoregressive moving average time series of orders* p *and* q is denoted by $ARMA(p, q)$ and defined by

$$X(t) = \sum_{k=1}^{p} \alpha_k \, X(t-k) + \sum_{m=0}^{q} \beta_m \, Y(t-m) \quad \text{for } t \in \mathbb{Z},$$

where $p > 0$, $q \geq 0$, $\alpha_1, \ldots, \alpha_p$, $\beta_1, \ldots, \beta_q$ are real parameters and $\beta_0 = 1$.

Definition 2. Consider a time series $\mathbb{X} = \{X(t), t \in \mathbb{Z}\}$. The *first order difference process* is denoted by $\nabla X(t)$ and defined by

$$\nabla X(t) = X(t) - X(t-1) \quad \text{for } t \in \mathbb{Z}.$$

Analogously, the *difference process of order* d can be defined as

$$\nabla^d X(t) = \underbrace{\nabla(\nabla(\ldots(\nabla}_{d-\text{times}} X)))(t) \quad \text{for } t \in \mathbb{Z}.$$

Definition 3.

A discrete-time process $\mathbb{X} = \{X(t), t \in \mathbb{Z}\}$ is called the *autoregressive integrated moving average time series*, denoted by $ARIMA(p, d, q)$, if $\nabla^d X(t)$ is an $ARMA(p, q)$ time series.

4 Practical application

As we mentioned in Section 2, pricing of life insurance products is generally made on the basis of technical interest rate ι. In this part we compare actuarial present value of a term insurance calculated by using the basic model and expected present value of future expenses computed with regard to the assumption of stochastic character of investment interest rates. Interest rates in the second approach are assumed to follow an $ARIMA$ time series.

Suppose that a person aged 60 has purchased a three-year term insurance for $5,000$ units. The benefit is payable at the end of the year of policyholder's death. The insurance company invests collected premium in assets with maturity of one year. If the person does not die, the insurer reinvests the accumulated premium back into the one-year assets. Probabilities of death are (see [7]): $q_{60} = 0.012353$; $q_{61} = 0.013612$, $q_{62} = 0.014531$. Firstly, we calculate the net single premium for this policy and then we estimate the present value of future payment under the assumption of stochastic development of interest rates.

Calculation of insurance premium

Technical rate of interest $\iota = 0.025$ p. a. converted to force of interest is $\delta_\iota = \log(1+\iota) = 0.02469261$ p. a. By applying probabilities of death and formula (3) we should obtain the net single premium $A^{\;1}_{60:\overline{3|}} = 189.97$ units.

Simulation study

Let us consider that force of interest is assumed to obey an $ARIMA$ time series. Euribor 12M interest rates from January 3, 2011 to April 26, 2013 (see [3]) served as reference data for the calibration of parameters of $ARIMA$ model. To calibrate the coefficients of $ARIMA$ model we applied the package `forecast` and the procedure `auto.arima()` in $\mathcal{R}$. Using the Akaike Information Criterion (AIC) the output of procedure was as follows:

```
# Series: euribor
# ARIMA(2,2,2)
#
# Coefficients:
#          ar1     ar2      ma1     ma2
#       0.4190  0.0883  -1.4043  0.4253
# s.e.  0.3105  0.0427   0.3102  0.2963
#
# sigma^2 estimated as 1.153e-08:  log likelihood=4576.9
# AIC=-9143.81   AICc=-9143.7   BIC=-9121.88
```

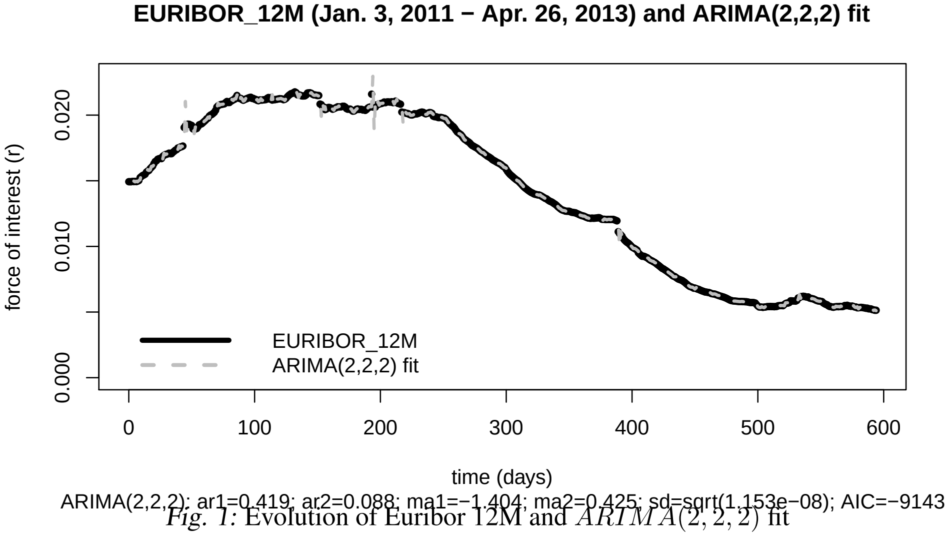


Fig. 1: Evolution of Euribor 12M and $ARIMA(2,2,2)$ fit

To calculate the present value of future payments related to the 3-year term insurance, we performed a simulation study which was carried out using the statistical software $\mathcal{R}$ [6]. We chose the following parameters: number of simulations $N = 5000$, starting value of the interest rate $\delta_0 = \log(1 + 0.00515)$ p. a. ($r_0 = 0.00515$ is the EURIBOR 12M interest rate from Apr. 26, 2013), length of the working year $y = 257$ days. We realized daily simulations for the full three year horizon, but technically were used only interest rates on the beginning of each year. Let us denote $\widetilde{A}^1_{60:\overline{3}|}$ the present value of future payments related to the abovementioned three-year term insurance. The mean of simulations was $\mathrm{E}\left[\widetilde{A}^1_{60:\overline{3}|}\right] = 198.10$ units, while the simulated 95% confidence interval (CI) for $\widetilde{A}^1_{60:\overline{3}|}$ was $(182.68, 215.51)$.

To make it fair and comparable with classical approach, we changed the starting value of simulations to $\delta_0 = \log(1+\iota) = \log(1+0.025)$ p. a. and carried out an additional simulation. The result was $\mathrm{E}\left[\widetilde{A}^1_{60:\overline{3}|}\right] = 190.40$ units with 95% CI $(176.23, 205.85)$.

5 Conclusions

The stochastic approach in previous example has shown that the basic net single premium wouldn't be enough to cover the expenses related to the chosen life insurance product. The different results are due to the more pessimistic prognosis for interest rates in $ARIMA$-model (in case of first simulation). From properties of $ARIMA$-process it follows that the simulated interest may take negative values. Another disadvantage of the stochastic approach is that the final result is a little inaccurate (confidence interval for the present value of future payments is too wide). Finally, it was shown that between two methods are only small differences, if the initial value of simulation was equal to the technical interest rate. Nevertheless, the stochastic approach may be useful for insurance companies, for example in finding an appropriate hedging strategy or in calculating the solvency capital requirement.

Acknowledgements: This research was supported by the Slovak grant VEGA No. 2/0038/12 and Grant of Comenius University No. UK/160/2013.

Bibliography

[1] Gerber, H. U. (1997). *Life Insurance Mathematics*, Third Edition, Springer-Verlag Berlin Heidelberg. ISBN 3-540-58858-3.

[2] Box, G. E. P., Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*, Revised Edition, Holden-Day. ISBN 0-8162-1104-3.

[3] Euribor Info. (2013). EURIBOR 12M - Daily rates, (June 26, 2013). URL http://www.euribor-info.com/euribor-intraday.csv

[4] Parker, G. (1998). *Stochastic Interest Rates with Actuarial Applications*, Appl. Stochastic Models & Data Anal. 14, 335-341.

[5] Potocký, R. (2012). *Modely v životnom a neživotnom poistení*, First Edition, STATIS Bratislava. ISBN 978-80-85659-71-9.

[6] R Core Team. (2012). *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/

[7] Statistical Office of the Slovak Republic (2012). *Life Tables 2011*, (January 28, 2013), URL http://portal.statistics.sk/showdoc.do?docid=33032

Statistical methodology in the scope of performance budgeting

Žiga Kotnik[*][1] **and Maja Klun**[1]

[1]*Faculty of Administration, University of Ljubljana*

Abstract

The environmental protection has become one of the main political priorities of the United Nations and the European Union. The environment is one of the areas where measurement of performance and efficiency is particularly difficult specially owing to lack of information and absence of traceability of actual effects on the environment. For this reason, environment requires its own approach that will properly evaluate environmental data and use them when planning the budget. Performance budgeting promises such solution as this approach investigates the linkage between spent public resources and planned public policy objectives. Realization of these objectives is measured through a set of indicators, attributed to each objective. The purpose of the paper is to present a brief theoretical and methodological framework of performance budgeting in the field of environmental policy and set a proper model for studying the linkage between environmental taxes, environmental expenditures and environmental impacts that are all interrelated. These will be estimated by a specifically tailored statistical model and tested in the case of the EU Member States.

Keywords: Performance budgeting, statistical methodology, the European Union, environment.
AMS subject classifications: 62M10.

1 Introduction

The current global financial and economic crisis is revealing the importance of the question of the effectiveness and efficiency of the public sector. Measuring performance in general applies to systematic efforts to assess government activity and enhance accountability for progress and outcomes in achieving results [6]. Especially among the OECD countries there is a trend for greater orientation toward effects in public sector management [3] as a consequence of international organizations' recommendations [15].

Environmental protection and pollution is becoming increasingly important issue of every society. The occurrence of negative environmental externalities that affect the society as a whole, reflect the growing public concern and need for effective control of environmental pollution [4, 10]. The article discusses an issue of effectiveness and efficiency of the public sector, namely the concept of performance budgeting in an environmental perspective. This concept helps us to ascertain the connection between allocated public funds and goals of specific policies we want to achieve by these means. Attainment of these goals is measured through a set of indicators attributed to individual goals. The purpose of the article is to present a brief theoretical and methodological framework of performance budgeting in the field of environmental policy and set a proper model for studying the linkage between interrelated groups, i.e. environmental taxes, environmental expenditures and environmental impacts.

[*] Corresponding author, e-mail: ziga.kotnik@fu.uni-lj.si

2 A brief literature review

Authors [18, 8, 16, 14, 9] agree there is no one single definition of performance budgeting. However, the review of the literature suggests what it means commonly. Experts on public budgeting agree that performance budgeting presents a promising tool for improving governance and accountability of public finance expenditures. Performance budgeting denotes the allocation of funds to achieve programmatic goals and objectives as well as some indications or measurements of work, efficiency, and/or effectiveness. In other words, the budget concept links the findings of performance measurement to budget allocations and investigates connection between spent public resources and planned public policy objectives [18, 14, 9].

Although no standard definition of performance budgeting exists Carter [5] states that it is a way to allocate resources to achieve specific objectives based on program goals and measured results. It differs from traditional approaches because it focuses on spending results rather than the money spent—on what the money buys rather than the amount that is made available.

Joyce [8] defined two utilitarian aims that performance budgeting wishes to fulfil, i.e. to improve decision-making and ameliorate service delivery. In fact, public budgeting is about making choices. To make better choices, decision-makers need qualitative and complete information and data. Performance budgeting is able to provide these through its various components or devices; e.g., the setting of goals and objectives, the prioritizing of these ends, and the measuring of performance levels (via the indices of efficiency and effectiveness) [18].

When defining suitable theoretical and methodological framework for researching connections between interconnected groups, i.e. environmental taxes, expenditures and impacts, it is important to include all three groups into the model. Performance budgeting model is accomplished only when all groups are taken under consideration.

3 Methodology framework for assessing efficiency and effectiveness of environmental policy

The usual reason for the failure of theoretical framework's concretization to measure the effectiveness and efficiency in the public sector is a lack of focus on defining goals needed to be achieved by public administration and indicators that measure achieved goals. The environment is such a case since measuring effectiveness and efficiency and allocation of resources can be very difficult because of lack of quality information, different goals between countries and difficult traceability of actual impact on the environment [16]. Therefore, when establishing performance budgeting the most important thing is good definition of the most important indicators and their target values, since in most cases indicators can be used as a basis for international comparison of comprehensive long-term social trends [1, 13].

In order to properly verify linkage between taxes, expenditures and impacts adequate simultaneous equations models (SEMs) [7, 17] will be used to evaluate effects of environmental taxes collected, environmental expenditures spent to achieve the environmental impacts, and effects of environmental impacts on environmental taxes collected after a certain period of time. According to Gujarati [7] SEMs are models where the interaction between all variables is taken into account that means multiple regression equations are estimated, one for each interdependent variable (taxes, expenditures, impacts).

Available environmental data for the European Union are combined in panels of time series from different cross-sectional units, i.e. using environmental indicators for taxes, expenditures and impacts. We will construct appropriate composite variables (composite indicators), apply different multivariate methods, e.g. factor analysis of each group of indicators taxes, expenditures and impacts, determine the time lags between three groups based on strong theory foundation and computation of correlations between the time series with lags. The OLS estimators and sensitivity tests will be used to evaluate regression functions.

We elaborate a performance budgeting model (Figure 1) and try to determine whether environmental expenditures and environmental impacts may be explained with environmental taxes. Finally, the validity of the proposed model on the basis of properly selected data will be verified.

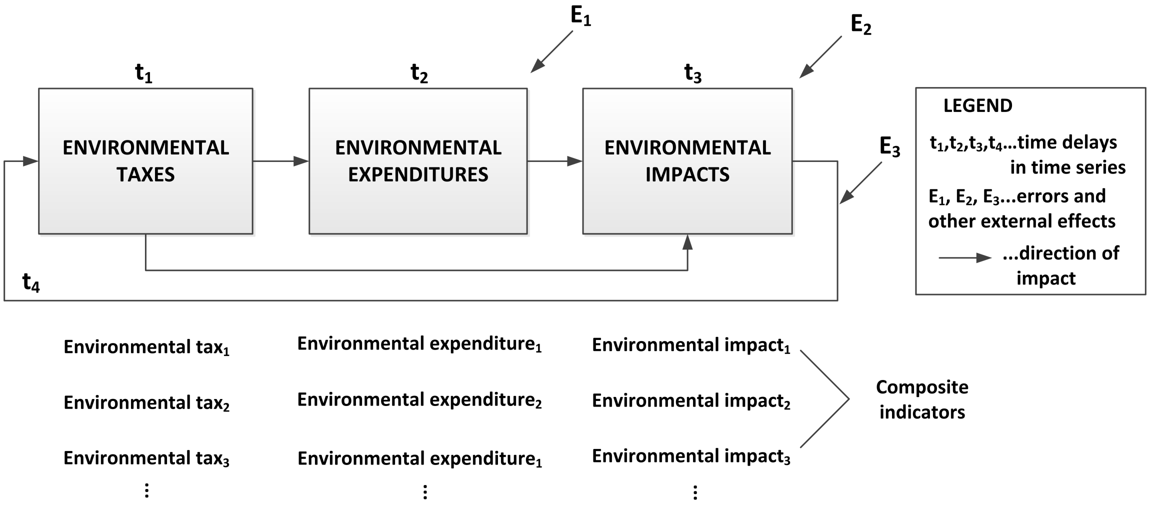


Figure 1: Performance budgeting model

The performance budgeting model is accomplished by connecting environmental taxes, environmental expenditures and environmental impact, and setting up a feedback loop between these three groups. In addition, we need to take into account the influence of environmental impact on environmental taxes after a certain period of time. Higher expenditures in one period influent environmental impacts which may affect the reduction of environmental taxes in the later period. Polluters begin to behave in an environmentally friendly manner that reduces the tax base. This presents a feedback loop in the model that should provide an additional test of the theoretical framework. Moreover, we will consider individual effects of time lags and proper steps in dealing with econometric environmental models.
Environmental model will be tested for three different environmental domains, namely:

- protection of ambient air and climate
- wastewater management
- waste management.

Furthermore, we present a set of environmental indicators, among others: environmental taxes and expenditures (Table 1), included in the proposed model that will evaluate influence of spent environmental taxes on achieving environmental impacts and the connection between the taxes, expenditures and impacts in the field of environment in general. Disposable environmental data for the European Union are available for all above-mentioned components, i.e. environmental taxes, environmental expenditures and environmental impacts. Data for selected indicators are attainable from international statistical database, i.e. the World Bank, the OECD, the UN, the European Commission (European Directorate for Taxation and Customs) and the Eurostat.
Apart from above-mentioned three groups a set of the control variables, e.g. growth rate of GDP, total investment over lagged GDP, lagged share of expenditures on public goods (as % of total government exp), proposed by several authors [11, 12, 2] will be used to inspect above-mentioned connections in our model.

Total environmental taxes (main subgroups)	**Environmental protection expenditure (main subgroups)**
Energy taxes Pollution taxes Resource taxes Transport taxes	Total investments Pollution treatment investments Pollution prevention investments Total current expenditure In-house current expenditure Fees and purchases Receipts from by-products Subsidies/transfers Revenues from sales of environmental services

Table 1: Environmental taxes and expenditures

4 Discussion

The short paper proposes a statistical performance budgeting model to be used in the field of the environment. Presented model contains all three key groups, i.e. environmental taxes, environmental expenditures and environmental impacts. Performance budgeting in the environmental perspective is realized only after all three groups are taken into consideration. In this way the model presents an upgrade to existing methodologies. Further, it includes a feedback loop between all above-mentioned groups by taking into account the effect of the environmental impact on collected environmental taxes after a certain period of time. By including such a relatively large number of environmental indicators in the model, this will substantiate the connection between environmental taxes, environmental expenditures and impacts on the environment.
The methodology devoloped here could be used in other similar research fields, e.g. macroeconomics and administration, and will help to develop other scientific fileds as well. The intertwine of statistics and social sciences will contribute to new knowledge and interdisciplinary developments in the field of public finance.

Bibliography

[1] Aristovnik, A. and Seljak, J. (2010). Performance budgeting: selected international experiences and some lessons for Slovenia. *Journal of Economics* 58(3), 271-–291.

[2] Bernauer, T. and Koubi, V. (2006). States as Providers of Public Goods: How Does Government Size Affect Environmental Quality? *Working Paper No 14*, Center for Comparative and International Studies, Zurich.

[3] Blöndal, J.R. and Curristine, T. (2004).Budgeting in Chile. *OECD Journal of Budgeting* 4(2), 7—45.

[4] Cardwell, M. (2006). The Polluter Pays Principle in European Community Law and Its Impact on United Kingdom Farmers. *Oklahoma Law Review* 58(89), 89—113.

[5] Carter, K. (1994). The performance budget revisited: A report on state budget reform. *Legislative Finance Paper No. 91* National Conference of State Legislatures, Denver, CO.

[6] Dawson, C.S. (1995). *Performance measurement and budgeting: Relearning old truths*, Legislative Commission on Government Administration, Albany, NY.

[7] Gujarati, D.N. (2003). *Basic Econometrics, Fourth Edition*, McGraw-Hill, New York.

[8] Joyce, P.G. (1999). Performance-Based Budgeting. *In Handbook of Government Budgeting*, Meyers, R.T., Editor, Jossey-Bass, San Francisco, 597—619.

[9] Joyce, P.G. (2003). *Linking Performance and Budgeting: Opportunities in the Federal Budget Process.*, IBM Center for the Business of Government, Arlington, Virginia.

[10] Klun M. (1997). Davki in okolje *In Zbornik znanstvenih razprav*, Abrahamsberg, N, Editor. Visoka upravna šola, Ljubljana, 149—162.

[11] López, R., Galinato, G.I. and Islam, A. (2011). Fiscal spending and the environment: Theory and empirics. *Journal of Environmental Economics and Management* 62(2), 180—198.

[12] López, R. and Islam, A. (2008). When Government Spendings Serves the Elites: Consequences of Economic Growth in a Context of Market Imperfections. *Working Paper 08–13* University of Maryland, College Park, Maryland.

[13] National Performance Review. (1993). *Mission Driven, Results Oriented Budgeting*, Office of the Vice President, Washington, DC.

[14] van Nispen, F.K.M. and Posseth, J.A. (2006). Performance Budgeting in the Netherlands: Beyond Arithmetic. *OECD Journal on Budgeting* 6(4), 37—62.

[15] OECD (2008). *Performance Budgeting: A user's guide*, OECD Publishing, Paris.

[16] Perrin, B. (2002). *Implementing the Vision: Addressing Challenges to Results-Focused Management and Budgeting*, OECD, Paris.

[17] Wooldridge, J.M. (2003). *Introductory Econometrics. A Modern Approach*, South-Western College Pub., Cincinnati, Ohio.

[18] Young, R.D. (2003). *Performance-Based Budget Systems*, USC Instutite for Public Services and Policy Research, Columbia, South Carolina.

Upper and lower bounds for ordered random variables

Nuria Torrado[*]

Department of Statistical Methods, University of Zaragoza

Abstract

Our aim was to examine upper and lower bounds for some reliability functions for independent but not identically distributed random variables. This problem was studied by different authors when the random variables are independent and identically distributed (see [3, 4, 7], among others).
In the article and in the presentation a short overview on the wide field of stochastic orderings is given, showing some results given by Torrado and Lillo [8] and also some of the current research the author is doing in moment. Some applications to multiple-outlier models will be briefly discussed. Multiple-outliers models are interesting due to applications in the study of the robustness of different estimators of parameters of a wide range of distributions, see e.g. Balakrishnan [1].

Keywords: Reliability Theory, Multiple-outlier Models, Ordered random variables, Stochastic Orderings.
AMS subject classifications: 60E15, 60K10, 62G30.

1 Introduction

Models of ordered random variables are widely used in statistical modelling and inference. In this section we review some models of ordered random variables, such as order statistics and spacings.
If the random variables $X_1, \ldots, X_n$ are arranged in ascending order of magnitude, then the i 'th smallest of X_i's is denoted by $X_{i:n}$. The ordered quantities

$$X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}\,, \tag{1}$$

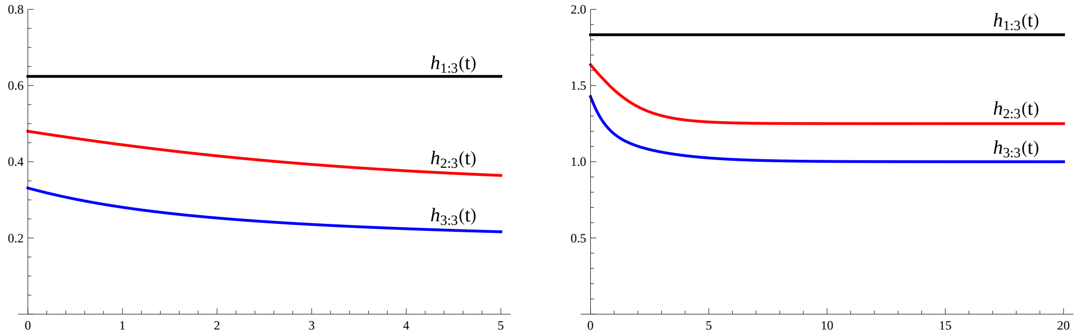
are called *order statistics* (OS), and $X_{i:n}$ is the i'th order statistic. These random variables are of great interest in many areas of statistics, specifically, there is a very interesting application of OS's in reliability theory. The $(n-k+1)$'th OS in a sample of size n represents the life length of a k-out-of-n system which is an important technical structure. It consists of n components of the same kind with independent and identically distributed life lengths. All n components start working simultaneously, and the system works, if at least k components function; i.e. the system fails, if $(n-k+1)$ or more components fail.
Another interesting random variables, which correspond to times elapsed between successive failures in the reliability context, are *simple spacings*. The i 'th simple spacing is defined as

$$D_{i:n} = X_{i:n} - X_{i-1:n}.$$

A lot of work has been done in the literature on stochastic comparisons of order statistics and spacings, see [5] for a recent review.
In the conventional modelling of these structures, the component lifetimes are supposed to be independent and identically distributed random variables. However, in many practical situations, like in reliability theory, the observations are not necessarily iid. For example, in software reliability, failure times of a software

[*] e-mail: nuria.torrado@gmail.com

program are modeled as order statistics of independent nonidentically distributed (i.ni.d) exponential random variables. According to Miller [6], these models are called EOS. It is well known that OS from heterogeneous exponential random variables are ordered with respect to various magnitude orderings, such as the hazard rate ordering. Thus, a natural question to ask is whether the spacings from exponential random variables with different scale parameters are also ordered according to some stochastic orderings, for instance with respect the hazard rate ordering. In Figure 1, we show two examples on this, when $\lambda_i = a\,b^i$, $a > 0$, $0 < b < 1$ and when $\lambda_i = a\,i^{-b}$, $a > 0$, $1 < b < \infty$, which are case 3 (geometric rates) and case 4 (power rates) in Miller [6], respectively.



(a) $\lambda_i = a\,b^i$, $a = 3$, $b = 0.4$

(b) $\lambda_i = a\,i^{-b}$, $a = 3$, $b = 1.1$

Figure 1: Hazard rate function of spacings for two EOS software reliability models

Specifically, Figure 1(a) and Figure 1(b) present the hazard rate function, $h_{i:3}(t)$, of normalized spacings from three heterogeneous exponential random variables having hazard rate $\lambda_i = a\,b^i$, $a = 3$, $b = 0.4$ and $\lambda_i = a\,i^{-b}$, $a = 3$, $b = 1.1$, respectively. As seen from these figures, the normalized spacings are ordered according to the hazard rate ordering in both cases.

The objective of this work is first to discuss some recent results on stochastic comparisons between simple spacings of heterogeneous samples and present some extensions. Specifically, we study stochastic orderings among spacings in the two sample problem, and also, show some applications to multiple-outlier models.

The article is organized as follows. In Section 2, we introduce some useful definitions which will be used in the following sections. We investigate, in Section 3.1, the likelihood ratio ordering of spacings of a sample from heterogeneous exponential random variables. Finally, we briefly discuss some applications to multiple-outlier models in Section 4.

2 Definitions of magnitude orders

In this section, we give briefly a review of stochastic orders related to the magnitude of random variables. Throughout, we shall use increasing to mean non-decreasing and decreasing to mean non-increasing.

Definition 2.1. *Let X and Y be univariate random variables with cumulative distribution functions (cdf's) F and G, respectively. We say that X is smaller than Y in the* usual stochastic order *if $\overline{F}(t) \leq \overline{G}(t)$, for all t and in this case, we write $X \leq_{st} Y$.*

Recall that the hazard rate function is a measure of the tendency to fail and it is also known as the instantaneous failure rate. The *hazard rate function*, h_X, of a random variable X at t is defined on the support of the distribution by

$$h_X(t) = \lim_{\Delta t \to 0} \frac{P\,(t < X \leq t + \Delta t \mid X > t)}{\Delta t}.$$

Definition 2.2. *Let h_Y be the hazard rate function of another random variable Y. We say that X is said to be smaller than Y in the* hazard rate order, *denoted by $X \leq_{hr} Y$, if $h_X(t) \geq h_Y(t)$, for all t, or if $\overline{G}(t)/\overline{F}(t)$ is increasing in t for which the ratio is well defined.*

Recall that the *reversed hazard rate function* r_X of a random variable X, at the point t is defined as

$$r_X(t) = \lim_{\Delta t \to 0} \frac{P\left(t - \Delta t \leq X < t \mid X < t\right)}{\Delta t}.$$

Definition 2.3. *Let r_Y be the reversed hazard rate function of another random variable Y. We say that X is smaller than Y in the* reversed hazard rate order *if $G(t)/F(t)$ is increasing in t for which the ratio is well defined, or if $r_X(t) \leq r_Y(t)$, for all t, denoted by $X \leq_{rh} Y$.*

Recall that the *Glaser's* function η_X of a random variable X (see [2]), at the point t is defined as

$$\eta_X(t) = -\frac{f'(t)}{f(t)} = -\big(\log f(t)\big)',$$

where f is the density function of X.

Definition 2.4. *Let η_Y be the Glaser's function of another random variable Y. We say that X is smaller than Y in the* likelihood ratio order *if $\eta_X(t) \geq \eta_Y(t)$ for all t, denoted by $X \leq_{lr} Y$.*

The relationships among the four first orders are illustrated in the following diagram.

$$\begin{tikzcd}
X \leq_{lr} Y \arrow[r, Rightarrow] \arrow[d, Rightarrow] & X \leq_{hr} Y \arrow[d, Rightarrow] \\
X \leq_{rh} Y \arrow[r, Rightarrow] & X \leq_{st} Y
\end{tikzcd}$$

3 Upper and lower bounds

In this section, we study conditions under which simple spacings are ordered in the likelihood ratio ordering. Here we consider a sequence of i.ni.d. random variables, $X_1, \ldots, X_n$, a set of independent exponential random variables with X_i having hazard rate λ_i, for $i = 1, \ldots, n$ and another set of independent and identically distributed exponential random variables with a common hazard rate.

3.1 Lower bounds

In the following result, we provide a lower bound for the Glaser's function of spacings $D_{1:n}, D_{2:n}, \ldots, D_{n:n}$ from the sequence of i.ni.d. random variables $X_1, \ldots, X_n$.

Theorem 3.1. *(see [8]) Let $X_1, \ldots, X_n$ be independent exponential random variables such that X_i has hazard rate λ_i, for $i = 1, \ldots, n$, and $Y_1, \ldots, Y_n$ be a random sample of size n from an exponential distribution with common hazard rate θ. If $\overline{\lambda} \leq \theta$, then*

$$C_{i:n} \leq_{\text{lr}} D_{i:n},$$

for $i = 1, \ldots, n$, where $D_{i:n}$ and $C_{i:n}$ are the i'th spacing from X_i's and Y_i's, respectively, and $\overline{\lambda} = \sum_{j=1}^{n} \lambda_j / n$.

An interesting special case, which is a consequence of the above result, is the following.

Proposition 3.1. *(see [8]) Let $X_1, \ldots, X_n$ be independent exponential random variables such that X_i has hazard rate λ_i, for $i = 1, \ldots, n$, $Y_1, \ldots, Y_n$ be a random sample of size n from an exponential distribution with common hazard rate $\theta = \max\{\lambda_1, \ldots, \lambda_n\}$. Then*

$$C_{i:n} \leq_{\text{lr}} D_{i:n},$$

for $i = 1, \ldots, n$, where $D_{i:n}$ and $C_{i:n}$ are the i 'th spacing from X_i's and Y_i's, respectively.

3.2 Upper bounds

In the following result, we provide an upper bound for the Glaser's function of spacings from the sequence of i.ni.d. random variables $X_1, \ldots, X_n$.

Theorem 3.2. *(see [8]) Let $X_1, \ldots, X_n$ be independent exponential random variables such that X_i has hazard rate λ_i, for $i = 1, \ldots, n$, and $Z_1, \ldots, Z_n$ be a random sample of size n from an exponential distribution with common hazard rate β. If $\beta \leq \frac{\sum_{j=1}^{n-i+1} \lambda_{(j)}}{n-i+1}$, then*

$$D_{i:n} \leq_{\text{lr}} H_{i:n},$$

for $i = 1, \ldots, n$, where $D_{i:n}$ and $H_{i:n}$ are the i 'th spacing from X_i's and Z_i's, respectively.

An interesting special case, which is a consequence of the above result, is the following.

Proposition 3.2. *(see [8]) Let $X_1, \ldots, X_n$ be independent exponential random variables such that X_i has hazard rate λ_i, for $i = 1, \ldots, n$, $Z_1, \ldots, Z_n$ be a random sample of size n from an exponential distribution with common hazard rate $\beta = \min\{\lambda_1, \ldots, \lambda_n\}$. Then*

$$D_{i:n} \leq_{\text{lr}} H_{i:n},$$

for $i = 1, \ldots, n$, where $D_{i:n}$ and $H_{i:n}$ are the i 'th spacing from X_i's and Z_i's, respectively.

4 Discussion

A few applications of, and complements to, the results of Section 3 are briefly described below. In this section, we consider a special case, the so called multiple-outlier exponential models. These models are defined as follows: Let $X_1, \ldots, X_n$ be a set of independent exponential random variables such that X_i has hazard rate λ for $i = 1, \ldots, p$ and X_j has hazard rate λ_* for $j = p+1, \ldots, n$. Some researchers have investigated these models of random variables, see [9] for a recent review. The simple spacings and normalized spacings from a multiple-outlier exponential model are, respectively, defined by

$$D_{i:n}(p, q; \lambda, \lambda_*) = X_{i:n} - X_{i-1:n}$$

and

$$D^*_{i:n}(p, q; \lambda, \lambda_*) = (n - i + 1) D_{i:n}(p, q; \lambda, \lambda_*),$$

for $i = 1, \ldots, n$, with $X_{0:n} \equiv 0$, $q = n - p \geq 1$ and $p \geq 1$.

Theorem 4.1. *(see [8]) Let $X_1, X_2, \ldots, X_n$ follow a multiple-outlier exponential model with parameters λ and λ_*. If $\lambda \geq \lambda_*$, $p \geq 1$ and $q \geq 1$, then*

$$D_{i:n}(p - k_2, q + k_2; \lambda, \lambda_*) \geq_{\text{lr}} D_{i:n}(p, q; \lambda, \lambda_*) \geq_{\text{lr}} D_{i:n}(p + k_1, q - k_1; \lambda, \lambda_*),$$

where $1 \leq k_1 \leq q$, $1 \leq k_2 \leq p$ and $i = 1, \ldots, n$.

Acknowledgements: The financial support of the Spanish Ministry of Education and Science under grant SEJ2007-64500 is acknowledged.

Bibliography

[1] Balakrishnan, N. (2007). Permanents, order statistics, outliers, and robustness, *Revista Matemática Complutense* 20, 7–107.

[2] Glaser, R.E. (1980). Bathtub and related failure rate characterizations, *Journal of the American Statistical Association* 75, 667–672.

[3] Kochar, S.C. and Korwar, R. (1996). Stochastic orders for spacings of heterogeneous exponential random variables, *Journal of Multivariate Analysis* 57, 69–83.

[4] Kochar, S.C. and Xu, M. (2011). Stochastic comparisons of spacings from heterogeneous samples, in M. Wells and A. Sengupta Ed., *Advances in Directional and Linear Statistics*, Festschrift Volume for J.S. Rao, Springer, p.p. 113–129.

[5] Kochar, S.C. (2012). Stochastic Comparisons of Order Statistics and Spacings: A Review, *ISRN Probability and Statistics*, doi:10.5402/2012/839473.

[6] Miller, D.R. (1986). Exponential order statistics models of software reliability growth, *IEEE Trans. Soft. Eng.* 12, 12–24.

[7] Pledger, G. and Proschan, F. (1971). Comparisons of order statistics from heterogeneous populations, with applications in reliability, in J.S. Rustagi Ed., *Optimizing Methods in Statistics*, Academic Press, New York, p.p. 89–113.

[8] Torrado, N. and Lillo, R.E. (2013). Likelihood ratio order of spacings from two heterogeneous samples. *Journal of Multivariate Analysis* 114, 338–348.

[9] Torrado, N. and Lillo, R.E. (2013). On stochastic properties of spacings with applications in multiple-outlier models, in H. Li and X. Li Ed., *Stochastic Orders in Reliability and Risk*, Springer, p.p. 103–123.

Developing statistical methodologies for anthropometry

Guillermo Vinué[*]

Department of Statistics and O.R., University of Valencia, Valencia, Spain

Abstract

Fitting Ready To Wear clothes is a basic problem for customer and apparel companies. One of the most important problems to develop new patterns and grade to all sizes is the lack of updated anthropometric data. In this context, in 2006 the Spanish Ministry of Health promoted a 3D anthropometric study of the Spanish female population. A sample of 10.415 Spanish females from 12 to 70 years old randomly selected was measured. The obtained anthropometric data constitute valuable information to understand the body shape of the population. A very important challenge is to define an optimal sizing system. A sizing system classifies a specific population into homogeneous subgroups based on some key body dimensions. Our research group has developed some clustering methodologies using some of the ideas of [9, 11], among others. In addition, the shape of the 10.415 women is described by using a set of correspondence points. In this case, we have used the statistical shape analysis [4] to divide the population into efficient sizes according to their shape. In the multivariate accommodation problem, a set of representative human models is commonly used to accommodate a certain percentage of the population. We use the archetypal analysis [2] to that end. The archetypes returned by the archetypal analysis are not necessarily observed individuals. However, in human modeling it is crucial that the archetypes are individuals of the target population. An algorithm inspired by the Partitioning Around Medoids (PAM) clustering algorithm to obtain necessarily observed individuals, which we call archetypoids, has been proposed. All the just mentioned statistical methodologies use the anthropometric data of the Spanish survey. Besides, the archetypal analysis is also applied to a well-known anthropometric database of aircraft pilots. The methodologies are also gathered together in an R package called *Anthropometry*, soon freely available.

Keywords: Anthropometric data, Clustering, Statistical shape analysis, Archetypal analysis.
AMS subject classifications: 62P30.

1 Introduction

Both apparel development process and human modelling require updated anthropometric data to develop new patterns and products adapted to the current target population. Physical measurements have been traditionally taken by using rudimentary methods like calipers, rulers or measuring tapes [8, 10]. These kinds of procedures are very easy to use and no particularly expensive. However, they present an important drawback: the set of measurements obtained and therefore the shape information, is imprecise and inaccuracy. In addition, this process always needs the interaction with real subjects with a consequent increment of time. The development of new 3D body scanner technology constitutes a step forward in the way of collecting anthropometric data. They capture the 3D shape images of the people being measured and provide accurate and reproducible anthropometric data [6, 7]. The great potential of the scanning systems for the digitization of the human body has contributed to promote new anthropometric surveys in different countries (USA, France, UK, Germany and Australia among others). In this context, the Spanish Ministry of Health promoted

[*] e-mail: Guillermo.Vinue@uv.es

a 3D anthropometric study of the Spanish female population. This survey aimed to generate anthropometric data from Spanish women for the clothing industry [1]. The scan anthropometric data are mainly used in two specific fields: in Anthropometry, the body measurements serves to define sizing systems for the apparel industry. In Ergonomics, representative human models of the population are searched; for example, to design aircraft cockpits. These data constitute valuable information to understand the body shape of the population. Therefore, rigorous statistical methodologies to deal with must be developed. The methodologies we have developed concerns clustering, the statistical shape analysis and the archetypal analysis. All of them analyze the data from the anthropometric study of the Spanish female population. In addition, the archetypal analysis is also applied to an anthropometric database of aircraft pilots. For a more efficient use of the anthropometric data, software tools must be introduced. For this reason, an R package called *Anthropometry* has been created to gather together all the mentioned methodologies. The outline of the paper is as follows: Section 2 describes the data sets used and the foundation of the statistical methods developed. Some illustrative results are given in Section 4. Finally, in Section 4 some conclusions end the paper.

2 Materials and Methods

In this section, the databases used for all the calculus is first presented. Next, each approach is shortly summarized.

2.1 Our datasets

Spanish anthropometric survey

In 2006 a 3D anthropometric study of the Spanish female population was organized by the Spanish Ministry of Health supported by the main Spanish companies in the garment industry and developed by the Biomechanics Institute of Valencia together with researchers from the statistical and nutritional areas. After finishing the study, a database was generated formed by 10.415 Spanish females from 12 to 70 years old randomly selected from the official Postcode Address File, 95 anthropometric measurements and 66 points (landmarks) on the woman's body representing its shape. Ref. [1] details the experimental design, subject recruitment, data collection and data processing. The website http://antropometria.ibv.org/ was also created as a query tool for companies (in Spanish only).

USAF survey

This data set comes from the 1967 United States Air Force (USAF) Survey. It was conducted during the first three months of 1967 under the direction of the Anthropology Branch of the Aerospace Medical Research Laboratory. A total of 202 variables (including body dimensions and background variables) were taken on 2420 Air Force personnel between 21 and 50 years of age. The data set is available from http://www.dtic.mil/dtic/.

2.2 Clustering

One of the most important issues in the apparel development process is to define a sizing system that provides a good fit to the majority of the population. A sizing system classifies a specific population into homogeneous subgroups based on some key body dimensions. Hence, clustering is the natural statistical approach to be applied. Our research group has developed some clustering methodologies using some of the ideas of [9, 11], among others.

2.3 Statistical shape analysis

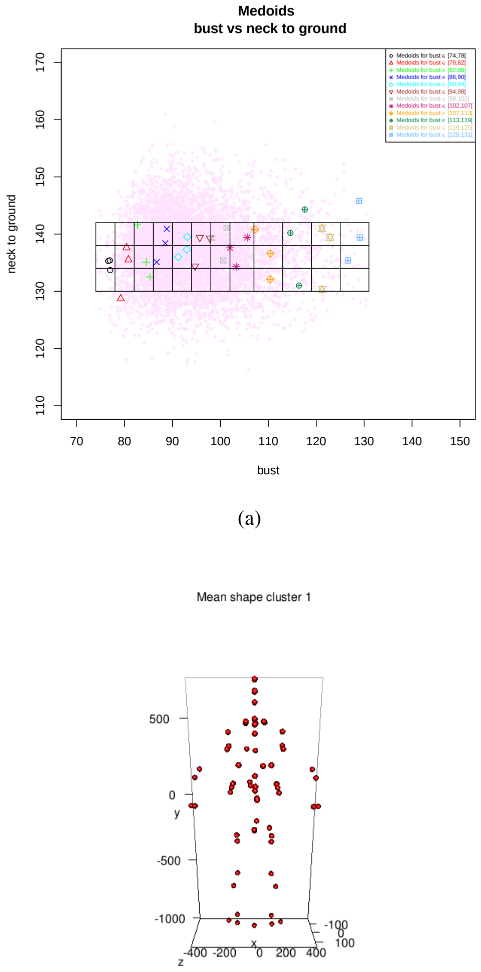
The k-means clustering algorithm has been widely used to divide the population into morphologies, see e.g. [3]. The basic foundation of k-means is that the sample mean is the value that minimizes the Euclidean distance from each point, to the centroid of the cluster to which it belongs. Two fundamental concepts of the statistical shape analysis are the Procrustes mean and the Procrustes distance. Therefore, it arises in a natural way the idea of integrating the Procrustes mean and the Procrustes distance into k-means. In this way, we can use the k-means algorithm in the shape analysis context. We propose to use the k-means algorithm to divide the population into efficient sizes according to their body shapes represented by landmarks, instead of using it by just employing a set of anthropometric variables as usual.

2.4 Archetypal Analysis

In the multivariate accommodation problem, a small group of representative cases (human models) which represents the anthropometric variability of the target population is commonly used. The appropriate selection of this small group is critical in order to accommodate a certain percentage of the population. We use the archetypal analysis [2] to that end. The archetypes returned by the archetypal analysis are a convex combination of the sampled individuals, but they are not necessarily observed individuals. However, in human modeling it is crucial that the archetypes are real people. We have developed an algorithm inspired by the PAM clustering method to obtain necessarily observed individuals (called archetypoids). We have applied this algorithm in a cockpit design problem and in an apparel design problem.

3 Results

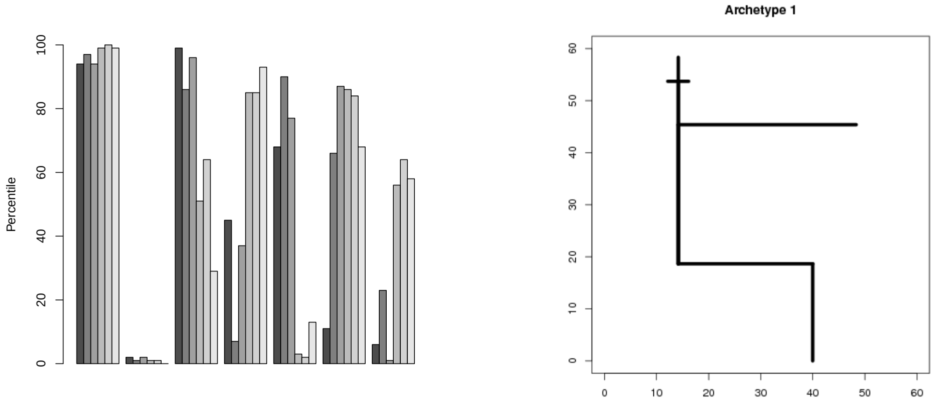
As an illustration, some graphical results provided by our methodologies are shown. Fig. 1 (left plot) shows the bust and neck to ground measurements of the women, jointly with the medoids provided by one of the clustering methodologies proposed and the prototypes defined by the European Normative to sizing system [5]. The 3D mean shape of one cluster returned by the k-means algorithm adapted to the statistical shape analysis can be also seen in Fig. 1 (right plot). Fig. 2 represents the percentiles and one skeleton plot of the archetypes obtained from the aircraft pilots database. Finally, Fig. 3 shows the 3D shape of the trunk of an archetypoid woman.



(a)

(b)

Figure 1: Medoids provided by one proposed clustering methodology jointly with the prototypes defined by the European Normative (figure a) and 3D mean shape returned by the k-means algorithm adapted to the statistical shape analysis (figure b).



(a)

(b)

Figure 2: Percentiles of the aircraft pilots archetypes and one illustrative archetype.

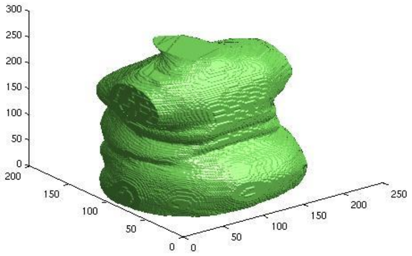


Figure 3: 3D shape of the trunk of an archetypoid woman.

4 Conclusions

Updated anthropometry data of the target population constitute valuable information to optimize sizing systems and reduce the design process cycle. Rigorous statistical methodologies including clustering, statistical shape analysis and archetypal analysis have been specially developed to deal with anthropometric data. They use the data obtained from a 3D anthropometric study of the Spanish female population and from an aircraft pilots survey. All these methodologies are gathered in an R package, soon freely available.

Acknowledgements: This research has been partially supported by grants TIN2009–14392–C02–01 and TIN2009–14392–C02–02.

Bibliography

[1] Alemany, S., González, J.C., Nácher, B., Soriano, C., Arnáiz, C. and Heras, A. (2010). Anthropometric survey of the Spanish female population aimed at the apparel industry. *Proceedings of the International Conference on 3D Body Scanning Technologies*, Lugano, Switzerland.

[2] Cutler, A. and Breiman, L. (1994). Archetypal Analysis. *Technometrics* 36 (4), 338–347.

[3] Chunga, M., Lina, H., and Wang, M.-J. J. (2007). The development of sizing systems for taiwanese elementary- and high-school students. *International Journal of Industrial Ergonomics*, 37, 707–716.

[4] Dryden, I.E. and Mardia, K.V. (1998). *Statistical Shape Analysis*, John Wiley & Sons, Chichester.

[5] European Committee for Standardization (2002). European Standard EN 13402-2: Size system of clothing. Primary and secondary dimensions.

[6] Istook, C.L. and Hwang, S-J.(2007). 3D body scanning systems with application to the apparel industry. *Journal of Fashion Marketing and Management*, 5, 120–132.

[7] Lerch, T., MacGillivray, M. and Domina, T. (2007). 3D Laser Scanning: A Model of Multidisciplinary Research. *Journal of Textile and Apparel, Technology and Management*, 5, 1–22.

[8] Lu, J.-M., and Wang, M-J.J. (2008). Automated anthropometric data collection using 3D whole body scanners. *Expert Systems with Applications* 35, 407–414.

[9] McCulloch, C., Paal, B. and Ashdown, S. (1998). An optimization approach to apparel sizing. *Journal of the Operational Research Society* 49, 492-499.

[10] Shu, C., Wuhrer, S. and Xi, P (2011). Geometric and Statistical Methods for Processing 3D Anthropometric Data. *In International Symposium on Digital Human Modeling*.

[11] van der Laan, M.J. and Pollard, K.S. (2003). A new algorithm for hybrid hierarchical clustering with visualization and the bootstrap. *Journal of Statistical Planning and Inference* 117, 275–303.

The zero area Brownian bridge

Maik Görgens[*]

Department of Mathematics, Uppsala University

Abstract

We consider the Brownian motion W on the interval $[0,1]$. The Brownian bridge B arises from the Brownian motion by pinning W_1 down to 0, i.e., the Brownian bridge arises by conditioning the Brownian motion to fulfill $W_1 = 0$. We condition the Brownian bridge further by requiring $\int_0^1 B_s ds = 0$. We call the resulting Gaussian process on $[0,1]$ zero area Brownian bridge and denote it by M. We study properties of M and give anticipative as well as non-anticipative representations.

Keywords: Brownian bridge, Conditioning, Gaussian processes, Series expansions
AMS subject classifications: 60G15, 60H10, 60J65

1 Introduction

In [2] the notion of conditioned Gaussian processes was introduced. The aim of this note is to explain what we mean by conditioned Gaussian processes, to present the main results of [2], and to apply them to a Brownian motion on $[0,1]$ conditioned to be zero at time one and having vanishing integral.

Let $(C([0,T]), \|\cdot\|_\infty)$ be the separable Banach space of continuous functions on $[0,T]$ equipped with the supremum norm $\|f\|_\infty = \sup_{0\le s\le T} |f(s)|, f \in C([0,T])$. Let $\mathcal{C}$ denote the Borel σ-algebra on $C([0,T])$. The dual space $C([0,T])^*$ of $C([0,T])$ can be identified with the space of signed finite Borel measures on $[0,T]$. For $f \in C([0,T])$ and $a \in C([0,T])^*$ we use the notation $a(f)$ and $\int f(s)\, a(ds)$ interchangeably. In particular, we use the second form if the integration only runs over a subset of $[0,T]$.

Let $X = (X_s)_{s\in[0,T]}$ be a continuous Gaussian process defined on a probability space $(\Omega, \mathfrak{A}, \mathbb{P})$. Assume $\mathbb{E}X_s = 0$ for all $s \in [0,T]$ and let $R_X : [0,T] \times [0,T] \to \mathbb{R}$ be the covariance function of X, $R_X(s,t) = \mathbb{E}X_s X_t$. A *condition* for X is an element $a \in C([0,T])^*$ and X *fulfills the condition* a if $a(X) = 0$, almost surely. Let $A \subset C([0,T])^*$ be a finite set of conditions. We define a probability measure $\mathbb{P}^{(A)}$ on $(\Omega, \mathfrak{A})$ by

$$\mathbb{P}^{(A)}(F) = \mathbb{P}(F \mid a(X) = 0 \quad \text{for all } a \in A), \quad F \in \mathfrak{A}, \tag{1}$$

and let $P_X^{(A)}$ be the induced measure on $(C([0,T]), \mathcal{C})$ of X under $\mathbb{P}^{(A)}$. Though we condition on an event of probability zero in (1), the measure $\mathbb{P}^{(A)}$ is well defined since $a(X)$ is Gaussian and we condition on $a(X) = 0$ for all $a \in A$ (see also Section 9.3 in [3]).

A continuous Gaussian process $X^{(A)} = (X_s^{(A)})_{s\in[0,T]}$ defined on a probability space $(\Omega', \mathfrak{A}', \mathbb{P}')$ is a *conditioned process of X with respect to the set of conditions A* if its induced measure $\mathbb{P}_{X^{(A)}}$ on $(C([0,T]), \mathcal{C})$ coincides with $P_X^{(A)}$. The conditioned process is thus only defined in law. The process X defined on $(\Omega, \mathfrak{A}, \mathbb{P}^{(A)})$ is a version of the conditioned Gaussian process of X (defined on $(\Omega, \mathfrak{A}, \mathbb{P})$) with respect to the conditions A.

We now introduce the zero area Brownian bridge.

[*] e-mail: maik@math.uu.se

Example (The zero area Brownian bridge)**.** *We consider the standard linear Brownian motion $W = (W_s)_{s\in[0,1]}$ and condition it by the set of conditions $A = \{a_1, a_2\} \subset C([0,1])^*$ defined by*

$$a_1(f) = f(1) \qquad \textit{and} \qquad a_2(f) = \int_0^1 f(s)ds, \qquad f \in C([0,1]).$$

We denote the conditioned process of W by the set of conditions A by $M = (M_s)_{s\in[0,1]}$, i.e., we put $M = W^{(A)}$, and call it the zero area Brownian bridge.

2 A series expansion and basic properties of the conditioned process

The following result will be crucial for our work.

Theorem 2.1 (Theorem 3.5.1 in [1])**.** *For every continuous Gaussian process $X = (X_s)_{s\in[0,T]}$ there is a separable Hilbert space H and a linear and bounded operator $u : H \to C([0,T])$ such that for every orthonormal basis $(h_i)_{i=1}^\infty \subset H$ the series*

$$\sum_{i=1}^\infty \omega_i(uh_i) \tag{2}$$

converges almost surely in $C([0,T])$ and

$$X_s = \sum_{i=1}^\infty \omega_i(uh_i)(s)$$

holds in the sense of finite-dimensional distributions, where $(\omega_i)_{i=0}^\infty$ is a sequences of independent standard normal random variables defined on a probability space $(\Omega, \mathfrak{A}, \mathbb{P})$.

We say that u is the *a*ssociated operator of X. The law of X is completely described by its associated operator. In particular, for the covariance function $R_X(s,t) = \mathbb{E}X_sX_t$ of X it holds

$$R_X(s,t) = \sum_{i=1}^\infty (uh_i)(s)(uh_i)(t) = \langle u^*\delta_s, u^*\delta_t\rangle, \tag{3}$$

where $u^* : C([0,T])^* \to H$ is the adjoint operator of u, i.e., $\langle u^*a, h\rangle = a(uh)$ for all $h \in H$ and $a \in C([0,T])^*$, δ_s is the point evaluation functional, i.e., $\delta_s(f) = f(s)$ for $f \in C([0,T])$, and $\langle\cdot,\cdot\rangle$ denotes the scalar product on H. Hence, a change of the orthonormal basis in (2) gives another process X', in general different from X, but, by (3), X and X' have the same finite-dimensional distributions.

Example (The zero area Brownian bridge – continued)**.** *The associated operator of the Brownian motion W is $u : L_2([0,1]) \to C([0,1])$ with*

$$(uh)(s) = \int_0^s h(x)\,dx$$

for $h \in L_2([0,1])$. The trigonometric basis in $L_2([0,1])$,

$$\{e_n : n \geq 0\} = \{1\} \cup \{\sqrt{2}\cos(\pi n x) : n \geq 1\}$$

yields the well known representation

$$W_s = \omega_0 s + \sqrt{2}\sum_{n=1}^\infty \omega_n \frac{\sin(\pi n s)}{\pi n}.$$

Given a finite set of conditions A we define the closed linear subspace

$$H^{(A)} = \{h \in H : a(uh) = 0 \quad \text{for all } a \in A\} \subset H$$

and call it the *reduced Hilbert space* with respect to A. Let $H_{(A)} \subset H$ be the orthogonal complement of $H^{(A)}$ (we write $H_{(A)} = H \ominus H^{(A)}$). We call $H_{(A)}$ the *detached subspace* of H with respect to A. By definition of u^*,

$$\begin{aligned} H^{(A)} &= \{h \in H : \langle u^* a, h\rangle = 0 \quad \text{for all } a \in A\} \\ &= \{h \in H : h \text{ is orthogonal to } u^* a \text{ for all } a \in A\} \subset H, \end{aligned}$$

and thus $H_{(A)}$ is spanned by the elements u^*a,

$$H_{(A)} = \operatorname{span}\{u^*a : a \in A\},$$

implying that $H_{(A)}$ is (at most) of dimension N.
Define

$$X^{(A)} = \sum_{i=1}^{\infty} \omega_i(uf_i), \tag{4}$$

where $(f_i)_{i=1}^{\infty} \subset H^{(A)}$ is an orthonormal basis in $H^{(A)}$. By (3), the law of $X^{(A)}$ is independent of the choice of the orthonormal basis in $H^{(A)}$ and since (4) differs from (2) only by a finite number of terms (given that we assume that $\{f_1, f_2, \ldots\}$ is a subset of $\{h_1, h_2, \ldots\}$) the series in (4) converges in $C([0,T])$ almost surely.

Theorem 2.2. *The process $X^{(A)}$ defined in (4) is a conditioned process of X with respect to A.*

Let $R_{X^{(A)}}$ be the covariance function of the conditioned process $X^{(A)}$ of X with respect to $A \subset C([0,T])^*$ and let $(e_i)_{i=1}^N \subset H_{(A)}$ be an orthonormal basis in the detached subspace $H_{(A)}$.

Proposition 2.1. *We have*

$$R_{X^{(A)}}(s,t) = R_X(s,t) - \sum_{i=1}^{N} (ue_i)(s)(ue_i)(t).$$

Example (The zero area Brownian bridge – continued)**.** *It holds*

$$(u^*a_1)(x) = 1 \qquad \text{and} \qquad (u^*a_2)(x) = 1 - x.$$

The detached subspace $H_{(A)}$ of $L_2([0,1])$ with respect to the set of conditions $A = \{a_1, a_2\} \subset C([0,1])^*$ is thus $H_{(A)} = \operatorname{span}\{1, 1-x\}$. An orthonormal basis in $H_{(A)}$ is $\{e_1, e_2\} = \{1, \sqrt{3}(1-2x)\}$. Hence, according to Proposition 2.1, the covariance of the zero area Brownian bridge $M = W^{(A)}$ is given by ($0 \le s,t \le 1$)

$$\begin{aligned} R_M(s,t) &= R_W(s,t) - (ue_1)(s)(ue_1)(t) - (ue_2)(s)(ue_2)(s) \\ &= \min\{s,t\} - st - 3(s - s^2)(t - t^2). \end{aligned}$$

Assume that the set $\{u^*a_i : 1 \le i \le N\} \subset H_{(A)}$ is linearly independent in H and define a matrix $B = (B_{ij})_{i,j=1}^N$ and a vector $b(X) = (b_1(X), \ldots, b_N(X))$ by $B_{ij} = a_i(ue_j)$ and $b_i = a_i(X)$.

Theorem 2.3. *The matrix B is invertible and an anticipative representation of the conditioned process $X^{(A)}$ is*

$$X^{(A)} = X - \sum_{i=1}^{N} \xi_i(X)(ue_i),$$

where $\xi(X) = (\xi_1(X), \ldots, \xi_N(X))^\tau$ is given by $\xi(X) = B^{-1}b(X)$.

Example (The zero area Brownian bridge – continued)**.** *In our example the matrix B and the vector b become*

$$B = \begin{pmatrix} 1 & 0 \\ 1/2 & 1/(2\sqrt{3}) \end{pmatrix} \quad \textit{and} \quad b = \begin{pmatrix} W_1 \\ I_1 \end{pmatrix},$$

where $I_1 = \int_0^1 W_x \, dx$. Solving the linear equation system $B\xi = b$ yields $\xi_1 = W_1$ and $\xi_2 = \sqrt{3}(2I_1 - W_1)$. By Theorem 2.3, an anticipative representation for M is

$$\begin{aligned} M_s &= W_s - W_1 s - \sqrt{3}(2I_1 - W_1)\sqrt{3}(s - s^2) \\ &= W_s - s(3s-2)W_1 - 6s(1-s)I_1. \end{aligned}$$

3 A non-anticipative representation

For $0 \leq s \leq T$, let $\mathfrak{F}_s \subset \mathcal{C}$ be the smallest σ-algebra on $C([0,T])$ such that all δ_r, $0 \leq r \leq s$, are $\mathfrak{F}_s$-$\mathfrak{B}(\mathbb{R})$-measurable, where $\mathfrak{B}(\mathbb{R})$ is the Borel σ-algebra on $\mathbb{R}$. A progressively measurable functional on $C([0,T])$ is a mapping $\beta : [0,T] \times C([0,T]) \to \mathbb{R}$ such that for each $0 \leq s \leq T$, the restriction of β to $[0,s] \times C([0,T])$ is $\mathfrak{B}([0,s]) \otimes \mathfrak{F}_s$-$\mathfrak{B}(\mathbb{R})$-measurable.

Theorem 3.1. *The probability measures $\mathbb{P}_X$ and $\mathbb{P}_{X^{(A)}}$ are equivalent on $\mathfrak{F}_s$ if and only if there exist $e'_i \in H^{(A)}$, $1 \leq i \leq N$, such that*

$$(ue'_i)(x) = (ue_i)(x), \quad 0 \leq x \leq s. \tag{5}$$

Otherwise $\mathbb{P}_X$ and $\mathbb{P}_{X^{(A)}}$ are orthogonal on $\mathfrak{F}_s$.

Assuming that there is a progressively measurable functional β on $C([0,T])$ such that X is a strong solution to a stochastic differential equation of the form

$$dX_s = \alpha dW_s + \beta(s, X)ds, \qquad X_0 = 0, \qquad 0 \leq s < T,$$

an application of Girsanov's Theorem yields the following result.

Theorem 3.2. *Assume that the supremum over all s for which (5) holds is T. Then there is a Brownian motion $W' = (W'_s)_{s\in[0,T]}$ defined on the probability space $(C([0,T]), \mathcal{C}, \mathbb{P}_{X^{(A)}})$ and a progressively measurable functional δ on $C([0,T])$ such that the conditioned process $X^{(A)}$ is a (strong) solution of the stochastic differential equation*

$$dX_s^{(A)} = \alpha dW'_s + \delta(s, X^{(A)})ds, \qquad X_0^{(A)} = 0, \qquad 0 \leq s < T. \tag{6}$$

Almost surely, for almost all $0 \leq s < T$, the drift term $\delta(s, X^{(A)})$ is given by

$$\delta(s, X^{(A)}) = \lim_{r \searrow 0} \frac{\mathbb{E}[X_{s+r}^{(A)} \mid \mathfrak{F}_s] - X_s^{(A)}}{r}.$$

If we further assume that X is a Markov process, we are able to calculate the drift term in (6) explicitly. Define Gaussian processes $I^{(A),i}$ by

$$I_s^{(A),i} = \int_0^s X_x^{(A)} \, a_i(dx), \quad 0 \leq s \leq T, \quad 0 \leq i \leq N.$$

Theorem 3.3. *The Gaussian process $(X^{(A)}, I^{(A),1}, \ldots, I^{(A),N})$ is an $(N+1)$-dimensional (in general time-inhomogeneous) Markov process.*

Define a matrix D_s and a vector d_s by

$$D_s = \begin{pmatrix} g(s) & (ue_1)(s) & \dots & (ue_N)(s) \\ \int_{s+}^{T} g(x)\, a_1(dx) & \int_{s+}^{T} (ue_1)(x)\, a_1(dx) & \dots & \int_{s+}^{T} (ue_N)(x)\, a_1(dx) \\ \vdots & \vdots & \ddots & \vdots \\ \int_{s+}^{T} g(x)\, a_N(dx) & \int_{s+}^{T} (ue_1)(x)\, a_N(dx) & \dots & \int_{s+}^{T} (ue_N)(x)\, a_N(dx) \end{pmatrix}$$

and

$$d_s = \begin{pmatrix} X_s^{(A)} \\ -I_s^{(A),1} \\ \vdots \\ -I_s^{(A),N} \end{pmatrix}.$$

Theorem 3.4. *For every $s < t$ there are $\mathfrak{F}_s^{X^{(A)}}$-measurable random variables $\xi_0, \dots, \xi_N$ such that*

$$\mathbb{E}[X_t^{(A)} | \mathfrak{F}_s^{X^{(A)}}] = \xi_0 g(t) + \sum_{i=1}^{N} \xi_i (ue_i)(t).$$

Assume that the matrix D_s is invertible. Then $\xi = (\xi_0, \dots, \xi_N)^\tau$ is given by $\xi = D_s^{-1} d_s$.

Example (The zero area Brownian bridge – continued)**.** *All assumptions of this section are fulfilled by the Brownian motion W and the conditions a_1 and a_2. Define $J_s = \int_0^s M_x\, dx$, $0 \le s \le 1$. Then $(M_s, J_s)_{s\in[0,1]}$ is a Markov process (Theorem 3.3) and M is a solution of the stochastic differential equation (Theorem 3.2)*

$$dM_s = dW_s + \delta(s, M)ds, \quad M_0 = 0, \quad 0 \le s < 1,$$

where δ is a progressively measurable functional on $C([0,1])$. By Theorem 3.4, for $0 \le s \le t < 1$, we have

$$\mathbb{E}[M_t \mid \mathfrak{F}_s^M] = \xi_0 + \xi_1 t + \xi_2 \sqrt{3}(t - t^2),$$

where $\xi = (\xi_0, \xi_1, \xi_2)$ is the solution of the system of linear equations $D_s \xi = d_s$ with $d_s = (M_s, 0, -J_s)$ and

$$D_s = \begin{pmatrix} 1 & s & \sqrt{3}(s - s^2) \\ 1 & 1 & 0 \\ 1 - s & (1 - s^2)/2 & \sqrt{3}(1 - s^2)/2 - (1 - s^3)/\sqrt{3} \end{pmatrix}.$$

Solving this system of linear equations yields

$$\xi_0 = \frac{M_s(2s^2 - s - 1) - 6J_s s}{(s-1)^3}, \qquad \xi_1 = -\frac{M_s(2s^2 - s - 1) - 6J_s s}{(s-1)^3}$$
$$\xi_2 = -\sqrt{3}\frac{M_s(s-1) - 2J_s}{(s-1)^3},$$

and thus

$$\mathbb{E}[M_t \mid \mathfrak{F}_s^M] = \frac{M_s(2s^2 - s - 1) - 6J_s s}{(s-1)^3} - t\, \frac{M_s(2s^2 - s - 1) - 6J_s s}{(s-1)^3} - 3(t - t^2)\, \frac{M_s(s-1) - 2J_s}{(s-1)^3}.$$

We have

$$\lim_{r \searrow 0} \frac{\mathbb{E}[M_{s+r} \mid \mathfrak{F}_s^M] - M_s}{r} = -\frac{4M_s}{1-s} - \frac{6J_s}{(1-s)^2}.$$

Hence, M has the stochastic differential

$$dM_s = dW_s - \frac{4M_s}{1-s} ds - \frac{6J_s}{(1-s)^2} ds, \quad M_0 = 0, \quad 0 \leq s < 1.$$

Acknowledgements: The author wishes to thank Ingemar Kaj and Svante Janson for their comments and suggestions regarding this work.

Bibliography

[1] Bogachev, Vladimir I. (1998). *Gaussian measures*, Mathematical Surveys and Monographs (62), American Mathematical Society.

[2] Görgens, M. (2013) Conditioning of Gaussian processes and a zero area Brownian bridge. Preprint, arXiv:1302.4186.

[3] Janson, S. (1997). *Gaussian Hilbert spaces*, Cambridge University Press.

[4] Mörters, P. and Peres, Y. (2010). *Brownian motion*, Cambridge University Press.

Directed random graphs and convergence to the Tracy-Widom distribution

Takis Konstantopoulos[1] **and Katja Trinajstić**[*1]

[1] *Department of Mathematics, Uppsala University, Sweden*

Abstract

We consider a directed random graph on the 2-dimensional integer lattice, placing independently, with probability p, a directed edge between any pair of distinct vertices (i_1, i_2) and (j_1, j_2), such that $i_1 \leq j_1$ and $i_2 \leq j_2$. Let $L_{n,m}$ denote the maximum length of all paths contained in an $n \times m$ rectangle. The asymptotic distribution for a centered/scaled version of $L_{n,m}$, for fixed m, as $n \to \infty$, was derived in [3]. Here, we address the problem of finding the limit when both n and m tend to infinity, so that $m \sim n^a$. We make a sequence of transformations in order to exhibit a resemblance of our model to a last passage percolation model. This requires the use of suitably defined regenerative points (called skeleton points), together with a number of pathwise and probabilistic bounds. Making use of a Komlós-Major-Tusnády coupling, as in [2], with a last-passage Brownian percolation model, we are able to prove that, for $a < 3/14$, the asymptotic distribution is the Tracy-Widom distribution.

Keywords: Random graph, Last passage percolation, Strong approximation, Tracy-Widom distribution
AMS subject classifications: 05C80, 60F17, 60K35, 06A06

1 Introduction

A directed version of a standard Erdős-Rényi random graph, sometimes called random acyclic directed graph, with vertex set $\{1, 2, \ldots, n\}$ is defined as follows: For each pair of vertices $\{i, j\}$ toss a coin with probability of heads equal to p, $0 < p < 1$, independently from pair to pair; if a head shows up then introduce an edge directed from $\min(i, j)$ to $\max(i, j)$. There is a natural extension of this graph to the whole of $\mathbb{Z}$ and, moreover, to $\mathbb{Z} \times \mathbb{Z}$ where the total order on the vertex set is replaced by the product order: $(i_1, i_2) \prec (j_1, j_2)$ if the two pairs are distinct and $i_1 \leq i_2$, $j_1 \leq j_2$. In the last model, coins are tossed only for pairs of vertices which are comparable in this partial order.

A path of length ℓ in the directed graph is a sequence $(i_0, i_1, \ldots, i_\ell)$ of vertices $i_0 \prec i_1 \prec \ldots \prec i_\ell$ such that there is an edge between any two consecutive vertices. Foss and Konstantopoulos [4] considered a random directed graph with vertex set $\mathbb{Z}$ and studied the maximum length of all paths with start and end points in the interval $[i, j]$, denoted by $L[i, j]$. They showed that there exists a deterministic constant $C = C(p)$ such that

$$\lim_{n\to\infty} L[1, n]/n = C \text{ a.s.} \tag{1}$$

A central limit theorem for $L[1, n]$ is established in [3], where is, in addition, proved a central limit theorem for the maximum length of all paths in the two-dimensional case. If $L_{n,m}$ denotes the maximum length of all paths of the graph on $\mathbb{Z} \times \mathbb{Z}$, restricted to $\{0, \ldots, n\} \times \{1, \ldots, m\}$, then there is a positive κ (depending on p and the fixed integer m), such that

$$\left(\frac{L_{[nt],m} - Cnt}{\kappa\sqrt{n}},\ t \geq 0 \right) \xrightarrow[n\to\infty]{\text{(d)}} (Z_{t,m},\ t \geq 0),$$

[*] Corresponding author, e-mail: katja.trinajstic@math.uu.se

where $Z_{\bullet,m}$ is the stochastic process defined in terms of m independent standard Brownian motions, $B^{(1)}, \dots, B^{(m)}$, via the formula

$$Z_{t,m} := \sup_{0=t_0<t_1\cdots<t_{m-1}<t_m=t} \sum_{j=1}^{m} [B^{(j)}_{t_j} - B^{(j)}_{t_{j-1}}], \quad t \geq 0.$$

One can speak of Z as a *Brownian directed percolation model*, the terminology stemming from the picture of a "weighted graph" on $\mathbb{R} \times \{1, \dots, m\}$ where the weight of a segment $[s,t] \times \{j\}$ equals the change $B^{(j)}_t - B^{(j)}_s$ of a Brownian motion. If a path from $(0,0)$ to (t,m) is defined as a union $\bigcup_{j=1}^{m} [t_{j-1}, t_j] \times \{j\}$ of such segments, then Z represents the maximum weight of all such paths.
Baryshnikov [1], answering an open question by Glynn and Whitt [5], showed that

$$Z_{1,m} \overset{\text{(d)}}{=} \lambda_m,$$

where λ_m is the largest eigenvalue of a GUE matrix of dimension m. Since $Z_{\bullet,m}$ is $1/2$-self-similar, we see that

$$Z_{t,m} \overset{\text{(d)}}{=} \sqrt{t}\lambda_m.$$

Fluctuations of λ_m around the centering sequence $2\sqrt{m}$ have been quantified by Tracy and Widom [7] who showed the existence of a limiting law, denoted by F_{TW}:

$$m^{1/6}(\lambda_m - 2\sqrt{m}) \xrightarrow[m\to\infty]{\text{(d)}} F_{\text{TW}}.$$

A natural question then, raised in [3], is whether one can obtain F_{TW} as a weak limit of $L_{n,m}$ when n and m tend to infinity simultaneously. Our paper is concerned with resolving this question. To see what scaling we can expect, rewrite the last display, for arbitrary $t > 0$, as

$$m^{1/6}(\frac{Z_{t,m}}{\sqrt{t}} - 2\sqrt{m}) \xrightarrow[m\to\infty]{\text{(d)}} F_{\text{TW}}.$$

A statement of the form $X(t,m) \xrightarrow[m\to\infty]{\text{(d)}} X$, where the distribution of $X(t,m)$ does not depend on the choice of $t > 0$, implies the statement $X(t, m(t)) \xrightarrow[t\to\infty]{\text{(d)}} X$, for any function $m(t)$ such that $m(t) \xrightarrow[t\to\infty]{} \infty$. Hence, upon setting $m = [t^a]$, we have

$$t^{a/6}\left(\frac{Z_{t,[t^a]}}{\sqrt{t}} - 2\sqrt{t^a}\right) \xrightarrow[t\to\infty]{\text{(d)}} F_{\text{TW}}. \tag{2}$$

It is reasonable to assume that an analogous limit theorem holds for a centered scaled version of the largest length $L_{n,[n^a]}$, namely that

$$n^{a/6}\left(\frac{L_{n,[n^a]} - c_1 n}{c_2\sqrt{n}} - 2\sqrt{n^a}\right) \xrightarrow[n\to\infty]{\text{(d)}} F_{\text{TW}}, \tag{3}$$

where c_1, c_2 are appropriate constants. Since we are talking about interchange of limits here, it is also reasonable to assume that (3) holds provided that a is small enough.

2 Skeleton points

In a directed random graph on $\mathbb{Z}$ exists, almost surely, a random integer sequence $\{\Gamma_r, r \in \mathbb{Z}\}$ with the property that for all r, all $i < \Gamma_r$, and all $j > \Gamma_r$, there is a path from i to Γ_r and a path from Γ_r to j. The existence of such points, referred to as *skeleton points*, is established in [3]. Since the directed Erdős-Rényi

graph is invariant under translations, so is the sequence of skeleton points, i.e., $\{\Gamma_r, r \in \mathbb{Z}\}$ has the same law as $\{n + \Gamma_r, r \in \mathbb{Z}\}$, for all $n \in \mathbb{Z}$. Moreover, it turns out that the sequence forms a stationary renewal process. If we enumerate the skeleton points according to $\cdots < \Gamma_{-1} < \Gamma_0 \leq 0 < \Gamma_1 < \cdots$, we have that $\{\Gamma_{r+1} - \Gamma_r, r \in \mathbb{Z}\}$ are independent random variables, whereas $\{\Gamma_{r+1} - \Gamma_r, r \neq 0\}$ are i.i.d. Stationarity implies that the law of the omitted difference $\Gamma_1 - \Gamma_0$ has a density which is proportional to the tail of the distribution of $\Gamma_2 - \Gamma_1$. In [3] it is shown that the distance $\Gamma_2 - \Gamma_1$ between two successive skeleton points has a finite 2nd moment. One can follow the same steps of the proof, to show that in our case, with constant edge probability p, this random variable has moments of all orders. Moreover, one can show that for some $\alpha > 0$ (the maximal such α depends on p) it holds that $Ee^{\alpha(\Gamma_2 - \Gamma_1)} < \infty$.
The rate λ of the sequence of skeleton points can be expressed as an infinite product:

$$\lambda_0 := \frac{1}{E(\Gamma_2 - \Gamma_1)} = \prod_{k=1}^{\infty} (1 - (1-p)^k)^2.$$

The most important property of the skeleton points is that if γ is a skeleton point, and if $i \leq \gamma \leq j$, then a path with length $L[i,j]$ (a maximum length path) must necessarily contain γ. This crucial property will be used several times, especially since, restriction of the graph on the interval between two successive skeleton points is independent of the restriction on the complement of the interval. Hence, for every $i < j$ the following equality holds

$$L[\Gamma_i, \Gamma_j] = L[\Gamma_i, \Gamma_{i+1}] + L[\Gamma_{i+1}, \Gamma_{i+2}] + \cdots + L[\Gamma_{j-1}, \Gamma_j],$$

i.e., $L[\Gamma_i, \Gamma_j]$ is a sum of $j - i$ i.i.d. random variables.
Consider now a directed random graph G with vertices $\mathbb{Z} \times \mathbb{Z}$. We refer to the set $\mathbb{Z} \times \{j\}$ as "line j" or "jth line", and note that the restriction of G onto $\mathbb{Z} \times \{j\}$ is a directed Erdős-Rényi random graph on $\mathbb{Z}$. We denote this restriction by $G^{(j)}$. Typically, a superscript (j) will refer to a quantity associated with this restriction. For example, for $a \leq b$,

$$L^{(j)}[a,b] := \text{ the maximum length of all paths in } G^{(j)} \text{ with vertices between } (a,j) \text{ and } (b,j).$$

Clearly, the $\{G^{(j)}, j \in \mathbb{Z}\}$ are i.i.d. random graphs, identical in distribution to the directed Erdős-Rényi random graph. Therefore, as in (1), for each $j \in \mathbb{Z}$,

$$\lim_{n \to \infty} L^{(j)}[1,n]/n = C \text{ a.s.}$$

In order to be able to resemble $L_{n,m}$ as a sum of i.i.d. random variables, we need to slightly change the definition of a skeleton point in G.

Definition 2.1 (Skeleton points in G). *A vertex (i,j) of the directed random graph G is called skeleton point if it is a skeleton point for $G^{(j)}$ (for any $i' < i < i''$, there is a path from (i',j) to (i,j) and a path from (i,j) to (i'',j)) and if there is an edge from (i,j) to $(i,j+1)$.*

Therefore, the skeleton points on line j are obtained from the skeleton point sequence of the directed Erdős-Rényi random graph $G^{(j)}$ by independent thinning with probability p. When we refer to skeleton points on line j, we shall be speaking of this thinned sequence. The elements of this sequence are denoted by

$$\cdots < \Gamma_{-1}^{(j)} < \Gamma_0^{(j)} \leq 0 < \Gamma_1^{(j)} < \Gamma_2^{(j)} < \cdots$$

and have rate

$$\lambda = \frac{1}{E(\Gamma_2^{(j)} - \Gamma_1^{(j)})} = p\lambda_0 = p \prod_{k=1}^{\infty} (1 - (1-p)^k)^2.$$

In addition, it is shown in [4] that C can also be expressed as

$$C = \frac{EL[\Gamma_1, \Gamma_2]}{E(\Gamma_2 - \Gamma_1)},$$

which is equivalent to

$$C = \frac{EL[\Gamma_1^{(j)}, \Gamma_2^{(j)}]}{E(\Gamma_2^{(j)} - \Gamma_1^{(j)})}.$$

We define the variance by

$$\sigma^2 := \mathrm{var}(L[\Gamma_1^{(j)}, \Gamma_2^{(j)}] - C(\Gamma_2^{(j)} - \Gamma_1^{(j)})).$$

For later use we need also the associated counting process of skeleton points on line j, which is defined by

$$\Phi^{(j)}(t) - \Phi^{(j)}(s) = \sum_{r\in\mathbb{Z}} \mathbf{1}(s < \Gamma_r^{(j)} \leq t), \quad s,t \in \mathbb{R}, \quad s \leq t,$$

together with the agreement that

$$\Phi^{(j)}(0) = 0.$$

In other words, using the counting process we can write the last skeleton point on the line j before the point (t, j) as $\Gamma^{(j)}_{\Phi^{(j)}(t)}$.

3 Convergence to the Tracy-Widom distribution

A different model which, *a priori*, seems to bear little resemblance to ours, is the *directed last passage percolation model* on $\mathbb{Z}^2$. We are given a collection of i.i.d. random variables indexed by elements of $\mathbb{Z}^2_+$. A path from the origin to the point $n \in \mathbb{Z}^2_+$ is a sequence of elements of $\mathbb{Z}^2_+$, starting from the origin and ending at n, such that the difference of successive members of the sequence is equal to the unit vector $(0, 1)$ or $(1, 0)$. The weight of a path is the sum of the random variables associated with its members. Let $L_{n,m}$ be the largest weight of all paths from $(0, 0)$ to (n, m). Assuming that the random variables have a finite moment of order larger than 2, Bodineau and Martin [2] approximated partial sums of i.i.d. with Brownian motions using the Komlós-Major-Tusnády (KMT) construction and showed that (3) holds for all sufficiently small positive a (the threshold depending on the order of the finite moment). Relating the ideas from the model above to the directed random graph, we are able to prove a similar result:

Theorem 3.1. *Consider the directed random graph on $\mathbb{Z} \times \mathbb{Z}$ and let $L_{n,m}$ be the maximum length of all paths between two vertices in $\{0, 1, \ldots, n\} \times \{1, 2, \ldots, m\}$. Let λ, C, σ^2 be defined as above. Then, for all $0 < a < 3/14$,*

$$n^{a/6}\left(\frac{L_{n,[n^a]} - Cn}{\sqrt{\lambda\sigma^2}\sqrt{n}} - 2\sqrt{n^a}\right) \xrightarrow[n\to\infty]{\text{(d)}} F_{TW}, \tag{4}$$

where F_{TW} is the Tracy-Widom distribution.

We begin the proof of Theorem 5 by introducing a quantity $S_{n,m}$ which resembles a last passage percolation path weight,

$$\frac{1}{\sigma} S_{n,m} = \sup_{0=t_0<t_1\cdots<t_{m-1}<t_m=t} \sum_{j=1}^{m} \sum_{k=\Phi^{(j)}(t_{j-1})+1}^{\Phi^{(j)}(t_j)} \chi_k^{(j)},$$

where

$$\chi_k^{(j)} := \frac{1}{\sigma}\{L^{(j)}[\Gamma_{k-1}^{(j)}, \Gamma_k^{(j)}] - C(\Gamma_k^{(j)} - \Gamma_{k-1}^{(j)})\}.$$

In our case $\{\chi_k^{(j)}, j \geq 1, k \geq 1\}$ represent weights in the last passage percolation on $\mathbb{Z}_+^2$. It can be proven that the maximum lenght of all paths in $\{0, 1, \ldots, n\} \times \{1, 2, \ldots, m\}$, $L_{n,m}$, is close enough to $S_{n,m}$, i.e.,

$$\frac{S_{n,[n^a]} - (L_{n,[n^a]} - Cn)}{n^{1/2-a/6}} \xrightarrow[n\to\infty]{\text{(p)}} 0.$$

Thus, taking into account (2), it remains to show that

$$\frac{\sigma^{-1} S_{n,[n^a]} - Z_{\lambda n,[n^a]}}{n^{1/2-a/6}} \xrightarrow[n\to\infty]{\text{(p)}} 0$$

to prove (4). Using the Komlós-Major-Tusnády strong approximation result [6] we can for every j jointly construct i.i.d. random variables $\{\chi_k^{(j)}, k \geq 1\}$ and $B^{(j)}$ so that they are close enough. In addition, in order to, independently of the joint construction, take care of the random indices that appear in the expresion for $S_{n,m}$ we prove and make use of a convergence rate result for the counting processes $\{\Phi^{(j)}, j \geq 1\}$.

Bibliography

[1] Baryshnikov, Y. (2001). GUEs and queues. *Probab. Theory Related Fields* 119, 256-274.

[2] Bodineau, T. and Martin, J. (2005). A universality property for last-passage percolation paths close to the axis. *Electron. Commun. Probab.* 10, 105-112.

[3] Denisov, D., Foss, S. and Konstantopoulos, T. (2012). Limit theorems for a random directed slab graph. *Ann. Appl. Probab.* 22, 702-733.

[4] Foss, S. and Konstantopoulos, T. (2003). Extended renovation theory and limit theorems for stochastic ordered graphs. *Markov Process. Related Fields* 9, n. 3, 413–468.

[5] Glynn, P.W. and Whitt, W. (1991). Departures from many queues in series. *Ann. Appl. Probab.* 1, 546-572.

[6] Komlós, J., Major, P. and Tusnády G. (1976). An approximation of partial sums of independent rv's and the sample df. II. *Z. Wahrsch. und Verw. Gebiete* 34, 33-58.

[7] Tracy, C.A. and Widom, H. (1994). Level-spacing distributions and the Airy kernel. *Comm. Math. Phys.* 159, 151-174.

Gibbs point process approximation based on Stein's method

Dominic Schuhmacher[1] **and Kaspar Stucki**[*2]

[1] *Institute for Mathematical Stochastics, University of Goettingen*
[2] *Institute of Mathematical Statistics and Actuarial Science, University of Bern*

Abstract

We develop Stein's method in the setting of Gibbs point processes. This yields upper bounds for the total variation distance between the distributions of two Gibbs point processes. Applications are provided to various well-known processes and settings from spatial statistics and statistical physics.

Keywords: Conditional intensity, pairwise interaction process, birth-death process, Stein's method, total variation distance.
AMS subject classifications: Primary 60G55; Secondary 60J75, 82B21.

1 Introduction

Gibbs processes form one of the most important classes of point processes in spatial statistics that may incorporate dependence between the points, see [8, Chapter 6]. They are furthermore, mainly in the special guise of pairwise interaction processes, one of the building blocks of modern statistical physics, see [9]. Up to the somewhat technical condition of hereditarity, see Section 2, a Gibbs (point) process on a compact metric space $\mathcal{X}$ is simply a point process whose distribution is absolutely continuous with respect to a "standard" Poisson process distribution. It is thus a natural counterpart in the point process world to a real-valued random variable that has a density with respect to some natural reference measure. A notorious difficulty with Gibbs processes is that in most interesting cases their densities can only be specified up to normalizing constants, which typically renders explicit calculations, e.g. of the total variation distance between two such processes, difficult.

Based on Stein's method we develop a theorem about upper bounds on the total variation distance between Gibbs process distributions in a very general setting. These bounds provide natural rates of convergence in many situations, and give explicit constants, which are small if one of the Gibbs processes is not too far away from a Poisson process.

This article presents a summary of the results from [11].

2 Preliminaries

Let $(\mathcal{X}, d)$ be a compact metric space, which serves as the state space for all our point processes. We equip $\mathcal{X}$ with its Borel σ-algebra $\mathcal{B} = \mathcal{B}(\mathcal{X})$. Let $\boldsymbol{\alpha} \neq 0$ be a fixed finite reference measure on $(\mathcal{X}, \mathcal{B})$. If $\mathcal{X}$ has a suitable group structure $\boldsymbol{\alpha}$ is typically chosen to be the Haar measure. If $\mathcal{X} \subset \mathbb{R}^D$, we tacitly use Lebesgue measure and write $|A| = \mathrm{Leb}^D(A)$ for $A \subset \mathcal{X}$. Denote by $(\mathfrak{N}, \mathcal{N})$ the space of finite counting measures ("point configurations") on $\mathcal{X}$ equipped with its canonical σ-algebra, see [7, Section 1.1]. A *point process* is simply a random element of $\mathfrak{N}$.

[*] Corresponding author, e-mail: kaspar.stucki@stat.unibe.ch

Throughout the paper let Π be the *Poisson process* with intensity measure $\boldsymbol{\alpha}$, i.e. for pairwise disjoint $A_1, \dots, A_n \in \mathcal{B}$ the random variables $\Pi(A_1), \dots, \Pi(A_n)$ are independent and for $1 \leq i \leq n$, $\Pi(A_i)$ has the Poisson distribution with mean $\boldsymbol{\alpha}(A_i)$. We use the definition of a Gibbs point process from spatial statistics. Call a function $u \colon \mathfrak{N} \to \mathbb{R}_+$ *hereditary* if for any $\xi, \eta \in \mathfrak{N}$ with $\xi \leq \eta$, we have that $u(\xi) = 0$ implies $u(\eta) = 0$. A point process Ξ on $\mathcal{X}$ is called a *Gibbs process* if it has a hereditary density u with respect to the Poisson process distribution $\mathscr{L}(\Pi)$, i.e. $\mathbb{E}f(\Xi) = \mathbb{E}(u(\Pi)f(\Pi))$ for all measurable $f : \mathfrak{N} \to \mathbb{R}_+$. It will be convenient to identify a Gibbs process by its conditional intensity. Let Ξ be a Gibbs process with density u. We call the function $\lambda(\cdot \,|\, \cdot) \colon \mathcal{X} \times \mathfrak{N} \to \mathbb{R}_+$,

$$\lambda(x \,|\, \xi) = \frac{u(\xi + \delta_x)}{u(\xi)} \tag{1}$$

the *conditional intensity (function)* of Ξ. For this definition we use the convention that $0/0 = 0$. It is well-known that the conditional intensity is the $\boldsymbol{\alpha} \otimes \mathscr{L}(\Xi)$-almost everywhere unique product measurable function that satisfies the *Georgii–Nguyen–Zessin equation*

$$\mathbb{E}\biggl(\int_{\mathcal{X}} h(x, \Xi - \delta_x)\, \Xi(\mathrm{d}x)\biggr) = \int_{\mathcal{X}} \mathbb{E}\bigl(h(x, \Xi)\lambda(x \,|\, \Xi)\bigr)\, \boldsymbol{\alpha}(\mathrm{d}x) \tag{2}$$

for every measurable $h \colon \mathcal{X} \times \mathfrak{N} \to \mathbb{R}_+$.

A Gibbs process Ξ on $\mathcal{X}$ is called a *pairwise interaction process (PIP)* if there exist $\beta : \mathcal{X} \to \mathbb{R}_+$ and symmetric $\varphi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ such that Ξ has the density

$$u(\xi) = c_u \prod_{1 \leq i \leq n} \beta(x_i) \prod_{1 \leq i < j \leq n} \varphi(x_i, x_j) \tag{3}$$

for any $\xi = \sum_{i=1}^n \delta_{x_i} \in \mathfrak{N}$, where c_u is the normalizing constant, which is usually not analytically computable. We then denote the distribution of Ξ by $\mathrm{PIP}(\beta, \varphi)$. The PIP is called *inhibitory* if $\varphi \leq 1$. The conditional intensity of $\Xi \sim \mathrm{PIP}(\beta, \varphi)$ is accordingly given by

$$\lambda(x \,|\, \xi) = \beta(x) \prod_{i=1}^n \varphi(x, x_i). \tag{4}$$

The following stability condition plays a crucial role for the proofs of our main results. A Gibbs process is called *locally stable* if there exists an integrable function $\psi^* \colon \mathcal{X} \to \mathbb{R}_+$ such that

$$\lambda(x \,|\, \xi) \leq \psi^*(x)$$

Local stability is satisfied for many point process distributions traditionally used in spatial statistics, see [8, p. 84 ff.]. However some processes from statistical physics, e.g. the *Lennard–Jones process*, are not locally stable.

The *total variation distance* between two point processes H and Ξ is defined as

$$d_{\mathrm{TV}}(\mathscr{L}(\mathrm{H}), \mathscr{L}(\Xi)) = \sup_{f \in \mathcal{F}_{TV}} |\mathbb{E}f(\mathrm{H}) - Ef(\Xi)|, \tag{5}$$

where $\mathcal{F}_{TV}$ is the set of measurable functions $f : \mathfrak{N} \to [0, 1]$.

3 Stein's method for Gibbs process approximation

Stein's method, originally conceived for normal approximation [12], has evolved over the last forty years to become an important tool in many areas of probability theory and for a wide range of approximating distributions. See [3] for an overview of the first thirty years of this history. A milestone in the evolution of Stein's method was the discovery in [2] that a natural Stein equation may often be set up by choosing as a right hand side the infinitesimal generator of a Markov process whose stationary distribution is the approximating distribution of interest. Many important developments stem from this so-called *generator approach* to Stein's method, and several of them concern point process approximation, such as [5], [4], [10], or [14].

In this section we develop the generator approach for Gibbs process approximation. For technical details and the proofs of the statement we refer the reader to [11]. Let $\mathrm{H} \sim \mathrm{Gibbs}(\lambda)$ and $\Xi \sim \mathrm{Gibbs}(\nu)$ be Gibbs processes. We assume that the *approximating process* H is locally stable. Define the generator

$$\mathcal{A}h(\xi) = \int_{\mathcal{X}} \big[h(\xi + \delta_x) - h(\xi)\big]\lambda(x \,|\, \xi)\, \boldsymbol{\alpha}(\mathrm{d}x) + \int_{\mathcal{X}} \big[h(\xi - \delta_x) - h(\xi)\big]\, \xi(\mathrm{d}x), \tag{6}$$

for all $h\colon \mathfrak{N} \to \mathbb{R}$ in $\mathscr{D}(\mathcal{A})$, the domain of $\mathcal{A}$. It can be shown that $\mathcal{A}$ is the generator of a spatial *birth-death process* Z with birth rate $\lambda(\cdot \mid \cdot)$ and unit per capita death rate. Denote Z_ξ for the birth death process with starting configuration $\xi \in \mathfrak{N}$, i.e. $Z_\xi(0) = \xi$. Furthermore $\mathscr{L}(\mathrm{H})$ is the unique stationary distribution, see [6, Section 4.2 and Section 4.11, Problem 5] for more details.

We set up the Stein equation as

$$f(\xi) - \mathbb{E}f(\mathrm{H}) = \mathcal{A}h(\xi), \tag{7}$$

and its solution is given by

$$h_f(\xi) = -\int_0^\infty \big[\mathbb{E}f(Z_\xi(t)) - \mathbb{E}f(\mathrm{H})\big]\, dt. \tag{8}$$

Since $Z_\xi(t)$ converges weakly to H as $t \to \infty$, the function h_f measures in some sense how long it takes for Z_ξ to "forget" the starting configuration ξ. By the Stein equation (7) we can rewrite the total variation distance between H and Ξ as

$$d_{\mathrm{TV}}(\mathscr{L}(\Xi), \mathscr{L}(\mathrm{H})) = \sup_{f \in \mathcal{F}_{TV}} |\mathbb{E}f(\Xi) - \mathbb{E}f(\mathrm{H})| = \sup_{f \in \mathcal{F}_{TV}} |\mathbb{E}\mathcal{A}h_f(\Xi)|. \tag{9}$$

Then the Georgii–Nguyen–Zessin equation (2) yields

$$\begin{aligned}
\mathbb{E}\mathcal{A}h_f(\Xi) &= \mathbb{E}\int_{\mathcal{X}} \big[h_f(\Xi + \delta_x) - h_f(\Xi)\big]\lambda(x \,|\, \Xi)\, \boldsymbol{\alpha}(\mathrm{d}x) \\
&\quad + \mathbb{E}\int_{\mathcal{X}} \big[h_f(\Xi - \delta_x) - h_f(\Xi)\big]\, \Xi(\mathrm{d}x) \\
&= \mathbb{E}\int_{\mathcal{X}} \big[h_f(\Xi + \delta_x) - h_f(\Xi)\big](\lambda(x \,|\, \Xi) - \nu(x \,|\, \Xi))\, \boldsymbol{\alpha}(\mathrm{d}x).
\end{aligned} \tag{10}$$

By constructing an explicit coupling between two birth-death processes, see [11], one can control the difference $h_f(\Xi + \delta_x) - h_f(\Xi)$ in (10), and it is then possible to obtain reasonable bounds on the last expression in (9), which yields the results of the next section.

4 Main Results

The general result is the following.

Theorem 4.1. *Let* $\Xi \sim \mathrm{Gibbs}(\nu)$ *and* $\mathrm{H} \sim \mathrm{Gibbs}(\lambda)$ *be Gibbs processes. Suppose that* H *is locally stable. Then there exists a constant* $c_1(\lambda) < \infty$ *such that*

$$d_{\mathrm{TV}}(\mathscr{L}(\Xi), \mathscr{L}(\mathrm{H})) \leq c_1(\lambda) \int_{\mathcal{X}} \mathbb{E}|\nu(x \,|\, \Xi) - \lambda(x \,|\, \Xi)|\, \boldsymbol{\alpha}(dx). \tag{11}$$

Details about the constant $c_1(\lambda)$ can be found in [11]. In particular, if H is a Poisson process, then $c_1(\lambda) = 1$. For inhibitory pairwise interaction processes (11) can be simplified.

Theorem 4.2. *Suppose that* $\Xi \sim \mathrm{PIP}(\beta, \varphi_1)$ *and* $\mathrm{H} \sim \mathrm{PIP}(\beta, \varphi_2)$ *are inhibitory. Let* $\nu(y) = \mathbb{E}(\nu(y \,|\, \Xi))$ *denote the intensity of* Ξ. *Then*

$$d_{\mathrm{TV}}(\mathscr{L}(\Xi), \mathscr{L}(\mathrm{H})) \leq c_1(\lambda) \int_{\mathcal{X}} \int_{\mathcal{X}} \beta(x)\nu(y)|\varphi_1(x,y) - \varphi_2(x,y)| \, \boldsymbol{\alpha}(\mathrm{d}x) \, \boldsymbol{\alpha}(\mathrm{d}y). \tag{12}$$

In general the intensity $\nu(y)$ is not known, but for inhibitory pairwise interaction processes one has always the crude estimate $\nu(y) \leq \beta(y)$. For stationary processes on $\mathbb{R}^d$ there are more elaborate bounds available, see [13].

A Gibbs process on a subset of $\mathbb{R}^d$ is an *area interaction process* if its conditional intensity is given by

$$\nu(x \,|\, \xi) = \tilde{\beta}\gamma^{-|\mathbb{B}(x,R/2) \setminus \bigcup_{y \in \xi} \mathbb{B}(y,R/2)|},$$

for some parameters $\tilde{\beta}, \gamma, R > 0$ and where $\mathbb{B}(x, R/2)$ denotes the open ball around x with radius $R/2$. In [1] it is shown that if $\tilde{\beta}, \gamma \to 0$ in such a way that $\tilde{\beta}\gamma^{-\alpha_D (R/2)^D} \to \beta$ (α_D denotes the volume of the unit ball in $\mathbb{R}^d$), the area interaction process converges weakly to a *Strauss hard core process*, i.e. a pairwise interaction process $\mathrm{PIP}(\beta, \varphi)$ with $\varphi(x,y) = \mathbf{1}\{|x-y| \geq R\}$. With the help of Theorem 4.1 one can show the convergence in the total variation norm and furthermore determine the exact rate of convergence, see [11].

To overcome the somehow restrictive local stability condition we use the following trick. Define

$$A_k = \{\xi \in \mathfrak{N} \colon \sup_{y \in \mathcal{X}} \xi(\mathbb{B}(y, \delta/2)) \leq k\}. \tag{13}$$

Let H_{A_k} denote the Gibbs process H conditioned on the event $\mathrm{H} \in A_k$, i.e. we require that the H has at most k points inside any ball with radius $\delta/2$. For non-inhibitory pairwise interaction processes satisfying some very standard condition from statistical physics, it can be shown that the conditioned process is then locally stable. Furthermore in [11] it is shown that

$$d_{\mathrm{TV}}(\mathscr{L}(\Xi), \mathscr{L}(\mathrm{H})) \leq d_{\mathrm{TV}}(\mathscr{L}(\Xi), \mathscr{L}(\mathrm{H}_{A_k})) + \mathbb{P}(\mathrm{H} \in A_k). \tag{14}$$

Thus one can apply Theorem 4.1 and by letting $\delta \to 0$ and or $k \to \infty$ one can make $\mathbb{P}(\mathrm{H} \in A_k)$ arbitrary small. This procedure allows e.g. the comparison of two Lennard–Jones processes.

Bibliography

[1] A. J. Baddeley and M. N. M. van Lieshout. Area-interaction point processes. *Ann. Inst. Statist. Math.*, 47(4):601–619, 1995.

[2] A. D. Barbour. Stein's method and Poisson process convergence. *J. Appl. Probab.*, 25A:175–184, 1988.

[3] A. D. Barbour and Louis H. Y. Chen, editors. *An introduction to Stein's method*, volume 4 of *Lecture Notes Series. Institute for Mathematical Sciences. National University of Singapore*. Singapore University Press, Singapore, 2005.

[4] A. D. Barbour and Marianne Månsson. Compound Poisson process approximation. *Ann. Probab.*, 30(3):1492–1537, 2002.

[5] Andrew D. Barbour and Timothy C. Brown. Stein's method and point process approximation. *Stochastic Process. Appl.*, 43(1):9–31, 1992.

[6] Stewart N. Ethier and Thomas G. Kurtz. *Markov processes*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, 1986. Characterization and convergence.

[7] Olav Kallenberg. *Random measures*. Akademie-Verlag, Berlin, fourth edition, 1986.

[8] Jesper Møller and Rasmus P. Waagepetersen. *Statistical inference and simulation for spatial point processes*, volume 100 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, Boca Raton, FL, 2004.

[9] David Ruelle. *Statistical mechanics: Rigorous results*. W. A. Benjamin, Inc., New York–Amsterdam, 1969.

[10] Dominic Schuhmacher. Distance estimates for dependent thinnings of point processes with densities. *Electron. J. Probab.*, 14(38):1080–1116, 2009.

[11] Dominic Schuhmacher and Kaspar Stucki. On bounds for Gibbs point process approximation. *Preprint*, 2012. Available at http://arxiv.org/abs/1207.3096.

[12] Charles Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), Vol. II: Probability theory*, pages 583–602, Berkeley, Calif., 1972. Univ. California Press.

[13] Kaspar Stucki and Dominic Schuhmacher. Bounds for the probability generating functional of a Gibbs point process. *Preprint*, 2012. Available at http://arxiv.org/abs/1210.4177.

[14] Aihua Xia and Fuxi Zhang. On the asymptotics of locally dependent point processes. *Stochastic Process. Appl.*, 122(9):3033–3065, 2012.

BCa-JaB method as a diagnostic tool for linear regression models

Ufuk Beyaztas[*1] **and Aylin Alin**[1]

[1] *Department of Statistics, Dokuz Eylul University*

Abstract

Jackknife-after-bootstrap (JaB) has first been proposed by [5] then used by [6] and [1] to detect influential observations in linear regression models. This method uses the percentile confidence interval to provide cut-off values for the measures. In order to improve JaB, we propose using Bias Corrected and accelerated (BCa) confidence interval introduced by [4]. In this study, the performance of BCa-JaB and conventional JaB methods are compared for DFFITS, Welsch's distance, modified Cook's distance and t-star statistics. Comparisons are based on both real world examples and simulation study. The results reveal that under considered scenarios proposed method provides more symmetric threshold values which give more accurate and reliable results.

Keywords: Bootstrap, BCa confidence interval, influential observation, regression diagnostics, robustness.
AMS subject classifications: 62F40; 62G09; 62J05; 62J20

1 Introduction

Detection and evaluation of influential observations is a critical part of data analysis in linear models. In this paper, we will work on four of the well known diagnostic measures used to detect influential observations for linear regression model: Welsch's distance, modified Cook's distance, likelihood distance and t-star (See [2] for more information about the statistics). The common idea in these measures is to identify the influential observations by comparing the results obtained from two models with and without i*th* observation.

With the increase in technology, computer intensive methods became very popular in statistics literature. Jackknife-after-bootstrap (JaB) technique is one of them. It has been developed by [5] and has been used by [6] and [1] to determine the cut-off values for various diagnostic measures in linear regression models. In JaB method, estimated cut-off values for the measures are determined from JaB distribution by using percentile quantiles, say $2.5\%th$ and $97.5\%th$ percentiles. Cut-off points obtained by this way are of the first order "*accuracy*" and are of the first order "*correctness*" where accuracy refers to the coverage errors, and correctness is a measure of the provision for a confidence interval to exact confidence interval. BCa confidence interval has been proposed by [4] to improve the performance of percentile interval. Unlike percentile cut-offs, BCa cut-offs are obtained to account for bias and skewness. Hence, they do not assume symmetric distribution. In this study, we propose replacing percentile confidence interval with BCa in JaB method. We also propose an adjustment on the BCa cut-offs to make them more robust. The performance of conventional and robust BCa-JaB methods are compared on both real world examples and simulated data sets for the diagnostic measures under consideration. The linear regression model used with influence measures throughout this study is $Y = \beta X + \varepsilon$ where Y is an $n \times 1$ column vector for response variable, X is an $n \times p$ $(p = k + 1)$ fixed full-rank design matrix, β is an $p \times 1$ vector of unknown parameters including β_0 , and ε is an $n \times 1$ error vector. Section 2 includes detailed information about the BCa-JaB methods, and numerical and simulation results will be discussed in detail in Section 3.

[*] Corresponding author, e-mail: ufuk.beyaztas@deu.edu.tr

2 Method

[5] described the idea behind the JaB method as follows: a sample of size *n* from $z_1, z_2, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n$ has the same distribution as a bootstrap sample from $z_1, z_2, \ldots, z_n$ in which none of the bootstrap values equals z_i. As described by [6] the rationale behind this approach is to generate a "null" bootstrap distribution of θ under the hypothesis that the i*th* data point is not influential. They propose that since the i*th* data point is not present in any of the resamples from which bootstrap distribution is generated, it cannot exert influence and thus the distribution generated is free from the influence of this point.This method is a very powerful method to detect unusual cases compared to traditional methods. In order to improve this method, we propose using BCa confidence interval to determine cut-off values. BCa method demonstrated by [4] is an automatic algorithm for producing highly accurate confidence limits from a bootstrap distribution ([3]). Suppose θ is a parameter of interest and $\hat{\theta}$ is the estimator from the original data, $\hat{\theta}^*$ is an estimate of $\hat{\theta}$ from the bootstrapped data.

Let $\hat{z}_{0-JaB}$ and $\hat{a}_{JaB}$ represent the bias correction and acceleration parameters for JaB distribution, respectively. $\hat{z}_{0-JaB}$ is calculated as follows:

$$\hat{z}_{0-JaB} = \Phi^{-1}\{\frac{\#\{\hat{\theta_b^*} < \hat{\theta}_{20\%trim}\}}{n^2B/e}\} \tag{1}$$

where $\hat{\theta}_{20\%trim}$ is the 20% trimmed mean of *n* diagnostic statistics calculated from the original data set and B is the number of bootstrap. By using the jackknife procedure, we set the acceleration parameter as

$$\hat{a}_{JaB} = \frac{\sum_{i=1}^{n}(\overline{\theta}_{(-i)20\%trim} - \theta_{(-i)})^3}{6\{\sum_{i=1}^{n}(\overline{\theta}_{(-i)20\%trim} - \theta_{(-i)})^2\}^{3/2}} \tag{2}$$

where $\theta_{(-i)})$ is the value of θ produced when the i*th* observation is deleted from the original sample and $\overline{\theta}_{(-i)20\%trim}$ is 20% trimmed mean of all $\theta_{(-i)})$ values. Normally, arithmetic mean is used instead of trimmed mean in equations (1) and (2). But, since mean is highly sensitive to unusual observations, we propose a robust version using 20% trimmed mean. Let $\hat{G}(c)_{JaB}$ represents the JaB distribution, then the BCa endpoint for αthquantile is computed as;

$$\hat{\theta}_{BCa}[\alpha]_{JaB} = \hat{G}^{-1}_{JaB}\Phi(\hat{z}_{0-JaB} + \frac{\hat{z}_{0-JaB} + z^{\alpha}}{1 - \hat{a}_{JaB}(\hat{z}_{0-JaB} + z^{\alpha})}) \tag{3}$$

Lower and upper limits are calculated for $\alpha/2$ and $1-\alpha/2$, respectively. Then, these BCa limits are used as cut-off points for detection of influential observations. The algorithm of BCa-JaB method can be described briefly as follows;

Step 1. Let θ_i be the diagnostic statistic that we study. The appropriate model is fitted for original data set, and the θ_i for $i = 1, 2, \ldots, n$ are calculated.

Step 2. Calculate the jackknife value of a measures of interest.

Step 3. Calculate the acceleration parameter $\hat{a}_{JaB}$ in equation (2) by using trimmed jackknife value calculated in Step 2.

Step 4. Construct B resamples with replacement from the original data set.

Step 5. For each data point within these B resamples, get a subset of the samples which do not contain that data point, so there are B/e resamples obtained for each data point. Calculate about n values of θ_i, for each of these resample, so nB/e values of θ_i are obtained. Collect all nB/e values of θ into a single vector.

Step 6. Calculate the bias correction parameter $\hat{z}_{0-JaB}$ in equation (1) using the vector created in Step 5.

Step 7. Calculate the adjusted quantiles of generated JaB distribution by using bias correction parameter calculated in Step 6. These quantiles are then compared to the original θ_i $i = 1, 2, \ldots, n$ values to flag the points as influential or not.

The steps 1-7 are repeated M times. Then, the average and standard error for the number of flagged points for all these M simulations can be calculated. It should be noted that this algorithm runs only once for the real data.

3 Numerical and simulation results

For real world examples and simulation studies, 3100 resamples were created from the original data set so that for each data point without corresponding point roughly 1000 resamples were produced. The calculations were carried out using R 2.15.2 on an Intel Core i7-2670QM 2.20 GHz PC.

As a real world example we used the soil evaporation data set which is available in ”TeachingDemos” R package to compare the performances of proposed BCa-JaB and conventional JaB methods. The set includes 46 observations and 10 explanatory variables. For this example, the normality of the resamples is deformed by points 2 and 33, and BCa-JaB adjusts the cut-off points to the right compared to conventional JaB. This adjustment is shown in our results in Table 1. Points 2, 31, 32 and 41 are seemed as influential in influential plot (to save the space, plot is not shown here). It seems that for this data set, BCa-JaB is more effective for modified Cook's distance and DFFITS.

Table 1: Regression influence diagnostics for the soil evaporation data, *n*=46, *p*=11

Method	*Welsch's dist.*	*Modified Cook's dist.*	*DFFITS*	*t-star*
Conventional JaB				
Low cut-off	-20.901	-4.083	-2.289	-2.793
High cut-off	10.512	2.323	1.302	2.089
Influential points	31	None	None	2, 8, 33, 41
BCa-JaB				
Low cut-off	-16.420	-3.362	-1.885	-2.417
	(3.38%)	(3.32%)	(3.32%)	(3.43%)
High cut-off	12.314	2.609	1.463	1.463
	(98.20%)	(98.16%)	(98.16%)	(98.22%)
Influential points	2, 31, 32	2, 31, 41	2, 31, 41	2, 31, 41

For the simulation study, we generated data under the regression model $Y = 1+2X_1+4X_2+3X_3+2X_4+\varepsilon$. The modeling scenarios are adapted in such a way that no clear influential data points were deliberately generated, and a clearly influential data point was inserted into the data set. For the model, X was generated i.i.d. $N(2, 1)$ variates and ε was generated with one of two error distributions: normal $N(0, 0.5625)$ and centered log-normal $(1.5[expN(0, 0.5625)-exp(1/2)]; skewed)$. The deliberately inserted influential point was at $(x_2 = 10, y = 10)$. For each statistic, M = 500 simulations with 3100 bootstrap resamples were performed. The average number of points flagged as influential for simulation is recorded as “Average no. of points” in the table. For deliberately inserted data point, the detection rate for simulations is recorded as “% point identified”. The standard deviations are given in brackets. The values in parenthesis below the

BCa cut-offs are the quantiles adjusted for bias and skewness. Table 2 presents the results under sample size $n = 50$.

When no deliberate influential data point is inserted into the original data set the cut-off points for both methods are almost same under both error distributions. Accordingly, average number of points flagged by both methods are nearly the same with non significant advantage of BCa-JaB. It might seem awkward to flag some points influential even though no influential point inserted deliberately. However, even if there are no deliberately inserted influential points, some influential points may occur randomly. When no points inserted deliberately, having points flagged may seem confusing and as an error. However, the issue of flagging points is reasonable in a sense that some point will have the most extreme value of the measure and the solution would be to put tests on these points (Martin, 2011 by personal contact). With inserted influential point, the results of both methods are different. The structure of the JaB distribution is distorted by the deliberately inserted influential data point. Skewness of the inserted influential observation is to the left and JaB method already adjust the cut-off points to handle this problem. However, BCa method makes further adjustment to cut-offs to correct both the skewness and bias so that we get more symmetric cut-offs. It seems that its performance is much significant under log-normal distribution.

4 Conclusion

We proposed a new approach based on robust BCa-JaB method to detect influential observations. Both real world and simulated data sets support our claim for improvement on conventional JaB method.

Bibliography

[1] Beyaztas, U. and Alin, A. (2013). Jackknife-after-bootstrap method for detection of influential observations in linear regression models. *Communications in Statistics: Simulation and Computation* 42, 1256–1267.

[2] Chatterjee, S. Hadi, A.S. (1986). Influential observations, high leverage points, and outliers in linear regression. *Statistical Science* 1, 379–416.

[3] DiCiccio, T.J. Efron, B. (1996). Bootstrap Confidence Intervals. *Statistical Science* 11, 189–228.

[4] Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the Statistical Association* 82, 171–185.

[5] Efron, B. (1992). Jackknife-after-bootstrap standard errors and influence functions. *Journal of the Royal statistical Society* 54, 83–127.

[6] Martin, M.A. and Roberts, S. (2010). Jackknife-after-bootstrap regression influence diagnostics. *Journal of Nonparametric Statistics* 22, 257–269.

Table 2: Simulation results, *n*=50, *p*=5 for all distribution of errors.

Distribution of errors	Normal				Log-normal			
Method	*Welsch's distance*	*Modified Cook's distance*	*DFFITS*	*t-star*	*Welsch's distance*	*Modified Cook's distance*	*DFFITS*	*t-star*
Influential point not present								
JaB								
Low cut-off	-5.315	-2.110	-0.703	-2.021	-4.058	-1.597	-0.532	-1.405
High cut-off	5.332	2.114	0.704	2.026	7.159	2.860	0.953	2.850
Average no. of points	2.544	2.510	2.510	2.468	2.270	2.258	2.258	1.864
(SD)	(0.908)	(0.876)	(0.876)	(0.835)	(0.803)	(0.817)	(0.817)	(0.809)
BCa-JaB								
Low cut-off	-5.323	-2.113	-0.704	-2.024	-4.036	-1.589	-0.529	-1.412
	(2.52%)	(2.52%)	(2.52%)	(2.50%)	(2.60%)	(2.59%)	(2.29%)	(2.46%)
High cut-off	5.317	2.107	0.702	2.021	7.175	2.862	0.954	2.799
	(97.44%)	(97.44%)	(97.44%)	(97.45%)	(97.51%)	(97.50%)	(97.50%)	(97.41%)
Average no. of points	2.582	2.560	2.560	2.520	2.330	2.304	2.304	1.900
(SD)	(0.892)	(0.880)	(0.880)	(0.804)	(0.840)	(0.863)	(0.863)	(0.838)
Influential point present								
JaB								
Low cut-off	-6.942	-2.693	-0.897	-2.179	-6.232	-2.411	-0.803	-1.875
High cut-off	4.856	1.934	0.644	1.863	5.509	2.198	0.732	2.149
Average no. of points	2.066	2.054	2.054	1.748	1.922	1.920	1.920	1.792
(SD)	(0.752)	(0.726)	(0.726)	(0.661)	(0.4740)	(0.747)	(0.747)	(0.624)
% point identified	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)
BCa-JaB								
Low cut-off	-5.678	-2.229	-0.743	-1.923	-4.095	-1.623	-0.541	-1.464
	(3.422%)	(3.42%)	(3.42%)	(3.49%)	(4.82%)	(4.82%)	(4.82%)	(4.79%)
High cut-off	5.406	2.143	0.7114	2.026	7.456	2.953	0.984	2.806
	(98.19%)	(98.20%)	(98.20%)	(98.26%)	(98.84%)	(98.84%)	(98.84%)	(98.83%)
Average no. of points	2.136	2.102	2.102	1.720	2.602	2.602	2.602	2.420
(SD)	(0.811)	(0.777)	(0.777)	(0.685)	(0.860)	(0.858)	(0.858)	(0.872)
% point identified	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)

Adaptive estimation in mixture models with varying mixing probabilities

Alexey V. Doronin[*]

Kyiv National Taras Shevchenko University, Kyiv, Ukraine

Abstract

Semiparametric estimation problems are considered for a model of finite mixture with mixing probabilities varying from observation to observation. We present estimators based on adaptive estimating equations, and compare them with estimators of two another types, namely the moment and quantile ones. Performance of these estimators is compared both analytically and by simulations.

Keywords: Finite mixture model, adaptive estimation, simulation, generalized estimating equation.
AMS subject classifications: 62F12, 62F35, 62G05, 62G35, 62G20.

1 Introduction

We consider a series of N subjects $O_{1;N}, ..., O_{N;N}$ belonging to M different populations (components of mixture), $N \geq 1$. Let $ind(O_{j;N})$ indicates the unknown true number of component to which the subject $O_{j:N}$ belongs. For each subject $O_{j;N}$ some numerical characteristics $\xi_{j;N} := \xi(O_{j;N}) \in \mathbb{R}$ are observed. So, we obtain a series of samples $\xi_{1;N}, ..., \xi_{N;N}$ from N observations.
We denote by $F_m(A) := P[\xi(O_{j;N}) \in A | ind(O_{j;N}) = m]$, $m = \overline{1, M}$ the CDF of $\xi(O_{j;N})$ under the condition that $O_{j;N}$ belongs to mth component of a mixture, and by $p^m_{j;N} := P[ind(O_{j;N})]$ the probability that $O_{j;N}$ belongs to mth component of a mixture (concentration of mth component). The set of concentrations $(p^m_{j;N})_{j=\overline{1,N}, m=\overline{1,M}}$ is assumed to be known. Thus, the CDF of $\xi_{j;N}$ can be expressed as $P[\xi_{j;N} \in A] = \sum\limits_{m=1}^{M} p^m_{j;N} F_m(A)$.
In what follows we assume that the CDF of the first component is parametrized with some Euclidean parameter $t \in \Theta \subset \mathbb{R}^d$ (i.e. $F_1(A) = F_1(A, t)$). We denote by $\vartheta \in \Theta$ the true value of parameter. The CDFs of the rest of the components are assumed to be fully unknown.
Moment, quantile and adaptive estimators for ϑ by the sample $\xi_{1;N}, ..., \xi_{N;N}$ are discussed in Sections 2-4. Performance of these estimates is assessed via simulations in Section 5.

2 Moment estimators

We denote by $\Gamma_N := (\langle p^k_{;N} p^l_{;N} \rangle_N)_{k,l=\overline{1,M}}$ the M-by-M Gramm's matrix for concentrations $p^m_{j;N}$ (symbol $\langle \cdot \rangle_N$ means averaging over index j), by e_m a vector from $\mathbb{R}^M$, which has a unit on the mth place, and the rest of its elements are zeros.

[*] e-mail: al_doronin@ukr.net

In [4] the set of minimax weight coefficients $a^m_{j;N} := (p_{j;N})^T(\Gamma_N)^{-1}e_m$ (under condition $\det\Gamma_N \neq 0$) is introduced, and the weighted empirical CDF

$$\hat{F}_{m;N}(x) := \frac{1}{N}\sum_{j=1}^{N} a^m_{j;N}\mathbb{I}_{\{\xi_{j;N}\leq x\}}$$

is considered as an estimate for $F_m(x)$.

In [3] improved weighted empirical CDF $\hat{F}^+_{m;N}(x) := \min\left\{1, \sup_{y\leq x}\hat{F}_{m;N}(y)\right\}$ is introduced.

Consider some measurable function $h : \mathbb{R} \to \mathbb{R}^d$.

As an estimate for $\int h(x)\hat{F}_{m;N}(dx)$ we consider the weighted moment of $h(\cdot)$

$$\hat{h}^m_N := \int h(x)\hat{F}_{m;N}(dx) = \frac{1}{N}\sum_{j=1}^{N} a^m_{j;N}h(\xi_{j;N}). \tag{1}$$

Unbiasedness, consistency and asymptotic normality for estimator defined in (1) are demonstrated in [4] under certain conditions.

We define moment estimator $\hat{\vartheta}^{simple}_N$ as a solution of moment equation

$$\int h(x)\hat{F}_{1;N}(dx) = \int h(x)F_1(dx, \hat{\vartheta}^{simple}_N). \tag{2}$$

Alternatively, we define improved moment estimator $\hat{\vartheta}^{impr}_N$ as a solution of equation

$$\int h(x)\hat{F}^+_{1;N}(dx) = \int h(x)F_1(dx, \hat{\vartheta}^{impr}_N). \tag{3}$$

Consistency and asymptotic normality for both $\hat{\vartheta}^{simple}_N$ and $\hat{\vartheta}^{impr}_N$ are demonstrated in [4] and in [3].

3 Quantile estimators

3.1 Estimators for quantiles

We denote by $Q_m(\alpha)$ the quantile for distribution $F_m(\cdot)$ of level α.

It is proposed in [4] to define an estimator $\hat{Q}_{m;N}(\alpha)$ for a quantile $Q_m(\alpha)$ as a value of a function, inversed to piece-wise linear interpolation of improved CDF $\hat{F}^+_{m;N}$ defined in section 2. Consistency and asymptotic normality of this estimator are demonstrated in [4].

3.2 Quantile estimator (for Gaussian distribution)

Let $F_1(x;t)$ be the CDF of a Gaussian distribution with the true value of a parameter $\vartheta = (\mu,\sigma)^T \in \mathbb{R}^2$.

We denote by $\gamma := Q^{\mathcal{N}(0,1)}(3/4) - Q^{\mathcal{N}(0,1)}(1/4)$ the interquartile range of standard Gaussian distribution (approximately 1.34898), and $E := \begin{pmatrix} 0 & 1 & 0 \\ -1/\gamma & 0 & 1/\gamma \end{pmatrix}$. In [2] the quantile estimator for Gaussian component is defined by

$$\hat{\vartheta}^{quant}_N := (\hat{\mu}^{quant}_N, \hat{\sigma}^{quant}_N)^T := E\cdot(\hat{Q}^1_N(1/4), \hat{Q}^1_N(1/2), \hat{Q}^1_N(3/4))^T. \tag{4}$$

Theorem 3.1. *(Theorem 1 from [2])*
Assume that

(i) $\sup_{j;N} |a^m_{j;N}| < \infty$.

(ii) The limits $\alpha^m_k := \lim_{N\to\infty} \langle p^k_{\cdot;N} (a^m_{\cdot;N})^2 \rangle_N$, $\alpha^{*;m}_{k,l} := \lim_{N\to\infty} \langle p^k_{\cdot;N} p^l_{\cdot;N} (a^m_{\cdot;N})^2 \rangle_N$ *exist,* $k, l = \overline{1, M}$.

(iii) $F_k(\cdot)$ *are continuous on* $\mathbb{R}$, $k = \overline{1, M}$.

(iv) The unbiasedness condition $\langle a^m p^k \rangle = \mathbb{I}_{k=m}$, $k = \overline{1, M}$ *holds.*

(v) Functions $F_k(\cdot)$, $k = \overline{1, M}$ *are monotone increasing in some neighborhoods* $I_1, ..., I_q$ *of points* $Q^{F_m}(\alpha_1), ..., Q^{F_m}(\alpha_q)$ *respectively.*

(vi) On $I_1, ..., I_q$ *the function* $F_m(\cdot)$ *has a continuous derivative* $f_m(\cdot)$, *and* $f_m(Q^{F_m}(\alpha_i)) \neq 0$, $i = \overline{1, q}$.

Then
1. $\sqrt{N} \cdot (\hat{Q}^m_N(\alpha_i) - Q^{F_m}(\alpha_i))_{i=\overline{1,q}} \xrightarrow{W} \mathcal{N}(\mathbb{O}_q, S)$, *where* $S = (S_{r,s})_{r,s=\overline{1,q}}$ *is q-by-q matrix with elements*

$$S_{r,s} = \frac{\sum\limits_{k=1}^{M} \alpha^m_k F_k(\min\{Q^{F_m}(\alpha_r), Q^{F_m}(\alpha_s)\}) - \sum\limits_{k,l=1}^{M} \alpha^{*;m}_{k,l} F_k(Q^{F_m}(\alpha_r)) F_l(Q^{F_m}(\alpha_s))}{f_m(Q^{F_m}(\alpha_r)) f_m(Q^{F_m}(\alpha_s))}$$

(here $\xrightarrow{W}$ *denotes the weak convergence).*
2. Assuming $(\alpha_i)_{i=\overline{1,q}} = (1/4, 1/2, 3/4)^T$, *we get* $\sqrt{N}(\hat{\vartheta}^{quant}_N - \vartheta) \xrightarrow{W} \mathcal{N}(\mathbb{O}_3, ESE^T)$.

4 Adaptive estimator (from GEE method)

In this section we present the adaptive estimator, derived from GEE method, introduced in [1].

4.1 GEE estimator

We denote by $\hat{g}^1_N(t)$ the moment estimator for value $\int g(x;t) F_1(dx;\vartheta)$ defined in (1).
The GEE (generalized estimating equation) estimator is considered in [1] as a solution of GEE $\hat{g}^1_N(\hat{\vartheta}^{GEE}_N) = \mathbb{O}_d$ with some estimating function $g(x;t) : \mathfrak{X} \times \Theta \to \mathbb{R}^d$.
Consistency and asymptotic normality of GEE estimator are demonstrated in [1].

4.2 Adaptive estimator

Adaptive estimator in [1] is constructed as a GEE estimator with the estimating function adapted by data to derive optimal dispersion matrices. For practical needs it is recommended in [1] to consider a vector of some predefined parametrized functions $u(x;t) \in \mathbb{R}^R$, and choose the estimating function as a linear combination of $u(x;t)$ (e.g. B-splines): $g(x;t) = B(t) \cdot u(x;t)$, where $B(t)$ is some d-by-R matrix. Approximate adaptive estimator is obtained from pilot estimator as one-step Newton type approximate solution of adapted estimating equation. Any $\sqrt{N}$-consistent estimator such as a moment or a quantile one can be used as the pilot estimator $\tilde{\vartheta}_N$. Thus, adaptive estimator takes form $\hat{\vartheta}^{adapt}_N := \tilde{\vartheta}_N - \hat{B}_N(\tilde{\vartheta}_N) \cdot \hat{u}_{1;N}(\tilde{\vartheta}_N)$ where $\hat{B}_N(\tilde{\vartheta}_N)$ and $\hat{u}_{1;N}(\tilde{\vartheta}_N)$ are estimations for the optimal coefficients matrix $B^*(\vartheta)$ and $\int u(x;\vartheta) F_1(dx;\vartheta)$ respectively.
Consistency and asymptotic normality of the adaptive estimator defined by (6) are demonstrated in [1].

4.3 Lower bound for dispersion matrix of adaptive estimator

Denote:

$$E_0 := (\mathbb{I}_{d\times d}, \mathbb{O}_{d\times 1}),$$

$$E_+ := \begin{pmatrix} \mathbb{I}_{(d+1)\times(d+1)} \\ \mathbb{O}_{(M-1)\times(d+1)} \end{pmatrix},$$

$$\mathbb{V} := -\int u(x;\vartheta)(f^*(x))^T \mu(dx) E_+,$$

$$f^*(x) := (\frac{\partial}{\partial\vartheta_1} f_1(x;\vartheta), ..., \frac{\partial}{\partial\vartheta_d} f_1(x;\vartheta), f_1(x), ..., f_M(x))^T,$$

$$\alpha_i := \lim_{N\to\infty} \langle (a^1_{\cdot;N})^2 p^i_{\cdot;N}\rangle_N,$$

$$\alpha^*_{i,k} := \lim_{N\to\infty} \langle (a^1_{\cdot;N})^2 p^i_{\cdot;N} p^k_{\cdot;N}\rangle_N,$$

$$r(x) := \sum_{i=1}^{M} \alpha_i f_i(x),$$

$$\mathbb{Z}_1 := \sum_{m,l=2}^{M} \alpha^*_{m,l} \int u(x;\vartheta) f_m(x)\mu(dx) \int u(x;\vartheta)^T f_l(x)\mu(dx),$$

$$\mathbb{Z}_2 := \int r(x) u(x;\vartheta) u(x;\vartheta)^T \mu(dx),$$

$$\mathbb{Z} := \mathbb{Z}_2 - \mathbb{Z}_1.$$

It is shown in [1] that the lower bound for dispersion matrix for an adaptive estimator of form $g(x;t) = B(t)\cdot u(x;t)$ is

$$S^* := S(B^*) := E_0(\mathbb{V}^T\mathbb{Z}^{-1}\mathbb{V})^{-1}E_0^T. \tag{5}$$

This lower bound is achieved on the matrix $B^*(\vartheta) := E_0(\mathbb{V}^T\mathbb{Z}^{-1}\mathbb{V})^{-1}\mathbb{V}^T\mathbb{Z}^{-1}$.

4.4 Empirical adaptive estimator

Denote

$$\alpha^*_{m,l;N} := \langle (a^1_{\cdot;N})^2 p^i_{\cdot;N} p^k_{\cdot;N}\rangle_N,$$

$$\hat{\mathbb{Z}}_{1;N}(t) := \sum_{m,l=2}^{M} \alpha^*_{m,l;N} \hat{u}^m_N(t) \hat{u}^l_N(t)^T,$$

$$\hat{\mathbb{Z}}_{2;N}(t) := \frac{1}{N}\sum_{j=1}^{N} (a^1_{j;N})^2 u(\xi_{j;N};t) u(\xi_{j;N};t)^T,$$

$$\hat{\mathbb{Z}}_N(t) := \hat{\mathbb{Z}}_{2;N}(t) - \hat{\mathbb{Z}}_{1;N}(t).$$

The estimate for $B^*(t)$ as a function of t is defined in [1] as

$\hat{B}_N(t) := E_0\left[\mathbb{V}(t)^T\hat{\mathbb{Z}}_N(t)^{-1}\mathbb{V}(t)\right]^{-1}\mathbb{V}(t)^T\hat{\mathbb{Z}}_N(t)^{-1}.$

For some $\sqrt{N}$-consistent pilot estimator $\tilde{\vartheta}_N$ the adaptive estimate takes the form

$$\hat{\vartheta}^{adapt}_N := \tilde{\vartheta}_N - \hat{B}_N(\tilde{\vartheta}_N)\cdot\hat{u}^1_N(\tilde{\vartheta}_N). \tag{6}$$

Consistency and asymptotic normality of $\hat{\vartheta}^{adapt}_N$ defined in (6) are demonstrated in [1].

Theorem 4.1. *(Theorem 4 from [1]) Assume that*

(i) $u(x;t)$ *is continuously differentiable by* t *for almost all* x *(*$\mathrm{mod}\ \mu$*).*

(ii) For some open ball B, $\vartheta \in B \subset \Theta$ *fulfills* $\int \sup\limits_{t\in B} \left\| \frac{\partial}{\partial t} u(x;t) \right\|^2 F_i(dx) < \infty, \quad i = \overline{1,M}$.

(iii) $\int \|u(x;t)\|^2 F_i(dx) < \infty$ *for all* $i = \overline{1,M}$.

(iv) $\det \Gamma \neq 0$.

(v) Limits α_i *and* $\alpha^*_{i,m}$ *exist.*

(vi) $\tilde{\vartheta}_N$ *is a consistent estimate for* ϑ.

(vii) $\mathbb{V}$ *is a matrix of full rank.*

(viii) $\tilde{\vartheta}_N$ *is a* $\sqrt{N}$*-consistent estimate of* ϑ.

Then $\sqrt{N}(\hat{\vartheta}^{adapt}_N - \vartheta) \xrightarrow{W} \mathcal{N}(\mathbb{O}_d, S^*)$ *with* S^* *defined in (5).*

5 Numerical examples

We assessed performance of the following estimators by simulations.

A. Simple estimate $\hat{\vartheta}^{simple}_N$ defined by (2) with $h(x) := (x, x^2)^T$.

B. Improved estimate $\hat{\vartheta}^{impr}_N$ defined by (3) with $h(x) := (x, x^2)^T$-

C. Quantile estimate $\hat{\vartheta}^{quant}_N$ defined by (4).

D. Adaptive estimate $\hat{\vartheta}^{adapt}_N$ defined by (6) with $\hat{\vartheta}^{impr}_N$ as a pilot.

E. Adaptive estimate $\hat{\vartheta}^{adapt}_N$ defined by (6) with $\hat{\vartheta}^{quant}_N$ as a pilot.

Experiments were conducted on two types of two-component mixture from Gaussian distributions with the following parameters:

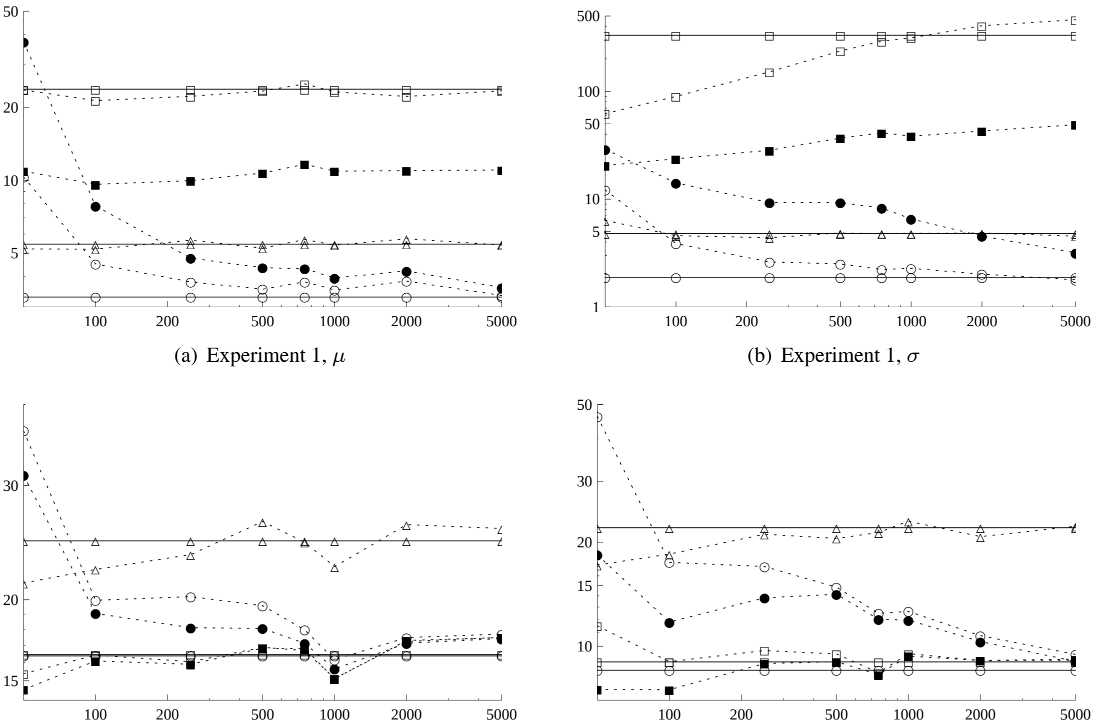
Experiment 1. Component 1: $\mu = -3$, $\sigma = 1$; component 2: $\mu = 3$, $\sigma = 2$.

Experiment 2. Component 1: $\mu = 0$, $\sigma = 2$; component 2: $\mu = 1$, $\sigma = 2$.

The estimates were calculated for different sizes of a sample (value N): 50, 100, 250, 500, 750, 1000, 2000, 5000. The dispersion of constructed estimates was calculated from 1000 samples (for each value of N). The set of concentration was uniform: $p^1_{j;N} := \frac{j}{N}$, $p^2_{j;N} := 1 - p^1_{j;N}$, $j = \overline{1,N}$.

For adaptive estimate as a vector $u(x;t)$ is taken a vector from 8 functions. First 5 of them are cubic B-splines with support $(t_1 - 4t_2, t_1 + 4t_2)$ and uniform subdivision of this support into 8 intervals. The last 3 functions: 1, $(x - t_1)/t_2$, $(x - t_1)^2/t_2^2$.

The results of simulation are presented in Figure 1.



(a) Experiment 1, μ

(b) Experiment 1, σ

(c) Experiment 2, μ

(d) Experiment 2, σ

Figure 1: The variance of estimates multiplied by the number of observations (N): □ – simple estimates, ■ – improved estimates, △ – quantile estimates, ● and ○ – adaptive estimates with improved and quantile as pilot ones respectively. Asymptotic values are presented by dotted lines.

So, in our experiments the adaptive estimators outperformed the other ones in almost all cases for sample sizes larger then 100.

Bibliography

[1] Maiboroda, R.E., Sugakova, O.V. and Doronin, A.V. (2013). Generalized estimating equations for mixtures with varying concentrations. *The Canadian Journal of Statistics*. 41, 2, 217–236.

[2] Doronin, A.V. (2012). Robust Estimates for Mixtures with Gaussian Component. *Bulletin of Taras Shevchenko National University of Kyiv. Series: Physics & Mathematics* (in Ukrainian). 1, 18–23.

[3] Maiboroda, R.E. and Kubaichuk, O.O. (2005). Improved estimators for moments constructed from observations of a mixture. *Theory of Probability and Mathematical Statistics*. 70, 83–92.

[4] Maiboroda, R.E. and Sugakova, O.V. (2008). *Estimation and classification by observations from mixtures*. Kyiv University Publishers, Kyiv (in Ukrainian).

Drift parameter estimation in models with fractional Brownian motion by discrete observations

Kostiantyn Ralchenko[*]

Taras Shevchenko National University of Kyiv

Abstract

We study a problem of estimating of unknown drift parameter in stochastic differential equation driven by fractional Brownian motion. Using Girsanov theorem, we can find the form of maximum likelihood ratio, and, moreover, represent it via the observable process. The form of this representation is rather complicated. In the simplest case it can be simplified, we can discretize it and establish the convergence a.s. of the discretized version of maximum likelihood ratio to the true value of parameter in the framework of "high frequency data".

Keywords: fractional Brownian motion, stochastic differential equation, parameter estimation, strong consistency, discretization.
AMS subject classifications: Primary: 60G22, 62F10. Secondary: 60H10, 62F12.

1 The explicit form of the likelihood ratio

Let $B^H = \left\{B_t^H, t \geq 0\right\}$ be a fractional Brownian motion with Hurst index $H \in (1/2, 1)$, defined on the probability space $(\Omega, \mathcal{F}, \mathsf{P})$. Denote by $(\mathcal{F}_t)_{t\geq 0}$ — the filtration generated by B^H. We study the problem of estimating of an unknown drift parameter from [1]. Consider the stochastic differential equation driven by fractional Brownian motion B^H:

$$\begin{aligned} dX_t &= \theta a(t, X_t)dt + b(t, X_t)dB_t^H, \quad 0 \leq t \leq T, \quad T > 0, \\ X\Big|_{t=0} &= X_0 \in \mathbb{R}. \end{aligned} \tag{1}$$

Here $\theta \in \mathbb{R}$ is the unknown parameter to be estimated.
Suppose that the following assumptions hold:

(I) there exist positive constants C_1, C_2 such that for all $t \in [0, T]$, $x, y \in \mathbb{R}$

$$|a(t,x) - a(t,y)| + |b(t,x) - b(t,y)| \leq C_1|x-y|,$$
$$|a(t,x)| + |b(t,x)| \leq C_2(1+|x|);$$

(II) there exist constants $C_3 > 0$ and $\rho \in \left(\frac{1}{H} - 1, 1\right)$ such that for all $t \in [0,T]$, $x, y \in \mathbb{R}$

$$|b'_x(t,x) - b'_y(t,y)| \leq C_3|x-y|^\rho;$$

(III) there exist constants $C_4 > 0$ and $\gamma \in (1-H, 1)$ such that for all $t, s \in [0,T]$, $x \in \mathbb{R}$

$$|b(t,x) - b(s,x)| + |b'_x(t,x) - b'_x(s,x)| \leq C_4|t-s|^\gamma.$$

[*] e-mail: k.ralchenko@gmail.com

According to [3, Theorem 2.1], under the conditions (I)–(III) there exists a unique solution X of the stochastic equation (1).
In addition, suppose that the following conditions hold:

(IV) $b(t,x) \neq 0$;

(V) $a \in C([0,\infty) \times \mathbb{R})$.

Denote $\alpha = H - \frac{1}{2}$, $\widetilde{\alpha} = (1-2\alpha)^{-1}$, $C_H = \left(\frac{\Gamma(2-2\alpha)}{2H\Gamma(1-\alpha)^3\Gamma(\alpha+1)}\right)^{\frac{1}{2}}$, $l_H(t,s) = C_H s^{-\alpha}(t-s)^{-\alpha} I_{\{0<s<t\}}$, $\psi(t,x) = \frac{a(t,x)}{b(t,x)}$, $\varphi(t) = \psi(t,X_t)$, $I(t) = \int_0^t l_H(t,s)\varphi(s)ds$. Under the conditions (I), (III), (IV), (V) $\varphi(t), t \in [0,T]$ is a continuous process with probability 1. Hence, it is Lebesgue integrable and for each $t \in [0,T]$ there exists an integral $\int_0^t l_H(t,s)\varphi(s)ds$.
Consider the new process $\widehat{B}_t^H := B_t^H + \theta\int_0^t \varphi(s)ds$. Suppose that the following assumptions hold.

(VI) there exist a function δ that for all $t \in [0,T]$ a. s. belongs to $L_1[0,t]$ and satisfy the equation

$$\theta\int_0^t l_H(t,s)\varphi(s)ds = (\widetilde{\alpha})^{-1/2}\int_0^t \delta_s ds;$$

(VII) $\mathsf{E}\int_0^t s^{2\alpha}\delta_s^2 ds < \infty, t \in [0,T]$;

(VIII) $\mathsf{E}\exp\left\{L_t - \frac{1}{2}\langle L\rangle_t\right\} = 1$, where $L_t = \int_0^t s^{\alpha}\delta_s d\widehat{B}_s$, and $\widehat{B}$ is Wiener process with respect to the probability measure $\mathsf{P}_0(t)$ corresponding to the zero drift such that

$$\int_0^t l_H(t,s)\, d\widehat{B}_s^H = \widetilde{\alpha}^{-1/2}\int_0^t s^{-\alpha}\, d\widehat{B}_s.$$

(The existence of this Wiener process follows from the representation of fractional Brownian motion via Wiener process on a finite interval introduced in [2].)

Then the likelihood ratio $\frac{d\mathsf{P}_\theta(t)}{d\mathsf{P}_0(t)}$ for the probability measure $\mathsf{P}_\theta(t)$ corresponding to our model and the probability measure $\mathsf{P}_0(t)$ corresponding to the model with zero drift is equal to

$$\frac{d\mathsf{P}_\theta(t)}{d\mathsf{P}_0(t)} = \exp\left\{L_t - \frac{1}{2}\langle L\rangle_t\right\}. \qquad (2)$$

Moreover L_t is a square-integrable martingale.
Note that the likelihood in (2) is not the likelihood of the observed process, but the likelihood of the unobserved driving noise. The following theorem allow us to present the likelihood ratio (2) as a function of the observed process X_t.
Assume that

(IX) $a, b, \psi \in C^{1,1}([0,\infty) \times \mathbb{R})$.

Theorem 1.1. *Suppose that the assumptions* (I)–(IV)*,* (VI)*,* (IX) *hold. Then*

$$\begin{aligned} L_t = {} & C_H \mathrm{B}(1-\alpha, 1-\alpha)\theta\psi(0,0)J_t \\ & + \theta\int_0^t \Big[\widetilde{\alpha}s^{2\alpha-1}\int_0^s l_H(s,u)\left(\psi'_t(u,X_u) + \theta\psi'_x(u,X_u)a(u,X_u)(s-u)\right)du \\ & - \int_0^s\int_0^u \Big(\widetilde{\alpha}\alpha s^{2\alpha}u^{-1} l_H(s,u)\psi'_t(v,X_v) \end{aligned}$$

$$+\theta\psi_x'(v,X_v)a(v,X_v)C_H\left(\widetilde{\alpha}\alpha s^{2\alpha}u^{-1-\alpha}(s-u)^{-\alpha}+u^{2\alpha-2}v^{1-\alpha}(u-v)^{-\alpha}\right)\Big)dvdu$$
$$+C_H\int_0^s u^{2\alpha-2}\int_0^u v^{1-\alpha}(u-v)^{-\alpha}\psi_x'(v,X_v)dX_vdu$$
$$+\widetilde{\alpha}C_Hs^{2\alpha-1}\int_0^s u^{1-\alpha}(s-u)^{-\alpha}\psi_x'(u,X_u)dX_u\Bigg]dJ_s,$$

where $J_t=\int_0^t l_H(t,s)b^{-1}(s,X_s)dX_s$.

Example 1.1. *Let* $a(t,x)=b(t,x)$. *Then* $\psi(t,x)=1$, $\psi_t'(t,x)=\psi_x'(t,x)=0$. *Therefore*

$$L_t=C_H\mathrm{B}(1-\alpha,1-\alpha)\theta J_t=\mathrm{B}(1-\alpha,1-\alpha)C_H\theta\int_0^t l_H(t,s)a^{-1}(s,X_s)dX_s$$

and the maximum likelihood estimator for θ *is*

$$\widehat{\theta}_t=\frac{\int_0^t s^{-\alpha}(t-s)^{-\alpha}a^{-1}(s,X_s)dX_s}{\mathrm{B}(1-\alpha,1-\alpha)t^{1-2\alpha}}.$$

2 Strong consistency of estimators by discrete observations

Let X_t be a solution of the equation

$$dX_t=\theta a(X_t)dt+b(X_t)dB_t^H, \tag{3}$$

where the coefficients a and b satisfy the following condition: there exist constants $\mu\in(0,1]$, $K>0$, $L>0$, $M>0$ and for every $N>0$ there exists $R_N>0$ such that the following assumptions hold:

(A) $|a(x)|+|b(x)|\le K$ for all $x,y\in\mathbb{R}$,

(B) $|a(x)-a(y)|+|b(x)-b(y)|\le L|x-y|$ for all $x,y\in\mathbb{R}$,

(C) $|b'(x)-b'(y)|\le R_N|x-y|^\mu$ for all $|x|\le N$, $|y|\le N$,

(D) $|a(x)|\ge M$, $|b(x)|\ge M$ for all $x\in\mathbb{R}$.

Suppose that we observe the values $X_{\frac{k}{2^n}}$, $k=0,1,\dots,2^{2n}$.

Theorem 2.1. *Let*

$$\hat{\theta}_n^{(1)}=\frac{\sum_{k=1}^{2^{2n}}\left(\frac{k}{2^n}\right)^{-\alpha}\left(2^n-\frac{k}{2^n}\right)^{-\alpha}b^{-1}\left(X_{\frac{k-1}{2^n}}\right)\left(X_{\frac{k}{2^n}}-X_{\frac{k-1}{2^n}}\right)}{\sum_{k=1}^{2^{2n}}\left(\frac{k}{2^n}\right)^{-\alpha}\left(2^n-\frac{k}{2^n}\right)^{-\alpha}b^{-1}\left(X_{\frac{k-1}{2^n}}\right)a\left(X_{\frac{k-1}{2^n}}\right)\frac{1}{2^n}}.$$

Then with probability one $\hat{\theta}_n^{(1)}\to\theta$, $n\to\infty$.

Remark 2.1. *When* $a(x)=b(x)$ *(for example, in the linear model) the estimator* $\hat{\theta}_n^{(1)}$ *is a discretized version of the maximum-likelihood estimator.*

Theorem 2.2. *Let*

$$\hat{\theta}_n^{(2)}=\frac{\sum_{k=1}^{2^{2n}}b^{-1}\left(X_{\frac{k-1}{2^n}}\right)\left(X_{\frac{k}{2^n}}-X_{\frac{k-1}{2^n}}\right)}{\frac{1}{2^n}\sum_{k=1}^{2^{2n}}b^{-1}\left(X_{\frac{k-1}{2^n}}\right)a\left(X_{\frac{k-1}{2^n}}\right)}.$$

Then with probability one $\hat{\theta}_n^{(2)}\to\theta$, $n\to\infty$.

Bibliography

[1] Mishura, Y. (2008). *Stochastic Calculus for Fractional Brownian Motion and Related Processes*, Lecture Notes in Mathematics, vol. 1929, Springer, Berlin.

[2] Norros, I., Valkeila, E. and Virtamo, J. (1999). An elementary approach to a Girsanov formula and other analytical results on fractional Brownian motions. *Bernoulli.* 5(4), 571–587.

[3] Nualart, D. and Răşcanu, A. (2002). Differential equations driven by fractional Brownian motion. *Collect. Math.* 53, 55–81.

Optimal designs for discriminating between functional linear models

Verity Fisher*[*] and David Woods

University of Southampton, UK

Abstract

Improvements in online measuring and monitoring have facilitated an increase in the number of observations that can be taken on each experimental unit in industrial and scientific experiments. Examples are from diverse areas such as biometry, chemistry, psychology and climatology. It can often be assumed that the application of a treatment to each unit generates a smooth functional response. A semi-parametric model is often used for the response when we are interested in how changes to the levels of the controllable factors influence these functions. Relatively simple polynomial models are chosen to describe the treatement effects. In this paper, we present methods for the design of experiments with functional data when the aim is to discriminate between linear models for the treatment effect. We develop an extension of the T-optimality criterion to functional data for discriminating between two competing models. The methodology is motivated by an example from Tribology and assessed via simulation studies to calculate the power of the resulting analyses.

Keywords: Functional data, model discrimination, T-optimality
AMS subject classifications: 62K05

1 Introduction

In many industrial and scientific experiments, each run can now produce a vast amount of data, collected using automatic monitoring and measurement systems. Often, it can be assumed that these data are generated by a smooth underlying function [5] and that the measurement processes are precise enough that the function can be accurately reconstructed, essentially without error. Then, the data can be assumed to be functional, with the output from each run of the experiment being a smooth function, typically not following a simple parametric form. Further, these functions may vary between runs of the experiment, potentially as the result of both aleatoric (i.e. random) variability and systematic variability resulting from application of different treatments, or combinations of values of the controllable factors. As with scalar regression, linear models may be used to partition this variability and assess how changes in treatment influence the shape of the functions.

This work is motivated by wear testing in Tribology, and in particular an experiment to study the wear in a pin and disc assembly for a given lubricant performed by the National Centre for Advanced Tribology, Southampton. The effects of six factors required investigation using a 20 run experiment, with each run defined by a different treatment. For each run, data on a number of functional responses were collected using automatic sensors, including the total wear of the pin and disc measured by a linear variable displacement transformer. The aim of the experiment was to understand which of the six factors had a substantive impact on each functional response; that is, to choose between contending functional linear models.

Here, we propose extending the criterion of T-optimality [2] to find designs that provide the most information for discriminating between two competing models. In Section 2, we introduce the functional linear

[*] Corresponding author, e-mail: V.Fisher@Southampton.ac.uk

model and, in Section 3, we go on to discuss T-optimal design that enables “best” discrimination between two competing functional linear models. In Section 4, we find a T-optimal design for a simple functional data example and perform a simulation study to assess the resulting sensitivity or power to detect the correct model.

2 Functional linear models

We assume the following linear model for the functional responses from an N-run experiment:

$$\text{M1}: \qquad \boldsymbol{Y}(t) = \mathbf{X}_1\boldsymbol{\beta}_1(t) + \boldsymbol{\varepsilon}(t)\,,$$

with $t \in \mathcal{T} \subset \mathbb{R}$, $\boldsymbol{Y}(t) = (Y_1(t), \dots, Y_N(t))^{\mathrm{T}}$, $\boldsymbol{\beta}_1(t) = (\beta_{11}(t), \dots, \beta_{1p_1}(t))^{\mathrm{T}}$, $\boldsymbol{\varepsilon}(t) = (\varepsilon_1(t), \dots, \varepsilon_N(t))^{\mathrm{T}}$ and $\mathbf{X}$ an $N \times p_1$ model matrix. The error functions $\varepsilon_j(t)$ are realisations from a Gaussian stochastic process with mean zero and covariance function $\gamma(t,u)$; for $i \neq j$, $\varepsilon_i(t)$ and $\varepsilon_j(u)$ are assumed independent. That is, the observed functions $Y_j(t)$ are assumed to be linear combinations of unknown functions $\beta_{1k}(t)$ with the addition of independent error functions $\varepsilon_j(t)$ ($j = 1, \dots, N$; $k = 1, \dots, p_1$).
The aim of the experiment is to discriminate between model M1 and a rival model

$$\text{M2}: \qquad \boldsymbol{Y}(t) = \mathbf{X}_2\boldsymbol{\beta}_2(t) + \boldsymbol{\eta}(t)\,,$$

with $\mathbf{X}_2$ an alternative $N \times p_2$ model matrix with corresponding vector of unknown functions $\boldsymbol{\beta}_2(t) = (\beta_{21}(t), \dots, \beta_{2p_2}(t))^{\mathrm{T}}$ and $\boldsymbol{\eta}(t)$ defined as $\boldsymbol{\varepsilon}(t)$. To discriminate between these two models, tests using the following quantity have been suggested [3]:

$$T = \int_t \left[\hat{\mathbf{Y}}_2(t) - \hat{\mathbf{Y}}_1(t)\right]^{\mathrm{T}} \left[\hat{\mathbf{Y}}_2(t) - \hat{\mathbf{Y}}_1(t)\right] \mathrm{d}t\,, \tag{1}$$

where $\hat{\mathbf{Y}}_i = (\hat{Y}_{i1}(t), \dots, \hat{Y}_{iN}(t))^{\mathrm{T}}$ are the fitted functions from model Mi (see Section 3). When model M1 is nested in model M2, under H_0 : *M1 is correct*, the distribution of the test statistic $\mathfrak{F} = \big\{(N - p_2)T\big\}/\big\{(p_2 - p_1)rss_2\big\}$ can be approximated by an F-distribution with adjusted degrees of freedom [6]. Here, rss_2 is the integrated residual sum of squares for model M2 and the adjustment to the degrees of freedom reflects the covariance function γ.

3 T-optimality for functional linear models

We find approximate optimal designs in f factors which are represented by a discrete probability measure ξ on the design region $\mathcal{X} = [-1, 1]^f$:

$$\xi = \left\{ \begin{array}{ccc} \boldsymbol{x}_1 & \dots & \boldsymbol{x}_n \\ w_1 & \dots & w_n \end{array} \right\}, \tag{2}$$

where $\boldsymbol{x}_j = (x_{j1}, \dots, x_{jf})^T \in \mathcal{X}$ are support points with associated weights $0 < w_j \leq 1$; $\sum_{j=1}^n w_j = 1$. Using data collected from design (2), we obtain fitted functions from model M1 as

$$\hat{\mathbf{Y}}_1(t) = \mathbf{X}_1\left(\mathbf{X}_1^{\mathrm{T}}\mathbf{W}\mathbf{X}_1\right)^{-1}\mathbf{X}_1^{\mathrm{T}}\mathbf{W}\boldsymbol{Y}(t) = \mathbf{H}\boldsymbol{Y}(t)\,.$$

Here $\mathbf{W} = \mathrm{diag}(w_1, \dots, w_n)$ and $\mathbf{X}_i$ is now defined for the n support points. Using (1), if we expect data from M2 where $E[\boldsymbol{Y}_2(t)] = \mathbf{X}_2\boldsymbol{\beta}_2(t)$, a T-optimal design $\xi^\star$ maximizes

$$
\begin{aligned}
\Phi(\xi) &= \int_t \left[E[\boldsymbol{Y}_2(t)] - \hat{\boldsymbol{Y}}_1(t)\right]^T \mathbf{W} \left[E[\boldsymbol{Y}_2(t)] - \hat{\boldsymbol{Y}}_1(t)\right] \mathrm{d}t \\
&= \int_t \left[\mathbf{X}_2\boldsymbol{\beta}_2(t) - \mathbf{H}\mathbf{X}_2\boldsymbol{\beta}_2(t)\right]^{\mathrm{T}} \mathbf{W} \left[\mathbf{X}_2\boldsymbol{\beta}_2(t) - \mathbf{H}\mathbf{X}_2\boldsymbol{\beta}_2(t)\right] \mathrm{d}t \\
&= \int_t \boldsymbol{\beta}_2^{\mathrm{T}}(t)\mathbf{X}_2^{\mathrm{T}} \left[\mathbf{I} - \mathbf{H}\right]^{\mathrm{T}} \mathbf{W} \left[\mathbf{I} - \mathbf{H}\right] \mathbf{X}_2\boldsymbol{\beta}_2(t) \, \mathrm{d}t \,. \qquad (3)
\end{aligned}
$$

Lemma 3.1. *Assume M1 is nested within M2, so $p_1 < p_2$, $\mathbf{X}_2 = [\mathbf{X}_1 : \mathbf{X}_{21}]$ and $\boldsymbol{\beta}_2^{\mathrm{T}}(t) = [\boldsymbol{\beta}_1^{\mathrm{T}}(t), \boldsymbol{\beta}_{21}^{\mathrm{T}}(t)]$ with $\mathbf{X}_{21}$ an $n \times (p_2 - p_1)$ model matrix and $\boldsymbol{\beta}_{21}$ an $(p_2 - p_1)$ vector of unknown functions. Then objective function (3) is given by*

$$
\Phi(\xi) = \int_t \boldsymbol{\beta}_{21}^{\mathrm{T}}(t)\mathbf{X}_{21}^{\mathrm{T}} \left(\mathbf{I} - \mathbf{H}\right)^{\mathrm{T}} \mathbf{W} \left(\mathbf{I} - \mathbf{H}\right) \mathbf{X}_{21}\boldsymbol{\beta}_{21}(t) \, \mathrm{d}t \,,
$$

and hence does not depend on the parameter vector $\boldsymbol{\beta}_1(t)$ which is common to both M1 and M2.

Proof. The proof is analogous to that for the scalar regression case [1]. □

Theorem 3.1. *Assume M1 is nested in M2, as in Lemma 1, and $p_2 = p_1 + 1$; that is, models M1 and M2 differ by only one term. Then the T-optimal design does not depend on the unknown function $\boldsymbol{\beta}_{21}(t)$.*

Proof. If $p_2 - p_1 = 1$, $\mathbf{X}_{21}$ is a $n \times 1$ vector and $\boldsymbol{\beta}_{21}(t)$ is a single function $\beta_{21}(t)$. From Lemma 1,

$$
\begin{aligned}
\Phi(\xi) &= \int_t \beta_{21}^2(t)\mathbf{X}_{21}^{\mathrm{T}} \left(\mathbf{I} - \mathbf{H}\right)^{\mathrm{T}} \mathbf{W} \left(\mathbf{I} - \mathbf{H}\right) \mathbf{X}_{21} \, \mathrm{d}t \\
&= \mathbf{X}_{21}^{\mathrm{T}} \left(\mathbf{I} - \mathbf{H}\right)^{\mathrm{T}} \mathbf{W} \left(\mathbf{I} - \mathbf{H}\right) \mathbf{X}_{21} \int_t \beta_{21}^2(t) \, \mathrm{d}t \\
&\propto \mathbf{X}_{21}^{\mathrm{T}} \left(\mathbf{I} - \mathbf{H}\right)^{\mathrm{T}} \mathbf{W} \left(\mathbf{I} - \mathbf{H}\right) \mathbf{X}_{21} \,, \qquad (4)
\end{aligned}
$$

where the constant of proportionality does not depend on ξ. Therefore, the T-optimal design that maximises (4) does not depend on the function $\beta_{21}(t)$. □

Corollary 3.1. *When M1 is nested in M2 and $p_2 = p_1 + 1$, it follows directly from (4) that the same design $\xi^\star$ is T-optimal for both the functional linear model and the scalar linear model.*

4 Example and simulation study

We construct a T-optimal design to compare the functional linear models

$$
Y(t) = \beta_{10}(t) + \beta_{11}(t)x + \epsilon(t) \,, \qquad (5)
$$

and

$$
Y(t) = \beta_{20}(t) + \beta_{21}(t)x + \beta_{22}(t)x^2 + \eta(t) \,. \qquad (6)
$$

That is, we find an optimal design to test if (5) is appropriate given data from (6). As the models differ by only one term, from Theorem 1, $\xi^\star$ will not depend on any of the unknown functions. Maximising (4) using the Nelder-Mead algorithm [4], we find the T-optimal design

$$
\xi^\star = \left\{ \begin{array}{ccc} -1 & 0 & 1 \\ 0.25 & 0.5 & 0.25 \end{array} \right\}, \qquad (7)
$$

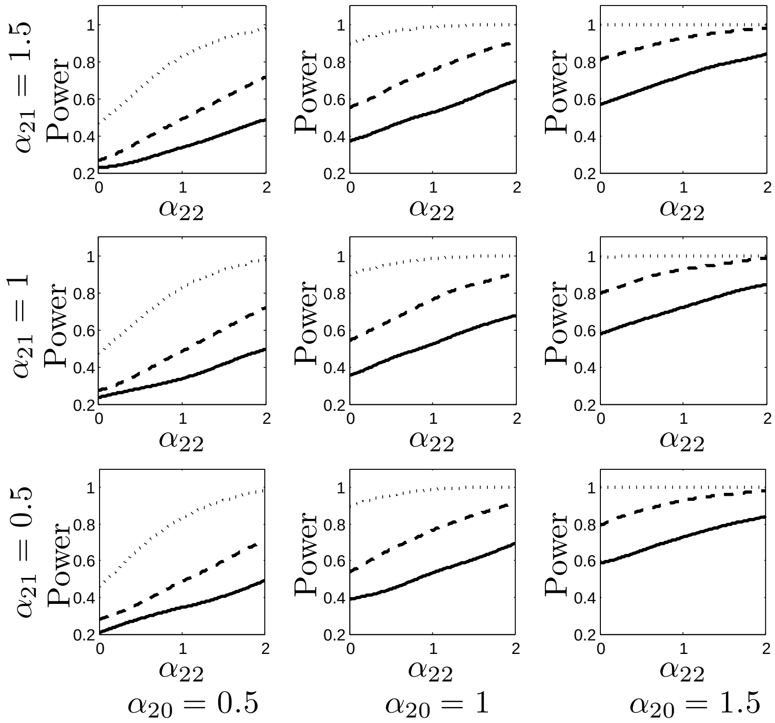


Figure 1: Power calculated from 1000 simulations using the functional T-optimal design for nine combinations of α_{20} and α_{21} values with $0 \leq \alpha_{22} \leq 2$, and number of runs $N = 12$ (–), $N = 24$ (−−) and $N = 72$ ($\cdots$).

which, from the corollary, is also T-optimal for comparing first- and second-order scalar regression models. To assess the power for rejecting H_0 : "model (5) is correct", we perform a simulation study using an exact T-optimal design with N runs obtained by rounding (7). Data is generated from model (6) assuming:

- the functions $Y_j(t)$ are observed at points $t_1, \dots, t_m \in [-1, 1]$, $j = 1, \dots, N$;
- $\beta_{2k}(t) = \alpha_{k0} + \alpha_{k1}t + \alpha_{k2}t^2$, $k = 0, 1, 2$;
- Cov $(\varepsilon_g(t_u), \varepsilon_h(t_v)) = \sigma_a^2 \rho^{|u-v|} + \sigma_b^2$ for $g = h$ and $0 < \rho < 1$, and 0 otherwise.

For each of $S = 1000$ generated data sets (with $\sigma_a^2 = 0.1$, $\sigma_b^2 = 2$, $\rho = 0.75$), we approximate (1) as

$$T \approx \sum_{j=1}^{m} \left[\hat{\mathbf{Y}}_2(t_j) - \hat{\mathbf{Y}}_1(t_j)\right]^{\mathrm{T}} \left[\hat{\mathbf{Y}}_2(t_j) - \hat{\mathbf{Y}}_1(t_j)\right] ,$$

calculate $\mathfrak{F} = (N - 3)T/rss_2$, where rss_2 is the residual sum of squares from model (6), and compare $\mathfrak{F}$ to the appropriate F-distribution [6]. We approximate the power as the proportion of simulations for which H_0 is rejected.

Figure 1 displays the results of this study. As the number, N, of design points increases, the power increases for all values of the other parameters considered, as expected. Further, (i) the power increases with the value of α_{22} for all other parameters; (ii) the larger the value of α_{20}, the higher the power over the whole range for α_{22}; and (iii) fixing α_{20} and increasing α_{21} (across rows in Figure 1) has little or no effect on the power. These observations can be explained by the form of the parameter function $\beta_{22}(t) = \alpha_{20} + \alpha_{21}t + \alpha_{22}t^2$ for $-1 \leq t \leq 1$. The magnitude of this function determines the difference between models (5) and (6). Increasing the value of either α_{20} or α_{22} clearly increases $\beta_{22}(t)$ for all $t \in [-1, 1]$, as $t^2 \geq 0$. However, the linear parameter α_{21} does not have a constant impact across t, and hence the value of this parameter has little overall effect on the size of $\beta_{22}(t)$. For larger α_{20}, there is a smaller difference in power between the different numbers of runs due to the more straightforward discrimination problem. Overall, for $N = 72$ runs and an appreciable difference between models (for example, $\alpha_{20} > 1$), the power to reject model (5) is greater than 90%.

5 Conclusions and future work

We have demonstrated the application of a model discrimination criterion for design selection with functional data and presented a series of results on the properties of the resulting designs. In particular, by establishing the equivalence of the functional and scalar linear model design problems for nested models differing by one term, we have made available well-known methods of T-optimality for a broader class of problems.

The near-ubiquity of functional data in many areas of science, engineering and industry encourages the further development of results for functional linear models. Possibilities for future work include optimality criteria for different experimental aims, and understanding the role of the reconstruction of functional data from discrete observations in the selection of optimal designs.

Acknowledgements: This work was supported by the UK Engineering and Physical Sciences Research Council and GlaxoSmithKline, the former through a PhD studentship (Fisher) and a Research Fellowship (Woods), the latter through a basic research grant. We would also like to express our thanks for the advice of Professor Susan Lewis (University of Southampton) throughout this work.

Bibliography

[1] Atkinson, A.C., Donev, A.N. and Tobias, R.D. (2007). *Optimum Experimental Designs, with SAS*, Oxford University Press, Oxford.

[2] Atkinson, A. C. and Fedorov, V. V. (1975). The design of experiments for discriminating between two rival models. *Biometrika*, 62, 57-70

[3] Faraway, J.J. (1997). Regression analysis for a functional response. *Technometrics*, 39, 254-261.

[4] Nelder, J.A. and Mead, R. (1965) A simplex method for function minimization. *The Computer Journal*, 7, 308-313.

[5] Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis*, Springer, New York.

[6] Shen, Q. and Faraway, J. (2004). An F test for linear models with functional responses. *Statistica Sinica*, 14, 1239-1257.

On sample selection models and skew distributions

Emmanuel O. Ogundimu[*1] **and Jane L. Hutton**[2]

[1]*Centre for Statistics in Medicine, University of Oxford, Botnar Research Centre*
[2]*Department of Statistics, University of Warwick, UK.*

Abstract

Scores arising from questionnaires often follow asymmetric distributions, on a fixed range. This can be due to scores clustering at one end of the scale or selective reporting. Sometimes, the scores are further subjected to sample selection, which is a class of missing data problem, resulting in partial observability. Thus, methods based on complete cases for skew data are inadequate for the analysis of such data and a general sample selection model is required. Heckman proposed a full maximum likelihood estimation method under the normality assumption for sample selection problems, and parametric and non-parametric extensions have been proposed. We generalize Heckman [5,6] to allow for underlying skew-normal distribution. Finite sample performance of the maximum likelihood estimator of the model is studied via Monte Carlo simulation. The model parameters are more precisely estimated under the new model, even in the presence of moderate to extreme skewness, than the Heckman selection models. Application to data from a study of neck injuries where the responses are substantially skew successfully discriminates between selection and inherent skewness.

Keywords: Generalized Sample selection, Missing data, Closed Skew-normal distribution.
AMS subject classifications: 62D99

1 Introduction

Scores arising from instruments designed to assess quality of life (QoL) (e.g. screening questionnaires) often follow asymmetric distributions due to skewness inherent in Likert-scale type instruments. In addition, the realized samples from the underlying discrete process are further subjected to selective reporting (e.g. the selection of maximum of correlated observations) and missing data, with the scores reflecting a selected population. Consequently, there is need for a general model for sample selection with inherent skewness.

A selection model was introduced by Heckman (see [4]). He proposed a full maximum likelihood estimation under the assumption of normality. His method was criticized on the ground of its sensitivity to normality assumption prompting him to develop the two-step estimator (see [5]). Sample selection models arise in practice as a result of the partial observability of the outcome of interest in a study. The data are missing not at random (MNAR) because the observed data do not represent a random sample from the population, even after controlling for covariates.

The two most common deviations from normality are heavier tails and skewness. In dealing with heavier tails in sample selection, [6] derived a model using links between hidden truncation and sample selection but with an underlying bivariate-t error distribution. They noted that a more appealing flexible parametric model is necessary that can accommodate heavy tails and skewness. Other researchers (e.g. [7]) noted that the effects of asymmetry on the normal theory methods are generally more serious than those of the nonnormal peakedness. We therefore propose the use of a skew-normal distribution for modeling asymmetry in sample selection framework.

[*] Corresponding author, e-mail: E.O.Ogundimu@warwick.ac.uk

A continuous random variable Z is said to have a standard skew-normal distribution with parameter $\lambda \in \mathbb{R}$ if its density is

$$f(z;\lambda) = 2\phi(z)\Phi(\lambda z), \quad z \in \mathbb{R}, \tag{1}$$

where ϕ and Φ denote the standard normal PDF (probability density function) and corresponding CDF (cumulative distribution function) respectively. The parameter λ is called the shape parameter because it regulates the shape of the density function.

2 Classical and general sample selection models

In this section, we review the classical selection normal model (SNM) and introduce a more general sample selection model with an underlying skew-normal error distribution.

2.1 Selection normal model (SNM)

Let $Y_i^\star$ be the outcome variable of interest, assumed linearly related to covariates x_i through the standard multiple regression

$$Y_i^\star = \beta' x_i + \sigma\varepsilon_{1i}, \quad i = 1, \dots, N.$$

Suppose the main model is supplemented by a selection (missingness) equation

$$S_i^\star = \gamma' x_i + \varepsilon_{2i}, \quad i = 1, \dots, N$$

where β and γ are unknown parameters and x_i are fixed observed characteristics not subject to missingness, the variance of $S_i^\star$ is fixed as 1 because the variance is not identifiable from sign alone. Selection is modeled by observing $Y_i^\star$ only when $S_i^\star > 0$, i.e. we observe $S_i = I(S_i^\star > 0)$ and $Y_i = Y_i^\star S_i$ for $n = \sum_{i=1}^N S_i$ of N individuals. Thus an observation has the conditional density

$$f(y|x, S^\star > 0) = \frac{f(y, S^\star > 0|x)}{P(S^\star > 0|x)} = \frac{f(y|x)P(S^\star > 0|y,x)}{P(S^\star > 0|x)}. \tag{2}$$

Equation (2) is the basis of the unification of selection problems as skew distributions given by [1]. The quantity $f(y|x)$ is a proper PDF, with a skewing function $P(S^\star > 0|y,x)$, and a normalizing function $P(S^\star > 0|x)$. It is straightforward to show that under the additional assumption

$$\begin{pmatrix}\varepsilon_{1i}\\ \varepsilon_{2i}\end{pmatrix} \sim N_2\left\{\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right\};$$

$$f(y|x, S=1;\Theta) = \frac{\frac{1}{\sigma}\phi\left(\frac{y-\beta' x}{\sigma}\right)\Phi\left(\frac{\gamma' x + \rho\left(\frac{y-\beta' x}{\sigma}\right)}{\sqrt{1-\rho^2}}\right)}{\Phi(\gamma' x)}, \tag{3}$$

(see [2]), where $\Theta = (\beta, \sigma, \gamma, \rho)$. The parameter $\rho \in$ [-1,1] determines the correlation of $Y_i^\star$ and $S_i^\star$, and hence the nature and severity of the selection process. The complete density of a sample selection model has a continuous component (the conditional density given by (3)), and a discrete component given by $P(S = 1|x)$. The marginal distribution of the selection equation determines the model to be fitted to the discrete process. In [2] and [4], a probit model $P(S = s) = \{\Phi(\gamma' x)\}^s\{1 - \Phi(\gamma' x)\}^{1-s}$ was used.

2.2 Selection Skew-normal model (SSNM)

Suppose we relax the assumption of bivariate normality given in section 2.1 such that the underlying error distribution is bivariate skew-normal. That is,

$$\begin{pmatrix}\varepsilon_{1i}\\ \varepsilon_{2i}\end{pmatrix} \sim SN_2\left\{\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix},\begin{pmatrix}\lambda\\0\end{pmatrix}\right\},$$

where λ is the skewness parameters for $Y_i^\star$. Then $f(y|x, S = 1; \Xi)$ (where $\Xi = \beta, \sigma, \gamma, \rho, \lambda$) is still defined as equation (2). The joint distribution of the outcomes and the selection process can be written in a closed skew-normal (CSN) distribution form (see [3] for details) as

$$\begin{pmatrix}Y^\star\\ S^\star\end{pmatrix} \sim CSN_{2,1}\left\{\boldsymbol{\mu} = (\beta' x, \gamma' x), \Sigma = \begin{pmatrix}\sigma^2 & \rho\sigma\\ \rho\sigma & 1\end{pmatrix}, D = (\lambda/\sigma, 0), \nu = 0, \Delta = 1\right\}.$$

Using the conditional and marginal distribution properties of the CSN distribution, we have

$$f(y|x, S = 1; \Xi) = \frac{\frac{2}{\sigma}\phi\left(\frac{y-\beta' x}{\sigma}\right)\Phi\left(\frac{\lambda(y-\beta' x)}{\sigma}\right)\Phi\left(\frac{\gamma' x+\rho\left(\frac{y-\beta' x}{\sigma}\right)}{\sqrt{1-\rho^2}}\right)}{\Phi_{SN}\left(\gamma' x; 0, 1, \frac{-\lambda\rho}{\sqrt{1+\lambda^2-\lambda^2\rho^2}}\right)}, \tag{4}$$

which is the continuous component of the SSNM log-likelihood function. The complete SSNM log-likelihood function is given by

$$l(\Xi) = \sum_{i=1}^{n} S_i\Big(\ln f(y_i|x_i, S_i = 1)\Big) + \sum_{i=1}^{n} S_i\Big(\ln \Phi_{SN}(\gamma' x_i; 0, 1, \lambda^\star)\Big) + $$
$$+ \sum_{i=1}^{n}(1 - S_i) \ln \Phi_{SN}(-\gamma' x; 0, 1, -\lambda^\star), \tag{5}$$

where $\lambda^\star = -\lambda\rho/\sqrt{1 + \lambda^2 - \lambda^2\rho^2}$ and $f(y|x, S = 1)$ by (4).

3 Simulation and Data example

In the section, simulation is used to study the finite samples properties of the MLEs for the SSNM and SNM models. The models are applied to the Neck disability index (NDI) scores.

3.1 Monte Carlo Simulation

We set the outcome and selection equations as $Y_i^\star = 0.5 + 1.5x_i + \varepsilon_{1i}$ and $S_i^\star = 1 + x_i + 1.5w_i + \varepsilon_{2i}$ respectively, where $i = 1, \ldots, N = 1000$. Thus, $\beta' = (0.5, 1.5)$, and γ'=(1, 1, 1.5). The covariates, x_i and $w_i \overset{iid}{\sim} N(0, 1)$, and are independent of ε_{1i} and ε_{2i} which are generated from bivariate skew-normal distribution with $\lambda = 0$ and 1 (note that the skewness parameter for the second equation is set to zero in both cases). The covariance matrix has $\sigma = 1$ & $\rho = 0.5$.

Table 1 shows that the SSNM model outperforms the SNM model for the skewness parameters considered, although the SNM model has a negligible advantage when $\lambda = 0$ with smaller bias in the intercept of the outcome equation. The SSNM gave consistently smaller bias as compared to the SNM model for the selection equation parts of the models when $\lambda = 0$ and 1. Since, the variance σ describes the variability of the probability distribution of the outcomes Y_i, and the bias in the estimation of the intercept and σ is less in the SSNM model, correct prediction intervals of new observations will be obtained under the model. The SNM model shows less variability in parameters estimation.

	$\lambda = 0$				$\lambda = 1.0$			
	Bias		Variance		Bias		Variance	
	SSNM	SNM	SSNM	SNM	SSNM	SNM	SSNM	SNM
β_0	16	-1	108	24	445	5620	341	14
β_1	-3	-3	19	19	4	10	12	12
γ_0	61	67	74	50	401	3516	266	82
γ_1	40	52	60	59	108	533	72	70
γ_2	80	98	94	92	201	835	134	122
σ	28	-9	17	9	-110	-1697	66	5
ρ	-7	-6	84	84	-72	-636	132	114
λ	-27	-	175	-	-501	-	1446	-

Table 1: Simulation results (in 1/10,000)

3.2 Data Application

We examine a longitudinal data set on neck injury which was collected using the NDI questionnaire, where two treatments are compared (Physiotherapy vs. Usual advice). The self-completed questionnaire assess pain-related activity restrictions in 10 areas including personal care, lifting, sleeping, driving, concentration, reading and work and results in a score between 0 and 50. The data were collected using questionnaires at regular intervals over a follow-up period at 4, 8 and 12 months after patient's emergency department attendance. We first identify predictors of dropout at each measurement occasion using probit regression. At month 8, age and sex of the patients are good predictors of missingness. We restricted attention to this measurement occasion to illustrate the new model.

A preliminary analysis shows that the effect of sex is not significant in the outcome equation of the models and it was removed.

	SSNM			SNM		
	Estimate	S.E.	p-value	Estimate	S.E	p-value
			Selection Equation			
int	0.208	0.177	0.239	0.835	0.100	0.000
age	0.021	0.005	0.000	0.024	0.006	0.000
sex(f)	0.309	0.126	0.014	0.335	0.129	0.009
			Outcome Equation			
int	-3.769	0.802	0.000	0.799	0.621	0.198
age	0.074	0.025	0.003	0.086	0.023	0.000
scores at Month 4	0.678	0.035	0.000	0.687	0.035	0.000
treatment(physio)	0.766	0.532	0.150	0.887	0.538	0.099
σ	7.723	0.563	0.000	6.166	0.292	0.000
ρ	0.758	0.174	0.000	0.802	0.072	0.000
λ	1.537	0.450	0.001	-	-	-

Table 2: Fit of selection skew-normal model (SSNM) and Selection-normal model (SNM) models to the NDI scores at 8 months.

Table 2 shows the results of fitting the SSNM and the SNM models to the NDI scores at month 8. As observed in the simulation study, the coefficients in the selection equations for the SNM model are consistently larger than the SSNM model. In particular, the estimate of the skewness parameter (λ = 1.537) is statistically significant in the SSNM model. This implies that neglecting the influence of λ in the model, although it leads to the same qualitative conclusions for the covariate effects in the outcome equation will lead to wrong predictive power of the model. In addition, the SSNM model has a better fit (log-likelihood = -1452.67) to the NDI data than the SNM model (log-likelihood =-1455.03) at the cost of 1 degree of freedom.

In conclusion the SSNM has good estimates of the intercept both in the selection and outcome equations and hence will give better predictions even when the underlying process is bivariate normal. The model is well identified in the sense that for any $\Theta_1 \neq \Theta_2$, $f(y, \Theta_1) \neq f(y, \Theta_2)$, where Θ_1 and Θ_2 are model parameters. Further, the observed information matrix is non-singular, although there is stationarity of the profile likelihood for λ at $\lambda = 0$. Further extension to bivariate skew-t distribution may be able to better handle heavy tails and skewness simultaneously, and it is currently being investigated.

Bibliography

[1] Arellano-Valle, R. B. and Marcia, D. B. and Genton, M. G. (2006). A unified view of skewed distributions arising from selections. *The Canadian Journal of Statistics* 34, 581–601.

[2] Copas, J. B. and Li, H. (1997). Inference for non-random samples. *Journal of the Royal Statistical Society: Series B* 59, 55–95.

[3] Gonzalez-Farias, G. and Dominguez-Molina, J. A. and Gupta, A. K. (2004) *The closed skew-normal. In M. G. Genton (Ed.), Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*, Chapman & Hall, CRC, Florida.

[4] Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement* 5, 475–492.

[5] Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica* 47, 153–161.

[6] Marchenko, Y. V. and Genton, M. G. (2012). A Heckman Selection-t Model. *Journal of the American Statistical Association* 107:497, 304–317.

[7] Mudholkar, G. S. and Hutson, A. D. (2000). The epsilon-skew normal distribution for analyzing near-normal data. *Journal of Statistical Planning and Inference* 83, 291–309.

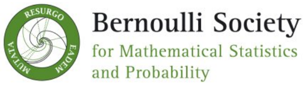
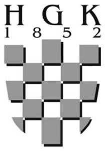
List of Corresponding Authors

COUNTRY	PARTICIPANT	AFFILIATION	E-MAIL ADDRESS
Austria	Peter Scheibelhofer	Graz University of Technology (PhD student) and ams AG	peter.scheibelhofer@ams.com
	Lukas Steinberger	Department of Statistics and OR, University of Vienna	lukas.steinberger@univie.ac.at
Belgium	Joke Durnez	Ghent University	joke.durnez@ugent.be
	Rudolf Schenk	Université catholique de Louvain	rudolf.schenk@uclouvain.be
Bulgaria	Bono Nonchev	Faculty of Mathematics and Informatics, Sofia University ”St. Kliment Ohridski”	bono@nonchev.info
	Teodosi Geninski	Faculty of Mathematics and Informatics, Sofia University ”St. Kliment Ohridski”	teodosi.g@gmail.com
Croatia	Danijel Grahovac	Department of Mathematics, J.J. Strossmayer University of Osijek	dgrahova@mathos.hr
Czech Republic	Petr Novák	Department of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University, Prague	novakp@karlin.mff.cuni.cz
Czech Republic	Katarína Starinská	Department of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University, Prague	starinskak@gmail.com
Denmark	Andreas Basse-O’Connor	Department of Mathematics, University of Aarhus	basse@imf.au.dk
	Alexander Sokol	Department of Mathematical Sciences, University of Copenhagen	alexander@math.ku.dk
Finland	Leena Annukka Pasanen	Department of Mathematical Sciences, University of Oulu	Leena.Pasanen@oulu.fi
France	Eric Sibony	Télécom ParisTech	esibony@gmail.com
Germany	Eike Christian Brechmann	Technische Universität München, Zentrum Mathematik, Lehrstuhl für Mathematische Statistik	brechmann@ma.tum.de
	Philip Preuß	Ruhr-Universität Bochum, Fakultät für Mathematik, Lehrstuhl für Stochastik	philip.preuss@ruhr-uni-bochum.de
Greece	Katerina Orfanogiannaki	Department of Statistics, Athens University of Economics and Business	korfanogiannaki@gmail.com
	Panagiotis Papastamoulis	Department of Statistics & Insurance Science, University of Pireaus	papapast@yahoo.gr
Italy	Manuela Cattelan	Department of Statistical Sciences, University of Padova	manuela.cattelan@stat.unipd.it
The Netherlands	Botond Szabó	Eindhoven University of Technology	b.szabo@tue.nl
Poland	Piotr Szulc	Institute of Mathematics and Computer Science, Wroclaw University of Technology	piotr.a.szulc@pwr.wroc.pl
Portugal	B.G. Manjunath	Department of Statistics and Applications, Faculty of Sciences, Lisbon University	bgmanjunath@gmail.com
	Laetitia Da Costa Teixeira	Faculty of Sciences and Institute of Biomedical Sciences Abel Salazar, University of Porto	laetitiateixeir@gmail.com

COUNTRY	PARTICIPANT	AFFILIATION	E-MAIL ADDRESS
Romania	Mircea Dragulin	Faculty of Mathematics and Informatics, University Bucharest	mircea.mate@yahoo.com
Russia	Ekaterina Krymova	Institute for Information Transmission Problems, Moscow	ekkrym@gmail.com.
Slovakia	Gábor Szücs	Faculty of Mathematics, Physics and Informatics, Department of Applied Mathematics and Statistics, Comenius University	szucs@fmph.uniba.sk
	Alena Bachratá	Faculty of Mathematics, Physics and Informatics, Department of Applied Mathematics and Statistics, Comenius University	alena.bachrata@fmph.uniba.sk
Slovenia	Žiga Kotnik	Faculty of Administration, University of Ljubljana	ziga.kotnik@fu.uni-lj.si
Spain	Nuria Torrado-Robles	Department of Statistical Methods, University of Zaragoza	nuria.torrado@gmail.com
	Guillermo Vinué	Department of Statistics and Operational Research, Faculty of Mathematics, Universidad de Valencia	Guillermo.Vinue@uv.es
Sweden	Katja Trinajstić	Department of Mathematics, Uppsala University	katja@math.uu.se
	Maik Görgens	Department of Mathematics, Uppsala University	maik@math.uu.se
Switzerland	Kaspar Stucki	Institut for Mathematical Statistics and Actuarial Science, University of Bern	kaspar@stucki.org
Turkey	Ufuk Beyaztas	Department of Statistics, Faculty of Science, Dokuz Eylul University	ufuk.beyaztas@deu.edu.tr
Ukraine	Kostiantyn Ralchenko	Dept. of Probability Theory, Statistics and Actuarial Mathematics, Mechanics and Mathematics Faculty, Taras Shevchenko National University of Kyiv	k.ralchenko@gmail.com
	Alexey Doronin	Dept. of Probability Theory, Statistics and Actuarial Mathematics, Mechanics and Mathematics Faculty, Taras Shevchenko National University of Kyiv	al_doronin@ukr.net
United Kingdom	Emmanuel Ogundimu	Centre for Statistics in Medicine, University of Oxford, Botnar Research Centre	O.E.Ogundimu@warwick.ac.uk
	Verity Fisher	Mathematics, University of Southampton	v.fisher@southampton.ac.uk

Author index

Sponsors

	Bernoulli Society for Mathematical Statistics and Probability
	Ministry of Science, Education and Sports of the Republic of Croatia
	Croatian Academy of Sciences and Arts
	Croatian Chamber of Economy
	Osijek-Baranja County

City of Osijek

AZ Pension Fund

Konzum

Mehanotehna d.o.o.

Tourist Board, City of Osijek