Proceedings of the 19th European Young Statisticians Meeting

August 31 – September 4, 2015
Prague

Stanislav Nagy
(editor)

Proceedings of the 19th European Young Statisticians Meeting.
Prague, August 31 – September 4, 2015
Stanislav Nagy, editor
Prague 2015.

ISBN 978-80-7378-301-3

Editor:
Stanislav Nagy, nagy@karlin.mff.cuni.cz

Department of Probability and Mathematical Statistics
Faculty of Mathematics and Physics
Charles University in Prague
Sokolovská 83, 186 75 Praha 8 – Karlín, Czech Republic.

Published by MATFYZPRESS
Publishing House of the Faculty of Mathematics and Physics
Charles University in Prague
Sokolovská 83, 186 75 Praha 8, Czech Republic
as the 495. publication.
The publication didn't pass the review or lecturer control.
Prague 2015

Scientific Programme Committee

Charles University in Prague, Czech Republic

- Marie Hušková
- Matúš Maciak
- Stanislav Nagy

Local Organizing Committee

Charles University in Prague, Czech Republic

- Zdeněk Hlávka
- Daniel Hlubinka
- Marie Hušková
- Matúš Maciak
- Marek Omelka
- Marie Turčičová

Conference Partners

Preface

The first European Young Statisticians Meeting was organized in 1978 (Wiltshire, Great Britain), the second one in 1981 (Bressanone, Italy), and since then regularly every two years in different European countries.

From the very beginning the idea of the event is that young researchers from different countries come together and establish new research contacts at the beginning of their scientific careers.

In line with previous meetings, each of the representatives from selected European countries suggested at most two young researchers to participate in the Meeting. Also, five distinguished researchers have been invited to give plenary lectures.

We hope that you find the Meeting interesting and useful.

Welcome to the 19th EYSM 2015 in Prague. Enjoy the city, its history, its architecture and culture, and have a great time!

Local Organizing Committee
Prague, July 2015

Contents

Optimal Bandwidths for Multiscale Local Polynomial Decompositions

Mohamed Amghar[*1] **and Maarten Jansen**[2]

[1]*Department of Mathematics and Statistics, Université Libre de Bruxelles, Belgium*
[2]*Departments of Mathematics and Computer Science, Université Libre de Bruxelles, Belgium*

Abstract: This paper discusses the choice of the bandwidths in a multiscale local polynomial data transform. The transform adopts the local polynomial smoothing paradigm for the construction of a multiresolution data decomposition, much like a wavelet transform or a Laplacian pyramid. The bandwidths depend on the resolution level, defining for each level the scale of the coefficients. As a result, the scale is not necessarily dyadic as in a discrete wavelet transform, nor is it grid dependent as in second generation wavelet transform. Unlike in a uniscale local polynomial smoothing scheme, the bandwidth in a multiscale data transform is not optimised for data processing, i.e. smoothing, but rather for data transformation. The bandwidth at each level should be chosen in a way that it makes the representation after transformation as suitable as possible for subsequent, non-linear processing. We argue that the choice typically amounts to maximisation of the L1-sparsity of the data in the absence of noise. We also investigate the multivariate optimisation problem of choosing bandwidths at successive scales.

Keywords: local polynomial, thresholding, sparsity, bandwidth, wavelet
AMS subject classifications: 62J07, 62G08, 62J02

1 Introduction

In a wavelet representation, data are decomposed into a basis that consists of basis functions that are all translations and dilations of a single mother function. As a consequence, each wavelet coefficient carries specific, local information about the data. More specifically, it describes the contribution at a local scale and at a local point in a time to the data. As all basis functions are translations and dilations, the data must be sampled on an equispaced, dyadic grid of locations.

The multiscale local polynomial decomposition [3] provides an alternative for wavelet transforms when the observations are non-equispaced. The decomposition combines the benefits of two approaches. On one hand, the local polynomial approach leads to a representation in which all coefficients carry information that is local in time. On the other hand, the scales in multiscale transformation are set by choosing a sequence of bandwidths. In this transform, the bandwidth is a scale parameter. It provides a natural way to deal with data of which the sampling

*Corresponding author: Mamghar@ulb.ac.be

rate fluctuates over time. As the bandwidths play a role in a data transformation, rather than in data processing, the bandwidths selection is driven by different objectives than in a uniscale smoothing setting. The bandwidths should be taken such that the data are represented in an optimal way for further nonlinear processing, however without processing the data at this stage.

Section 2 of this paper reviews the key elements of the multiscale local polynomial transform, highlighting the differences with a wavelet transform. Next, in Section 3, we discuss a model for sparsity leading to a criterion for the optimal selection of the bandwidths. Finally, in Section 4, we explore a few heuristics in the multivariate bandwidth selection problem.

2 Multiscale local polynomial decomposition

The multiscale local polynomial transform is based on a Laplacian pyramid scheme [4], which starts off by assigning the vector of observations $\boldsymbol{Y}$ to a finest scale vector $\boldsymbol{s}_J$. From there on, iterations over scales $j = J-1, J-2, \ldots, L$ proceed as

$$\boldsymbol{s}_j = (\widetilde{H}_j \boldsymbol{s}_{j+1})_e, \tag{1}$$

$$\boldsymbol{d}_j = D_j^{-1}(\boldsymbol{s}_{j+1} - P_j \boldsymbol{s}_j). \tag{2}$$

In this expression, index e stands for a subset of $\{0, \ldots, n_{j+1}-1\}$, where n_{j+1} is the length of the vector $\boldsymbol{s}_{j+1}$. The subset e typically (but not necessarily) contains the set of even numbers in $\{1, \ldots, n_{j+1}\}$, meaning that $\boldsymbol{s}_j = (\widetilde{H}_j \boldsymbol{s}_{j+1})_e$ contains the even subsamples of the vector $\widetilde{H}_j \boldsymbol{s}_{j+1}$. The matrix $\widetilde{H}_j$ is a square, not necessarily invertible matrix, aiming at some preprocessing of the data, which could be antialiasing for instance. In this paper, we take $\widetilde{H}_j = I_{n_{j+1}}$, and so $\boldsymbol{s}_j = \boldsymbol{s}_{j+1,e}$. If e is indeed the set of evens, then at coefficient level we have $s_{j,k} = s_{j+1,2k}$ and so $n_j = \lceil n_{j+1}/2 \rceil$. Furthermore, in (2), D_j is an optional diagonal matrix, used for normalisation or standardisation of the coefficients. More importantly, P_j is the local polynomial smoothing matrix, whose rows are filled in by $P_{j;\mathrm{row}k} = P_j(t_{j+1,2k+1}; \boldsymbol{t}_j)$. In this expression, $\boldsymbol{t}_{j+1}$ is the grid of locations or covariate values at scale $j+1$ and $\boldsymbol{t}_j = \boldsymbol{t}_{j+1,e}$ is the subsampled version of it.
The function $P_j(t; \boldsymbol{t}_j)$, not to be confused with the matrix P_j, carries out the smoothing, using a locally least squares polynomial of degree $\widetilde{p}-1$, i.e.,

$$P_j(t; \boldsymbol{t}_j) = \mathrm{T}^{(\widetilde{p})}(t) \left(\mathrm{T}_j^{(\widetilde{p})^T} \mathrm{W}_j(t) \mathrm{T}_j^{(\widetilde{p})} \right)^{-1} \left(\mathrm{T}_j^{(\widetilde{p})^T} \mathrm{W}_j(t) \right). \tag{3}$$

In this expression, $\mathrm{T}^{(\widetilde{p})}(t)$ is a row vector of power functions, $\mathrm{T}^{(\widetilde{p})}(t) = [1\, t\, \ldots\, t^{\widetilde{p}-1}]$. The matrix $\mathrm{T}_j^{(\widetilde{p})}$ replaces the power functions in each column of $\mathrm{T}^{(\widetilde{p})}(t)$ by a column of evaluations in the grid $\boldsymbol{t}_j$, leading to the construction $\mathrm{T}_j^{(\widetilde{p})} = [\mathbf{1}\, \boldsymbol{t}_j\, \ldots\, \boldsymbol{t}_j^{\widetilde{p}-1}]$. Finally, $\mathrm{W}_j(t)$ is a diagonal matrix of weight functions with on the diagonal $(\mathrm{W}_j)_{kk}(t) = K\left(\frac{t-t_{j,k}}{h_j}\right)$. The function $K(t)$ is the kernel function and h_j is the bandwidth.

Iterative application of (1) and (2) leads to a decomposition of $\boldsymbol{Y} = \boldsymbol{s}_J$ into $[\boldsymbol{s}_L, \boldsymbol{d}_L, \ldots \boldsymbol{d}_{J-1}]$. The inverse transform leading to the reconstruction of $\boldsymbol{s}_J$ can

be realized by the iteration

$$\boldsymbol{s}_{j+1} = D_j \boldsymbol{d}_j + P_j \boldsymbol{s}_j, \tag{4}$$

starting from $\boldsymbol{s}_L$ and $\boldsymbol{d}_L$. The decomposition is overcomplete, as the number of coefficients in the representation equals

$$\#\{s_{L,k}\} + \#\{d_{j,k}, j = L, \ldots, J-1\} = n_L + \sum_{j=L}^{J-1} n_{j+1} = \sum_{j=L}^{J} \lceil n/2^{J-j} \rceil = \mathcal{O}(2n).$$

We emphasize that although the transform is redundant, most of the $\mathcal{O}(2n)$ coefficients in the decomposition will be close to zero. This sparsity allows us to use the decomposition in a subsequent data compression scheme. The redundancy is in contrast to a fast wavelet transform, which is critically subsampled, meaning that n observations lead to n coefficients. The difference between a redundant and a critically downsampled transform comes from a different construction of the detail coefficients $\boldsymbol{d}_j$. This is best illustrated with a simple example of a wavelet transform whose form comes as close as possible to that of the multiscale local polynomial transform. That wavelet transform is proceeds as an iteration of

$$\boldsymbol{s}_j = \boldsymbol{s}_{j+1,e}, \tag{5}$$

$$\boldsymbol{d}_j = D_j^{-1}(\boldsymbol{s}_{j+1,e'} - P_j \boldsymbol{s}_j). \tag{6}$$

In each step, the index set $\{0, \ldots, n_{j+1}\}$ is partitioned into "even" and "non-even" (odd) complements e and e'. As a consequence the number of detail coefficients equals $\#\{d_{j,k}\} = n_{j+1} - n_j$, and hence $\#\{s_{L,k}\} + \#\{d_{j,k}, j = L, \ldots, J-1\} = n_L + \sum_{j=L}^{J-1}(n_{j+1} - n_j) = n_J = n$. In each step, one half of the data, $\boldsymbol{s}_{j+1,e}$, is used to predict the other half, $\boldsymbol{s}_{j+1,e'}$, using a prediction matrix P_j. The construction is a simple example of a lifting scheme [6]. All classical wavelets can be factored into this scheme. On the other hand, just as the multiscale local polynomial transform, the construction of a lifting scheme does not require equispaced, dyadic data nonequispaced data. When the scheme takes the irregularity of $\boldsymbol{t}_{j+1}$ into account, the resulting wavelets are termed second generation wavelets [7].

Unlike the multiscale local polynomial transform, however, the lifting scheme cannot use local polynomial or any other smoothing operation in its prediction matrix P_j. This is because the inverse transform from processed coefficients would lead to a fractal like reconstruction [3]. This can be understood by looking at the reconstruction from a coarse scale approximation where all details happen to be zero. In that case, the odd fine scale coefficients follow from $\boldsymbol{s}_{j+1,e'} = P_j \boldsymbol{s}_j + \boldsymbol{d}_j = P_j \boldsymbol{s}_{j+1,e}$. For a smooth reconstruction, it is necessary that if an odd point $t_{j+1,2k+1}$ is close to its even neighbour, then so should be the coefficients. This means that

$$\lim_{u \to t_{j,k}} P_j(u; \boldsymbol{t}_j) \cdot \boldsymbol{s}_j = s_{j,k}, \tag{7}$$

which is not the case if $P_j(u; \boldsymbol{t}_j)$ is a smoothing operation. Instead, wavelet transforms either use more complicated lifting schemes, or, if they use a simple prediction operation, then this must be interpolating.

Besides the overcompleteness and the sort of prediction operation, a third important difference between wavelets and multiscale local polynomial transforms lies in the definition of scales at each resolution level. In a wavelet transform, the scale of the prediction operator follows from the distance between adjacent points in $\boldsymbol{t}_j$. The prediction uses a fixed number of points from $\boldsymbol{t}_j$ that are close to a given point in $\boldsymbol{t}_{j+1,e'}$. The scale of the prediction operation thus depends on the local sample density. Moreover, in 2D, the lifting scheme needs a triangulation or some other system that describe neighbourhood. In the local polynomial transform, both scale and neighbourhood are fixed by the bandwidth, which leads to a more stable decomposition that is also easier to implement. The number of nonzeros in row k of the prediction matrix P_j depends on the number of adjacent points within the bandwidth around $t_{j+1,2k+1}$. A matrix P_j in a wavelet transform has a fixed number of nonzeros in each of its rows.

3 A model for bandwidth selection

In a multiscale local polynomial transform, the bandwidth represents the scale of resolution level j. It also defines the set of neighbours for each point in $\boldsymbol{t}_{j+1}$. Unlike in uniscale local polynomial smoothing [1, Chapter 3], or local polynomials with time varying bandwidths [5], the objective in the context of this paper is to control, but not to reduce the variance of the reconstruction. Variance reduction, denoising, or smoothing is left to the nonlinear processing within the sparse representation.

The nonlinear processing consists of a selection of large coefficients, for instance by thresholding. In this framework, the bandwidths h_j are chosen to make the selection as successfull as possible. The success of a nonlinear processing depends of course on the strategy and the criterion used in the selection, but also on the sparsity of the data representation. By sparsity we mean that the information available from n observations can be captured by a small subset of the coefficients, while most of the coefficients are close to zero. This is modelled by assuming that all coefficients $d_{j,k}$ come from a random variable plus noise [2]. The model for the random variable is a mixture distribution, imposing most observations to be near zero, while a few outliers carry all the essential information. The sparsity model becomes

$$\widetilde{D}_n = X_n D_{n,1} + (1 - X_n) D_{n,0} + \sigma Z. \tag{8}$$

The dependence on the sample size allows us to let sparsity increase for $n \to \infty$, thereby expressing a general principle that higher the sample sizes generally imply more redundancy in the observations. The variables $D_{n,x}$ for $x \in \{0, 1\}$ are modelled to have a double exponential (Laplacian) distribution with parameters $a_{n,x}$. The binary label X_n, with Bernoulli distribution, labels the class to which $\widetilde{D}_n$. With small probability $p_n = P(X_n = 1)$, we have $\widetilde{D}_n = D_{n,1} + \sigma Z$, i.e., $\widetilde{D}_n$ is a large coefficient with noise. This occurs if the prediction of $s_{j+1,2k+1}$ is far from the actual value, which is the case if $t_{j+1,2k+1}$ lies within a bandwidth h_j from a singularity in $f(t)$. In the other case, modelled by the even $\{x_{j,k} = 0\}$, the detail offset $d_{j,k}$ is small plus noise, which corresponds to $f(t)$ being Lipschitz $\widetilde{p}$ continuous on $[t_{j+1,2k+1} - h_j, t_{j+1,2k+1} + h_j]$.

In general, the bandwidths in a data transform will be smaller than in a context of data smoothing, as there is no reason to take the bandwidth any larger than strictly necessary for the purpose of the construction of a $\widetilde{p}-1$ degree polynomial around each point in $\boldsymbol{t}_{j+1}$. Unnecessarily large bandwidths would increase the number of coefficients that have a singularity within a bandwidth's distance. In practice, the bandwidth will even be smaller, leaving some of the points $\boldsymbol{t}_{j+1}$ with a lower degree local polynomial. If the detail coefficient in such a point is small, modelled by $D_{n,0}$, that is, it may see a slightly increasing value due to the lower degree of the polynomial. This increase is compensated by a reduced number of large coefficients, those modelled by $D_{n,1}$.

The objective is to select the bandwidths h_j so that the sparsity model (8) becomes as likely as possible. Let $\widehat{p}_n$, $\widehat{a}_{n,0}$, $\widehat{a}_{n,1}$ and $\widehat{\sigma}^2$ be the maximum likelihood estimators for the model parameters, given the coefficients $\boldsymbol{d} = [\boldsymbol{d}_j]$ for given choices of h_j, $j = L, \ldots, J-1$. Then we optimize the likelihood

$$L(\widehat{p}_n, \widehat{a}_{n,0}, \widehat{a}_{n,1}, \widehat{\sigma}^2; \boldsymbol{d}; \boldsymbol{h})$$

as a function of $\boldsymbol{h} = [h_L, \ldots, h_{J-1}]$. Unfortunately, the likelihood turns out to be very sensitive to misspecification of the model for the small coefficients, leading to poor estimations of $a_{n,0}$, and consequently, poor comparisons of maximum likelihood values for different choices of bandwidths. On the other hand, as the information is concentrated in the large coefficients, the quality of the representation depends primarily on these values. It can be formalized that under mild conditions [2], the soft-thresholded (**ST**) coefficients have a zero inflated Laplacian distribution whose parameter does not depend on the noise, i.e., for an appropriate threshold λ_n, we have

$$\mathrm{ST}(\widetilde{D}_n, \lambda_n) \xrightarrow{d} \widetilde{X}_n D_{n,1}, \tag{9}$$

where $\widetilde{X}_n = I(|\widetilde{D}_n| > \lambda_n)$. The optimization of the likelihood of the model for $\mathrm{ST}(\widetilde{D}_n, \lambda_n)$ amounts to a minimization of the ℓ_1 norm of the thresholded coefficients, i.e., find h_j so that

$$\sum_{j=L}^{J-1} \sum_{k=1}^{n_{j+1}} |\mathrm{ST}(d_{j,k}, \lambda_n)| = \sum_{j=L}^{J-1} \sum_{k=1}^{n_{j+1}} \mathrm{ST}(|d_{j,k}|, \lambda_n)$$

is minimized.

4 Bandwidths in a multiscale transform

Since the bandwidth operates as the scale in a multiscale decomposition, it can be optimised at each resolution level j. Our first simulations, illustrated in Figure 1 seem to suggest that the bandwidth at each scale roughly increases in a dyadic way, but not quite so. Our experiment was set up as follows: a test signal, commonly known as the heavisine test function, was sampled without error, at 1000 inequidistant points, which were uniformly distributed on the x-axis. Next, the multiscale local linear transform was carried out using 4 resolution levels, using a cosine kernel function. At each resolution level, the optimal bandwidth was defined as the

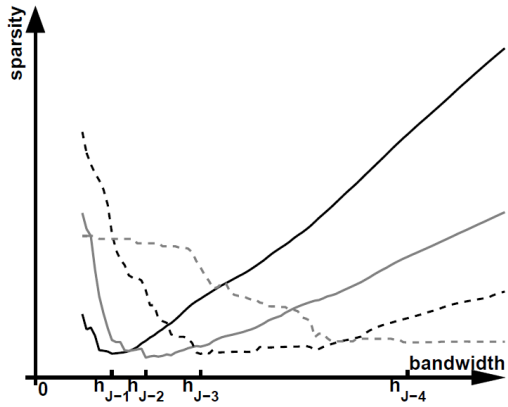


Figure 1: Sparsity, defined as $\|\boldsymbol{d}_j\|_1$, as a function of bandwidth h_j at scales $j = J-1, J-2, J-3, J-4$. The optimal bandwidths at finer scales are used when proceeding to the next coarser scale.

bandwidth that minimizes $\|\boldsymbol{d}_j\|_1$, and this bandwidth was used when proceeding to the next, coarser scale. Further experiments confirm that this scale-by-scale optimization finds a vector of bandwidths that is close to the globally optimal vector of bandwidths.

Acknowledgements: Research support by the IAP research network grant nr. P7/06 of the Belgian government (Belgian Science Policy) is gratefully acknowledged.

References

[1] J. Fan and I. Gijbels. Local Polynomial Modelling and its Applications. *Chapman and Hall, London,* 1996.

[2] M. Amghar and M. Jansen. Using bandwidths as scales in multiscale local polynomial decompositions. *In preparation,* 2015.

[3] M. Jansen. Multiscale local polynomial smoothing in a lifted pyramid for non-equispaced data. *IEEE Transactions on Signal Processing,* 61(3):545555, 2013.

[4] P. J. Burt and E. H. Adelson. Laplacian pyramid as a compact image code. *IEEE Trans. Commun,* 31(4):532540, 1983.

[5] P. Vieu. Nonparametric regression: Optimal local bandwidth choice. *Journal of the Royal Statistical Society, Series B,* 53(2):453464, 1991.

[6] W. Sweldens. The lifting scheme: A custom-design construction of biorthogonal wavelets. *Appl. Comput. Harmon. Anal*, 3(2):186200, 1996.

[7] W. Sweldens. The lifting scheme: a construction of second generation wavelets. *SIAM J. Math. Anal,*29(2):511546, 1998.

Modelling Company Performance Based on Financial Ratios

Slav Angelov*

New Bulgarian University, Department of Informatics, Bulgaria

Abstract: This research is based on the information gathered from the yearly financial reports of the companies from the bulgarian gas supplying industry. There are around 30 firms licensed to do such an activity. Analysing their behavior, stability and future is a part of the macroeconomic situation in the country.
The goal is to make a model which will predict their stability.
We will explore a few dozen financial ratios which are extracted from the data in the reports. To reach the goal regression analyses techniques will be used. Calculations and graphics are realized in R language. The model will be compared with existing econometric models e.g Altman Z-score model.

Keywords: regression models, financial ratios, Altman Z-score
AMS subject classifications: 62J02, 62J05

*Corresponding author: slav_angelov@abv.bg

Weibull-Poisson Regression Model with Shared Gamma Frailty

Oykum Esra Askin[*1] **and Deniz Inan**[2]

[1] *Yildiz Technical University, Department of Statistics,Turkey*
[2] *Marmara University, Department of Statistics, Turkey*

Abstract: Frailty models are extensively used in modeling unobserved heterogeneity. The hazard shape of lifetimes should be determined correctly in order to obtain unbiased parameter estimates. Weibull, Gompertz and Exponential distributions are the most popular on the choice of hazard function. In some cases, the occurrence of an event depends on several causes or latent risks and we can only observe the minimum lifetime. Based on this type of latent competing risk scenario, several distributions have been introduced as a particular case of homogeneous Poisson process. To the best of our knowledge, there is no work examines the convenience of such distributions in frailty models. In this study, we propose a bivariate Weibull-Poisson regression model with shared gamma frailty. The Particle Swarm Optimization (PSO) is performed to find MLEs of simulated data.

Keywords: frailty model, shared gamma frailty, particle swarm optimization
AMS subject classifications: 62N01, 62N02, 68T20

References

[1] D. Clayton. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrica*, (65):141-151, 1978.

[2] D. Hanagal, R. Sharma. Bayesian estimation of parameters for the bivariate Gompertz regression model with shared gamma frailty under random censoring. *Statistics & Probability Letters*, 82(7), 1310-1317, 2012.

[3] C. Kus. A new lifetime distribution. *Computational Statistics and Data Analysis*, (51):4497-4509,2007.

[4] W. Lu, D. Shi. A new compounding life distribution: the Weibull-Poisson distribution. *Journal of Applied Statistics*, 39(1):21-38, 2011.

[5] M. Macera, F. Louzada, V. Cnacho, C. Fontes. The exponential-Poisson model for recurrent event data: An application to a set of data on malaria in Brazil. *Biometric Journal*:doi:10.1002/bimj.201300116, 2014.

[6] J. Vaupel, K. Manton,E. Stallard. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Biometrics*, (16): 439-454,1978

*Corresponding author: oykumesra@gmail.com

A Mentenance Model with a Quasi Generalized Lindley Distribution

Irina Adriana Bancescu*

Doctoral School of Mathematics, University of Bucharest, Romania

Abstract: Lindley distribution has been recently in the attention of statistics that showed its flexible properties proving its better suitability for modelling lifetime data. This paper proposes a new generalization which has three submodels: the Lindley, exponential and gamma distribution. Several properties have been discussed and an application for mentenance models is proposed.

The Lindley distribution was introduced by Lindley (1958) as a new distribution useful to analyze lifetime data especially in applications modeling stress-strength reliability. Ghitany et al. (2008) have showed that this distribution is better than the exponential one when its come to modelling lifetime data. They also showed in a numerical example that the Lindley distribution gives better modeling for waiting times and survival times data than the exponential distribution.

In 2013 Rama and Mishra have introduced a new two-parameter Quasi Lindley distribution (QLD), of which the Lindley distribution (LD) is a particular case. The properties of OLD have been studied showing its better flexible than Lindley and exponential distributions.

We introduce a new quasi generalized Lindley distributions (QGL) which reduces not only to the Lindley distribution, but also to the gamma and exponential distribution, so being more suitable for modelling lifetime data.

With the development of industry mentenance models have also developed. Lam Y. and Zhang Y.L. (2003) have proposed a mentenance model for a intrinsec deteriorating system. Based on this model Lam. Y. has developed a mentenance model for a deteriorating system subjected to an random external damage. We consider the model by Lam. Y (2007) with functioning times and repair times independent identically distributed quasi generalized Lindley and not only.

Keywords: quasi generalized Lindley distribution, hazard function, mentenance model

AMS subject classifications: 60E15, 62F03, 62F10, 62P30

*Corresponding author: irina_adrianna@yahoo.com

The Weighted Log-Lindley Distribution and Its Applications to Lifetime Data Modeling

Bogdan Corneliu Biolan*

University of Bucharest, Doctoral School of Mathematics, Bucharest, Romania

Abstract: Modeling and analyzing data represent important issues in many applied sciences, including engineering, finance, actuarial science and medicine. The relevance and effectiveness of the methods used in statistical research are determined by the probability distribution used for modeling real data. The need to solve problems involving a large range of real data sets conducted to the development of various classes of new probability distributions.

Recently, a lot of distributions for modeling and analyzing data sets have been proposed. However, researchers often face a lot of critical situations when real data does not follow any of the existent probability distributions. Beside these, when data are recorded according to a certain stochastic model, the recorded observations will have the original distribution if and only if equal chance of being recorded is given to every observation. Biased data arise in all domains of science. Often, sampling units cannot be selected with equal probability for statistical studies. The importance of using weighted distributions arises in such kind of situations. Among the solutions for bias correction, weighted distribution theory gives a unified approach for modeling the biased data.

In this paper we introduce the new family of Weighted Log-Lindley distribution, which represents an extension of the Log-Lindley distribution. Its mathematical properties will be studied, including moments, quantile and generating functions, order statistics, Kullback-Leiber divergence and Shannon entropy. The inference with respect to the initial model will be studied in order to compare the performance between the new model and the Log-Lindley distribution in terms of adequacy for data modeling. Maximum likelihood estimators for the new model will be derived and compared with the estimators corresponding to the original model and to other related distributions. Also some applications will be developed, regarding stochastic dominance and Fisher information matrix. The conclusions drawn indicate that the Weighted Log-Lindley distribution represents a more flexible family, with powerful statistical performances for modeling a large range of data sets.

Keywords: weighted Log-Lindley distribution, stochastic ordering, Log-Lindley distribution, data modeling, insurance

AMS subject classifications: 60E15, 62F03, 62F10, 62P05

Acknowledgements: This paper has been financially supported withinthe project entitled Programe doctorale si postdoctorale - suport pentru cresterea competitivitatii cercetarii n domeniul Stiintelor exacte, contract number POSDRU/159/1.5/S/

*Corresponding author: bbiolanc@yahoo.com

137750. This project is co-financed by European Social Fund through Sectoral Operational Programme for Human Resources Development 2007-2013. Investing in people!

Partial Least Squares
A New Statistical Insight through Orthogonal Polynomials

Mélanie Blazère*[1], Fabrice Gamboa and Jean-Michel Loubes[1]

[1]*Institut de mathématiques de Toulouse, France*

Abstract: Partial Least Square (PLS) is nowadays a widely used dimension reduction technique in multivariate regression, especially when the explanatory variables are highly collinear or when they outnumber the observations. Originally designed to remove the problem of multicollinearity in the set of explanatory variables, PLS acts as a dimension reduction method by creating orthogonal latent components that maximize the variance and are also optimal for predicting the output variable. If the PLS method proved helpful in a large variety of situations (especially in chemical engineering and genetics), this iterative procedure is complex and still little is known about its theoretical properties. In this paper, we present a new approach (based on the connections between PLS and orthogonal polynomials) to analyse some statistical aspects of this method. First, we present the PLS method as it was initially introduced. Then, we explain the link between PLS and some specific discrete orthogonal polynomials, that we refer to as the residual polynomials. Thanks to the theory of orthogonal polynomials, we then derive an explicit analytical expression for the residual polynomials that clearly shows how the PLS estimator depends on the signal and noise. Based on this approach, new results are stated for the empirical risk and the mean square prediction error. The shrinkage properties of the PLS estimator are also investigated. At last, we show how this new approach, through polynomials, provides a unified framework to easily recover most of the already known PLS properties.

Keywords: Partial Least Squares regression, orthogonal polynomials, empirical risk, mean squares prediction error, shrinkage properties

AMS subject classifications: 62J05, 62J07, 62H12

1 Introduction

In this talk, I will present to you a new approach for PLS. If the PLS statistical properties are not fully understood, it is mainly because the PLS estimator depends in a non linear and complicated way on the response. Our work has mainly consisted in finding an explicit expression (with respect to the noise and to the eigenelements of the design matrix) of the dependency function that links the PLS estimator to the response. [1]. Then, we have taken advantage of this work to bring new elements in the study of the statistical properties of this estimator [2].

*Corresponding author: melanie.blazere@math.univ-toulouse.fr

2 Framework

2.1 The model

We consider the following regression model

$$Y = X\beta^* + \varepsilon \tag{1}$$

where $Y \in \mathbb{R}^n$ denotes the response, $X \in \mathbb{M}_{n\times p}$ is the design matrix, $\beta^* \in \mathbb{R}^p$ is the unknown target paremeter and $\varepsilon \in \mathbb{R}^n$ represents the noise. We allow p to be larger than n and we denote by r the rank of X^TX.

2.2 An important tool: the singular value decomposition

The singular value decomposition of X is given by

$$X = UDV^T$$

where the columns of $U \in \mathbb{M}_{n,n}$, denoted by $u_1, ..., u_p$, form an orthonormal basis of $\mathbb{R}^n$ and those of $V \in \mathbb{M}_{p,p}$, denoted by $v_1, ..., v_p$, an orthonormal basis of $\mathbb{R}^p$. The matrix $D \in \mathbb{M}_{n,p}$ contains $(\sqrt{\lambda_1}, ..., \sqrt{\lambda_r})$ on the diagonal and zero anywhere else. Without loss of generality, we assume that $\lambda_1 \geq \lambda_2 \geq .... \geq \lambda_r > 0$.

We define $\widetilde{\varepsilon}_i := \varepsilon^T u_i,\ i = 1, ..., n$ and $\widetilde{\beta}_i^* := {\beta^*}^T v_i,\ i = 1, ..., p$. The two following quantities are important and will appear many time in this talk.

1. $p_i := (X\beta^*)^T u_i,\ i = 1, ..., n.$
2. $\widehat{p}_i := Y^T u_i,\ i = 1, ..., n.$

2.3 The PLS method

The PLS method [3] at step k (where $k \leqslant r$) consists in finding $(w_k)_{1\leq k\leq K}$ and $(t_k)_{1\leq k\leq K}$ that maximise $\left[\mathrm{Cov}(Y, Xw_k)\right]^2$ under the constraints

$$\|w_k\|^2 = 1 \ \text{ and } \ t_k = Xw_k \text{ orthogonal to } t_1, ..., t_{k-1}.$$

The PLS estimator at step k denoted by $\widehat{\beta}_k$ is given by regressing Y on $t_1, ...t_k$.

In this paper, we do not consider the sequential construction of the PLS components. We rather use that PLS is the minimization of least squares over some Krylov subspaces.

Proposition 1. *[4].*
For $1 \leq k \leq r$, we have

$$\widehat{\beta}_k = \underset{\beta\in\mathcal{K}^k(X^TX, X^TY)}{argmin} \|Y - X\beta\|^2$$

where $\mathcal{K}^k(X^TX, X^TY) = \left\{X^TY, (X^TX)X^TY, ..., (X^TX)^{k-1}X^TY\right\}$.

We refer to [5, 6, 7] and to [8] for an overview of important known results on PLS.

3 Link between PLS and orthogonal polynomials

For every $k \in \mathbb{N}$, we denote by $\mathcal{P}_k$ the set of the polynomials of degree less than k and by $\mathcal{P}_{k,1}$ the set of the polynomial in $\mathcal{P}_k$ whose constant term equals 1.

3.1 PLS, a minimization problem over polynomials

Proposition 2 below shows that $\widehat{\beta}_k$ is of the form $\widehat{P}_k(X^TX)X^TY$, where $\widehat{P}_k \in \mathcal{P}_{k-1}$ is a kind of polynomial regularization of the inverse of X^TX.

Proposition 2. *[1]*
Let $k \leq r$. We have

$$\widehat{\beta}_k = \widehat{P}_k(X^TX)X^TY$$

where $\widehat{P}_k \in \mathcal{P}_{k-1}$ satisfies

$$\widehat{P}_k \in \underset{P \in \mathcal{P}_{k-1}}{argmin} \|Y - XP(X^TX)X^TY\|^2$$

and

$$\|Y - X\widehat{\beta}_k\|^2 = \|\widehat{Q}_k(XX^T)Y\|^2$$

where $\widehat{Q}_k(t) = 1 - t\widehat{P}_k(t) \in \mathcal{P}_{k,1}$ satisfies

$$\widehat{Q}_k \in \underset{Q \in \mathcal{P}_{k,1}}{argmin} \|Q(XX^T)Y\|^2.$$

The polynomials $\widehat{Q}_k$ are called the residual polynomials.

3.2 The residual polynomials

The sequence of residual polynomials $\left(\widehat{Q}_k\right)_{0 \leqslant k \leq r}$ is orthogonal with respect to a discrete measure.

Proposition 3. *[1]*
$\widehat{Q}_0 := 1, \widehat{Q}_1, ..., \widehat{Q}_r$ are orthogonal polynomials with respect to the discrete measure

$$d\widehat{\mu} = \sum_{i=1}^{r} \lambda_i \widehat{p}_i^2 \delta_{\lambda_i}.$$

4 Main result: an explicit analytical expression of the residual polynomials

We are now able to establish an explicit and exact formulation for the residual polynomials. This expression clearly shows how the disturbance on the observations and the distribution of the eigenelements impact on the residuals.

Theorem 1. *[1]*
Let $k \leq r$ and $I_k^+ = \{(j_1, ..., j_k) : r \geq j_1 > ... > j_k \geq 1\}$. We have

$$\widehat{Q}_k(x) = \sum_{(j_1,..,j_k)\in I_k^+} \left[\widehat{w}_{(j_1,..,j_k)} \prod_{l=1}^{k} (1 - \frac{x}{\lambda_{j_l}}) \right] \tag{2}$$

where

$$\widehat{w}_{j_1,..,j_k} := \frac{\widehat{p}_{j_1}^2 ... \widehat{p}_{j_k}^2 \lambda_{j_1}^2 ... \lambda_{j_k}^2 V(\lambda_{j_1}, ..., \lambda_{j_k})^2}{\sum_{(j_1,..,j_k)\in I_k^+} \widehat{p}_{j_1}^2 ... \widehat{p}_{j_k}^2 \lambda_{j_1}^2 ... \lambda_{j_k}^2 V(\lambda_{j_1}, ..., \lambda_{j_k})^2}.$$

$V(\lambda_{j_1}, ..., \lambda_{j_k})$ denotes the Vandermonde determinant associated to $\lambda_{j_1}, ..., \lambda_{j_k}$.

During the presentation, I will explain and give an interpretation of this formula. This formula is called the representation formula.

5 Application to the study of the PLS statistical properties

In this section, we further explore the statistical properties of PLS. We will see how well suited is the representation formula to the study of the PLS properties.

5.1 Approximation properties

We provide below a new expression for the empirical risk in terms of the eigenelements of X and of the noise on the observations.

Theorem 2. *[2]*
For $k < r$

$$\| Y - X\widehat{\beta}_k \|^2 =$$

$$\sum_{r>j_1>...>j_k\geq 1} \left[\widehat{w}_{j_1,..,j_k} \sum_{i=j_1+1}^{r} \left(\prod_{l=1}^{k} \left(1 - \frac{\lambda_i}{\lambda_{j_l}} \right)^2 \widehat{p}_i^2 \right) \right] + \sum_{i=r+1}^{n} \widehat{p}_i^2. \tag{3}$$

where by convention $\sum_{i=r+1}^{n} \widehat{p}_i^2 = 0$ si $r \geq n$

Proposition 4. *[2]*
Let $k < r$.

$$\| Y - X\widehat{\beta}_k \|^2 \leq \left(1 - \frac{\lambda_n}{\lambda_1} \right)^{2k} \sum_{i=k+1}^{r} \widehat{p}_i^2 + \sum_{i=r+1}^{n} \widehat{p}_i^2.$$

It should be noticed that if $\frac{\lambda_r}{\lambda_k} > 1 - \delta$ then $\sum_{i=k+1}^{r} \left[\prod_{l=1}^{k} \left(1 - \frac{\lambda_i}{\lambda_l} \right)^2 \widehat{p}_i^2 \right] \leq \delta \sum_{i=k+1}^{r} \widehat{p}_i^2$.

Furthermore, Proposition 4 allows to easily prove that PLS shrinks the residual faster than principal components regression (PCR), in the sense that

$$\| Y - X\widehat{\beta}_k \|^2 < \sum_{i=k+1}^{n} \widehat{p}_i^2 := \| Y - X\widehat{\beta}_{PCR}^k \|^2 .$$

5.2 Prediction properties

In this section, we investigate the predictive properties of the PLS estimator through the study of the mean squares prediction error.

5.2.1 A new MSPE decomposition

We study the PLS Mean Squares Prediction Error (MSPE) defined as

$$MSPE(\widehat{\beta}_k) := \mathbb{E}\left[\| X\beta^* - X\widehat{\beta}_k \|^2\right].$$

Proposition 5 below provides an interesting decomposition of $\| X\beta^* - X\widehat{\beta}_k \|^2$.

Proposition 5. *[2]*

$$\| X\beta^* - X\widehat{\beta}_k \|^2 = \sum_{i=1}^{r} \widehat{Q}_k(\lambda_i) p_i^2 + \sum_{i=1}^{r} \left(1 - \widehat{Q}_k(\lambda_i)\right) \widetilde{\varepsilon}_i^2. \tag{4}$$

It should be noticed that $\widehat{\beta}_k = \sum_{i=1}^{r} \left(1 - \widehat{Q}_k(\lambda_i)\right) \frac{\widehat{p}_i}{\sqrt{\lambda_i}} v_i$, so that the PLS estimator can be viewed as a shrinkage estimator. However, the PLS filter factors are random and not always in $[0, 1]$. In this talk, I will explain why an expansion in some of the eigendirections does not necessarily lead to an increase of the MSPE in case of PLS, by looking into details at Proposition 5,.

5.2.2 An upper bound for the MSPE under a low variance of the noise

Here, we aim at having a control of $\frac{1}{n} \| X\beta^* - X\widehat{\beta}_k \|^2$. The real variables $\varepsilon_1, ..., \varepsilon_n$ are assumed to be unobservable i.i.d centered gaussian random variables with common variance σ_n^2 and it is assumed that

- (H.1): $\sigma_n^2 = \mathcal{O}(\frac{1}{n})$ and (H.2): $\min_{1 \le i \le n} \{p_i^2\} \ge L_n := \frac{\log n}{n}$.

We get the following theorem

Theorem 3. *[1]*
Let $k \le r$. Assume (H.1) and (H.2).
With probability larger than $1 - n^{1-C}$ where $C > 1$, we have

$$\frac{1}{n} \| X\beta^* - X\widehat{\beta}_k \|^2 \le \frac{1}{n} \left(1 - \frac{\lambda_n}{\lambda_1}\right)^{2k} \sum_{i=k+1}^{r} p_i^2 + \frac{\log(n)}{n^2} \sum_{i=1}^{n} |1 - Q_k^*(\lambda_i)| + A.\frac{k}{n} \sqrt{\frac{\log n}{n L_n}} \sum_{i=1}^{n} \left[\max_{I_k^+} \left(\prod_{l=1}^{k} \left| \frac{\lambda_i}{\lambda_{j_l}} - 1 \right| \right)^2 p_i^2 \right],$$

where $A > 0$ is a constant and Q_k^* is the noise-free version of $\widehat{Q}_k$.

During the presentation, I will go into the details of the main ideas that enable to get this result.

References

[1] Blazère, M., Gamboa, F., and Loubes, J.-M. (2014). Pls: a new statistical insight through the prism of orthogonal polynomials. *arXiv preprint arXiv:1405.5900*.

[2] Blazère, M., Gamboa, F., and Loubes, J.-M. (2014). A unified framework for the study of the PLS estimator's properties. *arXiv preprint arXiv:1411.0229*.

[3] Wold, S., Ruhe, A., Wold, H., and Dunn, III, W. (1984). The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, 5(3):735–743.

[4] Helland, I. S. (1988). On the structure of partial least squares regression. *Communications in statistics-Simulation and Computation*, 17(2):581–607.

[5] Butler, N. A. and Denham, M. C. (2000). The peculiar shrinkage properties of partial least squares regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(3):585–593.

[6] Lingjaerde, O. C. and Christophersen, N. (2000). Shrinkage structure of partial least squares. *Scandinavian Journal of Statistics*, 27(3):459–473.

[7] Phatak, A. and de Hoog, F. (2002). Exploiting the connection between pls, lanczos methods and conjugate gradients: alternative proofs of some properties of pls. *Journal of Chemometrics*, 16(7):361–367.

[8] Rosipal, R. and Krämer, N. (2006). Overview and recent advances in partial least squares. In *Subspace, Latent Structure and Feature Selection*, pages 34–51. Springer.

The Selection of Relevant Groups of Explanatory Variables in GWA Studies

Damian Brzyski[*12]

[1] *Departments of Computer Science and Mathematics, Wrocław University of Technology, Poland*
[2] *Faculty of Mathematics and Computer Science, Jagiellonian University, Kraków, Poland*

Genome-wide association studies (GWAS) have become increasingly popular in investigating the genetic factors of human disease as well as studying the phenotype-genotype associations. Technological progress and development in the area of mathematical methods for large-scale data analysis allow to use these type of studies in a growing class of problems. One of the most common approaches is the quantitative trait locus (QTL) mapping which rely on the identification of regions on the genome which influence a quantitative phenotype data, e.g. height, biomarker concentrations or gene expression. The standard is to store genotype data in the form of matrix $X \in M(n,p)$, where p is the number of all considered regions and n represents the number of all individuals involved in the study. Under such circumstances, QTL maping is most commonly presented as a problem of explanatory variables selection in model $Y = X\beta + z$, for $z \sim N(0, \sigma I)$ being stochastic error and Y being the random variable indicating quantitative trait for which vector of observation (phenotype data), y, is given. Here β is vector of interest and the task is to find estimator of this parameter having relatively small number of nonzero coefficients (i.e. sparse solution) which corresponds to selection of relevant variables. According to the widely spread practice we assume that matrix is centered (sum of the elements of each column is equal to 0) and normalized so as each column has unit l_2-norm.

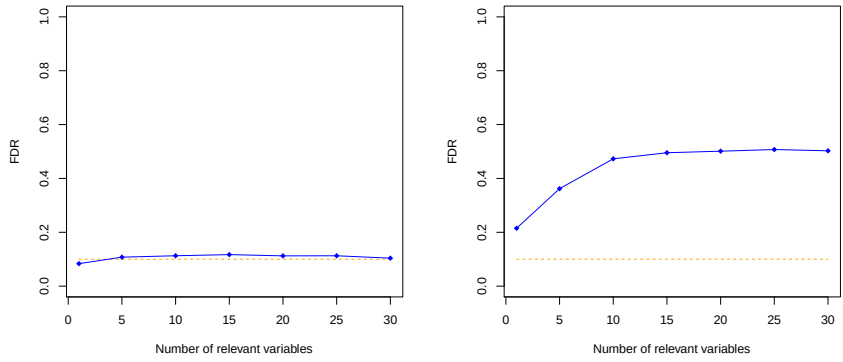
The aim of the project was to cluster explanatory variables into a groups and propose the method for relevant groups estimation (i.e. finding groups containing some response-related regressors) in such way, that the *group false discovery rate* (gFDR), defined as the expected proportion of truly irrelevant groups in all discovered groups, could be controlled below the assumed level.

To achieve this goal we will use the *group SLOPE* method, which defines estimate of relevant groups based on convex optimization problem in which information about data structure (clusters) is contained. Group SLOPE is an generalization of group LASSO [3], [4], [5] and it reduces to SLOPE [1], when size of each group is equal to one. The idea of control the fraction of falsely selected groups differs our method from others approaches, in which penalized methods are used to choose entire groups of predictors [6], [7], [8].

In single-variable-in-group scenario the authors of [1] showed that SLOPE can successfully control gFDR when correlations between variables are weak (under such scenario gFDR reduces to *false discovery rate*, FDR). In Figure 1 we present

*Corresponding author: damian.brzyski@uj.edu.pl

estimated FDR in case when entries of design matrix come from standard normal distribution (data were centered and normalized), and we compare this with GWAS data (in both cases $n = p = 1000$, target FDR level is equal to 0.1 and the same starting parameters are used). As it can be observed FDR grows rapidly beyond assumed level in right-hand side figure which is induced by specific structure of genetic data involving strongly correlated predictors which often are statistically indistinguishable. This effect shows the need for developing new method, which could be applied in the GWAS context, and provides direct motivation to this project.

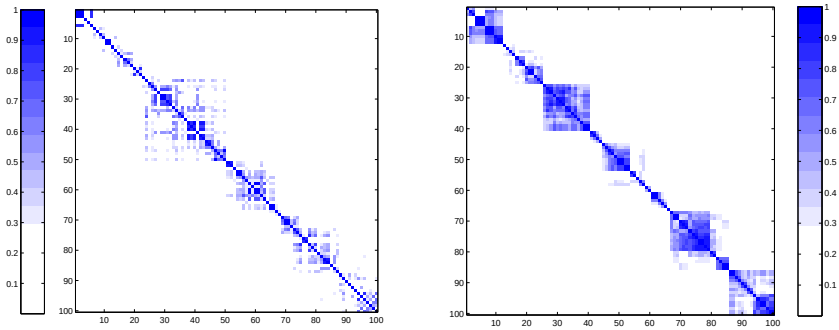


(a) Normally distributed entries (b) GWAS data

Figure 1: Estimated FDR for SLOPE

The method we used, group SLOPE, has been designed to transfer the SLOPE false discovery rate control property into groups level in situation when correlations between variables, included to different clusters, are weak. Under such circumstances, grouping data based on the strength of the linearity between them is natural procedure. Design matrices in GWA studies are structured in specific ways. There appears tendency of strong correlation between nearly located columns while columns of distant indices are generally weakly correlated. The dependency is not obvious, however, and groups of strongly correlated covariates are mixed rather than located one after another. To "disentangle" the explanatory variables, i.e. to permute them and cut into blocks corresponding to the various clusters, an algorithm based on hierarchical clustering (HCA) [2] was used. HCA clusters data using the similarity matrix which, in considered case, was defined based on correlation matrix. In Figure 2 we present heatmaps of correlations matrices (absolute values) for original data (2a) and for disentangled variables (2b). Clearly noticeable, block diagonal structure of the latter shows that data were properly clustered.

Suppose that $I = \{I_1, \ldots, I_m\}$ is some partition of set $\{1, \ldots, p\}$ defining the division of predictors into groups. Let l_i be the number of elements in group i for $i = 1, \ldots, m$. For a given data matrix, $X \in M(n, p)$, we will consider the linear



(a) Original (b) Disentangled

Figure 2: Heat maps of 100 rows and columns of correlation matrix (absolute values)

regression model with m groups of form

$$Y = \sum_{i=1}^{m} X_{I_i}\beta_{I_i} + z, \tag{1}$$

where $z \sim \mathcal{N}(0, \sigma\mathbf{I}_n)$. Here, the task is to identify groups containing at least one relevant variable or, equivalently, find the support of $\|\beta_I\|_2 := \left(\|\beta_{I_1}\|_2, \ldots, \|\beta_{I_1}\|_2\right)^T$.

Let L be diagonal, m by m matrix such as $L_{i,i} = l_i$ for $i = 1, \ldots, m$. For given sequence of nonincreasing, nonnegative starting parameters $\lambda_1, \ldots, \lambda_m$ we consider gSLOPE estimate, defined as solution to

$$\arg\min_{b} \quad \left\{\frac{1}{2}\left\|y - \sum_{i=1}^{m} X_{I_i} b_{I_i}\right\|_2^2 + \sigma J_\lambda\left(L^{\frac{1}{2}}\|b_I\|_2\right)\right\}, \tag{gSLOPE}$$

where J_λ is a norm defined as $J_\lambda(b) := \sum_{i=1}^{m} \lambda_i |b|_{(i)}$ for $|b|_{(i)}$ denoting the ith largest magnitude of b. For β_{gS} being the solution to above optimization problem, we define the estimate of $\|\beta_I\|_2$ support by the indices corresponding to nonzeros of $\|(\beta_{gS})_I\|_2$.

Naturally the performance of group SLOPE in the context of gFDR control is strongly influenced by the starting parameters. This issue will be discussed and concrete choice of these parameters, depending on assumed at the beginning gFDR level, will be proposed.

Keywords: group SLOPE, group LASSO, variables clustering, model selection, relevant SNPs

AMS subject classifications: 62J07, 62H30

Acknowledgements:

- This research received funding from European Union's 7th Framework Programme for research, technological development and demonstration under Grant Agreement no 602552.

- This article uses data from Northern Finland Birth Cohort 1966 (NFBC1966).

References

[1] M. Bogdan, R. van den Berg, W. Su, E. J. Candes. Statistical Estimation and Testing via the Ordered ℓ_1 Norm.

[2] Johnson S. C. Hierarchical clustering schemes. *Psychometrika*, Vol. 32, 241–254, 1967.

[3] Yuan M., Lin Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society*, Series B, 68(1):49-67, 2007.

[4] Yuan M., Lin Y. Consistent group selection in high-dimensional linear regression. *Bernoulli*, 16(4):1369-1384, 2007.

[5] Noah Simon, Jerome Friedman, Trevor Hastie, Rob Tibshirani A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 2013.

[6] By Peng Zhao, Guilherme Rocha, Bin Yu The composite absolute penalties family for grouped and hierarchical variable selection. *Annals of Statistics*, 37(6A):3468-3497, 2009.

[7] Howard D. Bondell, Brian J. Reich Simultaneous regression shrinkage, variable selection and clustering of predictors with OSCAR. *Biometrics*, 64(1):115–23, 2008.

[8] Guillaume Obozinski, Ben Taskar, Michael I. Jordan Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, 20(2):231–252, 2010.

Experience with Linear Programming for Experimental Design

Katarína Burclová[*1] **and Andrej Pázman**[1]

[1]*Department of Applied Mathematics and Statistics, Faculty of Mathematics, Physics and Informatics, Comenius University, Bratislava, Slovakia*

Abstract: In [4, 3] is proposed a method for computing optimal experimental design via linear programming (LP) as a modification of method of cutting planes [2]. We extend these results to a larger set of optimality criteria. The main idea is to rewrite the concave, positive homogeneous optimality criterion ϕ in a form:

$$\phi(\xi) = \min_{\mu \in \Xi} \sum_{x \in \mathcal{X}} H(\mu, x)\xi(x), \tag{1}$$

with a given function $H(\cdot;\cdot)$, $\xi \in \Xi$, the set of all probability measures on $\mathcal{X}$, where $\mathcal{X}$ is supposed to be finite design space. This reformulation allows us to interpret the problem of finding optimal design $\xi^* = \arg\max_{\xi \in \Xi} \phi(\xi)$ by the iterative algorithm which solves an LP problem at each iteration. For the criteria of D- and A-optimality and for the class of E_k criteria we obtained the required formula (1) using standard algebraic relations.

The proposed algorithm contains a stopping rule, but also the standard stopping rules following from the equivalence theorem can be used. The chief advantage of the algorithm is the possibility of combining optimality criteria or adding some supplementary (cost) constraints linear in ξ. By modifying the algorithm we can easily compute e.g. a D-optimal design under the condition that the A-optimality criterion attains a prescribed value a. Moreover, the computationally difficult problem of the "criterion robust" design (cf. [1]) can be approached by the LP method.

Keywords: optimum design, optimality criteria, cutting-plane method, cost constraints

AMS subject classifications: 62K05, 90C05

References

[1] R. Harman. Minimal efficiency of designs under the class of orthogonally invariant information criteria. *Metrika*, 60:137–153, 2004.

[2] J. Kelley. The cutting plane method for solving convex programs. *Journal of the Society for Industrial and Applied Mathematics*, 8(4):703–712, 1960.

[3] A. Pázman, and L. Pronzato. Optimum design accounting for the global nonlinear behavior of the model *Annals of Statistics*, 42(4):1426–1451, 2014.

*Corresponding author: katarina.burclova@fmph.uniba.sk

[4] L. Pronzato, and A. Pázman. *Design of Experiments in Nonlinear Models. Asymptotic Normality, Optimality Criteria and Small-Sample Properties.* Springer, 2013.

Propensity Score Matching with Clustered Data

Massimo Cannas[*1] **and Bruno Arpino**[2]

[1] *University of Cagliari, Italy*
[2] *Pompeu Fabra University, Barcelona, Spain*

Abstract: This paper focuses on the implementation of propensity score matching for clustered data. Different approaches to reduce bias due to cluster-level confounders are considered and compared using Monte Carlo simulations. We investigated methods that exploit the clustered structure of data in two ways: in the estimation of the propensity score model (through the inclusion of fixed or random effects) or in the implementation of the matching algorithm. In addition to a pure within-cluster matching, we also assessed the performance of a "preferential" within-cluster matching. This approach first searches for control units to be matched to treated units within the same cluster. If matching is not possible within-cluster, then the algorithm searches in other clusters. The preferential within-cluster matching approach, combining the advantages of within- and between-cluster matching, showed a relatively good performance both in the presence of big and small clusters and it was often t he best method.

Keywords: causal inference, hierarchical data, propensity score matching

1 Introduction

In observational studies, direct comparison of outcomes across treatment groups can give rise to biased estimates because groups being compared may be different due to lack of randomization. Subjects with certain characteristics may have higher probabilities than others to be exposed to the treatment. If these characteristics are also related to the outcome under investigation, an unadjusted comparison of the groups is likely to produce wrong conclusions about the treatment effect.

Propensity scores, defined as the probability to receive the treatment conditional on the set of observed variables, were introduced by Rosenbaum and Rubin [4] as a one-dimensional summary of the multidimensional set of covariates, such that when the propensity scores are balanced across the treatment and control groups, the distribution of all covariates are balanced across the two groups. In this way the problem of adjusting for a multivariate set of observed characteristics reduces to adjusting for the one-dimensional propensity score. (See Austin [1] for a review on the use of propensity score methods in the medical literature).

In this paper, we focus on propensity score matching and consider different approaches to take into account the clustered structure of the data with the aim of reducing the bias due to cluster-level confounders. We consider methods that exploit the information on the clusters to which units belong in two ways: in the

*Corresponding author: massimo.cannas@unica.it

estimation of the propensity score model *via* the inclusion of fixed or random effects; in the implementation of the matching algorithm.

When clusters sizes are big enough, within-cluster matching is a valid strategy but it can still imply the lost of many units that cannot find a match because the search is forced to be within clusters [2]. Discarding unmatched units is problematic because it may imply a change of the estimand [3]. In addition to a pure within-cluster matching, we also propose and assess the performance of an approach that has not been tested in previous studies. This approach first searches for control units to be matched to treated units within the same cluster. If matching is not possible within-cluster, then the algorithm searches in other clusters. This approach, that we define 'preferential' within-cluster matching, is expected to carry the benefits of pure within-cluster matching (in terms of bias reduction) and matching on the pooled dataset (in terms of minimizing the number of unmatched units).

2 Background

Consider a two-level data structure where N individual-level units, indexed by i ($i = 1, 2, ..., n_j$), are nested in J second-level units (clusters), indexed by j ($j = 1, 2, ..., J$). We consider a binary treatment administered at the individual level, T, and an outcome variable, Y also measured at the individual level. Confounders can be first (X) or second-level (Z) variables.

Usually, the Average Treatment effect on the Treated (ATT) is considered as an interesting summary of individual causal effects: $ATT = E(Y_{ij}(1) - Y_{ij}(0) \,|\, T_{ij} = 1)$. To identify the ATT with observational data, the following assumptions are often invoked:

- Unconfoundedness: $Y(1), Y(0) \perp T | (X, Z)$;
- Overlap: $0 < P(T = 1 \,|\, (X, Z)) < 1$.

Under the previous assumptions, adjustment on the propensity score is sufficient to eliminate bias due to observed confounders [4]. The propensity score, e, is defined for each unit as the probability to receive the treatment given its covariate values: $e_{ij} = P(T_{ij} = 1 \,|\, (X_{ij}, Z_j))$. Rosenbaum and Rubin [4] proved that the propensity score is a balancing score, i.e., $(X, Z) \perp T \,|\, e(X, Z)$, meaning that at each value of the propensity score the distribution of the covariates defining the propensity score is the same in the treated and control groups. They also showed that if unconfoundedness holds conditioning on covariates it also holds conditioning on the propensity score, i.e., $Y(1), Y(0) \perp T \,|\, e(X, Z)$. These results justify adjustment on the propensity score instead of on the full multivariate set of covariates.

Usually, in observational studies the propensity score is not known and must be estimated from the data. Parametric models, such as logit or probit models, with inclusion of interactions and higher order terms are commonly used. An incorrectly estimated propensity score may fail to respect the balancing property. Our focus is not on misspecification of the functional form of the propensity score model but on the bias caused by omitted cluster-level confounders. If one or more variables

affecting the selection into treatment and potential outcomes are not observed, then unconfoundedness is violated and ATT estimators based on the propensity score will be biased. the foll owing can be adapted when some observed cluster-level variables are observed and others are not.

Among propensity score methods available to adjust for an unbalanced distribution of covariates between treated and control groups, we consider propensity score matching (PSM). In particular, we consider one-to-one nearest neighbor matching within a maximum distance (caliper) of 0.20 standard deviations of the estimated propensity score. For each treated unit in the sample, the algorithm searches for the closest control unit in terms of propensity score. If no control unit is available in the range defined by the caliper, the treated unit is discarded from the working sample. We considered matching with replacement, where the same control unit can be used several times as a match. Matching with replacement is expected to improve the quality of matches and therefore to reduce bias [8]. However, a bias-variance trade-off emerges because matching with replacement increases variance of estimates [9]. Since our main focus is on the bias of the estimators we considered matching with replacement.

When the dataset has a 2-level structure one can consider different ways of implementing PSM. The methods we compare are as follows:

A) Single-level propensity score model; matching on the pooled dataset;

B) Single-level propensity score model; within-cluster matching;

C) Single-level propensity score model; preferential within-cluster matching;

D) Random-effects propensity score model; matching on the pooled dataset;

E) Fixed-effects propensity score model; matching on the pooled dataset.

3 Simulation scenario

In this section we describe our simulation experiments aimed at comparing the performance of the different matching strategies described above in the presence of unobserved confounders at the cluster-level.

3.1 Set-up

We designed our simulation experiments to mimic the observed data in several respects. First, we kept the same data structure observed in our dataset, i.e. the same number of clusters (hospitals) and the same clusters' sample sizes (see Table 1). In this way, in our simulations we consider a realistic case with a strongly unbalanced structure where some clusters are big and others have small sample sizes. Second, instead of generating values of covariates as realizations of random variables as typically done in simulation studies, we used the same covariates distribution as observed in the dataset. The only exception was for a cluster-level variable, Z, that we introduce to explore the confounding effect at the cluster (i.e., hospital) level. Finally, the coefficients of individual-level covariates in the true

models generating the treatment and the outcome were set to values similar to observed coefficients estimated on the real data.

Given the complete set of covariates (X, Z) the probability of being treated was generated according to:

$$e_{ij} = 1/[1 + exp(\beta_0 + \beta_1 X_{1ij} + \cdots + \beta_k X_{kij} + \beta_{k+1} Z_j)] \quad (1)$$

and the outcome was generated by the following model:

$$P(Y_{ij} = 1) = 1/[1 + exp(\gamma_0 + \gamma_1 X_{1ij} + \cdots + \gamma_k X_{kij} + \gamma_{k+1} Z_j + \alpha T_{ij})], \quad (2)$$

where $\boldsymbol{\beta} = [\beta_0, \cdots \beta_{k+1}]$ and $\boldsymbol{\gamma} = [\gamma_0, \cdots \gamma_{k+1}]$ are the vectors of coefficients, $X = (X_1, \cdots X_k)$ is the set of observed individual-level confounders and Z is the cluster-level confounder. Values of Z are generated as realizations of a normal variable with $\mu_Z = 0$ and $\sigma_Z = 0.25$, which is equal to the average standard deviation of the observed confounders.

Under each scenario, 500 datasets were generated from models (1) and (2). For each simulated dataset we employed the PSM methods described in the previous section to obtain a matched subset. The simulation experiments were implemented in R [10]. In particular, for methods A, D and E we obtained the matched subsets using the function `Match` in the package `Matching` [7]. At the time of writing neither this package nor others have an option for implementing within-cluster (B) and preferential within-cluster matching (C) so we programmed a routine that makes use of the `Match` function (the code is available from the authors upon request).

We summarized the results by averaging over the 500 replicates the following metrics calculated on each dataset: the number and the percentage of unmatched treated units, the absolute standardized bias (ASB) of each confounder, the estimated treatment effect ($\widehat{ATT}$), the percent bias of the estimated effect ($\%BIAS$) and the squared error (SE).

4 Results

Table 1 presents the results of the baseline simulation study introduced in the previous section. We considered three scenarios by varying the effect of the hospital-level unobserved confounder in the true treatment assignment model, that is $\beta_Z = \{0.2, 0.4, 0.8\}$. For each scenario, we compare the performance of the five PSM strategies described in section 3 (A-E) in terms of unmatched units, balance (ASB), percent bias and mean squared error (MSE). We also report in the first column the results obtained without any adjustment ("no matching").

In general, we notice that an unadjusted comparison between treated and control groups' outcomes gives strongly biased estimates (relative bias ranging from

57% to 66%). On the other hand, PSM methods guarantee a considerable reduction of the bias that tends to increase as the effect of Z increases. However, PSM methods that take clustering into account (B, C, D and E) achieve a lower bias.

Method C performs particularly well when the effect of Z is low. Otherwise, the performance of methods B and C is quite similar in terms of relative bias, but method C has the advantage of reducing the number of unmatched treated units compared to method B. The pure within-cluster matching, in fact, discards on average about 55 units (corresponding to about 1% of the treated units) as compared to less than 1 treated unit that, on average, remains unmatched with method C. Finally, we notice that there is no substantial difference with respect to the variability of ATT estimates as measured by the MSE.

Metrics	Strategy					
	No matching	A	B	C	D	E
$\beta_Z = 0.2$						
No. unmatched units	0.00	0.62	53.10	0.62	0.82	0.71
% unmatched units	0.00	0.01	0.90	0.01	0.01	0.01
ASB Z	17.90	18.49	0.00	0.25	0.88	1.23
ASB X	13.01	0.95	1.64	1.63	0.93	0.94
ASB All	13.28	1.93	1.55	1.55	0.92	0.96
% Bias	57.42	9.05	3.67	0.61	8.36	8.80
SE	0.0065	0.0035	0.0035	0.0033	0.0034	0.0025
$\beta_Z = 0.4$						
No. unmatched units	0.00	0.81	55.80	0.81	1.45	1.65
% unmatched units	0.00	0.01	0.93	0.01	0.02	0.03
ASB Z	35.72	36.32	0.00	0.35	0.83	0.95
ASB X	12.89	1.04	1.72	1.69	0.99	1.01
ASB All	14.16	3.00	1.62	1.62	0.98	1.00
% Bias	62.76	17.85	2.96	2.62	8.02	8.35
SE	0.0070	0.0038	0.0035	0.0037	0.0036	0.0036
$\beta_Z = 0.6$						
No. unmatched units	0.00	0.93	60.96	0.94	0.84	0.87
% unmatched units	0.00	0.01	0.10	0.01	0.01	0.01
ASB Z	53.03	53.47	0.00	0.62	0.78	0.79
ASB X	12.75	1.15	1.93	1.90	1.08	1.09
ASB All	15.00	4.06	1.83	1.83	1.06	1.07
% Bias	65.88	24.24	2.28	3.78	8.72	7.78
SE	0.0075	0.0042	0.0036	0.0038	0.0037	0.0037

Table 1: Simulation results after propensity score matching with replacement.

Acknowledgements: We would like to thank the Autonomous Region of Sardinia for providing the anonymized data used in the empirical application.

References

[1] Austin PC. A critical appraisal of propensity score matching in the medical literature between 1996 and 2003. *Statistics in medicine*, 27(12): 2037-2049, 2008.

[2] Gayat E, Thabut G, Christie JD, Mebazaa A, Mary JY and R Porcher. Within-center matching performed better when using propensity score matching to analyze multicenter survival data: empirical and Monte Carlo studies. *Journal of clinical epidemiology*, 66(9): 1029-1037, 2013.

[3] Crump RK, Hotz VJ, Imbens GW, Mitnik O. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96:187-195, 2009.

[4] PR, Rosenbaum, and DB Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70: 41-55, 1982.

[5] Arpino B, Mealli F. The specification of the propensity score in multilevel observational studies. *Computational Statistics and Data Analysis*, 55: 1770 -1780, 2011.

[6] Thoemmes FJ, West SG. The use of propensity scores for nonrandomized designs with clustered data. *Multivariate Behavioral Research*, 46(3): 514-543, 2011.

[7] Sekhon JS. Multivariate and Propensity Score Matching Software with Automated Balance Optimization. *Journal of Statistical Software*, 42(7): 1-52, 2011.

[8] EA Stuart Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1): 1, 2010.

[9] Caliendo M, Kopeinig S. Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys* 2008; **22(1)**: 31-72.

[10] R Core Team (2014). *R: A language and environment for statistical computing. R Foundation for Statistical Computing.* URL http://www.R-project.org/.

Estimating Whole Brain Dynamics Using Spectral Clustering

Ivor Cribben[1] **and Yi Yu**[*2]

[1]*Department of Finance and Statistical Analysis, Alberta School of Business, Canada*
[2]*Statistical Laboratory, University of Cambridge, UK*

Abstract: Recently, in functional magnetic resonance imaging (fMRI), there has been an increased interest in quantifying changes in connectivity between brain regions over an experimental time course to provide a deeper insight into the fundamental properties of brain networks. The application of graphical and network modelling has been instrumental in these analyses and has enabled the examination of the brain as an integrated system. In this work, we propose a new statistical method to provide important insights into the time-varying nature of the connectivity of brain regions while subjects are at rest. The novel method uses spectral clustering to study the network structure between brain regions and uses a non-parametric metric to detect the change in the structures across time course. The new method allows for situations where the number of brain regions is greater than the number of time points in the experimental time course ($n < p$). This method promises to offer deeper insight into the inner workings of the whole brain. We apply this new method to simulated data and to a resting-state fMRI data set. The temporal features of this novel connectivity method will provide a more accurate understanding of the large-scale characterisations of brain disorders such as Alzheimers disease and may lead to better diagnostic and prognostic indicators.

*Corresponding author: y.yu@statslab.cam.ac.uk

On the Estimation of the Central Core of a Set. Algorithms to Estimate the λ-Medial Axis

Antonio Cuevas[1], **Pamela Llop**[2] **and Beatriz Pateiro-López**[*3]

[1]*Universidad Autónoma de Madrid, Spain*
[2]*Facultad de Ingeniería Química (UNL) and Instituto de Matemática Aplicada del Litoral (UNL - CONICET), Argentina*
[3]*Universidad de Santiago de Compostela, Spain*

Abstract: The medial axis and the inner parallel body of a set C in the Euclidean space are different formal translations for the notions of the "central core" and the "bulk", respectively, of C. On the basis of their applications in image analysis, both notions (and especially the first one) have been extensively studied in the literature, from different points of view. A modified version of the medial axis, called λ-medial axis, has been recently proposed in order to get better stability properties. The estimation of these relevant subsets from a random sample of points is a partially open problem which has been considered only very recently. Our aim is to show that standard, relatively simple, techniques of set estimation can provide natural, consistent, easy-to-implement estimators for both the λ-medial axis $\mathcal{M}_\lambda(C)$ and the inner parallel body $I_\lambda(C)$ of C. The consistency of these estimators follows from two results of stability (i.e. continuity in the Hausdorff metric) of $\mathcal{M}_\lambda(C)$ and $I_\lambda(C)$ obtained under some, not too restrictive, regularity assumptions on C. As a consequence, natural algorithms for the approximation of the λ-medial axis and the λ-inner parallel body can be derived. The whole approach could be useful for some practical problems in image analysis where estimation techniques are needed.

Keywords: medial axis, set estimation, r-convexity

AMS subject classifications: 62G05

1 Introduction

There is a rich mathematical literature devoted to the study of the "central part" of a set C in the Euclidean space (which would represent, in statistical terms, the "median of C"); see [1]. Of course, the first step in any such study must be to give a precise meaning to this loose notion of "set median". Different definitions, closely related but not always equivalent, have been proposed. The most popular one is perhaps the *medial axis* of C, $\mathcal{M}(C)$, defined as the subset of points in C having at least two projections on the boundary ∂C. Other closely related (but not equivalent) usual notions are the *skeleton* , $\mathcal{S}(C)$, (the set of centers of maximal balls included in C) and the *cut locus* of C, defined as the topological closure of $\mathcal{M}(C)$; see below for further discussion on these notions. The medial axis was

*Corresponding author: beatriz.pateiro@usc.es

introduced by [2] as a tool in image analysis. The papers by [4], [3] and [5], among many others, analyze these ideas from different points of view.

We are especially concerned with a modified version of the medial axis, called *λ-medial axis*, $\mathcal{M}_\lambda(C)$, introduced in [3] to deal with the well-known problem of instability in the medial axis: the medial axis $\mathcal{M}(C)$ and $\mathcal{M}(D)$ might be far away from each other even if the original sets C and D and their boundaries are very close together; see [4] and references therein. The λ-medial axis leaves out those points of $\mathcal{M}(C)$ whose metric projections on ∂C are too close together.

Another, perhaps less popular, closely related concept is the so-called *λ-inner parallel body*, $I_\lambda(C)$, defined as the set of points in C whose distance to ∂C is at least λ. So far this concept has been mainly studied in the case where C is convex [see, e.g., [8]] but we will see that this assumption is not necessary to find a simple consistent estimator of $I_\lambda(C)$. The λ-inner parallel body has a simple intuitive interpretation and is obviously close to the notion of "core" of C. In some cases it provides an outer approximation to the λ-medial axis. The algorithm proposed in the paper by [7] for medial axis estimation (under a regression-type sampling model) relies on an estimate of the inner parallel set.

This work deals with the statistical problem of estimating the λ-medial axis, $\mathcal{M}_\lambda(C)$, and the λ-inner parallel body, $I_\lambda(C)$, from a random sample of points $X_1, \ldots, X_n$ drawn inside C. The whole approach relies on a simple plug-in idea: we will use methods of set estimation (see, e.g., [6] for a survey) to get a suitable estimator $C_n = C_n(X_1, \ldots, X_n)$ of C. Then the natural estimators of $\mathcal{M}_\lambda(C)$ and $I_\lambda(C)$ would be just $\mathcal{M}_\lambda(C_n)$ and $I_\lambda(C_n)$, respectively.

Whereas the theoretical and practical aspects of the medial axis (and associated notions) have received a considerable attention, the problem of estimating this set has been considered only very recently: we refer to the recent paper by [7], though the sampling model considered by these authors is a bit different to that we will use here. In short, we will show that imposing an additional shape restriction on C (called *r-convexity*) one can obtain, in return, a considerable simplification in the theory and practice of the estimation of $\mathcal{M}_\lambda(C)$ and $I_\lambda(C)$.

Acknowledgements: This work has been partially supported by Spanish Grant MTM2013-41383P from Ministry of Economy and Competitiveness, European Regional Development Fund (ERDF) and the IAP research network grant no. P6/03 from the Belgian government (third author).

References

[1] Attali, D., Boissonnat, J.-D. and Edelsbrunner, H. Stability and Computation of Medial Axis: a State-of-the-Art Report. In *Mathematical Foundations of Scientific Visualization, Computer Graphics, and Massive Data Exploration*, T. Moeller, B. Hamann and R. Russell, eds., pp 109–125. Springer-Verlag, Berlin, 2009.

[2] Blum, H. A Transformation for Extracting New Descriptors of Shape. In W. Wathen-Dunn, editor, *Models for the Perception of Speech and Visual Form*, Cambridge, MA, MIT Press, 362–380, 1967.

[3] Chazal, F. and Lieutier, A. The "λ-medial Axis". *J. Graphical Models*, 67:304–331, 2005.

[4] Chazal, F. and Soufflet, R. Stability and Finiteness Properties of Medial Axis and Skeleton. *J. Dynam. Control Systems*, 10:149–170, 2004.

[5] Chaussard, J., Couprie, M and Talbot, H. Robust Skeletonization Using the Discrete λ-medial Axis. *Pattern Recognition Letters*, 32:1384–1394, 2011.

[6] Cuevas, A. and Fraiman, R. Set Estimation. In *New Perspectives on Stochastic Geometry*, W.S. Kendall and I. Molchanov, eds., pp. 374–397. Oxford University Press, 2010.

[7] Genovese, C.R., Perone-Pacifico, M., Verdinelli, I. and Wasserman, L. The Geometry of Nonparametric Filament Estimation. *J. Amer. Statist. Assoc.*, 107:788–799, 2012.

[8] Sangwine-Yager, J.R. Bonnesen-style Inequalities for Minkowski Relative Geometry. *Trans. Amer. Math. Soc.*, 307:373–382, 1988.

Model Fitting for Space-Time Point Patterns Using Projection Processes

Jiří Dvořák[*12]

[1]*Department of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University in Prague, Czech Rep.*
[2]*Department of Image Processing, Institute of Information Theory and Automation, The Czech Academy of Sciences, Czech Rep.*

Abstract: Statistical inference for spatial and space-time point patterns recorded e.g. in ecological or epidemiological applications represents a challenging task. The data consists of a random collection of points $\{u_1, \dots, u_N\}$ observed in a compact observation window W, or, in the space-time setting, a random collection of space-time events $\{(u_1, t_1), \dots, (u_N, t_N)\}$ observed in W over a time period T. Both the number of observed points N and their locations are random.

We focus our attention on the space-time clustered point patterns. We discuss the possibility to use dimension-reduction techniques to fit different parts of a space-time model separately. Specifically, we define the projections of the process to the spatial and temporal domain, respectively, and introduce a step-wise estimation procedure based on these projections. We also discuss the problem of possible cluster overlapping and the resulting loss of information in the projections and the challenges it presents for parameter estimation.

Keywords: space-time point process, Cox process, minimum contrast estimation, K-function, projection process

AMS subject classifications: 62M30, 60G55

1 Introduction

Many fields of science deal with data that are point patterns, such as positions of trees in a rain forest, maps of disease cases or the locations of point-like defects in industrial materials. A point pattern consists of a finite set of points $\{u_1, \dots, u_N\}$ observed in a compact observation window $W \subset \mathbb{R}^d$. Figure 1 provides an example of different point patterns. A random process generating such point patterns is called a *point process*. Formally speaking, a point process X in $\mathbb{R}^d$ is a measurable mapping from an abstract probability space to the space of locally finite subsets of $\mathbb{R}^d$. This implies that the number N of points of $X \cap W$ is a random variable. Also, the locations of the points are random.

The necessity to analyse point patterns arising in practical applications lead to significant development of the point process theory in the past decades. For a comprehensive review see [2, 4] or [5]. Parametric point process models enable detailed statistical inference and hence are sought for in practice.

[*]Corresponding author: dvorak@karlin.mff.cuni.cz

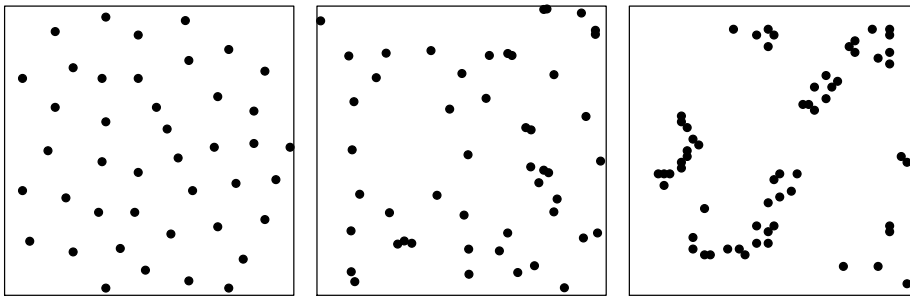


Figure 1: Examples of different point patterns. *Left:* the locations of the centers of 42 biological cells observed under optical microscopy in a histological section. *Middle:* simulated realization of a Poisson point process (see the text). *Right:* the locations of 62 seedlings and saplings of California redwood trees in a square sampling region. The two real point patterns are standard datasets available in the `spatstat` package [1] for `R`.

The basic point process model is the Poisson process. Consider a non-negative function $\lambda(u)$, $u \in \mathbb{R}^d$. The point process X is called the *Poisson process with intensity function* λ if the number of points of X in a Borel set $B \subset \mathbb{R}^d$ has a Poisson distribution with mean $\int_B \lambda(u)\,\mathrm{d}u$ and the point counts in disjoint Borel sets are independent random variables. This model exhibits no interaction between points. For a sample realization see the middle panel of Figure 1.

A natural generalization of the Poisson process is the *Cox process*. Let $\Lambda(u)$, $u \in \mathbb{R}^d$, be a non-negative random field such that $u \mapsto \Lambda(u)$ is a locally integrable function with probability 1. If the conditional distribution of the point process X given $\Lambda = \lambda$ is the distribution of the Poisson process with intensity function λ, X is said to be the Cox process with the driving field Λ [4]. A Cox process is the model of choice when it comes to analysis of clustered point patterns (such as the one in the right panel of Figure 1) as it is able to accomodate the attractive interactions among the points. However, Cox process models are not suitable for modelling regular patterns such as the one in the left panel of Figure 1.

In this contribution we will focus on the problem of model fitting for clustered space-time point patterns. In this case each point (sometimes also called event) of the space-time point process X has a spatial position $u \in \mathbb{R}^d$ and a temporal coordinate $t \in \mathbb{R}$. In the following we will use $(u,t) \in X$ to denote that a point of X occurs at location u at time t. We assume that the data is observed in a compact set $W \times T \subset \mathbb{R}^d \times \mathbb{R}$ with positive Lebesgue measure where W is a spatial region observed over the time period T.

We emphasize that we consider a continuous time domain T (not discrete time instances) and that the events have no duration. Hence $(u,t) \in X$ might e.g. in an epidemiological application correspond to an animal kept on a farm with location u being *reported* at time t to be infected.

Furthermore, we remark that analysis of space-time point patterns observed in

a compact subset of $\mathbb{R}^d \times \mathbb{R}$ requires dedicated methods. It is not appropriate to use methods for spatial point patterns in $\mathbb{R}^{d+1}$ – the temporal coordinate plays a distinct role and cannot be interchanged with the spatial coordinates. For example, it is not natural to measure the difference between two space-time events using the Eucliedean norm in $\mathbb{R}^{d+1}$.

2 Background

For ease of exposition we restrict our attention to stationary point processes, i.e. the processes with translation invariant distribution. At first we present the necessary definitions in the spatial setting and then extend them to the space-time setting. We also assume that the point process in question is *simple*, i.e. two points cannot occur at the same location. For further details [2] can be consulted.

Let X be a stationary point process in $\mathbb{R}^d$. If B is a Borel set in $\mathbb{R}^d$ we denote by $|B|$ its volume and by $X(B)$ the number of points of X in B. The *intensity* ρ of X is defined as the expected number of points of X in a set of unit volume, i.e. $\rho = \mathbb{E}X(B)/|B|$ for B with positive volume and ρ does not depend on the choice of B. We assume $\rho > 0$.

Furthermore, let $\mathrm{d}u$ be an infinitesimal region containing the point $u \in \mathbb{R}^d$. The *second-order intensity function* of X is defined by

$$\rho_2(u,v) = \lim_{|\mathrm{d}u|,|\mathrm{d}v|\to 0} \frac{\mathbb{E}\left[X(\mathrm{d}u)X(\mathrm{d}v)\right]}{|\mathrm{d}u|\,|\mathrm{d}v|}.$$

For $u \neq v$, $\rho_2(u,v)\,|\mathrm{d}u|\,|\mathrm{d}v|$ can be regarded as the approximate probability that $\mathrm{d}u$ and $\mathrm{d}v$ both contain a point of X. From the assumption of stationarity of X we see that, with a slight abuse of notation which is common in the field of spatial statistics, $\rho_2(u,v) = \rho_2(0, v-u) = \rho_2(v-u)$.

Useful point process characteristics are the *pair-correlation function* $g(u) = \rho_2(u)/\rho^2$, $u \in \mathbb{R}^d$, and the *K-function* defined by

$$K(r) = \int_{B(o,r)} g(w)\,\mathrm{d}w,\ r > 0,$$

where $B(o,r)$ is the ball centered at the origin o with radius r. We remark that the value $\rho K(r)$ can also be interpreted as the mean number of points from X in $B(o,r) \setminus \{o\}$ provided that there is a point of X in o.

Now we can extend our definitions to the space-time setting. We keep track of the temporal coordinate of points by explicitely writing $(u,t) \in \mathbb{R}^d \times \mathbb{R}$. It is straightforward to define the space-time intensity, second-order intensity function and the space-time pair-correlation function in the same way as above. Now $\rho_2((u,t),(v,s)) = \rho_2(v-u, s-t) = \rho^2 g(v-u, s-t)$.

The definition of the space-time K-function must take into account the principal difference between the spatial and temporal coordinates. Hence we define

$$K(r,t) = \int_{B(o,r)} \int_{-t}^{t} g(w,\tau)\,\mathrm{d}\tau\,\mathrm{d}w,\ r > 0,\ t > 0,$$

as a function of two arguments – the spatial distance r and the temporal lag t. Now $\rho K(r,t)$ can be interpreted as the mean number of points from X in the cylinder $(B(o,r)\times[-t,t])\setminus\{(o,o)\}$ provided that there is a point of X in (o,o).

3 Model fitting

While the extent of this contribution does not allow us to specify the model in detail, we have in mind the *shot-noise Cox processes* [4]. They constitute a wide and flexible class of models for clustered point patterns. In what follows it is crucial that the model allows the separation of the first-order parameters (intensity $\rho>0$) and interaction parameters (possibly a vector parameter ψ of the pair-correlation function $g(u,t;\psi)$). The model is fully parametrized by (ρ,ψ).

For a general Cox point process it is very difficult to obtain maximum likelihood estimate of the model parameters. The reason is that the likelihood involves an expectation of a complicated integral term with respect to the distribution of the random driving field Λ. One can of course take advantage of MCMC or other techniques and use approximations of the likelihood function [4]. This approach is usually computationally very demanding and thus faster, simulation-free alternatives based on moment properties are preferred.

For a space-time shot-noise Cox process there is an explicit expression for the space-time K-function $K(r,t;\psi)$. The non-parametric estimate of the space-time K-function is

$$\widehat{K}(r,t)=\frac{1}{(\widehat{\rho})^2\,|W\times T|}\sum_{(u_i,t_i),(u_j,t_j)\in X\cap(W\times T)}^{\neq}\frac{I(\|u_i-u_j\|\le r,|t_i-t_j|\le t)}{w_1(u_i,u_j)\,w_2(t_i,t_j)},$$

where I is the indicator function, $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^d$, w_1,w_2 are the spatial and temporal edge-correction factors [3], $\widehat{\rho}=X(W\times T)/|W\times T|$ is the natural non-parametric estimate of the intensity ρ and the summation is over distinct pairs of points of X.

Now we can use the classical minimum contrast method to find the estimate of the interaction parameters ψ by minimizing the discrepancy

$$\int_{r_{min}}^{r_{max}}\int_{t_{min}}^{t_{max}}(\widehat{K}(r,t)^q-K(r,t;\psi)^q)^2\,\mathrm{d}t\,\mathrm{d}r,$$

for some bounds $0<r_{min}<r_{max}$ and $0<t_{min}<t_{max}$. Here q is a variance stabilizing coefficient adjusting for the non-constant variance of $\widehat{K}(r,t)$. Typical values of q are $1/4$ or $1/2$.

However, this approach requires rather large amount of data in order to obtain stable estimates of $\widehat{K}(r,t)$. Based on the idea used in [3] we proposed in [6] to define the projections of the space-time process X to the spatial and temporal domain, respectively, and use these projections for estimation. We define

$$X_s=\{u:(u,t)\in X,t\in T\},$$
$$X_t=\{t:(u,t)\in X,u\in W\},$$

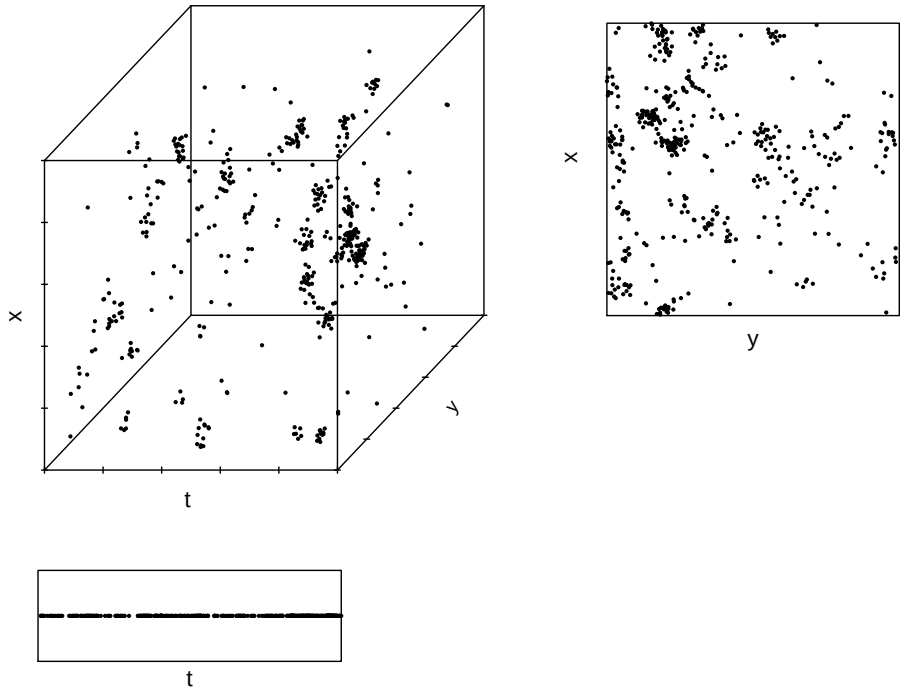


Figure 2: A sample realization of a space-time point process observed in $W \times T = [0,1]^2 \times [0,1]$ (upper left) together with the corresponding spatial projection (upper right) and temporal projection (below).

i.e. when defining the spatial projection process X_s we keep only the points occuring in the time period T and then disregard their temporal coordinate. Similarly for the temporal projection process X_t. For a graphical illustration see Figure 2.

In [6] we consider space-time shot-noise Cox processes and we require an additional assumption of a certain type of second-order separability. It in fact specifies that the (vector) interaction parameter $\psi = (\psi_s, \psi_t)$ involves a spatial interaction parameter ψ_s and a temporal interaction parameter ψ_t and that the space-time K-function $K(r,t)$ of the process X depends in a separable way on the spatial part (depending only on ψ_s and r) and the temporal part (depending only on ψ_t and t).

Under this assumption it is possible to derive explicit formulae for the K-function $K_s(r;\psi_s)$ of the projection process X_s and $K_t(\tau;\psi_t)$ of X_t, respectively. Then K_s and K_t can be estimated non-parametrically and we can use minimum contrast estimation based on K_s in order to estimate ψ_s and, similarly, minimum contrast estimation based on K_t in order to estimate ψ_t. In this way the interaction parameters are estimated separately and, more importantly, the non-parametric estimates of K_s and K_t are more stable because by projecting we have reduced the dimension of the corresponding space.

4 Challenges

An interesting problem is caused by possible overlapping of clusters of points in the data. Even clusters which were originally clearly separated in the space-time domain may overlap in the spatial or temporal projection, making it more difficult to estimate the interaction parameters. This problem is more pronounced in the temporal projection process where we project from $\mathbb{R}^d \times \mathbb{R}$ to $\mathbb{R}$ only.

A consequence of the cluster overlapping problem is that increasing the amount of observed data may not in general result in more precise estimates of the interaction parameters. Consider for example a fixed spatial domain W and an increasing sequence of time intervals $T_n = [0, n]$. If we use T_n to define a sequence of spatial projection processes $X_s^{(n)}$, the intensity $\rho_s^{(n)}$ of $X_s^{(n)}$ would increase unboundedly. This implies more and more cluster overlapping in $X_s^{(n)}$. It can be proved formally that in the limit the influence of the spatial interaction parameter ψ_s is lost and it cannot be identified from K_s. The situation is analogous if we consider a fixed time period T and an increasing sequence of spatial regions W_n.

This makes it difficult to find appropriate asymptotic regime for this estimation problem and to formulate conditions under which consistency and asymptotic normality of the resulting estimators hold. We conclude this contribution by a remark that it is in fact possible to formulate asymptotic results but different normalization (by $|W_n|^{1/2}$ or $|T_n|^{1/2}$) is required for $\widehat{\psi_s}$ and $\widehat{\psi_t}$, respectively.

Acknowledgements: The author would like to express his gratitude to his former supervisor RNDr. Michaela Prokešová, Ph.D., for her never-ending enthusiasm, patience and support.

References

[1] A. Baddeley and R. Turner. Spatstat: An R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12(6):1–42, 2005.

[2] J. Illian, A. Penttinen, H. Stoyan, and D. Stoyan. *Statistical Analysis and Modelling of Spatial Point Patterns*. Wiley, 2008.

[3] J. Møller and M. Ghorbani. Aspects of second-order analysis of structured inhomogeneous spatio-temporal point processes. *Statistica Neerlandica*, 66(4):472–491, 2012.

[4] J. Møller and R. P. Waagepetersen. *Statistical inference and simulation for spatial point processes*. Chapman & Hall/CRC, 2004.

[5] J. Møller and R. P. Waagepetersen. Modern statistics for spatial point processes. *Scandinavian Journal of Statistics*, 34(4):643–684, 2007.

[6] M. Prokešová and J. Dvořák. Statistics for inhomogeneous space-time shot-noise Cox processes. *Methodology and Computing in Applied Probability*, 16(2):433–449, 2014.

The Evolving Evolutionary Spectrum

Mark Fiecas[*1] **and Hernando Ombao**[2]

[1]*Department of Statistics, University of Warwick, Coventry CV4 7AL, United Kingdom*
[2]*Department of Statistics, University of California at Irvine, Bren Hall 2206, Irvine, CA 92697*

Abstract: In this work, we developed a novel time series model in the context of designed experiments that captures two sources of nonstationarities: 1) over time within a trial of the experiment and 2) across the trials of the experiment. Under the proposed model we construct spectral measures that change with respect to both sources of nonstationarities. To estimate the evolving evolutionary spectral density matrix, we used a two-stage procedure. In the first stage, we computed the within-trial time-localized periodogram matrix. In the second stage, we developed a data-driven approach for combining information across trials from the local periodogram matrices. We assessed the performance of our proposed method using simulated data.

Keywords: multivariate time series, cross-coherence, local stationarity, spectral analysis

AMS subject classifications: 62M10, 62P10

1 Introduction

The context of our work is in neuroscience experiments where nonstationary time series data are collected. Often in the experiment a stimulus is presented numerous times, and so the data set consists of a collection of many nonstationary multivariate time series across the trials of the experiment. However, existing methodology for spectral analysis of nonstationary time series data will assume that the data across the trials are identical realizations of the same underlying process. This is not necessarily the case in practice. Indeed, the variability across trials may arise from the subject changing performance strategy, reduced neuronal activity potentially caused by habituation to the stimulus, or other neurophysiological processes [1, 2]. While many statistical models take into account nonstationarity within a single trial, there is certainly a need for statistical methodologies to account for the nonstationarity over the course of the experiment in order to give a more accurate characterization of the underlying dynamics of the data.

*Corresponding author: m.fiecas@warwick.ac.uk

2 The evolving evolutionary spectrum

We begin with our statistical model that captures the two sources of nonstationarities. Our modeling framework is inspired by the work of Dahlhaus (2000) [3] on locally stationary processes (LSP) but we go further by modeling the dynamics across an *entire experiment* rather than just within a single trial. The following model was first developed by Fiecas and Ombao (2014) and used to study local field potentials collected during an associative learning experiment [4]. We refer the reader to that work for more details and discussion.

Definition 1. A sequence of locally stationary zero-mean P-variate time series $\mathbf{X}_{t,r}$ where $r = 1, \ldots, R$ denotes the trials within the entire experiment and $t = 1, \ldots, T$ denote the time points within a trial is said to follow a *slowly evolving locally stationary process* (SEv-LSP) if it admits the representation

$$\mathbf{X}_{t,r} = \int_{-0.5}^{0.5} \mathbf{A}(t/T, r/R, \omega) \exp(i2\pi\omega t) d\mathbf{Z}_r(\omega),$$

where:

1. $\mathbf{Z}_r(\omega)$ is a zero-mean P-variate orthogonal increment process on $[-0.5, 0.5]$ with $\mathbf{Z}_r(\omega) = \overline{\mathbf{Z}_r(-\omega)}$, $\mathbf{Z}_r(\omega)$ is uncorrelated with $\mathbf{Z}_{r'}(\omega)$ for $r \neq r'$, and

$$\operatorname{cum}(d\mathbf{Z}_r(\omega_1), \ldots, d\mathbf{Z}_r(\omega_k)) = \eta(\sum_{j=1}^{k} \omega_j) \Lambda_k(\omega_1, \ldots, \omega_{k-1}) d\omega_1 \cdots d\omega_k,$$

   where $\operatorname{cum}(\cdot)$ denotes the k-th order cumulant, $\Lambda_1 = 0, \Lambda_2 = \mathrm{Id}_P$, where Id_P is the $P \times P$ identity matrix, $|\Lambda_k(\omega_1, \ldots, \omega_{k-1})| \leq C_k$ where C_k is a constant, Λ_4 is continuous, and $\eta(\omega) = \sum_{j=-\infty}^{\infty} \delta(\omega + j)$ is the periodic extension of the Dirac delta function, and

2. For each $(u, v, \omega) \in [0, 1] \times [0, 1] \times [-0.5, 0.5]$, the complex-valued transfer function denoted $\mathbf{A}(u, v, \omega)$ (of dimension $P \times P$) has continuous second partial derivatives with respect to u, v, and ω, and $\partial^l \mathbf{A}(u, v, \omega)/\partial\omega^l = \overline{\partial^l \mathbf{A}(u, v, -\omega)/\partial\omega^l}$ for $l = 0, 1$.

Definition 2. The *evolving evolutionary spectral density matrix*, with dimension $P \times P$, defined at rescaled time $u \in [0, 1]$ within rescaled trial-time $v \in [0, 1]$ and at frequency $\omega \in [-0.5, 0.5]$ is

$$\mathbf{f}(u, v, \omega) = \mathbf{A}(u, v, \omega) \mathbf{A}(u, v, \omega)^*,$$

where (*) denotes the conjugate transpose.

In spectral analysis, a popular approach for investigating the linear dependence between two time series is via cross-coherence analysis. Cross-coherence is analogous to cross-correlation but is frequency specific, i.e., it measures the strength of linear association between two time series at a particular frequency.

Definition 3. The *evolving evolutionary coherence* between dimensions p and q of a SEv-LSP $\mathbf{X}_{t,r}$ is

$$\rho^2_{pq}(u,v,\omega) = \left| \frac{f(u,v,\omega)_{pq}}{\sqrt{f(u,v,\omega)_{pp} f(u,v,\omega)_{qq}}} \right|^2,$$

where $f(u,v,\omega)_{pq}$ is the (p,q)-th element of $\mathbf{f}(u,v,\omega)$.

Our estimation approach begins with a trial-specific periodogram matrix that is localized in time within a trial. Let $\{u_j\}_{j=1}^B$ be an increasing sequence in $(0,1)$ and $\epsilon \in (0,1)$ such that we have a collection of subintervals of rescaled time $[u_j - \epsilon, u_j + \epsilon] \subset [0,1]$. Next let $\mathcal{B}_j = \{[(u_j - \epsilon)T], \ldots, [(u_j + \epsilon)T]\}$, where $[\cdot]$ is the greatest integer function, be the j-th time block having midpoint $[u_j T]$ and size $|\mathcal{B}_j| = [2\epsilon T]$. Without loss of generality, we assume $|\mathcal{B}_j|$ to be even. To construct the r-th trial time-localized periodogram matrix, first let $\mathbf{d}_{j,r}(\omega)$ be the discrete Fourier transform of $\mathbf{X}_{t,r}$ restricted to the j-th block:

$$\mathbf{d}_{j,r}(\omega_k) = |\mathcal{B}_j|^{-1/2} \sum_{t \in \mathcal{B}_j} \mathbf{X}_{t,r} \exp(-i2\pi\omega_k t),$$

where $\omega_k = k/|\mathcal{B}_j|$, $k = -|\mathcal{B}_j|/2, \ldots, |\mathcal{B}_j|/2 - 1$, are the Fourier frequencies. Then the r-th trial time-localized periodogram matrix at rescaled time u_j and frequency ω_k is $\mathbf{I}_{j,r}(\omega_k) = \mathbf{d}_{j,r}(\omega_k)\mathbf{d}_{j,r}(\omega_k)^*$.

For any $u, v \in [0,1]$, $\mathbf{I}_{j,r}(\omega)$ is an asymptotically unbiased estimator, but not consistent, for $\mathbf{f}(u,v,\omega)$. For locally stationary time series, a popular solution is to smooth over frequencies. To obtain a consistent estimator while maintaining frequency resolution, we instead take advantage of the slow evolution of the process over the entire experiment by smoothing across trials, i.e., our estimator for the evolving evolutionary spectral density matrix is the time-localized periodogram matrix smoothed across trials, given by

$$\widehat{\mathbf{f}}^{(M_{jk})}_{j,r}(\omega_k) = (2M_{jk}+1)^{-1} \sum_{s=-M_{jk}}^{M_{jk}} \mathbf{I}_{j,r+s}(\omega_k),$$

for some positive integer M_{jk}. We pick each M_{jk}, $j = 1, \ldots, B$ and $k = -|\mathcal{B}_j|/2$, $\ldots$, $|\mathcal{B}_j|/2 - 1$ to optimize some criterion. We refer the reader to Fiecas and Ombao (2014) for a data-driven method for selecting each M_{jk} [4]. By smoothing across trials instead of over frequencies, we do not lose frequency resolution, and we will be able to investigate the evolution of the spectral properties of the data from one trial of the experiment to the next. From here, we can extract the approporiate elements of the matrix to obtain an estimator for the evolving evolutionary coherence. Under mild regularity conditions, one can show that our estimators are consistent [4].

3 Simulation study

In this simulation study we simulated bivariate time series for R trials that slowly evolved from one trial to the next. To this end, we first created a 2×2 matrix-valued

function $\mathbf{\Psi}(u, v, \omega)$, whose elements are

$$\begin{aligned}
\psi_{11}(u, v, \omega) &= (v + 0.5)[\{1.2\cos(\pi\omega)\}^2 + 0.4\sin(2\pi uT) + 0.7], \\
\psi_{21}(u, v, \omega) &= 0.6\cos(2\pi\omega) + 0.4\cos(2\pi uT) + 1 + \\
&\qquad i0.4\sin(2\pi\omega)\{4(uT - 0.5)^2 + 0.5\}, \\
\psi_{12}(u, v, \omega) &= 0 \\
\psi_{22}(u, v, \omega) &= \{1.3\cos(2\pi\omega)\}^2 + 0.4\sin(2\pi uT) + 0.8.
\end{aligned}$$

From $\mathbf{\Psi}(u, v, \omega)$, we simulated time series data with $T = 512$ or $T = 1024$, and $R = 100$ or 250 by

$$\mathbf{X}_{t,r} = \sum_{k=-T/2}^{T/2-1} \mathbf{\Psi}(t/T, r/R, k/T) \exp(i2\pi kt/T)\mathbf{Z}_r(k),$$

where $\{\mathbf{Z}_r(k), k = -T/2, \ldots, T/2 - 1,$ and $r = 1, \ldots, R\}$ are independent such that for $k/T \notin \{0, \pm 0.5\}$, the distribution of $\mathbf{Z}_r(k)$ was bivariate complex normal with zero mean and covariance matrix $T^{-1}\mathrm{Id}$, and for $k/T \in \{0, \pm 0.5\}$, the distribution of $\mathbf{Z}_r(k)$ was bivariate real normal with zero mean and covariance matrix $T^{-1}\mathrm{Id}$. For each $r = 1, \ldots, R$, $\mathbf{X}_{t,r}$ is a LSP with spectral density matrix $\mathbf{f}(t/T, r/R, k/T) = \mathbf{\Psi}(t/T, r/R, k/T)\mathbf{\Psi}(t/T, r/R, k/T)^*$ [5]. Note that the evolving evolutionary power spectrum $f_{11}(u, v, \omega)$ slowly changed over v, but the evolving evolutionary power spectrum of the second dimension $f_{22}(u, v, \omega)$ and the evolving evolutionary coherence $\rho^2_{12}(u, v, \omega)$ were constant over v.

The parameter settings in our estimation procedure to estimate the evolving evolutionary spectral properties of the simulated data were as follows. We constructed the local periodogram matrices using blocks in time that have 64 time points, and these blocks were overlapping by setting $[T(u_j - u_{j-1})] = 8$ for all blocks j. We also investigated the effects of block sizes and step sizes, and the results were similar.

We summarize the results of our simulation study by computing the mean squared error with respect to the Hilbert-Schmidt norm of the estimate of the evolving evolutionary spectral density matrix at each point on the time-frequency grid for each trial, and then averaging over each discrete time-frequency point and across the trials. These simulation results are in Table 1. The LSP model does not account for the evolution over trials, hence the high bias relative to the SEv-LSP model. On the other hand, the LSP model has lower variance because it averages over the entire experiment. The SEv-LSP model averages locally over trials, hence, the higher variance relative to the LSP model. The higher MSE of the LSP, therefore, was driven primarily by the biased estimates as a result of not accounting for the evolution of the spectral density matrix over the trials.

References

[1] A. Arieli, A. Sterkin, A. Grinvald, and A. Aertsen. Dynamics of Ongoing Activity: Explanation of the Large Variability in Evoked Cortical Responses. *Science*, 273:1868–1871, 1996.

		LSP			SEv-LSP		
R	T	Bias^2	Variance	MSE	Bias^2	Variance	MSE
100	512	4.630	0.861	5.491	1.587	2.772	4.358
	1024	4.602	0.663	5.265	1.552	2.868	4.420
250	512	4.612	0.277	4.889	1.536	1.779	3.315
	1024	4.607	0.281	4.888	1.525	1.807	3.332

Table 1: The squared bias, variance, and MSE averaged over all time-frequency points and over all trials.

[2] J.R. Duann, T.P. Jung, W.J. Kuo, T.C. Yeh, S. Makeig, J.C. Hsieh, and T. Sejnowski. Single-Trial Variability in Event-Related BOLD Signals. *NeuroImage*, 15:823–835, 2002.

[3] R. Dahlhaus. A Likelihood Approximation for Locally Stationary Processes. *The Annals of Statistics*, 28:1762–1794, 2000.

[4] M. Fiecas and H. Ombao. Modeling the evolution of dynamic brain processes during an associative learning experiment. Manuscript submitted for publication, 2014. http://works.bepress.com/mfiecas/8.

[5] W. Guo and M. Dai. Multivariate Time-Dependent Spectral Analysis Using Cholesky Decomposition. *Statistica Sinica*, 16:825–845, 2006.

Weak Rates of Approximation of Integral-Type Functionals of Markov Processes

Iurii Ganychenko[*1] **and Alexei Kulik**[2]

[1]*Department of Probability Theory, Statistics and Actuarial Mathematics, Kyiv National Taras Shevchenko University, Kyiv, Ukraine*
[2]*Institute of Mathematics, Ukrainian National Academy of Sciences, Kyiv, Ukraine*

Abstract: We study weak rates of approximations of integral functional of Markov process by integral sums. An assumption on the process is formulated only in terms of transition probability density and, therefore, current approach is not strongly dependent on the process structure. Obtained results allow to control the rate of approximations of Feynman-Kac semigroup or occupation time options price.

Keywords: weak rates, Monte-Carlo method, Feynman-Kac formula, occupation time options

AMS subject classifications: 60H07, 60H35

1 Introduction and main result

For $\mathbb{R}^d$-valued Markov process $X_t, t \geq 0$ and some function h the following objects are considered:

1) an integral-type functional

$$I_T(h) = \int_0^T h(X_t)\, dt;$$

2) an approximative sequence of integral sums

$$I_{T,n}(h) = \frac{T}{n} \sum_{k=0}^{n-1} h(X_{(kT)/n}), \quad n \geq 1.$$

In what follows, P_x denotes the law of the process X, conditioned by $X_0 = x$, E_x denotes the correspondent expectation w.r.t. this law. Both the absolute value of a real number and an Euclidean norm in $\mathbb{R}^d$ are denoted by $|\cdot|$. $\|\cdot\|$ stands for the sup-norm.

The only assumption on the process X is the following.

X. The process X possesses a transition probability density $p_t(x, y)$, which is differentiable w.r.t. t and satisfies

$$\Big|\partial_t p_t(x,y)\Big| \leq C_T t^{-1-d/\alpha} Q\Big(t^{-1/\alpha}(x-y)\Big), \quad t \leq T,\ C_T \geq 1, \tag{1}$$

with some $\alpha \in (0, 2]$ and some distribution density Q.

*Corresponding author: iurii_ganychenko@ukr.net

Remark 1. For instance, Brownian motion and multidimensional diffusions satisfy the assumption **X** for $\alpha = 2$ and $Q(x) = c_1 \exp(-c_2|x|^2)$ with some c_1, c_2. Symmetric alpha-stable process satisfies this assumption for any $\alpha \in (0, 2]$ and $Q(x) = \begin{cases} c_1 e^{-c_2|x|^2} & , \alpha = 2, \\ \frac{c}{1+|x|^{d+\alpha}}, & \alpha \in (0,2), \end{cases}$. We refer the reader to [1] for more details.

Our research is based on the following result (see Proposition 2.1 [1]).

Proposition 6. *Let assumption **X** holds and h is bounded. Then there exists a positive constant C, such that*

$$\left| E_x I_T(h) - E_x I_{T,n}(h) \right| \le C \frac{\log n}{n}. \tag{2}$$

In other words, under some conditions on the process X, the rate of approximations of expectations of a given integral functional of such a process is well-controlled. Such rates are called *weak* rates.

One can find some generalizations of Proposition 6 in recent researches. One of various approaches there is an obtaining of such a weak rate (2), but in the case, where the process X is also approximated, not only the limited functional. Such a result is obtained in [2] (see Theorem 2.5). Completely different approach to generalize the result of Proposition 6 is studied in [4] and [1]: so-called *strong* rates are obtained there, i.e. the bounds for

$$E_x \left| I_T(h) - I_{T,n}(h) \right|^p.$$

We introduce our main result below.

Theorem 4. *Let **X** hold and h is bounded. Then for any $k \in \mathbb{N}$ and bounded function f:*

$$\left| E_x (I_T(h))^k f(X_T) - E_x (I_{T,n}(h))^k f(X_T) \right| \le 6k^2 C_T T^k \|h\|^k \left(\frac{\log n}{n} \right) \cdot \|f\|.$$

Remark 2. Proposition 6 follows from Theorem 4 if $k = 1$ and $f \equiv 1$.

Therefore, Theorem 4 generalizes the result of Proposition 6 simultaneously in two ways. Firstly, we control an approximation rate of the integral functional powered by any k. Secondly, we add a test function of the last-time point.

The next corollary is quite important.

Let us consider any analytical function g, which is defined in some neighbourhood of 0, and constants $D_g, R_g > 0$, such that for any natural m : $\left| \frac{g^{(m)}(0)}{m!} \right| \le D_g \cdot \left(\frac{1}{R_g} \right)^m$.

Theorem 5. *Let **X** holds and f is such that $T\|h\| < R_g$. Then for each bounded f we have:*

$$\left| E_x g(I_T(h)) f(X_T) - E_x g(I_{T,n}(h)) f(X_T) \right| \le C_{T,h,D_g,R_g} \left(\frac{\log n}{n} \right) \cdot \|f\|, \tag{3}$$

where

$$C_{T,h,D_g,R_g} = 6 D_g C_T \frac{T\|h\|}{R_g} \cdot \left(1 + \frac{T\|h\|}{R_g} \right) \cdot \frac{1}{\left(1 - \frac{T\|h\|}{R_g} \right)^3}.$$

2 Applications

2.1 On approximation of Feynman-Kac semigroup

Let $X_\cdot \equiv Z_\cdot$ be a Brownian motion valued in $\mathbb{R}^d$. Then the assumption **X** holds for $\alpha = 2$ and $Q(x) = c_1 \exp(-c_2|x|^2)$ with some c_1, c_2 (see [1] for more details). Therefore, the transition probability density is defined by:

$$p_t(x,y) = \frac{1}{(2\pi t)^{d/2}} \exp\left(-\frac{|y-x|^2}{2t}\right), t > 0, x, y \in \mathbb{R}^d.$$

We define a semigroup in $C_b(\mathbb{R}^d)$ by the formula

$$E_x f(Z_t) := R_t f(x) = \int_{\mathbb{R}^d} p_t(x,y) f(y) dy$$

and denote its generator by $\mathcal{A}$. If additionally $f \in \mathcal{S}(\mathbb{R}^d)$ (the space of rapidly decreasing functions), then $\mathcal{A}$ is a Laplace operator.

Then the formula

$$R_t^h f(x) = E_x \left[f(Z_t) \exp\left\{ \lambda I_t(h) \right\} \right]$$

defines a semigroup on $C_b(\mathbb{R}^d)$ with its generator given in the form

$$\mathcal{A}_h f = \mathcal{A} f + \lambda h f$$

(see [6], Chapter 1).

We put

$$R_{t,n}^h f(x) = E_x \left[f(Z_t) \exp\left\{ \lambda I_{t,n}(h) \right\} \right]$$

and provide the following corollary of Theorem 4.

Theorem 6. *For each bounded functions f, h and $\lambda > 0$ the following estimate holds:*

$$\left| R_t^h f(x) - R_{t,n}^h f(x) \right| \le C_{T,\lambda,h} \left(\frac{\log n}{n} \right) \cdot \|f\|,$$

where

$$C_{T,\lambda,h} = 6 C_T \lambda \|h\| T \cdot (1 + \lambda \|h\| T) \cdot \exp\{\lambda \|h\| T\}.$$

Therefore, the main result allows to control the rate of approximations of Feynman-Kac semigroup with the accuracy $(\log n)/n$.

2.2 On approximation of the price of an occupation time option

We do not assume any smoothness conditions on the function h. Therefore, the important particular case of h being an indicator function and $I_T(h)$ being respectively an occupation time can be considered in scopes of our approach. It allows us to apply the main result to pricing the options, which price depends on the time spent by the process X in some defined set.

In what follows, we put $d = 1$.

Let the price of an asset $S = \{S_t, t \geq 0\}$ be of the form:

$$S_t = S_0 \exp(X_t).$$

Then the time spent by the process S in a set $J \subset \mathbb{R}$ (or the time spent by X in a set $J' \subset \mathbb{R}$), from time 0 to time T, is defined by

$$\int_0^T \mathbb{I}_{\{S_t \in J\}} dt = \int_0^T \mathbb{I}_{\{X_t \in J'\}} dt.$$

We consider the options, which price depends on the time that the process S spend in the set J. Such kind of options is introduced in [5] and called an *occupation time option*. Comparing to the standard barrier options, which activated or cancelled when S hits some defined level (barrier), the payoff of the occupation time option depends on a time that the process S stays below or above such a barrier.

For instance, for the strike price K, the barrier L and the knock-out rate ρ, the payoff of a down-and-out call option is given by:

$$\exp\left(-\rho \int_0^T \mathbb{I}_{\{S_t \leq L\}} dt\right) \cdot (S_T - K)_+.$$

Then its price $\mathbf{C}(\mathrm{T})$ is given by

$$\mathbf{C}(T) = \exp(-rT) E\left[\exp\left(-\rho \int_0^T \mathbb{I}_{\{S_t \leq L\}} dt\right) \cdot (S_T - K)_+\right],$$

where r is the risk-free interest rate (see [5]).

The problem of such a price estimate is solved in [3] for Lévy process with negative jumps. And the approach is strongly dependent on the process structure.

We put

$$\mathbf{C}_n(T) = \exp(-rT) E\left[\exp\left(-\rho T/n \sum_{k=0}^{n-1} \mathbb{I}_{\{S_{kT/n} \leq L\}} dt\right) \cdot (S_T - K)_+\right],$$

and provide another corollary of Theorem 4.

Theorem 7. *Let* $\boldsymbol{X}$ *holds and there exists* $u > 1$*, such that* $G := E\exp(uX_T) = ES_T^u < +\infty$*. Then*

$$\left|\boldsymbol{C}_n(T) - \boldsymbol{C}(T)\right| \leq 3\max\{C_{T,\rho,h}, G\} \exp(-rT) \left(\frac{\log n}{n^{1-1/u}}\right),$$

where $C_{T,\rho,h}$ *is the same as in Theorem 3 and* h *is an indicator function.*

Therefore, our main result allows to control the rate of approximations of $\mathbf{C}(T)$ by $\mathbf{C}_n(T)$ with the accuracy $(\log n)/n^{1-1/u}$ for the class of Markov processes which satisfy the assumption $\mathbf{X}$ with $E\exp(uX_T) < +\infty$.

References

[1] Iu. Ganychenko, A. Kulik, *Rates of approximation of nonsmooth integral-type functionals of Markov processes*, Modern Stochastics: Theory and Applications, **2** (2014), 117–126.

[2] E. Gobet, C. Labart, *Sharp estimates for the convergence of the density of the Euler scheme in small time*, Elect. Comm. in Probab., **13** (2008), 352–363.

[3] Helene Guerin, Jean-Francois Renaud, *Joint distribution of a spectrally negative Levy process and its occupation time, with step option pricing in view*, arXiv:1406.3130.

[4] A. Kohatsu-Higa, A. Makhlouf, H.L. Ngo, *Approximations of non-smooth integral type functionals of one dimensional diffusion precesses*. Stochastic Processes and their Applications, **124** (2014), issue 5, 1881–1909.

[5] V. Linetsky, *Step options*. Math. Finance, **9** (1999), no. 1, 55–96.

[6] A. Sznitman, *Brownian motion, obstacles and random media*, Springer, Berlin, 1998.

Recursive Estimation of the Median Covariation Matrix in Hilbert Spaces

Antoine Godichon*

Institut de Mathématiques de Bourgogne, Université de Bourgogne,
9 Rue Alain Savary, 21078 Dijon, France

Abstract: The geometric median, also called L^1-median is often used in statistics because of its robustness. Moreover, it is more and more usual to deal with large sample taking values in high dimensional spaces. In this context, a fast estimator of the median has been introduced by [2] which consists in an averaged stochastic gradient algorithm. We propose to give a deep study of these estimators. We also define a new robust dispersion matrix (closely related to the median) called Median Covariation Matrix as well as algorithms to estimate it. This matrix can be very interesting in robust Principal Components Analysis. Indeed, under assumptions, it has the same eigenspaces as the covariation matrix, but it is less sensitive to outliers.

Keywords: geometric median, high dimension, minimum covariance determinant, projection pursuit, recursive estimation

AMS subject classifications: 62H12, 62E17, 62E20, 62M15

1 Introduction

The geometric median, also called L^1-median, is a generalization of the real median introduced by [7]. In the multivariate case, it is closely related to the Fermat-Webber's problem, which consists in finding a point minimizing the sum of distances from given points. This is a well known convex optimization problem. Many properties of the median are given by [8] in Banach spaces, such as its existence, uniqueness and robustness. This last property is one of the principal factors of interest about the median. Moreover, [4] propose a deep study estimators of the median in the general case of Banach spaces.

Many algorithms for estimating the median exist in the literature, and one of the most used in the multivariate case, is the one introduced by [10], which consists in solving the Fermat-Webber's problem generated by the sample with Weizfeld's algorithm. Nevertheless, this last algorithm can be difficult to compute when we have to deal with large data taking values in high dimensional spaces. This is why we will focus on the algorithm introduced by [2] which has the same asymptotic distribution, and which consists in a stochastic gradient algorithm and its averaged version.

*Corresponding author: Antoine.Godichon@u-bourgogne.fr

Moreover, we will speak about the Median Covariation Matrix (MCM), which is studied by [9]. This is a robust indicator of dispersion closely related to the median, which can be used in robust PCA (see
[3] for example). Indeed, under weak conditions, if the distribution of the data we would like to study is symmetric, then the MCM has the same eigenspaces as the usual covariance matrix. As mentioned in
[3], many algorithms exist to estimate this indicator, but we will focus on a completely recursive one in order to be able to deal with big data. This algorithm consists in estimating both the median and the MCM with stochastic gradient algorithms and their averaged versions.

In this report, we first define the median and give some assumptions before recalling the algorithms introduced by [2]. We also give some results due to [1] and [6] on their asymptotic behavior. Finally, we will define the Median Covariation Matrix and recall the recursive algorithms introduced by [3], before giving their rates of convergence in quadratic mean.

2 Estimating the geometric median

2.1 Definition and assumptions

Let us consider a random variable X taking values in a separable Hilbert space H (not necessarily with a finite dimension). We denote by $\langle .,. \rangle$ its inner product and $\|.\|$ the associated norm. The geometric median m of X is defined by

$$m := \arg\min_{h \in H} \mathbb{E}\left[\|X - h\| - \|X\|\right].$$

We now introduce two assumptions:

(A1) X is not concentrated on a straight line: for all $h \in H$, there is $h' \in H$ such that $\langle h, h' \rangle = 0$ and

$$\mathrm{Var}\left(\langle X, h' \rangle\right) > 0.$$

(A2) X is not concentrated around single points: there is a positive constant C such that for all $h \in H$,

$$\mathbb{E}\left[\frac{1}{\|X - h\|}\right] \leq C.$$

Note that assumption **(A1)** ensures that the median is uniquely defined (see [8]), and assumption **(A2)** enables to give some convexity properties such as the twice differentiability of the function we would like to minimize.

2.2 The algorithms

Let G be the function we would like to minimize. It is defined for all $h \in H$ by

$$G(h) := \mathbb{E}\left[\|X - h\| - \|X\|\right].$$

An important fact is that under assumption **(A2)** G is convex and is Frchet-differentiable. Its gradient is given for all $h \in H$ by

$$\nabla G(h) = -\mathbb{E}\left[\frac{X-h}{\|X-h\|}\right].$$

This legitimates the fact to use a stochastic gradient (or Robbins-Monro) algorithm. Let us now consider independent random variables $X_1, ..., X_n, ...$ with the same law as X. We recall the Robbins-Monro algorithm introduced by [2]:

$$m_{n+1} = m_n + \gamma_n \frac{X_{n+1} - m_n}{\|X_{n+1} - m_n\|}, \tag{1}$$

with m_1 chosen bounded. The step sequence (γ_n) is a decreasing sequence of positive real numbers, and verifies the following usual conditions (see [5] for example)

$$\sum_{n\geq 1} \gamma_n = +\infty, \qquad \sum_{n\geq 1} \gamma_n^2 < \infty.$$

The averaged version of the algorithm is defined recursively by

$$\overline{m}_{n+1} = \overline{m}_n + \frac{1}{n+1}\left(m_{n+1} - \overline{m}_n\right), \tag{2}$$

with $\overline{m}_0 = 0$, which can be written as $\overline{m}_n = \frac{1}{n}\sum_{k=1}^{n} m_k$.

2.3 Rate of convergence

The strong consistency of these algorithms is given in [2]. We now consider a step sequence of the form $\gamma_n := c_\gamma n^{-\alpha}$, with $c_\gamma > 0$ and $\alpha \in (1/2, 1)$. Then, we have the following rates of convergence of the Robbins-Monro algorithm.

Theorem 8. *Suppose assumptions **(A1)** and **(A2)** are fulfilled. There is a positive constant C' such that for all $n \geq 1$,*

$$\mathbb{E}\left[\|m_n - m\|^2\right] \leq \frac{C'}{n^\alpha}.$$

More generally, for all integer $p \geq 1$, there is a positive constant K_p such that for all $n \geq 1$,

$$\mathbb{E}\left[\|m_n - m\|^{2p}\right] \leq \frac{K_p}{n^{p\alpha}}.$$

With the help of previous rates, one can obtain the following rates for the averaged algorithm.

Theorem 9. *Suppose assumptions **(A1)** and **(A2)** are fulfilled. There is a positive constant C'' such that for all $n \geq 1$,*

$$\mathbb{E}\left[\|\overline{m}_n - m\|^2\right] \leq \frac{C''}{n}.$$

More generally, for all integer $p \geq 1$, there is a positive constant $K_{p'}$ such that for all $n \geq 1$,

$$\mathbb{E}\left[\|\overline{m}_n - m\|^{2p}\right] \leq \frac{K'_p}{n^p}.$$

3 Estimating the Median Covariation Matrix

3.1 Definition and assumptions

We now consider a separable Hilbert space H and the space of linear operators mapping H to H, denoted by $\mathcal{S}(H)$. Let $(e_j)_{j\in J}$ be a basis of H, we equippe $\mathcal{S}(H)$ with the following inner product: let $A, B \in \mathcal{S}(H)$,

$$\langle A, B\rangle_F = \sum_{j\in J} \langle A(e_j), B(e_j)\rangle .$$

Thus, $\mathcal{S}(H)$ is also a separable Hilbert space and the norm associated to the previous inner product, denoted by $\|.\|_F$, is the well known Hilbert-Schmidt (or Froebenius) norm.

Let X be a random variable taking values in H, the Median Covariation Matrix Γ_m of X is defined by

$$\Gamma_m := \arg\min_{V\in\mathcal{S}(H)} \mathbb{E}\left[\left\|(X-m)^T(X-m) - V\right\|_F - \left\|(X-m)^T(X-m)\right\|_F\right], \quad (3)$$

where m is the median of X. The Median Covariation Matrix Γ_m can be seen as the geometric median of the random variable $(X-m)^T(X-m)$, and is so robust. We now introduce two assumptions:

(A3) For all $V \in \mathcal{S}(H)$, there is $V' \in \mathcal{S}(H)$ such that $\langle V, V'\rangle_H = 0$ and

$$\mathrm{Var}\left(\left\langle (X-m)^T(X-m), V'\right\rangle_F\right) > 0.$$

(A4) There is a positive constant C such that for all $h \in H$ and $V \in \mathcal{S}(H)$,

$$\mathbb{E}\left[\frac{1}{\|(X-h)^T(X-h) - V\|_F^2}\right] \le C.$$

Note that assumption **(A4)** implies assumption **(A2)**. Under assumptions **(A1)** and **(A3)**, the Median Covariation Matrix is uniquely defined.

3.2 The algorithms

In the particular case when the median m is known, the algorithms and their asymptotic properties are analogous to the ones for the estimation of the median. We consider from now that m is not known and for all $h \in H$, let G_h be the functional defined for all $V \in \mathcal{S}(H)$ by

$$G_h(V) := \mathbb{E}\left[\left\|(X-h)^T(X-h) - V\right\|_F - \left\|(X-h)^T(X-h)\right\|_F\right].$$

One can check that G_m is the function we would like to minimize. These functional are Frchet-differentiable and their gradients are defined for all $V \in \mathcal{S}(H)$ by

$$\nabla G_h(V) = -\mathbb{E}\left[\frac{(X-h)^T(X-h) - V}{\|(X-h)^T(X-h) - V\|_F}\right].$$

Then we can now introduce a "stochastic gradient algorithm" and its averaged version. Let $X_1, ..., X_n, ...$ be independent random variables with the same law as X. The stochastic gradient estimator V_n and its averaged version $\overline{V}_n$ are defined recursively by

$$
\begin{aligned}
m_{n+1} &= m_n + \gamma_n \frac{X_{n+1} - m_n}{\|X_{n+1} - m_n\|}, \\
\overline{m}_{n+1} &= \overline{m}_n + \frac{1}{n+1}\left(m_{n+1} - \overline{m}_n\right), \\
V_{n+1} &= V_n + \gamma_n \frac{(X_{n+1} - \overline{m}_n)^T (X_{n+1} - \overline{m}_n) - V_n}{\|(X_{n+1} - \overline{m}_n)^T (X_{n+1} - \overline{m}_n) - V_n\|_F}, \qquad (4) \\
\overline{V}_{n+1} &= \overline{V}_n + \frac{1}{n+1}\left(V_{n+1} - \overline{V}_n\right), \qquad (5)
\end{aligned}
$$

with m_1 and V_1 chosen bounded, $\overline{m}_0 = 0$, $\overline{V}_0 = 0$.

3.3 Rates of convergence

We now consider a step sequence of the form $\gamma_n := c_\gamma n^{-\alpha}$, with $c_\gamma > 0$ and $\alpha \in (1/2, 1)$. Then, we have the following rates of convergence for the stochastic gradient algorithm:

Theorem 10. *Suppose assumptions **(A1)** to **(A4)** hold. There is a positive constant K such that for all $n \geq 1$,*

$$\mathbb{E}\left[\|V_n - \Gamma_m\|_F^2\right] \leq \frac{K}{n^\alpha}.$$

Moreover, for all $\beta \in (\alpha, 2\alpha)$, there is a positive constant K_β such that for all $n \geq 1$,

$$\mathbb{E}\left[\|V_n - \Gamma_m\|_F^4\right] \leq \frac{K_\beta}{n^\beta}.$$

Finally, the following theorem gives the rate of convergence in quadratic mean of the averaged estimator.

Theorem 11. *Suppose assumptions **(A1)** to **(A4)** hold. There is a positive constant K' such that for all $n \geq 1$,*

$$\mathbb{E}\left[\|\overline{V}_n - \Gamma_m\|_F^2\right] \leq \frac{K'}{n}.$$

Acknowledgements: This work is a resume of three paper written in collaboration with Hervé Cardot and Peggy Cénac, the author would like to thank them for this as well as for their many advice.

References

[1] H. Cardot, P. Cénac, and A. Godichon. Online estimation of the geometric median in Hilbert spaces: non asymptotic confidence balls. Technical report, arXiv:1501.06930, 2015.

[2] H. Cardot, P. Cénac, and P.-A. Zitt. Efficient and fast estimation of the geometric median in Hilbert spaces with an averaged stochastic gradient algorithm. *Bernoulli*, 19:18–43, 2013.

[3] H. Cardot and A. Godichon. Robust principal compoents analysis based on the median covariation matrix. *arXiv:1504.02852*, 2015.

[4] A. Chakraborty and P. Chaudhuri. The spatial distribution in infinite dimensional spaces and related quantiles and depths. *The Annals of Statistics*, 42:1203–1231, 2014.

[5] M. Duflo. *Random iterative models*, volume 34 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1997. Translated from the 1990 French original by Stephen S. Wilson and revised by the author.

[6] A. Godichon. Estimating geometric median in hilbert spaces with stochastic gradient algorithms; l^p and almost sure rates of convergence. *arXiv:1504.02267*, 2015.

[7] J. B. S. Haldane. Note on the median of a multivariate distribution. *Biometrika*, 35(3-4):414–417, 1948.

[8] J. H. B. Kemperman. The median of a finite measure on a Banach space. In *Statistical data analysis based on the L_1-norm and related methods (Neuchâtel, 1987)*, pages 217–230. North-Holland, Amsterdam, 1987.

[9] D. Kraus and V. M. Panaretos. Dispersion operators and resistant second-order functional data analysis. *Biometrika*, 99:813–832, 2012.

[10] Y. Vardi and C.-H. Zhang. The multivariate L_1-median and associated data depth. *Proc. Natl. Acad. Sci. USA*, 97(4):1423–1426, 2000.

Statistical Inference for the Sparse Parameter of a Partially Linear Single-Index Model

Thomas Gueuning[*1] **and Gerda Claeskens**[1]

[1]*ORSTAT and Leuven Statistics Research Center*
KU Leuven, Faculty of Economics and Business
Naamsestraat 69, 3000 Leuven, Belgium

Abstract: We perform inference for the sparse and potentially high-dimensional parametric part of a partially linear single-index model. We construct a desparsified version of a penalized estimator for which asymptotic normality can be proven. This allows us to take the uncertainty associated with the variable selection process into account and to construct confidence intervals for all the components of the parameter.

Keywords: desparsifying, high-dimensional data, confidence intervals, penalized estimator, single-index model

AMS subject classifications: 62F12, 62F25, 62G08

1 Introduction

In the last decades, high-dimensional data have increasingly become available, leading to the development of new statistical methodologies. Such data are characterized by the number of variables being larger than the number of observations, which makes classical statistical tools such as least-squares estimation unsuitable. In order to tackle this difficulty, sparsity is often assumed, meaning that only a few variables suffice to explain the model. This sparsity assumption has led to the development of penalized regression techniques including the Lasso [6] and the SCAD [1].

Much work has been done on point estimation properties such as consistency in prediction and in variable selection, in particular for the Lasso estimator. A major drawback of the penalized estimators is that, due to their sparse structure, we can only characterize the distribution of their active set of variables. This means that the uncertainty associated with the variable selection process is completely ignored. Construction of confidence intervals is thus directly possible only for the components that were not shrunk to zero, which can lead to wrong inference. To tackle this problem, [5] construct a desparsified version of the Lasso estimator for which asymptotic normality is proven for the generalized linear model family.

In this paper we extend this idea of desparsifying the lasso to the partially linear single-index model $Y = \eta_0(Z^T\alpha) + X^T\beta + \epsilon$, where $\eta_0(\cdot)$ is a one-dimensional unknown function. This model is a natural extension of the linear model to include non-parametric effects. Compared to the additive non-parametric model, it has

*Corresponding author: thomas.gueuning@kuleuven.be

the advantage to require the estimation of only one unknown function, overcoming the so-called curse of dimensionality. Details of the usefulness of this model can be found in [3]. We describe in section 2.1 a method for the estimation of the parametric part (α, β) and of the non-parametric part $\eta_0(\cdot)$ introduced by [4]. In case of high-dimensional data or if some sparsity is expected, a penalized estimator can be obtained. In section 2.2, we show how to construct a desparsified version of the penalized estimator and state its asymptotic normality. This allows us to perform inference for all the components of the parametric part (α, β). Section 2.3 summarizes our methodology and section 3 assesses finite sample performance through a simulation study.

2 The partially linear single-index model: estimation and inference on the parametric part

Let $\{(Y_i, X_i, Z_i), i = 1, \ldots, n\}$ be a sample generated by the partially linear single-index model

$$Y = \eta_0(Z^T\alpha^0) + X^T\beta^0 + \epsilon,$$

where $(Z, X) \in \mathbb{R}^{p\times q}$ are the covariate vectors associated with the response variable Y, $\eta_0(\cdot)$ is the unknown link function of the single-index $Z^T\alpha^0$, $\epsilon \sim N(0, \sigma_\epsilon^2)$ is the error term and $\alpha^0 \in \mathbb{R}^p$ and $\beta^0 \in \mathbb{R}^q$ are unknown parameters to estimate. For identifiability reasons, we assume that $\left\|\alpha^0\right\| = 1$ and that the first non-zero entry of α^0 is positive. If $p + q > n$, the framework is high-dimensional. We use the notation $\xi^0 = (\alpha^0, \beta^0) \in \mathbb{R}^{p+q}$ for the parameter vector.

2.1 Estimation of the parameters and of the function η_0

Several estimation techniques have been introduced in the last two decades, including the backfitting algorithm. We use the profile least-squares approach introduced by [4] where the local linear regression technique is used to estimate the unknown function η_0. In concrete terms, given a value of (α, β) of the parameter vector we estimate η_0 and its derivative η_0' at a point $u \in \mathbb{R}$ as

$$(\widehat{\eta_0(u)}, \widehat{\eta_0'(u)}) = \underset{(a,b)\in\mathbb{R}^2}{\arg\min} \sum_{i=1}^{n} (a + b(Z_i^T\alpha - u) + X_i^T\beta - Y_i)^2 K_h(Z_i^T\alpha - u) \qquad (1)$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$ with $K(\cdot)$ a kernel function and h a bandwidth. The idea underlying (1) is to use the first order approximation $\eta_0(Z_i^T\alpha) \approx \eta_0(u) + \eta_0'(u)(Z_i^T\alpha - u)$ for $Z_i^T\alpha$ close to u. For every value of the parameter $\xi = (\alpha, \beta)$ we thus have an estimator $\widehat{\eta}(u, \xi)$ of $\eta_0(u)$, explicitly obtained as

$$\widehat{\eta}(u, \xi) = \widehat{a} = \frac{K_{20}(u, \xi)K_{01}(u, \xi) - K_{10}(u, \xi)K_{11}(u, \xi)}{K_{00}(u, \xi)K_{20}(u, \xi) - K_{10}^2(u, \xi)}, \qquad (2)$$

where $K_{jl}(u, \xi) = \sum_{i=1}^{n} K_h(Z_i^T\alpha - u)(Z_i^T\alpha - u)^j(X_i^T\beta - Y_i)^l$ for $j = 0, 1, 2$ and $l = 0, 1$. We obtain a profile least-squares estimator by the minimization of

$$Q(\alpha, \beta) = \frac{1}{2n} \sum_{i=1}^{n} (Y_i - \widehat{\eta}(Z_i^T\alpha; \alpha, \beta) - X_i^T\beta)^2. \qquad (3)$$

If the parameter vector (α, β) is expected to be sparse, we can add a penalty term such as used for the SCAD, the Lasso or the adaptive Lasso. We define the Lasso estimator $\widehat{\xi} = (\widehat{\alpha}, \widehat{\beta})$ as the minimizer of

$$L(\alpha, \beta) = \frac{1}{2n} \sum_{i=1}^{n} (Y_i - \widehat{\eta}(Z_i^T \alpha; \alpha, \beta) - X_i^T \beta)^2 + \lambda \|\alpha\|_1 + \lambda \|\beta\|_1 . \tag{4}$$

Note that different tuning parameters could be used to penalize differently on α and on β. It is also possible to linearize $\widehat{\eta}(Z_i^T \alpha; \alpha, \beta)$ using an initial value $(\widetilde{\alpha}, \widetilde{\beta})$ in order to make the optimization program computationally more efficient.

2.2 Inference on the parameter via a desparsifying process

With the Lasso penalty term replaced by the SCAD function, the asymptotic normality of the non-zero coefficients of the estimator obtained by the minimization of $L(\alpha, \beta)$ is proven by [4]. This allows to perform inference on the non-zero coefficients selected by the SCAD but does not take into account the variability associated with the variable selection process. Some coefficients could be wrongly shrunk to zero and no statistical inference could be done on these ones. To tackle this problem we propose to use a technique introduced by [5], which consists of desparsifying the penalized estimator. We establish the asymptotic normality of the desparsified estimator, which allows us to construct confidence intervals for all the components of the estimator, and not only for the active set of variables. We are then able to detect components which were wrongly set to zero. We now show how we construct our desparsified estimator.

Let $\widehat{\xi} = (\alpha, \beta)$ be the minimizer of (4) and let us define

$$X_{i,\widehat{\xi}} = \begin{bmatrix} \frac{\partial}{\partial \alpha} \widehat{\eta}(Z_i^T \alpha; \alpha, \beta)|_{(\widehat{\alpha}, \widehat{\beta})} \\ \frac{\partial}{\partial \beta} \widehat{\eta}(Z_i^T \alpha; \alpha, \beta)|_{(\widehat{\alpha}, \widehat{\beta})} \quad + \quad X_i \end{bmatrix}$$

and

$$\widehat{\Sigma}_{\widehat{\xi}} = \frac{1}{n} \sum_{i=1}^{n} X_{i,\widehat{\xi}} X_{i,\widehat{\xi}}^T.$$

We define $\widehat{\Theta}_{\widehat{\xi}}$ as being a relaxed inverse of $\widehat{\Sigma}_{\widehat{\xi}}$. Two ways to construct this relaxed inverse are described in [2]. We define the desparsified estimator $\widehat{\xi}^{\text{desp}}$ as follows:

$$\widehat{\xi}^{\text{desp}} = \widehat{\xi} + \widehat{\Theta}_{\widehat{\xi}} \left[\frac{1}{n} \sum_{i=1}^{n} X_{i,\widehat{\xi}} \left(Y_i - \widehat{\eta}(Z_i^T \widehat{\alpha}; \widehat{\alpha}, \widehat{\beta}) - X_i^T \widehat{\beta} \right) \right]. \tag{5}$$

The idea underlying this construction is the following. Given that $(\widehat{\alpha}, \widehat{\beta})$ minimizes (4), we have the Karash-Kuhn-Tucker condition

$$\frac{1}{n} \sum_{i=1}^{n} X_{i,\widehat{\xi}} (Y_i - \widehat{\eta}(Z_i^T \widehat{\alpha}; \widehat{\alpha}, \widehat{\beta}) - X_i^T \widehat{\beta}) = \lambda \widehat{\kappa}$$

with $\|\widehat{\kappa}\|_\infty \leq 1$ and $\widehat{\kappa}_j = \text{sign}(\widehat{\xi}_j)$. Then by using the equality $Y_i = X_i^T\beta_0 + \eta_0(Z_i^T\alpha) + \epsilon_i$ and the approximation $\widehat{\eta}(Z_i^T\widehat{\alpha};\widehat{\alpha},\widehat{\beta}) \approx \widehat{\eta}(Z_i^T\alpha_0;\alpha_0,\beta_0) + \frac{\partial\widehat{\eta}}{\partial\xi}|_{(\widehat{\alpha},\widehat{\beta})}.(\widehat{\xi} - \xi_0)$, we obtain

$$\frac{1}{n}\sum_{i=1}^{n} X_{i,\widehat{\xi}}\left[-X_{i,\widehat{\xi}}^T(\widehat{\xi}-\xi_0) + \epsilon_i + \eta_0(Z_i^T\alpha) - \widehat{\eta}(Z_i^T\alpha_0;\alpha_0,\beta_0)\right] \approx \lambda\widehat{\kappa}.$$

Now, using the fact that $\widehat{\Theta}_{\widehat{\xi}}$ is a relaxed inverse of $\frac{1}{n}\sum_{i=1}^n X_{i,\widehat{\xi}}X_{i,\widehat{\xi}}^T$, we have

$$\begin{aligned} -(\widehat{\xi}-\xi_0) \quad + \quad & \widehat{\Theta}_{\widehat{\xi}}\left[\frac{1}{n}\sum_{i=1}^{n} X_{i,\widehat{\xi}}\left(\epsilon_i + (\eta_0(Z_i^T\alpha^0) - \widehat{\eta}(Z_i^T\alpha_0;\alpha_0,\beta_0))\right)\right] \\ \approx \quad & \widehat{\Theta}_{\widehat{\xi}}\left[\frac{1}{n}\sum_{i=1}^{n} X_{i,\widehat{\xi}}\left(Y_i - \widehat{\eta}(Z_i^T\widehat{\alpha};\widehat{\alpha},\widehat{\beta}) - X_i^T\widehat{\beta}\right)\right]. \end{aligned}$$

We recognize the definition of $\widehat{\xi}^{\text{desp}}$ and see that $\widehat{\xi}^{\text{desp}} - \xi_0$ can be approximated by the sum of a Gaussian term and a term tending to zero.

Theorem 12. *Let $\widehat{\xi}$ be the Lasso penalized estimator, obtained by the minimization of the penalized profile least-squares (4). Let $\widehat{\xi}^{\text{desp}}$ be obtained by the desparsifying process (5) where $\widehat{\Theta}_{\widehat{\xi}}$ is obtained by the nodewise regression technique. Assume that conditions (C1) to (C10) of [2] do hold. Then, for each $j \in \{1,\ldots,p+q\}$ we have:*

$$\frac{\sqrt{n}(\widehat{\xi}_j^{\text{desp}} - \xi_j^0)}{\widehat{\sigma}_j} = V_j + o_P(1),$$

where V_j converges weakly to a $N(0,1)$ distribution and where

$$\widehat{\sigma}_j^2 := \sigma_\epsilon^2\left(\widehat{\Theta}_{\widehat{\xi}}\widehat{\Sigma}_{\widehat{\xi}}\widehat{\Theta}_{\widehat{\xi}}^T\right)_{j,j}.$$

A proof can be found in [2]. The construction of confidence intervals for each component of ξ is straightforward.

2.3 Summary of the methodology

We now summarize our methodology to perform inference on the parametric part of a partially linear single-index model.

1. Define the penalized estimator $\widehat{\xi} = (\widehat{\alpha},\widehat{\beta})$ as the minimizer of $L(\alpha,\beta) = \frac{1}{2n}\sum_{i=1}^n(Y_i - \widehat{\eta}(Z_i^T\alpha;\alpha,\beta) - X_i^T\beta)^2 + \lambda\|\alpha\|_1 + \lambda\|\beta\|_1$, where the function $\widehat{\eta}(Z_i^T\alpha;\alpha,\beta)$ is defined in equation (2).

2. Compute $\widehat{\eta}(Z_i^T\widehat{\alpha};\widehat{\alpha},\widehat{\beta})$ and $\frac{\partial\widehat{\eta}(Z_i^T\alpha;\alpha,\beta)}{\partial(\alpha,\beta)}|_{(\widehat{\alpha},\widehat{\beta})}$ by using equation (2).

3. Compute $X_{i,\widehat{\xi}}$, $\widehat{\Sigma}_{\widehat{\xi}}$, $\widehat{\theta}_{\widehat{\xi}}$ as described in section 2.2.

4. Compute $\widehat{\xi}^{\text{desp}} = \widehat{\xi} + \widehat{\Theta}_{\widehat{\xi}} \left[\frac{1}{n} \sum_{i=1}^{n} X_{i,\widehat{\xi}} (Y_i - \widehat{\eta}(Z_i^T \widehat{\alpha}; \widehat{\alpha}, \widehat{\beta}) - X_i^T \widehat{\beta}) \right]$.

5. Use Theorem 12 to perform inference. For example construct a confidence interval for ξ_j^0 at a confidence level $1-c$ as $CI_j = \left[\widehat{\xi}_j^{\text{desp}} \pm \frac{\widehat{\sigma}_j}{\sqrt{n}} \Phi^{-1}(1-c/2) \right]$, with $\widehat{\sigma}_j$ defined in the theorem and Φ the standard normal cumulative distribution function.

3 Simulation results

We now illustrate our method via a simulation study. We consider the model

$$Y = (Z^T \alpha^0 - 0.5)^2 + X^T \beta^0 + \epsilon$$

where ϵ is from a $N(0, 0.3^2)$ distribution and where Z and X are independently generated from a $N(0, I_{p\times p})$ and a $N(0, I_{q\times q})$ distribution for each observation. The parameters are defined as $\alpha^0 = (\frac{1}{\sqrt{s_0}} \cdot 1_{s_0}, 0 \cdot 1_{p-s_0})$ and $\beta^0 = (1_{s_0}, 0 \cdot 1_{q-s_0})$, with s_0 the sparsity level and 1_a a vector of ones of length a. We compute 95% univariate confidence intervals using the procedure described in the previous section. Figure 1 illustrates an example of univariate confidence intervals for one realization. Table 1 compares the average coverage obtained by using the non-penalized estimator, the penalized estimator and our desparsified estimator over 500 independent realizations for several settings. We observe good finite-sample performances.

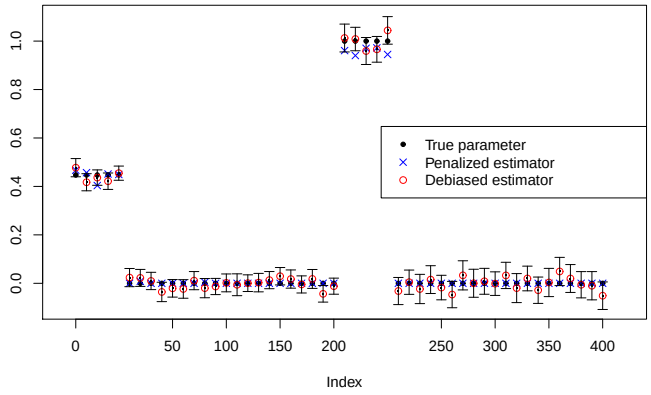


Figure 1: 95% confidence intervals for one realization with $(n, p, q, s_0) = (500, 200, 200, 5)$. For clarity only 10% of the components of the true non-active set are shown.

		$n = 200$			$n = 500$		
		Non-Penalized	Penalized	Desparsified	Non-Penalized	Penalized	Desparsified
Sparsity $s_0 = 2$	Avg cov	0.78	n/a	0.94	0.86	n/a	0.93
	Avg cov $S_{0,\alpha}$	0.69	0.66	0.92	0.80	0.87	0.90
	Avg cov $S^c_{0,\alpha}$	0.75	n/a	0.93	0.82	n/a	0.92
	Avg cov $S_{0,\beta}$	0.81	0.26	0.93	0.89	0.35	0.94
	Avg cov $S^c_{0,\beta}$	0.83	n/a	0.95	0.90	n/a	0.95
Sparsity $s_0 = 5$	Avg cov	0.78	n/a	0.97	0.81	n/a	0.93
	Avg cov $S_{0,\alpha}$	0.65	0.59	0.94	0.71	0.64	0.87
	Avg cov $S^c_{0,\alpha}$	0.75	n/a	0.96	0.74	n/a	0.91
	Avg cov $S_{0,\beta}$	0.83	0.26	0.97	0.88	0.31	0.94
	Avg cov $S^c_{0,\beta}$	0.82	n/a	0.97	0.88	n/a	0.96

Table 1: Average coverage of confidence intervals for nominal coverage of 0.95 over 500 simulation runs with 100 independent variables ($p = q = 50$) and $\sigma_\epsilon = 0.3$. We vary the number of observations n and the sparsity level s_0. The scaled lasso is used to compute the penalized estimator. $S_{0,\alpha}$ (resp. $S_{0,\beta}$) is the true active set of the components of α (resp. β), while $S^c_{0,\alpha}$ (resp. $S^c_{0,\beta}$) is the true non-active set.

4 Discussion

We obtained confidence intervals for all parametric components of a partially linear single-index model in a high-dimensional setting where sparsity is assumed. We show the desparsification process to work well for such models.

Acknowledgements: We acknowledge the support of the Fund for Scientific Research Flanders, KU Leuven grant GOA/12/14 and of the IAP Research Network P7/06 of the Belgian Science Policy. The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Hercules Foundation and the Flemish Government - department EWI.

References

[1] J. Fan, R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[2] T. Gueuning, G. Claeskens. Confidence intervals for high-dimensional partially linear single-index models *Technical report*, 2015.

[3] J.-L. Horowitz. Semiparametric methods in econometrics. Springer, 1998.

[4] H. Liang, X. Liu, R. Li, C.-L. Tsai. Estimation and testing for partially linear single-index models *Annals of statistics*, 38(6):3811, 2010.

[5] S. van de Geer, P. Bühlmann, Y. Ritov, R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models *Annals of Statistics*, 42(3):1166:1202, 2014.

[6] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, Series B (Methodological) : 267-288, 1996.

Random Matrix Models with Heavy Tails

Johannes Heiny[*1] **and Thomas Mikosch**[1]

[1] *University of Copenhagen, Denmark*

Abstract: Many fields of modern sciences are faced with high-dimensional data sets. In order to explore the structure in the data the sample covariance matrix can be used. Often dimension reduction techniques facilitate further analyzes of large data matrices in genetic engineering and finance. Principal Component Analysis for example makes a linear transformation of the data to obtain vectors of which the first few contain most of the variation in the data. These principal component vectors correspond to the largest eigenvalues of the sample covariance matrix. This motivates to study the eigenvalue decomposition of the sample covariance matrix.

Random Matrix Theory is concerned with the spectral properties of large dimensional random matrices. In this context both the distribution of the entries of a random matrix as well as their dependence structure play a crucial role. The case of heavy-tailed components is of particular interest and the theory is not as well developed as in the light-tailed case.

We consider the (random) sample covariance matrix of a multivariate time series. The p-dimensional time series constitutes a linear process across time and between components. The input noise of the linear process is assumed to have a regularly varying tail with index $\alpha \in (0, 4)$. For such distributions moments higher than α cease to exist. If we have n observations from this time series, we can calculate the corresponding sample covariance matrix. In classical multivariate statistics p is fixed and relatively small in comparison to the sample size n. In our setting both the dimension and the sample size tend to infinity simultaneously.

If the multivariate time series is iid across time and between the components, Auffinger et al. [1] showed that the point process of properly normalized eigenvalues of the sample covariance matrix converges in distribution to a Poisson point process that only depends on the index of regular variation. Davis et al. [2] provided an extension by dropping the independence assumption across time. We also drop the independence between components. In particular, we study the asymptotic behavior of the largest eigenvalues of such a sample covariance matrix by approximation results.

Keywords: regular variation, sample covariance matrix, largest eigenvalues, dependent entries

AMS subject classifications: 60B20, 60F05, 60F10, 60G10, 60G70

Acknowledgements: Johannes Heiny's research is supported by the Danish Research Council Grant DFF-4002-00435 "Large random matrices with heavy tails and dependence".

[*]Corresponding author: johannes.heiny@math.ku.dk

References

[1] Auffinger, A., Ben Arous, G. and Péché, S. (2009) Poisson convergence for the largest eigenvalues of heavy tailed random matrices. *Ann. Inst. H. Poincaré, Probab. Statist.* **45**, 589–610.

[2] Davis, R.A., Pfaffel, O. and Stelzer, R. (2014) Limit theory for the largest eigenvalues of sample covariance matrices with heavy tails. *Stoch. Proc. Appl.* **124**, 18–50.

Comparison of Methods for Variable Selection in High-Dimensional Linear Mixed Models

Jozef Jakubík*

Institute of Measurement Science, Slovak Academy of Sciences, Bratislava, Slovakia

Abstract: Currently is the analysis of high-dimensional data a popular field of research, thanks to many applications e.g. in genetics. At the same time, the type of problems that tend to arise in genetics, can often be modelled using LMMs in conjunction with high-dimensional data. In this paper we introduce two new methods and briefly compare them to existing methods, which can be used for variable selection in high-dimensional linear mixed models. We compare the methods on "small dimension" high-dimensional data, because some of the compared methods are not suitable for very high dimensions. As we will show in a simulation study, both methods perform well compared to existing methods.

Keywords: linear mixed model, variable selection, high-dimensional data
AMS subject classifications: 62J07

1 Introduction

The linear mixed model (LMM) allows us to specify the covariance structure of the model, which enables us to capture relationships in data, for example population structure, family relatedness etc. Therefore, LMMs are often preferred to linear regression models. Consider a LMM of the form

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\varepsilon},$$

where

- $\boldsymbol{Y}$ is $n \times 1$ vector of observations,
- $\boldsymbol{X}$ is $n \times p$ matrix of regressors,
- $\boldsymbol{\beta}$ is $p \times 1$ vector of unknown fixed effects,
- $\boldsymbol{Z}$ is $n \times q$ matrix of predictors,
- $\boldsymbol{u}$ is $q \times 1$ vector of random effects with the distribution $\mathcal{N}(0, \boldsymbol{D})$,
- $\boldsymbol{\varepsilon}$ is $n \times 1$ error vector with the distribution $\mathcal{N}(0, \boldsymbol{R} = \sigma^2 \boldsymbol{I})$ and independent from $\boldsymbol{u}$.

*Corresponding author: jozef.jakubik.jefo@gmail.com

In genome-wide association studies in genetics, one studies the dependence of phenotype on the genotype. Genetic information can consist of up to 10^6 variables, but only information about the genotype of a small group of subjects is available. Variable selection in high-dimensional data refers to the selection of a small group of variables (denote it S^0, and $s^0 = |S^0|$ the number of relevant variables) which influence observations. In our case we assume, that matrix $\boldsymbol{X}$ is high-dimensional and we select only variables from matrix $\boldsymbol{X}$.

More information about the model can be found in section 3.

2 Methods

In this paper we compare five methods for variable selection in high-dimensional LMMs.

All of the mentioned methods are primarily $\boldsymbol{\beta}$ estimation methods, not selection methods. However they can be thought of as selection methods if we define selected variables to be those for which $\boldsymbol{\beta}_i \neq 0$ for $i = 1, \ldots, p$.

2.1 LASSO

Least absolute shrinkage and selection operator [4, 1] is an established method for selecting variables in linear regression models. LASSO corresponds to the ℓ_1-penalized ordinary least squares estimate:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left[\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right],$$

where λ is a fixed penalized parameter.

In this study we use the LASSO as the reference, as it ignores LMM data structure. For the LASSO method we use the built-in `MATLAB` function `lasso()`.

2.2 LMMLASSO

Authors in [3] suggest a method based on the minimization of the non-convex objective function consisting of the ℓ_1 penalized maximum likelihood estimate of parameter $\boldsymbol{\beta}$ from $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{V}(= \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}^\mathsf{T} + \boldsymbol{R}))$:

$$(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{D}}, \widehat{\boldsymbol{R}}) = \arg\min_{\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{R}} \left[\frac{1}{2} \log |\boldsymbol{V}| + \frac{1}{2} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\mathsf{T} \boldsymbol{V}^{-1} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1 \right],$$

where λ is a fixed parameter. For this method we used the language `R` package `lmmlasso`, which uses the coordinate gradient descent algorithm to the optimize objective function.

2.3 LASSOP

In paper [2] authors introduce a method based on the log-likelihood of the complex data $(\boldsymbol{Y}^\mathsf{T}, \boldsymbol{u}^\mathsf{T})^\mathsf{T}$ penalized with the ℓ_1 penalization:

$$(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{D}}, \widehat{\boldsymbol{R}}) = \underset{\boldsymbol{\beta},\boldsymbol{D},\boldsymbol{R}}{\arg\min} \left[\log|\boldsymbol{R}| + (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})^{\mathsf{T}}\boldsymbol{R}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u}) + \log|\boldsymbol{D}| + \boldsymbol{u}^{\mathsf{T}}\boldsymbol{D}^{-1}\boldsymbol{u} + \lambda\|\boldsymbol{\beta}\|_1\right],$$

where λ is a fixed parameter. The objective function is non-convex, like in LMMLASSO. This method is implemented in language `R` in package `MMS`. The optimize problem in this implementation is solved by the adjusted EM algorithm.

2.4 New approach one

The first approach consists in a transformation that removes group effects from data. The principle of this transformation is widely used in data analysis, for example in restricted/residual maximum likelihood (REML). In our case we transform the data as follows

$$\widetilde{\boldsymbol{X}} = (\boldsymbol{I} - \boldsymbol{Z}\boldsymbol{Z}^{+})\boldsymbol{X},$$
$$\widetilde{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{Z}\boldsymbol{Z}^{+})\boldsymbol{Y},$$

where $\boldsymbol{Z}^{+}$ is the pseudoinverse matrix. The transformation eliminates random segments of the problem (associated with matrix $\boldsymbol{Z}$), which allows us to use the LASSO method for linear regression model.

2.5 New approach two

Methods LMMLASSO and LASSOP are based on non-convex optimization problems with one penalty parameter. For problems of dimension higher than 10^4 are methods based on non-convex optimization problems almost unusable, because their computational complexity is beyond the capabilities of current computers. One of the possible solutions to this problem is the simplification of the optimized function to a convex function. Therefore, we have proposed a method based on the solution to the following convex problem

$$(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{u}}) = \underset{\boldsymbol{\beta},\boldsymbol{u}}{\arg\min} \left[\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u}\|_2^2 - \lambda\|\boldsymbol{\beta}\|_1 - \Lambda\sum_{i=1}^{q^*}\|\,{}_i\boldsymbol{u}\|_2^2\right],$$

where λ and Λ are fixed parameters, q^* is the number of variance components (without error) and ${}_i\boldsymbol{u}$ is the part of vector $\boldsymbol{u}$ belonging to the i-th variance component.

Basically we are exchanging computational complexity for the need to inspect a two-dimensional parameter space. We implement the method in `MATLAB` using the modeling system `CVX` and the solver `Mosek`.

3 Simulation study

This study compares the presented methods on high-dimensional data with "small dimension", because the curent implementations of methods LMMLASSO and LASSOP are commonly unable to solve problems of dimension higher then $p = 10^3$.

Data in our simulation study are divided into twenty groups of six observations. Together we have $n = 120$ observations. For each observation we observe $p = 150$ variables, but only $s^0 = \{1, \ldots, 10\}$ variables influence the observations. Relevant variables are randomly selected from all variables and the effect of relevant variables is one. The effect of other variables is zero. Matrix $\boldsymbol{Z}$ captures group structure of the data. For every group we observe two variables, therefore we consider two variance components and the error variance component. $\boldsymbol{Z}$ is a block diagonal matrix and $\boldsymbol{u}$ consist of two parts, each for one variance component. Both parts of the random effects $\boldsymbol{u}$ are randomly selected from $\mathcal{N}(0, \boldsymbol{D} = 2 \cdot \boldsymbol{I})$. Errors are from $\mathcal{N}(0, \boldsymbol{I})$.

For all mentioned methods we get different sets of selected variables for different parameters λ or Λ. We generate a hundred problems as described in the previous paragraph. As a correctly solved problem we consider only a problem for which the method gives for at least one parameter or parameter combination as the selected variable set exactly set S^0. Figure 1 shows the number of correctly solved problems for all five methods for different numbers of relevant variables (from 1 to 10).

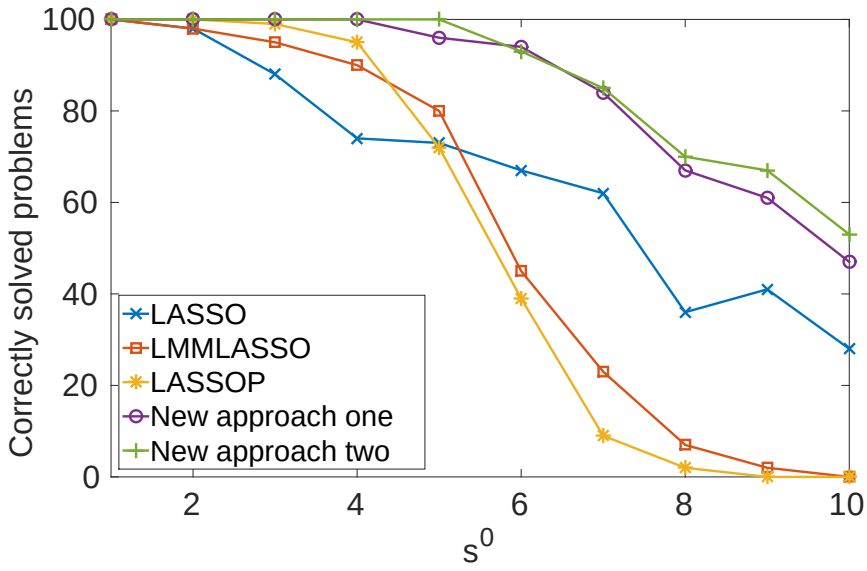


Figure 1: Comparison of the number of correctly solved problems for different s^0 with five different methods.

4 Conclusion

In Figure 1 we can see that for small numbers of relevant variables all methods are almost infallible. With the increasing number of relevant variables, the accuracy of methods decreases. This is understandable, because with more relevant variables the correlation of each relevant variable with the vector of observations decreases. Therefore, it is more difficult to identify correctly the exact set of variables. The accuracy of methods LMMLASSO and LASSOP decreases faster. The seemingly weak performance of these methods is caused by the very strict condition of correctness. Often, the sets of selected variables identify by these methods contained only a few unnecessary variables. This may be due to the implementation in different languages.

This study hints at the potential of the newly proposed methods to significantly outperform both the LMMLASSO and the LASSOP. The newly proposed methods are also suitable for high-dimensional data with dimension up to 10^5.

However, this is only a preliminary study and one of the first addressing the question. A more extensive analysis can be expected in the future.

Acknowledgements: The work was supported by the Scientific Grant Agency VEGA of the Ministry of Education of the Slovak Republic and the Slovak Academy of Sciences, by the projects VEGA 2/0047/15 and VEGA 2/0043/13.

References

[1] P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[2] F. Rohart, M. San Cristobal, and B. Laurent. Selection of fixed effects in high dimensional linear mixed models using a multicycle ECM algorithm. *Computational Statistics & Data Analysis*, 80:209–222, Dec. 2014.

[3] J. Schelldorfer, P. Bühlmann, and S. van De Geer. Estimation for High-Dimensional Linear Mixed-Effects Models Using ℓ_1-Penalization. *Scandinavian Journal of Statistics*, 38(2):197–214, June 2011.

[4] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58:267–288, 1996.

Confidence Regions for High-Dimensional Sparse Models

Jana Janková[*1] **and Sara van de Geer**[1]

[1]*ETH Zürich, Switzerland*

Abstract: In many important statistical applications the number of parameters may be much larger than the sample size. One of the main approaches to deal with this situation is to assume that the underlying true model is sparse so that only a sufficiently small number of parameters are important. Methods for point estimation in such sparse high-dimensional settings have been extensively studied in the past years (see [1]). Less work has yet been done on developing methodology for inference for the parameters of interest in this setting. We propose an estimator for low-dimensional parameters of the high-dimensional vector and show it is asymptotically normal and regular, leading to confidence regions for low-dimensional parameters. Our approach is based on ℓ_1-penalized M-estimators which serve as initial estimates for construction a one-step corrected estimator. We show validity of the suggested approach in several situations including quantile regression where the loss function is not differentiable or precision matrix estimation [2], and further consider general high-dimensional models. Under a sparsity assumption on the high-dimensional parameter, smoothness conditions on the expected loss and an entropy condition we show that the proposed de-sparsified estimator is asymptotically normal. This leads to uniformly valid confidence regions and hypothesis testing for low-dimensional parameters.

Keywords: high-dimensional models, sparsity, inference, lasso

AMS subject classifications: 62F12

References

[1] Bühlmann, P. and van de Geer, S. Statistics for high-dimensional data. *Springer* (2011).

[2] Jankova, J. and van de Geer, S. Confidence intervals for high-dimensional inverse covariance estimation. *ArXiv :1403.6752* (2014).

[3] van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, **42**(3) (2014), 1166-1202.

[4] Zhang, C.-H. and Zhang, S. S. Confidence intervals for low-dimensional parameters in high-dimensional linear models. *Journal of the Royal Statistical Society: Series B*, **76** (2014), 217–242.

[*]Corresponding author: jankova@stat.math.ethz.ch

Asymptotic Theory for Copula Rank-Based Periodograms

Tobias Kley*

Ruhr-Universität Bochum, Fakultät für Mathematik, Germany

Abstract: Alternative spectral concepts for the analysis of time series recently have been considered by various authors. The *copula spectral density kernels* introduced in [3] provide a full characterization of the copulas associated with the pairs (X_t, X_{t-k}) in a process $(X_t)_{t\in\mathbb{Z}}$, and account for important dynamic features, such as changes in the conditional shape (skewness, kurtosis), time-irreversibility, or dependence in the extremes, that their traditional counterparts cannot capture. Despite various proposals for estimation strategies, only quite incomplete asymptotic distributional results are available so far for the proposed estimators, which constitutes an important obstacle for their practical application. This paper contains motivation and definitions for the class of smoothed rank-based cross-periodograms considered in [13]. An asymptotic analysis is presented: for a very general class of (possibly non-linear) processes, properly scaled and centered smoothed versions of those cross-periodograms, indexed by couples of quantile levels, converge weakly, as stochastic processes, to Gaussian processes. Note that the present paper is a very condensed version of [13], where technical details and simulation results can be found.

Keywords: time series, spectral analysis, quantiles, copulas, ranks

AMS subject classifications: 62M15, 62G35

1 Introduction

Frequency domain methods play a central role in the nonparametric analysis of time series. The classical approach is based on the *spectral density* which is traditionally defined as the Fourier transform of the autocovariance function of the process under study. Common tools for the estimation of spectral densities are the *periodogram* and its *smoothed* versions. The success of periodograms in time series analysis is rooted in their fast and simple computation (through the *fast Fourier transform* algorithm) and their interpretation in terms of cyclic behavior, both of a stochastic and of deterministic nature, which in specific applications are more illuminating than time-domain representations.

Being intrinsically connected to means and covariances, the traditional spectral analysis inherits both the nice features (such as optimality properties in the analysis of Gaussian series) of L^2-methods, but also their weaknesses: they are lacking robustness against outliers and heavy tails, and are unable to capture important

*Corresponding author: tobias.kley@ruhr-uni-bochum.de

dynamic features such as changes in the conditional shape (skewness, kurtosis), time-irreversibility, or dependence in the extremes.

Various extensions and modifications of the traditional periodogram were proposed to remedy those drawbacks (see for example [6], [9], [10], [14], Chapter 8 of [20], and [21]). While the objective of those attempts is a robustification of classical tools, they essentially aim at protecting existing spectral methods against the impact of possible outliers or violations of distributional assumptions.

Recently, alternative spectral concepts that account for more general dynamic features were proposed. A first step in that direction was taken by [7], who proposes a generalized spectral density with covariances replaced by joint characteristic functions. In the specific problem of testing pairwise independence, [8] introduces a test statistic based on the Fourier transforms of (empirical) joint distribution functions and copulas at different lags. The strand of literature indicating the renewed surge of interest in that type of concept includes [2], [3], [5], [15], [16], [17], and [18]. A more detailed account of these and some further contributions are given in [13].

The objective of the present paper is to provide a short, comprehensive presentation of the general class of smoothed rank-based copula cross-periodograms discussed at length in [13]. Their asymptotic properties are discussed in Section 3. For a detailed discussion, applications and a simulation study the reader shall be referred to the full version of the paper (i. e. [13]).

2 Copula spectral density kernels and rank-based copula periodograms

Let $(X_t)_{t\in\mathbb{Z}}$ denote a strictly stationary, real-valued process, of which we observe a finite stretch $X_0, ..., X_{n-1}$. Denote by F the marginal distribution function of X_t, and by $q_\tau := \inf\{x \in \mathbb{R} : \tau \leq F(x)\}$, $\tau \in [0,1]$ the corresponding quantile function, where we use the convention $\inf \emptyset = \infty$. Note that if $\tau \in \{0,1\}$ then $-\infty$ and ∞ are possible values for q_τ. Our main object of interest is the *copula spectral density kernel*

$$\mathfrak{f}_{q_{\tau_1},q_{\tau_2}}(\omega) := \frac{1}{2\pi}\sum_{k\in\mathbb{Z}} \mathrm{e}^{-\mathrm{i}\omega k}\gamma_k^U(\tau_1,\tau_2), \qquad \omega \in \mathbb{R},\ (\tau_1,\tau_2) \in [0,1]^2, \tag{1}$$

based on the *copula cross-covariances*

$$\gamma_k^U(\tau_1,\tau_2) := \mathrm{Cov}\Big(I\{F(X_t) \leq \tau_1\}, I\{F(X_{t-k}) \leq \tau_2\}\Big), \qquad k \in \mathbb{Z}.$$

Note that $\mathfrak{f}_{q_{\tau_1},q_{\tau_2}}(\omega)$ exists under mild mixing assumptions. These quantities were introduced in [3], and generalize the *τ-th quantile spectral densities* of [5], with which they coincide for $\tau_1 = \tau_2 = \tau$; an integrated version actually was first considered by [8].

It can be shown (cf. [3]) that the copula spectral densities provide a complete description of the pairwise copulas of a time series. Thus, by accounting for much more than the covariance structure of a series, it extends and supplements the classical L^2-spectral density.

It is important to observe that $\gamma_k^U(\tau_1,\tau_2)$ and $\mathfrak{f}_{q_{\tau_1},q_{\tau_2}}(\omega)$ can be seen as the cross-covariance and cross-spectrum of the bivariate process

$$\Big(I\{F(X_t)\le\tau_1\}, I\{F(X_t)\le\tau_2\}\Big). \tag{2}$$

Thus, an estimator for the cross spectrum of (2) is also an estimator for $\mathfrak{f}_{q_{\tau_1},q_{\tau_2}}(\omega)$. If F were known and $(X_t)_{t\in\mathbb{Z}}$ was observed, we could determine (2) and follow the lines of [1] to construct an estimator. Observe, that (2) has binary-valued component processes, such that classical assumptions regarding the dependency structure (e. g. linearity of the process) can fail to hold. The estimation of cross spectra of *general non-linear processes* has been and remains to be an active domain of research (see [1] for early results, and [4], [19] or [22] for more recent references).

To handle the case where F is unknown, it is reasonable to estimate the unknown F with the empirical distribution function $\widehat{F}_n(x) := n^{-1}\sum_{t=0}^{n-1} I\{X_t\le x\}$. More precisely, we define $I_{n,R}$ as the collection

$$I_{n,R}^{\tau_1,\tau_2}(\omega) := \frac{1}{2\pi n} d_{n,R}^{\tau_1}(\omega) d_{n,R}^{\tau_2}(-\omega), \quad d_{n,R}^{\tau}(\omega) := \sum_{t=0}^{n-1} I\{\widehat{F}_n(X_t)\le\tau\}\mathrm{e}^{-\mathrm{i}\omega t}, \tag{3}$$

$\omega\in\mathbb{R}$, $(\tau_1,\tau_2)\in[0,1]^2$, $\tau\in[0,1]$. Note that $n\widehat{F}_n(X_t)$ is the rank of X_t among $X_0,\dots,X_{n-1}$. Hence, we will refer to $I_{n,R}$ as the *rank-based copula periodogram*; shortly, the *CR-periodogram*. Let $\rightsquigarrow$ denote the *Hoffman–Jørgensen convergence*, namely, the weak convergence in the space of bounded functions $[0,1]^2\to\mathbb{C}$, which we denote by $\ell^\infty_{\mathbb{C}}([0,1]^2)$. Note that results in empirical process theory are typically stated for spaces of real-valued, bounded functions; see Chapter 1 of [23]. By identifying $\ell^\infty_{\mathbb{C}}([0,1]^2)$ with the product space $\ell^\infty([0,1]^2)\times\ell^\infty([0,1]^2)$ these results transfer immediately. We will see (Theorem 13) that, under suitable assumptions,

$$\Big(I_{n,R}^{\tau_1,\tau_2}(\omega)\Big)_{(\tau_1,\tau_2)\in[0,1]^2} \rightsquigarrow \Big(\mathbb{I}(\tau_1,\tau_2;\omega)\Big)_{(\tau_1,\tau_2)\in[0,1]^2} \quad \text{as } n\to\infty,$$

for any fixed frequencies $\omega\neq 0 \mod 2\pi$, where the limit $\mathbb{I}$ is such that

$$\mathbb{E}[\mathbb{I}(\tau_1,\tau_2;\omega)] = \mathfrak{f}_{q_{\tau_1},q_{\tau_2}}(\omega) \quad \text{for all } (\tau_1,\tau_2)\in[0,1]^2 \text{ and } \omega\neq 0 \mod 2\pi$$

and $\mathbb{I}(\cdot,\cdot;\omega_1)$ and $\mathbb{I}(\cdot,\cdot;\omega_2)$ are independent as soon as both $\omega_1-\omega_2\neq 0 \mod 2\pi$ and $\omega_1+\omega_2\neq 0 \mod 2\pi$. In view of this asymptotic independence at different frequencies, it seems natural to consider smoothed versions of $I_{n,R}^{\tau_1,\tau_2}(\omega)$, namely, for $(\tau_1,\tau_2)\in[0,1]^2$ and $\omega\in\mathbb{R}$, averages of the form

$$\widehat{G}_{n,R}(\tau_1,\tau_2;\omega) := \frac{2\pi}{n}\sum_{s=1}^{n-1} W_n\big(\omega-2\pi s/n\big) I_{n,R}^{\tau_1,\tau_2}(2\pi s/n), \tag{4}$$

where W_n denotes a sequence of weighting functions.

3 Asymptotic properties

We make the following assumption regarding the process $(X_t)_{t\in\mathbb{Z}}$:

(A) Assume that the process $(X_t)_{t\in\mathbb{Z}}$ is strictly stationary and exponentially α-mixing, i. e., there exists constants $K<\infty$ and $\kappa\in(0,1)$, such that,

$$\alpha(n) := \sup_{\substack{A\in\sigma(X_0,X_{-1},\ldots)\\ B\in\sigma(X_n,X_{n+1},\ldots)}} |\mathbb{P}(A\cap B)-\mathbb{P}(A)\mathbb{P}(B)| \le K\kappa^n\ ,\ n\in\mathbb{N}. \tag{5}$$

Theorem 13. *Assume that F is continuous and that $(X_t)_{t\in\mathbb{Z}}$ satisfies Assumption (A). Then, for every fixed $\omega\neq 0 \mod 2\pi$,*

$$\Big(I_{n,R}^{\tau_1,\tau_2}(\omega)\Big)_{(\tau_1,\tau_2)\in[0,1]^2} \rightsquigarrow \Big(\mathbb{I}(\tau_1,\tau_2;\omega)\Big)_{(\tau_1,\tau_2)\in[0,1]^2} \quad \text{in } \ell^\infty_{\mathbb{C}}([0,1]^2).$$

The (complex-valued) limiting processes $\mathbb{I}$, are of the form

$$\mathbb{I}(\tau_1,\tau_2;\omega)=\frac{1}{2\pi}\mathbb{D}(\tau_1;\omega)\overline{\mathbb{D}(\tau_2;\omega)},\quad (\tau_1,\tau_2)\in[0,1]^2,$$

with $\mathbb{D}(\tau;\omega)=\mathbb{C}(\tau;\omega)+\mathrm{i}\mathbb{S}(\tau;\omega)$ where $\mathbb{C}$ and $\mathbb{S}$ denote two centered jointly Gaussian processes. For $\omega\in\mathbb{R}$, their covariance structure takes the form

$$\mathbb{E}\Big[(\mathbb{C}(\tau_1;\omega),\mathbb{S}(\tau_1;\omega))'(\mathbb{C}(\tau_2;\omega),\mathbb{S}(\tau_2;\omega))\Big]=\pi\begin{pmatrix}\Re\mathfrak{f}_{q_{\tau_1},q_{\tau_2}}(\omega) & -\Im\mathfrak{f}_{q_{\tau_1},q_{\tau_2}}(\omega)\\ \Im\mathfrak{f}_{q_{\tau_1},q_{\tau_2}}(\omega) & \Re\mathfrak{f}_{q_{\tau_1},q_{\tau_2}}(\omega)\end{pmatrix}.$$

Moreover, $\mathbb{D}(\tau;\omega)=\mathbb{D}(\tau;\omega+2\pi)=\overline{\mathbb{D}(\tau;-\omega)}$, and the family $\{\mathbb{D}(\cdot;\omega):\omega\in[0,\pi]\}$ is a collection of independent processes.

In order to establish the convergence of the smoothed CR-periodogram process (4), we require the weights W_n in (4) to satisfy the following assumption:

(W) The weight function W is real-valued and even, with support $[-\pi,\pi]$; moreover, it has bounded variation, and satisfies $\int_{-\pi}^{\pi}W(u)\mathrm{d}u=1$.

Denoting by $b_n>0$, $n=1,2,\ldots$, a sequence of scaling parameters such that $b_n\to 0$ and $nb_n\to\infty$ as $n\to\infty$, define

$$W_n(u):=\sum_{j=-\infty}^{\infty} b_n^{-1}W(b_n^{-1}[u+2\pi j]).$$

Theorem 14. *Let Assumptions (A) and (W) hold. Assume that X_0 has a continuous distribution function F and that there exist constants $\kappa>0$, $k\in\mathbb{N}$, s. t.*

$$b_n=o(n^{-1/(2k+1)}) \quad \text{and} \quad b_n n^{1-\kappa}\to\infty.$$

Then, for any fixed $\omega\in\mathbb{R}$, the process

$$\sqrt{nb_n}\Big(\widehat{G}_{n,R}(\tau_1,\tau_2;\omega)-\mathfrak{f}_{q_{\tau_1},q_{\tau_2}}(\omega)-B_n^{(k)}(\tau_1,\tau_2;\omega)\Big)_{\tau_1,\tau_2\in[0,1]}\rightsquigarrow H(\cdot,\cdot;\omega) \tag{6}$$

in $\ell^\infty_{\mathbb{C}}([0,1]^2)$, where the bias $B_n^{(k)}$ is given by

$$B_n^{(k)}(\tau_1,\tau_2;\omega):=\begin{cases}\displaystyle\sum_{j=2}^{k}\frac{b_n^j}{j!}\int_{-\pi}^{\pi}v^jW(v)dv\frac{d^j}{d\omega^j}\mathfrak{f}_{q_{\tau_1},q_{\tau_2}}(\omega) & \omega\neq 0 \mod 2\pi,\\ n(2\pi)^{-1}\tau_1\tau_2 & \omega=0 \mod 2\pi.\end{cases} \tag{7}$$

The process $H(\cdot,\cdot;\omega)$ in (6) is a centered Gaussian process characterized by

$$\mathrm{Cov}\big(H(u_1,v_1;\omega),H(u_2,v_2;\omega)\big) = 2\pi\Big(\int_{-\pi}^{\pi} W^2(w)dw\Big) \times \Big(\mathfrak{f}_{q_{u_1},q_{u_2}}(\omega)\mathfrak{f}_{q_{v_2},q_{v_1}}(\omega) + \mathfrak{f}_{q_{u_1},q_{v_2}}(\omega)\mathfrak{f}_{q_{v_1},q_{u_2}}(\omega)I\{\omega = 0 \mod \pi\}\Big).$$

Moreover, $H(\omega) = H(\omega+2\pi) = \overline{H(-\omega)}$, and the family $\{H(\omega),\ \omega\in[0,\pi]\}$ is a collection of independent processes. In particular, the weak convergence (6) holds jointly for any finite fixed collection of frequencies ω.

Theorem 14 can be used to conduct asymptotic inference in various ways. An important example is the construction of asymptotic confidence intervals (see Remark 3.4 and Section 5 in [13] for details). An R package is available (see [11, 12]).

Acknowledgements: This work has been supported by the Sonderforschungsbereich "Statistical modelling of nonlinear dynamic processes" (SFB 823) of the Deutsche Forschungsgemeinschaft. The author was supported by a PhD Grant of the Ruhr-Universität Bochum and by the Ruhr-Universität Research School funded by Germany's Excellence Initiative [DFG GSC 98/1]. This short note is a summary, essentially, of results obtained in collaboration with Holger Dette, Marc Hallin, and Stanislav Volgushev, whom I thank very much.

References

[1] D. R. Brillinger. *Time Series: Data Analysis and Theory.* Holt, Rinehart and Winston, Inc., New York, 1975.

[2] R. A. Davis, T. Mikosch, and Y. Zhao. Measures of serial extremal dependence and their estimation. *Stoch. Proc. Appl.*, 123:2575–2602, 2013.

[3] H. Dette, M. Hallin, T. Kley, and S. Volgushev. Of copulas, quantiles, ranks and spectra: An L_1-approach to spectral analysis. *Bernoulli*, forthcoming, 2015+.

[4] L. Giraitis and H. L. Koul. On asymptotic distributions of weighted sums of periodograms. *Bernoulli*, 19(5B):2389–2413, 2013.

[5] A. Hagemann. Robust spectral analysis (arxiv:1111.1965v2). *ArXiv*, 2013.

[6] J. B. Hill and A. McCloskey. Heavy tail robust frequency domain estimation. Available online: `http://www.econ.brown.edu/fac/adam_mccloskey/Research_files/FDTTQML.pdf`, 2013.

[7] Y. Hong. Hypothesis testing in time series via the empirical characteristic function: a generalized spectral density approach. *J. Am. Stat. Assoc.*, 94(448):1201–1220, 1999.

[8] Y. Hong. Generalized spectral tests for serial dependence. *J. Roy. Stat. Soc. B*, 62(3):557–574, 2000.

[9] V. Katkovnik. Robust M-periodogram. *IEEE T. Signal Proces.*, 46(11):3104–3109, 1998.

[10] R. Kleiner, R. D. Martin, and D. J. Thomson. Robust estimation of power spectra. *J. Roy. Stat. Soc. B*, 41(3):313–351, 1979.

[11] T. Kley. Quantile-based spectral analysis in an object-oriented framework and a reference implementation in R: The **quantspec** package. *J. Stat. Softw.*, forthcoming, 2015+.

[12] T. Kley. *quantspec: Quantile-based spectral analysis functions*, 2015. R package version 1.0-3.

[13] T. Kley, S. Volgushev, H. Dette, and M. Hallin. Quantile Spectral Processes: Asymptotic Analysis and Inference. *Bernoulli*, forthcoming, 2015+.

[14] C. Klüppelberg and T. Mikosch. Some limit theory for the self-normalised periodogram of stable processes. *Scand. J. Stat.*, 21:485–491, 1994.

[15] J. Lee and S. S. Rao. The quantile spectral density and comparison based tests for nonlinear time series (arxiv:1112.2759v2). *ArXiv*, 2012.

[16] T.-H. Li. Laplace periodogram for time series analysis. *J. Am. Stat. Assoc.*, 103(482):757–768, 2008.

[17] T.-H. Li. Quantile periodograms. *J. Am. Stat. Assoc.*, 107(498):765–776, 2012.

[18] T.-H. Li. *Time Series with Mixed Spectra: Theory and Methods*. CRC Press, Boca Raton, 2013.

[19] W. Liu and W. B. Wu. Asymptotics of spectral density estimates. *Econometric Theory*, 26(4):1218–1245, 2010.

[20] R. Maronna, D. Martin, and V. Yohai. *Robust Statistics: Theory and Methods*. Wiley Series in Probability and Statistics. Wiley, 2006.

[21] T. Mikosch. Periodogram estimates from heavy-tailed data. In R. A. Adler, R. Feldman, and M. S. Taqqu, editors, *A Practical Guide to Heavy Tails: Statistical Techniques for Analysing Heavy-Tailed Distributions*, pages 241–258. Birkhäuser, Boston, 1998.

[22] X. Shao and W. B. Wu. Asymptotic spectral theory for nonlinear time series. *Ann. Stat.*, 35(4):1773–1801, 2007.

[23] A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York, 1996.

Application of Dividend Policies to Finite Difference Methods in Option Pricing

Dessislava Koleva[*1] **and Mariyan Milev**[2]

[1] *Sofia University, Bulgaria*
[2] *UFT-Plovdiv, Bulgaria*

Abstract: The purpose of this short paper is to outline some practical issues that arise when one computes the fair value of a derivative on an underlying that pays out discrete dividends. In particular, the case of American options will be presented. We will show that as opposed to the benchmark Crank-Nicolson method, the positivity and smoothness of the numerical solution are preserved when discrete dividend payments are applied to an exponentially fitted scheme. Then, we will discuss some approaches to dividend policies in extreme scenarios and will see how our suggested scheme fits these cases.

Keywords: finite differences, dividends, American options
AMS subject classifications: 35K10, 62P05

1 Introduction

American options are widely employed in the financial industry. As there are no closed form solutions for this type of derivatives, one has to rely on numerical approximations in their pricing. Here we will focus of the application of finite difference methods for pricing these contracts. The methods employed often turn out to be sensitive towards special conditions, such as dividends. That is, the approximations may result in a solution with spurious oscillations or even negative prices. Also, when applying a numerical method, border cases need to be taken into account - low volatility levels, low asset price levels. Special attention will be given the case when the underlying asset price is lower than the dividend declared by the company.

2 Theoretical background

We consider a standard *geometric Brownian motion* diffusion process with constant coefficients r and σ for the evolution of the underlying asset price S

$$dS/S = rdt + \sigma dW_t, \tag{1}$$

where r and σ denote, respectively, the interest rate and volatility in percentages and belong to the interval $[0, 1]$. If t is the time to expiry T of the contract, $0 \leq t \leq T$, the price $V(S, t)$ of the option satisfies the Black-Scholes PDE ([1])

$$-\frac{\partial V}{\partial t} + r\,S\,\frac{\partial V}{\partial S} + \frac{1}{2}\,\sigma^2 S^2\,\frac{\partial^2 V}{\partial S^2} - r\,V = 0, \tag{2}$$

*Corresponding author: koleva_dn@yahoo.com

endowed by its initial and boundary conditions. The solution $V(S,t)$ depends on the two independent variables S and t. It should be noted that the option price can move on the positive real axis interval $[0,+\infty)$.

Finite difference schemes aim in approximating the solution of a PDE by solving a set of discretized equations. A discussion on different methods can be found in [7]. The most famous schemes are the explicit, implicit and Crank-Nicolson. They all rely on taking central differences when discretizing the partial derivatives with respect to the underlying asset S, i.e.

$$\frac{\partial V}{\partial S} = \frac{V_{i,j+1} - V_{i,j-1}}{2h}, \qquad \frac{\partial^2 V}{\partial S^2} = \frac{V_{i,j+1} + V_{i,j-1} - 2V_{i,j}}{h^2}. \tag{3}$$

Here, $V_{i,j}$ denotes the value of the option at i-th point in time and j-th point in the discretized interval $[0, S_{max}]$. The time step has a size of k and the asset step has a size of h.

The *explicit scheme* then takes the time derivative as forward difference of t (backward in physical time)

$$\frac{\partial V}{\partial t} = \frac{V_{i,j} - V_{i-1,j}}{k}. \tag{4}$$

It is accurate at order $O(k,h)$. However, it is stable and convergent only for

$$\lambda = \frac{k}{h^2} \leq 0.5. \tag{5}$$

That is, problems may occur for small time steps.

The *implicit scheme* takes the time derivative as backward difference of t (forward in physical time)

$$\frac{\partial V}{\partial t} = \frac{V_{i+1,j} - V_{i,j}}{k}. \tag{6}$$

It is accurate at order $O(k,h)$. Also, the implicit finite difference scheme is unconditionally stable.

The *Crank-Nicolson scheme* is the average of the explicit and implicit. It is accurate at order $O(k^2,h^2)$. As the implicit method, Crank-Nicolson is unconditionally stable. However, in the presence of points of discountinuity in the initial conditions or with special boundary values, small asset steps may introduce spurious ocsillations. The latter may even lead to obtaining negative option prices. A discussion on these effects can be found in [4], for example.

The scheme introduced by [3] is an interesting alternative to the previously described methods. It is exponentially fitted, based on a hyberbolic cotangent function. Consider the operator L defined as:

$$LV = -\frac{\partial V}{\partial t} + \mu(S,t)\frac{\partial V}{\partial S} + \sigma(S,t)\frac{\partial^2 V}{\partial S^2} + b(S,t)V, \tag{7}$$

where $\mu(S,t) = rS$, $\sigma(S,t) = \frac{1}{2}\sigma^2 S^2$ and $b(S,t) = -r$.

Replacing the derivatives, the fitted operator is defined by

$$L_k^h U_j^n = -\frac{U_j^{n+1} - U_j^n}{k} + \mu_j^{n+1}\frac{U_{j+1}^{n+1} - U_{j-1}^{n+1}}{2h} + \rho_j^{n+1}\frac{\delta_x^2 U_j^{n+1}}{h^2} + b_j^{n+1}U_j^{n+1}, \tag{8}$$

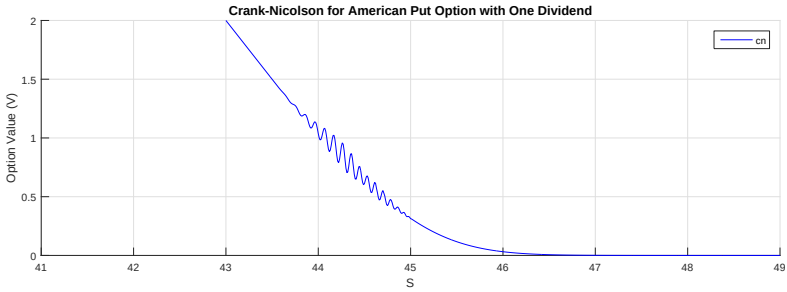


Figure 1: Crank-Nicolson scheme for an American put option.

where h and k are the space and time step, respectively. The factor ρ is defined as

$$\rho_j^{n+1} = \frac{\mu_j^{n+1} h}{2} coth \frac{\mu_j^{n+1} h}{2\sigma_j^{n+1}}. \tag{9}$$

The scheme is stable and consistent. Moreover, it converges regardless of the volatility size. It does not suffer from oscillations for extreme parameter values and behaves good for special conditions (such as barriers) and low volatility environment ([6]). The method is of order $O(k, h)$.

In what follows we apply dividends to the Duffy scheme and explore its compatibility with a few dividend policies.

3 Results

With an implementation based on [2], the discrete dividend extension was applied to the scheme suggested by Duffy. It handles low volatility levels and provides a smooth and oscillation-free solution when pricing American options. An example is run on an American put option with current underlying value of 40, strike = 40, rate = 0.1, volatility = 0.05, time to maturity of one year, and an expected dividend of size 5 in 0.6 years from evaluation date.

Figure 1 plots the result obtained by the Crank-Nicolson finite difference scheme. As one can easily see, the price function suffers from spurious oscillations. Then, on Figure 2 is provided the outcome of employing the Duffy scheme on the very same option. The undesired price fluctuations are gone and the result is a smooth function.

4 Discussion on Dividend Policies

When pricing derivatives on an underlying asset which pays out dividends, one has to deal with the shifts these payments cause to the underlying price. A broad review of different dividend policies can be found in [5]. Here we discuss some of them and show how the Duffy scheme fits into these policies.

Since the classical Black-Scholes option pricing model has become popular, practitioners and academics are suggesting various approaches to applying dividends to

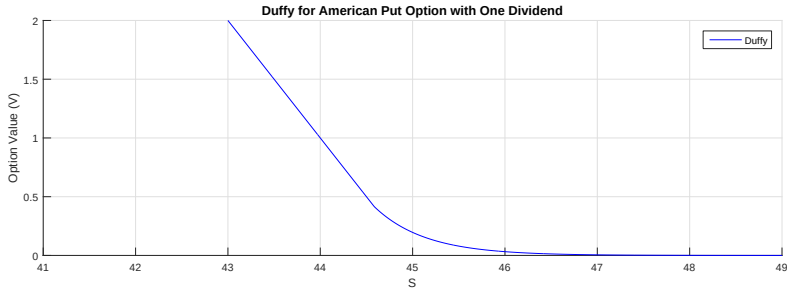


Figure 2: Duffy scheme for an American put option

the underlying asset price. The continuous dividends assumption is oversimplifying as longer term dividends are less predictable. Another simple model suggests updating the stock series with the discounted expected dividend which again shifts the price inaccurately for dividends far in the future. A separate group of policies is based on applying an adjustment to the volatility term. An approach which is inefficient in terms of computational time is employing non-recombining binomial trees.

It is interesting to review the policies suggested in case of a serious drop in the underlying value. That is, what should the numerical approach be if the asset price becomes lower than the declared dividend? If S denotes the underlying price at the time the dividend should be paid out and D is the declared dividend size, in the scenario just described we have that $S < D$. The two suggested policies for this case are *survivor* and *liquidator*.

According to the survivor policy, the company will pay no dividend to its shareholders. In other words, the company will have no ability to fulfill the declared payment but will still survive.

On the other hand, the liquidator policy adopts a more strict view. If the company finds itself in a situation when its share market price is lower than the declared dividend, it should pay out an amount equal to the current asset price. This is the policy encorporated in the Duffy scheme.

Despite having a marginal effect, the choice of a dividend policy in case of a severe drop in asset prices is a step towards having a fully specified model.

5 Conclusion

This short paper presented a brief outline of some of the most popular finite difference methods employed in option pricing, together with the main problems in their application. Then it was shown that the Duffy scheme is suitable for applying discrete dividends, as it does not suffer from numerical issues that may result in spurious oscillations. Finally, different dividend policies in option pricing approximations were discussed.

References

[1] F. Black and M. Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, (81):637–654, 1973.

[2] P. Brandimarte. *Numerical Methods in Finance: a MATLAB - based Introduction*. John Wiley & Sons, New York, 2001.

[3] D. Duffy. A critique of the Crank-Nicolson scheme, strengths and weaknesses for financial instrument pricing. *Wilmott Magazine*, pages 68–74, July 2004.

[4] D. Duffy. *Finite difference methods in financial engineering: A partial differential equation approach*. John Wiley & Sons, Chichester, 2006.

[5] E. Haug, J. Haug, and A. Lewis. Back to basics: A new approach to the discrete dividend problem. *Wilmott Magazine*, pages 37–47, September 2003.

[6] M. Milev and A. Tagliani. Low volatility options and numerical diffusion of finite difference schemes. *Serdica Mathematical Journal*, 36(3):223–236, 2010.

[7] P. Wilmott. *Derivatives: The theory and practice of financial engineering*. John Wiley & Sons, Chichester, University edition, 1998.

Estimation of the Offspring Mean Matrix in 2-Type Critical Galton-Watson Processes

Kristóf Körmendi*[1] and Gyula Pap[2]

[1] *MTA-SZTE Analysis and Stochastics Research Group, University of Szeged, Aradi vértanúk tere 1, H–6720 Szeged, Hungary.*
[2] *Bolyai Institute, University of Szeged, Aradi vértanúk tere 1, H–6720 Szeged, Hungary.*

Abstract: The well known Galton–Watson process can be generalized the following way; suppose the population consists of multiple types of individuals and allow the offsprings to have a different type than their parents. Denote the number of types by d. To describe this process we have to assign a d-dimensional offspring distribution to each of the types. The offspring mean matrix is a matrix whose columns consist of the expectation of these distributions.

In the single-type case we distinguish between subcritical, critical and supercritical processes based on the relation of the offspring mean to 1. For multi-type Galton–Watson processes we make the same distinction based on the spectral radius of the offspring mean matrix. Immigration can be introduced to the model exactly as in the single-type case, by adding i.i.d. random vectors representing the number of immigrants in each generation.

In Ispány et al. [1] we describe the asymptotic properties of the conditional least squares estimate of the offspring mean matrix for critical 2-type Galton-Watson processes with immigration under heavy restrictions on the structure of the matrix. This talk will focus on how can one replace the restrictions with more sensible assumptions, namely positive regularity of the matrix. These results are contained in [2].

Keywords: Galton–Watson process with immigration, conditional least squares estimator

AMS subject classifications: 60J80, 62F12

References

[1] Ispány, M., Körmendi, K., Pap, G. Asymptotic behavior of CLS estimators for 2-type doubly symmetric critical Galton-Watson processes with immigration *Bernoulli*, 20(4):2247–2277, 2014.

[2] Körmendi, K., Pap, G. Statistical inference of 2-type critical Galton-Watson processes with immigration
available on arXiv: arxiv.org/abs/1502.04900

*Corresponding author: kormendi@math.u-szeged.hu

Invariance Principle Under Self-Normalization for AR(1) Process

Jurgita Markevičiūtė*

Faculty of Mathematics and Informatics, Vilnius University
Naugarduko str. 24, LT-03225 Vilnius, Lithuania

Abstract: We investigate the polygonal line process built on the observations of the first order autoregressive process. We prove the functional limit theorem in the continuous functions space C[0, 1] under the certain self-normalization, assuming that innovations of the autoregressive process are in domain of attraction of normal distribution.

Keywords: autoregressive process, nearly nonstationary process, functional limit theorem, self-normalization, domain of attraction of normal distribution

AMS subject classifications: 60F17, 62M10

1 Introduction

We investigate the first order autoregressive process $(y_{n,k})$ defined by

$$y_{n,k} = \phi_n y_{n,k-1} + \varepsilon_k, \quad y_{n,0} = 0, \quad n \geq 0, \quad 0 \leq k \leq n, \tag{1}$$

where (ε_k) are i.i.d. random variables and $\varepsilon_1 \in DAN$[1], $E\varepsilon_1 = 0$ and $n(1 - \phi_n) \xrightarrow[n\to\infty]{} \infty$. If ϕ_n is a constant, then $|\phi_n| = |\phi| \leq 1$ and we have a stationary autoregressive process. If we set $\phi_n = 1 - \gamma_n/n$, $\gamma_n/n \to 0$ and $\gamma_n \to \infty$, as $n \to \infty$, then $\phi_n \xrightarrow[n\to\infty]{} 1$ and such process is called nearly nonstationary first order autoregressive process (see [1]). Though $(y_{n,k})$ is a triangular array, but for simplicity, we will omit index n and we will write $y_k = \phi_n y_{k-1} + \varepsilon_k$.

The aim of this paper is to investigate the convergence of polygonal line process built on observations (y_k) under the self-normalization. Such polygonal line processes, assuming that innovations have finite second moment, have been investigated by Markevičiūtė, Račkauskas and Suquet [2]. Here we assume that second moment does not exist and we will use certain self-normalization. Let us define polygonal line processes built on y_k's

$$S_n(t) = \sum_{k=1}^{[nt]} y_k + (nt - [nt])y_{[nt]+1}, \quad S_n(0) = 0, \quad t \in [0, 1]. \tag{2}$$

*Corresponding author: jurgita.markeviciute@mif.vu.lt

[1] DAN denotes domain of attraction of normal distribution.

and ε_k's

$$W_n(t) = \sum_{k=1}^{[nt]} \varepsilon_k + (nt - [nt])\varepsilon_{[nt]+1}, \quad W_n(0) = 0, \quad t \in [0,1]. \tag{3}$$

Processes $S_n(t)$ and $W_n(t)$ are defined in the continuous functions space $\mathrm{C}[0,1]$ endowed with the uniform norm

$$\|f\|_\infty = \sup_{0\le r\le 1} |f(r)|, \quad \text{for every} \quad f \in \mathrm{C}[0,1].$$

In what follows, $\xrightarrow[n\to\infty]{\mathcal{D}}$ denotes convergence in distribution and $\xrightarrow[n\to\infty]{\mathrm{P}}$ denotes convergence in probability.

We assume that $\varepsilon_1 \in DAN$, this means that there exists a sequence $b_n \to \infty$ such that

$$b_n^{-1} \sum_{k=1}^{n} \varepsilon_k \xrightarrow[n\to\infty]{\mathcal{D}} \mathfrak{N}(0,1). \tag{4}$$

Then by Gnedenko-Raikov's Theorem,

$$b_n^{-2} V_n^2 \xrightarrow[n\to\infty]{\mathrm{P}} 1, \quad \text{where} \quad V_n^2 = \sum_{k=1}^{n} \varepsilon_k^2.$$

Račkauskas and Suquet [3] proved that

$$V_n^{-1} W_n \xrightarrow[n\to\infty]{\mathcal{D}} W \quad \text{in} \quad \mathrm{C}[0,1],$$

where $W = (W(t), t \in [0,1])$ is a standard Wiener process. So from two latter results and Slutsky's Theorem, obviously follows that

$$b_n^{-1} W_n \xrightarrow[n\to\infty]{\mathcal{D}} W \quad \text{in} \quad \mathrm{C}[0,1]. \tag{5}$$

Further we state few useful facts (see for example [4]). If $\varepsilon_1 \in DAN$, then with normalizing sequence b_n as in (4), for each $\tau > 0$, one has

$$nP(|\varepsilon_1| > \tau b_n) \xrightarrow[n\to\infty]{} 0, \tag{6}$$

$$\frac{n}{b_n^2} E\varepsilon_1^2 \mathbf{1}_{\{|\varepsilon_1| \le \tau b_n\}} \xrightarrow[n\to\infty]{} 1, \tag{7}$$

$$\frac{n}{b_n} E\,|\varepsilon_1|\, \mathbf{1}_{\{|\varepsilon_1| > b_n\}} \xrightarrow[n\to\infty]{} 0. \tag{8}$$

Note that one may put $b_n = n^{1/2}\ell_n$, where ℓ_n is a slowly varying sequence.

2 Limit theorems

The main result of the paper is given in Theorem 15. We use as the normalization $\frac{1}{n}\sum_{k=1}^{n} S_n^2(k/n)$. As the limit we get the functional depending on the Wiener process.

Theorem 15. *Suppose* (y_k) *is defined by* (1) *and* (ε_k) *are i.i.d. random variables with* $\varepsilon_1 \in DAN$, $E\varepsilon_1 = 0$, b_n *is defined by* (4) *and* $n(1-\phi_n) \xrightarrow[n\to\infty]{} \infty$, *then*

$$\frac{S_n}{\sqrt{\frac{1}{n}\sum_{k=1}^{n} S_n^2(k/n)}} \xrightarrow[n\to\infty]{\mathcal{D}} \frac{W}{\int_0^1 W^2(s)\mathrm{d}s} \quad in \quad \mathrm{C}[0,1]. \tag{9}$$

To prove Theorem 15, we need few additional results that might be of independent interest. We start with the functional limit theorem in the continuous function space C[0, 1]. We prove that polygonal line process converge to Wiener process under the normalization $(1-\phi_n)/b_n$.

Theorem 16. *Suppose* (y_k) *is defined by* (1) *and* (ε_k) *are i.i.d. random variables with* $\varepsilon_1 \in DAN$, $E\varepsilon_1 = 0$, b_n *is defined by* (4) *and* $n(1-\phi_n) \xrightarrow[n\to\infty]{} \infty$, *then*

$$\frac{1-\phi_n}{b_n} S_n \xrightarrow[n\to\infty]{\mathcal{D}} W \quad in \quad \mathrm{C}[0,1]. \tag{10}$$

Next, we give another useful result.

Theorem 17. *Suppose* (y_k) *is defined by* (1) *and* (ε_k) *are i.i.d. random variables with* $\varepsilon_1 \in DAN$, $E\varepsilon_1 = 0$, b_n *is defined by* (4) *and* $n(1-\phi_n) \xrightarrow[n\to\infty]{} \infty$, *then*

$$\frac{1}{n}\sum_{k=1}^{n} \left(\frac{1-\phi_n}{b_n} S_n(k/n)\right)^2 \xrightarrow[n\to\infty]{\mathcal{D}} \int_0^1 W^2(s)\mathrm{d}s. \tag{11}$$

Finally we turn to a key point of the Theorem 16, that is to control the behaviour of $\max_{1\le k\le n} |y_k|$.

Lemma 1. *Suppose* (y_k) *is defined by* (1) *and* (ε_k) *are i.i.d. random variables with* $\varepsilon_1 \in DAN$, $E\varepsilon_1 = 0$, b_n *is defined by* (4) *and* $n(1-\phi_n) \xrightarrow[n\to\infty]{} \infty$, *then*

$$b_n^{-1} \max_{1\le k\le n} |y_k| \xrightarrow[n\to\infty]{\mathrm{P}} 0. \tag{12}$$

3 Proofs

First we will prove Lemma 1.

Proof. Lemma 1. We need to prove

$$P(b_n^{-1} \max_{1\le k\le n} |y_k| > \delta) \xrightarrow[n\to\infty]{} 0.$$

For this, let us introduce truncated random variables:

$$\begin{aligned} \varepsilon_j' &= \varepsilon_j \mathbf{1}_{\{|\varepsilon_j|\le b_n\}}, \quad \widehat{\varepsilon}_j = \varepsilon_j' - E\varepsilon_j' \\ \varepsilon_j'' &= \varepsilon_j \mathbf{1}_{\{|\varepsilon_j|> b_n\}}, \quad \widetilde{\varepsilon}_j = \varepsilon_j'' - E\varepsilon_j''. \end{aligned}$$

Then we obtain

$$P(b_n^{-1} \max_{1\le k\le n} |y_k| > \delta) \le nP(|\varepsilon_1| > \delta b_n) + P'_n,$$

where

$$P'_n = P\left(b_n^{-1} \max_{1\le k\le n} |y'_k| > \delta\right) \quad \text{and} \quad y'_k = \sum_{j=1}^{k} \phi_n^{k-j} \varepsilon'_j.$$

Now using (6), it only remains to show that $P'_n \xrightarrow[n\to\infty]{} 0$. For this we need to center all ε'_j and we get

$$P'_n \le P_n^1 + P_n^2,$$

where

$$P_n^1 = b_n^{-1} \max_{1\le k\le n} \left|\sum_{j=1}^{k} \phi_n^{k-j} E\varepsilon'_j\right|, \quad P_n^2 = P\left(b_n^{-1} \max_{1\le k\le n} \left|\sum_{j=1}^{k} \phi_n^{k-j} \widehat{\varepsilon}_j\right| > \delta\right).$$

Note that $E\varepsilon'_j = E\varepsilon''_j$, so

$$\begin{aligned} P_n^1 &= b_n^{-1} \max_{1\le k\le n} \left|\sum_{j=1}^{k} \phi_n^{k-j} E\varepsilon''_j\right| \le E\,|\varepsilon''_1|\, b_n^{-1} \max_{1\le k\le n} \left|\sum_{j=1}^{k} \phi_n^{k-j}\right| \\ &\le E\,|\varepsilon_1|\,\mathbf{1}_{\{|\varepsilon_1|>b_n\}} b_n^{-1}(1-\phi_n)^{-1} = \frac{n}{b_n} E\,|\varepsilon_1|\,\mathbf{1}_{\{|\varepsilon_1|>b_n\}} \cdot (n(1-\phi_n))^{-1} \xrightarrow[n\to\infty]{} 0, \end{aligned}$$

because of the conditions (8) and $n(1-\phi_n) \xrightarrow[n\to\infty]{} \infty$.

Next let us turn to P_n^2. Notice that $E(\widehat{\varepsilon}_1)^2 \le 4E\varepsilon_1^2 \mathbf{1}_{\{|\varepsilon_1|\le b_n\}}$, so we obtain

$$\begin{aligned} P_n^2 &\le \frac{1}{b_n^2\delta^2} E\left(\max_{1\le k\le n} \left|\sum_{j=1}^{k} \phi_n^{k-j} \widehat{\varepsilon}_j\right|\right)^2 \le \frac{1}{b_n^2\delta^2} \max_{1\le k\le n} \sum_{j=1}^{k} \phi_n^{2(k-j)} E(\widehat{\varepsilon}_j)^2 \\ &\le \frac{E(\widehat{\varepsilon}_1)^2}{b_n^2\delta^2}(1-\phi_n^2)^{-1} \le C b_n^{-2}(1-\phi_n)^{-1} E\varepsilon_1^2 \mathbf{1}_{\{|\varepsilon_1|\le b_n\}} \\ &= C\frac{n}{b_n^2} E\varepsilon_1^2 \mathbf{1}_{\{|\varepsilon_1|\le b_n\}} (n(1-\phi_n))^{-1} \xrightarrow[n\to\infty]{} 0, \end{aligned}$$

where C is a constant and convergence follows from conditions $n(1-\phi_n) \xrightarrow[n\to\infty]{} \infty$ and (7). So the proof is finished. □

Now we are ready to prove Theorem 16.

Proof. Theorem 16. We have (1), then summing up both sides by k we obtain:

$$\frac{1-\phi_n}{b_n} \sum_{k=1}^{[nt]} y_k = (y_0 - y_{[nt]}) b_n^{-1} + b_n^{-1} \sum_{k=1}^{[nt]} \varepsilon_k.$$

Taking into account (5) we obtain that it is enough to prove

$$\left\| \frac{1-\phi_n}{b_n} S_n - \frac{1}{b_n} W_n \right\|_\infty \xrightarrow[n\to\infty]{\mathrm{P}} 0.$$

It is easy to see

$$\left\| \frac{1-\phi_n}{b_n} S_n - \frac{1}{b_n} W_n \right\|_\infty = \sup_{0\le r\le 1} \left| \phi_n (y_0 - y_{[nr]}) b_n^{-1} \right| \le \max_{0\le k\le n} \left| \phi_n b_n^{-1} (y_0 - y_k) \right|.$$

Note that $|\phi_n| \le 1$, $\forall n$, also $y_0 = 0$ and

$$b_n^{-1} \max_{1\le k\le n} |y_k| \xrightarrow[n\to\infty]{\mathrm{P}} 0$$

by Lemma 1, so the proof is finished. □

The proof of Theorem 17 is based on the integral approximation by sums.

Proof. Theorem 17. Let us denote

$$Z_n^1 = \frac{1}{n} \sum_{k=1}^n \left(\frac{1-\phi_n}{b_n} S_n(k/n) \right)^2 \quad \text{and} \quad Z_n^2 = \int_0^1 \left(\frac{1-\phi_n}{b_n} S_n(t) \right)^2 \mathrm{d}t.$$

Then we have for any function $f \in \mathrm{C}[0,1]$

$$\left| \frac{1}{n} \sum_{k=1}^n f(k/n) - \int_0^1 f(s)\mathrm{d}s \right| = \left| \sum_{k=1}^n \int_{(k-1)/n}^{k/n} (f(k/n) - f(s))\mathrm{d}s \right|$$
$$\le \sum_{k=1}^n \int_{(k-1)/n}^{k/n} |f(k/n) - f(s)| \, \mathrm{d}s \le \omega_0 \left(f, \frac{1}{n} \right),$$

where $\omega_0\left(f, \frac{1}{n}\right)$ is modulus of continuity defined by

$$\omega_0 \left(f, \frac{1}{n} \right) = \sup_{|t-s|\le 1/n} |f(t) - f(s)|.$$

So we have

$$\left| Z_n^1 - Z_n^2 \right| \le \frac{1}{n} \omega_0 \left(\left(\frac{1-\phi_n}{b_n} S_n(t) \right)^2, 1 \right).$$

By Theorem 16 we have that $\left(\frac{1-\phi_n}{b_n} S_n(t) \right)$ is tight, thus

$$\left| Z_n^1 - Z_n^2 \right| \xrightarrow[n\to\infty]{\mathrm{P}} 0.$$

Finally, since integral is a continuous function, so we have by Theorem 17

$$Z_n^2 \xrightarrow[n\to\infty]{\mathcal{D}} \int_0^1 W^2(s)\mathrm{d}s$$

and the proof is finished. □

Finally we can prove the main result.

Proof. Theorem 15. The result (9) follows from Theorems 16 and 17 and a continuous mapping theorem. Note that (9) exclude the degenerated case $P(\varepsilon_1 = 0) = 1$, so that almost surely $\sum_{k=1}^{n} S_n^2(k/n) > 0$ for large enough n. To use continuous mapping theorem correctly, let us define functional

$$F(x)(t) = \frac{x(t)}{\sqrt{\int_0^1 x^2(s)\mathrm{d}s}}$$

and denote D_F the set of discontinuity points of F. We need to show that $P(W \in D_f) = 0$. For F the only possible discontinuities appears when $X = 0$, thus $P(W \in D_f) = 0$ is equivalent to $P(W = 0) = 0$ and this is a well known result. So the proof is finished. □

References

[1] L. Giraitis and P. Phillips. Uniform limit theory for stationary autoregression. *Journal of time series analysis*, 27(1):51–60, 2006.

[2] J. Markevičiūtė, and A. Račkauskas, and Ch. Suquet. Functional limit theorems for sums of nearly nonstationary processes. *Lithuanian Mathematical Journal*, 52(3):282–296, 2012.

[3] A. Račkauskas, and Ch. Suquet. Invariance principles for adaptive self-normalized partial sums processes. *Stochastic Processes and their Applications*, 95(1):63–81, 2001.

[4] A. Račkauskas, and Ch. Suquet. Functional central limit theorems for self-normalized partial sums of linear processes. *Lithuanian Mathematical Journal*, 51(2):251–259, 2011.

ICA Based on Fourth Moments

Jari Miettinen[*1], **Klaus Nordhausen**[2], **Hannu Oja**[2] **and Sara Taskinen**[1]

[1]*Department of Mathematics and Statistics, University of Jyvaskyla, Finland*
[2]*Department of Mathematics and Statistics, University of Turku, Finland*

Abstract: In independent component analysis it is assumed that the components of the observed random vector are linear combinations of latent independent random variables, and the aim is then to find an estimate for a transformation matrix back to these independent components. This paper summarizes some contents of [8], where the statistical properties of four well-known estimation procedures based on the use of fourth moments are studied in detail.

Keywords: affine equivariance, FastICA, FOBI, JADE, kurtosis
AMS subject classifications: 62H05, 62H10, 62H12

1 Introduction

The basic independent component (IC) model assumes that the components of the p-variate observed vector $\boldsymbol{x}_i$ are linear combinations of the p mutually independent latent components of $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ip})$. The model is written as

$$\boldsymbol{x}_i = \boldsymbol{\mu} + \boldsymbol{\Omega}\boldsymbol{z}_i, \quad i = 1, \ldots, n, \tag{1}$$

and the full rank $p \times p$ matrix $\boldsymbol{\Omega}$ is called the mixing matrix. The location vector $\boldsymbol{\mu}$ is a nuisance parameter and the aim is, using only a random sample $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, to find an estimate of an unmixing matrix $\boldsymbol{W}$ such that $\boldsymbol{W}\boldsymbol{x}$ has independent components. Then $\boldsymbol{z}$ and $\boldsymbol{W}\boldsymbol{x}$ would differ only by order, signs, and scales of the components. General assumptions for the latent components to be identifiable are

(A1) the components $z_{i1}, \ldots, z_{ip}$ of $\boldsymbol{z}_i$ are independent,

(A2) second moments exist, $E(\boldsymbol{z}_i) = \boldsymbol{0}$ and $E(\boldsymbol{z}_i\boldsymbol{z}_i') = \boldsymbol{I}_p$, and

(A3) at most one of the components $z_{i1}, \ldots, z_{ip}$ of $\boldsymbol{z}_i$ has a normal distribution.

The last assumption is needed because of the fact that the components of any orthogonal transformation of two independent Gaussian random variables are still Gaussian and mutually independent. The assumption on the existence of second moments can be replaced with some other way of fixing the scales of the components, see, e.g., [7]. When dealing with the fourth moments, (A2) is of course not restrictive at all. For more discussion on model (1) and on more general IC models, see [6].

Under (A2), a natural first step is so called whitening, i.e., standardization $\boldsymbol{x} \to \boldsymbol{x}_{st} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu})$, where $\boldsymbol{\Sigma} = \boldsymbol{\Omega}\boldsymbol{\Omega}'$ is the covariance matrix of the random

*Corresponding author: jari.p.miettinen@jyu.fi

variable $\boldsymbol{x}$ from model (1) . Then $\boldsymbol{z} = \boldsymbol{U}\boldsymbol{x}_{st}$ for some orthogonal matrix $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p)'$. Thus after estimating $\boldsymbol{\Sigma}$, the estimation problem is simplified to the estimation problem of an orthogonal matrix $\boldsymbol{U}$ only, and then $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Sigma}^{-1/2}$.

In this paper we review the use of fourth moments in ICA [8]. Section 2 presents four well-known estimators and Section 3 displays statistical properties of one of them. The performance of the four estimators is discussed briefly in Section 4.

2 Estimators

2.1 Independent component functionals

In the following definition, the population quantity which we wish to estimate is defined as the value of an independent component functional $\boldsymbol{W}(F_{\boldsymbol{x}})$, where $F_{\boldsymbol{x}}$ denotes the distribution function of $\boldsymbol{x}$.

Definition 4. The $p \times p$ matrix-valued functional $\boldsymbol{W}(F)$ is said to be an independent component (IC) functional if (i) $\boldsymbol{W}(F_{\boldsymbol{x}})\boldsymbol{x}$ has independent components in the IC model and (ii) $\boldsymbol{W}(F_{\boldsymbol{x}})$ is affine equivariant in the sense that

$$\boldsymbol{W}(F_{\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}}) = \boldsymbol{W}(F_{\boldsymbol{x}})\boldsymbol{A}^{-1}$$

for all nonsingular $p \times p$ matrices $\boldsymbol{A}$ and for all p-vectors $\boldsymbol{b}$.

Notice that (ii) implies that in the independent component model, $\boldsymbol{W}(F_{\boldsymbol{x}})\boldsymbol{x}$ does not depend on the specific choices of $\boldsymbol{z}$ and $\boldsymbol{\Omega}$, up to the signs and the order of the components. The sample version is $\boldsymbol{W}(\boldsymbol{X}) = \boldsymbol{W}(F_n)$, where F_n denotes the empirical distribution function of $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$.

2.2 Univariate kurtosis and ICA

The classical kurtosis measures of a random variable x with mean μ and variance σ^2 are $\beta = E\left([x-\mu]^4\right)/\sigma^2$ and $\kappa = \beta - 3$. For standardized variable $z = (x-\mu)/\sigma$, then $\beta = E(z^4)$. Normally distributed x has $\kappa = 0$, and thus, the deviation of κ from zero indicates non-normality of x. The succes of many ICA estimators is motivated heuristically by the central limit theorem, stating that a mixture of mutually independent random variables is more normal than the original variables, and the idea is to find such orthogonal $\boldsymbol{U}$ that the components of $\boldsymbol{U}\boldsymbol{x}_{st}$ are maximally non-normal. When the fourth moments are used as the measure of non-normality, the consistency of estimators can be proved as in the following theorem, see [8].

Theorem 18. *Let the components of $\boldsymbol{z} = (z_1, \ldots, z_p)'$ be independent and standardized so that $E(\boldsymbol{z}) = \boldsymbol{0}$ and $Cov(\boldsymbol{z}) = \boldsymbol{I}_p$, and assume that at most one of the kurtosis values $\kappa_i = E(z_i^4) - 3$, $i = 1, \ldots, p$, is zero. Then the following inequalities hold true.*

(i) $$|E((\boldsymbol{u}'\boldsymbol{z})^4) - 3| \leq \max\left\{|E(z_1^4) - 3|, \ldots, |E(z_p^4) - 3|\right\}$$

for all $\boldsymbol{u}$ such that $\boldsymbol{u}'\boldsymbol{u} = 1$. The equality holds only if $\boldsymbol{u} = \boldsymbol{e}_i$ for i such that $|E(z_i^4) - 3| = \max\left\{|E(z_1^4) - 3|, \ldots, |E(z_p^4) - 3|\right\}$, and

(ii) $$|E[(\boldsymbol{u}_1'\boldsymbol{z})^4] - 3| + \cdots + |E[(\boldsymbol{u}_p'\boldsymbol{z})^4] - 3| \leq |E[z_1^4] - 3| + \cdots + |E[z_p^4] - 3|$$

for all orthogonal matrices $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p)'$. *The equality holds only if* $\boldsymbol{U} = \boldsymbol{JP}$ *for some sign-change matrix* $\boldsymbol{J}$ *and permutation matrix* $\boldsymbol{P}$.

Items (i) and (ii) suggest the deflation-based [4] and symmetric [5] FastICA functionals, respectively.

Definition 5. The deflation-based FastICA functional is $\boldsymbol{W}(F_{\boldsymbol{x}}) = \boldsymbol{U}\boldsymbol{\Sigma}^{-1/2}$, where $\boldsymbol{\Sigma} = Cov(\boldsymbol{x})$ and the rows of an orthogonal matrix $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p)'$ are found one by one by maximizing $|E((\boldsymbol{u}_k'\boldsymbol{x}_{st})^4) - 3|$ under the constraint that $\boldsymbol{u}_k'\boldsymbol{u}_k = 1$ and $\boldsymbol{u}_j'\boldsymbol{u}_k = 0$, $j = 1, \ldots, k-1$.

Definition 6. The symmetric FastICA functional is $\boldsymbol{W}(F_{\boldsymbol{x}}) = \boldsymbol{U}\boldsymbol{\Sigma}^{-1/2}$, where $\boldsymbol{\Sigma} = Cov(\boldsymbol{x})$ and $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p)'$ maximizes

$$|E((\boldsymbol{u}_1'\boldsymbol{x}_{st})^4) - 3| + \ldots + |E((\boldsymbol{u}_p'\boldsymbol{x}_{st})^4) - 3|$$

under the constraint that $\boldsymbol{U}\boldsymbol{U}' = \boldsymbol{I}_p$.

Later, new estimators with different non-normality measures have been introduced under the name of FastICA, but not all of them have been proved to be consistent. If the sample size is small, finding the FastICA estimates is occasionally difficult due to convergence problems.

2.3 Multivariate kurtosis and ICA

Define, for any $p \times p$ matrix $\boldsymbol{A}$, $\boldsymbol{B}(\boldsymbol{A}) = E\left(\boldsymbol{x}_{st}\boldsymbol{x}_{st}'\boldsymbol{A}\boldsymbol{x}_{st}\boldsymbol{x}_{st}'\right)$, and further,

$$\boldsymbol{B}^{ij} = \boldsymbol{B}(\boldsymbol{E}^{ij}), \quad i, j = 1, \ldots, p, \quad \text{and} \quad \boldsymbol{B} = \boldsymbol{B}(\boldsymbol{I}_p) = \sum_{i=1}^{p} \boldsymbol{B}^{ii},$$

where $\boldsymbol{E}^{ij}$ denotes the $p \times p$ matrix with ijth element one and others zero.

The kurtosis matrix $\boldsymbol{B}$ is diagonal for $\boldsymbol{x}_{st} = \boldsymbol{z}$ with independent components, and one of the earliest solutions to the independent component problem, FOBI (fourth order blind identification) [1], uses simultaneous diagonalization of the covariance matrix $\boldsymbol{\Sigma}$ and the kurtosis matrix $\boldsymbol{B}$.

Definition 7. The FOBI functional is $\boldsymbol{W}(F_{\boldsymbol{x}}) = \boldsymbol{U}\boldsymbol{\Sigma}^{-1/2}$, where $\boldsymbol{\Sigma} = Cov(\boldsymbol{x})$ and the rows of $\boldsymbol{U}$ are the eigenvectors of $\boldsymbol{B} = E\left(\boldsymbol{x}_{st}\boldsymbol{x}_{st}'\boldsymbol{x}_{st}\boldsymbol{x}_{st}'\right)$.

FOBI is a fast and simple method, but it has a major weakness that it is consistent only if the kurtosis values of the independent components are distinct.

Unlike the kurtosis matrix $\boldsymbol{B}$ in FOBI, the matrices $\boldsymbol{B}^{ij}$ are not diagonal at $\boldsymbol{x}_{st} = \boldsymbol{z}$. However, the fourth order cumulant matrices

$$\boldsymbol{C}^{ij} = \boldsymbol{B}^{ij} - \boldsymbol{E}^{ij} - (\boldsymbol{E}^{ij})' - \mathrm{tr}(\boldsymbol{E}^{ij})\boldsymbol{I}_p, \quad i, j = 1, \ldots, p,$$

are diagonal. The IC functional based on Joint Approximate Diagonalization of these (Eigen)matrices is called JADE [2], and it is defined as follows.

Definition 8. The JADE functional is $\boldsymbol{W}(F_{\boldsymbol{x}}) = \boldsymbol{U}\boldsymbol{\Sigma}^{-1/2}$ where $\boldsymbol{\Sigma} = Cov(\boldsymbol{x})$ and the orthogonal matrix U maximizes

$$\sum_{i=1}^{p}\sum_{j=1}^{p} ||\mathrm{diag}(\boldsymbol{U}\boldsymbol{C}^{ij}\boldsymbol{U}')||^2,$$

where $\mathrm{diag}(A)$ is a diagonal matrix with the same diagonal elements as A, and $||\cdot||$ is the matrix norm.

The affine equivariance of the JADE estimate is not obvious, because the matrices $\boldsymbol{C}^{ij}$ are not orthogonal equivariant. Interestingly, the joint diagonalization of $\boldsymbol{C}^{ij},\ i,j = 1,\ldots,p$ is orthogonal equivariant, which was rigorously proved in [8], and the affine equivariance of the estimate folllows.

There are several algorithms for an approximate diagonalization of several symmetric matrices, but the statistical properties of the corresponding estimates are not known. The most popular algorithm is perhaps the Jacobi rotation algorithm suggested in [3]. It appeared in our simulations that the Jacobi rotation algorithm is computationally much faster and always provided the same solution as our algorithm based on the estimating equations, see Section 3.

The JADE estimate requires computation of $p(p+1)/2$ fourth moment matrices, which means that the computational load grows quickly with the number of components. Recently, a faster and asymptotically equivalent estimate with JADE was found in [9], where the joint use of third and fourth moments in ICA was studied. The squared symmetric FastICA is obtained, when the absolute values of the regular symmetric FastICA are replaced by squares, i.e., we maximize, under the orthogonality constraint,

$$(E((\boldsymbol{u}_1'\boldsymbol{x}_{st})^4) - 3)^2 + \ldots + (E((\boldsymbol{u}_p'\boldsymbol{x}_{st})^4) - 3)^2.$$

3 Asymptotic properties of the JADE estimate

Let $\boldsymbol{Z} = (\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n)$ be a random sample from a distribution satisfying (A1), (A2), and in addition to (A3), at least one of the following assumptions.

(A4) The fourth moments of $\boldsymbol{z}$ exist, and at most one of the kurtosis values $\kappa_k = E(z_{ik}^4)$, $k = 1,\ldots,p$, is zero.

(A5) The fourth moments of $\boldsymbol{z}$ exist and are distinct.

For the independent components $z_{ik},\ k = 1,\ldots,p$, write also $\gamma_k = E(z_{ik}^3)$ and $\sigma_k^2 = Var(z_{ik}^3)$.

The limiting distributions of all four unmixing matrix estimates based on fourth moments depend on the joint limiting distribution of

$$\sqrt{n}\,\widehat{s}_{kl} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_{ik}z_{il} \quad \text{and} \quad \sqrt{n}\,\widehat{r}_{kl} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(z_{ik}^3 - \gamma_k)z_{il},$$

$k \neq l = 1, \ldots, p$. For the FOBI estimate we would also need

$$\sqrt{n}\,\widehat{r}_{mkl} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} z_{im}^2 z_{ik} z_{il},$$

for distinct $k, l, m = 1, \ldots, p$.

The JADE estimate is $\widehat{\boldsymbol{W}} = \widehat{\boldsymbol{U}}\widehat{\boldsymbol{\Sigma}}^{-1/2}$, where $\widehat{\boldsymbol{U}}$ solves the estimating equations $\boldsymbol{u}_i'\boldsymbol{T}(\boldsymbol{u}_j) = \boldsymbol{u}_j'\boldsymbol{T}(\boldsymbol{u}_i)$ and $\boldsymbol{u}_i'\boldsymbol{u}_j = \delta_{ij},\ i, j = 1, \ldots, p$. Here δ_{ij} is the Kronecker delta and

$$\boldsymbol{T}(\boldsymbol{u}) = \sum_{i=1}^{p}\sum_{j=1}^{p} (\boldsymbol{u}'\boldsymbol{C}^{ij}\boldsymbol{u})\boldsymbol{C}^{ij}\boldsymbol{u}.$$

The following theorem for the JADE estimate was given in [8], where you can also find the corresponding results for the other estimates, and the references to the articles where they were first given in. We consider the distribution of $\widehat{\boldsymbol{W}} = \boldsymbol{W}(\boldsymbol{Z})$, i.e., the case $\boldsymbol{\Omega} = \boldsymbol{I}_p$. Due to affine equivariance of the estimate, we have $\boldsymbol{W}(\boldsymbol{X}) = \boldsymbol{W}(\boldsymbol{Z})\boldsymbol{\Omega}^{-1}$ in the general case.

Theorem 19. *Let $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ be a random sample from a distribution of $\boldsymbol{z}$ with bounded eighth moments satisfying the assumptions (A1), (A2) and (A4). Then there is a sequence of solutions $\widehat{\boldsymbol{W}}$ such that $\widehat{\boldsymbol{W}} \to_P \boldsymbol{I}_p$ and*

$$\begin{aligned}
\sqrt{n}\,(\widehat{w}_{kk} - 1) &= -1/2\sqrt{n}\,(\widehat{s}_{kk} - 1) + o_P(1), \quad k = l, \quad \textit{and} \\
\sqrt{n}\,\widehat{w}_{kl} &= \frac{\kappa_k\sqrt{n}\,\widehat{r}_{kl} - \kappa_l\sqrt{n}\,\widehat{r}_{lk} + \left(3\kappa_l - 3\kappa_k - \kappa_k^2\right)\sqrt{n}\,\widehat{s}_{kl}}{\kappa_k^2 + \kappa_l^2} + o_P(1), \\
&\quad k \neq l.
\end{aligned}$$

Notice that the limiting distribution of $\sqrt{n}\,\widehat{w}_{kl}$ depends only on kth and lth independent components. This holds true also for deflation-based and symmetric FastICA. On the contrary, $\sqrt{n}\,\widehat{w}_{kl}$ of the FOBI estimate depends also on the other components through $\sqrt{n}\,\widehat{r}_{mkl},\ m \neq k, l$. The limiting distributions of the diagonal elements are the same for all four estimates.

Under the assumptions of Theorem 19, the central limit theorem gives the asymptotic normality of the estimate, and the componentwise asymptotic variances are

$$\begin{aligned}
ASV(\widehat{w}_{kk}) &= (\kappa_k + 2)/4, \quad \text{and} \\
ASV(\widehat{w}_{kl}) &= \left(\kappa_k^2 + \kappa_l^2\right)^{-2}\left(\kappa_k^2(\sigma_k^2 - \kappa_k^2 - 6\kappa_k - 9) + \kappa_l^2(\sigma_l^2 - 6\kappa_l - 9)\right), \quad k \neq l
\end{aligned}$$

Since the effect of the term $\sqrt{n}\,\widehat{r}_{mkl}$ is rather minor in FOBI, the comparison of the estimates can be reduced to the comparison of the sum $ASV(\widehat{w}_{kl}) + ASV(\widehat{w}_{lk})$ for selected kth and lth component distributions.

4 Comparison of the estimates

Besides the fact that JADE and squared symmetric FastICA are asymptotically equivalent, we have the following facts, when $\boldsymbol{\Omega} = \boldsymbol{I}_p$.

1. $\sqrt{n}\,\widehat{w}_{kl}$ of the symmetric FastICA estimate and that of the JADE estimate are asymptotically equivalent if the kth and lth marginal distributions are the same,
2. If the independent components are identically distributed, then the symmetric FastICA and JADE estimates are asymptotically equivalent. In this case, the sum of the asymptotic variances of the off-diagonal elements is one half of that of the deflation-based FastICA estimate. The FOBI estimate fails in this case.
3. $ASV(\widehat{w}_{kl})$ of the FOBI estimate is always larger than or equal to that for symmetric FastICA, $k \neq l$. The variances are equal when $p = 2$ and $\kappa_k > 0 > \kappa_l$.
4. $\sqrt{n}\,\widehat{w}_{kp}$ of the deflation-based FastICA estimate and of the JADE estimate are asymptotically equivalent if the pth marginal distribution is normal.

Wide comparisons of the asymptotic variances in [8] expressed that JADE and the symmetric FastICA perform best in most cases, and their asymptotic variances are quite close to each other. The main difference between these two estimators appears when one of the components has a normal distribution. Then JADE outperforms symmetric FastICA.

Acknowledgements: This research was supported by the Academy of Finland (grants 251965, 256291 and 268703).

References

[1] J. F. Cardoso. Source separation using higher order moments. *Proceedings of IEEE International Conference on Accoustics, Speech and Signal Processing, Glasgow, UK*, 2109–2112, 1989.

[2] J.F. Cardoso and A. Souloumiac. Blind beamforming for non Gaussian signals. *IEE Proceedings-F*, 140:362–370, 1993.

[3] D. B. Clarkson. A least squares version of algorithm AS 211: The F-G diagonalization algorithm. *Applied Statistics*, 37:317–321, 1988.

[4] A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9:1483–1492, 1997.

[5] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10:626–634, 1999.

[6] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis.* John Wiley and Sons, New York, 2001.

[7] P. Ilmonen and D. Paindaveine. Semiparametrically efficient inference based on signed ranks in symmetric independent component models. *The Annals of Statistics*, 39:2448–2476, 2011.

[8] J. Miettinen, S. Taskinen, K. Nordhausen, and H. Oja. Fourth moments and independent component analysis. *Accepted for publication in Statistical Science. Preprint available as arXiv*, 2015.

[9] J. Virta, K. Nordhausen, and H. Oja, Joint use of third and fourth moments in independent component analysis. Submitted, 2015.

Computational Aspects of Parameter Estimation in Ordinary Differential Equation Systems

Frederik Riis Mikkelsen*

Department of Mathematical Sciences, University of Copenhagen, Denmark

Abstract: Ordinary differential equation (ODE) systems are widely applicable in many branches of the natural sciences. They are especially valuable for analysing entire networks of processes with no internal noise. Though simple from a statistical point of view, the applicability of these models are usually hindered by their computational complexity. In this work I present a selection of current methods to cope with the computational aspects of estimating parameters in ODE systems. Based on some of these methods, I present an algorithm for finding maximum likelihood estimates (MLE) with certain computational qualities.

Keywords: ODE systems, parameter estimation, non-linear least squares, computational statistics

AMS subject classifications: 62J02

1 Introduction

We have in mind a d-dimensional ordinary differential equation system:

$$\dot{x} = f(x, \theta), \qquad x \in \mathbb{R}^d \tag{1}$$

parametrised by a p-dimensional vector $\theta \in \mathbb{R}^p$. For given θ, a solution to (1) is a function $\psi_\theta : \mathbb{R} \to \mathbb{R}^d$, such that

$$\psi_\theta(t) = \psi_\theta(0) + \int_0^t f(\psi_\theta(s), \theta)\, ds, \qquad \text{for all } t \in \mathbb{R}. \tag{2}$$

We observe the state of the system at discrete time points $0 = t_1 < t_2 < ... < t_n$ with independent Gaussian noise:

$$y_j = \psi_\theta(t_j) + \varepsilon_j, \qquad \varepsilon_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2 I_d) \tag{3}$$

for $j = 1, ..., n$. The negative log-likelihood is directly available (σ^2 is omitted):

$$\ell_y(\theta) = \frac{1}{2} \sum_{j=1}^{n} \|y_j - \psi_\theta(t_j)\|_2^2. \tag{4}$$

This modelling framework for ODE systems is therefore quite simple from a statistical point of view. However, as explained below, optimising (4) is rather difficult from a computational angle.

*Corresponding author: frm@math.ku.dk

2 The optimisation problem

Finding the maximum likelihood estimator reduces to solving a non-linear least squares problem, due to (4). However, evaluating the likelihood requires solutions of the underlying ODE system. Various numerical methods for finding approximative solutions exist, but are relatively time consuming. Specifically, employing an explicit Runge-Kutta scheme of size s (see e.g. section 17 in [5] for details) the number of evaluations of f is $\frac{sT}{\delta}$. Here δ is the step size and T is the time span. For such a scheme the global truncation error is $\mathcal{O}\left(T\delta^p\right)$, for some scheme-dependent $p \leq s$. Using an implicit Runge-Kutta scheme, instead, leads to a smaller global truncation error, but raises the number of evaluations of f and is $\mathcal{O}\left(\frac{dsT}{\delta}\right)$ in best case scenarios.

The number of f-evaluations is substantial for assessing the computational complexity of evaluating ℓ_y. Though linear in each variable (considering $1/\delta$ as measuring the mesh), the number of f-evaluations is typically large. Consequently, in order to optimise (4) efficiently, a minimal number of evaluations of ℓ_y is preferable, especially when the observed time points cover a large time span or the ODE system is stiff.

3 Methods

3.1 Gauss-Newton approach (shooting)

From a numerical optimisation perspective, ℓ_y has the valuable property of being a sum of squares. Thus the classical Gauss-Newton algorithm is typically a first choice for the optimisation scheme. The Gauss-Newton algorithm has the same rate of convergence as most second order approximation algorithms, but requires no computations of the hessian matrix (see e.g. section 10 in [6] for details). However, calculating the gradient of ℓ_y:

$$\nabla_\theta \ell_y(\theta) = -\sum_{j=1}^{n} (y_j - \psi_\theta(t_j))' D_\theta \psi(t_j) \tag{5}$$

amounts to deriving $D_\theta \psi$. This differential is typically only available as a solution to the matrix differential equation system

$$\dot{D_\theta \psi} = \frac{\partial f}{\partial x}(\psi(t), \theta) D_\theta \psi + \frac{\partial f}{\partial \theta}(\psi(t), \theta). \tag{6}$$

Consequently, employing the Gauss-Newton algorithm requires solving (1) and (6) simultaneously at each step. Using an explicit Runge-Kutta scheme of size s, this amounts to evaluating f, $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial \theta}$, $\frac{sT}{\delta}$ times at each step of the optimisation. Subsequently, for large and complex systems, evaluating the gradient of ℓ_y is either extremely time consuming or close to impossible.

There are numerous variations of the above approach (see, e.g., [3] for a more sophisticated version). They are often referred to as *shooting* methods, inspired by the shooting method from boundary value problems. These methods typically

rely on general optimisation algorithms and will therefore not exploit all essential features of ODE systems. The remaining algorithms incorporate these features by considering, e.g., the functional nature of the data.

3.2 Generalised smoothing approach (collocation)

Certain implicit Runge-Kutta schemes are so-called *collocation* methods. They rely on the principle that an approximative solution to (1) can be found in some finite dimensional function space, typically spanned by a set of spline functions. The collocation method therefore amounts to finding an element of the function space that satisfy (1) in some pre-specified time points, called *collocation points*. This approach is the inspiration to various parameter estimation methods in ODE systems. The following method is due to Ramsay et. al, see [4]:

This approach relies on the approximation of ψ_θ given by

$$\psi_\theta(t) \approx \varphi(t)'\widehat{c}_\theta \qquad \text{for } t \in [0, t_n]. \tag{7}$$

Here φ is a vector of univariate basis functions, which combined with the vector $\widehat{c}_\theta$ of coefficients yields an approximative solution to (1). The parameter dependence $\theta \mapsto \psi_\theta$ is therefore passed on to $\theta \mapsto \widehat{c}_\theta$. The least squares criterion:

$$J(c, \theta) = \sum_j \|y_j - \varphi(t_j)'c\|_2^2 + \lambda \int_{t_1}^{t_n} \|\dot{\varphi}(t)'c - f(\varphi(t)'c, \theta)\|_2^2 \, dt \tag{8}$$

is proposed, where $\lambda > 0$ is a tuning parameter shifting the weight between the data fitting criterion and the so-called *fidelity measure* of $\varphi' c$. Applying profiling methods to (8), $\widehat{c}_\theta$ appears as the minimum of $c \mapsto J(c, \theta)$. If f is linear in x, the minimisation problem reduces to a linear least squares problem, thus providing an analytical expression for $\theta \mapsto \widehat{c}_\theta$.

By introducing this approach, some of the tools of functional data analysis is suddenly available, which provide new insightful views on the estimation problem. However, it is worth considering the influence of the choice of φ on the inference. Moreover, the relation between optimising a family of semi-norms parametrised by λ (the criterion J in (8)) and the actual MLE defined through (4) is not completely clear. Finally, for non-linear systems, optimising $c \mapsto J(c, \theta)$ and $\theta \mapsto J(\widehat{c}_\theta, \theta)$ using gradient based methods still require evaluating $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial \theta}$ many times (depending on how the integral in (8) is approximated).

3.3 Gradient/integral matching

The core principle of this method is: if the whole noiseless curve ψ is observed, then θ can be inferred by minimising

$$\int_{t_1}^{t_n} \left\| \dot{\psi}(t) - f(\psi(t), \theta) \right\|_2^2 dt, \quad \text{or} \quad \int_{t_1}^{t_n} \left\| \psi(t) - \psi(t_1) - \int_{t_1}^{t} f(\psi(s), \theta) ds \right\|_2^2 dt. \tag{9}$$

We therefore consider the estimator, that takes a non-parametric estimate of ψ, $\widehat{\psi}$, and returns the value of θ minimising (9):

$$\widehat{\psi} \mapsto \arg\min_{\theta} \int_{t_1}^{t_n} \left\| \widehat{\psi}(t) - \psi(t_1) - \int_{t_1}^{t} f(\widehat{\psi}(s), \theta) ds \right\|_2^2 dt. \tag{10}$$

If $\psi(t_1)$ is unknown, it can be included in the parameter vector θ. In [1] the author proves that if the above map is applied to a consistent non-parametric estimator, the resulting estimator of θ is also consistent, under mild regularity assumptions. Additionally, he also finds conditions for asymptotic normality.

This approach truly flourishes when applied to systems in which f is linear in θ (and not necessarily linear in x). In such cases (10) reduces to a linear least squares problem, which can be solved even for very large and complex systems, i.e., for large n and p. Furthermore, one can introduce, e.g., ℓ^1-penalties to (10) and apply the method to systems with $p >> nd$ and still have computationally stable methods for finding solutions.

Brewer et al. ([2]) proposed an iterative procedure exploiting the qualities of this type of gradient matching. More precisely, they consider a fitting criterion resembling that of [4]:

$$\sum_j \|y_j - \varphi(t_j)'c\|_2^2 + \lambda \sum_r \|\dot{\varphi}(t_r)'c - f(\varphi(t_r)'\widetilde{c}, \theta)\|_2^2 \tag{11}$$

where r is allowed to run over a finer (or coarser) grid than j. The iterations consist of letting $\widetilde{c}$ be fixed and then estimate (c, θ) as the linear least squares estimates of (11). The estimate of c then enters as $\widetilde{c}$ in the next iteration. By applying gradient matching iteratively one avoids choosing a specific $\widehat{\psi}$, as opposed to a non-iterative gradient matching.

Similarly to the generalised smoothing approach, this method has the following important strength: the optimisation problem and the ODE-solution problem are separated. Thus evaluating the fitting criterion is inexpensive and evaluating the gradient (typically) requires less calculations of $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial \theta}$.

The method described above is applicable to many non-trivial systems and can handle large and sparse models. However, there are things to consider: the iterative procedure is still dependent on the choice of φ. It is also unclear how optimising the criterion (11) is related to the MLE given through (4). Finally, it is nontrivial whether this sequence of iterative estimates of (c, θ) converge to the optimum of (11) (if it converges at all).

4 Combining algorithms

Returning to the original problem of minimising (4), we required relatively few evaluations of ℓ_y and $\nabla \ell_y$. In this section we consider a new algorithm based on the above, which yields the actual MLE (a quality of the shooting methods) and still exploits the computationally attractive aspects of gradient matching.

Firstly, given a current estimate of θ, denoted θ_k in the iterative procedure, we calculate an approximative solution curve ψ_{θ_k}. Then we perform integral matching

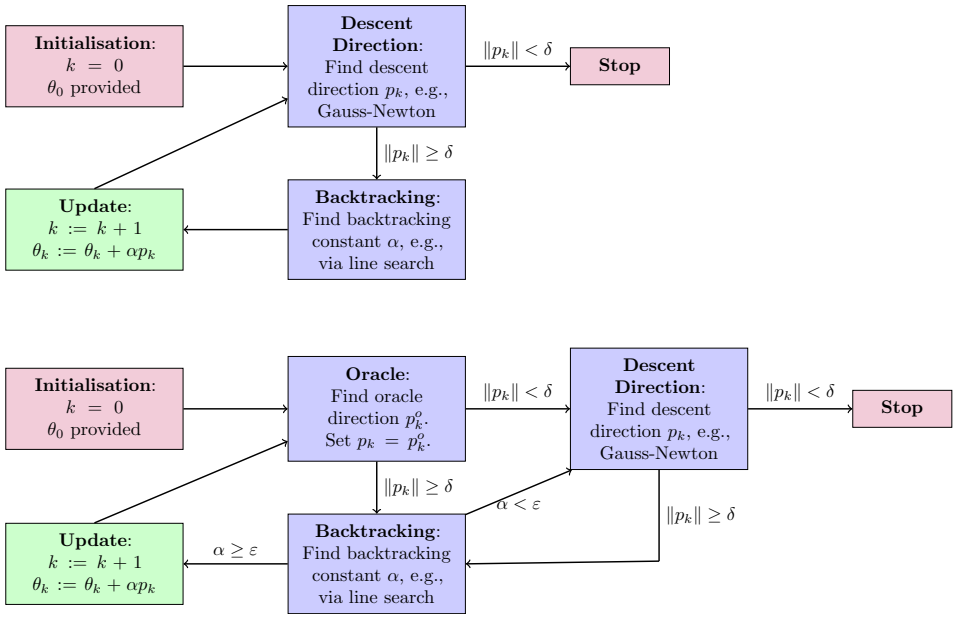


Figure 1: Flowcharts of a generic line search algorithm with and without an oracle.

between the curve and the observations. The resulting estimate of θ:

$$\theta_k^{\text{oracle}} = \arg\min_{\theta} \sum_{j=1}^{n} \left\| y_j - \psi(t_1) - \int_{t_1}^{t_j} f(\psi_{\theta_k}(s), \theta) ds \right\|_2^2. \tag{12}$$

is called the *oracle* estimate. We denote $p_k^{\text{oracle}} = \theta_k^{\text{oracle}} - \theta_k$ the *oracle* direction. In general it is not certain that $\ell_y(\theta_k^{\text{oracle}}) < \ell_y(\theta_k)$, hence a backtracking of p_k^{oracle} must be employed in order to gain a descent:

$$\theta_{k+1} = \theta_k + \alpha p_k^{\text{oracle}}$$

for some $\alpha \in [0, 1]$. However, it is not even certain that p_k^{oracle} is a descent direction! In which case, the backtracking will fail to find a positive α within numerical tolerance. In this case, no benefit from the oracle is gained. The algorithm then passes on to some classic optimisation scheme, e.g., Gauss-Newton. Once a single Gauss-Newton update is done, the algorithm returns to the oracle for the next iterate. A generic line search optimisation algorithm with and without an oracle are visualised by two flowcharts in figure 1.

This algorithm maintains the convergence properties of the Gauss-Newton algorithm, while benefiting from computational advantages possessed by the oracle. In practice the oracle mostly provides excellent descent directions and the Gauss-Newton part will only be invoked to verify that the final iterate is an approximative local minima.

5 Discussion and further work

The new combined algorithm presented above has been implemented and tested on simulated data from mass action kinetics models. The results look promising both for large and small σ^2, along with high and low frequency data. In these studies the oracle always provided a descent direction. Intuitively this is not true in general, as the algorithm will perform poorly for *stiff* or *chaotic* systems. The computationally heavy part of the algorithm is the Gauss-Newton part. Consequently, it is of high interest to find conditions that ensure the oracle alone provides the descent directions necessary to find MLE.

Additionally, the algorithm can be extended to parameter estimation with forced sparsity, e.g., using ℓ^p penalties. This is relevant for estimating unknown model structures. However, when introducing such penalties the algorithm has to be revised in order to accommodate potential lack of smoothness.

Acknowledgements: A great thanks to Professor Niels Richard Hansen and Martin Vincent of Department of Mathematical Sciences at University of Copenhagen for guidance and many insightful discussions.

References

[1] N. J-B. Brunel. Parameter estimation of ODE's via nonparametric estimators. *Electronic Journal of Statistics*, 2:1242–1267, 2008.

[2] D. Brewer, M. Barenco, R. Collard, M. Hubank and J. Stark. Fitting ordinary differential equations to short time course data. *Philos. Transact. A Math. Phys. Eng. Sci.*, 366:519–544, 2008.

[3] Z. Li, M. R. Osborne and T. Prvan. Parameter estimation of ordinary differential equations. *IMA Journal of Numerical Analysis*, 25(2):264–285, 2005.

[4] J. O. Ramsay, G. Hooker, D. Campbell and J. Cao. Parameter estimation for differential equations: a generalised smoothing approach. *J. R. Statist. Soc. B*, 69(5):741–796, 2007.

[5] W. H. Press, B. P. Flannery, S. A. Teukolsky and W. T. Vetterling. *Numerical Recipes: The Art of Scientific Computing* (3rd ed). Cambridge University Press, 2007.

[6] J. Nocedal and S. J. Wright. *Numerical Optimization* (2nd ed). Springer, 2006.

On Calculation of the Integrals Depending on a Parameter by Monte-Carlo Method

Yuriy Mlavets[*1] **and Yuriy Kozachenko**[2]

[1] *Uzhhorod National University, Ukraine*
[2] *Taras Shevchenko National University of Kyiv, Ukraine*

Abstract: In this work we use the theory $\mathbf{F}_\psi(\Omega)$ spaces in order to find the accuracy and reliability for the calculation of the improper integrals depending on a parameter t by Monte Carlo method.

Keywords: $\mathbf{F}_\psi(\Omega)$ space of random variables, condition **H**, Monte Carlo method, stochastic process

AMS subject classifications: 65C05, 60G07

1 Introduction

In this paper we developed a theory for finding the reliability and accuracy for the calculation of integrals depending on a parameter by Monte-Carlo method in $L_p(T)$ metrics.

There are many works devoted to the usage of the Monte-Carlo methods for calculation of integrals. Among them are the books by Yermakov [1] and Yermakov & Mikhailov [2].

The paper by Kurbanmuradov & Sabelfeld [7] contains the estimate for the accuracy in the space $C(T)$ and reliability for the calculation of integrals depending on a parameter if the set of integration is bounded. To obtain these results the theory of sub-Gaussian processes had been used.

The space $\mathbf{F}_\psi(\Omega)$ was introduced by Yermakov & Ostrovsky in the paper [3]. The paper [5] is devoted to studying the properties of such spaces and there had been found the conditions of fulfilling the condition **H** (see Definition 10) in this spaces. The condition **H** is necessary for finding the reliability and accuracy when we calculate integrals by Monte-Carlo method.

The choice of the space depends on particular integral and allows to find better accuracy. In this paper, the accuracy is defined via the norm in $L_p(T)$ space.

2 $\mathbf{F}_\psi(\Omega)$ – space

Definition 9. [6] Let $\psi(u) > 0$, $u \geq 1$ be monotonically increasing, continuous function for which $\psi(u) \to \infty$ as $u \to \infty$. A random variable ξ belongs to the space $\mathbf{F}_\psi(\Omega)$ if

$$\sup_{u \geq 1} \frac{\left(E\left|\xi\right|^u\right)^{1/u}}{\psi(u)} < \infty.$$

*Corresponding author: yura-mlavec@ukr.net

The similar definition was formulated in the paper by S. M. Yermakov & Ye. I. Ostrovskii [3]. But there was required that $E\xi = 0$ as $\xi \in \mathbf{F}_\psi(\Omega)$. Moreover, there were considered the random variables for which $E\left|\xi\right|^u = \infty$ for some $u > 0$.

It had been proved in [3] that $\mathbf{F}_\psi(\Omega)$ is a Banach space with a norm

$$\left\|\xi\right\|_\psi = \sup_{u \geq 1} \frac{\left(E\left|\xi\right|^u\right)^{1/u}}{\psi(u)}.$$

Theorem 20. *[6] If a random variable ξ belongs to the space $\mathbf{F}_\psi(\Omega)$, then for any $\varepsilon > 0$ the following inequality holds true:*

$$P\left\{|\xi| > \varepsilon\right\} \leq \inf_{u \geq 1} \frac{\left\|\xi\right\|_\psi^u \left(\psi(u)\right)^u}{\varepsilon^u}.$$

Theorem 21. *[6] If a random variable ξ belongs to the space $\mathbf{F}_\psi(\Omega)$ and $\psi(u) = u^\alpha$, where $\alpha > 0$, then for any $\varepsilon \geq e^\alpha \left\|\xi\right\|_\psi$ the following inequality is true:*

$$P\left\{|\xi| > \varepsilon\right\} \leq \exp\left\{-\frac{\alpha}{e}\left(\frac{\varepsilon}{\left\|\xi\right\|_\psi}\right)^{1/\alpha}\right\}.$$

Definition 10. [5] We say that the condition **H** for the Banach spaces $B(\Omega)$ of random variables is fulfilled if there exists such an absolute constant C_B that for any centered and independent random variables $\xi_1, \xi_2, \ldots, \xi_n$ from $B(\Omega)$ the following is true:

$$\left\|\sum_{i=1}^n \xi_i\right\|^2 \leq C_B \sum_{i=1}^n \left\|\xi_i\right\|^2.$$

The constant C_B is called a scale constant for the space $B(\Omega)$. For space $\mathbf{F}_\psi(\Omega)$ we shall denote the constants $C_{\mathbf{F}_\psi(\Omega)}$ as C_ψ.

Theorem 22. *[8] For the space $\mathbf{F}_\psi(\Omega)$, where $\psi(u) = u^\alpha$, $\alpha \geq \frac{1}{2}$ the condition* **H** *is fulfilled and it is true the following inequality:*

$$\left\|\sum_{i=1}^n \xi_i\right\|_\psi^2 \leq 4 \cdot 9^\alpha \sum_{i=1}^n \left\|\xi_i\right\|_\psi^2.$$

Note, that when $\alpha < \frac{1}{2}$ then the condition **H** *is not fulfilled for this space.*

3 Estimates in the norm $L_p(T)$ for the stochastic processes from the spaces $\mathbf{F}_\psi(\Omega)$

Theorem 23. *Let ν be the σ-finite measure on the compact metric space (T, ρ) and $X = \{X(t),\ t \in T\}$ be a measurable stochastic process from the space $\mathbf{F}_\psi(\Omega)$. If for some $p \geq 1$ the following condition is true*

$$\int_T \left\|X(t)\right\|_\psi^p \, d\nu(t) < \infty,$$

then

1) the integral $\int\limits_T |X(t)|^p \, d\nu(t)$ *exists with probability one and the inequality holds true:*

$$\left\| \left(\int\limits_T |X(t)|^p \, d\nu(t) \right)^{1/p} \right\|_\psi \le \frac{\psi(p)}{\psi(1)} \left(\int\limits_T \|X(t)\|_\psi^p \, d\nu(t) \right)^{1/p};$$

2) for any $\varepsilon > 0$ *the following inequality holds:*

$$P\left\{ \left(\int\limits_T |X(t)|^p \, d\nu(t) \right)^{1/p} > \varepsilon \right\} \le$$

$$\le \inf_{u \ge 1} \frac{\left(\frac{\psi(p)}{\psi(1)}\right)^u \left(\int\limits_T \|X(t)\|_\psi^p \, d\nu(t) \right)^{u/p} (\psi(u))^u}{\varepsilon^u}.$$

Example 1. *Consider the space* $\mathbf{F}_\psi(\Omega)$*, where* $\psi(u) = u^\alpha$*,* $\alpha > 0$*. It follows from the Theorems 23 and 21 that for* $\varepsilon \ge (ep)^\alpha \left(\int\limits_T \|X(t)\|_\psi^p \, d\nu(t) \right)^{1/p}$

$$P\left\{ \left(\int\limits_T |X(t)|^p \, d\nu(t) \right)^{1/p} > \varepsilon \right\} \le$$

$$\le \exp\left\{ -\frac{\alpha}{ep} \left(\frac{\varepsilon}{\left(\int\limits_T \|X(t)\|_\psi^p \, d\nu(t) \right)^{1/p}} \right)^{1/\alpha} \right\}.$$

Theorem 24. *Let* ν *be a* σ*-finite measure on a compact metric* (T, ρ) *and* $Y = \{Y(t),\ t \in T\}$ *be the stochastic process from the space* $\mathbf{F}_\psi(\Omega)$ *and the condition* $\mathbf{H}$ *is fulfilled for this space with the constant* C_ψ*. Let* $EY(t) = m(t)$*,* $Z_n(t) = \frac{1}{n} \sum\limits_{k=1}^n Y_k(t) - m(t) = \frac{1}{n} \sum\limits_{k=1}^n (Y_k(t) - m(t))$*, where* $Y_k(t)$ *are the independent copies of* $Y(t)$*. Then the following inequality holds for all* $p \ge 1$

$$\left\| \left(\int\limits_T |Z_n(t)|^p \, d\nu(t) \right)^{1/p} \right\|_\psi \le \frac{2\sqrt{C_\psi}}{\sqrt{n}} \cdot \frac{\psi(p)}{\psi(1)} \left(\int\limits_T \|Y(t)\|_\psi^p \, d\nu(t) \right)^{1/p}$$

and for every $\varepsilon > 0$ the following estimate is true

$$P\left\{\left(\int_T |Z_n(t)|^p \, d\nu(t)\right)^{1/p} > \varepsilon\right\} \leq$$

$$\leq \inf_{u \geq 1} \frac{\left(\frac{2\sqrt{C_\psi}}{\sqrt{n}} \cdot \frac{\psi(p)}{\psi(1)}\right)^u \left(\int_T \|Y(t)\|_\psi^p \, d\nu(t)\right)^{u/p} (\psi(u))^u}{\varepsilon^u}.$$

Example 2. *Let us consider the space $\mathbf{F}_\psi(\Omega)$, where $\psi(u) = u^\alpha$, $\alpha > 0$ then it follows from the Theorems 24 and 21 that if $\varepsilon \geq (ep)^\alpha \frac{2\sqrt{C_\psi}}{\sqrt{n}} \left(\int_T \|Y(t)\|_\psi^p \, d\nu(t)\right)^{1/p}$*

$$P\left\{\left(\int_T |Z_n(t)|^p \, d\nu(t)\right)^{1/p} > \varepsilon\right\} \leq$$

$$\leq \exp\left\{-\frac{\alpha}{ep}\left(\frac{\varepsilon}{\frac{2\sqrt{C_\psi}}{\sqrt{n}}\left(\int_T \|Y(t)\|_\psi^p \, d\nu(t)\right)^{1/p}}\right)^{1/\alpha}\right\}.$$

4 Reliability and accuracy in the space $L_p(T)$ for the calculation of integrals depending on a parameter

Let $\{\mathcal{S}, \mathcal{A}, \mu\}$ be a measurable space, μ be a σ-finite measure and $p(s) \geq 0, s \in \mathcal{S}$ be such measurable function that $\int_{\mathcal{S}} p(s) d\mu(s) = 1$. Let $m(A)$, $A \in \mathcal{A}$ be the measure $m(A) = \int_A p(s) d\mu(s)$. $m(A)$ is a probability measure and the space $\{\mathcal{S}, \mathcal{A}, m\}$ is a probability space.

Let us consider the integral $\int_{\mathcal{S}} f(s,t)p(s)d\mu(s) = I(t)$ assuming that it exists. Let the function $f(s,t)$ depend on the parameter $t \in T$, where (T, ρ) is some compact set and the function $f(s,t)$ be continuous with regard to t.

Suppose $f(s,t)$ is the stochastic process on $\{\mathcal{S}, \mathcal{A}, m\}$ and we denote it as $\xi(s,t) = \xi(t)$ and $I(t) = \int_{\mathcal{S}} f(s,t)p(s)d\mu(s) = \int_{\mathcal{S}} f(s,t)dm(s) = E\xi(t)$.

Let $\xi_i(t)$, $i = 1, 2, \ldots, n$ be the independent copies of the stochastic process $\xi(t)$, $Z_n(t) = \frac{1}{n}\sum_{i=1}^n \xi_i(t)$. So, according to the strong law of large numbers $Z_n(t) \to E\xi(t) = I(t)$ with probability one for any $t \in T$.

Definition 11. We say that $Z_n(t)$ approximates $I(t)$ in the space $L_p(T)$ with reliability $1-\delta>0$ and accuracy $\varepsilon>0$ if the following inequality holds true:

$$P\left\{\left(\int\limits_T |Z_n(t)-I(t)|^p\,d\mu(t)\right)^{1/p} > \varepsilon\right\} \le \delta.$$

Theorem 25. *Let $I(t)=E\xi(t)=\int\limits_S f(s,t)p(s)d\mu(s)$, $\xi(t)$ be the stochastic process which belongs to the space $\mathbf{F}_\psi(\Omega)$ satisfying the condition* **H** *with constant C_ψ, $\widetilde{Z}_n(t)=\frac{1}{n}\sum\limits_{i=1}^{n}(\xi_i(t)-I(t))$, $\xi_i(t)$ be the independent copies of the stochastic process $\xi(t)$.*

Then, for all $p\ge 1$ the following inequality holds true

$$\left\|\left(\int\limits_T \left|\widetilde{Z}_n(t)\right|^p d\mu(t)\right)^{1/p}\right\| \le \frac{2\sqrt{C_\psi}}{\sqrt{n}}\cdot\frac{\psi(p)}{\psi(1)}\left(\int\limits_T \|\xi(t)\|_\psi^p\,d\mu(t)\right)^{1/p},$$

and $\widetilde{Z}_n(t)$ approximates $I(t)$ with reliability $1-\delta$ and accuracy ε in the space $L_p(T)$ for such n that

$$\inf_{u\ge 1}\frac{\left(\frac{2\sqrt{C_\psi}}{\sqrt{n}}\cdot\frac{\psi(p)}{\psi(1)}\right)^u\left(\int\limits_T\|\xi(t)\|^p\,d\mu(t)\right)^{u/p}(\psi(u))^u}{\varepsilon^u}\le\delta. \tag{1}$$

Example 3. *Consider the space $\mathbf{F}_\psi(\Omega)$, where $\psi(u)=u^\alpha$, $\alpha>\frac{1}{2}$. Then the Theorem 22 implies that the condition* **H** *is fulfilled for this space with the constant $C_\psi=4\cdot 9^\alpha$. It follows from the Example 2 and the Theorem 25 that if $\varepsilon\ge\frac{4(3pe)^\alpha\left(\int\limits_T\|\xi(t)\|_\psi^p d\mu(t)\right)^{1/p}}{\sqrt{n}}$, then*

$$\inf_{u\ge 1}\frac{\left(\frac{2\sqrt{C_\psi}}{\sqrt{n}}\cdot\frac{\psi(p)}{\psi(1)}\right)^u\left(\int\limits_T\|\xi(t)\|^p\,d\mu(t)\right)^{u/p}(\psi(u))^u}{\varepsilon^u}\le$$

$$\le\exp\left\{-\frac{\alpha}{e}\left(\frac{\sqrt{n}\varepsilon}{4(3pe)^\alpha\left(\int\limits_T\|\xi(t)\|_\psi^p\,d\mu(t)\right)^{1/p}}\right)^{1/\alpha}\right\}.$$

So, the inequality (1) *holds if it is true that*

$$\exp\left\{-\frac{\alpha}{e}\left(\frac{\sqrt{n}\varepsilon}{4(3pe)^\alpha\left(\int\limits_T\|\xi(t)\|_\psi^p\,d\mu(t)\right)^{1/p}}\right)^{1/\alpha}\right\}\le\delta,$$

as

$$n \geq \left(\frac{4(3pe)^{\alpha} \left(\int\limits_{T} \|\xi(t)\|_{\psi}^{p} \, d\mu(t) \right)^{1/p}}{\varepsilon} \right)^{2} \left((-\ln \delta) \frac{e}{\alpha} \right)^{2\alpha}.$$

Then

$$n \geq \left(\frac{4(3p)^{\alpha} \left(\int\limits_{T} \|\xi(t)\|_{\psi}^{p} \, d\mu(t) \right)^{1/p}}{\varepsilon} \right)^{2} \max \left(1, \left(-\frac{\ln \delta}{\alpha} \right)^{2\alpha} \right).$$

References

[1] S. M. Ermakov. The Monte Carlo Method and Contiguous Questions. *Moscow, Nauka*, 163, 1975.

[2] S. M. Ermakov and G. A. Mikhailov. The Course of Statistical Modelling. *Moscow, Nauka*, 319, 1976.

[3] S. V. Ermakov and E. I. Ostrovskii. Continuity conditions, exponential estimates, and the central limit theorem for random fields. *Dep. VINITI, Moscow*, 3752-B.86.0:42, 1986.

[4] Yu. V. Kozachenko and Yu. Yu. Mlavets. Probability of large deviations of sums of random processes from Orlicz space. *Monte Carlo Methods Applications*, 17:155–168, 2011.

[5] Yu. V. Kozachenko and Yu. Yu. Mlavets. The Banach spaces $\mathbf{F}_{\psi}(\Omega)$ of random variables . *Theory of Probability and Mathematical Statistics*, 86:105–121, 2013.

[6] Yuriy Kozachenko and Yuriy Mlavets. Stochastic processes from $\mathbf{F}_{\psi}(\Omega)$ spaces. *Contemporary Mathematics and Statistics*, 2(1):55–75, 2014.

[7] O. Kurbanmuradov and K. Sabelfeld. Exponential bounds for the probability deviations of sums of random fields. *Monte Carlo Methods and Applications*, 12(3-4):211–229, 2006.

[8] Yu. Yu. Mlavets. A relationship between the spaces $\mathbf{F}_{\psi}(\Omega)$ and Orlicz spaces random variables. *Scientific Bulletin of the Uzhhorod University. Series Mathematics and Informatics*, 25(1):77–84, 2014.

Behavior of Rank Tests and R-Estimates in Measurement Error Models

Radim Navrátil*

Masaryk University, Brno, Czech Rep.

Abstract: Behavior of rank tests and R-estimates in presence of measurement errors is studied. It is showed that rank tests of some hypotheses stay valid in measurement error models. The presence of measurement errors only decreases their power. Unlike that R-estimates in measurement error models are biased. Unfortunately this bias cannot be corrected without any knowledge of the distribution of measurement errors.

Keywords: aligned rank tests, measurement error models, rank tests, R-estimates

AMS subject classifications: 62G10, 62G30

1 Introduction

Measurement error models (also called *errors-in-variables models*) are regression models that account for measurement errors in the independent variables (regressors). These models occur very commonly in practical data analysis, where some variables cannot be observed exactly, usually due to instrument or sampling error. Sometimes ignoring measurement error may lead to correct conclusions, however in some situations it may have dramatic consequences.

The most of the literature about measurement error models uses parametric approach with its restrictive normality assumptions or a knowledge of some additional information about error distribution (see e.g. [1]). We avoided this and introduced a class of rank tests that is valid even if measurement errors are present. The main goal of this paper is to investigate the behavior of standard rank procedures in measurement error models - both tests and estimates. Proofs of the theorems are omitted, but they may be found in [3] and[4].

2 Behavior of rank tests in measurement error models

Consider classical linear regression model

$$Y_i = \beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta} + e_i, \quad i = 1, \ldots, n, \tag{1}$$

*Corresponding author: navratil@math.muni.cz

where $\beta_0 \in \mathbb{R}$ and $\boldsymbol{\beta} \in \mathbb{R}^p$ are unknown parameters, model errors e_i are assumed to be independent identically distributed (i.i.d.) with an unknown distribution function F and density f, $\mathbf{x}_i$ are vectors of known regressors, such that

$$\mathbf{Q}_n = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^\top, \quad \text{with} \quad \overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$$

is a positive definite matrix (further on we will tacitly assume that this holds). Our aim is to test the hypothesis $\mathbf{H}_0 : \boldsymbol{\beta} = \mathbf{0}$ against $\mathbf{K}_0 : \boldsymbol{\beta} \neq \mathbf{0}$.

Choose a nondecreasing, nonconstant, square integrable score generating function
$\varphi : (0,1) \mapsto \mathbb{R}$ and define

$$\begin{aligned} a_n(i) &= \varphi\left(\frac{i}{n+1}\right), \quad i = 1, \ldots, n, \\ A^2(\varphi) &= \int_0^1 (\varphi(t) - \overline{\varphi})^2 dt, \quad \overline{\varphi} = \int_0^1 \varphi(t)dt. \end{aligned}$$

Let R_i be the rank of Y_i among $Y_1, \ldots, Y_n$ and define vector of linear rank statistics

$$\mathbf{S}_n = n^{-1/2}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})a_n(R_i).$$

Test criterion for $\mathbf{H}_0$ is then

$$T_n^2 = A^{-2}(\varphi)\mathbf{S}_n^\top \mathbf{Q}_n^{-1}\mathbf{S}_n. \tag{2}$$

Assume that f has finite Fisher information with respect to the location

$$0 < I(f) = \int \left(\frac{f'(x)}{f(x)}\right)^2 f(x)dx < \infty \tag{3}$$

and there exists a positive definite matrix $\mathbf{Q}$, such that as $n \to \infty$

$$\mathbf{Q}_n \to \mathbf{Q}, \tag{4}$$

$$\frac{1}{n}\max_{i=1,\ldots,n}(\mathbf{x}_i - \overline{\mathbf{x}})^\top \mathbf{Q}_n^{-1}(\mathbf{x}_i - \overline{\mathbf{x}}) \to 0. \tag{5}$$

Remark 3. Generally, Fisher information is defined for parametric family $d(x,\theta)$ of densities as

$$I(d,\theta) = \int \left(\frac{\frac{\partial}{\partial\theta}d(x,\theta)}{d(x,\theta)}\right)^2 d(x,\theta)dx.$$

If θ is a location parameter, i.e. $d(x,\theta) = f(x-\theta)$, then $I(d,\theta) = I(f)$.

Theorem 26. *Assume that (3) – (5) hold. Then in model (1) under* $\mathbf{H}_0$ *test statistic* T_n^2 *has asymptotically as* $n \to \infty$ χ^2 *distribution with* p *degrees of freedom and under sequence of local alternatives*

$$\mathbf{K}_{0,n} : \boldsymbol{\beta} = n^{-1/2}\boldsymbol{\beta}^*, \quad \mathbf{0} \neq \boldsymbol{\beta}^* \in \mathbb{R}^p \text{ fixed}$$

T_n^2 has asymptotically as $n \to \infty$ noncentral χ^2 distribution with p degrees of freedom and noncentrality parameter

$$\eta^2 = \boldsymbol{\beta}^{*\top}\mathbf{Q}\boldsymbol{\beta}^* \frac{\gamma^2(\varphi, f)}{A^2(\varphi)}, \quad \gamma(\varphi, f) = \int_0^1 \varphi(t)\widetilde{\varphi}(t, f)dt, \quad \widetilde{\varphi}(t, f) = -\frac{f'(F^{-1}(t))}{f(F^{-1}(t))}.$$

2.1 Model with errors in regressors

Measurement error model assumes that regressors $\mathbf{x}_i$ are not observed accurately, but only with an additive, unobservable, error $\mathbf{v}_i$, i.e. we observe $\mathbf{w}_i = \mathbf{x}_i + \mathbf{v}_i$ instead of $\mathbf{x}_i$, where $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are i.i.d. random vectors independent of $e_1, \ldots, e_n$. In other words, we may write

$$\begin{aligned} Y_i &= \beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta} + e_i, \\ \mathbf{w}_i &= \mathbf{x}_i + \mathbf{v}_i, \quad i = 1, \ldots, n. \end{aligned} \tag{6}$$

Denote $T_{w,n}^2$ test statistic (2) based on observed values $(Y_i, \mathbf{w}_i)$. It is easy to show that $T_{w,n}^2$ has under $\mathbf{H}_0$ the same distribution as T_n^2. The presence of measurement errors only decreases power of the test.

2.2 Aligned rank tests

However, we are often more interested in testing hypothesis only about a component of the parameter $\boldsymbol{\beta}$, identify regressors that have influence on response variable. Denote

$$\begin{aligned} \boldsymbol{\beta} &= (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top, \qquad \mathbf{x}_i = (\mathbf{x}_{1,i}^\top, \mathbf{x}_{2,i}^\top)^\top, \\ \mathbf{v}_i &= (\mathbf{v}_{1,i}^\top, \mathbf{v}_{2,i}^\top)^\top, \qquad \mathbf{w}_i = (\mathbf{w}_{1,i}^\top, \mathbf{w}_{2,i}^\top)^\top, \end{aligned}$$

where $\boldsymbol{\beta}_1 \in \mathbb{R}^{p-q}$, $\boldsymbol{\beta}_2 \in \mathbb{R}^q$, $\mathbf{x}_{1,i} \in \mathbb{R}^{p-q}$, $\mathbf{x}_{2,i} \in \mathbb{R}^q$, $\mathbf{v}_{1,i} \in \mathbb{R}^{p-q}$, $\mathbf{v}_{2,i} \in \mathbb{R}^q$, $\mathbf{w}_{1,i} \in \mathbb{R}^{p-q}$, $\mathbf{w}_{2,i} \in \mathbb{R}^q$, $1 \le q < p$. Then model (6) can be rewritten as

$$\begin{aligned} Y_i &= \beta_0 + \mathbf{x}_{1,i}^\top \boldsymbol{\beta}_1 + \mathbf{x}_{2,i}^\top \boldsymbol{\beta}_2 + e_i, \\ \mathbf{w}_{1,i} &= \mathbf{x}_{1,i} + \mathbf{v}_{1,i}, \\ \mathbf{w}_{2,i} &= \mathbf{x}_{2,i} + \mathbf{v}_{2,i}, \quad i = 1, \ldots, n. \end{aligned} \tag{7}$$

Our goal is to test the hypothesis $\mathbf{H}_1 : \boldsymbol{\beta}_2 = \mathbf{0}$ against $\mathbf{K}_1 : \boldsymbol{\beta}_2 \neq \mathbf{0}$, considering β_0 and $\boldsymbol{\beta}_1$ as nuisance parameters.

Rank tests are invariant with respect to the location, but not to the nuisance regression. That is why we have to first estimate the nuisance parameter $\boldsymbol{\beta}_1$ and then apply the standard test on residuals. Due to the absence of knowledge of distribution of model errors and to preserve robust properties we use an R-estimator of parameter $\boldsymbol{\beta}_1$.

Model (7) under $\mathbf{H}_1$ reduces to $Y_i = \beta_0 + \mathbf{w}_{1,i}^\top \boldsymbol{\beta}_1 + e_i^*$, where $e_i^* = e_i - \mathbf{v}_{1,i}^\top \boldsymbol{\beta}_1$ are i.i.d. random variables with density $f^* = f_{\beta_1}^*$.

Choose a nondecreasing, nonconstant, square integrable score generating function $\psi : (0,1) \mapsto \mathbb{R}$ that is skew-symmetric, i.e. $\psi(1-t) = -\psi(t), \quad \forall\, 0 < t < 1$

and define $\widetilde{a}_n(i) = \psi\left(\frac{i}{n+1}\right), \quad i = 1, \ldots, n$. Following [2] we define the rank (pseudo)estimator $\widehat{\boldsymbol{\beta}}_{1,n}$ of $\boldsymbol{\beta}_1$ as a minimizer of

$$\mathcal{D}_n(\mathbf{b}) = \sum_{i=1}^{n} \left(Y_i - \mathbf{w}_{1,i}^\top \mathbf{b}\right) \widetilde{a}_n(R_i(\mathbf{b}))$$

with respect to $\mathbf{b} \in \mathbb{R}^{p-q}$, where $R_i(\mathbf{b})$ is the rank of $(Y_i - \mathbf{w}_{1,i}^\top \mathbf{b})$ among $(Y_1 - \mathbf{w}_{1,1}^\top \mathbf{b}), \ldots, (Y_n - \mathbf{w}_{1,n}^\top \mathbf{b})$.

Now, consider residuals

$$\widehat{e}_i = Y_i - \mathbf{w}_{1,i}^\top \widehat{\boldsymbol{\beta}}_{1,n}, \quad i = 1, \ldots, n$$

and apply the test described in Section 2 on residuals $\widehat{e}_1, \ldots, \widehat{e}_n$. Note that unlike the situation in Section 2 residuals $\widehat{e}_i$ are not independent, because they depend on the R-estimate of nuisance parameter $\boldsymbol{\beta}_1$. However under some assumptions this fact does not affect the asymptotic distribution.

Hence choose a nondecreasing, nonconstant, square integrable score generating function $\varphi : (0,1) \mapsto \mathbb{R}$ (it may differ from ψ) and define $a_n(i) = \varphi\left(\frac{i}{n+1}\right), \quad i = 1, \ldots, n$ and compute

$$\widehat{\mathbf{S}}_n = n^{-1/2} \sum_{i=1}^{n} (\mathbf{w}_{2,i} - \overline{\mathbf{w}}_2) a_n(R_i(\widehat{\boldsymbol{\beta}}_{1,n})),$$

where $R_i(\widehat{\boldsymbol{\beta}}_{1,n})$ is the rank of $\widehat{e}_i$ among $\widehat{e}_1, \ldots, \widehat{e}_n$. Denote

$$\mathbf{D}_{1,n} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{w}_{1,i} - \overline{\mathbf{w}}_1)(\mathbf{w}_{1,i} - \overline{\mathbf{w}}_1)^\top, \quad \mathbf{D}_{2,n} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{w}_{2,i} - \overline{\mathbf{w}}_2)(\mathbf{w}_{2,i} - \overline{\mathbf{w}}_2)^\top.$$

Finally, consider test statistic

$$\widehat{T}_n^2 = A^{-2}(\varphi) \widehat{\mathbf{S}}_n^\top \mathbf{D}_{2,n}^{-1} \widehat{\mathbf{S}}_n.$$

Assume that there exist positive definite matrices $\mathbf{Q}_1, \mathbf{Q}_2, \mathbf{C}_1, \mathbf{C}_2$, such that

$$\mathbf{Q}_{1,n} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_{1,i} - \overline{\mathbf{x}}_1)(\mathbf{x}_{1,i} - \overline{\mathbf{x}}_1)^\top \to \mathbf{Q}_1, \tag{8}$$

$$\mathbf{C}_{1,n} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{v}_{1,i} - \overline{\mathbf{v}}_1)(\mathbf{v}_{1,i} - \overline{\mathbf{v}}_1)^\top \xrightarrow{p} \mathbf{C}_1, \tag{9}$$

$$\frac{1}{n} \max_{i=1,\ldots,n} (\mathbf{w}_{1,i} - \overline{\mathbf{w}}_1)^\top \mathbf{D}_{1,n}^{-1} (\mathbf{w}_{1,i} - \overline{\mathbf{w}}_1) \to 0, \tag{10}$$

$$\mathbf{Q}_{2,n} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_{2,i} - \overline{\mathbf{x}}_2)(\mathbf{x}_{2,i} - \overline{\mathbf{x}}_2)^\top \to \mathbf{Q}_2, \tag{11}$$

$$\mathbf{C}_{2,n} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{v}_{2,i} - \overline{\mathbf{v}}_2)(\mathbf{v}_{2,i} - \overline{\mathbf{v}}_2)^\top \xrightarrow{p} \mathbf{C}_2, \tag{12}$$

$$\frac{1}{n} \max_{i=1,\ldots,n} (\mathbf{w}_{2,i} - \overline{\mathbf{w}}_2)^\top \mathbf{D}_{2,n}^{-1} (\mathbf{w}_{2,i} - \overline{\mathbf{w}}_2) \to 0. \tag{13}$$

Finally, we are able to describe asymptotic distribution of $\widehat{T}_n^2$.

Theorem 27. *Assume that (3), (8) – (13) hold. Then in model (7) under* $\mathbf{H}_1$ *test statistic* $\widehat{T}_n^2$ *has asymptotically as* $n \to \infty$ χ^2 *distribution with* q *degrees of freedom and under local alternative*

$$\mathbf{K}_{1,n}: \ \boldsymbol{\beta}_2 = n^{-1/2}\boldsymbol{\beta}_2^*, \quad \mathbf{0} \neq \boldsymbol{\beta}_2^* \in \mathbb{R}^q \text{ fixed}$$

$\widehat{T}_n^2$ *has asymptotically noncentral* χ^2 *distribution with* q *degrees of freedom and noncentrality parameter*

$$\widehat{\eta}^2 = \boldsymbol{\beta}_2^{*\top}\mathbf{Q}_2(\mathbf{Q}_2 + \mathbf{C}_2)^{-1}\mathbf{Q}_2\boldsymbol{\beta}_2^* \frac{\gamma^2(\varphi, f^*)}{A^2(\varphi)}.$$

Recall that f^* depends on unknown nuisance parameter $\boldsymbol{\beta}_1$ and distribution of measurement errors $v_{1,i}$, hence the asymptotic power of the test does depend on the nuisance parameter $\boldsymbol{\beta}_1$ unlike the situation without measurement errors.

3 Behavior of R-estimates in measurement error models

Now, we are interested in R-estimator of the slope vector $\boldsymbol{\beta}$ in model (6), considering β_0 as nuisance parameter. In the previous section we already needed to estimate the nuisance slope parameter. Hence, there was developed the idea to use R-estimates in measurement error models. Unfortunately, like other classical estimates they are also (asymptotically) biased.

Remind briefly the approach already used in the previous section. Let $R_i(\mathbf{b})$ be the rank of the residual $Y_i - \mathbf{w}_i^\top \mathbf{b}, \quad i = 1, \ldots, n$ and denote the vector of linear rank statistics

$$\mathbf{S}_n(\mathbf{b}) = n^{-1/2} \sum_{i=1}^{n} (\mathbf{w}_i - \overline{\mathbf{w}}) a_n(R_i(\mathbf{b})),$$

where the scores $a_n(i) = \psi\left(\frac{i}{n+1}\right)$ are generated by square integrable score function ψ that is skew-symmetric on $(0,1)$. [2] defined the rank estimator $\widehat{\boldsymbol{\beta}}_n$ of $\boldsymbol{\beta}$ as a minimizer of

$$\mathcal{D}_n(\mathbf{b}) = \sum_{i=1}^{n} \left(Y_i - \mathbf{w}_i^\top \mathbf{b}\right) a_n(R_i(\mathbf{b})) \quad \text{with respect to} \quad \mathbf{b} \in \mathbb{R}^p.$$

We are able to study asymptotic properties of $\widehat{\boldsymbol{\beta}}_n$ in the presence of measurement errors and find its local asymptotic bias only in a neighborhood of true value of the parameter $\boldsymbol{\beta}$, i.e. under local alternative $\boldsymbol{\beta}_n = n^{-1/2}\boldsymbol{\beta}^*$ with a fixed $\boldsymbol{\beta}^* \in \mathbb{R}^p$.

In the sequel, all limits are taken as $n \to \infty$, unless mentioned otherwise. We shall now describe the needed assumptions on the underlying entities.

F.1 F has an absolutely continuous density f and derivative f' a.e. and has positive and finite Fisher information $I(f)$.

F.2 For every $u \in \mathbb{R}$, $\int \left(|f'(x-tu)|^j / f^{j-1}(x)\right)dx \to \int \left(|f'(x)|^j / f^{j-1}(x)\right)dx < \infty$, as $t \to 0$, $j = 2, 3$.

V.1 The measurement errors $\mathbf{v}_i$ are independent of e_i and have p-dimensional distribution function $\mathbf{G}$ with a continuous density $\mathbf{g}$.

V.2 $\mathbb{E}\mathbf{C}_n \to \mathbf{C}$, where $\mathbf{C}_n = n^{-1}\sum_{i=1}^n (\mathbf{v}_i - \overline{\mathbf{v}})(\mathbf{v}_i - \overline{\mathbf{v}})^\top$ and $\mathbf{C}$ is a positive definite matrix. Moreover, $\sup_{n\geq 1} \mathbb{E}\big(\|\mathbf{v}_n\|^3 + \|\mathbf{x}_n\|^3\big) < \infty$.

V.3 $\mathbb{E}\left[n^{-1}\sum_{i=1}^n (\mathbf{v}_i - \overline{\mathbf{v}})(\mathbf{x}_i - \overline{\mathbf{x}})^\top\right] \to \mathbf{0}$.

X.1 If the regressors $\mathbf{x}_i$ are nonrandom, then assume that $\mathbf{Q}_n \to \mathbf{Q}$, where

$$\mathbf{Q}_n = \frac{1}{n}\sum_{i=1}^n (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^\top,$$

and $\mathbf{Q}$ is a positive definite matrix. Moreover,

$$\frac{1}{n}\max_{1\leq i\leq n} (\mathbf{x}_i - \overline{\mathbf{x}})^\top \mathbf{Q}_n^{-1} (\mathbf{x}_i - \overline{\mathbf{x}}) \to 0.$$

X.2 If the regressors $\mathbf{x}_i$ are random, then assume that they are independent of e_i, $\mathbf{v}_i$, $i = 1, \ldots, n$, and

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}}_n)^\top\right] \to \mathbf{Q},$$

where $\mathbf{Q}$ is a positive definite matrix.

Theorem 28. *Under the conditions* $\mathbf{F.1} - \mathbf{F.2}$, $\mathbf{V.1} - \mathbf{V.3}, \mathbf{X.1} - \mathbf{X.2}$ *and under the local alternative*

$$\boldsymbol{\beta}_n = n^{-1/2}\boldsymbol{\beta}^*,\ \boldsymbol{\beta}^* \in \mathbb{R}^p \ \textit{fixed}$$

the R-estimator $\widehat{\boldsymbol{\beta}}_n$ *in model (6) is asymptotically normally distributed with the bias* $\mathbf{B} = -(\mathbf{Q} + \mathbf{C})^{-1}\mathbf{C}\ \boldsymbol{\beta}^*$, *i.e.*

$$n^{1/2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n) \xrightarrow{d} \mathcal{N}_p\left(\mathbf{B}, (\mathbf{Q} + \mathbf{C})^{-1}\frac{A^2(\psi)}{\gamma^2(\psi, f)}\right).$$

Remark 4. In measurement error model R-estimator $\widehat{\boldsymbol{\beta}}_n$ is asymptotically biased. This bias depends on asymptotic variance matrix of unobserved regressors $\mathbf{x}_i$ and estimated parameter $\boldsymbol{\beta}$. Hence it is impossible to correct it without any additional knowledge about distribution of measurement errors $\mathbf{v}_i$.

Acknowledgements: The research was supported by Student Project Grant at MU (specific research, rector's programme) MUNI/A/1441/2014.

References

[1] W. A. Fuller. *Measurement error models.* John Wiley and Sons, New York, 1987.

[2] L. A. Jaeckel. Estimating regression coefficients by minimizing the dispersion of the residuals. *Ann. Math. Statist.*, 43:1449-1459, 1972.

[3] J. Jurečková, H. L. Koul, R. Navrátil and J. Picek. Behavior of R-estimators under measurement errors. *To apper in Bernoulli.*

[4] R. Navrátil. Tests of statistical hypotheses in measurement error models. *PhD. thesis*, Charles Univerity in Prague, 2014.

Recognition of the Objects in Digital Images Using Weighted Fuzzy C-Means Clustering Algorithm for Directional Data (W-FCM4DD)

Eda Özkul[*1] **and Orhan Kesemen**[1]

[1]*Department of Statistics and Computer Sciences, Faculty of Science, Karadeniz Technical University, 61080 Trabzon, Turkey*

Abstract: Object recognition is one of the most important issues of image analysis. Dominant points of the objects in digital images provide important clues about the recognition of the objects. Many dominant point detection algorithms have been developed. They can be classified into two categories: corner detection approaches and polygonal approximation approaches [2, 7, 8, 9, 10].

Dominant point detection is used for data reduction in many pattern recognition applications. Although the small number of dominant points of an object provides reduction of memory volume and computing time, it cannot represent the objects sufficiently [1, 3, 4]. Specifically, if there are noises in the digital image, they can be detected as an element of the object. They can be removed by image processing techniques; however, this can cause to lose some of the features of the object. In this case, some problems can occur in the determination of the dominant points of the object.

In this study, a clustering algorithm is used to solve the problems caused by noises. Each edge pixel of any object in the digital images can be a corner (dominant) point, and its probability can be calculated various methods. The probability of corner point of each edge pixel gives the possibility of the point. This study aims to cluster the index of each edge pixel. For this a weighted clustering algorithm should be used. On the other hand, indices of edge pixels are circular structure, and this makes impossible to use linear clustering. Therefore, these pixels are converted into circular data, and their indices are clustered with weighted fuzzy C-means clustering algorithmn for directional data (W-FCM4DD) [5, 6]. The possibilities of corner point of each pixel are calculated and taken as a weighted parameter in the algorithm. Thus, these possibilities can contribute to more effectively determine dominant point of the object. Therefore, more effective results are obtained in the determination of the dominant points that can represent the object. Furthermore, the optimal number of corners of the object can be determined.

Keywords: dominant point detection, W-FCM4DD algorithm, image analysis, directional data

AMS subject classifications: 68T10, 62H30, 62H11

*Corresponding author: edaozkul@ktu.edu.tr

References

[1] A. Garrido, and M. Garcia-Sivente. Boundary simplification using a multiscale dominant-point detection algorithm. *Pattern Recognition*, 31(6):791–804, 1998.

[2] C-H. Teh, and R.T. Chin. On the detection of dominant points on digital curves. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 11(8):859–872, 1989.

[3] M. Marji, and P. Siy. A new algorithm for dominant points detection and polygonization of digital curves. *Pattern Recognition*, 36(10):2239–2251, 2003.

[4] M. S. Tahaei, S. N. Hashemi, A. Mohades, and A. Gheibi. Geometric algorithm for dominant point extraction from shape contour. *Pattern Analysis and Applications*, 17(3):481–496, 2014.

[5] Ö. Tezel. *Fuzzy C-means clustering algorithm for directional data.* Master Thesis, Department of Statistics and Computer Sciences, Karadeniz Technical University, Trabzon, Turkey, 2014.

[6] S. Vefaeinahr. *Hue-Based Segmentation of Color Images Using Weighted Directional Clustering Algorithms.* Master Thesis, Department of Statistics and Computer Sciences, Karadeniz Technical University, Trabzon, Turkey, 2015

[7] T. P. Nguyen, and I. Debled-Rennesson. A discrete geometry approach for dominant point detection. *Pattern Recognition*, 44(1):32–44, 2011.

[8] W-Y. Wu. A dynamic method for dominant point detection. *Graphical Models*, 64(5):304–315, 2002.

[9] W-Y. Wu. An adaptive method for detecting dominant points. *Pattern Recognition*, 36(10):2231–2237, 2003.

[10] W-Y. Wu. Dominant point detection using adaptive bending value. *Image and Vision Computing*, 21(6):517–525, 2003.

Polynomial Approach to Distributions via Sampling

Ioanna Papatsouma*

Department of Mathematics, Aristotle University of Thessaloniki, Greece

Abstract: A new use of Coefficient of Variation (CV) is presented in order to define distribution models [5, 11] via sampling [8], connected with a great deal of socio-economic, political, medical or biological issues. The cases of a random variable (rv) X following increasing [9], descending or symmetric [7] probability density function (pdf) are studied and the suitably obtained models are presented. It is of great interest what happens when the rv X does not take values in the whole set of real numbers or in an infinite subset of it, but the range becomes finite and the distribution becomes truncated [1, 2, 3, 4]. In order to verify and validate the polynomial distribution model, we check the correspondence between sample data and models outputs [6, 10].

Keywords: coefficient of variation, polynomial, random variable, sampling, truncated

AMS subject classifications: 62D05, 62E17

References

[1] A. Muhammad et al. SkSP-V Sampling Plan for The Exponentiated Weibull Distribution. *Journal of Testing and Evaluation*, 42(3):687–694, 2013.

[2] D. R. Clark. A Note on the Upper-Truncated Pareto Distribution. *Casualty Actuarial Society E-Forum, Winter 2013*, 1–22, 2013.

[3] G. Nanjundan. Estimation of Parameter in a New Truncated Distribution. *Open Journal of Statistics*, 3(4):221–224, 2013.

[4] J. W. Jawitz. Moments of truncated continuous univariate distributions. *Advances in Water Resources*, 27(3):269–281, 2004.

[5] K. Krishnamoorthy. *Handbook of Statistical Distribution with Applications.* CRC Press, 2006.

[6] L. Sachs. *Applied Statistics: A Handbook of Techniques, 2nd edition.* Verlag Inc., New York, 1984.

[7] N. Farmakis. Estimation of Coefficient of Variation: Scaling of Symmetric Continuous Distributions. *Statistics in Transition*, 6(1):83–96, 2003.

*Corresponding author: ioannapapatsouma@gmail.com

[8] N. Farmakis. A Unique Expression for the Size of Samples in Several Sampling Procedures. *Statistics in Transition*, 7(5):1031–1043, 2006.

[9] N. Farmakis. Coefficient of Variation: Connecting Sampling with some Increasing Distribution Models *Proceedings of Stochastic Modelling Techniques Data Analysis International Conference (SMTDA2010), Chania Crete Greece*, 259–267, 2010.

[10] P. S. Levy, S. Lemeshow. *Sampling of Populations: Methods and Applications, 4th edition.* John Wiley Sons, Inc., New York, 2008.

[11] T. A. Severini. *Elements of Distribution Theory.* Cambridge University Press, New York, 2005.

A Random Graph Evolution Procedure and Asymptotic Results

Bettina Porvázsnyik*[1], István Fazekas[1], Csaba Noszály[1] and Attila Perecsényi[1]

[1]*Department of Applied Mathematics and Probability Theory, University of Debrecen, P.O. Box 12, 4010 Debrecen, Hungary*

Abstract: Examination of random networks is one of the most popular topics in discrete probability theory. A large number of random graph models has been proposed and investigated to describe complex networks (see [2] and [3] for an overview). During the last two decades, many types of real world networks were studied by several researchers. It was shown that a main common characteristic of the most of real-world networks is their scale-free nature ([1]). According to Barabási and Albert, networks are called scale-free when the degree distribution has a power-law tail.

In our paper, we introduce a random graph model evolving over discrete time. The evolution of the graph is based on the interaction of N vertices. During the evolution both the preferential attachment rule and the uniform choice of vertices are allowed. In our model every vertex is characterized by three main parameters: its degree and its two weights. The weights of a given vertex describes the number and the type of its interactions. Asymptotic results for the model are presented. Besides mathematical proof, numerical evidence is also given for the power-law distribution.

Keywords: random graph, preferential attachment, scale-free, power law, submartingale

AMS subject classifications: 05C80, 60G42

Acknowledgements: István Fazekas and Bettina Porvázsnyik were supported by the TÁMOP-4.2.2.C-11/1/KONV-2012-0001 project. The project has been supported by the European Union, co-financed by the European Social Fund.

References

[1] A. L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[2] R. Durrett. *Random graph dynamics*. Cambridge University Press, Cambridge, UK, 2007.

[3] R. van der Hofstad. *Random Graphs and Complex Networks*. Eindhoven University of Technology, The Netherlands, 2013.

*Corresponding author: porvazsnyik.bettina@inf.unideb.hu

Finite Sample Properties of Tests Based on Prewhitened Nonparametric Covariance Estimators

David Preinerstorfer*

Department of Statistics and Operations Research, University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria

Abstract: We analytically investigate size and power properties of a popular family of procedures for testing linear restrictions on the coefficient vector in a linear regression model with temporally dependent errors. The tests considered are autocorrelation-corrected F-type tests based on prewhitened nonparametric covariance estimators that possibly incorporate a data-dependent bandwidth parameter, e.g., estimators as considered in Andrews and Monahan (1992), Newey and West (1994), or Rho and Shao (2013). For design matrices that are generic in a measure theoretic sense we prove that these tests either suffer from extreme size distortions or from strong power deffciencies. Despite this negative result we demonstrate that a simple adjustment procedure based on artificial regressors can often resolve this problem.

Keywords: autocorrelation robustness, size distortion, power deficiency, artificial regressors, prewhitening

AMS subject classifications: 62F03, 62J05, 62F35, 62M10, 62M15

References

[1] Andrews, D. W. K. and Monahan, J. C. An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator. *Econometrica*, 60:953–966, 1992.

[2] Newey, W. K. and West, K. D. Automatic lag selection in covariance matrix estimation. *The Review of Economic Studies*, 61:631–653, 1994.

[3] Rho, Y. and Shao, X. Improving the bandwidth-free inference methods by prewhitening. *Journal of Statistical Planning and Inference*, 143:1912–1922, 2013.

*Corresponding author: david.preinerstorfer@univie.ac.at

On the High Energy Behavior of Nonlinear Functionals of Random Eigenfunctions on $\mathbb{S}^d$

Maurizia Rossi*

Dipartimento di Matematica, Università di Roma "Tor Vergata", Italy

Abstract: In this short survey we recollect some of the recent results on the high energy behavior (i.e., for diverging sequences of eigenvalues) of nonlinear functionals of Gaussian eigenfunctions on the d-dimensional sphere $\mathbb{S}^d$, $d \geq 2$. We present a quantitative Central Limit Theorem for a class of functionals whose Hermite rank is two, which includes in particular the empirical measure of excursion sets in the non-nodal case. Concerning the nodal case, we recall a CLT result for the defect on $\mathbb{S}^2$. The key tools are both, the asymptotic analysis of moments of all order for Gegenbauer polynomials, and so-called Fourth-Moment theorems.

Keywords: Gaussian eigenfunctions, high energy asymptotics, quantitative central limit theorems, excursion volume, Gegenbauer polynomials

AMS subject classifications: 60G60, 42C10, 60D05, 60B10

1 Introduction

Let us consider a compact Riemannian manifold $(\mathcal{M}, g)$ and denote by $\Delta_{\mathcal{M}}$ its Laplace-Beltrami operator. There exists a sequence of eigenfunctions $\{f_j\}_{j\in\mathbb{N}}$ and a corresponding non-decreasing sequence of eigenvalues $\{E_j\}_{j\in\mathbb{N}}$

$$\Delta_{\mathcal{M}} f_j + E_j f_j = 0 \ ,$$

such that $\{f_j\}_{j\in\mathbb{N}}$ is a complete orthonormal basis of $L^2(\mathcal{M})$, the space of square integrable measurable functions on $\mathcal{M}$. One is interested in the high energy behavior i.e., as $j \to +\infty$, of eigenfunctions f_j, related to the geometry of both *level sets* $f_j^{-1}(z)$ for $z \in \mathbb{R}$, and connected components of their complement $\mathcal{M} \setminus f_j^{-1}(z)$. One can investigate e.g. the Riemannian volume of these domains: a quantity that can be formally written as a *nonlinear functional* of f_j.

The nodal case corresponding to $z = 0$ has received great attention (for motivating details see [11]).

At least for "generic" chaotic *surfaces* $\mathcal{M}$, Berry's Random Wave Model allows to compare the eigenfunction f_j to a "typical" instance of an isotropic, monochromatic *random* wave with wavenumber $\sqrt{E_j}$ (see [11]). In view of this, much effort has been first devoted to 2-dimensional manifolds such as the torus $\mathbb{T}^2$ (see e.g. [4]) and the sphere $\mathbb{S}^2$ (see e.g. [3], [2], [8], [12]). Spherical random fields have attracted a growing interest, as they model several data sets in Astrophysics and Cosmology, e.g. on Cosmic Microwave Background ([5]).

More recently random eigenfunctions on higher dimensional manifolds have been investigated: e.g. on the hyperspheres ([6]).

*Corresponding author: rossim@mat.uniroma2.it

1.1 Random eigenfunctions on $\mathbb{S}^d$

Let us fix some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, denote by $\mathbb{E}$ the corresponding expectation and by $\mathbb{S}^d \subset \mathbb{R}^{d+1}$ the unit d-dimensional sphere ($d \geq 2$); μ_d stands for the Lebeasgue measure of the hyperspherical surface. By real random field on $\mathbb{S}^d$ we mean a real-valued measurable map defined on $(\Omega \times \mathbb{S}^d, \mathcal{F} \otimes \mathcal{B}(\mathbb{S}^d))$, where $\mathcal{B}(\mathbb{S}^d)$ denotes the Borel σ-field on $\mathbb{S}^d$. Recall that the eigenvalues of the Laplace-Beltrami operator $\Delta_{\mathbb{S}^d}$ on $\mathbb{S}^d$ are integers of the form $-\ell(\ell+d-1) =: -E_\ell$, $\ell \in \mathbb{N}$.

The ℓ-th random eigenfunction T_ℓ on $\mathbb{S}^d$ is the (unique) centered, isotropic real Gaussian field on $\mathbb{S}^d$ with covariance function

$$K_\ell(x,y) := G_{\ell;d}(\cos \tau(x,y)) \quad x, y \in \mathbb{S}^d \ ,$$

where $G_{\ell;d}$ stands for the ℓ-th Gegenbauer polynomial normalized in such a way that $G_{\ell;d}(1) = 1$ and τ is the usual geodesic distance. More precisely, setting $\alpha_{\ell;d} := \binom{\ell+\frac{d}{2}-1}{\ell}$, we have $G_{\ell;d} = \alpha_{\ell;d}^{-1} P_\ell^{(\frac{d}{2}-1,\frac{d}{2}-1)}$, where $P_\ell^{(\alpha,\beta)}$ denote standard Jacobi polynomials. By isotropy (see e.g. [5]) we mean that for every $g \in SO(d+1)$, the random fields $T_\ell = (T_\ell(x))_{x\in\mathbb{S}^d}$ and $T_\ell^g := (T_\ell(gx))_{x\in\mathbb{S}^d}$ have the same law in the sense of finite-dimensional distributions. Here $SO(d+1)$ denotes the group of real $(d+1)\times(d+1)$-matrices A such that $AA' = I$ the identity matrix and $\det A = 1$.

Random eigenfunctions naturally arise as they are the Fourier components of those isotropic random fields on $\mathbb{S}^d$ whose sample paths belong to $L^2(\mathbb{S}^d)$.

Let us consider now functionals of T_ℓ of the form

$$S_\ell(M) := \int_{\mathbb{S}^d} M(T_\ell(x))\, dx \ , \tag{1}$$

where $M : \mathbb{R} \to \mathbb{R}$ is some measurable function such that $\mathbb{E}[M(Z)^2] < +\infty$, $Z \sim \mathcal{N}(0,1)$ a standard Gaussian r.v. In particular, if $M(\cdot) = 1(\cdot > z)$ is the indicator function of the interval $(z, +\infty)$ for $z \in \mathbb{R}$, then (1) coincides with the empirical measure $S_\ell(z)$ of the *z-excursion set* $A_\ell(z) := \{x \in \mathbb{S}^d : T_\ell(x) > z\}$.

1.2 Aim of the survey

We first present a quantitative CLT as $\ell \to +\infty$ for nonlinear functionals $S_\ell(M)$ in (1) on $\mathbb{S}^d$, $d \geq 2$, under the assumption that $\mathbb{E}[M(Z)H_2(Z)] \neq 0$, where $H_2(t) := t^2 - 1$ is the second Hermite polynomial.

For instance the above condition is fullfilled by the empirical measure $S_\ell(z)$ of z-excursion sets for $z \neq 0$. For the nodal case which corresponds to the defect

$$D_\ell := \int_{\mathbb{S}^d} 1(T_\ell(x) > 0)\, dx - \int_{\mathbb{S}^d} 1(T_\ell(x) < 0)\, dx \ , \tag{2}$$

we present a CLT for $d = 2$. Quantitative CLTs for D_ℓ on $\mathbb{S}^d$, $d \geq 2$, will be treated in a forthcoming paper.

We refer to [7], [8] and [6] for the spherical case $d = 2$ and to [6] for all higher dimensions. The mentioned results rely on both, the asymptotic analysis of moments of all order for Gegenbauer polynomials, and Fourth-Moment theorems (see [9], [1]).

2 High energy behavior via chaos expansions

For a function $M : \mathbb{R} \to \mathbb{R}$ as in (1), the r.v. $S_\ell(M)$ admits the chaotic expansion

$$S_\ell(M) = \sum_{q=0}^{+\infty} \frac{J_q(M)}{q!} \int_{\mathbb{S}^d} H_q(T_\ell(x))\, dx \tag{3}$$

(see [9]) in $L^2(\mathbb{P})$ (the space of finite-variance r.v.'s), where H_q is the q-th Hermite polynomial (see e.g. [10]) and $J_q(M) := \mathbb{E}[M(Z)H_q(Z)]$, $Z \sim \mathcal{N}(0,1)$. We have $\mathbb{E}[S_\ell(M)] = J_0(M)\mu_d$; w.l.o.g. $J_0(M) = 0$.

The main idea is first to investigate the asymptotic behavior of each chaotic projection, i.e. of each (centered) r.v. of the form

$$h_{\ell;q,d} := \int_{\mathbb{S}^d} H_q(T_\ell(x))\, dx \tag{4}$$

and then deduce the asymptotic behavior of the whole series (3). Note that $h_{\ell;1,d} = 0$, as T_ℓ has zero mean on $\mathbb{S}^d$. By the symmetry property of Gegenbauer polynomials ([10]), from now on we can restrict ourselves to even multiples ℓ, for which some straightforward computations yield

$$\mathrm{Var}[h_{\ell;q,d}] = 2q!\mu_d\mu_{d-1} \int_0^{\pi/2} G_{\ell;d}(\cos\vartheta)^q (\sin\vartheta)^{d-1}\, d\vartheta\ . \tag{5}$$

2.1 Asymptotics for moments of Gegenbauer polynomials

The proof of the following is in [7], [8] for $d = 2$ and in [6] for $d \geq 3$.

Proposition 7. *As $\ell \to \infty$, for $d = 2$ and $q = 3$ or $q \geq 5$ and for $d, q \geq 3$,*

$$\int_0^{\frac{\pi}{2}} G_{\ell;d}(\cos\vartheta)^q (\sin\vartheta)^{d-1}\, d\vartheta = \frac{c_{q;d}}{\ell^d}(1 + o(1))\ . \tag{6}$$

The constants $c_{q;d}$ are given by the formula

$$c_{q;d} := \left(2^{\frac{d}{2}-1}\left(\frac{d}{2}-1\right)!\right)^q \int_0^{+\infty} J_{\frac{d}{2}-1}(\psi)^q \psi^{-q\left(\frac{d}{2}-1\right)+d-1} d\psi\ , \tag{7}$$

where $J_{\frac{d}{2}-1}$ is the Bessel function ([10]) of order $\frac{d}{2}-1$. The r.h.s. integral in (7) is absolutely convergent for any pair $(d,q) \neq (2,3),(3,3)$ and conditionally convergent for $d = 2, q = 3$ and $d = q = 3$. Moreover for $c_{4;2} := \frac{3}{2\pi^2}$

$$\int_0^{\frac{\pi}{2}} G_{\ell;2}(\cos\vartheta)^4 \sin\vartheta\, d\vartheta = \ c_{4;2}\frac{\log\ell}{\ell^2}(1 + o(1))\ . \tag{8}$$

From [10], as $\ell \to +\infty$,

$$\int_0^{\frac{\pi}{2}} G_{\ell;d}(\cos\vartheta)^2 (\sin\vartheta)^{d-1}\, d\vartheta = 4\mu_d\mu_{d-1}\frac{c_{2;d}}{\ell^{d-1}}(1 + o(1))\ ,\ c_{2;d} := \frac{(d-1)!\mu_d}{4\mu_{d-1}}\ . \tag{9}$$

Clearly for any $d, q \geq 2$, $c_{q;d} \geq 0$ and $c_{q;d} > 0$ for all even q. Moreover we can give explicit expressions for $c_{3;2}, c_{4;2}$ and $c_{2;d}$ for any $d \geq 2$. We conjecture that the above strict inequality holds for every pair (d,q), and leave this issue as an open question for future research.

2.2 Fourth-Moment Theorems for chaotic projections

Let us recall the usual Kolmogorov d_K, total variation d_{TV} and Wasserstein d_W distances between r.v.'s X, Y: for $\mathcal{D} \in \{K, TV, W\}$

$$d_{\mathcal{D}}(X,Y) := \sup_{h \in H_{\mathcal{D}}} |\mathbb{E}[h(X)] - \mathbb{E}[h(Y)]| \ ,$$

where $H_K = \{1(\cdot \leq z), z \in \mathbb{R}\}$, $H_{TV} = \{1_A(\cdot), A \in \mathcal{B}(\mathbb{R})\}$ and H_W is the set of Lipschitz functions with Lipschitz constant one.

The r.v. $h_{\ell;q,d}$ in (4) belongs to the so-called qth Wiener chaos. The Fourth-Moment Theorem ([9]) states that if $Z \sim \mathcal{N}(0,1)$, for $\mathcal{D} \in \{K, TV, W\}$ we have

$$d_{\mathcal{D}}\left(\frac{h_{\ell;q,d}}{\sqrt{\mathrm{Var}[h_{\ell;q,d}]}}, Z\right) \leq C_{\mathcal{D}}(q)\sqrt{\frac{\mathrm{cum}_4(h_{\ell;q,d})}{\mathrm{Var}[h_{\ell;q,d}]^2}} \ , \tag{10}$$

where $C_{\mathcal{D}}(q) > 0$ is some explicit constant and $\mathrm{cum}_4(h_{\ell;q,d})$ is the fourth cumulant of the r.v. $h_{\ell;q,d}$. An application of (10) together with upper bounds for cumulants leads to the following result (see [6]).

Theorem 29. *For all $d, q \geq 2$ and $\mathcal{D} \in \{K, TV, W\}$ we have, as $\ell \to +\infty$,*

$$d_{\mathcal{D}}\left(\frac{h_{\ell;q,d}}{\sqrt{\mathrm{Var}[h_{\ell;q,d}]}}, Z\right) = O\left(\ell^{-\delta(q;d)}(\log \ell)^{-\eta(q;d)}\right) \ , \tag{11}$$

where $\delta(q;d) \in \mathbb{Q}$, $\eta(q;d) \in \{-1, 0, 1\}$ and $\eta(q;d) = 0$ but for $d = 2$ and $q = 4, 5, 6$.

The exponents $\delta(q;d)$ and $\eta(q;d)$ can be given explicitly (see [6]), turning out in particular that if $(d,q) \neq (3,3), (3,4), (4,3), (5,3)$ and $c_{q;d} > 0$,

$$\frac{h_{\ell;q,d}}{\sqrt{\mathrm{Var}[h_{\ell;q,d}]}} \stackrel{\mathcal{L}}{\to} Z \ , \qquad \text{as } \ell \to +\infty \ , \tag{12}$$

where from now on, $\to^{\mathcal{L}}$ denotes convergence in distribution and $Z \sim \mathcal{N}(0,1)$.

Remark 5. For $d = 2$, the CLT (12) was already proved in [8]; nevertheless Theorem 29 improves the existing bounds on the rate of convergence to the asymptotic Gaussian distribution.

2.3 Quantitative CLTs for Hermite rank 2 functionals

Proposition 7 states that whenever M is such that $J_2(M) \neq 0$ in (3), i.e. the functional $S_\ell(M)$ in (1) has Hermite rank two, then

$$\lim_{\ell \to +\infty} \frac{\mathrm{Var}[S_\ell(M)]}{\mathrm{Var}\left[\frac{J_2(M)}{2} h_{\ell;2,d}\right]} = 1 \ . \tag{13}$$

Hence, loosely speaking, $S_\ell(M)$ and its 2nd chaotic projection $\frac{J_2(M)}{2} h_{\ell;2,d}$ have the same high energy behaviour. The main result presented in this survey is the following, whose proof is given in [6].

Theorem 30. *Let $M : \mathbb{R} \to \mathbb{R}$ in (1) be s.t. $\mathbb{E}[M(Z)H_2(Z)] =: J_2(M) \neq 0$, then*

$$d_W\left(\frac{S_\ell(M)}{\sqrt{\mathrm{Var}[S_\ell(M)]}}, Z\right) = O\left(\ell^{-\frac{1}{2}}\right) , \qquad \text{as } \ell \to \infty , \tag{14}$$

where $Z \sim \mathcal{N}(0,1)$. In particular, as $\ell \to +\infty$,

$$\frac{S_\ell(M)}{\sqrt{\mathrm{Var}[S_\ell(M)]}} \xrightarrow{\mathcal{L}} Z . \tag{15}$$

3 Geometry of high energy excursion sets

Consider the empirical measure $S_\ell(z)$ of the z-excursion set $A_\ell(z)$ for $z \in \mathbb{R}$, as in §1.1. It is easy to check that in (3) $\mathbb{E}[S_\ell(z)] = \mu_d(1 - \Phi(z))$ and for $q \geq 1$, $J_q(1(\cdot > z)) = H_{q-1}(z)\phi(z)$, where Φ and ϕ denote respectively the cdf and the pdf of the standard Gaussian law. Since $J_2(1(\cdot > z) = z\phi(z)$, Theorem 30 immediately entails that, as $\ell \to \infty$, if $z \neq 0$

$$d_W\left(\frac{S_\ell(z) - \mu_d(1 - \Phi(z))}{\sqrt{\mathrm{Var}[S_\ell(z)]}}, Z\right) = O\left(\ell^{-\frac{1}{2}}\right) .$$

The nodal case $z = 0$ requires different arguments: in the chaos expansion for the defect (2) D_ℓ only odd chaoses occur but each of them "contributes" by Proposition 7. Asymptotics for the defect variance on $\mathbb{S}^2$ have been given in [7]:

$$\mathrm{Var}[D_\ell] = \frac{C}{\ell^2}(1 + o(1)) , \quad \text{as } \ell \to +\infty ,$$

for $C > \frac{32}{\sqrt{27}}$. Moreover in [8] a CLT has been proved: as $\ell \to +\infty$,

$$\frac{D_\ell}{\sqrt{\mathrm{Var}[D_\ell]}} \xrightarrow{\mathcal{L}} Z ,$$

where $Z \sim \mathcal{N}(0,1)$. In a forthcoming paper, we will provide quantitative CLTs for the defect on $\mathbb{S}^d$, $d \geq 2$.

Remark 6. The volume of excursion sets is just one instance of Lipschtz-Killing curvatures. In the 2-dimensional case, these are completed by the Euler-Poincaré characteristic ([3]) and the length of level curves ([4],[12] for the nodal variances). In forthcoming papers jointly with D. Marinucci, G. Peccati and I. Wigman, we will investigate the asympotic distribution of the latter on both the sphere $\mathbb{S}^2$ and the 2-torus $\mathbb{T}^2$. Our big proposal for the future is to characterize the high energy behavior of all Lipschitz-Killing curvatures on every "nice" compact manifold.

Acknowledgements: We thank D. Marinucci for valuable suggestions, P. Baldi, S. Campese and S. Cipolla for a careful reading of an earlier version of this work.

This research is supported by ERC Grant *Pascal* n.277742.

References

[1] E. Azmoodeh, S. Campese, and G. Poly. Fourth Moment Theorems for Markov diffusion generators. *Journal of Functional Analysis*, 266(4):2341–2359, 2014.

[2] V. Cammarota, and D. Marinucci. On the Limiting behaviour of needlets polyspectra. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, in press.

[3] V. Cammarota, D. Marinucci, and I. Wigman. Fluctuations of the Euler-Poincaré characteristic for random spherical harmonics. *Preprint*, arXiv:1504.01868.

[4] M. Krishnapur, P. Kurlberg, and I. Wigman. Nodal length fluctuations for arithmetic random waves. *Annals of Mathematics*, 177(2):699–737, 2013.

[5] D. Marinucci, and G. Peccati. *Random fields on the sphere: Representations, Limit Theorems and Cosmological Applications*. London Mathematical Society Lecture Notes, Cambridge University Press, 2011.

[6] D. Marinucci, and M. Rossi. Stein-Malliavin approximations for nonlinear functionals of random eigenfunctions on $\mathbb{S}^d$. *Journal of Functional Analysis*, 268(8):2379–2420, 2015.

[7] D. Marinucci, and I. Wigman. The defect variance of random spherical harmonics. *Journal of Physics A: Mathematical and Theoretical*, 44(35):355206, 2011.

[8] D. Marinucci, and I. Wigman. On nonlinear functionals of random spherical eigenfunctions. *Communications in Mathematical Physics*, 327(3):849–872, 2014.

[9] I. Nourdin, and G. Peccati. *Normal Approximations Using Malliavin Calculus: From Steins Method to Universality*. Cambridge University Press, 2012.

[10] G. Szego. *Orthogonal Polynomials, 4th Edition*. Colloquium Publications of the American Mathematical Society, 1975.

[11] I. Wigman. On the nodal lines of random and deterministic Laplace eigenfunctions. *Proceedings of the International Conference on Spectral Geometry, Dartmouth College*, 84:285-298, 2012.

[12] I. Wigman. Fluctuations of the nodal length of random spherical harmonics. *Communications in Mathematical Physics*, 298(3):787–831, 2010.

Network Sparsity Selection and Robust Estimation via Bootstrap with Applications to Genomic Data

José Sánchez*[1], Alexandra Jauhiainen[2], Sven Nelander[3] and Rebecka Jörnsten[4]

[1] *Bioinformatics CF, University of Gothenburg, Sweden*
[2] *Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Sweden*
[3] *IGP and Science for Life Laboratory, Uppsala University, Sweden*
[4] *Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, Sweden*

Abstract: Network estimation methods are popular tools to analyze different types of omics data. Most network methods require the selection of a tuning parameter (threshold) that controls the overall sparsity of the network estimate. Given sample size and signal strength, the goal should be to assemble a reproducible network with few false positives. Controlling the number of false positives is crucial in order to avoid overinterpretation and erroneous conclusions in the scientific disciplines where the network models are applied.

We present a bootstrap based method to address the sparsity selection problem. Our method: (i) selects an appropriate sparsity level controlling the presence of false positive edges, and (ii) constructs a final network estimate that improves over naive bootstrap threshold methods. To demonstrate these properties, we employ a comprehensive simulation study using varying true sparsities, network and sample sizes. Finally, we illustrate our procedure on a large-scale ovarian tumor sample data from The Cancer Genome Atlas (TCGA).

Keywords: network, bootstrap, high-dimension, lasso, cancer

AMS subject classifications: 62H12, 62F40, 62P10

1 Introduction

Network methods are increasingly favoured for analyzing large scale omics data. Their popularity stems both from their appealing visualization of complex data, and their potential to highlight mechanistic hypotheses and identify key hub variables. Such network structures may aid in the discovery of new drug targets essential for therapeutic development [2].

Several methods are available for network estimation. They all have in common the need of tuning parameters that control the sparsity of the resulting estimate. A large (conservative) parameter usually corresponds to a sparse network, which

*Corresponding author: jose.sanchez.lopez@gu.se

means that the network has few edges, while a small (liberal) value for the threshold will give a more dense network.

Here we propose an innovative use of the marginal statistics collected by bootstrap and present a method not only to select a sparsity level, but also to improve a single estimate generating a more robust estimate. Choosing a network size that matches that of the true network (which most often is unknown) does not guarantee a quality estimate, since it may include a large number of false positives. Instead, we propose a novel method, employing bootstrap and frequency statistics, that accurately controls the false positive rate.

The paper is structured as follows. In the methods section we describe our framework. In the results section we present a simulation study that illustrates the performance of our method and apply it to a genomic dataset on ovarian cancer extracted from The Cancer Genome Atlas (TCGA). We conclude the paper with a future work section.

2 Methods

The first step in our framework is to perform bootstrap of network estimates, producing a set of bootstrap graphs (networks) for a number of sparsity levels. Consider the data set X, applying network estimation method M_λ (here λ denotes the penalty parameter which controls the sparsity level) to bootstrap data set X^b, $b = 1, 2, \ldots, B$, we obtain a set of bootstrap networks $\{\widehat{\Theta}^1_\lambda, \widehat{\Theta}^2_\lambda, \ldots, \widehat{\Theta}^B_\lambda\}$.

The next step is to summarize the bootstrap networks into frequency statistics. The frequency statistic, $h_{ij,\lambda}$, for edge (i, j), is given by $h_{ij,\lambda} = \frac{1}{B}\sum_{b=1}^{B} \mathrm{I}\left(\left|\widehat{\theta}^b_{ij,\lambda}\right|\right)$, where $\mathrm{I}(x) = 1$ if $x > 0$ and 0 otherwise. Thus $h_{ij,\lambda}$ is the number of times the edge is present across bootstrap estimates, divided by the total number of bootstraps. Here, $\widehat{\theta}_{ij,\lambda}$ is the b-th bootstrap estimate of (i, j). We assemble the frequency statistics in a matrix $H_\lambda = [h_{ij,\lambda}]$ and inspect their distributions.

For very sparse networks, edge absence (zero counts) dominates histograms for H_λ. As the threshold parameter decreases, more and more edges are present across bootstrap networks and the histograms start to resemble a bimodal U-shape, where one mode corresponds to edges that are consistently absent across bootstraps and the other one to edges that are consistently present. For smaller values of the threshold parameter, we observe the left mode in the histogram (corresponding to absent edges) shifting to the right, which we consider to be a sign of overfitting.

The presence of these two edge populations; negatives (N, edges not present in the true network) and positives (P, edges present in the true network), motivates the use of a mixture to model the frequency statistics. The negative component has a natural parameter of interest, the false positive rate (FPR); here defined as the average failure rate (N $\rightarrow$ FP) across the negative population. Similarly, a success in the negative component corresponds to a negative edge being correctly classified as negative (N $\rightarrow$ TN). Likewise, the natural parameter of interest for the positive component is the average true positive rate (TPR), i.e. the power, here defined as 1 minus the average failure rate (P $\rightarrow$ FN) across the positive population. Similarly, a success in the positive population corresponds to a positive edge being correctly classified as positive (P $\rightarrow$ TP).

Due to the fact that the signal strength of each edge has a complex dependency on other edges present in the network, the failure rate of a single edge may vary substantially around the average for the edge population. This makes a simple binomial model for the number of failures in each population unsuitable. Instead, we propose the use of a Beta-Binomial mixture model to capture the inflated variability arising from the edge-specific failure rates in the two populations.

Each negative edge (i, j) thus has an edge-specific failure rate of becoming a false positive, $p_0(i, j)$. Similarly, each positive edge (k, l) has an edge-specific failure rate, $p_1(k, l)$ of becoming a false negative. Considering the network level population of negatives, we define the *average* failure rate as μ_0, and likewise for the population of positives, the *average* failure rate as μ_1. The mixture parameter π_0 of the two components in the model then represents the proportion of negatives and its complement, $1 - \pi_0$, the proportion of positives.

In the Beta-Binomial framework, $p_0(i, j)$ is assumed to come from a (prior) $\text{Beta}(\alpha_0, \beta_0)$ distribution, where the FPR can be defined as $\mu_0 = \frac{\alpha_0}{\alpha_0+\beta_0}$. The number of failures for edge (i, j), N $\rightarrow$ FP, is then modelled as an observation from a $\text{Bin}(B, p_0(i, j))$ distribution, where B is the number of bootstrap estimates. Similarly, $p_1(k, l)$ comes from $\text{Beta}(\alpha_1, \beta_1)$ distribution, where $\mu_1 = \frac{\alpha_1}{\alpha_1+\beta_1}$ = 1-TPR (average power). The number of failures for edge (k, l), P $\rightarrow$ FN, is then an observation from a $\text{Bin}(B, p_1(k, l))$ distribution.

In detail, the number of times edge (i, j) is present across bootstrap estimates $\widehat{\theta}^b_{ij,\lambda}$, given by $x_{ij,\lambda} = \sum_{b=1}^{B} \text{I}\left(\left|\widehat{\theta}^b_{ij,\lambda}\right|\right)$, is an observation of the random variable $X_{ij,\lambda}$ with distribution

$$f_{X_{ij,\lambda}}(k) = \pi_0 \binom{B}{k} \frac{\text{Be}(k + \alpha_0, k + \beta_0)}{\text{Be}(\alpha_0, \beta_0)} + (1 - \pi_0) \binom{B - k}{k} \frac{\text{Be}(B - k + \alpha_1, B - k + \beta_1)}{\text{Be}(\alpha_1, \beta_1)},$$

where Be is the Beta function, defined as $\text{Be}(x, y) = \int_0^1 t^{x-1}(1 - t)^{y-1} dt$.

The prior parameter μ_0 is the average false positive rate of the negative population. If the sparsity parameter is very stringent, the estimated μ_0 approaches 0, and if the sparsity constraint is relaxed, the μ_0 estimate can be quite large, thus reflecting the inclusion of a large number of false positives.

As the Beta-Binomial model offers a way to estimate the average false positive rate, we propose to select a sparsity level such that this estimate, $\widehat{\mu}_0 = \widehat{\mu}_0(\lambda)$, is below some predetermined value t ($t = 0.05$, for example). The optimal estimated network size, N^*, becomes thus the one corresponding to $\lambda^* = \max\{\lambda : \widehat{\mu}_0(\lambda) < t\}$.

Once N^* has been selected, network estimates can be post-processed to construct more robust estimates. The Beta-Binomial model produces a natural estimate for a network at any given sparsity level λ, as described below.

Let Δ_{ij} be the unobserved true class (positive or negative) of edge (i, j). The expected value of Δ_{ij}, given the parameters of the model and the edge presence counts, $\gamma_{ij}(\alpha_0, \beta_0, \alpha_1, \beta_1, \pi_0) = E(\Delta_{ij}|\alpha_0, \beta_0, \alpha_1, \beta_1, \pi_0; X_{ij,\lambda})$, is estimated by the EM-Algorithm by

$$\widehat{\gamma}_{ij} = \frac{\widehat{\pi}_0 \text{BetaBin}(x_{ij,\lambda}; \widehat{\alpha}_0, \widehat{\beta}_0)}{\widehat{\pi}_0 \text{BetaBin}(x_{ij,\lambda}; \widehat{\alpha}_0, \widehat{\beta}_0) + (1 - \widehat{\pi}_0)\text{BetaBin}(x_{ij,\lambda}; \widehat{\alpha}_1, \widehat{\beta}_1)}.$$

That is, the $\widehat{\gamma}_{ij}$ is an estimate of the class expected value of edge (i, j). The final network is thus constructed by removing edges where $\widehat{\gamma}_{ij} < 0.5$.

3 Results

Simulation study

We compare our proposed method on simulated data with BIC, CV and Hartigans' DIP statistic [1]. The DIP statistic compares each histogram with a uniform distribution, with a large test statistic indicating a substantial deviation from unimodality. The simulated data is generated from a network comprising 100 genes, constructed from expression data for glioblastoma from TCGA. We consider a sparse scenario where the true network contains about 13% non-zeros, and a dense one, where the true network contains about 21% non-zeros.

In the sparse setting both the Beta-Binomial model and the DIP select network sizes closely matching the true sparsity, which also corresponds to the size where the F1 measure is maximized (results not shown). At the same time, the Beta-Binomial model controls the FPR at the chosen level, while DIP is a bit less stringent (FPR < 0.1), as illustrated in Figure 1, top row. Both BIC and CV overestimate the network size, which means that both more true positives and false positives are included in the network estimate. However, the fact that the F1 is lower indicates a higher inclusion rate of false positives than true positives.

Estimating a dense network is a more difficult task compared to the sparse setting. This is reflected in the fact that the maximum of the F1 measure is less distinct (results not shown). Also, it is not possible to simultaneously control the FPR while maximizing the F1 measure (which was the case in the sparse setting). This is clearly shown for BIC and CV, which have higher TPR and F1, but at the price of including a higher number of false positives, see Figure 1 (bottom row). The Beta-Binomial and DIP methods control the FPR around 0.05 and 0.1, respectively.

Application to cancer genomic data

We apply our methodology to expression data from ovarian cancer tumors from The Cancer Genome Atlas (TCGA). The data set comprises 266 samples and ∼20500 genes. After pre-processing to take into account variability and the estimated network's connected components, the data set is reduced to ∼2000 genes. A total of 1000 bootstrap networks are estimated for a sequence of 10 values of the glasso threshold parameter. As we aim to model reasonably strong signals, with our screening process we have filtered out undesired noise prior to modeling. We need thus to focus only on threshold parameters between 0.7 (otherwise a larger set of genes should have been included) and 0.9 (since larger ones produce empty networks).

The results for network size selection are shown in Figure 2, top row. We include the Beta-Binomial models that control the FPR at 0.01 and 0.05. The former selects a very sparse network of about 4000 edges while the former opts for one of size 11,000. DIP selects a network with 16,000 edges, however, this still

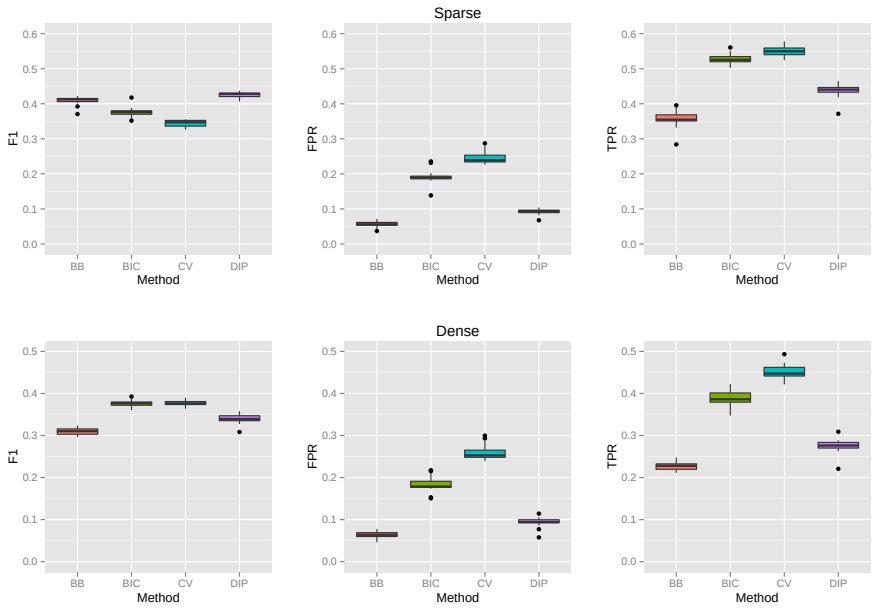


Figure 1: **Network size selection performance for simulated data.**

corresponds to a very sparse network with about only 3.5% of all the potential edges present. BIC and CV perform poorly; BIC (rescaled to be comparable with CV) is almost constant but with a minimum at the smallest network. CV, on the other hand, shows a steady decrease and chooses the largest network.

For further validation we compute the overlap of our final network assemblies with pathways from the PathwayCommons database. The overlap is computed as a fold enrichment of hits from the HPRD, Intact, NCI Nature and Reactome data bases. Figure 2, bottom-left, shows the results for Beta-Binomial classification and two alternative methods to assemble final estimaes (permuted and degree-preserved permuted networks, which offer a more principled way to select a bootstrap threshold as in [3], details omitted). The values for the x-axis correspond to the average network sizes across bootstraps for the selected values of λ. The vertical lines indicate the network sizes selected by each method. Degree-preserved network and Beta-Binomial classification have the best fold enrichment levels, suggesting that the constructed networks contain relevant connections which overlap with known biological pathways. Post-processing has an effect in the final network size though. In Figure 2, bottom-right, we show the fold enrichment levels plotted against the post-processed network sizes. Beta-Binomial classification and degree-preserved networks reduce the network sizes in order to correct for the inclusion of false positives. This effect is seen as a shift in the corresponding curves.

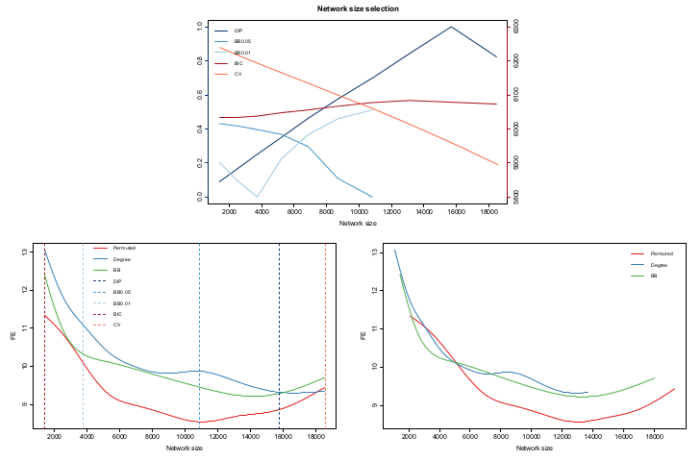


Figure 2: **Network size selection performance for real data.**

4 Future work

Our method can be used to select the tuning parameter for any network property that can be collected as a frequency statistic. Joint modeling of networks for groups of samples belonging to different cancer types, can require fusing of edges (the edge values are estimated to be equal across classes). This problem can be solved by an extension of the graphical lasso, the so called fused graphical lasso ([4]), which requires tuning of a fusing parameter as well as a sparsity parameter. Fusing here, is an example of another network property for which frequency statistics can be collected. We plan to expand the applicability of our framework to include the selection of fusing parameters in the same principled way as selection of sparsity parameters.

Another line of work is to couple our framework to modularized networks, where we choose the sparsity level individually for the modular components. With this approach we control the false positive rate in each modular component separately, but while boosting the true positive rate in some modules with a higher signal-to-noise ratio. In this way, we also control the false positive rate globally, but allowing for some modules to be denser (have more true positives) than what would be allowed with a global sparsity criterion.

Future work also includes a substantial decrease of computational burden by performing a linesearch for a suitable sparsity threshold. The idea is to let the Beta-Binomial estimated parameters to guide the search aiming for a particular FPR control, e.g. 0.01 and stop once it has been achieved. This approach improves execution time by reducing the number of threshold parameters for which the network has to be estimated and by avoiding estimation of very dense networks.

Acknowledgements: The work presented here was supported by grants from the Swedish Research Council and Wallenberg Foundation.

References

[1] J. Hartigan, and P.M. Hartigan. The dip test of unimodality. *The Annals of Statistics*, 70–84, 1985.

[2] R. Jornsten, T. Abenius, T. Kling, L. Schmidt, E. Johansson, T. Nordling, B. Nordlander, C. Sander, P. Gennemark, K. Funa, B. Nilsson, L. Lindahl, and S. Nelander. Network modeling of the transcriptional effects of copy number aberrations in glioblastoma. *Mol. Syst. Biol.*, 7(3):e33624, 2012.

[3] R. de Matos Simoes, and F. Emmert-Streib. Bagging statistical network inference from large-scale gene expression data. *PLoS ONE*, 7:486, 2011.

[4] P. Danaher, and P. Wang, and D. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2): 373–397, 2014.

Adaptive Confidence Bands for Markov Chains and Diffusions: Estimating the Invariant Measure and the Drift

Jakob Söhl[1] **and Mathias Trabs**[*2]

[1] *University of Cambridge, Statistical Laboratory, CB3 0WB, Cambridge, UK.*
[2] *Humboldt-Universität zu Berlin, Institut für Mathematik, Unter den Linden 6, D-10099 Berlin, Germany.*

Abstract: As a starting point we prove a functional central limit theorem for estimators of the invariant measure of a geometrically ergodic Harris-recurrent Markov chain in a multi-scale space. Its proof is inspired by [2]. It allows to construct confidence bands for the invariant density with optimal (up to undersmoothing) L^∞-diameter by using wavelet projection estimators.

In addition our setting applies to the drift estimation for a diffusion

$$dX_t = b(X_t)dt + dW_t, \quad t \geq 0,$$

observed discretely with fixed observation distance. We prove a functional central limit theorem for estimators of the drift function. Finally, adaptive confidence bands for the drift are constructed by using a data-driven estimator and based on the ideas by [3]. Due to the Markovian structure of the observations, the proofs rely on a non-standard concentration inequality by [1].

Keywords: adaptive confidence bands, diffusion, drift estimation, ergodic Markov chain, functional central limit theorem

AMS subject classifications: 62G15, 60F05, 60J05, 60J60, 62M05

References

[1] R. Adamczak and W. Bednorz. Exponential concentration inequalities for additive functionals of Markov chains. *ESAIM. Probability and Statistics*, to appear, 2015.

[2] I. Castillo and R. Nickl. On the Bernstein–von Mises phenomenon for nonparametric Bayes procedures. *The Annals of Statistics*, 42(5):1941–1969, 2014.

[3] E. Giné and R. Nickl. Confidence bands in density estimation. *The Annals of Statistics*, 38(2):1122–1170, 2010.

*Corresponding author: trabs@math.hu-berlin.de

The c-Loss Function: Balancing Total and Individual Risk in the Simultaneous Estimation of Poisson Means

Emil Aas Stoltenberg[*1] **and Nils Lid Hjort**[1]

[1]*Department of Mathematics, University of Oslo, Norway*

Abstract: This paper is devoted to the simultaneous estimation of the means of $p \geq 2$ independent Poisson distributions. A novel loss function that penalizes bad estimates of each of the means *and* the sum of the means is introduced. Under this loss function, a class of minimax estimators that uniformly dominate the maximum likelihood estimator (MLE) is derived. Estimators in this class are shown to also be minimax and uniformly dominating under the commonly used weighted squared error loss function. Estimators in this class can be fine-tuned to limit shrinkage away from the MLE, thereby avoiding implausible estimates of means anticipated to be bigger than the others. Further light is shed on this new class of estimators by showing that it can be derived by Bayesian and empirical Bayesian methods. Moreover, a class of prior distributions for which the Bayes estimators uniformly dominate the MLE under the new loss function is derived.

Keywords: Poisson, Bayes, simultaneous, estimation

AMS subject classifications: 60G09, 62G09

1 Introduction

Let $Y = (Y_1, \ldots, Y_p)$ be independent Poisson random variables with means $\theta = (\theta_1, \ldots, \theta_p)$, and let $\gamma = \sum_{i=1}^p \theta_i$. [1] showed that the MLE, $\delta^o(Y) = Y$ (which is also the uniformly minimum variance unbiased estimator), is inadmissible under the weighted squared error loss function $L_1(\delta, \theta) = \sum_{i=1}^p \theta_i^{-1}(\delta_i - \theta_i)^2$, provided that $p \geq 2$. [1] derived the estimator

$$\delta^{CZ}(Y) = \left(1 - \frac{p-1}{p-1+Z}\right) Y, \tag{1}$$

where $Z = \sum_{i=1}^p Y_i$, and showed that this estimator possesses uniformly smaller risk than the MLE under L_1.

The estimator in (1), which "shrinks" the MLE towards the zero boundary of the parameter space, guarantees a reduction in the total risk $R(\delta, \theta) = E_\theta L_1(\delta, \theta)$ relative to the MLE. With regard to other objectives, however, it may perform poorly. First, it may yield implausible estimates of those θ_i with unusually high values. Second, it is likely to perform poorly in estimating the sum of the individual Poisson means, γ. Recall that Z, the sum of p independent Poisson random variables, is itself Poisson with mean γ. In many situations where one is interested in estimating an ensemble of Poisson means, one may also be interested in

*Corresponding author: emilas@math.uio.no

a good estimate of the sum of the means, or equivalently the mean of the means, $\bar{\theta} = p^{-1}\gamma$. Think, for example, of a decision maker having to make budgetary decisions concerning each of the boroughs of a city *and* the city as a whole.

The purpose of this paper is to develop a class of estimators that compromise between the Clevenson-Zidek estimator in (1) and the MLE. That is, a class of estimators that has good ensemble properties with respect to the weighted squared error loss function, and has good individual properties when it comes to estimating the individual θ_i and γ. The problem of shrinking the MLE towards non-zero points in the parameter space has been treated elsewhere, see e.g. [2].

2 The c-Loss function

Consider the following extension of the weighted squared error loss function,

$$L_c(\delta, \theta) = \sum_{i=1}^{p} \frac{1}{\theta_i} (\delta_i - \theta_i)^2 + \frac{c}{\gamma} \left(\sum_{i=1}^{p} \delta_i - \gamma \right)^2 . \tag{2}$$

This loss function is equal to the weighted squared error loss function L_1 plus an extra term that penalizes for bad estimates of γ, where the weight accorded to this extra term is a function of the user-defined constant c. The risk of the MLE under L_c is constant $R(Y, \theta) = p + c$.

We will now develop a class of estimators with uniformly smaller risk than $p + c$. Consider estimators $\delta^* = (\delta_1^*, \ldots, \delta_p^*)$ of the form $\delta^* = (1 - \phi(Z))Y$, where $Z = \sum_{i=1}^{p} Y_i$, $E_\theta|\phi(Z)| < \infty$ and $\phi(z+1)(z+1) \geq \phi(z)z$. Using that $Eg(Y)/\theta = Eg(Y+1)/(Y+1)$ (for all integrable functions such that $g(y) = 0$, $\forall\, z \leq 0$), and that conditional on Z the vector Y is multinomial with cell probabilities θ_i/γ, $1 \leq i \leq p$, we obtain an expression for the difference in risk $E_\theta D = R(\delta^*, \theta) - R(Y, \theta)$,

$$\begin{aligned} E_\theta[L_c(\delta^*, \theta) - L_c(Y, \theta)] &= E_\theta E[L_c(\delta^*, \theta) - L_c(Y, \theta)|Z] \\ &= E_\theta \left\{ (\phi^2(Z) - 2\phi(Z)) \frac{Z[(p-1) + (1+c)Z]}{\gamma} + 2(1+c)\phi(Z)Z \right\} \\ &= E_\theta \left\{ (\phi^2(Z+1) - 2\phi(Z+1))[(p-1) + (1+c)(Z+1)] \right. \\ &\qquad \left. + 2(1+c)\phi(Z)Z \right\} \\ &= E_\theta D_c(Z). \end{aligned}$$

Here D_c is independent of the parameters, so a function $\phi(z)$ that ensures that $D_c(z) \leq 0$ for all z with strict inequality for at least one datum z yields an estimator that uniformly dominates the MLE. A class of such functions is $\phi(z) = \psi(z)/(p - 1 + (1+c)z)$ where the function ψ is such that $0 < \psi(z) < 2(p-1)$ and is non-decreasing for all $z \geq 0$. This gives a class of estimators, denoted $\mathcal{D}_c$, that uniformly dominate the MLE under (2), namely

$$\delta^c(Y) = \left(1 - \frac{\psi(Z)}{p - 1 + (1+c)Z} \right) Y. \tag{3}$$

Theorem 31. *For all $\theta \in \Theta$, estimators $\delta^c \in \mathcal{D}_c$ have smaller risk than the MLE $\delta^o(Y) = Y$ when loss is given by L_c in (2).*

Proof. Use the expression for D_c above, then

$$\begin{aligned}
R(\delta^c,\theta) &= p + c + E_\gamma D_c \\
&= p + c - \frac{1}{\gamma} E_\gamma \{\phi(Z) Z [2(p-1+(1+c)(Z-\gamma)) \\
&\qquad\qquad - \phi(Z)[p-1+(1+c)Z]]\} \\
&= p + c - \frac{1}{\gamma} E_\gamma \{\phi(Z) Z [2(p-1+(1+c)(Z-\gamma)) - \psi(Z)]\} \\
&\le p + c - \frac{2}{\gamma} E_\gamma \{\phi(Z) Z (1+c)(Z-\gamma)\} \\
&= p + c - 2(1+c) E_\gamma \left[\frac{\phi(Z) Z^2}{\gamma} - \phi(Z) Z\right] \\
&= p + c - 2(1+c) E_\gamma \left[\phi(Z+1)(Z+1) - \phi(Z) Z\right] \\
&\le p + c = R(Y,\theta),
\end{aligned}$$

for all θ because $\phi(z)z$ is a strictly increasing function of z. □

The optimal choice of ψ in terms of minimizing risk is $p-1$. The estimator in $\mathcal{D}_c$ with $\psi(z) = p-1$ is denoted δ_1^c. As seen from the proof above, the risk of estimators in $\mathcal{D}_c$ depends on θ_i only through the sum γ, and is therefore easy to compute numerically. The savings in risk relative to the MLE are substantial for small values of γ, and decrease as γ grows.

The difference in risk between estimators in $\mathcal{D}_c$ and the MLE under L_1 is $E_\gamma D(Z) = E_\gamma \left\{(\phi^2(Z) - 2\phi(Z)) Z[(p-1)+Z]\gamma^{-1} + 2\phi(Z)Z\right\}$. Inserting this expression in the proof of Theorem 31 it is straightforward to show that estimators in $\mathcal{D}_c$ uniformly dominate the MLE under L_1.

3 Bayes, minimax and admissibility

Let $\theta_1, \ldots, \theta_p$ be independent and identically distributed (iid) Gamma random variables with mean a/b and variance a/b^2 (denoted $\mathcal{G}(a,b)$). The Bayes solution under the c-Loss function is then

$$\delta_j^B(y) = \frac{1+c}{1+cg(z)} \frac{a+y_j-1}{b+1}, \tag{4}$$

where $g(z) = (p(a-1)+z)/(pa-1+z)$. With the prior sequence $\{\mathcal{G}(1, b/n)\}_{n=1}^{\infty}$ the minimum Bayes risk converges to $p+c$, i.e. the maximum risk of the MLE. This implies that the MLE is minimax under (2), hence (3) must also be minimax (see e.g. [5]).

In the following we sketch how estimators in the class $\mathcal{D}_c$ can be derived by Bayesian methods in three different ways. First, the results of [3] in a L_1-setting are extended to the c-Loss function. Let $\theta_i, 1 \le i \le p$ be iid $\mathcal{G}(1,b)$ and let $b \sim \pi_2(b) \propto b^{\alpha-1}(b+1)^{-(\alpha+\beta)}$. It can then be shown that the expectation of θ_i given b and the data are as in (4) with $b+1$ replaced by $E[b+1 \,|\, Z]$. This

expectation is given by

$$E[b+1 \mid Z] = \frac{p+\alpha+\beta+z-1}{z+\beta-1}.$$

Theorem 32. *Assume that $p > 2+c$ and consider the family of prior distributions $\pi_2(b) \propto b^{\alpha-1}\,(b+1)^{-(\alpha+\beta)}$ where*

$$0 < \alpha \leq \frac{p-2-c}{1+c}$$

and $\beta > 0$. Then the Bayes solution under the c-Loss function is a member of $\mathcal{D}_c$.

Proof. Inserting the expression for $E[b+1 \mid Z]$ in the Bayes solution in (4) with $a = 1$ we obtain

$$\delta_j(Y) = \frac{(1+c)(p-1+Z)}{p-1+(1+c)Z}\,\frac{z+\beta-1}{p+\alpha+\beta+z-1}Y_i. \tag{5}$$

Recall that the estimators in $\mathcal{D}_c$ are of the form $(1-\psi(Z)/(p-1+(1+c)Z))Y_i$ where ψ is non-decreasing and $0 \leq \psi(z) \leq 2(p-1)$ for all z. By some algebra we obtain that for the Bayes solution we here consider

$$\begin{aligned}\psi(z) &= p-1+(1+c)z-(1+c)(p-1+z)\frac{z+\beta-1}{p+\alpha+\beta+z-1}\\ &= (1+c)\frac{(p-1+z)(p+\alpha)}{p-1+\alpha+\beta+z} - c(p-1).\end{aligned}$$

This function is non-decreasing for all $z \geq 0$. Moreover, we see that it is bounded above by

$$\sup_{z\geq 0}\psi(z) = (1+c)(p+\alpha) \leq 2(p-1),$$

since $\alpha \leq (p-2-c)/(1+c)$. This means that the class of Bayes solutions in (5), where α satisfies the condition of the theorem, is in $\mathcal{D}_c$. □

Second, in an empirical Bayes setup the Poisson parameters are assumed iid $\mathcal{G}(1,b)$ and the parameter b is estimated from the data. An unbiased estimator of b is $z/(p-1+z)$. Inserting this estimator in (4) (and setting $a = 1$) we obtain the estimator δ_1^c given by

$$\delta_1^c(Y) = \left(1 - \frac{p-1}{p-1+(1+c)Z}\right)Y. \tag{6}$$

This estimator is in $\mathcal{D}_c$ and, as mentioned, it is the optimal estimator in $\mathcal{D}_c$ in terms of minimizing risk.

Finally, we show that the estimator δ_1^c in (6) can be derived as a generalized Bayes estimator. Reparametrize the Poisson means as $\theta_i = \alpha_i\lambda, 1 \leq i \leq p$, and let $(\alpha_1,\ldots,\alpha_p)$ be Dirichlet distributed with parameters $(a_1,\ldots,a_p)$. Define $a_0 = \sum_{i=1}^p a_i$. Let λ have the improper prior that is flat on the positive real line, $\lambda \sim \pi(\lambda) \propto I(\lambda > 0)$.

Using that $Y \mid Z$ is multinomial the Poisson likelihood can be factorized as the product of a multinomial and the marginal of Z, this gives that the posterior distribution $\theta_1, \ldots, \theta_p$ is

$$\begin{aligned}\pi(\theta_1, \ldots, \theta_p \mid Y) &\propto P(Y = y \mid Z) P(Z = z)\, \text{Dirichlet}(a_1 \ldots, a_p)\, \pi(\lambda) \\ &= \frac{z!}{y_1! \cdots y_p!} \alpha_1^{a_1+y_1-1} \cdots \alpha_p^{a_p+y_p-1} \lambda^z e^{-\lambda}\, \pi(\lambda) \qquad (7) \\ &\propto \text{Dirichlet}(a_1 + y_1, \ldots, a_p + y_p)\, \mathcal{G}(z+1, 1)\, \pi(\lambda),\end{aligned}$$

which also shows that $(\alpha_1, \ldots, \alpha_p)$ and λ are independent. With this parametrization the Bayes solution under the c-Loss function is

$$\delta_j^B(Y) = \frac{1+c}{1 + cE[\lambda^{-1} \mid Y] \sum_{i=1}^p \{E[\theta_i^{-1} \mid Y]\}^{-1}} \{E[\theta_j^{-1} \mid Y]\}^{-1}. \qquad (8)$$

With respect to the posterior distribution in (7), the expectation $E[\theta_j^{-1} \mid Y]$ in this expression is given by

$$\begin{aligned}E[\theta_j^{-1} \mid Y] &= \int_0^\infty \int_S \frac{1}{\alpha_j \lambda}\, \pi(\theta_1, \ldots, \theta_p \mid Y)\, d\alpha\, d\lambda \\ &= \int_0^\infty \int_S \frac{1}{\alpha_j \lambda} \left\{ \frac{\Gamma(a_0 + z)}{\prod_{i=1}^p \Gamma(a_i + y_i)} \prod_{i=1}^p \alpha_i^{a_i+y_i-1} \right\} \mathcal{G}(z+1, 1) \pi(\lambda)\, d\alpha\, d\lambda \\ &= \int_0^1 \mathcal{G}(z+1, 1) \pi(\lambda)\, d\lambda \int_S \alpha_j \frac{\Gamma(a_0 + z)}{\prod_{i=1}^p \Gamma(a_i + y_i)} \prod_{i=1}^p \alpha_i^{a_i+y_i-1}\, d\alpha.\end{aligned}$$

Here the expectation of α_j over the simplex S is $E[\alpha_j \mid Y] = (a_0+z-1)/(a_j+y_j-1)$. Inserting this in the posterior expectation of θ_j^{-1} gives

$$E[\theta_j^{-1} \mid Y] = \frac{a_0 + z - 1}{a_j + y_j - 1} \int_0^\infty \frac{1}{\lambda} \mathcal{G}(z+1, 1) \pi(\lambda)\, d\lambda,$$

for $j = 1, \ldots, p$. Morover, we have that

$$E[\lambda^{-1} \mid Z] = \int_0^\infty \frac{1}{\lambda} \mathcal{G}(z+1, 1) I(\lambda > 0)\, d\lambda = \frac{1}{z},$$

which gives

$$E[\theta_j^{-1} \mid Y] = \frac{a_0 + z - 1}{a_j + y_j - 1} \frac{1}{z}.$$

In addition, the sum in (8) equals

$$\sum_{i=1}^p \{E[\theta_i^{-1} \mid Y]\}^{-1} = z \sum_{i=1}^p \frac{a_j + y_j - 1}{a_0 + z - 1} = \frac{(a_0 + z - p)z}{a_0 + z - 1}.$$

Now, let $\alpha_1, \ldots, \alpha_p$ be uniformly distributed over the simplex S. This is achieved by setting $a_1 = \cdots = a_p = 1$. Then the sum $a_0 = p$. In summary, with λ uniform

over $\mathbb{R}_+$ and the $(\alpha_1, \ldots, \alpha_p)$ uniform on the simplex $S = [0,1]^p$, the Bayes solution under the c-Loss function equals

$$\delta_j^B(Y) = \frac{1+c}{1+c\frac{1}{Z}\frac{Z^2}{p-1+Z}} \frac{Y_j}{p-1+Z} = \left(1 - \frac{p-1}{p-1+(1+c)Z}\right) Y_j = \delta_1^c(Y).$$

This means that in addition to being an empirical Bayes estimator, the new estimator δ_1^c is also a generalized Bayes estimator.

We have yet to find out whether the estimator δ_1^c in (6) is admissible. If we in the prior distribution $\pi_2(b) \propto b^{\alpha-1}(b+1)^{-(\alpha+\beta)}$ considered above set $\alpha = m-1$ and $\beta = 1$, the estimator in (5) is admissible (since it is proper Bayes) for all $m > 1$. For $m = 0$ the estimator is equal to δ_1^c, but it is then no longer proper Bayes. Thus, in a sense, we are "one unit" away from proving that the optimal estimator in terms of minimizing risk (under L_c) is admissible. See [6] for more details.

4 Conclusion

In this paper we have derived a class of minimax estimators that uniformly dominate the MLE under the c-Loss function in (2). Estimators in this class are also minimax and uniformly dominant relative to the MLE under the weighted squared error loss function L_1. Importantly, estimators in this class can be fine-tuned in order to achieve the desired amount of balancing between two conflicting desiderata: good total risk and good individual risk in estimating individual θ_i and the sum γ.

References

[1] Clevenson, M.L. and J.V. Zidek Simultaneous Estimation of the Means of Independent Poisson Laws *Journal of the American Statistical Association*, 70:698–705, 1975.

[2] Ghosh, M. and J.T. Hwang and K. Tsui Construction of Improved Estimators in Multiparameter Estimation for Discrete Exponential Families *The Annals of Statistics*, 11:351–367, 1983.

[3] Ghosh, M. and A.Parsian Bayes Minimax Estimation of Multiple Poisson Parameters *Journal of Multivariate Analysis*, 11:280–288, 1981.

[4] Peng, J. *Simultaneous Estimation of the Parameters of Independent Poisson Disitributions* Technical Report No. 48, Stanford University, Department of Statistics, 1975

[5] Robert, C.P. *The Bayesian Choice. From Decision-Theoretic Foundations to Computational Implementation.* Springer, 2007.

[6] Stoltenberg, E. Aa. *The c-Loss Function: Balancing Total and Individual Risk in the Simultaneous Estimation of Poisson Means* Master's thesis, University of Oslo, Department of Mathematics, 2015

Selection Consistency of Generalized Information Criterion for Sparse Logistic Model

Hubert Szymanowski[*1] **and Jan Mielniczuk**[12]

[1]*Institute of Computer Science Polish Academy of Sciences, Poland*
[2]*Warsaw University of Technology, Poland*

Abstract: We consider selection rule for small-n-large-P logistic regression which consists in choosing a subset of predictors minimizing Generalized Information Criterion over all subsets of variables of size not exceeding k. Consistency of such rule under weak conditions were established in [3]. This is a generalization of the results of [2] to much broader regression scenario which allows also for a more general criterion function than considered there and k depending on a sample size. We will discuss possibility of further weakening of the assumptions.

Keywords: consistency, information criterion, feature selection, sparse logistic regression

AMS subject classifications: 62J12, 62F07

Acknowledgements: Study was supported by research fellowship within "Information technologies: research and their interdisciplinary applications" (agreement number POKL.04.01.01-00-051/10-00)

References

[1] Chen J, and Chen Z. Extended Bayesian Information Criteria for model selection with large model spaces. *Biometrika*, 95:759-771, 2008.

[2] Chen J., and Chen Z. Extended BIC for small-n-large-p sparse GLM. *Statist Sinica*, 22:555-574, 2012.

[3] Mielniczuk J., and Szymanowski H. Selection consistency of Generalized Information Criterion for small-n-large-p sparse logistic model. In *Springer Proceedings in Mathematics & Statistics: Stochastic Models, Statistics and Their Applications*, Steland A., Rafajłowicz E., Szajowski K., Eds., Vol. 122, 111-119, Springer, 2015.

*Corresponding author: h.szymanowski@ipipan.waw.pl

k-Sample Tests for Multivariate Censored Data

Måns Thulin*

Department of Statistics, Uppsala University, Sweden

Abstract: The one-way MANOVA problem of testing whether the mean vectors of several populations differ is common in many fields of research. In medicine, the rapid advance of high-throughput technologies has led to an increased interest in multivariate sets of biomarkers in e.g. blood samples. Such biomarkers can be used to understand different diseases and how they are affected by treatments and covariates. Biomarker data is often left-censored because some measurements fall below the laboratory's detection limit. I will discuss how censoring affects multivariate two-sample and one-way MANOVA tests, in terms of size and power. Classical parametric tests are found to perform better than nonparametric alternatives, which means that the current recommendations for analysis of censored multivariate data have to be revised. The good performance of the classical parametric tests can at least partially be explained by some asymptotic results related to multivariate skewness and kurtosis. An expansion of the size of the classical tests shows that up to $o(n^{-1})$ the size is determined by skewness and kurtosis, and closer inspection of the censoring process reveals that moderate censoring only has a mild effect on these quantities. If the underlying distribution is approximately normal then moderate censoring will therefore not affect the size of the tests much.

Keywords: censored data, MANOVA, nondetects, two-sample test

AMS subject classifications: 62H15, 62F05, 62N03

*Corresponding author: mans.thulin@statistik.uu.se

Forecasting Extreme Events in Agricultural Commodity Markets

Athanasios Triantafyllou[1], **George Dotsis**[*1] **and Alexander H. Sarris**[1]

[1]*Department of Economics, University of Athens, Greece*

Abstract: In this paper we empirically examine the predictive power of the risk neutral option-implied distribution on sudden (extreme) price spikes in agricultural commodity markets. We use as predictors of extreme returns the skewness of the risk neutral distribution, the variance risk premium and the tail risk measure. We find that our option-implied risk measures are robust and statistically significant predictors of the extreme events in the agricultural commodity markets. Our option-implied risk measures forecast both the magnitude and the probability of occurrence of a crash in these markets

Keywords: risk neutral moments, tail risk measures, extreme value theory, agricultural commodities

1 Introduction

In this paper we empirically estimate the forecasting power of the option-implied risk measures when used as predictors of extreme events in maize, wheat and soybeans markets. With the term extreme event, we define what in the relevant literature is called upward price spikes. While in the equity market the extremely unlikely event is a sudden market crash, in agricultural markets the unlikely extreme event is defined as a sudden upward price spike, since sudden increases in commodity prices are most significant not only for commodity investors (at micro level), but also for food security and policy issues in a macroeconomic level. In simpler words, while the equity investors are fearful of the left tail of the distribution of returns, the commodity investors are fearful of the shape of right tail. Our option-implied measures are constructed using some results of Extreme Value Theory (EVT) and of risk neutral valuation. Our primary motivations come from the literature in equity markets, in which the option-implied tail loss measures add significant forecasting power when used as predictors of stock-market crashes (Bollerslev et.al [3], Hamidieh [6], Vilkov et.al [8]). To the best of our knowledge, this is the first paper which deals with the forecasting of extreme events in agricultural commodity markets using option-implied information. While the empirical works in the relevant literature (e.g. Morgan et.al [7]) use the moments and the tails of the physical distributions (the distribution of the realized returns in agricultural commodity markets) in order to forecast extreme upward spikes in these markets, we use the moments and the (right) tails of the risk neutral option-implied distribution instead. Our contribution in the field is twofold: firstly, our empirical findings

*Corresponding author: gdotsis@econ.uoa.gr

show that the option-implied information in agricultural markets is extremely useful, not only when forecasting the variance of agricultural prices (see Triantafyllou et.al [9], Wang, Fausti and Quasmi (2012)), but also when we forecast the sudden upward price movements. Our empirical findings implicitly reveal that the option-implied risk neutral distribution can (and must) be a risk management tool for agricultural commodity investors, since the risk management techniques are most needed and appreciated not in normal times, but in times of turbulence. By our forecasting regressions, we find that our option-implied risk measures (namely risk neutral skewness, variance risk premia and Tail Risk Measures) add statistically significant forecasting power when used as predictor of agricultural commodity returns. In addition, our option-implied risk measures contain all the forecasting information of the physical (realized) tail risk measure existed in the relevant literature (Morgan et.al[7]). Secondly, when we define the extreme event in agricultural markets as a 2 standard deviation rise (above the expected one) in monthly returns, our forecasting binary (probit) regressions show that our option-implied tail-risk measure captures and forecasts in a statistically significant manner these extreme price spikes and the probability of the occurrence of these.

2 Methodology

2.1 Tail risk measure

In order to compute the tail risk measure we apply some results-tools of Extreme Value Theory (EVT) on the risk neutral option-implied distribution. We apply the second theorem of EVT, known as the Pickands-Balkema-de Haan theorem, to describe the distribution of a commodity price X above an extreme (unusually high) threshold value h by a Generalized Pareto distribution of the form:

$$G_{\beta,\xi}(h-x) = \begin{cases} l - (l + \xi \frac{h-x}{b})^{-\frac{l}{\xi}} & \xi \neq 0 \\ l - \exp(-\frac{h-x}{b}) & \xi = 0. \end{cases} \tag{1}$$

The tail risk measure $TLR_{h,t}$ at time t given a specific pre-determined threshold h is the expected excess tail value given in equation (2) relative to the current value x_t, and it is the following formula:

$$E(h - x/h > x) = \frac{b}{l - x} \tag{2}$$

The tail risk measure $TLR_{h,t}$ at time t given a specific pre-determined threshold h is the expected excess tail value given in equation (2) relative to the current value x_t, and it is the following formula:

$$TLR_{h,t} = \frac{E_t(h - x/h > x)}{x_t} \tag{3}$$

Under the risk neutral probability measure Q, we assume that the corresponding risk neutral option-implied distribution belongs to the maximum domain of attrac-

tion of an extreme-value distribution H_ξ. Then, the tail loss measure for a given threshold is unique[1].

2.2 Model free option-implied moments

We compute the model-free version of option implied moments using the method of Bakshi et.al. [1].Under the risk-neutral probability measure Q, the analytical formulas for conditional risk neutral moments are given below:

$$VAR = E_t^q(R^2) - [E_t^q(R)]^2 \tag{4}$$

$$SKEW = \frac{E_t^q(R^3) - 3E_t^q(R)E_t^q(R^2) + [E_t^q(R)]^3}{VAR^{3/2}} \tag{5}$$

In accordance with Bakshi et.al. [1], we define the Quad and Cubic contracts[2] as follows:

$$Quad = \exp(-r(T-t))E_t^q(R^2) \tag{6}$$

$$Qubic = \exp(-r(T-t))E_t^q(R^3) \tag{7}$$

In the equations (6) and (7), r is the risk-free interest rate (3-month US-Treasury Bill), t is the trading date and T is the expiration date of a given contract and consequently $T-t$ defines time to maturity. If we substitute $Quad$ and $Cubic$ expressions given in equations (6) and (7) into equations (4) and (5), we get the model free version of option implied variance ($MFIV$) and implied skewness ($MFIS$) given below :

$$MFIV = \exp(-r(T-t))Quad - [E_t^q(R)]^2 \tag{8}$$

$$MFIS = \frac{\exp(-r(T-t))Cubic - 3E_t^q(R)\exp(-r(T-t))Quad + 2[E_t^q(R)]^3}{MFIV^{3/2}} \tag{9}$$

Furthermore, Bakshi et.al [1] show that under the risk-neutral pricing measure Q, the $Quad$ and $Cubic$ contracts can be expressed as continuous functions of out-of-the-money European calls $C(t,T,K)$ and out-of-the-money European puts $P(t,T,K)$ in the form given below:

$$Quad = \int_F^\infty \frac{2(1-\ln[\frac{K}{F}])}{K^2}C(t,T,K)dK + \int_0^F \frac{2(1+\ln[\frac{F}{K}])}{K^2}P(t,T,K)dK \tag{10}$$

$$Qubic = \int_F^\infty \frac{6\ln[\frac{K}{F}] - 3\ln[\frac{K}{F}]^2}{K^2}C(t,T,K)dK - \int_0^F \frac{6\ln[\frac{F}{K}] + 3\ln[\frac{F}{K}]^2}{K^2}P(t,T,K)dK \tag{11}$$

K is the strike price of the option contract, F is the price of the underlying futures contract, t is the trading date and T is the expiration date of the option contract.

[1] See Vilkov et.al [8] for analytical proof of this proposition

[2] If we define with R the logarithmic returns of the underlying asset with price S_t [$R = \ln((S_t + \frac{1}{\ln(S_t)})$], then a $Quad$ (or volatility) contract is a theoretical contract with risk neutral quadratic expected return-payoff $E_t^Q(R^2)$ and a $Cubic$ contract is a contract with risk neutral cubic expected return-payoff $E_t^Q(R^3)$. Bakshi et.al. [1] prove that quadratic and cubic expected risk neutral returns are continuous functions of Out of the Money (OTM) call and put option prices.

In addition, Bakshi et.al. [1] prove that the expected (conditional on information at time t) risk-neutral returns $E_t^q(R)$ can be approximated by the following expression:

$$E_t^q(R) = \exp(r(T-t)) - 1 - \frac{\exp(r(T-t))}{2} Quad - \frac{\exp(r(T-t))}{6} Qubic \quad (12)$$

Knowing the analytical forms of *Quad* and *Cubic* contracts from equations (6) and eqrefQubic, and the approximating quantity of conditional risk neutral expected returns $E_t^q(R)$ from equation (12), we can compute by using numerical integration the model free option-implied moments given in equations (8) and (9).

2.3 Variance risk premium

The variance risk premium represents the compensation demanded by investors for bearing variance risk and it is defined as the difference between realized variance and a risk-neutral model-free implied variance ($MFIV_t$). According to Bliss et.al [2] and Carr et.al [4] is a reliable measure of risk aversion in financial markets. More specifically, following Carr et.al [4] and Christoffersen et.al [5], we define the variance risk premium as the difference between the $P-$measure expected variance and the Qmeasure expected variance, using the following formula:

$$VRP(t,T) = E_t^P(RV(t,T)) - E_t^Q(RV(t,T)) \quad (13)$$

References

[1] Bakshi, G., Kapadia, N. & D. Madan (2003). Stock Return Characteristics, Skew Laws, and the Differential Pricing of Individual Equity Options, *Review of Financial Studies,15,*101-143.

[2] Bliss, R. R. & N. Panigirtzoglou (2004). Option-Implied Risk Aversion Estimates.*Journal of Finance,LIX,*407-446.

[3] Bollerslev, T. & G. Todorov (2011). Tails, Fears and Variance Risk Premia, *Journal of Finance,66,*2165-2211.

[4] Carr, P. & L. Wu (2009). Variance Risk Premiums. *Review of Financial Studies,22,*1311-1341.

[5] Christoffersen, P., Kang, S. B. & Pan, X. (2010). Does Variance Risk Premium Predicts Futures Returns? Evidence in the Crude Oil Market. *Working Paper*

[6] Hamidieh, K. (2011). Recovering the Tail Shape Parameter of the Risk Neutral Density from Option Prices. *Working paper*

[7] Morgan, W., Cotter, J. & K. Dowd (2012). Extreme Measures of Agricultural Financial Risk. *Journal of Agricultural Economics*, *63,*65-82.

[8] Vilkov, G. & Y. Xiao (2013). Option-Implied Information and Predictability of Extreme Returns. *SAFE Working Paper 5*

[9] Triantafyllou, A., Dotsis, G & A.H. Sarris (2015). Volatility Forecasting and Time-Varying Variance Risk Premia in Grains Commodity Markets. *Journal of Agricultural Economics*, *(forthcoming)*

Overview of Some Interesting Statistical Problems in Biochemical Analysis of Glycans

Ivo Ugrina*

University of Zagreb, Croatia

Abstract: Glycomics is rapidly emerging field in high-throughput biology that aims to systematically study glycan structures of a given protein, cell type or organic system. As within other high-throughput methods in biology (microarrays, metabolomics, proteomics), accuracy of high-throughput methods is highly affected by complicated experimental procedures leading to differences between replicates and the existence of batch effects, among others. Study of appropriate methods for normalization, appropriate designs of experiments and batch removal methods tailored to the needs of glycomics is therefore a necessity.

Keywords: batch effects, normalization, noise, glycomics
AMS subject classifications: 62P10, 92C40

1 Introduction

Glycans are important structural and functional components of the majority of proteins. However, their structural complexity and the absence of a direct genetic template impedes understanding of the role of glycans in biological processes. A recent comprehensive report endorsed by the US National Academies concluded that glycans are directly involved in the pathophysiology of every major disease and that additional knowledge from glycoscience will be needed to realize the goals of personalized medicine [11].

This conclusion gave importance to the already existing field of biochemical analysis of glycans. Currently, numerous studies in the development of biochemical analysis of glycans are conducted. Successful implementation of high-throughput analytical techniques for glycan analysis resulted in publication of GWAS of the human glycome [9, 7, 8, 6]. A number of studies have investigated the role of glycans in human disease, including autoimmune diseases and cancer [2, 10].

Biochemical analysis of glycans is usually conducted with one of the following methods: UPLC-FLR, ultraperformance liquid chromatography with fluorescence detection; CGE-LIF, multiplex capillary gel electrophoresis with laser induced fluorescence detection; MALDI-TOF-MS, matrix assisted laser desorption/ionization time of flight MS; LC-ESI-MS, liquid chromatography electrospray MS.

Complexity of these methods, together with the complexity of glycans usually leads to different effects like batch effects or multiplicative errors making subsequent analysis of results more cumbersome. Therefore, the need for greater expertise in statistical and computer science methods becomes obvious.

*Corresponding author: ivo@iugrina.com

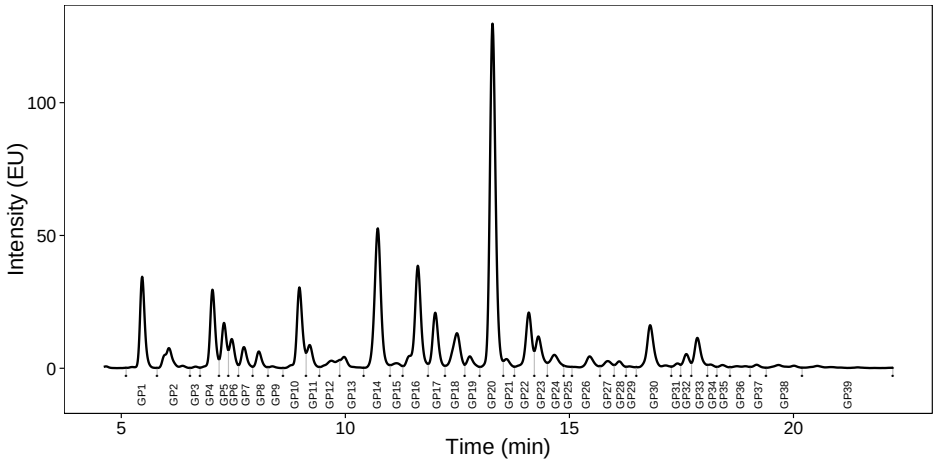


Figure 1: An example of a chromatogram from the UPLC analysis of glycan composition from the blood plasma sample.

2 Problems

Especially problematic is the multiplicative error as a consequence of laboratory conditions and the current practice of removing that error. An example of a result from the biochemical analysis of glycans with UPLC method from blood plasma is given in Figure 1. This is an example of a *chromatogram*. If one denotes the areas under the chromatogram between appropriate borders with GP_* the graph can be represented as a random vector $GP = (GP_1, \ldots, GP_{39})$. Multiplicative error then means that for every run of the biochemical analysis the results will be of the form $\overline{GP} = C \cdot GP$ where C denotes a random variable. Therefore, every run will give different intensities.

Current methodology within the filed of glycomics approaches the problem of the multiplicative error by the usage of Total Area Normalization. Normalization refers to the creation of shifted and scaled versions of statistics, where the intention is that these normalized values allow the comparison of corresponding normalized values for different datasets. Total Area Normalization (TAN), also called percentage normalization, is given by the transformation

$$GP_i^{TAN} = \frac{GP_i}{\sum_j GP_j} \ .$$

Although this normalization procedure works fine for the basic quality control it introduces many problems like spurious correlations (making network/pathway analysis problematic) or other problems with the constrained data (see [12, 1] for more details).

Another big problem in the analysis of glycans is the existence of batch effects similarly to the fields of metabolomics or microarray experiments. Conducting a glycan analysis on a big population means that the analysis will almost always be

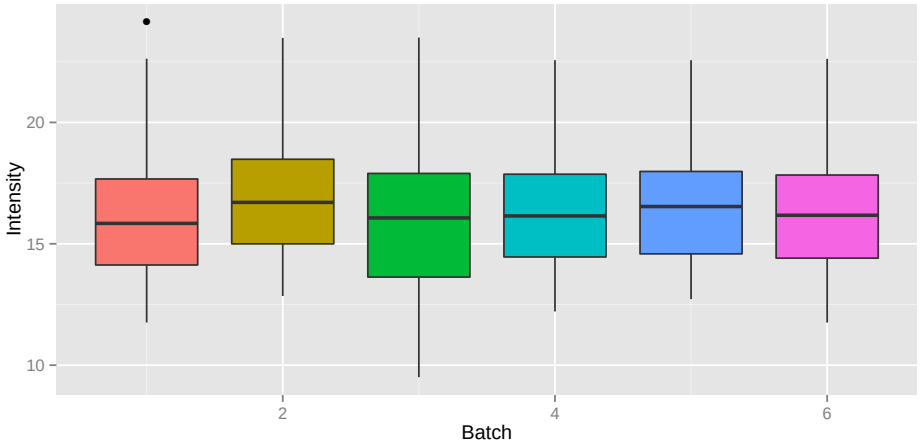


Figure 2: An example of batch effects (differences in medians and spread) within the UPLC analysis of blood plasma samples presented with boxplots.

conducted in batches of 100 to 1000 samples. As the time between the analysis of batches increases the probability of the introduction of some random effect increases also. This could be due to the changes on the machine where the analysis is conducted or by the change of a lab analyst.

Unfortunately, the current practice in glycomics research often ignores the problem of batch effects and therefore increases the probability of false results in latter statistical analysis.

An example of batch effects on an experiment designed to infer the experimental variability is given in Figure 2. Since all batches consist of the same (replicated) samples the results should behave the same. However, the batches were prepared by different lab analysts at different time points introducing batch effects non-intrinsic to the underlying samples.

Appropriate design of experiments is also sometimes ignored by the current practice in the field of biochemical analysis of glycans. This is closely connected with the aforementioned existence of batch effects since the prerequisite for batch correction procedures is an effective design of the experiment.

3 Improvements

The current practice can be improved by the introduction of appropriate methods for the design of experiments like randomized block designs [3], exploration of the effects of other normalization procedures like quantile normalization or median normalization [4] and the introduction of batch correction techniques like linear mixed effect models and empirical Bayes methods [5].

Acknowledgements: This research was supported by EU projects: MIMOmics (contract #305280) and IntegraLife (contract #315997).

References

[1] Wiley: Compositional Data Analysis: Theory and Applications - Vera Pawlowsky-Glahn, Antonella Buccianti.

[2] B. Adamczyk, T. Tharmalingam, and P. M. Rudd. Glycans as cancer biomarkers. *Biochimica Et Biophysica Acta*, 1820(9):1347–1353, Sept. 2012.

[3] S. Addelman. The Generalized Randomized Block Design. *The American Statistician*, 23(4):35–36, Oct. 1969.

[4] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics (Oxford, England)*, 19(2):185–193, Jan. 2003.

[5] C. Chen, K. Grennan, J. Badner, D. Zhang, E. Gershon, L. Jin, and C. Liu. Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods. *PLoS ONE*, 6(2):e17238, Feb. 2011.

[6] J. E. Huffman, A. Knezevic, V. Vitart, J. Kattla, B. Adamczyk, M. Novokmet, W. Igl, M. Pucic, L. Zgaga, . Johannson, I. Redzic, O. Gornik, T. Zemunik, O. Polasek, I. Kolcic, M. Pehlic, C. A. M. Koeleman, S. Campbell, S. H. Wild, N. D. Hastie, H. Campbell, U. Gyllensten, M. Wuhrer, J. F. Wilson, C. Hayward, I. Rudan, P. M. Rudd, A. F. Wright, and G. Lauc. Polymorphisms in B3gat1, SLC9a9 and MGAT5 are associated with variation within the human plasma N-glycome of 3533 European adults. *Human Molecular Genetics*, 20(24):5000–5011, Dec. 2011.

[7] Z. Kutalik, B. Benyamin, S. Bergmann, V. Mooser, G. Waeber, G. W. Montgomery, N. G. Martin, P. A. F. Madden, A. C. Heath, J. S. Beckmann, P. Vollenweider, P. Marques-Vidal, and J. B. Whitfield. Genome-wide association study identifies two loci strongly affecting transferrin glycosylation. *Human Molecular Genetics*, 20(18):3710–3717, Sept. 2011.

[8] G. Lauc, A. Essafi, J. E. Huffman, C. Hayward, A. Kneevi, J. J. Kattla, O. Polaek, O. Gornik, V. Vitart, J. L. Abrahams, M. Pui, M. Novokmet, I. Redi, S. Campbell, S. H. Wild, F. Boroveki, W. Wang, I. Koli, L. Zgaga, U. Gyllensten, J. F. Wilson, A. F. Wright, N. D. Hastie, H. Campbell, P. M. Rudd, and I. Rudan. Genomics Meets GlycomicsThe First GWAS Study of Human N-Glycome Identifies HNF1 as a Master Regulator of Plasma Protein Fucosylation. *PLoS Genet*, 6(12):e1001256, Dec. 2010.

[9] G. Lauc, J. E. Huffman, M. Pui, L. Zgaga, B. Adamczyk, A. Muini, M. Novokmet, O. Polaek, O. Gornik, J. Kriti, T. Keser, V. Vitart, B. Scheijen, H.-W. Uh, M. Molokhia, A. L. Patrick, P. McKeigue, I. Koli, I. K. Luki, O. Swann, F. N. van Leeuwen, L. R. Ruhaak, J. J. Houwing-Duistermaat, P. E. Slagboom, M. Beekman, A. J. M. de Craen, A. M. Deelder, Q. Zeng, W. Wang, N. D. Hastie, U. Gyllensten, J. F. Wilson, M. Wuhrer, A. F. Wright, P. M. Rudd, C. Hayward, Y. Aulchenko, H. Campbell, and I. Rudan. Loci Associated with

N-Glycosylation of Human Immunoglobulin G Show Pleiotropy with Autoimmune Diseases and Haematological Cancers. *PLoS Genet*, 9(1):e1003225, Jan. 2013.

[10] Y. Mechref, Y. Hu, A. Garcia, and A. Hussein. Identifying cancer biomarkers by mass spectrometry-based glycomics. *Electrophoresis*, 33(12):1755–1767, July 2012.

[11] National Research Council (US) Committee on Assessing the Importance and Impact of Glycomics and Glycosciences. *Transforming Glycoscience: A Roadmap for the Future*. The National Academies Collection: Reports funded by National Institutes of Health. National Academies Press (US), Washington (DC), 2012.

[12] R. Pincus. Aitchison, J.: The Statistical Analysis of Compositional Data. Chapman and Hall, London - New York 1986, XII, 416 pp., 25,00. *Biometrical Journal*, 30(7):794–794, Jan. 1988.

The Horseshoe and More General Sparsity Priors

Stéphanie L. van der Pas*

Leiden University, Netherlands

Abstract: We study the sparse normal means problem, where the mean vector is sparse in the nearly black sense. Adopting the frequentist framework where the data is generated according to some fixed mean vector, we use the posterior mean resulting from the horseshoe prior to recover the mean vector. We show that the posterior rate of convergence is at most of the order of the minimax rate. The horseshoe prior is not unique in this regard, as some recent extensions of the work on the horseshoe prior show.

Keywords: horseshoe, sparsity, Bayesian inference, normal means problem
AMS subject classifications: 62F15

1 Introduction

A common test case for sparsity methods is the sparse normal means problem, where we observe a vector $Y^n \in \mathbb{R}^n$, $Y^n = (Y_1, \ldots, Y_n)$, such that

$$Y_i = \theta_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

for independent standard normal random variables ε_i. The vector $\theta = (\theta_1, \ldots, \theta_n)$ is the vector of interest, and is assumed to be sparse in the *nearly black* sense, meaning that the number of nonzero entries of θ,

$$p_n := \#\{i \ : \ \theta_i \neq 0\}$$

is $o(n)$ as $n \to \infty$. Our goal is to *recover* θ, and to provide *uncertainty quantification*.

We focus in this paper on Bayesian methods. Thus, we use the posterior distribution to achieve both of our goals. The typical choice for our goal of recovery is to use a measure of centre, such as a median, mean, or mode as an estimator. For uncertainty quantification, a credible set is a natural object to use from a Bayesian point of view. To achieve realistic uncertainty quantification, the posterior should contract to its center at the same rate at which the estimator approaches the true parameter.

A well-studied approach has been to induce sparsity through a *spike and slab* prior [6], which is a mixture of a Dirac measure at zero (to account for the zero means) and a continuous distribution (to account for the nonzero means). An empirical Bayes version of this approach, where the mixing weight is obtained

*Corresponding author: svdpas@math.leidenuniv.nl

through marginal maximum likelihood, was found to work well for recovery [5], but no measure of uncertainty quantification was studied. For a fully Bayesian version, several combinations of priors on the mixing weight and on the nonzero coefficients (the 'slab') were found such that both goals are achieved [2]. Unfortunately, a spike and slab approach runs into computational difficulties, as it may require exploration of a model space of size 2^n.

Therefore, there is a need for priors that are not only suitable for recovery and uncertainty quantification, but are also feasible computationally on large data sets. The horseshoe prior is one such prior. In this paper, we review the theoretical results for the horseshoe prior, as described in [7]. The work has recently been extended in [4], and in forthcoming joint work with J. Schmidt-Hieber and J.-B. Salomond, which will be discussed during the presentation, and briefly in Section 4. We first review the horseshoe prior in Section 2, then discuss some posterior contraction results from [7] for the horseshoe prior in Section 3.

2 The horseshoe prior

The *horseshoe prior* was introduced by [1] and has the following hierarchical formulation:

$$\theta_i \mid \lambda_i, \tau \sim \mathcal{N}(0, \tau^2\lambda_i^2), \quad \lambda_i \sim C^+(0,1),$$

for $i = 1, \ldots, n$, where $C^+(0,1)$ is a standard half-Cauchy distribution. We use the coordinatewise posterior mean $T_\tau(y_i)$ as our estimator of θ_i. Figure 1 shows the prior density on each θ_i, and the posterior mean as a function of the observation y_i. It illustrates the role of the global parameter τ: decreasing τ leads to more mass near zero in the prior on θ_i and a stronger shrinkage effect in the posterior mean $T_\tau(y)$.

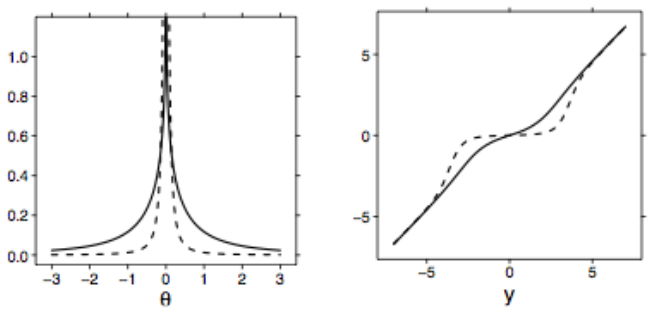


Figure 1: (Based on Figure 1 in [7]) The effect of decreasing τ on the prior on θ (left) and the posterior mean $T_\tau(y)$ (right). The solid line corresponds to $\tau = 1$, the dashed line to $\tau = 0.05$. Decreasing τ results in a higher prior probability of shrinking the observations towards zero.

The parameter τ controls the sparsity, and results from [7] show that the optimal choice for τ is the proportion of nonzero means p_n/n (up to a log factor). The role

of the parameters λ_i is to counteract the shrinkage effect at a local level.

The horseshoe estimator offers a computational advantage over the spike-and-slab approach, as it can be expressed as follows:

$$T_\tau(y_i) = y_i \frac{\int_0^1 z^{1/2} \frac{1}{\tau^2+(1-\tau^2)z} e^{y_i^2/2} dz}{\int_0^1 z^{-1/2} \frac{1}{\tau^2+(1-\tau^2)z} e^{y_i^2/2} dz}.$$

In case τ is estimated empirically, the posterior mean can be computed by plugging this estimate into the above expression, circumventing the need to use MCMC. It can be evaluated via a quadrature routine, or by a representation in terms of confluent hypergeometric functions, as discussed in [8].

3 Posterior contraction results for the horseshoe

In this section, some posterior contraction results from [7] are discussed. The main result is Theorem 33, which provides upper bounds on the rate of contraction of the posterior distribution around the true mean vector, and around the posterior mean. Denote the class of nearly black vectors by $\ell_0[p_n] = \{\theta \in \mathbb{R}^n \ : \ \#\{i \ : \ \theta_i \neq 0\} \leq p_n\}$.

Theorem 33 (Theorem 3.3 in [7]). *Suppose* $Y^n \sim \mathcal{N}(\theta_0, I_n)$, $n, p_n \to \infty$, $p_n = o(n)$ *and* $\tau = (p_n/n)^\alpha$, $\alpha \geq 1$. *Then:*

$$\sup_{\theta_0 \in \ell_0[p_n]} \mathbb{E}_{\theta_0} \Pi_\tau \left(\theta \ : \ \|\theta - \theta_0\|^2 > M_n p_n \log \frac{n}{p_n} \ \middle| Y^n \right) \to 0,$$

and

$$\sup_{\theta_0 \in \ell_0[p_n]} \mathbb{E}_{\theta_0} \Pi_\tau \left(\theta \ : \ \|\theta - T_\tau(Y^n)\|^2 > M_n p_n \log \frac{n}{p_n} \ \middle| Y^n \right) \to 0,$$

for any $M_n \to \infty$.

These upper bounds are equal, up to a multiplicative constant, to the minimax risk [3]. The contraction rate around the true mean vector θ_0 is therefore sharp, but this is not necessarily the case for the rate of contraction around the posterior mean $T_\tau(Y^n)$.

Further investigation into the role of τ (Theorems 3.4 and 3.5 of [7]) shows that if $\tau = (p_n/n)^\alpha$ for $\alpha \in (0, 1)$, the posterior variance may exceed the minimax rate, indicating suboptimal spread of the posterior. If $\alpha > 1$, there exists a sequence $\theta_{0,n} \in \ell_0[p_n]$ for which the mean square error and the posterior variance are of different orders. If $\alpha = 1$, the posterior variance and ℓ_2 risk only differ by a factor $\sqrt{\log(n/p_n)}$, and the gap can even be closed by taking $\tau = (p_n/n)\sqrt{\log(n/p_n)}$.

Thus, for our goals of recovery and uncertainty quantification, the best choice for τ is $\tau = (p_n/n)\sqrt{\log(n/p_n)}$, as in that case both the worst case ℓ_2 risk and the posterior variance are at the order of the minimax risk (Theorems 3.1, 3.2 and 3.4 in [7]).

In practice, the number of nonzero means p_n is typically unknown, and hence the value $\tau = (p_n/n)\sqrt{\log(n/p_n)}$ cannot be used. However, an empirical Bayes

procedure, in which τ is estimated from the data, can still yield the (near) minimax rate for the worst case risk. For example, consider for any $c_1 > 2, c_2 > 1$, the estimator

$$\widehat{\tau} = \max\left\{\frac{\#\{|y_i| \geq \sqrt{c_1 \log n},\ i = 1, \ldots, n}{c_2 n}, \frac{1}{n}\right\}.$$

This estimator estimates the number of nonzero means by counting those observations that are past the *universal threshold* $\sqrt{2 \log n}$. It is bounded below by $1/n$ for computational reasons, and because it corresponds to the assumption that there is at least one nonzero mean. By Theorem 4.1 combined with Lemma A.7 from [7], the horseshoe estimator combined with this estimator of τ will have a worst case risk of order $p_n \log n$, which is close to the minimax rate unless the truth is not very sparse.

4 Other sparsity priors

In [4], priors of the form

$$\theta_i \mid \lambda_i^2, \tau^2 \sim \mathcal{N}(0, \lambda_i^2 \tau^2), \quad \lambda_i^2 \sim \pi(\lambda_i^2), \quad i = 1, \ldots, n$$

are considered, for priors π with density given by

$$\pi(\lambda_i^2) = K \frac{1}{(\lambda_i^2)^{a+1}} L(\lambda_i^2),$$

where $K > 0$ is a constant and $L : (0, \infty) \to (0, \infty)$ is a non-constant, *slowly varying* function, where 'slowly varying' means that there exist $c_0, M \in (0, \infty)$ such that $L(t) > c_0$) for all $t \geq t_0$ and $\sup_{t \in (0,\infty)} L(t) \leq M$. The horseshoe is contained in this class of priors, by taking $a = 1/2$, $L(t) = t/(1 + t)$ and $K = 1/\pi$. The authors prove a posterior contraction theorem of the same type as Theorem 33 for this class of priors, provided $\alpha \in [1/2, 1]$. Although their results for lower bounds on the posterior variance are limited to the case $a = 1/2$, their results extend those for the horseshoe, showing that the horseshoe is not unique in its desirable posterior concentration properties. Forthcoming work with J. Schmidt-Hieber and J.-B. Salomond provides some more general conditions on sparsity priors such that the posterior concentrates at least as fast as the minimax rate.

Acknowledgements: Research supported by Netherlands Organization for Scientific Research NWO. This paper is based on joint work with Aad van der Vaart, Bas Kleijn, Johannes Schmidt-Hieber and Jean-Bernard Salomond.

References

[1] C. M. Carvalho, N. G. Polson and J. G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465-480, 2010.

[2] I. Castillo and A. W. van der Vaart. Needles and straw in a haystack: posterior concentration for possibly sparse sequences. *The Annals of Statistics*, 40(4):2069–2101, 2012.

[3] D. L. Donoho, I. M. Johnstone, J. C. Hoch and A. S. Stern. Maximum entropy and the nearly black object (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, 54:41–81, 1992.

[4] P. Ghosh and A. Chakrabarti. Posterior concentration properties of a general class of shrinkage estimators around nearly black vectors. arXiv:1412.8161, 2013.

[5] I. M. Johnstone and B. W. Silverman. Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *The Annals of Statistics*, 32(4):1594–169, 2004.

[6] T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032, 1988.

[7] S. L. van der Pas, B. J. K. Kleijn and A. W. van der Vaart. The horseshoe estimator: posterior concentration around nearly black vectors. *Electronic Journal of Statistics*, 8:2585–2618, 2014.

[8] N. G. Polson and J. G. Scott. Good, great, or lucky? Screening for firms with sustained superior performance using heavy-tailed priors. *The Annals of Applied Statistics*, 6:161–185, 2012.

Goodness-Of-Fit Tests for Exponentiality Based on Yanev-Chakraborty Characterization and Their Efficiencies

Ksenia Volkova*

Saint-Petersburg State University, Russia

Abstract: Two scale-free goodness-of-fit tests for exponentiality based on the recent characterization of exponential law by Yanev and Chakraborty are proposed. Test statistics are functionals of U-empirical processes. The first of these statistics is of integral type, it is similar to the classical statistics ω_n^1. The second one is a Kolmogorov type statistic. The limiting distribution and large deviations asymptotic of new statistics under null hypothesis are described. Their local Bahadur efficiency for parametric alternatives is calculated. The Kolmogorov type statistic is not asymptotically normal, therefore we evaluate the critical values by using Monte-Carlo methods. For small sample size efficiencies are compared with simulated powers of new tests. Also conditions of local asymptotic optimality of new statistics in the sense of Bahadur are discussed and examples of such special alternatives are given.

Keywords: testing of exponentiality, order statistics, U-statistics, Bahadur efficiency

AMS subject classifications: 60F10, 62G10, 62G20, 62G30

1 Introduction

We develop goodness-of-fit tests for exponentiality exploiting a characterization based on property of order statistics. The problem formulation is as follows: let $X_1, X_2, \ldots, X_n$ be i.i.d. observations having the continuous df F. Consider testing of composite hypothesis of exponentiality $H_0 : F \in \mathcal{E}(\lambda)$, where $\mathcal{E}(\lambda)$ denotes the class of exponential distributions with the density $f(x) = \lambda e^{-\lambda x}, x \geq 0$, where $\lambda > 0$ is some unknown parameter.

Suppose that the df F belongs to the class of distributions $\mathcal{F}$, if the corresponding density f has derivatives of all orders in the neighbourhood of zero.

Arnold and Villasenor in [1] conjectured, and Yanev and Chakraborty in [8] proved that the following characterized the exponential law within the class $\mathcal{F}$:

Let $X_1, \ldots, X_n$ be non-negative i.i.d. rv's with df F from class $\mathcal{F}$. Then the statistics $\max(X_1, X_2, X_3)$ *and* $\max(X_1, X_2) + \frac{X_3}{3}$ *are identically distributed if and only if the df F is exponential.*

*Corresponding author: efrksenia@gmail.com

Consider the usual empirical df $F_n(t) = n^{-1}\sum_{i=1}^{n} \mathbf{1}\{X_i < t\}, t \in R^1$, based on the observations $X_1, \ldots, X_n$. According to the characterization we construct for $t \geq 0$ the U-empirical df's by the formulae

$$H_n(t) = \binom{n}{3}^{-1} \sum_{1\leq i_1<i_2<i_3\leq n} \mathbf{1}\{\max(X_{i_1}, X_{i_2}, X_{i_3}) < t\}, \quad t \geq 0,$$

$$G_n(t) = \frac{1}{3}\binom{n}{3}^{-1} \sum_{1\leq i_1<i_2<i_3\leq n} [\mathbf{1}\{\max(X_{i_1}, X_{i_2}) + \frac{X_{i_3}}{3} < t\} +$$

$$+ \mathbf{1}\{\max(X_{i_2}, X_{i_3}) + \frac{X_{i_1}}{3} < t\} + \mathbf{1}\{\max(X_{i_3}, X_{i_1}) + \frac{X_{i_2}}{3} < t\}], \quad t \geq 0.$$

It is known that the properties of U-empirical df's are similar to the properties of usual empirical df's, see [2]. Hence for large n the df's H_n and G_n should be close under H_0, and we can measure their closeness by using some test statistics.

We suggest two scale-invariant statistics

$$I_n = \int_0^\infty (H_n(t) - G_n(t))\, dF_n(t), \tag{1}$$

$$D_n = \sup_{t\geq 0} \mid H_n(t) - G_n(t) \mid, \tag{2}$$

assuming that their large values are critical.

2 Integral statistic I_n

Without loss of generalization we can assume that $\lambda = 1$. The statistic I_n is asymptotically equivalent to the U-statistic of degree 4 with the centered kernel $\Psi(X_1, X_2, X_3, X_4)$ given by

$$\Psi(X_1, X_2, X_3, X_4) = \frac{1}{4} \sum_{\pi(i_1,\ldots,i_4)} \mathbf{1}\{\max(X_{i_1}, X_{i_2}, X_{i_3}) < X_{i_4}\} -$$

$$-\frac{1}{24} \sum_{\pi(i_1,\ldots,i_4)} \mathbf{1}\{\max(X_{i_1}, X_{i_2}) + \frac{X_{i_3}}{3} < X_{i_4}\},$$

where $\pi(i_1, \ldots, i_4)$ means all permutations of different indices from $\{i_1, \ldots, i_4\}$.

Theorem 34. *Under null hypothesis as $n \to \infty$ the statistic I_n is asymptotically normal with asymptotic variance given by*

$$\sqrt{n} I_n \xrightarrow{d} \mathcal{N}(0, \frac{23}{10920}).$$

2.1 Large deviations of the statistic I_n

The kernel Ψ is centered, bounded and non-degenerate. Hence according to the theorem of large deviations for such statistics from [5], we obtain the following result.

Theorem 35. *For* $a > 0$

$$\lim_{n\to\infty} n^{-1} \ln P(I_n > a) \sim \frac{5460}{23} a^2, \; as \, a \to 0.$$

3 Kolmogorov-type statistic D_n

Now we consider the Kolmogorov type statistic (2). Its indisputable merit is consistency against any alternative that follows directly from the characterization as such, while the integral statistic I_n is not always consistent.

In our case for fixed $t \geq 0$ the difference $H_n(t) - G_n(t)$ is a family of U-statistics with the kernels, depending on $t \geq 0$:

$$\Xi(X, Y, Z; t) = \mathbf{1}\{\max(X, Y, Z) < t\} - \frac{1}{3}\mathbf{1}\{\max(X, Y) + \frac{Z}{3} < t\} -$$
$$-\frac{1}{3}\mathbf{1}\{\max(Y, Z) + \frac{X}{3} < t\} - \frac{1}{3}\mathbf{1}\{\max(X, Z) + \frac{Y}{3} < t\}.$$

Limiting distribution of the statistic D_n is unknown. Using the methods of [6], one can show that the U-empirical process

$$\eta_n(t) = \sqrt{n}\,(H_n(t) - G_n(t))\,, \; t \geq 0,$$

weakly converges in $D(0, \infty)$ as $n \to \infty$ to certain centered Gaussian process $\eta(t)$ with calculable covariance. Then the sequence of statistics $\sqrt{n} D_n$ converges in distribution to the rv $\sup_{t\geq 0} |\eta(t)|$ but it is impossible to find explicitly its distribution. Hence it is reasonable to determine the critical values for statistics D_n by simulation.

3.1 Large deviations of the statistic D_n

The family of kernels $\{\Xi(X, Y, Z; t)\}, t \geq 0$ is not only centered but bounded. Using the results from [4] on large deviations for the supremum families of non-degenerate U-statistics, we obtain the following result.

Theorem 36. *For* $a > 0$

$$\lim_{n\to\infty} n^{-1} \ln P(D_n > a) \sim 4.966 a^2, \; as \; a \to 0.$$

4 Local Bahadur efficiencies of statistics I_n and D_n

We present the following alternatives against exponentiality which will be considered for all tests when $x \geq 0$:

i) Makeham distribution with the density $g_1(x, \theta) = \big(1 + \theta(1 - e^{-x})\big) \exp\big(- x - \theta(e^{-x} - 1 + x)\big), \theta > 0$;

ii) Weibull distribution with the density $g_2(x, \theta) = (1 + \theta) x^{\theta} \exp(-x^{1+\theta}), \theta > 0$;

iii) gamma distribution with the density $g_3(x,\theta) = \frac{x^\theta}{\Gamma(\theta+1)} e^{-x}, \theta > 0$;

iv) exponential mixture with negative weights (EMNW(β)) (see [3]) $g_4(x) = (1+\theta)e^{-x} - \theta\beta e^{-\beta x}, \theta \in \left[0, \frac{1}{\beta-1}\right], \beta > 1$.

We calculate the efficiencies of both statistics against common alternatives from the class $\mathcal{F}$. The statistic D_n has the non-normal limiting distribution, hence we use the Bahadur approach as a method of calculation of asymptotic efficiency, while the classical Pitman approach to efficiency is not applicable. However, it is known that the local Bahadur efficiency and the limiting Pitman efficiency usually coincide, see [7]. Finally, we analyze the conditions of local asymptotic optimality of our tests and describe the "most favorable" alternatives for them.

We supplement our research with simulated powers which principally support the theoretical values of efficiency.

Acknowledgements: The research of the author was supported by the grants RFBR 13-01-00172, NSh. 2504.2014.1, and SPbGU 6.38.672.2013.

References

[1] B. C. Arnold, J. A. Villasenor. Exponential characterizations motivated by the structure of order statistics in samples of size two. *Statistical and Probability Letters*, 83:596–601, 2013.

[2] R. Helmers, P. Janssen, R. Serfling. Glivenko-Cantelli properties of some generalized empirical DF's and strong convergence of generalized L-statistics. *Probability Theory and Related Fields*, 79:75–93, 1988.

[3] V. Jevremović. A note on mixed exponential distribution with negative weights. *Statistical and Probability Letters*, 11(3):259–265, 1991. doi: 10.1016/0167-7152(91)90153-I.

[4] Ya. Yu. Nikitin. Large deviations of U-empirical Kolmogorov-Smirnov tests, and their efficiency. *Journal of Nonparametric Statistics*, 22:649–668, 2010.

[5] Ya. Yu. Nikitin, E. V. Ponikarov. Rough large deviation asymptotics of Chernoff type for von Mises functionals and U-statistics. *Proceedings of the St. Petersburg Mathematical Society*, 7:124–167, 1999. English translation in AMS Translations, 2(203):107–146, 2001.

[6] B. W. Silverman. Convergence of a class of empirical distribution functions of dependent random variables. *The Annals of Probability*, 11:745–751, 1983.

[7] H. S. Wieand. A condition under which the Pitman and Bahadur approaches to efficiency coincide. *The Annals of Mathematical Statistics*, 4:1003–1011, 1976.

[8] G. P. Yanev, S. Chakraborty. Characterizations of exponential distribution based on sample of size three. *Pliska Studia Mathematica Bulgarica*, 23:237-244, 2013.

Simultaneous Perturbation Gradient Approximation Based Metropolis Adjusted Langevin Markov Chain Monte Carlo for Inference of Ordinary Differential Equations

Ivan Vujačić[*1] **and Mathisca de Gunst**[1]

[1] *VU University Amsterdam, Netherlands*

Abstract: The problem of parameter estimation for models defined by a system of ordinary differential equations (ODEs) is considered. The most efficient way to explore the parameter space is by using derivative information. Usual approaches for obtaining the gradient in ODEs setting like solving sensitivity equations and using finite difference formulas are computationally costly and not scalable to large scale systems. In this paper we use simultaneous perturbation gradient approximation (SPGA), originally proposed in stochastic optimization literature, as a substitute for the gradient in Metropolis adjusted Langevin algorithm (MALA). The obtained algorithm, called Simultaneous Perturbation Gradient Approximation based Metropolis Adjusted Langevin Markov chain Monte Carlo (SPGA MALA), requires at most three integration of the ODE system per MCMC step, regardless of the dimension of the system. This fixed computational costs makes SPGA MALA applicable to large scale systems. On the other hand, its efficiency is comparable to that of MALA. We demonstrate the performance of via simulations.

Keywords: Metropolis adjusted Langevin Markov Chain Monte Carlo methods, simultaneous perturbation gradient approximation, ordinary differential equations, parameter estimation

AMS subject classifications: 60J22, 65C40, 62F15

1 Introduction

Systems of ordinary differential equations (ODEs) are widely used in science and engineering for the mathematical modelling of various dynamic processes. We consider the system of the form

$$\begin{cases} \boldsymbol{x}'(t) = \boldsymbol{f}(\boldsymbol{x}(t), t; \boldsymbol{\theta}), \ t \in [0, T], \\ \boldsymbol{x}(0) = \boldsymbol{\xi}, \end{cases} \tag{1}$$

where $\boldsymbol{x}(t) = (x_1(t), \ldots, x_d(t))^\top \in \mathbb{R}^d$ is a state vector, $\boldsymbol{\xi}$ in $\Xi \subset \mathbb{R}^d$ is the initial condition, $\boldsymbol{\theta}$ in $\Theta \subset \mathbb{R}^p$ is a parameter and $\boldsymbol{f}$ is a known function. Given the values of $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$, we denote the solution of (1) by $\boldsymbol{x}(t; \boldsymbol{\theta}, \boldsymbol{\xi})$. Let us assume that a process is modelled by the system (1) with $\boldsymbol{\xi}_0$ known and $\boldsymbol{\theta}_0$ unknown. For

*Corresponding author: i.vujacic@vu.nl

simplicity, assume that we have noisy observations $y_i(t_j)$, $j = 1, \ldots, n$ of all the states $x_i(t; \boldsymbol{\theta}_0, \boldsymbol{\xi}_0)$, $i = 1, \ldots, d$ at time points $t_j \in [0, T]$, $j = 1, \ldots, n$:

$$y_i(t_j) = x_i(t_j; \boldsymbol{\theta}_0, \boldsymbol{\xi}_0) + \varepsilon_i(t_j), \quad i = 1, \ldots, d; j = 1, \ldots, n,$$

where $\varepsilon_i(t_j) \sim \mathcal{N}(0, \sigma_i^2)$. The problem is to estimate $\boldsymbol{\theta}_0$ from the data $\mathbf{Y} = (y_i(t_j))_{ij}$. The methodology presented here can also be used if $\boldsymbol{\xi}_0$ is unknown and some of the states are unobserved.

In this paper, we adopt Bayesian approach to inference. For some prior density π of $\boldsymbol{\theta}$ the posterior density is

$$p(\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\xi}_0, \boldsymbol{\sigma}) = \pi(\boldsymbol{\theta}) \prod_{j=1}^{d} \mathcal{N}\{\mathbf{Y}_{j,\cdot}|\mathbf{X}(\boldsymbol{\theta}, \boldsymbol{\xi}_0)_{j,\cdot}, \sigma_j \mathbf{I}_n\},$$

where $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_d)$, $\mathbf{X}(\boldsymbol{\theta}, \boldsymbol{\xi}_0) = (x_i(t_j; \boldsymbol{\theta}, \boldsymbol{\xi}_0))_{ij}$ and $\mathbf{I}_n$ is an identity matrix of order n. For exploring the parameter space there is an advantage in using gradient information in MCMC and optimization methods [3, 4]; for concrete examples in ODE estimation setting see [4, 6]. In the problem we consider, the gradient of the log-likelihood can be obtained by solving sensitivity equations, which are of order dp or via the finite difference formulas, which require solving the ODE system at least p times. Both approaches are computationally costly and not scalable to large scale systems.

In this paper, we avoid huge computational burden by using simultaneous perturbation stochastic approximation (SPGA), introduced by Spall [7]. To obtain SPGA, the system of the form (1) need be solved at most 2 times, regardless of the dimension of the system. By using SPGA instead of the gradient in Metropolis adjusted Langevin Markov Chain Monte Carlo (MALA) we obtain a method, which we call SPGA MALA, that can be used for large scale systems. Although there is some loss in efficiency of SPGA MALA due to using an approximation of the derivative this is outweighed by huge computational savings achieved.

The rest of the paper is organized as follows. In sections 2 and 3 reviews of MALA and SPGA are provided, respectively. Section 4 introduces the proposed method. In Section 5 we compare performance of MALA and SPGA MALA on simulated data for various models.

2 Metropolis adjusted Langevin Markov chain Monte Carlo (MALA)

For the probability density $p(\boldsymbol{\theta})$ let $\mathcal{L}(\boldsymbol{\theta}) = \log\{p(\boldsymbol{\theta})\}$ denote the log-density. The MALA proposal [4, p.130] is

$$\boldsymbol{\theta}^* = \boldsymbol{\theta}^k + \epsilon^2 \mathbf{M} \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}^k)/2 + \epsilon \sqrt{\mathbf{M}} \boldsymbol{z}^k, \tag{2}$$

where $\boldsymbol{\theta}^k$ is the value at k-th step, $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{z}|\mathbf{0}, \mathbf{I}_p)$, $\epsilon > 0$ is the step size and $\mathbf{M}$ is the weight matrix. The proposal density and acceptance probability are

$$\begin{aligned} &q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^k) = \mathcal{N}(\boldsymbol{\theta}^*|\boldsymbol{\mu}(\boldsymbol{\theta}^k, \epsilon), \epsilon^2 \mathbf{M}), \\ &\alpha = \min\{1, p(\boldsymbol{\theta}^*) q(\boldsymbol{\theta}^k|\boldsymbol{\theta}^*)/p(\boldsymbol{\theta}^k) q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^k)\}, \end{aligned} \tag{3}$$

respectively, where $\boldsymbol{\mu}(\boldsymbol{\theta}^k, \epsilon) = \boldsymbol{\theta}^k + \epsilon^2 \mathbf{M} \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}^k)/2$. The advantage of MALA over random walk Metropolis algorithm is that it uses the gradient information which leads to better exploration of the parameter space. Disadvantage is that it requires selection of the weight matrix $\mathbf{M}$. In [4], a fully automated algorithm is proposed for this but it cannot be used in our setting because it requires derivatives. For more details regarding MALA see [2, 4].

3 Simultaneous perturbation gradient approximation (SPGA)

In order to estimate partial derivatives via finite difference (FD) approximation the parameter perturbations are performed along each coordinate separately. For example, the estimate of the j-th partial derivative of $\mathcal{L}(\boldsymbol{\theta}^k)$ via the central difference formula is

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta}^k)}{\partial \theta_j} \approx \frac{\mathcal{L}(\boldsymbol{\theta}^k + h\boldsymbol{e}_j) - \mathcal{L}(\boldsymbol{\theta}^k - h\boldsymbol{e}_j)}{2h},$$

where $\boldsymbol{e}_j$ is the j-th unit vector and h is sufficiently small. This requires $2p$ evaluations of $\mathcal{L}$. With simultaneous perturbation (SP), introduced by Spall [7], all elements of $\boldsymbol{\theta}^k$ are randomly perturbed together. The two sided simultaneous perturbation gradient approximation (SPGA) is

$$\widehat{\nabla}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}^k) = \frac{\mathcal{L}(\boldsymbol{\theta}^k + h\Delta) - \mathcal{L}(\boldsymbol{\theta}^k - h\Delta)}{2h} (\Delta_1^{-1}, \Delta_2^{-1}, \ldots, \Delta_p^{-1})^\top, \tag{4}$$

where $\Delta = (\Delta_1, \Delta_2, \ldots, \Delta_p)^\top$ is usually a random vector of independent Bernoulli random variables that take values -1 and 1 with probability 0.5, although other choices are possible. Two sided SPGA requires *two* evaluations of $\mathcal{L}$ regardless of the dimension p. FD approximation is superior to SP approximation as an estimator of the gradient. However, Spall [7] showed that when used in stochastic optimization setting they achieve the same level of statistical accuracy for a given number of iterations in terms of estimation of the optimum of the objective function. In the next section, we follow the same idea but in the MCMC setting.

4 Simultaneous perturbation gradient approximation based Metropolis adjusted Langevin Markov chain Monte Carlo (SPGA MALA)

SPGA MALA proposal is obtained by substituting the gradient in MALA proposal (2) with its SPGA, defined in (4):

$$\boldsymbol{\theta}^* = \boldsymbol{\theta}^k + \epsilon^2 \mathbf{M} \widehat{\nabla}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}^k)/2 + \epsilon \sqrt{\mathbf{M}} \boldsymbol{z}^k. \tag{5}$$

In view of the MALA proposal density in (3), we require that the density of $\boldsymbol{\theta}^*$ given $\boldsymbol{\theta}^k$ and Δ is $q(\boldsymbol{\theta}^* | \boldsymbol{\theta}^k, \Delta) = \mathcal{N}(\boldsymbol{\theta}^* | \widehat{\boldsymbol{\mu}}(\boldsymbol{\theta}^k, \epsilon, \Delta), \epsilon^2 \mathbf{M})$, where $\widehat{\boldsymbol{\mu}}(\boldsymbol{\theta}^k, \epsilon, \Delta) = \boldsymbol{\theta}^k + \epsilon^2 \mathbf{M} \widehat{\nabla}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}^k)/2$. Since Δ can take 2^p values with equal probability it follows that

the density of $\boldsymbol{\theta}^*$ given $\boldsymbol{\theta}^k$ is the mixture density $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^k) = \frac{1}{2^p}\sum_\Delta q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^k, \Delta)$. The proposal mechanism (5) with proposal density q and standard accepetance probability $\alpha = \min\{1, p(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^k|\boldsymbol{\theta}^*)/p(\boldsymbol{\theta}^k)q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^k)\}$ defines a valid Markov chain; it is simply Metropolis Hastings (MH) algorithm where the proposal is a mixture density q. However, evaluating α is intractable for large p. Because of this, instead of α we use

$$\alpha_\Delta = \min\{1, p(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^k|\boldsymbol{\theta}^*, \Delta)/p(\boldsymbol{\theta}^k)q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^k, \Delta)\}.$$

In other words, instead of using q which involves calculation of each $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^k, \Delta)$ for 2^p possible values of Δ, we use the acceptance ratio which involves calculation of $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^k, \Delta)$ only for the drawn value of Δ.
This algorithm defines a valid Markov chain since it can be viewed as Metropolis-Hastings-Green(MHG) algorithm [2, p.41]. MHG algorithm allows *state-dependent mixing* or *random proposals* [1], meaning that on each step the proposal distribution need not be fixed but can belong to a countable family of proposal distributions. In our case it is a finite family $\{q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^k, \Delta) : \Delta = (\Delta_1, \ldots, \Delta_p), \Delta_i \in \{-1, 1\}\}$. Using the random proposal instead of the mixture proposal comes with a price. As pointed out in the discussion section of the article [1], the random proposal method is less efficient because it accepts fewer proposals. This reduces efficiency of SPGA MALA. The second reason for reduced efficiency of SPGA MALA is that instead of the gradient its approximation is used. However, it is clear that SPGA MALA will be much faster than MALA for large scale systems.

5 Numerical results

In this section we compare the described algorithm to MALA on simulated data generated from the following models.
Fitz Hugh Nagumo (FHN) example. Fitz-Hugh Nagumo system [4] models the behaviour of spike potentials in the giant axon of squid neurons. It has the form

$$\begin{aligned} x_1'(t) &= \theta_3\{x_1(t) - x_1(t)^3/3 + x_2(t)\}, \\ x_2'(t) &= -\tfrac{1}{\theta_3}\{x_1(t) - \theta_1 + \theta_2 x_2(t)\}. \end{aligned}$$

We used different notation than in [4], namely (x_1, x_2) for (V, R) and $(\theta_1, \theta_2, \theta_3)$ for (a, b, c). We set $\boldsymbol{\theta} = (0.2, 0.2, 3)$ and $\boldsymbol{\xi} = (-1, 1)$.
α- pinene example. The following model describes the thermal isomerization of α-pinene [8].

$$\begin{aligned} x_1'(t) &= -(\theta_1 + \theta_2)x_1(t), \\ x_2'(t) &= \theta_1 x_1(t), \\ x_3'(t) &= \theta_2 x_1(t) - (\theta_3 + \theta_4)x_3(t) + \theta_5 x_5(t), \\ x_4'(t) &= \theta_3 x_3(t), \\ x_5'(t) &= \theta_4 x_3(t) - \theta_5 x_5(t). \end{aligned}$$

The values of the parameters that we used are $\boldsymbol{\theta} = (0.1, 0.1, 0.3, 0.1, 0.3)$ and $\boldsymbol{\xi} = (1, 0, 0, 0, 0)$.
Hockin model. In [5], a model of the extrinsic blood coagulation is developed and consists of 34 differential equations and 42 rate constants. Due to lack of space

we do not present the model but refer the reader to the aforementioned reference. We fixed 32 parameters and estimated the remaining 10. The value of the selected parameter was set to $\boldsymbol{\theta} = (0.1, 0.4, 0.1, 0.32, 0.2, 1.05, 2.4, 6, 1.8, 8.2)$.

From each of the models presented above we generated 200 data points on the interval $[0, 20]$ and added Gaussian-distributed noise with standard deviation equal to 0.5. In SPGA (see (4)) we set $h = 10e{-}5$ while the gradient in MALA is obtained by solving sensitivity equations. Ideally, the tuning parameters in MALA should be chosen in such a way that acceptance rate is between 40% and 70%. As it was pointed out in Section 2, tuning of MALA is an issue. To simplify, for both MALA and SPGA MALA we set $\mathbf{M} = \mathbf{I}_p$, $\epsilon = 0.0002p^{-1/3}$ in all the simulations. This selection achieves the desired acceptance rate in FHN model and was used in [4]. For the other two models this is not the case. However, the most important thing here is to compare the performance of these two methods for the same selection of tuning parameters. For comparing sampling efficiency we followed approach used in [4]. A single Markov chain was initialized on the true mode and 5000 posterior samples were collected. The effective sample size (ESS) for each parameter was calculated; the minimum of ESS was used to calculate the time per effectively independent sample. For each method we ran 10 simulations, using the same data set. The methods were implemented in the interpreted language MATLAB and all computations were carried out on an Intel Core i5 computer with 1.3 GHz processor speed and 4 GB of memory. The results of our simulations are presented in Table 1.

The results of FHN example demonstrate loss in efficiency of SPGA MALA with respect to MALA; see Section 4 for the discussion. In α-pinene example MALA is still faster even though the sensitivity equations are of order 25. This is because the original system and the system of sensitivity equations are both linear. The example of Hockin model show the advantage of SPGA MALA. Sensitivity equations are of order 340 and this heavily affects the computatation time of MALA. On the other hand, the computation time of SPGA MALA is much smaller compared to that of MALA, making it much better in terms of the relative speed per effectively independent sample.

Acknowledgements: This research is supported by the Dutch Technology Foundation STW, which is part of the Netherlands Organisation for Scientific Research (NWO), and which is partly funded by Ministry of Economic Affairs. Mark Girolami and Ben Calderhead are acknowledged for making their MATLAB code used in [4] freely available. We thank Itai Dattner and Bart Bakker for useful comments.

References

[1] J. Besag, P. Green, D. Higdon, and K. Mengersen. Bayesian computation and stochastic systems. *Statistical science*, pages 3–41, 1995.

[2] S. Brooks, A. Gelman, G. Jones, and X.-L. Meng. *Handbook of Markov Chain Monte Carlo*. CRC press, 2011.

[3] A. R. Conn, K. Scheinberg, and L. N. Vicente. *Introduction to derivative-free optimization*, volume 8. Siam, 2009.

Model	Sampling method	Time (s)	Mean ESS ($\boldsymbol{\theta}$)	Total time /minimum mean ESS	Relative speed
$d = 2$ $p = 3$			$(\theta_1, \theta_2, \theta_3)$		
FHN	MALA	363.6	145, 30, 109	12.12	3.4
	SPGA	623.2	84 , 15, 48	41.55	1
	MALA				
$d = 5$ $p = 5$			$(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$		
α-Pinene	MALA	63.7	59, 58, 17, 11, 6	10.62	2.54
	SPGA	134.8	187, 96, 6, 23 , 5	26.96	1
	MALA				
$d = 34$ $p = 10$			$(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5,$ $\theta_6, \theta_7, \theta_8, \theta_9, \theta_{10})$		
Hockin	MALA	1.03e+04	5 6 8 7 7 7 6 5 6 8	2060	1
	SPGA	180.5	7 6 8 8 7 6 6 5 4 7	45.13	45.65
	MALA				

Table 1: Summary of results for 10 runs of the model parameter sampling schemes for different models with 5000 posterior samples.

[4] M. Girolami and B. Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

[5] M. F. Hockin, K. C. Jones, S. J. Everse, and K. G. Mann. A model for the stoichiometric regulation of blood coagulation. *Journal of Biological Chemistry*, 277(21):18322–18333, 2002.

[6] A. Raue, M. Schilling, J. Bachmann, A. Matteson, M. Schelke, D. Kaschek, S. Hug, C. Kreutz, B. D. Harms, F. J. Theis, et al. Lessons learned from quantitative dynamical modeling in systems biology. *PloS one*, 8(9):e74335, 2013.

[7] J. C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *Automatic Control, IEEE Transactions on*, 37(3):332–341, 1992.

[8] I. B. Tjoa and L. T. Biegler. Simultaneous solution and optimization strategies for parameter estimation of differential-algebraic equation systems. *Industrial & Engineering Chemistry Research*, 30(2):376–385, 1991.

Author Index