Proceedings of the 20th European Young Statisticians Meeting

Tilo Wiklund (Editor)

Held at the Department of Mathematics
Uppsala University, Uppsala, Sweden
14–18 August 2017

Local Organising Committee

Måns Thulin Department of Statistics, Uppsala University

Tilo Wiklund Department of Mathematics, Uppsala University

Held at the Department of Mathematics, Uppsala University and sponsored by the Departments of Mathematics and Statistics at Uppsala University.

UPPSALA
UNIVERSITET

Bernoulli Society
for Mathematical Statistics
and Probability

Preface

The European Young Statisticians Meetings are held every two years under the auspices of the European Regional Committee of the Bernoulli Society. In 2017, between the 14th and 18th of August, the 20th EYSM was held Uppsala, Sweden.

The idea of the meeting is to provide young researchers (less than thirty years of age or two to eight years of research experience) with an introduction to the international scene within the broad subject area, from pure probability theory to applied statistics. Participation is by invitation and every participant submits contribution and gives a talk. There are no parallel sessions.

Apart from the speakers included in these proceedings the meeting was guested by five keynote speakers: Jimmy Olsson from the Royal Institute of Technology in Stockholm, Jenny Wadsworth from Lancaster University, Jane Hillston from the University of Edinburgh, Hannu Oja from University of Turku and Svante Janson from Uppsala University.

We would like to express our gratitude to these speakers as well as to all participants of the conference. Moreover we wish to thank to the Department of Mathematics and the Department of Statistics at Uppsala University for sponsoring the conference and the European Regional Committee of the Bernoulli Society for entrusting us with organising the meeting.

September, 2017
Tilo Wiklund & Måns Thulin

Contents

Contributed Papers

Delete or Merge Regressors algorithm

Agnieszka Prochenka [*1] **and Piotr Pokarowski**[1]

[1]*Faculty of Mathematics, Informatics and Mechanics, University of Warsaw*

This paper addresses a problem of linear and logistic model selection in the presence of both continuous and categorical predictors. In the literature two types of algorithms dealing with this problem can be found. The first one is the well known `group lasso` ([3]) which selects a subset of continuous and a subset of categorical predictors. Hence, it either deletes or not an entire factor. An improvement of the `group lasso` regularization is `group MCP` (using Minimax Concave Penalty) described in [6]. It assumes a concave penalty and therefore uses more difficult optimization algorithms. The second type is `CAS-ANOVA` ([1]) which selects a subset of continuous predictors and partitions of factors. Therefore, it merges levels within factors. Similar method with different optimization method is called `gvcm` and is described in [5].

In the article an algorithm called `DMR` (Delete or Merge Regressors) is described. Like `CAS-ANOVA` it selects a subset of continuous predictors and partitions of factors. However, instead of using regularization, it is based on a stepwise procedure, where in each step either one continuous variable is deleted or two levels of a factor are merged. The order of accepting consecutive hypotheses is based on sorting Wald statistics. Some of the preliminary results for `DMR` are described in [2].

`DMR` algorithm works only for data sets where $p < n$ (number of columns in the model matrix is smaller than the number of observations). In the paper a modification of `DMR` called `DMRnet` is introduced that works also for data sets where $p \gg n$. `DMRnet` uses regularization in the screening step and `DMR after` decreasing the model matrix to $p < n$.

Theoretical results prove that `DMR` for linear and logistic regression are consistent model selection methods even when p tends to infinity with n. Furthermore, upper bounds on the error of selection were calculated. However, in this paper the focus is on description of the algorithm and real data example, for which `DMRnet` chooses smaller models with not higher prediction error than the competitive methods.

Keywords: factorial selection, logistic regression, linear regression, Wald statistics, hierarchical clustering

[*]Corresponding author: a.prochenka@phd.ipipan.waw.pl

1 Factorial selection

We consider n data points $(y_1, \mathbf{x}_{1\cdot}^T)$, $(y_2, \mathbf{x}_{2\cdot}^T)$, ..., $(y_n, \mathbf{x}_{n\cdot}^T)$ with univariate responses y_i and p-dimensional covariates $\mathbf{x}_{i\cdot}^T$. Denote by $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$, $\mathbf{x}_j = (x_{1j}, x_{2j}, \ldots, x_{nj})^T$, $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p)$ the n times p model matrix. We assume that $\mathbf{X}$ is a full rank matrix.

Let y_i be independent, such that $y_i \sim f_{\eta_i,\sigma^2}(\cdot)$ and $\eta_i = \mathbf{x}_{i\cdot}^T\boldsymbol{\beta}$ for $\boldsymbol{\beta} \in \mathbb{R}^p$, where f_{η_i,σ^2} is the density function of some distribution in the exponential family. Let us denote $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_n)^T$ and

$$\boldsymbol{\eta}^* = \mathbf{X}\boldsymbol{\beta}^* = \mathbf{1}\beta_{00}^* + \mathbf{X}_0\boldsymbol{\beta}_0^* + \mathbf{X}_1\boldsymbol{\beta}_1^* + \ldots + \mathbf{X}_l\boldsymbol{\beta}_l^*, \tag{1}$$

where

1. $\mathbf{X} = [\mathbf{1}, \mathbf{X}_0, \mathbf{X}_1, \ldots, \mathbf{X}_l]$ is a model matrix organized as follows: $\mathbf{X}_0$ is a matrix corresponding to continuous regressors and $\mathbf{X}_1, \ldots, \mathbf{X}_l$ are zero-one matrices encoding corresponding factors with the first level set as the reference.

2. $\boldsymbol{\beta}^* = [\beta_{00}^*, \boldsymbol{\beta}_0^{*T}, \boldsymbol{\beta}_1^{*T}, \ldots, \boldsymbol{\beta}_l^{*T}]^T \in \mathbb{R}^p$ is a parameter vector organized as follows: β_{00}^* is the intercept, $\boldsymbol{\beta}_0^* = [\beta_{10}^*, \ldots, \beta_{p_0 0}^*]^T$ is a vector of coefficients for continuous variables and $\boldsymbol{\beta}_k^* = [\beta_{2k}^*, \ldots, \beta_{p_k k}^*]^T$ is a vector of parameters corresponding to the k-th factor, $k = 1, \ldots, l$, hence the length of the parameter vector is $p = 1 + p_0 + (p_1 - 1) + \ldots + (p_l - 1)$.

Denote sets of indexes: $N = \{0, 1, \ldots, l\}$, $N_0 = \{0, 1, \ldots, p_0\}$ and $N_k = \{2, 3, \ldots, p_k\}$ for $k \in N \setminus \{0\}$. Let us define an elementary constraint for model (1) as a linear constraint of one of two types:

$$\mathcal{H}_{jk} : \ \beta_{jk}^* = 0 \text{ where } j \in N_k \setminus \{0\}, \ k \in N, \tag{2}$$

$$\mathcal{H}_{ijk} : \ \beta_{ik}^* = \beta_{jk}^* \text{ where } i, j \in N_k, \ i \neq j, \ k \in N \setminus \{0\}. \tag{3}$$

1.1 Feasible models

A feasible model can be defined as a sequence $M = (P_0, P_1, ..., P_l)$, where P_0 denotes a subset of indexes of continuous variables and P_k is a particular partition of levels of the k-th factor. Such a model can be encoded by a set of elementary constraints. A set of all feasible models is denoted by $\mathcal{M}$. Let us denote a model $F \in \mathcal{M}$ without constraints of types (2) or (3) as the full model.

Example. For illustration, let us consider a linear predictor with one factor and one continuous variable:

$$\mathbf{X}\boldsymbol{\beta}^* = \mathbf{1}\cdot 1 + \mathbf{X}_0\cdot 2 + \mathbf{X}_1\cdot \begin{bmatrix} -2 \\ -2 \\ 0 \end{bmatrix}$$
$$= \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \cdot 1 + \begin{bmatrix} -0.96 \\ -0.29 \\ 0.26 \\ -1.15 \\ 0.2 \\ 0.03 \\ 0.09 \\ 1.12 \end{bmatrix} \cdot 2 + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} -2 \\ -2 \\ 0 \end{bmatrix}$$

Then $\beta^* = [1, 2, -2, -2, 0]^T$. The full model $F = (P_0 = \{1\}, P_1 = \{\{1\}, \{2\}, \{3\}, \{4\}\})$ with $p_0 = 1, p_1 = 4, p = 5$. The true model is $(P_0 = \{1\}, P_1 = \{\{1, 4\}, \{2, 3\}\})$ and is the same as the full model with two elementary constraints: $\beta^*_{41} = 0$ and $\beta^*_{21} = \beta^*_{31}$.

Our goal is to find the best feasible model according to Generalized Information Criterion (GIC) or estimated prediction error using cross-validation, taking into account that the number of feasible models grows faster than exponentially with p. In order to significantly reduce the amount of computations, we propose a greedy backward search.

2 `DMR` and `DMRnet` algorithms

`DMR` for generalized linear models is described in details in Algorithm 1. `DMRnet` is a generalization of `DMR` to high-dimensional data where $p \gg n$ by adding screening step using `group lasso`. After reduction of the dimension of the model to $p < n$, `DMR` algorithm is used. In order to make the screening step more accurate and to better balance the impact of screening and the `DMR` selection steps, the screening is done multiple times.

Example. Example 1.1 continued, `DMR` algorithm with GIC:

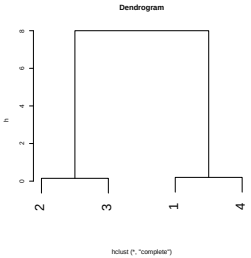
$$w^2_{110} = 9.35, \mathbf{D_1} = \begin{bmatrix} 0 & w^2_{121} & w^2_{131} & w^2_{141} \\ w^2_{121} & 0 & w^2_{231} & w^2_{241} \\ w^2_{131} & w^2_{231} & 0 & w^2_{341} \\ w^2_{141} & w^2_{241} & w^2_{341} & 0 \end{bmatrix} = \begin{bmatrix} 0 & 8.01 & 4.52 & 0.20 \\ 8.01 & 0 & 0.15 & 3.09 \\ 4.52 & 0.15 & 0 & 2.91 \\ 0.20 & 3.09 & 2.91 & 0 \end{bmatrix},$$

cutting heights for agglomerative clustering illustrated in Figure 1:

$$\mathbf{h} = [0, 0.15, 0.20, 8.01, 9.35]^T,\ \mathbf{A_0} = \begin{array}{c} \begin{matrix} \beta_{00} & \beta_{10} & \beta_{21} & \beta_{31} & \beta_{41} \end{matrix} \\ \begin{bmatrix} 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix} \end{array},$$

$\mathbf{GIC} = [28.33, 26.65, 25.36, 34.68, 39.59]^T$. The selected model according to GIC is the third one (GIC = 25.36) with two elementary constraints: $\beta^*_{41} = 0$ and $\beta^*_{21} = \beta^*_{31}$, which is the true model.

Figure 1: Dendrogram for hierarchical clustering used in Example 1.1.



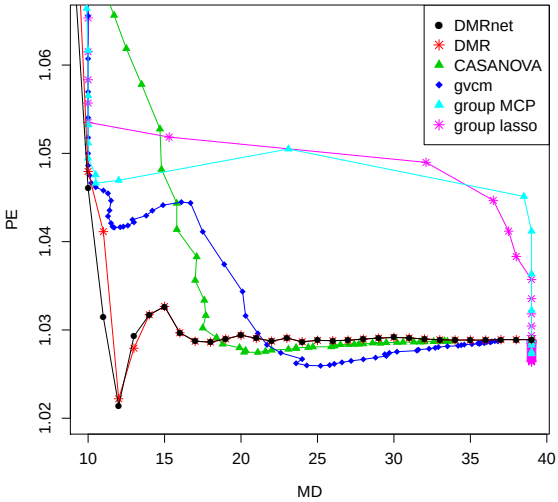
3 Real data example: Miete

The data set Miete comes from `http://www.statistik.lmu.de/service/datenarchiv`. The data consists of $n = 2053$ households interviewed for the Munich rent standard 2003. The response is monthly rent per square meter in Euros, data is described in detail in [4]. 8 categorical and 2 continuous variables give 36 and 3 (including the intercept) parameters. This gives $p = 39$.

In Figure 2 a plot of prediction error (PE) vs model dimension (MD) calculated by 10-fold C-V for 100 λ values for `CAS-ANOVA`, `gvcm`, `group MCP` and `group lasso` and from 1 to p for `DMR` and from 1 to $\min\{p, \frac{n}{2}\}$ for `DMRnet` is shown. For every algorithm we can find a global minimum: for `DMRnet` and `DMR` these are when MD = 12, for `CAS-ANOVA` when MD = 21.1, for `gvcm` when MD = 25.2 and for `group MCP` and `group lasso` for the full model, MD=39. If we chose models with the lowest prediction error, `DMRnet` would have both the smallest error and the smallest number of parameters.

Acknowledgements: The research is supported by the Polish National Science Center grant 2015/17/B/ST6/01878.

Figure 2: PE vs MD calculated by 10-fold C-V for Miete data set.



References

[1] Bondell, Howard D., and Brian J. Reich. "Simultaneous factor selection and collapsing levels in ANOVA." Biometrics 65.1 (2009): 169-177.

[2] Maj-Kańska, Aleksandra, Piotr Pokarowski, and Agnieszka Prochenka. "Delete or merge regressors for linear model selection." Electronic Journal of Statistics 9.2 (2015): 1749-1778.

[3] Yuan, Ming, and Yi Lin. "Model selection and estimation in regression with grouped variables." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68.1 (2006): 49-67.

[4] Tutz, Gerhard. Regression for categorical data. Vol. 34. Cambridge University Press, 2011.

[5] Oelker, Margret-Ruth, Jan Gertheiss, and Gerhard Tutz. "Regularization and model selection with categorical predictors and effect modifiers in generalized linear models." Statistical Modelling 14.2 (2014): 157-177.

[6] Breheny, Patrick, and Jian Huang. "Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors." Statistics and computing 25.2 (2015): 173-187.

Algorithm 1: DMR (Delete or Merge Regressors for generalized linear models)

Input: $\mathbf{y}$, $\mathbf{X}$

1. Computation of Wald statistics

Calculate Wald statistics for all elementary constraints defined in (2):

for $j \in N_k \setminus \{0\},\ k \in N$ **do**

$$w_{1jk}^2 = \frac{\widehat{\beta}_{jk}^2}{\widehat{Var}(\widehat{\beta}_{jk})}$$

end for

Calculate Wald statistics for all elementary constraints defined in (3):

for $i, j \in N_k,\ i \neq j,\ k \in N \setminus \{0\}$ **do**

$$w_{ijk}^2 = \frac{(\widehat{\beta}_{ik} - \widehat{\beta}_{jk})^2}{\widehat{Var}(\widehat{\beta}_{ik} - \widehat{\beta}_{jk})}$$

end for

2. Agglomerative clustering for factors (using complete linkage clustering)

For each factor perform agglomerative clustering using $\mathbf{D}_k = [d_{ijk}]_{ij}$ as dissimilarity matrix.

for $k \in N \setminus \{0\}$ **do**

$d_{1jk} = d_{j1k} = w_{1jk}$ for $j \in N_k$,

$d_{ijk} = w_{ijk}$ for $i, j \in N_k,\ i \neq j$,

$d_{iik} = 0$ for $i \in N_k$.

end for

Denote cutting heights obtained from the clusterings of l factors as $\mathbf{h}_1^T, \mathbf{h}_2^T, \ldots, \mathbf{h}_l^T$.

3. Sorting constraints (hypotheses) according to the likelihood ratio test statistics

Combine vectors of cutting heights: $\mathbf{h} = [0, \mathbf{h}_0^T, \mathbf{h}_1^T, \ldots, \mathbf{h}_l^T]^T$, where $\mathbf{h}_0$ is a vector of likelihood ratio test statistics for constraints concerning continuous variables and 0 corresponds to the full model. Sort elements of $\mathbf{h}$ in increasing order and construct a corresponding $(p-1) \times p$ matrix $\mathbf{A}_0$ of consecutive constraints.

4. Computation of log-likelihood for models on the nested path

for $m = 0, \ldots, p-1$ **do**

$L_{M_m} = \ell(\widehat{\boldsymbol{\beta}}_{M_m})$, where M_m is the model with m first constraints from A_0 accepted.

end for

Output: $\mathscr{M}^{\texttt{DMR}} = \{M_0, \ldots, M_{p-1}\}$, $\mathbf{L}^{\texttt{DMR}} = (L_{M_0}, \ldots, L_{M_{p-1}})^T$.

Matrix Independent Component Analysis

Joni Virta[*1]

[1] *University of Turku, Finland*

Independent component analysis (ICA) is a popular means of dimension reduction for vector-valued random variables. In this short note we review its extension to arbitrary tensor-valued random variables by considering the special case of two dimensions where the tensors are simply matrices.

Keywords: FOBI, Kronecker structure, Kurtosis

1 Matrix independent component model

For an introduction to classical vector-valued independent component analysis (ICA) the reader is referred to [3]. The tensorial ICA theory we next review was first introduced in [6] and further investigated in [7].

Let $\mathbf{X} \in \mathbb{R}^{p_1 \times p_2}$ be a random matrix from the matrix location-scale model

$$\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\Omega}_1 \mathbf{Z} \boldsymbol{\Omega}_2^T, \tag{1}$$

where the location matrix $\boldsymbol{\mu} \in \mathbb{R}^{p_1 \times p_2}$ and the non-singular mixing matrices $\boldsymbol{\Omega}_1 \in \mathbb{R}^{p_1 \times p_1}$ and $\boldsymbol{\Omega}_2 \in \mathbb{R}^{p_2 \times p_2}$ are unknown parameters and $\mathbf{Z} \in \mathbb{R}^{p_1 \times p_2}$ is an unobserved random matrix with finite joint fourth moments. Defining $vec : \mathbb{R}^{p_1 \times p_2} \to \mathbb{R}^{p_1 p_2}$ as the function that stacks the columns of its argument into a vector, the model (1) can be written as

$$vec\left(\mathbf{X}\right) = vec\left(\boldsymbol{\mu}\right) + \left(\boldsymbol{\Omega}_2 \otimes \boldsymbol{\Omega}_1\right) vec\left(\mathbf{Z}\right), \tag{2}$$

where $\otimes$ is the Kronecker product. Thus (1) can also be thought as a structured location-scale model (Kronecker model) for random vectors.

We will next describe conditions under which the model (1) is well-defined. For any non-singular $\mathbf{A}_1 \in \mathbb{R}^{p_1 \times p_1}$ and $\mathbf{A}_2 \in \mathbb{R}^{p_2 \times p_2}$ it can be written as

$$\mathbf{X} = \boldsymbol{\mu} + \left(\boldsymbol{\Omega}_1 \mathbf{A}_1^{-1}\right) \left(\mathbf{A}_1 \mathbf{Z} \mathbf{A}_2^T\right) \left(\boldsymbol{\Omega}_2 \mathbf{A}_2^{-1}\right)^T = \boldsymbol{\mu} + \boldsymbol{\Omega}_1^* \mathbf{Z}^* \left(\boldsymbol{\Omega}_2^*\right)^T,$$

showing that the parameters are not identifiable as such. Note that we can never achieve full identifiability as for any non-zero scalar β the maps $\boldsymbol{\Omega}_1 \mapsto \beta \boldsymbol{\Omega}_1$ and

*Corresponding author: joni.virta@utu.fi

$\boldsymbol{\Omega}_2 \mapsto \beta^{-1}\boldsymbol{\Omega}_2$ preserve the model. In the following we will refer to identifiability up to this proportionality as *proportional identifiability*. As a first step towards proportional identifiability we set the following constraints for $\mathbf{Z}$.

$$E\left[vec\left(\mathbf{Z}\right)\right] = \mathbf{0}_{p_1p_2} \quad \text{and} \quad Cov\left[vec\left(\mathbf{Z}\right)\right] = \mathbf{I}_{p_1p_2}.$$

The first constraint fixes the location matrix $\boldsymbol{\mu}$ and the second makes both $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$ proportionally identifiable up to orthogonal $\mathbf{A}_1$ and $\mathbf{A}_2$.

To impose more structure the model can be equipped with additional assumptions on the latent matrix $\mathbf{Z}$. The classical choice is to assume that $vec(\mathbf{Z}) \sim \mathcal{N}(\mathbf{0}_{p_1p_2}, \mathbf{I}_{p_1p_2})$, resulting in a general matrix normal distribution for $\mathbf{X}$. The normal model can further be generalized in two directions. Focusing on the orthogonal invariance of the standard normal distribution leads us to consider the class of spherical random matrices satisfying $\mathbf{Z} \sim \mathbf{U}_1\mathbf{Z}\mathbf{U}_2^T$ for all orthogonal $\mathbf{U}_1 \in \mathbb{R}^{p_1\times p_1}, \mathbf{U}_2 \in \mathbb{R}^{p_2\times p_2}$ and this in turn yields a matrix elliptical distribution for $\mathbf{X}$, see [4] for the previous two models.

The second generalization is based on another key characteristic of the standard multivariate normal distribution, the equivalence of uncorrelatedness and independence, and equips $\mathbf{Z}$ with the following assumption.

A1. The components of $\mathbf{Z}$ are mutually independent.

While assumption A1 is rather strong, actually strong enough to guarantee the proportional identifiability of $\mathbf{Z}$ in (1) up to some trivialities when paired with A2 below, it is still a natural choice in applications where the components of $\mathbf{Z}$ can each be thought to model one separate aspect of the phenomenon which then combine independently to produce the observation $\mathbf{X}$.

The Skitovich-Darmois theorem [5] states that if a set of independent random variables can be combined to yield non-trivial linear combinations that are itself independent they must all be normally distributed. Thus we must further restrict the presence of multivariate normal distribution in the latent matrix to avoid $\mathbf{A}_1\mathbf{Z}\mathbf{A}_2^T$ having independent components for non-trivial $\mathbf{A}_1$ and $\mathbf{A}_2$.

A2. At most one row of $\mathbf{Z}$ has a multivariate normal distribution and at most one column of $\mathbf{Z}$ has a multivariate normal distribution.

Assumptions A1 and A2 now jointly guarantee that $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$ are proportionally identifiable up to $\mathbf{A}_1$ and $\mathbf{A}_2$ containing a single ± 1 in each of their rows and columns. Consequently the matrix $\mathbf{Z}$ can be estimated up to the order and signs of its rows and columns, a defect that is usually of no consequence in practice.

Definition 1. We say that $\mathbf{X} \in \mathbb{R}^{p_1\times p_2}$ obeys the matrix independent component model (MICM) if it satisfies (1) along with assumptions A1 and A2.

To wrap everything up, in matrix independent component analysis we assume that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ is a random sample from the distribution of $\mathbf{X}$ obeying MICM and our objective is the estimation of the matrices $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$.

2 The estimation of Z

Let $\mathbf{X}$ obey MICM. Centering the random matrix as $\mathbf{X} \mapsto \mathbf{X} - E[\mathbf{X}]$ shows that without loss of generality we may assume in the following that $vec(\boldsymbol{\mu}) = \mathbf{0}_{p_1 p_2}$.

A key notion in the model is that, instead of treating the elements of $\mathbf{Z}$ separately, we consider them in an aggregate sort of way via their corresponding rows and columns. As an example take assumptions A1 and A2, the first of which can be written equivalently as "the rows of $\mathbf{Z}$ are mutually independent and the columns of $\mathbf{Z}$ are mutually independent". The same thought is also reflected in our definitions of the row and column covariance matrices,

$$\boldsymbol{\Sigma}_1(\mathbf{X}) = \frac{1}{p_2} E\left[\mathbf{X}\mathbf{X}^T\right] \quad \text{and} \quad \boldsymbol{\Sigma}_2(\mathbf{X}) = \frac{1}{p_1} E\left[\mathbf{X}^T\mathbf{X}\right].$$

The matrices $\boldsymbol{\Sigma}_1(\mathbf{X})$ and $\boldsymbol{\Sigma}_2(\mathbf{X})$ can be interpreted as the average covariance matrices of the p_2 columns and p_1 rows of $\mathbf{X}$, respectively. Under the independent component model they further enjoy the "equivariance property" described by the next lemma.

Lemma 1. *Let $\boldsymbol{X}$ obey MICM. Then the inverse square roots of the row and column covariance matrix satisfy*

$$\boldsymbol{\Sigma}_1(\boldsymbol{X})^{-1/2} = \frac{p_2^{1/2}}{\|\boldsymbol{\Omega}_2\|_F} \boldsymbol{U}_1 \boldsymbol{\Omega}_1^{-1} \quad \textit{and} \quad \boldsymbol{\Sigma}_2(\boldsymbol{X})^{-1/2} = \frac{p_1^{1/2}}{\|\boldsymbol{\Omega}_1\|_F} \boldsymbol{U}_2 \boldsymbol{\Omega}_2^{-1},$$

for some orthogonal matrices $\boldsymbol{U}_1 \in \mathbb{R}^{p_1 \times p_1}$ and $\boldsymbol{U}_2 \in \mathbb{R}^{p_2 \times p_2}$, where $\|\cdot\|_F$ is the Frobenius (Euclidean) norm.

For the proof of Lemma 1 and all other results in this review see [6]. Lemma 1 immediately yields the first step towards the estimation of $\mathbf{Z}$:

Lemma 2. *Let $\boldsymbol{X}$ obey MICM. Then we have*

$$\left(\boldsymbol{\Sigma}_1(\boldsymbol{X})^{-1/2}\right) \boldsymbol{X} \left(\boldsymbol{\Sigma}_2(\boldsymbol{X})^{-1/2}\right)^T = \gamma \boldsymbol{U}_1 \boldsymbol{Z} \boldsymbol{U}_2^T,$$

with orthogonal $\boldsymbol{U}_1 \in \mathbb{R}^{p_1 \times p_1}$ and $\boldsymbol{U}_2 \in \mathbb{R}^{p_2 \times p_2}$ and $\gamma = (p_1 p_2)^{1/2} \|\boldsymbol{\Omega}_2 \otimes \boldsymbol{\Omega}_1\|_F^{-1}$.

According to Lemma 2 the two-sided standardization of $\mathbf{X}$ reduces the problem of estimating $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$ to the easier task of estimating two orthogonal

matrices. In the following we denote by $\mathbf{X}^{st}$ the standardized matrix on the left-hand side of Lemma 2.

Our method for estimating $\mathbf{U}_1$ and $\mathbf{U}_2$ is based on an extension of a multivariate ICA method called *fourth order blind identification* (FOBI) [1] and will hereafter be referred to as MFOBI. Heuristically ICA can be thought of as the maximization of non-normality and FOBI achieves it via considering a matrix measuring kurtosis, a classical indicator of non-normality. Sure enough, we define both row and column versions of the matrix.

$$\mathbf{B}_1(\mathbf{X}) = \frac{1}{p_2} E\left[\mathbf{X}\mathbf{X}^T\mathbf{X}\mathbf{X}^T\right] \quad \text{and} \quad \mathbf{B}_2(\mathbf{X}) = \frac{1}{p_1} E\left[\mathbf{X}^T\mathbf{X}\mathbf{X}^T\mathbf{X}\right].$$

A key property of $\mathbf{B}_1(\mathbf{X})$ and $\mathbf{B}_2(\mathbf{X})$ with respect to our problem, diagonality under independence, is described in the next lemma.

Lemma 3. *Let the random matrix $\boldsymbol{Z} \in \mathbb{R}^{p_1 \times p_2}$ have mutually independent components with zero means, unit variances and finite joint fourth moments. Then we have*

$$\begin{aligned}\boldsymbol{B}_1(\boldsymbol{Z}) &= (p_1 + p_2 + 1)\,\boldsymbol{I}_{p_1} + diag(\kappa_{1\bullet}, \ldots, \kappa_{p_1\bullet}) \\ \boldsymbol{B}_2(\boldsymbol{Z}) &= (p_1 + p_2 + 1)\,\boldsymbol{I}_{p_2} + diag(\kappa_{\bullet 1}, \ldots, \kappa_{\bullet p_2}),\end{aligned}$$

where $\kappa_{i\bullet}$ is the ith row mean and $\kappa_{\bullet j}$ is the jth column mean of the kurtosis matrix $\boldsymbol{\kappa} = \left(E\left[z_{ij}^4 - 3\right]\right)_{ij}$.

Both matrices $\mathbf{B}_1(\mathbf{X})$ and $\mathbf{B}_2(\mathbf{X})$ are orthogonally equivariant and we obtain the following.

Lemma 4. *Let $\boldsymbol{X}$ obey MICM. Then we have*

$$\boldsymbol{B}_1(\boldsymbol{X}^{st}) = \gamma^4\,\boldsymbol{U}_1\boldsymbol{B}_1(\boldsymbol{Z})\,\boldsymbol{U}_1^T \quad \text{and} \quad \boldsymbol{B}_2(\boldsymbol{X}^{st}) = \gamma^4\,\boldsymbol{U}_2\boldsymbol{B}_2(\boldsymbol{Z})\,\boldsymbol{U}_2^T,$$

where $\boldsymbol{B}_1(\boldsymbol{Z})$ and $\boldsymbol{B}_2(\boldsymbol{Z})$ are diagonal by Lemma 3.

The two equations in Lemma 4 are the eigendecompositions of $\mathbf{B}_1(\mathbf{X}^{st})$ and $\mathbf{B}_2(\mathbf{X}^{st})$ and to guarantee the consistent estimation of $\mathbf{U}_1$ and $\mathbf{U}_2$, the corresponding eigenspectra must be distinct. In the light of Lemma 3 this requirement takes the following form.

A3. The row means of $\boldsymbol{\kappa}$ are distinct and the column means of $\boldsymbol{\kappa}$ are distinct, where $\boldsymbol{\kappa} = \left(E\left[z_{ij}^4 - 3\right]\right)_{ij}$ is the kurtosis matrix of the latent $\mathbf{Z}$.

Assumption A3 is a stronger version of assumption A2 and in particular says that no two rows or columns of $\mathbf{Z}$ may consist solely of random variables with identical distributions. Our main result is then the following.

Theorem 1. *Let* $\boldsymbol{X}$ *obey MICM and satisfy assumption A3. Further let* $\boldsymbol{V}_1 \in \mathbb{R}^{p_1 \times p_1}$ *and* $\boldsymbol{V}_2 \in \mathbb{R}^{p_2 \times p_2}$ *contain the eigenvectors of* $\boldsymbol{B}_1(\boldsymbol{X}^{st})$ *and* $\boldsymbol{B}_2(\boldsymbol{X}^{st})$*, respectively, as their columns. Then we have*

$$\boldsymbol{V}_1^T \boldsymbol{X}^{st} \boldsymbol{V}_2 = \gamma \boldsymbol{Z} \propto \boldsymbol{Z}.$$

The MFOBI solution of Theorem 1 enables the estimation of $\mathbf{Z}$ up to the scaling factor γ which is usually satisfactory enough, the shape and other higher-order properties of the components being of greater interest than their scales. In practice the MFOBI solution is obtained by replacing the expected values by the corresponding sample estimates. After the estimation of $\mathbf{Z}$ a further problem is the choosing of the most "interesting" components among the $p_1 p_2$ elements of $\mathbf{Z}$. Our kurtosis-based approach immediately leads to consider the components with extremal kurtosis, or to stay more in line with the spirit of the method, the rows and columns with the highest and lowest mean kurtoses. However, as the classical kurtosis is a very non-robust statistic the choice of a suitable criterion is still an open question.

3 Discussion

The naïve approach to model (1) is to vectorize it, resulting into (2), and proceed with standard methods of vector-valued ICA. However, this completely ignores the Kronecker structure of the mixing matrix $\mathbf{\Omega}_2 \otimes \mathbf{\Omega}_1$ and the price we pay for our negligence further comes in the form of stronger assumptions and increased computational cost. As an example, consider applying MFOBI to (1) versus applying FOBI to (2). Assumption A1 takes the same form for both methods but the counterpart of assumption A2 for FOBI is much more strict. Namely, it requires that at most one element of $vec(\mathbf{Z})$ has a normal distribution while in MFOBI the majority of the elements of $\mathbf{Z}$ can be normal if conveniently located. Similarly our assumption A3 and its vector-valued analogy, stating that the kurtoses of the elements of $vec(\mathbf{Z})$ are distinct, share the same relationship.

In order to compare the computational costs of the two methods we focus for simplicity only on the computationally most intensive part of the algorithms, the eigendecompositions. For $\mathbf{X} \in \mathbb{R}^{p_1 \times p_2}$ FOBI has to perform two eigendecompositions of a $p_1 p_2 \times p_1 p_2$ matrix while MFOBI requires the eigendecompositions of two $p_1 \times p_1$ and two $p_2 \times p_2$ matrices. In essence MFOBI "divides" the computational load into a larger number of smaller operations, lessening the overall complexity.

In [7] an extension of a second classical ICA method, *joint approximate diagonalization of eigen-matrices* (JADE) [2] for tensor-valued data was intro-

duced. Called TJADE, the method shares the standardization step of MFOBI (or more accurately, of TFOBI, its general tensor-valued extension) but approaches the estimation of the orthogonal matrices differently. In TJADE, instead of diagonalizing a single kurtosis matrix, we diagonalize several of them at once, essentially using more information in the estimation (and consequently increasing the computational burden as well). The implementations of both methods along with several other tensor extensions of classical methods can be found in the R-package *tensorBSS* [8].

Interestingly, restricting to matrix-valued observations only in this review serves more than just instructional purposes. In [6], [7] it is shown that the general tensor versions of the methods can be reduced to the matrix case. Similarly it can be shown that for the limiting distributions of the corresponding estimators it is sufficient to consider only the matrix case.

This note has been left devoid of examples and applications of the discussed methodology and instead several can be found, for example, in [6, 7].

References

[1] J.-F. Cardoso. Source separation using higher order moments. In *International Conference on Acoustics, Speech, and Signal Processing, 1989. ICASSP-89.*, pages 2109–2112. IEEE, 1989.

[2] J.-F. Cardoso and A. Souloumiac. Blind beamforming for non-Gaussian signals. In *IEE Proceedings F (Radar and Signal Processing)*, volume 140, pages 362–370. IET, 1993.

[3] P. Comon and C. Jutten. *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press, 2010.

[4] A. K. Gupta and D. K. Nagar. *Matrix variate distributions*, volume 104. CRC Press, 1999.

[5] I. Ibragimov. On the Ghurye-Olkin-Zinger theorem. *Journal of Mathematical Sciences*, 199(2), 2014.

[6] J. Virta, B. Li, K. Nordhausen, and H. Oja. Independent component analysis for tensor-valued data. *Preprint in arXiv:1602.00879*, 2016.

[7] J. Virta, B. Li, K. Nordhausen, and H. Oja. JADE for tensor-valued observations. *Preprint in arXiv:1603.05406*, 2016.

[8] J. Virta, B. Li, K. Nordhausen, and H. Oja. *tensorBSS: Blind Source Separation Methods for Tensor-Valued Observations*, 2016. R package version 0.3.

Nonparametric estimation of gradual change points in the jump behaviour of an Itō semimartingale

Michael Hoffmann*[1]

[1]*Ruhr-Universität Bochum*

In applications the properties of a stochastic feature often change gradually rather than abruptly, that is: after a constant phase for some time they slowly start to vary. In this paper we discuss the localisation of a gradual change point in the jump characteristic of a discretely observed Itō semimartingale. We propose a new measure of time variation for the jump behaviour of the process. Based on weak convergence of a suitable stochastic process we derive an estimator for the first point in time where the jump characteristic changes.

Keywords: Lévy measure, jump compensator, empirical processes, weak convergence, gradual changes

1 Introduction

Stochastic processes in continuous time are widely used in science nowadays, as they allow for a flexible modeling of the evolution of various real-life phenomena over time. Speaking of mathematical finance, of particular interest is the family of semimartingales, which is theoretically appealing as it satisfies a certain condition on the absence of arbitrage in financial markets and yet is rich enough to reproduce stylized facts from empirical finance such as volatility clustering, leverage effects or jumps. For this reason, the development of statistical tools modeled by discretely observed Itō semimartingales has been a major topic over the last years, both regarding the estimation of crucial quantities used for model calibration purposes and with a view on tests to check whether a certain model fits the data well. For a detailed overview of the state of the art we refer to the recent monographs by [4] and [1].

In the following, we are interested in the evolution of the jump behaviour over time in a completely non-parametric setting where we assume only structural

*michael.hoffmann@rub.de

conditions on the characteristic triplet of the underlying Itō semimartingale. To be precise, let $X = (X_t)_{t\geq 0}$ be an Itō semimartingale with a decomposition

$$X_t = X_0 + \int_0^t b_s\,ds + \int_0^t \sigma_s\,dW_s + \int_0^t \int_{\mathbb{R}} z1_{\{|z|\leq 1\}}(\mu - \bar{\mu})(ds,dz) + \int_0^t \int_{\mathbb{R}} z1_{\{|z|>1\}}\mu(ds,dz), \quad (1.1)$$

where W is a standard Brownian motion, μ is a Poisson random measure on $\mathbb{R}_+ \times \mathbb{R}$, and the predictable compensator $\bar{\mu}$ satisfies $\bar{\mu}(ds,dz) = ds\,\nu_s(dz)$. The main quantity of interest is the kernel ν_s which controls the number and the size of the jumps around time s. In [2] the authors are interested in the detection of abrupt changes in the jump measure of X. Based on high-frequency observations $X_{i\Delta_n}$, $i = 0,\dots,n$, with $\Delta_n \to 0$ they construct a test for a constant ν against the alternative

$$\nu_s^{(n)} = 1_{\{s<\lfloor n\theta_0\rfloor\Delta_n\}}\nu_1 + 1_{\{s\geq\lfloor n\theta_0\rfloor\Delta_n\}}\nu_2.$$

In the sequel, we will deal with gradual (smooth, continuous) changes of ν_s which basically means that ν_s is a non-constant function in $s \in \mathbb{R}_+$. We discuss how and how well the first point in time where the jump behaviour changes (gradually) can be estimated. To this end, we introduce the formal setup in Section 2 where we also define a measure of time variation which is used to detect changes in the jump characteristic. Section 3 is concerned with weak convergence of a standardized version of an estimator for this measure. In Section 4 we use this result to derive an estimator of the first change point for the jump behaviour. The proofs of the results presented in this paper can be found in [3].

2 The basic assumptions and a measure of gradual changes

In the sequel let $X^{(n)} = (X_t^{(n)})_{t\geq 0}$ be an Itō semimartingale of the form (1.1) with characteristic triplet $(b_s^{(n)}, \sigma_s^{(n)}, \nu_s^{(n)})$ for each $n \in \mathbb{N}$. We are interested in investigating gradual changes in the evolution of the jump behaviour and we assume throughout this paper that there is a driving law behind this evolution which is common for all $n \in \mathbb{N}$. Formally, we introduce a transition kernel $g(y,dz)$ from $([0,1],\mathbb{B}([0,1]))$ into $(\mathbb{R},\mathbb{B})$ such that

$$\nu_s^{(n)}(dz) = g\Big(\frac{s}{n\Delta_n}, dz\Big)$$

for $s \in [0, n\Delta_n]$. This transition kernel shall be an element of the set $\mathcal{G}$ to be defined below. Throughout the paper $\mathbb{B}(A)$ denotes the trace σ-algebra on a set $A \subset \mathbb{R}$ with respect to the Borel σ-algebra $\mathbb{B}$ of $\mathbb{R}$.

Assumption 1. *Let $\mathcal{G}$ denote the set of all transition kernels $g(\cdot, dz)$ from $([0,1], \mathbb{B}([0,1]))$ into $(\mathbb{R}, \mathbb{B})$ such that*

(1) For each $y \in [0,1]$ the measure $g(y, dz)$ does not charge $\{0\}$, i.e. $g(y, \{0\}) = 0$.

(2) The function $y \mapsto \int (1 \wedge z^2) g(y, dz)$ is bounded on the interval $[0,1]$.

(3) If

$$\mathcal{I}(z) := \begin{cases} [z, \infty), & \text{for } z > 0 \\ (-\infty, z], & \text{for } z < 0 \end{cases}$$

denotes one-sided intervals and

$$g(y,z) := g(y, \mathcal{I}(z)) = \int_{\mathcal{I}(z)} g(y, dx); \quad (y,z) \in [0,1] \times \mathbb{R} \setminus \{0\},$$

then for every $z \in \mathbb{R} \setminus \{0\}$ there exists a finite set $M^{(z)} = \{t_1^{(z)}, \dots, t_{n_z}^{(z)} \mid n_z \in \mathbb{N}\} \subset [0,1]$, such that the function $y \mapsto g(y,z)$ is continuous on $[0,1] \setminus M^{(z)}$.

(4) For each $y \in [0,1]$ the measure $g(y, dz)$ is absolutely continuous with respect to the Lebesgue measure with density $z \mapsto h(y,z)$, where the measurable function $h \colon ([0,1] \times \mathbb{R}, \mathbb{B}([0,1]) \otimes \mathbb{B}) \to (\mathbb{R}, \mathbb{B})$ is continuously differentiable with respect to $z \in \mathbb{R} \setminus \{0\}$ for fixed $y \in [0,1]$. The function $h(y,z)$ and its derivative will be denoted by $h_y(z)$ and $h_y'(z)$, respectively. Furthermore, we assume for each $\varepsilon > 0$ that

$$\sup_{y \in [0,1]} \sup_{z \in M_\varepsilon} \Big(h_y(z) + |h_y'(z)| \Big) < \infty,$$

where $M_\varepsilon = (-\infty, -\varepsilon] \cup [\varepsilon, \infty)$.

In order to investigate gradual changes in the jump behaviour of the underlying process we follow [5] and consider a measure of time variation for the jump behaviour which is defined by

$$D(\zeta, \theta, z) := \int_0^\zeta g(y,z)dy - \frac{\zeta}{\theta} \int_0^\theta g(y,z)dy, \tag{2.1}$$

where $(\zeta, \theta, z) \in C \times \mathbb{R} \setminus \{0\}$ and

$$C := \{(\zeta, \theta) \in [0,1]^2 \mid \zeta \le \theta\}. \tag{2.2}$$

Here and throughout this paper we use the convention $\frac{0}{0} := 1$. The time varying measure defined in (2.1) is indeed suitable for the detection of gradual changes in the jump characteristic of the underlying process, because one can show that the jump behaviour corresponding to the first $\lfloor n\theta \rfloor$ observations is identical for some $\theta \in [0,1]$ if and only if $D(\zeta, \theta, z) \equiv 0$ for all $0 \leq \zeta \leq \theta$ and $z \in \mathbb{R} \setminus \{0\}$ (see [3]).

We conclude this section with the main assumption for the characteristics of an Itō semimartingale which will be used throughout this paper.

Assumption 2. *For each $n \in \mathbb{N}$ let $X^{(n)}$ denote an Itō semimartingale of the form (1.1) with characteristics $(b_s^{(n)}, \sigma_s^{(n)}, \nu_s^{(n)})$ defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ that satisfies*

(a) There exists a $g \in \mathcal{G}$ such that

$$\nu_s^{(n)}(dz) = g\Big(\frac{s}{n\Delta_n}, dz\Big)$$

holds for all $s \in [0, n\Delta_n]$ and all $n \in \mathbb{N}$.

(b) The drift $b_s^{(n)}$ and the volatility $\sigma_s^{(n)}$ are predictable processes and satisfy

$$\sup_{n\in\mathbb{N}} \sup_{s\in\mathbb{R}_+} \Big(\mathbb{E}|b_s^{(n)}|^\alpha \vee \mathbb{E}|\sigma_s^{(n)}|^p\Big) < \infty,$$

for some $p > 2$, with $\alpha = 3p/(p+4)$.

(c) The observation scheme $\{X_{i\Delta_n}^{(n)} \mid i = 0, \dots, n\}$ satisfies

$$\Delta_n \to 0, \quad n\Delta_n \to \infty, \quad \text{and} \quad n\Delta_n^{1+\tau} \to 0,$$

for $\tau = (p-2)/(p+1) \in (0,1)$.

3 An estimator for the measure of time variation and weak convergence

In order to estimate the measure of time variation introduced in (2.1) we use the sequential empirical tail integral process defined by

$$U_n(\theta, z) = \frac{1}{n\Delta_n} \sum_{j=1}^{\lfloor n\theta \rfloor} 1_{\{\Delta_j^n X^{(n)} \in \mathcal{I}(z)\}},$$

where $\Delta_j^n X^{(n)} = X^{(n)}_{j\Delta_n} - X^{(n)}_{(j-1)\Delta_n}$, $\theta \in [0,1]$ and $z \in \mathbb{R} \setminus \{0\}$. An estimate for the measure of time variation defined in (2.1) is then given by

$$\mathbb{D}_n(\zeta,\theta,z) := U_n(\zeta,z) - \frac{\zeta}{\theta} U_n(\theta,z), \quad (\zeta,\theta,z) \in C \times \mathbb{R} \setminus \{0\}, \tag{3.1}$$

where the set C is defined in (2.2). The following theorem establishes consistency of $\mathbb{D}_n$ as it shows weak convergence of the process

$$\mathbb{H}_n(\zeta,\theta,z) := \sqrt{n\Delta_n}(\mathbb{D}_n(\zeta,\theta,z) - D(\zeta,\theta,z)). \tag{3.2}$$

with values in $\ell^\infty(B_\varepsilon)$, where $B_\varepsilon = C \times M_\varepsilon$.

Theorem 1. *If Assumption 2 is satisfied, then the process* $\mathbb{H}_n$ *defined in* (3.2) *satisfies* $\mathbb{H}_n \rightsquigarrow \mathbb{H}$ *in* $\ell^\infty(B_\varepsilon)$ *for any* $\varepsilon > 0$*, where* $\mathbb{H}$ *is a tight mean zero Gaussian process with covariance function*

$$\begin{aligned}
&\mathrm{Cov}(\mathbb{H}(\zeta_1,\theta_1,z_1),\mathbb{H}(\zeta_2,\theta_2,z_2)) = \\
&= \int_0^{\zeta_1\wedge\zeta_2} g(y,\mathcal{I}(z_1)\cap\mathcal{I}(z_2))dy - \frac{\zeta_1}{\theta_1}\int_0^{\zeta_2\wedge\theta_1} g(y,\mathcal{I}(z_1)\cap\mathcal{I}(z_2))dy \\
&\quad - \frac{\zeta_2}{\theta_2}\int_0^{\zeta_1\wedge\theta_2} g(y,\mathcal{I}(z_1)\cap\mathcal{I}(z_2))dy + \frac{\zeta_1\zeta_2}{\theta_1\theta_2}\int_0^{\theta_1\wedge\theta_2} g(y,\mathcal{I}(z_1)\cap\mathcal{I}(z_2))dy.
\end{aligned}$$

4 A consistent estimator for the gradual change point

If one defines

$$\mathcal{D}^{(\varepsilon)}(\theta) := \sup_{|z|\geq\varepsilon} \sup_{0\leq\zeta\leq\theta'\leq\theta} |D(\zeta,\theta',z)|,$$

for some pre-specified constant $\varepsilon > 0$, one can characterize the existence of a change point as follows: There exists a gradual change in the behaviour of the jumps larger than ε of the process (1.1) if and only if $\mathcal{D}^{(\varepsilon)}(1) > 0$. Our aim is to construct an estimator for the first point where the jump behaviour changes (gradually). For this purpose we define

$$\theta_0^{(\varepsilon)} := \inf\left\{\theta \in [0,1] \mid \mathcal{D}^{(\varepsilon)}(\theta) > 0\right\},$$

where we set $\inf\emptyset := 1$. We call $\theta_0^{(\varepsilon)}$ the change point of the jumps larger than ε of the underlying process (1.1). Intuitively, the estimation of $\theta_0^{(\varepsilon)}$ becomes more difficult the flatter the curve $\theta \mapsto \mathcal{D}^{(\varepsilon)}(\theta)$ is at $\theta_0^{(\varepsilon)}$. Therefore, we describe

the curvature of $\theta \mapsto \mathcal{D}^{(\varepsilon)}(\theta)$ by a local polynomial behaviour of the function $\mathcal{D}^{(\varepsilon)}(\theta)$ for values $\theta > \theta_0^{(\varepsilon)}$. More precisely, we assume throughout this section that $\theta_0^{(\varepsilon)} < 1$ and that there exist constants $\lambda, \eta, \varpi, c^{(\varepsilon)} > 0$ such that $\mathcal{D}^{(\varepsilon)}$ admits an expansion of the form

$$\mathcal{D}^{(\varepsilon)}(\theta) = c^{(\varepsilon)}\big(\theta - \theta_0^{(\varepsilon)}\big)^{\varpi} + \aleph(\theta) \tag{4.1}$$

for all $\theta \in [\theta_0^{(\varepsilon)}, \theta_0^{(\varepsilon)} + \lambda]$, where the remainder term satisfies $|\aleph(\theta)| \le K\big(\theta - \theta_0^{(\varepsilon)}\big)^{\varpi+\eta}$ for some $K > 0$. By Theorem 1 the process $\mathbb{D}_n(\zeta, \theta, z)$ from (3.1) is a consistent estimator of $D(\zeta, \theta, z)$. Therefore we set

$$\mathbb{D}_n^{(\varepsilon)}(\theta) := \sup_{|z| \ge \varepsilon} \sup_{0 \le \zeta \le \theta' \le \theta} |\mathbb{D}_n(\zeta, \theta', z)|.$$

The construction of an estimator for $\theta_0^{(\varepsilon)}$ utilizes the fact that $(n\Delta_n)^{1/2}\mathbb{D}_n^{(\varepsilon)}(\theta) \to \infty$ in probability for any $\theta \in (\theta_0^{(\varepsilon)}, 1]$. Moreover, for $\theta \in [0, \theta_0^{(\varepsilon)}]$ we have $(n\Delta_n)^{1/2}\mathbb{D}_n^{(\varepsilon)}(\theta) = O_{\mathbb{P}}(1)$ since this quantity converges weakly. Therefore, we consider the statistic

$$r_n^{(\varepsilon)}(\theta) := 1_{\{(n\Delta_n)^{1/2}\mathbb{D}_n^{(\varepsilon)}(\theta) \le \varkappa_n\}},$$

for a deterministic sequence $\varkappa_n \to \infty$. From the previous discussion we expect

$$r_n^{(\varepsilon)}(\theta) \to \begin{cases} 1, & \text{if } \theta \le \theta_0^{(\varepsilon)} \\ 0, & \text{if } \theta > \theta_0^{(\varepsilon)} \end{cases}$$

in probability if the threshold level $\varkappa_n$ is chosen appropriately. Consequently, we define the estimator for the change point by

$$\hat{\theta}_n^{(\varepsilon)} = \hat{\theta}_n^{(\varepsilon)}(\varkappa_n) := \int_0^1 r_n^{(\varepsilon)}(\theta) d\theta.$$

The following result establishes consistency of the estimator $\hat{\theta}_n^{(\varepsilon)}$ under rather mild assumptions on the sequence $(\varkappa_n)_{n\in\mathbb{N}}$.

Theorem 2. *If Assumption 2 is satisfied, $\theta_0^{(\varepsilon)} < 1$, and (4.1) holds for some $\varpi > 0$, then*

$$\hat{\theta}_n^{(\varepsilon)} - \theta_0^{(\varepsilon)} = O_{\mathbb{P}}\Big(\big(\frac{\varkappa_n}{\sqrt{n\Delta_n}}\big)^{1/\varpi}\Big),$$

for any sequence $\varkappa_n \to \infty$ with $\varkappa_n/\sqrt{n\Delta_n} \to 0$.

Theorem 2 makes the heuristic argument above more precise. A lower degree of smoothness in $\theta_0^{(\varepsilon)}$ yields a better rate of convergence of the estimator. Moreover, the slower the threshold level $\varkappa_n$ converges to infinity the better the rate of convergence. In [3] the authors discuss a data-driven choice of the threshold $\varkappa_n$ for which the probability for over- and underestimation of $\theta_0^{(\varepsilon)}$ can be controlled.

Acknowledgements: This work has been supported in part by the Collaborative Research Center "Statistical modeling of nonlinear dynamic processes" (SFB 823, Projects A1 and C1) and the Research Training Group "High-dimensional Phenomena in Probability - Fluctuations and Discontinuity" (RTG 2131) of the German Research Foundation (DFG).

References

[1] Y. Aït-Sahalia and J. Jacod. *High-Frequency Financial Econometrics*. Princeton University Press, 2014.

[2] A. Bücher, M. Hoffmann, M. Vetter, and H. Dette. Nonparametric tests for detecting breaks in the jump behaviour of a time-continuous process. *Bernoulli*, 23(2):1335–1364, 2017.

[3] M. Hoffmann, M. Vetter, and H. Dette. Nonparametric inference of gradual changes in the jump behaviour of time-continuous processes. *submitted to: Stochastic Processes and their Applications*, 2017. arXiv: 1704.04040.

[4] J. Jacod and P. Protter. *Discretization of Processes*. Springer, 2012.

[5] M. Vogt and H. Dette. Detecting gradual changes in locally stationary processes. *The Annals of Statistics*, 43(2):713–740, 2015.

Can humans be replaced by computers in taxa recognition?

Johanna Ärje*[1], Ville Tirronen[2], Jenni Raitoharju[3], Kristian Meissner[4], and Salme Kärkkäinen[5]

[1,5]*Department of Mathematics and Statistics, University of Jyvaskyla*
[2]*Department of Mathematical Information Technology, University of Jyvaskyla*
[3]*Department of Signal Processing, Tampere University of Technology*
[4]*Finnish Environment Institute*

Biomonitoring of waterbodies is vital as the number of anthropogenic stressors on aquatic ecosystems keeps growing. However, the continuous decrease in funding makes it impossible to meet monitoring goals or sustain traditional manual sample processing. In this paper, we review what kind of statistical tools can be used to enhance the cost efficiency of biomonitoring: We explore automated identification of freshwater macroinvertebrates which are used as one indicator group in biomonitoring of aquatic ecosystems. We present the first classification results of a new imaging system producing multiple images per specimen. Moreover, these results are compared with the results of human experts. On a data set of 29 taxonomical groups, automated classification produces a higher average accuracy than human experts.

Keywords: Biomonitoring, classification, image analysis, macroinvertebrates.

1 Introduction

Benthic macroinvertebrates are a diverse group of species that quickly react to changes in their environment [15]. Their community composition can reflect even subtle human-induced changes in their environment, making them an ideal indicator group for aquatic biomonitoring [7]. In many countries, biomonitoring of benthic macroinvertebrates is a key part of ecological status assessment of surface waters required by the European Union's Water Framework Directive [17].

The traditional process of macroinvertebrate biomonitoring is the following: First, macroinvertebrates are sampled, usually by using a kick-net method. Second, the specimens are sorted out from the detritus and identified manually

*Corresponding author: johanna.arje@jyu.fi

by an expert. Third, the observed taxa abundancies are used to calculate biological indices indicating changes compared to previous sampling or a reference community. Finally, the index values are combined to evaluate the ecological status of the sampled waterbody.

In macroinvertebrate biomonitoring a large proportion of the total cost and time is spent on manual identification by highly trained experts. It takes several years to train an expert and manually identifying a sample of few thousand individuals can take hours. The monitoring process could be expedited substantially by shifting to automated identification and in recent years there have been many studies on the automated identification of benthic macroinvertebrates [10, 8, 6, 2, 1]. Many biologists tend to oppose the shift to automated identification of macroinvertebrates due to fear of it not being accurate enough. However, manual identification has been found to be surprisingly error prone as well [4]. While there exist studies on the automated classification of macroinvertebrates, to our knowledge, none of them include a comparison between manual and automated identification accuracy.

In this article we introduce a new imaging system producing multiple images per specimen and present classification results on the new multiple image data base. We also compare the accuracy of automated classification to that of human experts.

2 Automated classification

There has been increasing interest in automated classification of benthic macroinvertebrates as continuing budget cuts disable the use of manual identification. In order to use automated classification, the specimens need to be imaged onto a computer and the classification methods need to be trained with data first keyed traditionally by several taxonomic experts. In our analyses, we have used both single image data and multiple image data.

2.1 Single image data

In the first phase of the study, the specimens were scanned onto a computer in single taxon batches using VueScan(c) software (http://www.hamrick.com/, Phoenix, Arizona, USA) with an HP Scanjet4850 flatbed scanner at an optical resolution of 2400 d.p.i. The scanned images were normalized to the same intensity range and color balance by using a calibration target. Individual specimens were segmented from the batch image, and each specimen was saved as a single posture image. A set of 64 simple geometry and intensity-based features were extracted for each specimen from the single posture images using

ImageJ [14]. A large data set of 35 taxonomical groups and 6418 specimens was imaged by the Finnish Environment Institute.

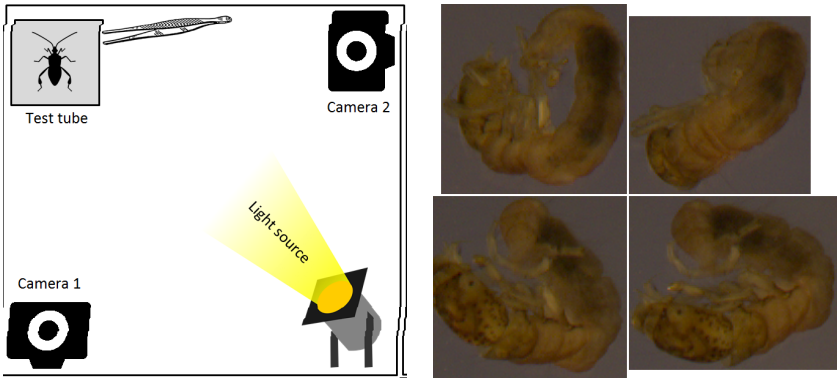
There exists a vast amount of different algorithms and models for classification. For the single image data, we compared a group of classification methods and presented a novel Bayesian classifier, RBA, that achieved classification accuracy of 81.2 % [2]. In another work, we used the image data with few modifications: With one taxon excluded from the study and four taxa combined into two, the data comprised of 32 taxonomic classes [1]. In addition to previously explored classifiers, a kernel extension of Extreme Learning Machine [5] was employed and it achieved the highest classification accuracy of 84.1 %.

2.2 Multiple image data

In the second phase of the study, a new imaging system was built to enable multiple images per specimen. The system is described in Figure 1a. It consists of two Basler ACA1920-155UC cameras (frame rate of 150 fps) with Megapixel Macro Lens (f=75mm, F:3.5-CWD<535mm) and a high power LED light. The cameras are placed at a 90 degree angle to each other to ensure multiple postures of each specimen. The software builds a model of the background and sets off the cameras when a significant change in the view of the camera is detected. A specimen is dropped into a cuvette filled with alcohol. As it sinks, both cameras take multiple shots of it and the resulting images are stored onto a computer (See example images in Figure 1b). The number of images per specimen depends on the size and weight of each specimen: Heavier specimen sink to the bottom of the cuvette faster, leading to a smaller number of images.

Using the described imaging device, the Finnish Environment Institute compiled an image data base of 126 lotic freshwater macroinvertebrate taxa and over 2.6 million images. For the current work, we restricted the number of classes to 29 taxa present at a human proficiency test to compare the classification results with human experts. We also restricted the number of images per specimen to a maximum of 50 images for computational reasons. If a specimen had more images from both cameras combined, we randomly selected 50 of them. The final data comprises of 7742 observations and a total of 367341 images. Using ImageJ, the same set of 64 features was extracted from the images as for the phase one data.

With the 64 features, extracted from the multiple image data, we explored automated classification using MLP, RF and SVM. We split the observations randomly into training (70 %), validation (10 %) and test (20 %) data 10 times. With each data split, we used the training data to build the model and the validation data to select optimal parameter values. Once the parameters were fixed, the training and validation data were combined to train the model



(a) (b)

Figure 1: (a): Schematic of the imaging system for macroinvertebrates pictured from above. (b): Example images of a *Polycentropus flavomaculatus* specimen from two cameras. The top row images are from camera 1 and the bottom row images from camera 2.

again. Each image of the test data was classified and the final class for each observation was based on majority vote among all the images of the specimen. All the models were built using R [13]. The results are shown in Table 1.

The highest classification accuracy is achieved with SVM. *Protonemura sp.*, *Hydropsyche saxonica*, *Diura sp.* and *Capnopsis schilleri* have high error rates due to a low number of observations in the training data. The hardest taxa to identify with adequate amount of training data are *Baetis vernus group* which is usually confused with *Baetis rhodani*, and *Kageronia fuscogrisea* and *Polycentropus irroratus* that are confused with several other taxa.

Table 1: Classification accuracy for multiple image data of 29 taxonomic classes. The means and standard deviations are computed over ten splits into training (80 %) and test (20 %) data.

Classifier	$\overline{acc}$	$sd(acc)$
RF	0.713	0.012
MLP	0.770	0.011
SVM	0.865	0.006

The classification results presented in Table 1 were obtained with very simple features and higher classification accuracy could be obtained using more refined features. In fact, even a simple principal component transformation that makes the features uncorrelated already slightly improves the classification accuracy for SVM ($\overline{acc} = 87.4\%, sd(acc) = 0.005$) and MLP ($\overline{acc} = 79.7\%, sd(acc) = 0.011$). With a convolutional neural network [CNN, 9] that uses the original images as input instead of features, the classification accuracy is even higher. We applied the MatConvNet [16] implementation of the Alexnet model with batch size 256 and 60 training epochs in Matlab [11] and achieved an average classification accuracy of 93.4 % ($sd(acc) = 0.006$).

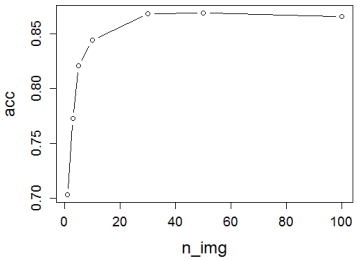


Figure 2: Classification accuracy of SVM plotted against the maximum number of images per specimen.

We also studied, how the maximum number of images per observation affects the classification accuracy with SVM. From Figure 2, it is clear that accuracy increases with the number of images per specimen. However, the difference in accuracy between a maximum of 30, 50 or 100 images is quite small while the difference in computational costs is much greater. It is crucial to consider both when deciding on the number of images to use per specimen. The classification accuracy achieved for the multiple image data presented here is not directly comparable with the results for the single image data of Section 2.1 as the data sets have only 14 taxa in common. However, from Figure 2 it is evident that having more than one image per specimen clearly improves the classification accuracy.

3 Manual classification

In order to compare automated and manual classification, we need classification results on the same set of taxa for both. The Finnish Environment institute

organized a proficiency test on taxonomic identification of boreal freshwater lotic, lentic, profundal and North-Eastern Baltic benthic macroinvertebrates in March 2016. The aim of the test was to assess the reliability of professional and semi-professional identification of macroinvertebrate taxa routinely encountered during North-Eastern Baltic coastal or boreal lake and river monitoring [12]. A part of the proficiency test included 10 experts each identifying 50 specimens of lotic freshwater macroinvertebrates belonging to a total of 46 taxonomic groups, of which 29 are in common with the multiple image data introduced in Section 2.2. The average accuracy for the 46 taxa data was 93.2 % ($sd = 0.061$) and for the 29 taxa in common with the image data, the average accuracy was 92.7 % ($sd = 0.064$). The hardest taxon to be identified was *Hydropsyche saxonica* as half of the specimen were confused with *Hydropsyche angustipennis*.

4 Discussion

The mismatch between funding and biomonitoring goals calls for more efficient monitoring processes. One way to lower the cost of macroinvertebrate biomonitoring is to shift from manual to automated identification of samples. The material costs of the imaging system described in Section 2.2 are approximately 4-5K €, while the price of a high quality stereo microscope traditionally used for macroinvertebrate identification is twice as much. The imaging system is more affordable and it fits well into the work flow of sample processing. Whether using manual or automated identification, the sampled specimens need to be sorted from the detritus. As a natural extension of this, the operator can drop a specimen into the cuvette of the imaging system before placing it into a vial for storing.

Automated classification can enhance the cost-effeciency of the macroinvertebrate sample processing also due to its speed. Training a human expert takes years while – depending on the choice of a classifier and the size of the training data – training a classification model can take 1–5 hours. Predicting taxonomic groups for a sample of 1600 specimens only takes a few minutes of computing compared to the 1–12 hours of manual labor. Also, once a classifier is trained, using it does not require expertise.

Of course, the viability of shifting to automated classification depends on the classification accuracy above all. In this paper, we presented classification results on an image data comprising 29 taxa also present at a human expert proficiency test. The achieved classification accuracy (87.4 % for SVM and 93.4 % for CNN) is in the range of human accuracy of the proficiency test ($82.4 - 100\%$) with the same taxa. The proficiency test included one or two specimen per taxon for each participant while the image test data for automated

classification comprised a total 1557 observations. When using the exact same amount of observations per taxon for testing as in the proficiency test, the classification accuracy for automated classifiers decreases but this is due to the fact that the image data is not a balanced data set and some of the 29 taxa have very few observations for training.

In order to adopt the automated identification process in practice, we need to achieve similarly high classification accuracy with a larger number of taxonomic groups. In this paper, we restricted the taxa to 29 in order to provide a comparison to human experts. Typically, 30–75 macroinvertebrate taxa are encountered at individual sites. While there is need for extending the classifiers to more taxa, the results so far are very promising.

Acknowledgements: We thank the Academy of Finland for the grants of Ärje (284513, 289076), Tirronen (289076, 289104) Kärkkäinen (289076), Meissner (289104) and Raitoharju (288584). The authors would like to thank CSC for computational resources. We kindly thank Moncef Gabbouj, Alexandros Iosifidis and Serkan Kiranyaz for collaboration.

References

[1] J. Ärje, S. Kärkkäinen, K. Meissner, A. Iosifidis, T. Ince, M. Gabbouj, and S. Kiranyaz. The effect of automated taxa identification errors on biological indices. *Expert systems with applications*, 72:108–120, 2017.

[2] J. Ärje, S. Kärkkäinen, T. Turpeinen, and K. Meissner. Breaking the curse of dimensionality in quadratic discriminant analysis models with a novel variant of a bayes classifier enhances automated taxa identification of freshwater macroinvertebrates. *Environmetrics*, 24(4):248–259, 2013.

[3] P. Haase, S. U. Pauls, K. Schindehütte, and A. Sunderman. First audit of macroinvertebrate samples from an EU Water Framework Directive monitoring program: human error greatly lowers precision of assessment results. *Journal of the North American Benthological Society*, 29(4):1279–1291, 2010.

[4] A. Iosifidis, A. Tefas, and I. Pitas. Graph embedded extreme learning machine. *IEEE Transactions on Cybernetics*, 2015. D.O.I. 10.1109/TCYB.2015.2401973.

[5] H. Joutsijoki, K. Meissner, M. Gabbouj, S. Kiranyaz, J. Raitoharju, J. Ärje, S. Kärkkäinen, V. Tirronen, T. Turpeinen, and M. Juhola. Evaluating the

performance of artificial neural networks for the classification of freshwater benthic macroinvertebrates. *Ecological Informatics*, 20:1–12, 2014.

[6] J. R. Karr and E. W. Chu. Sustaining living rivers. *Hydrobiologia*, 422/423:1–14, 2000.

[7] S. Kiranyaz, T. Ince, J. Pulkkinen, M. Gabbouj, J. Ärje, S. Kärkkäinen, V. Tirronen, M. Juhola, T. Turpeinen, and K. Meissner. Classification and retrieval on macroinvertebrate image databases. *Computers in Biology and Medicine*, 41(7):463–472, 2011.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105, 2012.

[9] D. A. Lytle, G. Martínez-Muñoz, W. Zhang, N. Larios, L. Shapiro, R. Paasch, A. Moldenke, E. N. Mortensen, S. Todorovic, and T. G. Dietterich. Automated processing and identification of benthic invertebrate samples. *Journal of the North American Benthological Society*, 29(3):867–874, 2010.

[10] MATLAB. *version 7.10.0 (R2010a)*. The MathWorks Inc., Natick, Massachusetts, 2010.

[11] K. Meissner, H. Nygård, K. Björklöf, M. Jaale, M. Hasari, L. Laitila, J. Rissanen, and M. Leivuori. Proficiency test 04/2016: Taxonomic identification of boreal freshwater lotic, lentic, profundal and north-eastern baltic benthic macroinvertebrates. *Reports of the Finnish Environment Institute*, 2, 2017.

[12] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[13] W. S. Rasband. *ImageJ*. U.S. National Institutes of Health, Bethesda, Maryland, USA, 1997-2010.

[14] D. M. Rosenberg and V. H. Resh, editors. *Freshwater Biomonitoring and Benthic Macroinvertebrates*. Chapman & Hall, 1993.

[15] A. Vedaldi and K. Lenc. MatConvNet: Convolutional neural networks for matlab. In *International Conference on Multimedia*, pages 689–692, 2015.

[16] WFD. Directive 2000/60/EC of the European Parliament and the Council of 23, October 2000. A framework for community action in the field of water policy. *Off. J. Eur. Commun.*, L327:72, 2000.

AIC post-selection inference in linear regression

Ali Charkhi*[1] and Gerda Claeskens[1]

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

Post-selection inference has been considered a crucial topic in data analysis. In this article, we develop a new method to obtain correct inference after model selection by the Akaike's information criterion [1] in linear regression models. Confidence intervals can be calculated by incorporating the randomness of the model selection in the distribution of the parameter estimators which act as pivotal quantities. Simulation results show the accuracy of the proposed method.

Keywords: Post-selection inference; Confidence intervals; Akaike information criterion.

1 Introduction

Consider the linear regression setting where the true model is of the form

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$ and we assume that σ^2 is known. For a given predictor matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) \in \mathbb{R}^{n \times p}$, we wish to model $\boldsymbol{\mu}$ by a linear function of all predictors, $\boldsymbol{X}\boldsymbol{b}$, or just a subset of predictors, $\boldsymbol{X}_M \boldsymbol{b}_M$, where $\boldsymbol{X}_M$ contains as columnsthe predictors with indices in $M \subseteq \{1, \ldots, p\}$. This setting can be considered as a nonparametric setting because there is no assumption about whether the true model is also linear for a true coefficients vector $\boldsymbol{\beta}^0$. The least squares estimator in linear regression is defined as $\widehat{\boldsymbol{\beta}}_M = (\boldsymbol{X}_M^t \boldsymbol{X}_M)^{-1} \boldsymbol{X}_M^t \boldsymbol{Y}$ which minimizes the expected squared error. In other words, $\widehat{\boldsymbol{\beta}}_M$ is the estimator of $\boldsymbol{\beta}_M = (\boldsymbol{X}_M^t \boldsymbol{X}_M)^{-1} \boldsymbol{X}_M^t \boldsymbol{\mu}$.

Regarding the inference, one can easily use classical confidence intervals (in any submodel) based on the normality of the observations. The difficulty arises when one selects a model based on a criterion from a collection of potential

*Corresponding author: ali.charkhi@kuleuven.be

models $\mathcal{M}$ and wants to do inference for the parameters in the selected model. Since this selection is data-driven, it is random. Ignoring this randomness may lead to incorrect inference. One way to incorporate the selection randomness in inference is using conditional inference, by conditioning on the selected model.

When one imposes the assumption that there exist a true model with parameters $\boldsymbol{\beta}^0$, [7, 8] showed that the distribution of a post-selection estimator can not be estimated in a uniform way. Considering model (1), [2] proposed a method to calculate confidence intervals which are valid irrespective of the selection criterion (Posi method), hence their confidence intervals are conservative for a specific model selection criterion. Their confidence intervals are for parameters in the selected model rather than the true value of the parameters. [6] studied post-selection inference for lasso in high dimensional data. [9] generalized the results to sequential regression procedures such as forward stepwise regression and least angle regression. [3] used the asymptotic distribution to calculate confidence intervals for the model parameters in general likelihood models when they assumed that there exits a true model (Asymp-AIC method).

In this article, we study post-selection inference for the population parameters after using AIC for model selection without assuming a true model to exist. Assuming σ^2 is known, AIC for model M is defined as

$$AIC(M) = \|\boldsymbol{Y} - \boldsymbol{X}_M\widehat{\boldsymbol{\beta}}_M\|^2 + 2\sigma^2|M|. \tag{2}$$

Knowledge about σ^2 may seem restrictive, but [5] showed that in this setting inference without knowing σ^2 is impossible. The main reason is that taking the variance estimation into account leads to insufficient information about the parameters for inference. Our simulations show that even we estimate the σ^2 using the same data, the results are still valid. When σ^2 is unknown, the AIC score for each model is different from the score in (2). In that case, one estimates σ^2 within each model by $\widehat{\sigma}^2 = \|\boldsymbol{Y} - \boldsymbol{X}_M\widehat{\boldsymbol{\beta}}_M\|^2/n$ which leads to the following formula for AIC for model M:

$$AIC(M, \sigma^2) = \log(\|\boldsymbol{Y} - \boldsymbol{X}_M\widehat{\boldsymbol{\beta}}_M\|^2) + \frac{2(|M|+1)}{n}. \tag{3}$$

In a set of models $\mathcal{M}$ the model with the smallest value of (2), or (3), is the best model according to AIC in the considered case.

2 Post-selection inference

When AIC selects a model, it defines an event which we call the *selection event*. If AIC selects model M, i.e. $M_{aic} = M$, then

$$AIC(M) \leq AIC(M_i), \qquad \forall M_i \in \mathcal{M}.$$

Define $\boldsymbol{P}_M = \boldsymbol{X}_M(\boldsymbol{X}_M^t\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M^t$. Using (2) we can represent the selection event as

$$\begin{aligned}\mathcal{S}_M(\mathcal{M}) &= \bigcap_{M_i\in\mathcal{M}} \left\{ \|(\boldsymbol{I}_n - \boldsymbol{P}_{M_i})\boldsymbol{Y}\|^2 + 2\sigma^2|M_i| - \|(\boldsymbol{I}_n - \boldsymbol{P}_M)\boldsymbol{Y}\|^2 - 2\sigma^2|M| \geq 0 \right\} \\ &= \bigcap_{M_i\in\mathcal{M}} \left\{ \boldsymbol{Y}^t(\boldsymbol{P}_M - \boldsymbol{P}_{M_i})\boldsymbol{Y} - 2\sigma^2(|M| - |M_i|) \geq 0 \right\}. \qquad (4)\end{aligned}$$

Similarly when (3) is used for selection, the event can be expressed as

$$\begin{aligned}\mathcal{S}_M^{\sigma^2}(\mathcal{M}) &= \bigcap_{M_i\in\mathcal{M}} \left\{ \log\left(\frac{\left\|\boldsymbol{Y} - \boldsymbol{X}_{M_i}\widehat{\boldsymbol{\beta}}_{M_i}\right\|^2}{\left\|\boldsymbol{Y} - \boldsymbol{X}_M\widehat{\boldsymbol{\beta}}_M\right\|^2} \right) \geq \frac{2(|M| - |M_i|)}{n} \right\} \\ &= \bigcap_{M_i\in\mathcal{M}} \left\{ \boldsymbol{Y}^t(\boldsymbol{I}_n - \boldsymbol{P}_{M_i})\boldsymbol{Y}\cdot\kappa_{M_i} - \boldsymbol{Y}^t(\boldsymbol{I}_n - \boldsymbol{P}_M)\boldsymbol{Y}\cdot\kappa_M \geq 0 \right\}, \qquad (5)\end{aligned}$$

where $\kappa_{M_i} = \exp\left(2(|M_i|)/n\right)$.

To obtain correct confidence intervals after model selection, we use conditional inference. In other words, for inference for a parameter of the form $\boldsymbol{\eta}_M^t\boldsymbol{\mu}$ in the selected model where $\boldsymbol{\eta}_M \in \mathbb{R}^n$ and is specified, we need to investigate the distribution of $\boldsymbol{\eta}_M^t\boldsymbol{Y} \mid \{M_{aic} = M\}$ which is equivalent to working with

$$\boldsymbol{\eta}_M^t\boldsymbol{Y} \mid \mathcal{S}_M(\mathcal{M}).$$

It is possible to rewrite $\mathcal{S}_M(\mathcal{M})$ in terms of $\boldsymbol{\eta}_M^t\boldsymbol{Y}$. Proofs for the following results can be found in [4].

Lemma 1. *Define $T = \boldsymbol{\eta}_M^t\boldsymbol{Y}$ and $\boldsymbol{Z} = \boldsymbol{Y} - \boldsymbol{w}T$ where $\boldsymbol{w} = \boldsymbol{\eta}_M(\boldsymbol{\eta}_M^t\boldsymbol{\eta}_M)^{-1}$ (T and $\boldsymbol{Z}$ are independent). Then*

$$\begin{aligned}\mathcal{S}_M(\mathcal{M}) = \bigcap_{M_i\in\mathcal{M}} \{ & \quad T^t\boldsymbol{w}^t\boldsymbol{D}_{M_i}\boldsymbol{w}T + 2T^t\boldsymbol{w}\boldsymbol{D}_{M_i}\boldsymbol{Z} \\ & + \boldsymbol{Z}^t\boldsymbol{D}_{M_i}\boldsymbol{Z} - 2\sigma^2(|M| - |M_i|) \geq 0\} \qquad (6)\end{aligned}$$

and

$$\mathcal{S}_M^{\sigma^2}(\mathcal{M}) = \bigcap_{M_i \in \mathcal{M}} \{ \quad T^t \boldsymbol{w}^t \boldsymbol{R}_{M_i} \boldsymbol{w} T \kappa_{M_i} - T^t \boldsymbol{w}^t \boldsymbol{R}_M \boldsymbol{w} T \kappa_M \\ + 2T^t \boldsymbol{w} \boldsymbol{R}_{M_i} \boldsymbol{Z} \kappa_{M_i} - 2T^t \boldsymbol{w} \boldsymbol{R}_M \boldsymbol{Z} \kappa_M \\ + \boldsymbol{Z}^t \boldsymbol{R}_{M_i} \boldsymbol{Z} \kappa_{M_i} - \boldsymbol{Z}^t \boldsymbol{R}_M \boldsymbol{Z} \kappa_M \geq 0 \} \tag{7}$$

where $\boldsymbol{D}_{M_i} = \boldsymbol{P}_M - \boldsymbol{P}_{M_i}$ *and* $\boldsymbol{R}_{M_i} = \boldsymbol{I}_n - \boldsymbol{P}_{M_i}$.

As expressions (6) and (7) show, the selection event can be written via quadratic functions of T. For the selection event in (6), define

$$a_i = \boldsymbol{w}^t \boldsymbol{D}_{M_i} \boldsymbol{w}, \quad b_i = 2\boldsymbol{w} \boldsymbol{D}_{M_i} \boldsymbol{Z}, \quad c_i = \boldsymbol{Z}^t \boldsymbol{D}_{M_i} \boldsymbol{Z} - 2\sigma^2(|M| - |M_i|),$$

and for the selection event in (7),

$$a_i = \boldsymbol{w}^t \boldsymbol{R}_{M_i} \boldsymbol{w} \kappa_{M_i} - \boldsymbol{w}^t \boldsymbol{R}_M \boldsymbol{w} \kappa_M, \quad b_i = 2(\boldsymbol{w} \boldsymbol{R}_{M_i} \boldsymbol{Z} \kappa_{M_i} - \boldsymbol{w} \boldsymbol{R}_M \boldsymbol{Z} \kappa_M), \\ c_i = \boldsymbol{Z}^t \boldsymbol{R}_{M_i} \boldsymbol{Z} \kappa_{M_i} - \boldsymbol{Z}^t \boldsymbol{R}_M \boldsymbol{Z} \kappa_M.$$

For both selection events in (6) and (7), it is obvious that the selection event can be written as

$$\bigcap_{M_i \in \mathcal{M}} \{a_i T^2 + b_i T + c_i \geq 0\}.$$

These inequalities lead to allowable values for T, namely, of the form $I_M^{\boldsymbol{Z}}(\mathcal{M}) = \cup_{i=1}^{l} (a_i(\boldsymbol{Z}), b_i(\boldsymbol{Z}))$. So, the estimator T for the population parameter $\boldsymbol{\eta}^t \boldsymbol{\mu}$ is a normal random variable which is restricted in $I_M^{\boldsymbol{Z}}(\mathcal{M})$.

Denote the standard normal CDF by $\Phi(x)$ and also denote the CDF of a $N(\mu, \sigma^2)$ random variable truncated to $D = \cup_{i=1}^{l} (a_i, b_i)$ by $F(\cdot; \mu, \sigma^2, D)$ which can be written as, for $x \in (a_r, b_r)$

$$F(x; \mu, \sigma^2, D) = \frac{\sum_{i=1}^{r-1} p_i + \Phi((x-\mu)/\sigma) - \Phi((a_r - \mu)/\sigma)}{\sum_{i=1}^{l} p_i}, \tag{8}$$

where $p_i = \Phi((b_i - \mu)/\sigma) - \Phi((a_i - \mu)/\sigma)$. The following result shows how we can use (8) as a pivotal quantity.

Result 1: *Let* $\boldsymbol{\eta} \in \mathbb{R}^n$ *be fixed,* $T = \boldsymbol{\eta}^t \boldsymbol{Y}$ *and the selection event is* $\mathcal{S}_M(\mathcal{M})$*, Then*

$$F\left(T; \boldsymbol{\eta}^t \boldsymbol{\mu}, \sigma^2 \|\boldsymbol{\eta}\|^2, I_M^{\boldsymbol{Z}}(\mathcal{M})\right) \mid \mathcal{S}_M(\mathcal{M}) \sim \ \textit{Unif}\,(0,1). \tag{9}$$

In post-selection inference, we are interested in constructing confidence intervals for parameters in the selected model. We mainly focus on a one-dimensional parameter. For parameters in the selected model, we construct confidence intervals for each parameter separately. In general, for $\boldsymbol{\eta}^t\boldsymbol{\mu} \in \mathbb{R}$ we are interested in obtaining a confidence interval $[L, U]$ such that $P(L \leq \boldsymbol{\eta}^t\boldsymbol{\mu} \leq U | I_M^{\boldsymbol{Z}}(\mathcal{M})) = 1 - \alpha$. We can use (9) to construct confidence intervals based on the method of pivotal quantities.

Result 2 *Let $\boldsymbol{\eta} \in \mathbb{R}^n$ and $T = \boldsymbol{\eta}^t\boldsymbol{Y}$. Define L and U such that*

$$F(T; L, \sigma^2 \|\boldsymbol{\eta}\|^2, I_M^{\boldsymbol{Z}}(\mathcal{M})) = 1 - \frac{\alpha}{2}, \qquad F(T; U, \sigma^2 \|\boldsymbol{\eta}\|^2, I_M^{\boldsymbol{Z}}(\mathcal{M})) = \frac{\alpha}{2},$$

then $[L, U]$ is a confidence interval for $\boldsymbol{\eta}^t\boldsymbol{\mu}$ conditional on $M_{aic} = M$ such that $P(\boldsymbol{\eta}^t\boldsymbol{\mu} \in [L, U] \mid M_{aic} = M) = 1 - \alpha$.

Result 2 is a general result, because $\boldsymbol{\eta} \in \mathbb{R}^n$ can be defined by the user. For instance, considering $\boldsymbol{\eta}^t = \boldsymbol{e}_i(\boldsymbol{X}_M^t\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M^t$ as the direction of interest for inference, Result 2 provides a confidence interval for the ith parameter in the selected model.

If the true model is indeed linear, i.e. there exist a $\boldsymbol{\beta}^0$ such that $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}^0$, and AIC selects a model M which does not contain all non-zero parameters, then $\widehat{\boldsymbol{\beta}}$ is an unbiased estimator not for the true parameters but for

$$\boldsymbol{\beta}_M = \boldsymbol{\beta}^0[M] + (\boldsymbol{X}_M^t\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M^t\boldsymbol{X}_{M^c}\boldsymbol{\beta}^0[M^c] \tag{10}$$

where M^c denotes the parameters not in the model M and $\boldsymbol{\beta}^0[M]$ represents the true coefficients in the model M. Result 2 can be used to calculate the confidence intervals for the components of $\boldsymbol{\beta}_M$.

3 Simulation study

Consider

$$Y_i = \sin(2x_i) + \epsilon_i, \qquad i = 1, \ldots, n,$$

where $x_i \sim N(0, 4)$ and $\epsilon_i \overset{i.i.d}{\sim} N(0, 9)$ for $i = 1, \ldots, 50$. In the models, consider orthogonal polynomials of degree 8. We include the intercept and the first order of the polynomial in all models and we fit all possible models with the other 7 terms (2^7 models). Denote the orthogonal polynomials by $\boldsymbol{g}(x) = (g_1(x), \ldots, g_8(x))$, we want to approximate $\sin(2x)$ by orthogonal polynomials. We run the simulation until the model with $M = (\beta_0, \beta_1, \beta_3, \beta_5)$ has been selected 1000 times. Denote $\boldsymbol{g_1} = (1_n, \boldsymbol{g})$, including a unit column for the intercept. The confidence intervals are calculated for the components of $(\boldsymbol{g_{1M}}^t\boldsymbol{g_{1M}})^{-1}\boldsymbol{g_{1M}}^t \sin(2\boldsymbol{x})$.

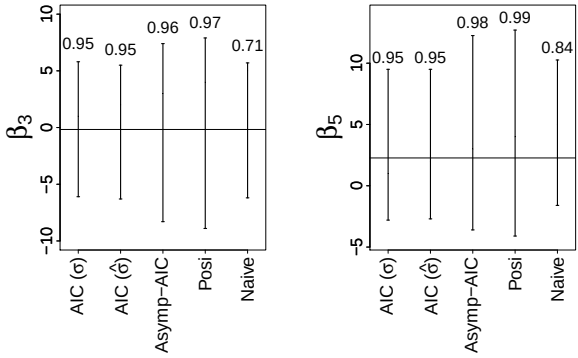


Figure 1: Mean of confidence intervals and their coverage probabilities over 1000 replications for different methods.

Figure 1 shows the mean of the confidence intervals over 1000 simulation runs for different methods along with their coverage probabilities for β_3 and β_5. We denote the proposed method by AIC(σ) when we use the knowledge about the σ and denote by AIC($\widehat{\sigma}$) where we estimate the variance in the full model. The results for [3] (Asymp-AIC) and [2] (Posi) are also presented. Both AIC(σ) and AIC($\widehat{\sigma}$) outperform other methods in terms of confidence interval lengths. The naive method leads to confidence intervals with similar length but the coverage probability is lower than the nominal value.

4 Conclusion

We proposed a new method for considering the selection randomness in inference by AIC for linear regression. In contrast the Asymp-AIC proposed by [3] which holds asymptotically, we do not need to simulate from the constrained multivariate normal distribution and the results are exact even in small sample sizes. The method performs better than PostAIC when the linear model is not the correct model. For normal linear regression models this method can be considered as a complement for PostAIC. Because if we assume the selected model is correct, the PostAIC can generate accurate confidence intervals; otherwise, the proposed method in this chapter can be used.

Acknowledgements: Your acknowledgements.

References

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory*, 267–281, 1973.

[2] R. Berk, L. Brown, A. Buja, K. Zhang and L. Zhao. Valid post-selection inference. *The Annals of Statistics*, 41:802–837, 2013.

[3] A. Charkhi and G. Claeskens. Asymptotic post-selection inference for Akaike's information criterion. *Technical report.*

[4] A. Charkhi and G. Claeskens. Exact post-selection inference for AIC in linear regression. *Technical report.*

[5] W. Fithian, D. L. Sun and J. Taylor. Optimal inference after model selection. *Technical report.*

[6] J. D. Lee, D. L. Sun, Y. Sun and J. Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

[7] H. Leeb and B. M. Pötscher. Model selection and inference: Facts and fiction. *Econometric Theory*, 21:22–59, 2005.

[8] H. Leeb and B. M. Pötscher. Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics*, 34:2554–2591, 2006.

[9] J. Taylor, R. Lockhart, R. J. Tibshirani, R. Tibshirani. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111:600–620, 2016.

The Elicitation Problem

Tobias Fissler*[1]

[1] *University of Bern*

Competing point forecasts for functionals such as the mean, a quantile, or a certain risk measure are commonly compared in terms of loss functions. These should be incentive compatible, i.e., the expected score should be minimized by the correctly specified functional of interest. A functional is called *elicitable* if it possesses such an incentive compatible loss function. With the squared loss and the absolute loss, the mean and the median possess such incentive compatible loss functions, which means they are elicitable. In contrast, variance or Expected Shortfall are not elicitable. Besides investigating the elicitability of a functional, it is important to determine the whole class of incentive compatible loss functions as well as to give recommendations which loss function to use in practice, taking into regard secondary quality criteria of loss functions such as order-sensitivity, convexity, or homogeneity.

Keywords: Consistency, Elicitability, Expected Shortfall, Scoring functions, Value at Risk

1 Evaluating and comparing forecasts

"From the cradle to the grave, human life is full of decisions. Due to the inherent nature of time, decisions have to be made today, but at the same time, they are supposed to account for unknown and uncertain future events. However, since these future events cannot be *known* today, the best thing to do is to base the decisions on *predictions* for these unknown and uncertain events. The call for and the usage of predictions for future events is literally ubiquitous and even dates back to ancient times." [2] Today, elaborated forecasts are present in a variety of different disciplines: government, business, finance, the energy market, agriculture, or everyday life.

Assume we have $m \in \mathbb{N}$ competing experts issuing their forecasts for time $t = 1, \dots, N$. Then, one has *prediction-observation-sequences*

$$\left(x_t^{(i)}, y_t\right)_{t=1,\dots,N} \qquad i \in \{1, \dots, m\}. \tag{1}$$

*Corresponding author: tobias.fissler@stat.unibe.ch

The values y_t are *ex post* realizations of a time series $(Y_t)_{t\in\mathbb{N}}$, taking values in an *observation domain* O, whereas $x_t^{(i)}$ are *ex ante* forecasts taking values in an *action domain* A. Assessing the quality of the forecasts, one can ask two main questions: (i) How good is the forecast at hand in *absolute* terms? And (ii) How good is the forecast at hand in *relative* terms? Question (i) deals with *forecast validation*, whereas question (ii) is concerned with *forecast selection*, *forecast comparison*, or *forecast ranking*. The concept of elicitability – and the elicitation problem in particular – focuses on question (ii).

1.1 Consistent scoring functions and elicitability

To introduce the abstract decision-theoretic framework of forecast comparison, there is no need to specify the observation domain O and the action domain A. In particular, the observations can be real-valued, vector-valued, but also functional-valued or even set-valued. Acknowledging the uncertainty of future outcomes, the forecasts can be probabilistic in nature, taking the form of probability distributions or densities. In this case, the action domain A coincides with a class of probability distributions $\mathcal{F}$ where one assumes that $\mathcal{F}$ contains the (conditional) distributions F_t of Y_t. On the other hand, one is often interested in certain statistical properties of the underlying distribution $F_t \in \mathcal{F}$ of Y_t such as the mean, the median, or a certain risk measure. Mathematically speaking, such a property can be specified in terms of a *functional* $T\colon \mathcal{F} \to \mathsf{A}$. In this situation, one speaks about *point forecasts*, and typically, A coincides with O (e.g. in case of the mean) where $\mathsf{A} = \mathbb{R}^k$, but might also be functional-valued or set-valued. Interestingly, the concept of probabilistic forecasts can be covered by the latter upon considering the identity map on $\mathcal{F}$ as the functional T. For most of the forthcoming results, we focus on vector-valued point forecasts, meaning $\mathsf{A} = \mathbb{R}^k$, and $\mathsf{O} = \mathbb{R}^d$.

Commonly, competing forecasts are assessed in terms of loss or *scoring functions* $S\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}$, with the most popular choices $S(x, y) = |x - y|$, or $S(x, y) = (x - y)^2$ when $\mathsf{A} = \mathsf{O} = \mathbb{R}$. Thus, if a forecaster reports the quantity $x \in \mathsf{A}$ and $y \in \mathsf{O}$ materializes, she is *penalized* by $S(x, y) \in \mathbb{R}$. Given the competing prediction-observation-sequences at (1), the ranking is done in terms of the *realized scores* $\bar{\mathbf{S}}_N^{(i)} = \frac{1}{N}\sum_{t=1}^{N} S(x_t^{(i)}, y_t)$, $i \in \{1, \dots, m\}$. That is, a forecaster is deemed to be the better the lower her realized score is. However, this ranking depends on the choice of the scoring function S. To incentivize truthful and hones forecasts, the *Bayes act* $\arg\min_{x\in\mathsf{A}} \mathbf{E}_F[S(x, Y)]$ should coincide with the correctly specified forecast $T(F)$, hence, the scoring function must be chosen *in line* with the functional T. If $T(F) = \arg\min_{x\in\mathsf{A}} \mathbf{E}_F[S(x, Y)]$ for all $F \in \mathcal{F}$, S is called *strictly $\mathcal{F}$-consistent* for $T\colon \mathcal{F} \to \mathsf{A}$. Following the terminology of [5, 8], a functional $T\colon \mathcal{F} \to \mathsf{A}$ is called *elicitable* if it possesses

a strictly $\mathcal{F}$-consistent scoring function S. Besides meaningful forecast comparison and ranking, the elicitability of a functional opens the possibility to do M-estimation. That is, under certain regularity conditions on the sequence $(Y_t)_{t\in\mathbb{N}}$ detailed e.g. in [6], $\widehat{T}_n = \arg\min_{x\in\mathsf{A}} \frac{1}{n}\sum_{t=1}^{n} S(x, Y_t)$ is a consistent estimator for T, if S is strictly consistent for T. Similarly, elicitability leads the way to generalized regression such as quantile regression or expectile regression; see [7, 9].

2 The elicitation problem

Having settled the basic definitions, one can formulate a threefold *elicitation problem* with respect to a fixed functional $T\colon \mathcal{F} \to \mathsf{A}$.

(i) Is T elicitable?

(ii) What is the class of strictly $\mathcal{F}$-consistent scoring functions for T?

(iii) What are *good choices* of strictly $\mathcal{F}$-consistent scoring functions?

The rest of this abstract summarizes some important ideas, contributions, and results concerning the elicitation problem.

2.1 Which functionals are elicitable?

One natural way to show the elicitability of a functional is by directly providing a strictly consistent scoring function. In particular, one can show that under certain regularity assumptions, the piecewise linear loss $S_\alpha(x, y) = (\mathbf{1}\{y \le x\} - \alpha)(x - y)$ is strictly consistent for the α-quantile, and that the piecewise squared loss $S_\tau(x, y) = |\mathbf{1}\{y \le x\} - \tau|(x - y)^2$ is strictly consistent for the τ-expectile (in particular, the mean and the median, as well as all moments, are elicitable, subject to mild regularity assumptions). [11] has provided a powerful necessary condition in terms of the level sets of the functional at hand, which is often relatively easy to check in practice.

Proposition 1 (Convex level sets [11])**.** *Let $T\colon \mathcal{F} \to \mathsf{A}$ be elicitable. Then, for any $F_0, F_1 \in \mathcal{F}$ such that $T(F_0) = T(F_1) = t$ and for any $\lambda \in (0, 1)$ such that $F_\lambda = (1 - \lambda)F_0 + \lambda F_1 \in \mathcal{F}$ it holds that $T(F_\lambda) = t$.*

Remarkably, the proof works independently of the specific choice of A. The result shows that variance and Expected Shortfall (ES) are generally not elicitable [5]. If $\mathsf{A} = \mathbb{R}$ and if the functional T fulfills some continuity conditions, [12] showed the sufficiency of convex level sets for elicitability. Similar results for sufficiency lack for the case $\mathsf{A} = \mathbb{R}^k$ when $k > 1$.

In case of vector-valued functionals, a functional $T = (T_1, \dots, T_k)$ consisting of elicitable components is again elicitable. If S_m is strictly consistent for T_m, then $S(x_1, \dots, x_k, y) = \sum_{m=1}^{k} S_m(x_m, y)$ is a strictly consistent scoring function for T. This observation provokes the questions (a) whether strictly consistent scoring functions must be necessarily of this form, and (b) whether functionals consisting only of elicitable components are the only vector-valued functionals. The *revelation principle* [11] gives a negative answer to question (b). It asserts that any bijection of an elicitable functional is elicitable. Since the pair (mean, variance) is a bijection of the first two moments, which are elicitable, this shows the elicitability of the pair (mean, variance), even though variance itself is not elicitable. This somehow unexpected result leads to the natural question: *Are bijections of functionals with elicitable components the only elicitable functionals?* It turns out that this is not the case: The two risk measures Expected Shortfall (ES) and Value at Risk (VaR) are, as a pair, jointly elicitable even though ES itself is not elicitable; see Theorem 1. Moreover, there is generally no (known) bijection between (VaR, ES) and a vector consisting only of elicitable components.

2.2 Determine the class of strictly consistent scoring functions

Interestingly, strictly consistent scoring functions for a functional T are not unique. E.g., if S is strictly consistent for T, then $(x, y) \mapsto \lambda S(x, y) + a(y)$ is also strictly consistent for T for any $\lambda > 0$ and any 'offset-function' $a\colon \mathsf{O} \to \mathbb{R}$. Moreover, the class of strictly consistent scoring functions is convex. However, there is far more flexibility in the class. A powerful tool is the so-called *Osband's principle* [11, 3]. It connects the gradient of an expected score with the expectation of an *identification function*. An identification function for a functional $T\colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$ is a function $V\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}^k$ such that $\mathbf{E}_F[V(x, Y)] = 0$ if and only if $x = T(F)$ for all $F \in \mathcal{F}$. Examples are $V(x, y) = x - y$ for the mean and $V(x, y) = \mathbf{1}\{y \leq x\} - \alpha$ for the α-quantile. If a functional $T\colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$ is elicitable and possesses an identification function, then, under some richness conditions on the class $\mathcal{F}$, there exists a matrix-valued function $h\colon \mathsf{A} \to \mathbb{R}^{k \times k}$ such that

$$\nabla_x \, \mathbf{E}_F[S(x, Y)] = h(x) \, \mathbf{E}_F[V(x, Y)] \qquad \forall x \in \mathsf{A}, \ \forall F \in \mathcal{F}. \tag{2}$$

One can also derive a *second order* Osband's principle considering the Hessian $\nabla_x^2 \, \mathbf{E}_F[S(x, Y)]$ of the expected score. Under appropriate smoothness conditions, the Hessian must be symmetric for all $F \in \mathcal{F}$ and positive semi-definite at $x = T(F)$. This implies further necessary conditions on the matrix-function

h often even leading to sufficient conditions for strict consistency.[1] Exploiting Osband's principle, one can show that – under some regularity conditions – $S\colon \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ is a strictly consistent scoring function for the mean if and only if S is of *Bregman type*, that is,

$$S(x,y) = \phi'(x)(x-y) - \phi(x) + a(y), \tag{3}$$

where $\phi\colon \mathbb{R}\to\mathbb{R}$ is strictly convex. Similarly, S is strictly consistent for the α-quantile if and only if

$$S(x,y) = \big(\mathbf{1}\{y\le x_1\} - \alpha\big)g(x_1) - \mathbf{1}\{y\le x_1\}g(y) + a(y), \tag{4}$$

where $g\colon \mathbb{R}\to\mathbb{R}$ is strictly increasing. Indeed, taking derivatives of the expected score, (3) becomes $\partial_x \mathbf{E}_F\left[S(x,Y)\right] = \phi''(x)(x-\mathbf{E}_F\left[Y\right])$ such that ϕ'' plays the role of h in (2). For (4), one obtains $\partial_x \mathbf{E}_F\left[S(x,Y)\right] = g'(x)(F(x)-\alpha)$, such that $g' = h$ in (2).

Expected Shortfall is jointly elicitable with Value at Risk

VaR and ES are the most popular risk measures in practice. For a financial position Y with distribution F and a level $\alpha\in(0,1)$, they are defined as

$$\begin{aligned}\mathrm{VaR}_\alpha(F) &:= F^{-1}(\alpha) = \inf\{x\in\mathbb{R} : F(x)\ge\alpha\},\\ \mathrm{ES}_\alpha(F) &:= \frac{1}{\alpha}\int_0^\alpha \mathrm{VaR}_\beta(F)\,\mathrm{d}\beta = \mathbf{E}_F\left[Y \mid Y\le \mathrm{VaR}_\alpha(F)\right].\end{aligned}$$

That means risky positions yield large *negative* values of VaR_α or ES_α. Intuitively, VaR_α gives the worst loss out of the best $(1-\alpha)\times 100\%$ of all cases, whereas ES_α gives the average loss given one exceeds VaR_α. There is an ongoing debate in academia and industry which risk measure to use. The debate mainly concentrates on ES_α and VaR_α. The latter, as a quantile, is elicitable under mild regularity conditions, it fails to be superadditive, thus violating the coherence property of risk measures. Moreover, it fails to take into account the size of losses beyond the level α. Conversely, ES_α considers the whole tail of the distribution beyond the level α, it fulfills the coherence property, but fails to be elicitable. In this light, the following result is crucial and opens the possibility to meaningful forecast comparison of joint (VaR, ES)-forecasts which is of particular importance in the context of quantitative risk management and especially the question of backtestability [4, 10].

[1] Using second order Osband's principle, one can show for example, that any vector of different quantiles and / or expectiles only possesses strictly consistent scoring functions that are additively separable. On the other hand, vectors of expectations allow for a more flexible structure similar to (3). This gives answers to the previous question (a).

Theorem 1 ([3]). *Let $\alpha \in (0,1)$. Let $\mathcal{F}$ be a class of distribution functions on $\mathbb{R}$ with finite first moments and unique α-quantiles.*
(i) If $\phi\colon \mathbb{R} \to \mathbb{R}$ is strictly convex and if for any $x_2 \in \mathbb{R}$, the function

$$[x_2, \infty) \to \mathbb{R}, \quad x_1 \mapsto g(x_1) + \phi'(x_2)\frac{x_1}{\alpha} \tag{5}$$

is strictly increasing, then the scoring function $S\colon \mathsf{A}_0 \times \mathbb{R} \to \mathbb{R}$, where $\mathsf{A}_0 := \{(x_1, x_2) \in \mathbb{R}^2 \colon x_1 \geq x_2\}$, of the form

$$\begin{aligned} S(x_1, x_2, y) = {} & \big(\mathbf{1}\{y \leq x_1\} - \alpha\big)g(x_1) - \mathbf{1}\{y \leq x_1\}g(y) + a(y) \\ & + \phi'(x_2)\Big(x_2 + \big(\mathbf{1}\{y \leq x_1\} - \alpha\big)\frac{x_1}{\alpha} - \mathbf{1}\{y \leq x_1\}\frac{y}{\alpha}\Big) - \phi(x_2), \end{aligned} \tag{6}$$

is strictly $\mathcal{F}$-consistent for $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$.
(ii) Conversely, under some regularity conditions, all strictly consistent scoring functions for $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$ are of the form given at (6).

Part (ii) of Theorem 1 asserting the necessity of the form at (6) can be shown using Osband's principle with the joint two-dimensional identification function $V(x_1, x_2, y) = \big(\mathbf{1}\{y \leq x_1\} - \alpha, x_2 + \big(\mathbf{1}\{y \leq x_1\} - \alpha\big)\frac{x_1}{\alpha} - \mathbf{1}\{y \leq x_1\}\frac{y}{\alpha}\big)'$. Part (i) can be proved by anticipating that for fixed x_1 the function $(x_2, y) \mapsto S(x_1, x_2, y)$ is of Bregman-type with minimum at $V_2(x_1, x_2, y) = 0$. On the other hand, for fixed x_2, due to the condition at (5), the function $(x_1, y) \mapsto S(x_1, x_2, y)$ is a strictly consistent scoring function for the α-quantile.

2.3 Secondary quality criteria besides strict consistency

Facing the multitude of strictly consistent scoring functions illustrated at (3), (4), and (6), this burden of choice calls for new concepts such as the notion of forecast dominance introduced in [1]. Alternatively, it motivates the introduction of secondary quality criteria besides strict consistency giving guidance which scoring function to use. This line of research is pursued in [2]. Generalizations of the concept of order-sensitivity [8] to the higher dimensional setting are introduced, ensuring meaningful forecast comparison of possibly misspecified predictions in particular settings. Convexity of scoring functions can show to be beneficial for optimization purposes, but also shed new light on the paradigm of maximizing the sharpness of a forecast subject to calibration as well as on incentives for cooperation between competing forecasters. Finally, equivariance properties of functionals motivate the notion of order-preserving scoring functions, nesting concepts such as homogeneity or translation invariance of scoring functions.

References

[1] W. Ehm, T. Gneiting, A. Jordan, and F. Krüger. Of quantiles and expectiles: consistent scoring functions, Choquet representations and forecast rankings. *JRSSB*, 78(3):505–562, 2016.

[2] T. Fissler. *On Higher Order Elicitability and Some Limit Theorems on the Poisson and Wiener Space*. PhD thesis, University of Bern, 2017.

[3] T. Fissler and J. F. Ziegel. Higher order elicitability and Osband's principle. *Ann. Statist.*, 44(4):1680–1707, 2016.

[4] T. Fissler, J. F. Ziegel, and T. Gneiting. Expected shortfall is jointly elicitable with value-at-risk: implications for backtesting. *Risk Magazine*, pages 58–61, January 2016.

[5] T. Gneiting. Making and Evaluating Point Forecasts. *J. Amer. Statist. Assoc.*, 106:746–762, 2011.

[6] P. J. Huber and E. M. Ronchetti. *Robust Statistics*. John Wiley & Sons, Inc., Hoboken, New Jersey, second edition, 2009.

[7] R. Koenker. *Quantile Regression*. Cambridge University Press, Cambridge, 2005.

[8] N. Lambert, D. M. Pennock, and Y. Shoham. Eliciting properties of probability distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pages 129–138, Chicago, Il, USA, 2008. ACM.

[9] W. K. Newey and J. L. Powell. Asymmetric Least Squares Estimation and Testing. *Econometrica*, 55:819–847, 1987.

[10] N. Nolde and J. F. Ziegel. Elicitability and backtesting. *Preprint on arXiv*, 2017.

[11] K. H. Osband. *Providing Incentives for Better Cost Forecasting*. PhD thesis, University of California, Berkeley, 1985.

[12] I. Steinwart, C. Pasin, R. Williamson, and S. Zhang. Elicitation and Identification of Properties. *JMLR: Workshop and Conference Proceedings*, 35:1–45, 2014.

Multilevel Functional Principal Component Analysis for Unbalanced Data

Zuzana Rošťáková*[1] and Roman Rosipal[1]

[1] *Institute of Measurement Science, Slovak Academy of Sciences, Slovakia*

Functional principal component analysis (FPCA) is the key technique for dimensionality reduction and detection of main directions of variability present in functional data. However, it is not the most suitable tool for the situation when analysed dataset contains repeated or multiple observations, because information about repeatability of measurements is not taken into account. Multilevel functional principal component analysis (MFPCA) is the modified version of FPCA developed for data observed at multiple visits. The original MFPCA method was designed for balanced data only, where for each subject the same number of measurements is available. In this article we propose the modified MFPCA algorithm which can be applied for unbalanced functional data. The modified algorithm is validated and tested on real–world sleep data.

Keywords: multilevel functional principal component analysis, functional data with multiple observations, sleep probabilistic curves

Introduction

Functional principal component analysis (FPCA) is an appropriate tool for detecting main directions of variability and dimensionality reduction of functional data [1]. On the other hand, FPCA considers each curve as a single observation and therefore it is not appropriate for detecting sources of variability in datasets with multiple observations. These multiple observations can be represented by repeated collection of data at multiple visits.

To address this repeated observations data design, the multilevel functional principal component analysis (MFPCA) method was developed [1]. MFPCA decomposes observed functional data into three parts i) the overall mean, common for all subjects, ii) the subject–specific deviation from the overall mean, and iii) the remaining deviation from a subject–specific profile. Moreover, the method is able to transform high dimensional functional data (possibly

*Corresponding author: zuzana.rostakova@gmail.com

infinite) into finite dimensional vector spaces of principal components at two levels.

The original MFPCA method was proposed and validated only for data with the same number of observations per subject. In this article we demonstrate that in its original form the method is not able to properly detect subject–specific profiles when the number of observations among subjects is different. Therefore we propose the modification of the original MFPCA method which can better deal with the unbalanced data situation.

The article is organised in the following way. The general description of MFPCA is given in the first section. The modified MFPCA method for unbalanced data is proposed in Section 2. In Section 3 the method is validated on real–world sleep data. Finally, Section 4 provides discussion and a few concluding remarks.

1 Multilevel functional principal component analysis

MFPCA deals with functional data with repeated observations in order to detect sources of variability at two levels; the between– and within–subject variability [1].

Let consider I subjects with J observations $X_{ij}, i = 1, \ldots, I; j = 1, \ldots, J$. For simplicity we assume that observed functional data are defined at the same time grid within a closed interval T and are sufficiently smooth. Moreover the observations or visits within each subject should have natural ordering. In [1], the authors used a two–way functional ANOVA model in order to decompose X_{ij} into a fixed and random part

$$X_{ij}(t) = \mu(t) + \eta_j(t) + Z_i(t) + W_{ij}(t), \qquad t \in T. \tag{1}$$

The overall mean μ and the visit–specific deviation from the overall mean $\eta_j, j = 1, \ldots, J$ are fixed effects. For identifiability we assume $\sum_{j=1}^{J} \eta_j(t) = 0$, $t \in T$. The subject–specific deviation from the visit–specific mean Z_i and the remaining deviation from the subject– and visit–specific profiles W_{ij} are uncorrelated stochastic processes with mean 0 and covariance functions $S_1 : T \times T \to \mathbb{R}$ and $S_2 : T \times T \to \mathbb{R}$.

According to the Karhunen-Loewe expansion the stochastic processes Z_i and W_{ij} can be decomposed in the following way

$$Z_i(t) = \sum_{k=1}^{\infty} \alpha_{ik} \phi_k^{(1)}(t) \qquad W_{ij}(t) = \sum_{l=1}^{\infty} \beta_{ijl} \phi_l^{(2)}(t)$$

where α_{ik} and β_{ijl} are random variables with mean 0 and

$$\mathrm{E}(\alpha_{ik}\alpha_{il}) = \begin{cases} 0, & \text{if } k \neq l, \\ \lambda_k^{(1)}, & \text{if } k = l, \end{cases} \qquad \mathrm{E}(\beta_{ijk}\beta_{ijl}) = \begin{cases} 0, & \text{if } k \neq l, \\ \lambda_k^{(2)}, & \text{if } k = l. \end{cases}$$

Moreover, $\{\alpha_{ik}, k = 1, 2, \dots\}$ are uncorrelated with $\{\beta_{ijl}, l = 1, 2, \dots\}$. We call them the level 1 and level 2 principal component scores. Two sets of orthonormal functional bases of the L^2 space

$$\{\phi_k^{(1)}, k = 1, 2, \dots\} \qquad \text{and} \qquad \{\phi_l^{(2)}, l = 1, 2, \dots\}$$

which represents the functional principal components (FPCs) at level 1 and level 2 are not necessarily mutually orthogonal.

In [1], the following three covariance functions are considered in order to estimate functional principal components at both levels

$$\begin{aligned} K_T(s,t) &= Cov\left(X_{ij}(s), X_{ij}(t)\right) = S_1(s,t) + S_2(s,t), \\ K_B(s,t) &= Cov\left(X_{ij}(s), X_{ik}(t)\right) = S_1(s,t), \\ K_W(s,t) &= K_T(s,t) - K_B(s,t) = \frac{1}{2} Cov\left(X_{ij}(s) - X_{ik}(s), X_{ij}(t) - X_{ik}(t)\right) = S_2(s,t). \end{aligned}$$

In other words, FPCs at level 1 are eigenfunctions of K_B and FPCs at level 2 are eigenfunctions of K_W.

Using the method of moments, the following estimators of unknown quantities are proposed in [1]

$$\widehat{\mu}(t) = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} X_{ij}(t), \qquad \widehat{\eta_j}(t) = \frac{1}{I}\sum_{i=1}^{I} X_{ij}(t) - \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} X_{ij}(t), \qquad t \in T$$

$$\widehat{K_T}(s,t) = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} \left(X_{ij}(s) - \widehat{\mu}(s) - \widehat{\eta_j}(s)\right)\left(X_{ij}(t) - \widehat{\mu}(t) - \widehat{\eta_j}(t)\right), \tag{2}$$

$$\widehat{K_B}(s,t) = \frac{1}{IJ(J-1)}\sum_{i=1}^{I}\sum_{j\neq l}^{J} \left(X_{ij}(s) - \widehat{\mu}(s) - \widehat{\eta_j}(s)\right)\left(X_{il}(t) - \widehat{\mu}(t) - \widehat{\eta_l}(t)\right), \tag{3}$$

$$\widehat{K_W}(s,t) = \widehat{K_T}(s,t) - \widehat{K_B}(s,t), \tag{4}$$

where $\widehat{\mu}$ and $\widehat{\eta_j}$ are estimated similarly as in the standard ANOVA model [1].

The way of selecting the number of functional principal components at each level separately, as well as the procedure for computing principal component scores at both levels are described in details in [1].

2 MFPCA for unbalanced data design

The original MFPCA algorithm was designed for balanced data with ordered visits. However, the authors state that this assumption is not restrictive and the method is able to deal with unbalanced data as well.

Let consider I subjects with $J_i, i = 1, \ldots, I$ observations. In this case, the number of observations may differ among subjects and we assume that the order of observations within each subject is exchangeable. Therefore the visit–specific deviations η_j from the overall mean are set to zero. The model (1) changes into one–way functional ANOVA

$$X_{ij}(t) = \mu(t) + Z_i(t) + W_{ij}(t), \quad t \in T, \quad j = 1, \ldots, J_i, \quad i = 1, \ldots I. \tag{5}$$

By computing the expected values of the covariance functions estimators (2), (3) and (4) for data with unbalanced design and $\widehat{\eta}_j \equiv 0$ we obtain

$$\begin{aligned}
\mathrm{E}\left(\widehat{K_T}(s,t)\right) &= \left(1 - \frac{A_2}{A_1^2}\right) S_1(s,t) + \left(1 - \frac{1}{A_1}\right) S_2(s,t), \\
\mathrm{E}\left(\widehat{K_B}(s,t)\right) &= \left(1 - \frac{2}{A_1}\frac{A_3 - A_2}{A_2 - A_1} + \frac{A_2}{A_1^2}\right) S_1(s,t) - \frac{1}{A_1} S_2(s,t), \\
\mathrm{E}\left(\widehat{K_W}(s,t)\right) &= \left(\frac{2}{A_1}\frac{A_3 - A_2}{A_2 - A_1} - 2\frac{A_2}{A_1^2}\right) S_1(s,t) + S_2(s,t),
\end{aligned} \tag{6}$$

where $A_k = \sum_{i=1}^{I} J_i^k, k = 1, 2, 3$. It means, that for $I \to \infty$ and a bounded number of observations for each subject $1 \leq J_i \leq M, M \in \mathbb{N}$, the matrices $\widehat{K_B}$ and $\widehat{K_W}$ are only asymptotically unbiased estimators of S_1 and S_2.

Therefore, when data are unbalanced, we propose the following modification of the covariance functions estimators. First, let define

$$\begin{aligned}
\widehat{K_W}^{UU}(s,t) &= \frac{1}{\sum_{i=1}^{I} J_i} \sum_{i=1}^{I} \sum_{j:J_i>1}^{J_i} \left(X_{ij}(s) - \widehat{\mu}(s)\right)\left(X_{ij}(t) - \widehat{\nu}_i^{(-j)}(t)\right), \\
\widehat{\nu}_i^{(-j)}(t) &= \frac{1}{J_i - 1} \sum_{l \neq j}^{J_i} X_{il}(t), \quad t \in T.
\end{aligned}$$

While $\mathrm{E}\left(\widehat{K_W}^{UU}(s,t)\right) = S_2(s,t)$ which holds also for unbalanced data, we can estimate FPCs at level 2 directly from $\widehat{K_W}^{UU}$. The estimator (2) for K_T remains the same with expected value (6). Therefore FPCs at level 1 can be estimated as eigenfunctions of the following function

$$\widehat{K_B}^{UU} = \frac{A_1^2}{A_1^2 - A_2}\left(\widehat{K_T} - \frac{A_1 - 1}{A_1}\widehat{K_W}^{UU}\right), \qquad \mathrm{E}\left(\widehat{K_B}^{UU}(s,t)\right) = S_1(s,t).$$

3 Application to sleep data

Sleep is a continuous process which can be described by a finite number of sleep stages. Probabilistic sleep model (PSM) characterises sleep with probability values of 20 sleep microstates [3]. Considering the probability values as a function of time we obtain a curve.

In the first step we took 292 probabilistic sleep curves of the PSM applied to sleep recordings from the SIESTA database [2]. These curves represent the sleep microstate similar to REM (or rapid eye movement sleep stage). Using the two–step clustering approach [4], the curves were divided into 12 clusters depicted in Figure 1. Objective of this study is to identify cluster representatives, which can be used for the further analysis of the sleep process. With this aim in mind, we applied model (5) to the clustered curves. Effectively this means that we have 12 clusters (or 'subjects') with a different number of observations, in this case the number of curves in each cluster. The number of curves varied from 4 (cluster 9) to 117 (cluster 12).

Using the original and modified MFPCA algorithms the cluster–specific profiles $P_i(t) = \widehat{\mu}(t) + \widehat{Z_i}(t),\ \ t \in T$ were computed for each cluster. The superior performance of the modified MFPCA algorithm is visible especially for clusters 2, 5 or 9 consisting of a smaller number of curves. Taking into account that the original sleep probabilistic curves are strictly positive, the cluster–specific profiles estimated by the original MFPCA method reached for short time subintervals unexpected negative values.

4 Conclusion

In this article we described modified version of the multilevel functional principal component analysis method [1]. MFPCA is an appropriate tool for detection of main direction of variability for functional data with repeated observations. Original MFPCA was developed only for balanced data where each subject has the same number of observations and the observations have natural order.

However, we found and demonstrated on real sleep data, that in its original form the algorithm applied to unbalanced data leads to inferior results because the estimators of covariance functions described in [1] are biased.This is especially true for datasets with a small sample size.

In this article we proposed the modified estimators of covariance functions for unbalanced data. These leads to the unbiased estimation of functional principal components at level 1 and 2. We proved good performance of the proposed modified version of MFPCA on the analysed sleep data.

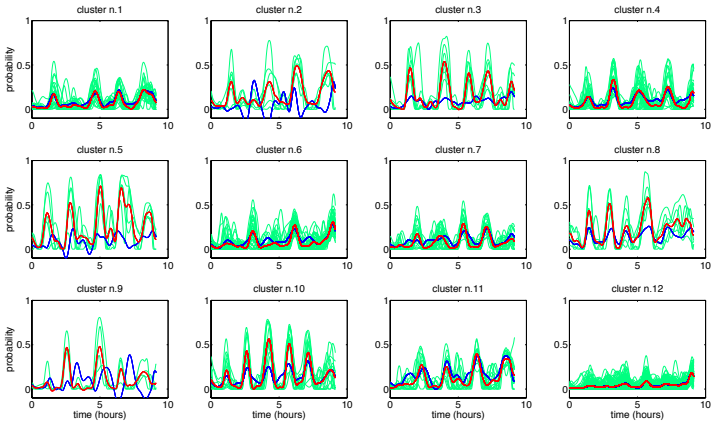


Figure 1: Cluster analysis of the sleep microstate similar to the REM sleep stage with 292 sleep probabilistic curves (light green) divided into 12 clusters. Cluster–specific profiles estimated by the modified MF-PCA algorithm (red) form better cluster representatives than their counterparts estimated by the original MFPCA algorithm (blue).

Acknowledgements: This work has been supported by the Slovak Research and Development Agency (grant APVV-0668-12) and by the VEGA grant 2/0011/16.

References

[1] C.-Z. Di, C. M. Crainiceanu, B. S. Caffo, and N. M. Punjabi. Multilevel functional principal component analysis. *The Annals of Applied Statistics*, 3(1):458–488, 2009.

[2] G. Klösch, B. Kemp, T. Penzel, A. Schlögl, P. Rappelsberger, E. Trenker, G. Gruber, J. Zeitlhofer, B. Saletu, W. Herrmann, S. Himanen, D. Kunz, M. Barbanoj, J. Röschke, A. Varri, and G. Dorffner. The SIESTA project polygraphic and clinical database. *Medicine and Biology Magazine*, 20(3):51–57, 2001.

[3] A. Lewandowski, R. Rosipal, and G. Dorffner. Extracting more information from EEG recordings for a better description of sleep. *Computer Methods and Programs in Biomedicine*, 108(3):961 – 972, 2012.

[4] Z. Rošťáková and R. Rosipal. A novel two-step iterative approach for clustering functional data. In *22nd International Conference on Computational Statistics (COMPSTAT 2016): Book of Abstracts*, page 61, 2016.

Mallows' Model Based on Lee Distance

Nikolay I. Nikolov[*1] **and Eugenia Stoimenova**[1,2]

[1] *Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Acad. G.Bontchev str., block 8, 1113 Sofia, Bulgaria*
[2] *Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Acad. G.Bontchev str., block 25A, 1113 Sofia, Bulgaria*

In this paper the Mallows' model based on Lee distance is considered and compared to models induced by other metrics on the permutation group. As an illustration, the complete rankings from the American Psychological Association election data are analyzed.

Keywords: Rank data analysis, Mallows' models, Lee distance, Metrics on permutations

1 Mallows' models

A full ranking of N items simply assigns a complete ordering to the items. Any such ranking vector can be viewed as an element π of the permutation group $\mathcal{S}_N$ generated by the first N natural integers. Thus the notation

$$\pi =< \pi^{-1}(1), \pi^{-1}(2), \ldots, \pi^{-1}(N) >$$

is used for the permutation π which corresponds to listing the various items in their ranked order. There are various nonparametric methods for modelling rank data. Some models have larger probabilities for rankings that are "close" to a "modal" ranking π_0. An example of such probability model is given by

$$P_{\theta,\pi_0}(\pi) = e^{\theta d(\pi,\pi_0) - \psi(\theta)} \quad \text{for } \pi \in \mathcal{S}_N, \tag{1}$$

where θ is a real parameter ($\theta \in \mathbb{R}$), $d(\cdot,\cdot)$ is a metric on $\mathcal{S}_N$, π_0 is a fixed ranking and $\psi(\theta)$ is a normalizing constant. When $\theta > 0$, π_0 is the modal ranking, for $\theta < 0$, π_0 is an antimode, and for $\theta = 0$, P_{θ,π_0} is the uniform distribution. More general model, with $d(\cdot,\cdot)$ being a discrepancy function, is suggested by Diaconis [4], but since all distances used in this paper are metrics, $d(\cdot,\cdot)$ could be regarded as a metric. Deza and Huang [3] considered some metrics on $\mathcal{S}_N$ which are widely used in applied scientific and statistical problems.

*Corresponding author: n.nikolov@math.bas.bg

$$F(\pi,\sigma) = \sum_{i=1}^{N} \mid \pi(i) - \sigma(i) \mid \qquad \text{Spearman's footrule}$$

$$R(\pi,\sigma) = \left(\sum_{i=1}^{N} (\pi(i) - \sigma(i))^2 \right)^{1/2} \qquad \text{Spearman's rho}$$

$$M(\pi,\sigma) = \max_{1 \le i \le N} \mid \pi(i) - \sigma(i) \mid \qquad \text{Chebyshev metric}$$

$$K(\pi,\sigma) = \#\{(i,j) : 1 \le i,j \le N, \ \pi(i) < \pi(j), \sigma(i) > \sigma(j)\} \qquad \text{Kendall's tau}$$

$$C(\pi,\sigma) = N \text{ minus number of cycles in } \sigma\pi^{-1} \qquad \text{Cayley's distance}$$

$$U(\pi,\sigma) = N \text{ minus length of the longest increasing subsequence in } \sigma\pi^{-1} \qquad \text{Ulam's distance}$$

$$H(\pi,\sigma) = \#\{i \in \{1,2,\ldots,N\} : \pi(i) \neq \sigma(i)\} \qquad \text{Hamming distance}$$

$$L(\pi,\sigma) = \sum_{i=1}^{N} \min(\mid \pi(i) - \sigma(i) \mid, N - \mid \pi(i) - \sigma(i) \mid) \qquad \text{Lee distance}$$

Easily can be shown that all of the presented metrics possess the following important property.

Definition 1. The metric d on $\mathcal{S}_N$ is called right-invariant, if and only if $d(\pi,\sigma) = d(\pi \circ \tau, \sigma \circ \tau)$ for all $\pi, \sigma, \tau \in \mathcal{S}_N$.

Critchlow [1] pointed that the right-invariance of metric is necessary requirement since it means that the distance between rankings does not depend on the labelling of the items. More properties for these metrics can be found in Critchlow [1, 2], Diaconis [4] and Marden [7].

If $d(\cdot,\cdot)$ is right-invariant, then (1) can be defined by the random variable $D(\pi) = d(\pi,\pi_0) = d(\pi\pi_0^{-1}, e_N)$, where $\pi \sim Uniform(\mathcal{S}_N)$ and e_N is the identity permutation ($e_N =< 1,2,\ldots,N >$). Notice that the distribution of D does not depend on π_0 and it could be assumed that $D(\pi) = d(\pi, e_N)$. Let's use the notation $D_{[*]}$ for the random variable D induced by some distance $[*]$ from the listed above. The special cases of (1) with $D = D_K$ and $D = D_{R^2}$ are first investigated by Mallows [6]. Models based on D_C and D_H can be found in Fligner and Verducci [5].

Model (1) could be significantly simplified if the distribution of D is known and can be written explicitly. Let $m(t)$ be the moment generating function of D. Then, as shown in [5],

$$e^{\psi(\theta)} = \sum_{\pi \in S_N} e^{\theta D(\pi)} = N! \sum_{d_i} P(D = d_i) e^{\theta d_i} = N!\, m(\theta)$$

$$\Rightarrow \quad \psi(\theta) = \log(N!\, m(\theta))\,. \tag{2}$$

For D_F, D_R, D_M, D_K, D_C, D_U and D_H numerical characteristics, exact distributions, asymptotic approximations and statistical applications can be found in Diaconis [4] and Marden [7]. The goal of this paper is to study the Mallows' model based on D_L and compare it to the models induced by the other given metrics. The rest of the paper is organized as follows. In Section 2 some properties of the distribution of Lee distance are derived under uniformity assumption. Maximum likelihood estimations and testing procedure for deviation from the Uniform distribution are proposed in Section 3. In Section 4 a comparison between the models based on the eight distances is made.

2 Lee distance

Let's first notice that $D_L(\pi) = L(\pi, e_N)$ can be decomposed in N terms:

$$D_L(\pi) = \sum_{i=1}^{N} \min\left(\mid \pi(i) - i \mid, N - \mid \pi(i) - i \mid\right) = \sum_{i=1}^{N} c_N(\pi(i), i)\,. \tag{3}$$

There is an interpretation of $c_N(i,j) := \min\left(\mid i - j \mid, N - \mid i - j \mid\right)$ in terms of graph theory. Let G be a simple cycle graph with nodes $\{i\}_{i=1}^{N}$ and edges $\bigcup_{i=1}^{N-1} \{i, i+1\}$ and $\{N, 1\}$. Then $c_N(i,j)$ is the minimum distances over G between the nodes i and j. Obviously, $0 \le c_N(i,j) \le N/2$ for even N and $0 \le c_N(i,j) \le (N-1)/2$ for odd N, i.e.

$$0 \le c_N(i,j) \le \left[\frac{N}{2}\right], \quad \text{for all } i, j = 1, 2, \ldots, N\,, \tag{4}$$

where $[x]$ is the greatest integer less than or equal to x. From (3) and (4) it follows that

$$0 \le D_L(\pi) \le N \left[\frac{N}{2}\right], \quad \text{for all } \pi \in \mathcal{S}_N\,. \tag{5}$$

The lower limit in (5) is reached only for $\pi = e_N$. When N is even the upper limit is reached only for π equals to

$$e_N^* := < \frac{N}{2} + 1, \frac{N}{2} + 2, \ldots, N-1, N, 1, 2, \ldots, \frac{N}{2} - 1, \frac{N}{2} > \ ,$$

and in the case of odd integers N the maximum value of D_L is reached when π is equal to e'_N or e''_N, where

$$e'_N :=< \frac{N+1}{2}, \frac{N+1}{2}+1, \ldots, N-1, N, 1, \ldots, \frac{N+1}{2}-2, \frac{N+1}{2}-1 >$$
$$e''_N :=< \frac{N+1}{2}+1, \frac{N+1}{2}+2, \ldots, N-1, N, 1, \ldots, \frac{N+1}{2}-1, \frac{N+1}{2} > .$$

Since

$$c_N(\pi(i), e_N(i)) + c_N(\pi(i), e^*_N(i)) = \min(\mid \pi(i) - i \mid, N - \mid \pi(i) - i \mid) +$$
$$\min\left(\mid \pi(i) - \frac{N}{2} - i \mid, N - \mid \pi(i) - \frac{N}{2} - i \mid\right) = \frac{N}{2} \text{ , for } i = 1, 2, \ldots, \frac{N}{2} \text{ ,}$$

and

$$c_N(\pi(i), e_N(i)) + c_N(\pi(i), e^*_N(i)) = \min(\mid \pi(i) - i \mid, N - \mid \pi(i) - i \mid) +$$
$$\min\left(\mid \pi(i) - i + \frac{N}{2} \mid, N - \mid \pi(i) - i + \frac{N}{2} \mid\right) = \frac{N}{2} \text{ , for } i = \frac{N}{2}+1, \ldots, N \text{ ,}$$

the relation

$$L(\pi, e_N) + L(\pi, e^*_N) = \sum_{i=1}^{N} c_N(\pi(i), e_N(i)) + c_N(\pi(i), e^*_N(i)) = \frac{N^2}{2} \text{ , } \quad (6)$$

is true for all $\pi \in \mathcal{S}_N$. The right-invariant property of L implies that $L(\pi, e_N)$ and $L(\pi, e^*_N)$ have the same distribution when $\pi \sim Uniform(\mathcal{S}_N)$. From that fact and (6) it follows that

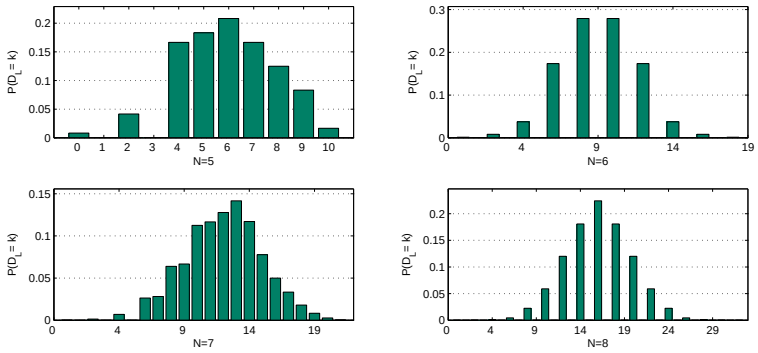
$$P(D_L = k) = P\left(D_L = \frac{N^2}{2} - k\right) \text{ , for } k = 0, 1, \ldots, \frac{N^2}{2} \text{ , i.e.}$$

the distribution of D_L is symmetric when N is even. Furthermore D_L can take only even values since

$$D_L(\pi) \equiv \sum_{i=1}^{N} \min(\mid \pi(i) - i \mid, N - \mid \pi(i) - i \mid) \pmod 2$$
$$\Rightarrow D_L(\pi) \equiv \sum_{i=1}^{N} \mid \pi(i) - i \mid \equiv 0 \pmod 2$$

for even integers N.

The probability mass function of D_L for $N = 5, 6, 7, 8$ is shown on the figure below.



3 Parameters estimation and tests for uniformity

Formula (2) can be used to find estimations for the unknown parameters in (1). Suppose that there are n observed complete rankings $\pi^{(1)}, \pi^{(2)}, \ldots, \pi^{(n)}$ and the mode π_0 in (1) is unknown. Then the loglikelihood function is given by

$$l(\theta, \pi_0, n) = \theta S(\pi_0) - n\psi(\theta),$$

where $S(\pi_0) = \sum_{i=1}^{n} d(\pi^{(i)}, \pi_0)$. In order to find the maximum likelihood estimations (MLE's), first it is necessary to calculate

$$\hat{\pi}_{min} = \underset{\pi \in \mathcal{S}_N}{\operatorname{argmin}}\, S(\pi) \quad \text{and} \quad \hat{\pi}_{max} = \underset{\pi \in \mathcal{S}_N}{\operatorname{argmax}}\, S(\pi)\,.$$

For $\theta < 0$, let $\hat{\theta}_{min}$ be the value for which $l(\theta, \hat{\pi}_{min}, n)$ is maximized. For $\theta > 0$, let the maximum of $l(\theta, \hat{\pi}_{max}, n)$ occurs for $\theta = \hat{\theta}_{max}$. Finally, the MLE's

$$\left(\hat{\theta}, \hat{\pi}_0\right) = \begin{cases} \left(\hat{\theta}_{min}, \hat{\pi}_{min}\right), & \text{if } l(\hat{\theta}_{min}, \hat{\pi}_{min}, n) \geq l(\hat{\theta}_{max}, \hat{\pi}_{max}, n) \\ \left(\hat{\theta}_{max}, \hat{\pi}_{max}\right), & \text{otherwise}\,. \end{cases}$$

If $\hat{\theta} = 0$ then (1) is the uniform model and $\hat{\pi}_0$ is not unique since for all $\pi \in \mathcal{S}_N$ the loglikelihood, $l(0, \pi, n)$, is the same. For Spearman's rho $R(\cdot,\cdot)$ and Kendall's tau $K(\cdot,\cdot)$ it can be shown that $\hat{\theta}_{min} = -\hat{\theta}_{max}$. From (6) it follows that $\hat{\theta}_{min} = -\hat{\theta}_{max}$ is also valid for Lee distance $L(\cdot,\cdot)$, when N is even. In these cases $l(\hat{\theta}_{min}, \hat{\pi}_{min}, n) = l(\hat{\theta}_{max}, \hat{\pi}_{max}, n)$ and it is enough to find just $\hat{\pi}_{min}$ and $\hat{\theta}_{min}$. The described MLE's and other methods for estimating θ and π_0 can be found in [7].

For testing the null hypothesis $H_0 : \theta = 0$ (*uniform model*) against the alternative $H_A : \theta \neq 0$, Marden [7] considered the likelihood ratio statistic (LRS) given by

$$LRS = 2\left[l_A(\hat{\theta}, \hat{\pi}_0, n) - l_0(0, \pi, n)\right] = 2\left[\hat{\theta}S(\hat{\pi}_0) - n\psi(\hat{\theta}) + n\log(N!)\right],$$

where l_0 and l_A are the loglikelihood functions under H_0 and H_A, respectively, and $\left(\hat{\theta}, \hat{\pi}_0\right)$ are the MLE's. Let $k(\pi)$ be the number of observations that are equal to $\pi \in \mathcal{S}_N$. Then the empirical probability for π is $\dfrac{k(\pi)}{n}$ and a quantity, which measures the total nonuniformity of the data, could be defined by

$$TNU = 2\sum_{\pi \in \mathcal{S}_N} k(\pi)\left[\log\left(\frac{k(\pi)}{n}\right) - \log\left(\frac{1}{N!}\right)\right].$$

Similarly to the multiple correlation coefficient in the linear regression, Marden [7] considered the coefficient

$$R^2 = \frac{LRS}{TNU},$$

which can be used to measure the percentage of nonuniformity in the data that is explained by the fitted model. Thus $R^2 = 1$ when the model exactly fits the data, and $R^2 = 0$ if it performs no better than the uniform model.

4 Comparison between the distance based models

In 1980, the American Psychological Association (APA) conducted an election in which five candidates were running for president and voters were asked to rank order all of the candidates. The complete rankings of 5738 voters are given in [4, p. 96]. The average ranks received by candidates A, B, C, D and E are 2.84, 3.16, 2.92, 3.09, and 2.99, respectively, and the total nonuniformity of the data is $TNU = 1717.51$. The fitted Mallows' models based on the eight distances considered are given in Table 1.

Since the theoretical distribution of LRS is unknown, it is approximated via simulations with 1000 trials for each distance, and the results for the mean and the 95% critical values of LRS's are presented in the last two columns. Notice that for all models the hypothesis of uniform distribution is rejected since LRS's are much larger than the simulated critical values. In fact all LRS's are larger than the maximum simulated values. However, all models explain less than a third of the nonuniformity, where the model based on D_L

Distance	$\hat{\theta}$	$\hat{\pi}_0$	**Ordering**	LRS	R^2	LRS_{sim} **mean**	LRS_{sim} **95% c.v.**
D_F	0.0828	51324	**BDCEA**	282.26	0.1643	4.63	9.62
D_{R^2}	−0.0163	15243	**ACEDB**	150.78	0.0878	3.49	7.95
D_M	−0.2639	15243	**ACEDB**	379.54	0.2210	5.49	10.43
D_K	−0.0722	15243	**ACEDB**	124.28	0.0723	3.83	8.43
D_C	−0.2483	23154	**CABED**	304.21	0.1771	7.22	11.98
D_U	−0.2505	23154	**CABED**	181.52	0.1057	6.80	12.06
D_H	0.2437	51324	**BDCEA**	290.16	0.1689	6.74	11.41
D_L	0.1656	51324	**BDCEA**	524.39	0.3053	5.52	10.73

Table 1: *Fitted Mallows' models for APA data*

has the highest $R^2 = 30.53\%$, and the lowest $R^2 = 7.23\%$ is obtained when using D_K.

The estimated "modal" orderings (antimodes for $\hat{\theta} > 0$) are given in the forth column. The ordering of D_{R^2}, D_M and D_K coincides with the "modal" ordering based on the average ranks. As mentioned in [7], there are definite camps within APA: candidates A and C are research psychologists, D and E are clinical psychologists, and B is a community psychologist. These groups can also be noticed from the orderings of D_{R^2}, D_M, D_K, D_C and D_U. Since the number of candidates $N = 5$ is odd, the maximum value of $L(e_N, \pi)$ is reached for $\pi = e'_N$ and $\pi = e''_N$. Thus the interpretation of the antimode ordering of D_L is more complex.

Candidate B is ranked last in all "modal" rankings, except in models based on D_C and D_U, where B separates the groups {A,C} and {D,E}. The rankings, which are constructed without considering candidate B, could be used to study the influence of B over the complete rankings models. The MLE's of the models' parameters for the new rankings are given in Table 2.

Distance	$\hat{\theta}$	$\hat{\pi}_0$	**Ordering**	LRS_{new}	R^2_{new}	LRS_{diff}	**Simulated** LRS_{diff} **cdf**
D_F	−0.0698	2143	**CAED**	126.22	0.1268	156.05	0.591
D_{R^2}	−0.0177	1243	**ACED**	59.79	0.0601	90.99	0.769
D_M	−0.2239	2143	**CAED**	226.30	0.2273	153.23	0.001
D_K	−0.0663	1243	**ACED**	54.64	0.0549	69.64	0.742
D_C	−0.2532	2143	**CAED**	251.79	0.2529	52.42	0.000
D_U	−0.2319	2143	**CAED**	124.75	0.1253	56.77	0.006
D_H	−0.1832	2143	**CAED**	217.59	0.2186	72.58	0.000
D_L	−0.1311	2143	**CAED**	265.39	0.2666	259.00	0.007

Table 2: *Fitted models without candidate B*

The total nonuniformity of the new data is $TNU_{new} = 995.58$. The "modal" orderings of the remaining four candidates are not changed in the models based on D_{R^2}, D_K, D_C and D_U, whereas there are new "modal" orderings in the other models. For D_C, D_U and D_H the coefficient R^2_{new} increases, while for D_F, D_{R^2}, D_K and D_L it decreases. R^2_{new} is almost the same as R^2 for the model based on D_M. The quantity $LRS_{diff} = LRS - LRS_{new}$ can be used to measure the influence of candidate B over the explanatory power of the models. The value of LRS_{diff} is simulated 1000 times for all complete models with parameters given in Table 1. The observed value of LRS_{diff} and the simulated empirical cumulative distribution function (taken at the observed value of LRS_{diff}) are given in the last two columns. There is a significant decrease in the explanatory power of the models based on D_F, D_{R^2} and D_K, since the values of LRS_{diff} are significant for these models. Thus it can be suggested that the models based on D_M, D_C, D_U, D_H and D_L are more "robust". Similar conclusion is made in [7, p. 30] by analyzing sport related rank data.

Acknowledgements: This work was supported by the grant I02/19 of the Bulgarian National Science Fund.

References

[1] D. E. Critchlow. *Metric Methods for Analyzing Partially Ranked Data.* Lecture Notes in Statistics, 34. Springer, New York, 1985.

[2] D. E. Critchlow. On rank statistics: an approach via metrics on the permutation group. *J. Statist. Plann. Inference*, 32(3):325–346, 1992.

[3] M. Deza and T. Huang. Metrics on permutations, a survey. *Journal of Combinatorics, Information and System Sciences*, 23:173–185, 1998.

[4] P. Diaconis. *Group Representations in Probability and Statistics.* IMS Lecture Notes - Monograph Series, 11. Hayward, Carifornia, 1988.

[5] M. Fligner and T. Verducci. Distance based ranking models, *Journal of the Royal Statistical Society*, 48(3):359–369, 1986.

[6] C. M. Mallows. Non-null ranking models. I. *Biometrika*, 44(1):114–130, 1957.

[7] J. I. Marden. *Analyzing and Modeling Rank Data.* Monographs on Statistics and Applied Probability, 64. Chapman & Hall, London, 1995.

Some recent characterization based goodness of fit tests

Bojana Milošević[*1]

[1] *University of Belgrade, Faculty of Mathematics*

In this paper some recent advances in goodness of fit testing are presented. Special attention is given to goodness of fit tests based on equidistribution and independence characterizations. New concepts are described through some modern exponentiality tests. Their natural generalizations are also proposed. All tests are compared in Bahadur sense.

Keywords: asymptotic efficiency, order statistics, independence,V-statistic

1 Introduction

Goodness of fit testing occupy a significant part of nonparametric statistic. Most of classical tests are based on distance between the assumed distribution function (d.f.) and its consistent estimate, empirical d.f. Symmetry tests are analogously constructed. A new approach, that is especially attractive in recent years, is making tests based on characterizations of different types. Those tests mostly use U-empirical d.f.'s, U-empirical integral transforms, U-empirical moments etc. The main advantage of these tests is that they are often free of some distribution parameters. Therefore they are suitable for testing composite hypothesis. In addition, there is an abundance of characterization theorems for some families of distributions, in particular for exponential and other life distributions, uniform, normal distribution, and characterizations of the family of symmetric distributions around zero. An extensive survey is given in classical monograph by Galambos and Kotz (see [4]) as well as in recent monograph by Ahsanullah (see [1]). Hence, many different modern goodness of fit tests (GOF) can be constructed (see [17]).

For purpose of comparison of tests, the Bahadur efficiency has become very popular. One of the reasons is that, unlike Pitman efficiency, it does not require the asymptotic normality of test statistics. For more details we refer to [14].

[*]Corresponding author: bojana@matf.bg.ac.rs

Bahadur efficiency can be expressed as the ratio of the Bahadur exact slope $c(\theta)$, a function describing the rate of exponential decrease for the attained level under the alternative, and $2K(\theta)$, the double Kullback-Leibler distance between the alternative and the set of null distributions. Under closed alternatives we use the *local Bahadur efficiency* given by

$$e = \lim_{\theta \to 0} \frac{c(\theta)}{2K(\theta)}.$$

According to Bahadur's theory slopes can be calculated in the following way. Suppose that, under an alternative indexed by a parameter θ, the sequence T_n converges in probability to some finite function $b(\theta)$. Suppose also that the large deviation limit

$$\lim_{n \to \infty} n^{-1} P_{H_0}\{T_n > t\} = -f(t)$$

exists for any t in an open interval I, on which f is continuous and $\{b(\theta), \theta > 0\} \subset I$. Then the Bahadur exact slope is

$$c(\theta) = 2f(b(\theta)).$$

Very often, the main obstacle is to find large deviation function. Crucial results concerning this problem are given in [15], [16] and [10].

In the following section we describe two types of characterization theorems and show how, using them, we can construct an integral and a Kolmogorov type statistic. In order to illustrate the general idea we provide two examples.

2 Characterizations and tests statistics

In this section we focus on so called "equidistribution based" and "independence based" characterizations and GOF tests based on them.

The first group contains the characterization of the following form.

Let $X_1, ..., X_{\max(m,p)}$ be i.i.d with d.f. F, $\omega_1 : R^m \mapsto R^1$ and $\omega_2 : R^p \mapsto R^1$ two sample functions such. Then the following equivalence hold

$$\omega_1(X_1, ..., X_m) \stackrel{d}{=} \omega_2(X_1, ..., X_p)$$

if and only if F belongs to some family $\mathcal{F}_0$.

Natural estimators of d.f.'s of ω_1 and ω_2 are V-empirical d.f.'s given by

$$G_{n1}(x) = \frac{1}{n^m} \sum_{1 \le i_1, ..., i_m \le n} I\{\omega_1(X_{i_1}, ..., X_{i_m}) < x\}$$

$$G_{n2}(x) = \frac{1}{n^p} \sum_{1 \le i_1, ..., i_p \le n} I\{\omega_2(X_{i_1}, ..., X_{i_p}) < x\}.$$

Alternatively, one can consider corresponding symmetrized U-empirical d.f.'s.

Therefore GOF tests can be constructed based on the difference between functions G_{n1} and G_{n2}. Mostly used in last few years (see e.g [18],[11],[8], etc.) are those of integral type

$$I_n = \int_{-\infty}^{\infty} (G_{n1}(x) - G_{n2}(x)) dF_n(x),$$

as well as Kolmogorov type statistic

$$K_n = \sup_x |G_{n1}(x) - G_{n2}(x)|.$$

Usually large values of statistics are significant. Under some additional conditions both statistics are often free of some distribution parameter. For example, in case of testing exponentiality sufficient condition is that ω_1 and ω_2 are homogenuous functions of sample elements.

The group of independence based characterization contains the characterizations of the following form.

Let $X_1, ..., X_m$ be i.i.d with d.f. F, $\omega_1 : R^m \mapsto R$ and $\omega_2 : R^p \mapsto R$ two sample function such that $p \leq m$. Then the following equivalence hold

$$\omega_1(X_1, ..., X_m) \text{ and } \omega_1(X_1, ..., X_p) \text{ are independent}$$

if and only if F belongs to some family $\mathcal{F}_0$.

Thus we may reformulate our null hypothesis into

$$H_0 : \quad H(x_1, x_2) = G_1(x_1)G(x_2), \quad \text{for all } x_1, x_2 \in R$$

where G_1, G_2 i H are marginal and joint d.f.'s of ω_1 and ω_2, respectively. Natural choice for test statistics is

$$\begin{aligned} I_n &= \int_{x_1,x_2} (G_n(x_1, x_2) - H_n(x_1, x_2)) dF_n(x_1) dF_n(x_2) \\ K_n &= \sup_{x_1,x_2} |G_n(x_1, x_2) - H_n(x_1, x_2)|, \end{aligned} \tag{1}$$

where

$$G_n(x_1, x_2) = G_{n1}(x_1)G_{n2}(x_2)$$

$$H_n(x_1, x_2) = \frac{1}{n^m} \sum_{1 \leq i_1, ..., i_m \leq n} I\{\omega_1(X_{i_1}, ..., X_{i_m}) < x_1, \omega_2(X_{i_1}, ..., X_{i_p}) < x_2\}$$

are suitable V-empirical d.f.'s. As previously, large values of proposed statistics are significant. Those type of GOF tests have been firstly considered by Milošević and Obradović in [10].

Notice that all integral type statistics are U-statistics, V-statistics or hybrid U-statistic. It turns up that in most cases their kernels are bounded and non-degenerate. Hence, the limiting distributions of these statistics, appropriately normalized, are normal. In case of Kolmogorov type statistics we have supremum over some family of U-statistics, usually non-degenerate in the sense of [10, 16]. Therefore their limiting distributions, appropriately normalized, coincide with that of a supremum of of absolute value of some centered Gaussian process (field). The critical values can be found using Monte Carlo simulations.

3 Examples and discussion

Recently Milošević and Obradović proved the following characterization theorem (see [6]). The theorem generalizes results from [12] based on original ideas from [2] and [19].

Let $X_1, \ldots, X_m$ be a random sample from the distribution whose density $f(x)$ has the Maclaurin expansion for $x > 0$, and let X_0 be a random variable independent of the sample that follows the same distribution. Let k be a fixed number such that $1 < k \leq m$. X is exponentially distributed if and only if one of the following three statement holds:

$$X_{(k;m)} \stackrel{d}{=} X_{(k-1;m-1)} + \frac{1}{m} X_m \tag{2}$$

$$X_{(k;m)} \stackrel{d}{=} X_{(k-1;m)} + \frac{1}{m-k+1} X_0 \tag{3}$$

$$X_{(k;m)} \stackrel{d}{=} \frac{1}{m} X_1 + \frac{1}{m-1} X_2 + \cdots + \frac{1}{m-k+1} X_k \tag{4}$$

In the spirit of the previous section many GOF tests based on this characterization theorem can be constructed. Denote with $I_k^{(j)}$ integral type, and with $K_k^{(j)}$ $(j = 1, 2, 3)$ Kolmogorov type tests based on j-th part of the theorem for $m = k$.

The tests $I_3^{(1)}$ and $K_3^{(1)}$ were proposed in [18], $I_2^{(2)}$ and $K_2^{(2)}$ in [7], $I_3^{(2)}$ and $K_3^{(2)}$ in [9], while the test $I_k^{(3)}$, $K_2^{(3)}$ and $K_3^{(3)}$ were proposed in [5]. Notice that the $I_2^{(1)}$ and $K_2^{(1)}$ coincide with $I_2^{(3)}$ and $K_2^{(3)}$, respectively. We propose $I_k^{(1)}$ and $I_k^{(2)}$ for arbitrary k and show that they are asymptotically equivalent to

U-statistics with nondegenerate symmetric kernel. We derive their asymptotic d.f.'s as well as large deviation functions.

We compare tests against some closed alternatives, namely Weibull, Makeham, Gamma, mixture of exponential distributions with negative weights (EMNW(3)) and linear failure rate (LFR) distribution. Their densities can be found e.g. in [9, 10]. The results are summarized in Tables 1 and 2.

Table 1: Bahadur efficiencies of integral type tests

	$j=1$			$j=2$			$j=3$		$I^{\mathcal{E}}$
Alt.	$e_2^{(1)}$	$e_3^{(1)}$	$\max_k e_k^{(1)}, k$	$e_2^{(2)}$	$e_3^{(2)}$	$\max_k e_k^{(2)}, k$	$e_3^{(3)}$	$\max_k e_k^{(3)}, k$	$e^{\mathcal{E}}$
Weibull	.621	.649	(.649, 3)	.750	.746	(.750, 2)	.664	(.710, 8)	.419
Makeham	.488	.654	(.783, 6)	.625	.772	(.872, 6)	.573	(.876, 14)	.714
Gamma	.723	.638	(.723, 2)	.796	.701	(.796, 2)	.708	(.723, 2)	.701
EMNW(3)	.694	.835	(.835, 3)	.844	.916	(.916, 3)	.799	(.885, 6)	.542
LFR	.104	.206	(.613, 20)	.208	.308	(.712, 23)	.159	(.804, 88)	.535

Table 2: Bahadur efficiencies of Kolmogorov type tests

	$j=1$		$j=2$		$j=3$	$K^{\mathcal{E}}$
Alt.	$e_2^{(1)}$	$e_3^{(1)}$	$e_2^{(2)}$	$e_3^{(2)}$	$e_3^{(3)}$	$e^{\mathcal{E}}$
Weibull	.092	.079	.277	.258	.152	.200
Makeham	.125	.123	.342	.370	.216	.375
Gamma	.093	.066	.267	.212	.138	.131
EMNW(3)	.149	.122	.396	.364	.230	.334
LFR	.052	.067	.155	.213	.106	.235

Now we pass to the example of GOF tests via independence based characterization. Fisz in [3] proved following theorem.

Let X and Y be i.i.d. random variables with continuous distribution function F. Then $\min\{X, Y\}$ *and* $|X - Y|$ *are independent if and only if* $F(x) = 1 - e^{-\lambda x}$*, for some positive constant* λ.

The Kolmogorov type test $K^{\mathcal{E}}$ based on this characterization have been proposed in [10]. Beside this test we propose the corresponding integral type test $I^{\mathcal{E}}$ of the form (1). We prove that the limiting distribution is normal and found the large deviation function. Bahadur efficiencies of tests are shown in Tables 1 and 2.

From the tables we can conclude that the "order" of proposed tests differ with alternative, the order of corresponding U-statistic and type of characterization and characterization itself. As far as Milošević-Obradović characterization based tests are considered, in case $m = k$ it can be noticed that for fixed k tests based on the second case are most efficient. However, it does not necessary hold for some other choices of m and k and alternative distributions.

The results for tests based on Fisz's characterizations are reasonably good in comparison to considered tests based on Milošević-Obradović characterization. It confirms that this rather new approach has a potential.

General conclusion is that the integral type tests are much more efficient then the corresponding Kolmogorov ones. On the other hand the Kolmogorov type tests are consistent against all alternatives, while the integral type tests can be made consistent against almost all alternatives of practical purpose considering their two-tailed versions.

Acknowledgements: The research was supported by MNTRS, Serbia, Grant No. 174012.

References

[1] M. Ahsanullah, *Characterizations of Univariate Continuous Distributions.* Atlantis Studies in Probability and Statistics ser. 7., Atlantis Press, 2017.

[2] B. C. Arnold, J. A. Villasenor, Exponential characterizations motivated by the structure of order statistics in samples of size two. *Statist. Probab. Lett.* 83(2):596–601, 2013.

[3] M. Fisz, Characterization of some probability distributions. *Skand. Aktuarietidskr* 41(1-2):65–67, 1958.

[4] J. Galambos, S. Kotz, *Characterizations of Probability Distributions*, Springer-Verlag, Berlin-Heidelberg-New York, 1978.

[5] M. Jovanović, B. Milošević, Ya. Yu. Nikitin, M. Obradović and K. Yu. Volkova, Tests of exponentiality based on Arnold-Villasenor characterization, and their efficiencies. *Comput. Statist. Data Anal.* 90:100–113, 2015.

[6] B. Milošević, M. Obradović, Some characterizations of the exponential distribution based on order statistics. *Appl. Anal. Discrete Math.* 10:394–407, 2016.

[7] B. Milošević, M. Obradović, Some characterization based exponentiality tests and their Bahadur Efficiencies. *Publ. Inst. Math* 100(114):107–117, 2016.

[8] B. Milošević, M. Obradović, Characterization based symmetry tests and their asymptotic efficiencies. *Statist. Probab. Lett.* 119:155–162, 2016.

[9] B. Milošević, Asymptotic Efficiency of New Exponentiality Tests Based on a Characterization, *Metrika.* 79(2): 221–236, 2016.

[10] B. Milošević, M. Obradović, Two-dimensional Kolmogorov-type Goodness-of-fit Tests Based on Characterizations and their Asymptotic Efficiencies. *J. Nonparametr. Stat.* 28(2):413–427, 2016.

[11] M. Obradović, On Asymptotic Efficiency of Goodness of Fit Tests for Pareto Distribution Based on Characterizations. *Filomat* 9(10):2311–2324, 2015.

[12] M. Obradović, Three Characterizations of Exponential Distribution Involving Median of Sample of Size Three, *J. Stat. Theory Appl.* 14(3):257–264, 2015.

[13] M. Obradović, M. Jovanović, B. Milošević, Goodness-of-fit tests for Pareto distribution based on a characterization and their asymptotics. *Statistics* 49(5):1026–1041, 2015.

[14] Ya. Yu. Nikitin, *Asymptotic Efficiency of Nonparametric Tests*, Cambridge University Press, New York, 1995.

[15] Ya. Yu. Nikitin, E. V. Ponikarov, Rough large deviation asymptotics of Chernoff type for von Mises functionals and U-statistics. *Proceedings of Saint-Petersburg Mathematical Society* 7 (1999) 124–167; English translation in AMS Translations, ser. 2, 203:107–146, 2001.

[16] Ya. Yu. Nikitin, Large Deviations of U-empirical Kolmogorov-Smirnov Test, and Their Efficiency. *J. Nonparametr. Stat.* 22(5):649–668, 2010.

[17] Ya. Yu. Nikitin, Tests based on characterizations, and their efficiencies: a survey. *Acta Comment. Univ. Tartu. Math.* 21(1):4-24, 2017.

[18] K. Volkova, Goodness-Of-Fit Tests for Exponentiality Based on Yanev-Chakraborty Characterization and Their Efficiencies. *19th EYSM* 2015.

[19] G. P. Yanev, S. Chakraborty. Characterizations of exponential distribution based on sample of size three. *Pliska Stud. Math. Bulgar.* 22(1): 237–244, 2013.

Confidence regions in Cox proportional hazards model with measurement errors and unbounded parameter set

Oksana Chernova[*1]

[1] *Taras Shevchenko National University of Kyiv, Ukraine*

Cox proportional hazards model with measurement errors in covariates is considered. It is the ubiquitous technique in biomedical data analysis. In Kukush et al. (2011) and Chimisov & Kukush (2014) asymptotic properties of a simultaneous estimator $(\lambda_n; \beta_n)$ for the baseline hazard rate $\lambda(\cdot)$ and the regression parameter β were studied, at that the parameter set $\Theta = \Theta_\lambda \times \Theta_\beta$ was assumed bounded.

In Kukush & Chernova (2017) we dealt with the simultaneous estimator $(\lambda_n; \beta_n)$ in the case, where the Θ_λ was unbounded from above and not separated away from 0. The estimator was constructed in two steps: first we derived a strongly consistent estimator and then modified it to provide its asymptotic normality.

In this talk, we construct the confidence region for β. We reach our goal in each of the three cases: (a) the measurement error is bounded, (b) it is normally distributed, or (c) it is a shifted Poisson random variable. The censor is assumed to have a continuous pdf. In future research we intend to elaborate a method for heavy tailed error distributions and construct the confidence interval for an integral functional of $\lambda(\cdot)$.

Keywords: asymptotic normality, confidence region, consistent estimator, Cox proportional hazards model, measurement errors, simultaneous estimation of baseline hazard rate and regression parameter.

1 Model formulation and estimation

Consider the Cox proportional hazards model (Cox, 1972), where a lifetime T has the following intensity function

$$\lambda(t|X; \lambda, \beta) = \lambda(t) \exp(\beta^T X), \quad t \geq 0.$$

[*]Corresponding author: chernovaoksan@gmail.com

A covariate X is a given random vector distributed in $\mathbb{R}^m$, β is a parameter belonging to $\Theta_\beta \subset \mathbb{R}^m$, and $\lambda(\cdot) \in \Theta_\lambda \subset C[0,\tau]$, $\tau > 0$, is a baseline hazard function.

Instead of lifetime T one can usually observe a censored lifetime $Y := \min\{T, C\}$ and the censorship indicator $\Delta := I_{\{T \le C\}}$. The censor C is distributed on $[0,\tau]$. Its survival function $G_C(u) = 1 - F_C(u)$ is unknown, while we know τ. The conditional pdf of T given X is

$$f_T(t|X,\lambda,\beta) = \lambda(t|X;\lambda,\beta)\exp\left(-\int_0^t \lambda(t|X;\lambda,\beta)ds\right).$$

We consider an additive error model, i.e., instead of X a surrogate variable

$$W = X + U$$

is observed, where a random error U has known moment generating function $M_U(z) := \mathsf{E}e^{z^T U}$. A couple (T, X), censor C, and measurement error U are stochastically independent.

Consider independent copies of the model $(X_i, T_i, C_i, Y_i, \Delta_i, U_i, W_i)$, $i = 1, ..., n$. Based on triples (Y_i, Δ_i, W_i), $i = 1, ..., n$, we estimate true parameters β_0 and $\lambda_0(t)$, $t \in [0,\tau]$. Our model is presented in Augustin (2004) where baseline hazard function is assumed to belong to a parametric space, while we consider $\lambda_0(\cdot)$ from a closed convex subset of $C[0,\tau]$. Following the idea from Augustin (2004), we use the objective function

$$Q_n^{cor}(\lambda,\beta) := \frac{1}{n}\sum_{i=1}^n q(Y_i,\Delta_i,W_i;\lambda,\beta),$$

with

$$q(Y,\Delta,W;\lambda,\beta) := \Delta\cdot(\log\lambda(Y) + \beta^T W) - \frac{\exp(\beta^T W)}{M_U(\beta)}\int_0^Y \lambda(u)du.$$

Introduce the basic assumptions.

(i) $\Theta_\lambda \subset C[0,\tau]$ is the following closed convex set of nonnegative functions

$$\Theta_\lambda := \{\ f : [0,\tau] \to \mathbb{R} \mid f(t) \ge 0, \forall t \in [0,\tau] \text{ and}$$

$$|f(t) - f(s)| \le L|t-s|, \forall t, s \in [0,\tau]\ \},$$

where $L > 0$ is a fixed constant.

(ii) $\Theta_\beta \subset \mathbb{R}^m$ is a compact set.

(iii) $\mathsf{E}U = 0$ and for some constant $\epsilon > 0$,

$$\mathsf{E}e^{D\|U\|} < \infty, \text{ where } D := \max_{\beta \in \Theta_\beta} \|\beta\| + \epsilon.$$

(iv) $\mathsf{E}e^{D\|X\|} < \infty$, with D defined in (iii).

(v) $\mathsf{P}(C > \tau) = 0$ and for all $\epsilon > 0$, $\mathsf{P}(C > \tau - \epsilon) > 0$.

(vi) The covariance matrix of random vector X is positive definite.

Denote

$$\Theta = \Theta_\lambda \times \Theta_\beta. \tag{1}$$

(vii) True parameters (λ_0, β_0) belong to Θ, which is given in (1), and moreover $\lambda_0(t) > 0$, $t \in [0, \tau]$.

(viii) β_0 is an interior point of Θ_β.

(ix) $\lambda_0 \in \Theta_\lambda^\epsilon$ for some $\epsilon > 0$, where

$$\Theta_\lambda^\epsilon := \{\ f : [0,\tau] \to \mathbb{R} \mid f(t) \geq \epsilon,\ \forall t \in [0,\tau],$$
$$|f(t) - f(s)| \leq (L - \epsilon)|t - s|, \forall t, s \in [0,\tau]\ \}.$$

(x) $\mathsf{P}(C > 0) = 1$.

(xi) For some $\epsilon > 0$, $\mathsf{E}e^{2D\|U\|} < \infty$ where D is defined in (iii).

(xii) $\mathsf{E}e^{2D\|X\|} < \infty$ where D is defined in (iii).

Definition 1. Fix a sequence $\{\varepsilon_n\}$ of positive numbers, with $\varepsilon_n \downarrow 0$, as $n \to \infty$. The corrected estimator $\left(\hat{\lambda}_n^{(1)}, \hat{\beta}_n^{(1)}\right)$ of (λ, β) is a Borel measurable function of observations (Y_i, Δ_i, W_i), $i = 1, ..., n$, with values in Θ and such that

$$Q_n^{cor}\left(\hat{\lambda}_n^{(1)}, \hat{\beta}_n^{(1)}\right) \geq \sup_{(\lambda,\beta)\in\Theta} Q_n^{cor}(\lambda, \beta) - \varepsilon_n.$$

Theorem 3 from Kukush & Chernova (2017) proves that under conditions (i) to (vii), the couple $\left(\hat{\lambda}_n^{(1)}, \hat{\beta}_n^{(1)}\right)$ is a strongly consistent estimator of the true parameters (λ_0, β_0).

Based on $\left(\hat{\lambda}_n^{(1)}, \hat{\beta}_n^{(1)}\right)$, we derive a modified estimator $\left(\hat{\lambda}_n^{(2)}, \hat{\beta}_n^{(2)}\right)$ to be consistent and asymptotically normal.

Definition 2. The modified corrected estimator $\left(\hat{\lambda}_n^{(2)}, \hat{\beta}_n^{(2)}\right)$ of (λ, β) is a Borel measurable function of observations (Y_i, Δ_i, W_i), $i = 1, ..., n$, with values in Θ and such that

$$\left(\hat{\lambda}_n^{(2)}, \hat{\beta}_n^{(2)}\right) = \begin{cases} \arg\max \left\{ Q_n^{cor}(\lambda, \beta) \mid (\lambda, \beta) \in \Theta,\ \mu_\lambda \geq \frac{1}{2}\mu_{\hat{\lambda}_n^{(1)}} \right\}, & \text{if } \mu_{\hat{\lambda}_n^{(1)}} > 0, \\ \left(\hat{\lambda}_n^{(1)}, \hat{\beta}_n^{(1)}\right), & \text{otherwise,} \end{cases}$$

with $\mu_\lambda := \min\limits_{t \in [0,\tau]} \lambda(t)$.

Introduce notations:

$$a(t) = \mathsf{E}[Xe^{\beta_0^T X} G_T(t|X)], \quad b(t) = \mathsf{E}[e^{\beta_0^T X} G_T(t|X)], \quad \Lambda(t) = \int_0^t \lambda_0(t)dt,$$

$$p(t) = \mathsf{E}[XX^T e^{\beta_0^T X} G_T(t|X)], \quad T(t) = p(t)b(t) - a(t)a^T(t), \quad K(t) = \frac{\lambda_0(t)}{b(t)},$$

$$M = \int_0^\tau T(u)K(u)G_c(u)du.$$

For $i = 1, 2, \ldots,$ introduce random variables

$$\zeta_i = -\frac{\Delta_i \cdot a(Y_i)}{b(Y_i)} + \frac{\exp(\beta_0^T W_i)}{M_U(\beta_0)} \int_0^{Y_i} a(u)K(u)du + \frac{\partial q}{\partial \beta}(Y_i, \Delta_i, W_i, \beta_0, \lambda_0),$$

with

$$\frac{\partial q}{\partial \beta}(Y, \Delta, W; \lambda, \beta) = \Delta \cdot W - \frac{M_U(\beta)W - \mathsf{E}(Ue^{\beta^T U})}{M_U(\beta)^2} \exp(\beta^T W) \int_0^Y \lambda(u)du.$$

Let

$$\Sigma_\beta = 4 \cdot \mathsf{Cov}(\zeta_1).$$

The following theorem from Kukush & Chernova (2017) states asymptotic normality of $\hat{\beta}_n^{(2)}$.

Theorem 1. *Assume conditions (i), (ii), and (v) – (xii). Then M is nonsingular and*

$$\sqrt{n}(\hat{\beta}_n^{(2)} - \beta_0) \xrightarrow{d} N_m(0, M^{-1}\Sigma_\beta M^{-1}). \tag{2}$$

2 Confidence region for the regression parameter

Based on Theorem 1, we construct a confidence region for the regression parameter. For $t \in [0, \tau]$ and $\beta \in \Theta_\beta$, denote

$$B(W,t;\lambda,\beta) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k! M_U((k+1)\beta)} \Lambda^k(t) e^{(k+1)\beta^T W},$$

$$A(W,t;\lambda,\beta) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k! M_U((k+1)\beta)} \Lambda^k(t) \left[W - \frac{\mathsf{E}[U e^{(k+1)\beta^T U}]}{M_U((k+1)\beta)} \right] e^{(k+1)\beta^T W},$$

$$P(W,t;\lambda,\beta) = \sum_{k=0}^{\infty} \frac{(-1)^k \Lambda^k(t)}{k! M_U((k+1)\beta)} \left[W W^T e^{(k+1)\beta^T W} - 2 \frac{\mathsf{E}[U e^{(k+1)\beta^T U}]}{M_U((k+1)\beta)} W^T e^{(k+1)\beta^T W} - \right.$$
$$\left. - \left(\frac{\mathsf{E}[U U^T e^{(k+1)\beta^T U}]}{M_U((k+1)\beta)} - 2 \frac{\mathsf{E}[U e^{(k+1)\beta^T U}] \mathsf{E}[U^T e^{(k+1)\beta^T U}]}{M_U^2((k+1)\beta)} \right) e^{(k+1)\beta^T W} \right].$$

Theorem 2. *Suppose that*

$$\sum_{k=0}^{\infty} \frac{\tilde{c}_{k+1}(\beta)}{k!} e^{k\beta^T z} < \infty, \qquad z \in \mathbb{R}^m, \qquad \beta \in \Theta_\beta,$$

where $\tilde{c}_{k+1}(\beta)$ *is substituted with each of the following expressions*

$$\frac{\mathsf{E}[||U|| e^{(k+1)\beta^T U}]}{M_U((k+1)\beta)}, \quad \frac{\mathsf{E}[||U||^2 e^{(k+1)\beta^T U}]}{M_U((k+1)\beta)}, \quad \left(\frac{\mathsf{E}[||U|| e^{(k+1)\beta^T U}]}{M_U((k+1)\beta)} \right)^2.$$

Then for all $t \in [0, \tau]$,

$$\hat{b}(t) = \frac{1}{n} \sum_{i=1}^{n} B(W_i, t; \hat{\lambda}_n^{(2)}, \hat{\beta}_n^{(2)}),$$

$$\hat{a}(t) = \frac{1}{n} \sum_{i=1}^{n} A(W_i, t; \hat{\lambda}_n^{(2)}, \hat{\beta}_n^{(2)}),$$

$$\hat{p}(t) = \frac{1}{n} \sum_{i=1}^{n} P(W_i, t; \hat{\lambda}_n^{(2)}, \hat{\beta}_n^{(2)})$$

are consistent estimators of $b(t)$, $a(t)$ *and* $p(t)$, *respectively.*

We point out that the conditions of Theorem 2 are fulfilled in the next cases: (a) U is bounded, (b) it is normally distributed, or (c) it is a shifted Poisson random variable.

Denote

$$\hat{M} = \int_0^{Y_{(n)}} \hat{T}(u)\hat{K}(u)\hat{G}_C(u)du, \tag{3}$$

where $\hat{G}_C$ is the Kaplan-Meier estimator of the survival function of censor C, and

$$\hat{\Sigma}_\beta = \frac{4}{n-1}\sum_{i=1}^{n} \hat{\zeta}_i \hat{\zeta}_i^T, \quad \text{with} \tag{4}$$

$$\hat{\zeta}_i := -\frac{\Delta_i \cdot \hat{a}(Y_i)}{\hat{b}(Y_i)} + \frac{\exp(\hat{\beta}_n^{(2)T} W_i)}{M_U(\hat{\beta}_n^{(2)})} \int_0^{Y_i} \hat{a}(u)\hat{K}(u)du + \frac{\partial q}{\partial \beta}(Y_i, \Delta_i, W_i; \hat{\beta}_n^{(2)}, \hat{\lambda}_n^{(2)}).$$

Theorem 3. *Assume the conditions of Theorem 2. The estimators $\hat{M}$ and $\hat{\Sigma}_\beta$ defined in (3) and (4) are consistent estimators of matrices M and Σ_β, respectively.*

Given a confidence probability $1-\alpha$, $0 < \alpha < 1/2$, let $\left(\chi^2_m\right)_\alpha$ be the upper α-quantile of the χ^2_m distribution. Based on Theorem 2 and 3, we take as an asymptotic confidence ellipsoid for $\hat{\beta}_n^{(2)}$ the set

$$E_n := \left\{ z \in \mathbb{R}^m \mid \left(z - \hat{\beta}_n^{(2)}\right)^T \left(\hat{M}^{-1}\hat{\Sigma}_\beta \hat{M}^{-1}\right)^{-1} \left(z - \hat{\beta}_n^{(2)}\right) \le \frac{1}{n}\left(\chi^2_m\right)_\alpha \right\}.$$

It holds $\mathsf{P}(\beta_0 \in E_n) \to 1-\alpha$ as $n \to \infty$.

Acknowledgements: I would like to express my sincere gratitude to my advisor Prof. Alexander Kukush for the support of my PhD research.

References

[1] Augustin, T. (2004). An exact corrected log-likelihood function for Cox's proportional hazards model under measurement error and some extensions. *Scandinavian Journal of Statistics* **31**, 1, 43–50.

[2] Chimisov, C., and Kukush, A. (2014). Asymptotic normality of corrected estimator in Cox proportional hazards model with measurement error. *Modern Stochastics: Theory and Applications* **1**, 1, 13–32.

[3] Kukush, A., Baran, S., Fazekas, I., and Usoltseva, E. (2011). Simultaneous estimation of baseline hazard rate and regression parameters in Cox proportional hazards model with measurement error. *Journal of Statistical Research* **45**, 2, 77–94.

[4] Kukush, A., and Chernova, O. (2017). Consistent estimation in Cox proportional hazards model with measurement errors and unbounded parameter set. *Theory of Probability and Mathematical Statistics* **96**, 100–109.

E-optimal approximate block designs for treatment-control comparisons

Samuel Rosa[*1]

[1]*Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava, Slovakia*

We study E-optimal block designs for comparing a set of test treatments with a control treatment. We provide the complete class of all E-optimal approximate block designs and we show that these designs are characterized by simple linear constraints. Employing the provided characterization, we obtain a class of E-optimal exact block designs with unequal block sizes for comparing test treatments with a control.

Keywords: Optimal design, Block design, Approximate design, Control treatment, E-optimality

1 Introduction

Consider a blocking experiment for comparing a set of test treatments with a control. As noted in [4], the experimental objective of comparing the test treatments with a control arises, for instance, in screening experiments or in experiments in which it is desired to assess the relative performance of new test treatments with respect to the standard treatment. Such objective is also quite natural for medical studies involving placebo (e.g., see [12], [11]).

Formally, we have

$$Y_j = \mu + \tau_{i(j)} + \theta_{k(j)} + \varepsilon_j, \quad j = 1, \ldots, n,$$

where μ is the overall mean, τ_i is the effect of the i-th treatment ($0 \leq i \leq v$), θ_k is the effect of the k-th block ($1 \leq k \leq d$), and the random errors $\varepsilon_1, \ldots, \varepsilon_n$ are uncorrelated, with zero mean and variance $\sigma^2 < \infty$. Treatment 0 denotes the control, and the test treatments are numbered $1, \ldots, v$. By τ, we denote the vector of treatment effects and by θ the vector of block effects. The assumed objective of the experiment is to estimate the comparisons of the test treatments with the control $\tau_i - \tau_0$ ($1 \leq i \leq v$) or comparisons with the control in short. Let $Q := (-1_v, I_v)^T$, where 1_v is the column vector of ones of

*Corresponding author: samuel.rosa@fmph.uniba.sk

length v and I_v is the identity matrix. Then, the experimental objective can be expressed as the estimation of $Q^T\tau$.

There is a large amount of literature on optimal exact designs for test treatment-control comparisons, mostly considering the A- and MV-optimality criteria; for a survey, see [4] or [5]. The E-optimality criterion also received some attention; see [6], [8], [7].

In this paper, we provide the class of *all* E-optimal approximate block designs for comparisons with the control. Based on the obtained class of optimal approximate designs, we provide a class of E-optimal exact designs, which extends the known results on E-optimality to the case of unequal block sizes.

1.1 Experimental design

An *exact design* ξ_e determines in each block the numbers of trials that are performed with the various treatments. Thus, ξ_e can be expressed as a function $\xi_e : \{0,\dots,v\} \times \{1,\dots,d\} \to \{0,1,2,\dots,n\}$ such that $\sum_{i,k} \xi_e(i,k) = n$. The value $\xi_e(i,k)$ determines the number of trials performed with treatment i in block k. Suppose that the blocks $1,\dots,d$ have pre-specified non-zero sizes $m_1,\dots,m_d$. We denote the class of all block designs for $v+1$ treatments and d blocks of sizes $m = (m_1,\dots,m_d)^T$ by $D(v,d,m)$.

An *approximate design* (or simply a *design*) is a function $\xi : \{0,\dots,v\} \times \{1,\dots,d\} \to [0,1]$, such that $\sum_{i,k} \xi(i,k) = 1$. The value $\xi(i,k)$ represents the proportion of all trials that are performed with treatment i in block k. For a given design ξ, let us denote the design matrix $X(\xi) := (\xi(i,k))_{i,k}$, let $r(\xi) := X(\xi)1_d$ be the vector of total treatment proportions and let $s(\xi) := X^T(\xi)1_v$ be the vector of relative block sizes. Because we consider non-zero block sizes, we always have $s(\xi) > 0$.

The information matrix of a design ξ for estimating all pairwise comparisons of treatments is $M(\xi) := \mathrm{diag}(r(\xi)) - X(\xi)\mathrm{diag}^{-1}(s(\xi))X^T(\xi)$, where $\mathrm{diag}^{-1}(s(\xi)) := \mathrm{diag}(s_1^{-1}(\xi),\dots,s_d^{-1}(\xi))$. The parameter system $Q^T\tau$ is said to be estimable under an approximate design ξ if $\mathcal{C}(Q) \subseteq \mathcal{C}(M(\xi))$, where $\mathcal{C}$ denotes the column space. In such a case, we say that ξ is feasible and we have $\mathrm{rank}(M(\xi)) = v$. The information matrix $N(\xi) := (Q^T M^-(\xi) Q)^{-1}$ of a feasible design ξ for estimating $Q^T\tau$ is obtained by deleting the first row and column of $M(\xi)$ (see [1], [3]). Let us partition $X(\xi)$ as $X^T(\xi) = (z(\xi), Z^T(\xi))$, where $z(\xi)$ is a $d \times 1$ vector; i.e., $Z(\xi) = (\xi(i,k))_{i>0,k}$. Then, the information matrix for comparing the test treatments with the control is

$$N(\xi) = \mathrm{diag}(r_1(\xi),\dots,r_v(\xi)) - Z(\xi)\mathrm{diag}^{-1}(s(\xi))Z^T(\xi). \tag{1}$$

Note that $N(\xi)$ is proportional to the inverse of the covariance matrix of the

least squares estimator of $\tau_1 - \tau_0, \ldots, \tau_v - \tau_0$. A design is said to be Ψ-optimal if it minimizes $\Psi(N(\xi))$ for some function Ψ.

We say that a design ξ that satisfies $\xi(i,k) = r_i s_k$ is a *product design* of r and s. We denote such design as $\xi = r \otimes s$.

2 *E*-optimality

We denote the largest (smallest) eigenvalue of a symmetric matrix A by $\lambda_{\max}(A)$ ($\lambda_{\min}(A)$). A design is E-optimal if it minimizes $\lambda_{\max}(N^{-1}(\xi))$ or, equivalently, if it maximizes $\lambda_{\min}(N(\xi))$. Such a design minimizes the maximum variance for the linear combinations $\sum_{i>0} x_i\tau_i - (\sum_{i>0} x_i)\tau_0$ over all normalized $x \in \mathbb{R}^v$. In the following theorem, we provide the complete characterization of E-optimal block designs for comparing the test treatments with the control: an approximate design ξ^* is E-optimal for the comparisons with the control if and only if

(i) in each block, ξ^* assigns one half of the trials to the control and

(ii) ξ^* is equireplicated in the test treatments.

Theorem 1. *An approximate block design ξ is E-optimal for the comparisons with the control if and only if it satisfies*

$$\xi(0,k) = \frac{s_k(\xi)}{2} \quad \textit{and} \quad r_1(\xi) = \ldots = r_v(\xi) = \frac{1}{2v}. \tag{2}$$

Proof. Let ξ be E-optimal. From Theorems 1 and 6 of [10] it follows that an E-optimal design must satisfy $r_0(\xi) = 1/2$ and $r_i(\xi) = 1/(2v)$ for $i > 0$, and that the optimal value of $\lambda_{\min}(N(\xi))$ is $\lambda^*_{\min} = 1/(4v)$. Moreover,

$$\begin{aligned}\lambda_{\min}(N(\xi)) &= \min_{x^T x=1} x^T N(\xi) x \le \frac{1}{v} 1_v^T N(\xi) 1_v \\ &= \frac{1}{v}\left(\sum_{i>0} r_i(\xi) - \sum_{k=1}^{d} \frac{1}{s_k(\xi)} (\sum_{i>0} \xi(i,k))^2\right) \\ &= \frac{1}{2v} - \frac{1}{v}\sum_{k=1}^{d} \frac{(q_k)^2}{s_k(\xi)},\end{aligned}$$

where $q_k := \sum_{i>0} \xi(i,k)$ $(1 \le k \le d)$. Because $\sum_k \xi(0,k) = 1/2$, we have $\sum_k q_k = 1/2$. Therefore, for fixed $s(\xi)$, the following holds (which can be seen by finding the minimum of the function on the left-hand side):

$$\sum_{k=1}^{d} \frac{q_k^2}{s_k(\xi)} \ge \sum_{k=1}^{d} \frac{(s_k(\xi)/2)^2}{s_k(\xi)} = \frac{1}{4},$$

$i\backslash k$	1	2	3	4
0	1/6	1/6	1/12	1/12
1	1/6	1/12	0	0
2	0	1/12	1/12	1/12

Table 1: E-optimal approximate block design ξ for comparing two test treatments with the control in 4 blocks of relative sizes $s = (1/3, 1/3, 1/6, 1/6)^T$. The value on position (i, k) represents $\xi(i, k)$.

using the fact that $\sum_k s_k(\xi) = 1$. The inequality is attained as equality if and only if $q_k = s_k(\xi)/2$ for all $k = 1, \ldots, d$. Hence,

$$\lambda_{\min}(N(\xi)) \leq \frac{1}{2v} - \frac{1}{4v} = \frac{1}{4v} = \lambda^*_{\min}. \tag{3}$$

Since ξ is E-optimal, the inequality is attained as equality, and thus $\xi(1, k) = s_k(\xi)/2$ for all $k = 1, \ldots, d$.

For the converse part, let ξ satisfy (2). Then, ξ is connected (see [2]) and therefore feasible. Moreover, $Z^T(\xi)1_v = s(\xi)/2$ and $Z(\xi)1_d = (2v)^{-1}1_v$. Therefore,

$$N(\xi)1_v = \frac{1}{2v}1_v - \frac{1}{2}Z(\xi)\mathrm{diag}^{-1}(s(\xi))s(\xi) = \frac{1}{2v}1_v - \frac{1}{2}Z(\xi)1_d.$$

Thus, $N(\xi)1_v = [1/(2v) - 1/(4v)]1_v = (4v)^{-1}1_v$. That is, $\lambda^* = 1/(4v)$ is an eigenvalue of $N(\xi)$ corresponding to the eigenvector 1_v. Therefore, it suffices to prove that λ^* is the smallest eigenvalue of $N(\xi)$.

Let $N(\xi) = (n_{ij})_{i,j}$. We note that $n_{ij} \leq 0$ for $i \neq j$. Using an argument similar to that in Theorem 3.1 of [6], let x be an eigenvector of $N(\xi)$. Let us denote the eigenvalue that corresponds to x as λ. By multiplying x by an appropriate constant, we obtain $\max_j |x_j| = 1$. Thus, $x_j \leq 1$ for all $1 \leq j \leq v$. Let i be the index that satisfies $|x_i| = 1$. Then, by multiplying x by ± 1, we obtain $x_i = 1$. Now, we can write

$$(N(\xi)x)_i = n_{ii}x_i + \sum_{j \neq i} n_{ij}x_j \geq n_{ii} + \sum_{j \neq i} n_{ij} = (N(\xi)1_v)_i,$$

where the inequality follows from $n_{ij} \leq 0$ for $j \neq i$, and $x_j \leq 1$ for $1 \leq j \leq v$. Because $(N(\xi)x)_i = \lambda x_i = \lambda$ and $(N(\xi)1_v)_i = \lambda^*$, we have $\lambda^* \leq \lambda$ for any eigenvalue λ. □

Table 1 gives an E-optimal block design provided by Theorem 1. Theorem 1 is a generalization of Theorems 1 and 2 of [11], where E-optimal block designs

$i\backslash k$	1	2	3
0	1	1	2
1	1	0	1
2	0	1	1

Table 2: Exact block design ξ_e for given block sizes $m = (2, 2, 4)^T$, which is E-optimal for comparing two test treatments with the control. The value on position (i, k) represents $\xi_e(i, k)$.

for comparisons with the placebo (control) for specific experimental settings are provided.

3 Exact Designs

For a strictly convex criterion, the only optimal approximate block designs are product designs with optimal treatment proportions (see [10]). For example, for given relative block sizes s, the product design $\xi^* = r^* \otimes s$, where

$$r_0^* = \frac{\sqrt{v} - 1}{v - 1}, \quad r_1^* = \ldots = r_v^* = \frac{\sqrt{v} - 1}{\sqrt{v}(v - 1)},$$

is the single A-optimal, as well as the single MV-optimal design, see [3], [10]. It is rather difficult to obtain optimal or efficient exact designs from such designs, e.g., by rounding methods (see Chapter 12 of [9]).

However, since E-optimality lacks strict convexity, the class of E-optimal designs is richer, and efficient exact designs can be obtained by the rounding methods more easily. Moreover, this allows for a simple construction of optimal exact designs for a wide range of experimental settings. We easily obtain the following theorem that provides a class of E-optimal exact designs for unequal block sizes.

Theorem 2. *If there exists an exact design $\xi_e^* \in D(v, d, m)$ that satisfies $\xi_e^*(0, k) = m_k/2$, $1 \le k \le d$, and ξ_e^* is equireplicated in the test treatments, then ξ_e^* is E-optimal for test treatment-control comparisons in $D(v, d, m)$.*

Proof. The approximate version ξ_e^*/n of ξ_e^* is in fact an E-optimal approximate design, because it satisfies the conditions of Theorem 1. Then, ξ_e is clearly an E-optimal exact design, because the class of approximate designs is a relaxation of the class of the "normalized" exact designs ξ_e/n. □

Theorem 2 generalizes Theorem 3.1 of [6], which provides E-optimal block designs for blocks of equal size, to blocks of unequal sizes. An E-optimal design given by Theorem 2 is provided in Table 2.

Acknowledgements: This work was supported by the Slovak Scientific Grant Agency [grant VEGA 1/0521/16].

References

[1] R. E. Bechhofer and A. C. Tamhane. Incomplete block designs for comparing treatments with a control: General theory. *Technometrics*, 23(1):45–57, 1981.

[2] J. A. Eccleston and A. Hedayat. On the theory of connected designs: characterization and optimality. *Ann. Stat.*, 2(6):1238–1255, 1974.

[3] A. Giovagnoli and H. P. Wynn. Schur-optimal continuous block designs for treatments with a control. In L. M. Le Cam and R. A. Olshen, editors, *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, pages 651–666, California, 1985. Wadsworth.

[4] A. S. Hedayat, M. Jacroux, and D. Majumdar. Optimal designs for comparing test treatments with controls. *Stat. Sci.*, 3(4):462–476, 1988.

[5] D. Majumdar. Optimal and efficient treatment-control designs. In S. Ghosh and C. R. Rao, editors, *Handbook of statistics 13: Design and Analysis of Experiments*, pages 1007–1053. North Holland, Amsterdam, 1996.

[6] D. Majumdar and W. I. Notz. Optimal incomplete block designs for comparing treatments with a control. *Ann. Stat.*, 11(1):258–266, 1983.

[7] J. P. Morgan and X. Wang. E-optimality in treatment versus control experiments. *J. Stat. Theory Pract.*, 5(1):99–107, 2011.

[8] W. I. Notz. Optimal designs for treatment-control comparisons in the presence of two-way heterogeneity. *J. Stat. Plan. Infer.*, 12:61–73, 1985.

[9] F. Pukelsheim. *Optimal design of experiments*. SIAM, Philadelphia, 2006.

[10] S. Rosa and R. Harman. Optimal approximate designs for estimating treatment contrasts resistant to nuisance effects. *Stat. Pap.*, 57(4):1077–1106, 2016.

[11] S. Rosa and R. Harman. Optimal approximate designs for comparison with control in dose-escalation studies. *TEST*, 2017 (to appear).

[12] S. J. Senn. *Statistical Issues in Drug Development*. Wiley, Chichester, 1997.

Information criteria for structured sparse variable selection

Bastien Marquis[*1] **and Maarten Jansen**[1]

[1] *Université Libre de Bruxelles, departments of Mathematics*

In contrast to the low dimensional case, variable selection under the assumption of sparsity in high dimensional models is strongly influenced by the effects of false positives. The effects of false positives are tempered by combining the variable selection with a shrinkage estimator, such as in the lasso, where the selection is realized by minimizing the sum of squared residuals regularized by an ℓ_1 norm of the selected variables. Optimal variable selection is then equivalent to finding the best balance between closeness of fit and regularity, i.e., to optimization of the regularization parameter with respect to an information criterion such as Mallows's Cp or AIC. For use in this optimization procedure, the lasso regularization is found to be too tolerant towards false positives, leading to a considerable overestimation of the model size. Using an ℓ_0 regularization instead requires careful consideration of the false positives, as they have a major impact on the optimal regularization parameter. As the framework of the classical linear model has been analysed in previous work, the current paper concentrates on structured models and, more specifically, on grouped variables. Although the imposed structure in the selected models can be understood to somehow reduce the effect of false positives, we observe a qualitatively similar behavior as in the unstructured linear model.

Keywords: variable selection, structured data, sparsity, lasso, Mallows's Cp.

1 Introduction

Recent literature has had considerable attention for the uncertainties that follow from the process of model or variable selection. On one hand, it has been realized that the selection of variables should look forward, focussing on the application in which the selected model will be used, so as not to waste degrees of freedom on variables that are of little importance in the application [2]. On the other hand, post-model selection inference is looking backwards,

*Corresponding author: bastien.marquis@ulb.ac.be

investigating the effects of the model selection uncertainty on the inference in the selected model [7, 6].

The contribution of this paper is, however, situated on the effect of the uncertainty on the variable selection process itself. The numerous insignificant components in sparse, high dimensional models lead to false positives being a main source of uncertainty. Well established methods for high dimensional variable selection are explicitly based on controlling the false discovery rate [1] or even the absolute number of false positives [4]. The methods in this class tend to be minimax oriented, rather than data driven. Another way to deal with false positives is to reduce the impact of a false positive by using shrinkage selection. This is realized, for instance, in the lasso, where the variable selection objective is formulated as a trade off between the sum of the residual squares and the ℓ_1 norm of the selected variables. The ℓ_1 norm, i.e., the sum of the absolute values of the selected variables, should be seen as an alternative for the ℓ_0 norm, measuring the size of the selected set. Finding the minimum sum of squared residuals, regularized by the number of selected variables, is a combinatorial problem, and therefore intractable from the computational point of view. The ℓ_1 alternative leads to a quadratic programming problem whose solution is still a proper variable selection, as it contains many exact zeros. The nonzeros, however, are not found by least squares projection, but rather by shrunk versions of the least squares estimators. The intuition behind this is that dubious parameters can be included into the model, but with a value close to zero. If such a parameter happens to be a false positive, its inclusion into the model has a limited impact on any inference in that model. With a much faster algorithm than its ℓ_0 counterpart, the ℓ_1 regularized variable selection, equiped with an appropriate choice of the regularization parameter, is able to find a model with a similar degree of sparsity [3].

Existing variable selection consistency results do not consider the case where the regularization parameter has to be optimized in a data dependent way, using an information criterion. While for fixed or minimax values of the parameter, ℓ_1 regularization provides a valid alternative for ℓ_0, the equivalence holds no longer through the optimization process. This is explained by the ℓ_1 tolerance towards false positives: since the ℓ_1 procedure reduces the impact of a false positive, the optimal balance between the sum of the residual squares and the regularization shifts towards larger models.

In searching for the optimal regularization, ℓ_1 can still be used to actually come up with a selection, but for the evaluation of the quality of the selection, it makes a difference whether the estimation within the selection keeps the shrinkage of the ℓ_1 regularization. If the shrinkage estimator is replaced by a least squares projection, then the optimal balance should shift back towards smaller models. It is obvious that the estimation of the ℓ_0 balance requires

a different expression of the information criterion. The compensation for the difference between ℓ_0 and ℓ_1 regularization has been described as a "mirror" effect [5], further explained in Section 2. It has been explored in the context of unstructured selection in a linear model. In this paper, we extend the scope to structured selection, presented in Section 3. The actual contribution of the paper then follows in Section 4.

2 Mirror effect in unstructured selection in linear models

Consider the sparse linear model

$$\mathbf{Y} = \mathbf{K}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where the design matrix $\mathbf{K}$ has size $n \times m$ with n smaller than m, and the number of nonzeros in $\boldsymbol{\beta}$ is unknown but smaller than n. Also, let $A_s \subset \{1, 2, \ldots, m\}$ be a selection with s nonzeros, obtained by a procedure, $\mathcal{S}(\mathbf{Y}, s)$, that selects among all possible subsets of size s. As an example, $\mathcal{S}(\mathbf{Y}, s)$ could be an implementation of the lasso, finetuned to have s nonzeros as result. Furthermore, let $\mathbf{K}_{A_s}$ denote the $n \times s$ submatrix consisting of the s columns in $\mathbf{K}$ corresponding to the selection. We investigate the quality of the least squares projection $\widehat{\boldsymbol{\beta}}_{A_s} = (\mathbf{K}_{A_s}^T \mathbf{K}_{A_s})^{-1} \mathbf{K}_{A_s}^T \mathbf{Y}$, assuming that $\mathbf{K}_{A_s}$ is non-singular. As a measure for quality, we adopt the prediction error, but a similar discussion would hold for any distance between selected and true model. The prediction error is defined as $\mathrm{PE}(\widehat{\boldsymbol{\beta}}_{A_s})$, where

$$\mathrm{PE}(\widehat{\boldsymbol{\beta}}_A) = \frac{1}{n} E\left(\|\mathbf{K}\boldsymbol{\beta} - \mathbf{K}_A \widehat{\boldsymbol{\beta}}_A\|_2^2 \right). \tag{1}$$

Let A_s^o be the selection provided by an oracle observing $\mathbf{K}\boldsymbol{\beta}$ without noise, using the same procedure, i.e., $A_s^o = \mathcal{S}(\mathbf{K}\boldsymbol{\beta}, s)$. Then the least squares projection, $\widehat{\boldsymbol{\beta}}_{A_s^o} = (\mathbf{K}_{A_s^o}^T \mathbf{K}_{A_s^o})^{-1} \mathbf{K}_{A_s^o}^T \mathbf{Y}$, depends on the observations through $\mathbf{Y}$, but not through A_s^o. The prediction error $\mathrm{PE}(\widehat{\boldsymbol{\beta}}_{A_s^o})$ is estimated unbiasedly by $\Delta_p(A_s^o)$, where $\Delta_p(A)$ is a non studentized version of Mallows's Cp criterion,

$$\Delta_p(\widehat{\boldsymbol{\beta}}_A) = \frac{1}{n} \|\mathbf{Y} - \mathbf{K}_A \widehat{\boldsymbol{\beta}}_A\|_2^2 + \frac{2|A|}{n} \sigma^2 - \sigma^2. \tag{2}$$

The selection $A_s = \mathcal{S}(\mathbf{Y}, s)$, however, depends on $\mathbf{Y}$. The expectation of (2) will not be equal to $\mathrm{PE}(\widehat{\boldsymbol{\beta}}_{A_s})$. As the second and third term of (2) are constants, this is explained by the behavior of $\|\mathbf{Y} - \mathbf{K}_{A_s} \widehat{\boldsymbol{\beta}}_{A_s}\|_2^2$. In the case where the procedure consists of minimizing (2) on all selections of size s, i.e., $\mathcal{S}(\mathbf{Y}, s) = \arg\min_{|A|=s} \Delta_p(A)$, the deviation of the information criterion from

the error curve can be described as a reflection with respect to the oracular mirror $\mathrm{PE}(\widehat{\boldsymbol{\beta}}_{A_s^o}) = E\Delta_p(\widehat{\boldsymbol{\beta}}_{A_s^o})$ [5], meaning that

$$\mathrm{PE}(\widehat{\boldsymbol{\beta}}_{A_s}) - \mathrm{PE}(\widehat{\boldsymbol{\beta}}_{A_s^o}) \approx \mathrm{PE}(\widehat{\boldsymbol{\beta}}_{A_s^o}) - E\Delta_p(\widehat{\boldsymbol{\beta}}_{A_s}) \tag{3}$$

An intuitive explanation follows by assuming that s is large enough to catch all really important variables into both A_s and A_s^o. Once the important variables are in the model, the remainder of the s variables are chosen to further minimize the distance between $\mathbf{K}_{A_s}\widehat{\boldsymbol{\beta}}_{A_s}$ and $\mathbf{Y}$. Among the remaining candidates, these variables perform best in fitting the signal $\mathbf{K}\boldsymbol{\beta}$ with the errors, and thus perform worst in staying close to signal without the errors. The contrast between the better-than-average appearance $E\Delta_p(\widehat{\boldsymbol{\beta}}_{A_s})$ and worse-than-average true prediction error follows from the fact that the optimisation over random variables $\Delta_p(\widehat{\boldsymbol{\beta}}_A)$ affects the statistics of the selected values. The oracle curve $\mathrm{PE}(\widehat{\boldsymbol{\beta}}_{A_s^o})$ acts as mirror, because the selection A_s^o does not depend on $\mathbf{Y}$, thus leaving the statistics of the selected values unchanged.

3 Structured selection with grouped variables

The lasso, in addition to providing us with a selection A_s considering an appropriate regularisation parameter, can be extented or used to take into account structured models such as grouped variables [9], graphical models [10, 8] or even hierarchical information [11]. When the variables are under the hypothesis to have a natural group structure, the coefficients within a group should all be nonzero (or zero).

In its Lagrangian form, the lasso problem of a linear model is expressed as

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{Y} - \mathbf{K}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \tag{4}$$

with λ being a regularisation parameter which can be adjusted to obtain the desired degree of sparsity. When $\mathbf{K}$ is orthogonal, the solution of (4) is simply a soft-thresholded version of the least-squares estimate whose threshold is λ. For the remainder of this paper, we consider the signal-plus-noise model $\mathbf{Y} = \boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $m = n$ and $\mathbf{K} = \mathbf{I}_n$. Then the best s term unstructured selection, mesured by the Cp-value, consists of the s largest elements from $\mathbf{Y}$.

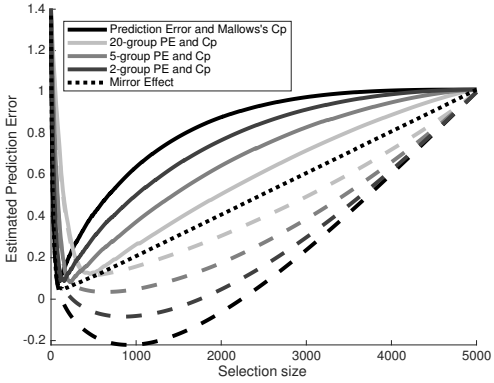
For group selection, the penalty in (4) can be modified to become the sum of the ℓ_2 norms of each group. This is known as group lasso and it aims to optimise the following expression, for the signal-plus-noise model with n_g groups,

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{Y} - \boldsymbol{\beta}\|_2^2 + \lambda\sum_{j=1}^{n_g}\|\boldsymbol{\beta}_j\|_2 \tag{5}$$

where $\boldsymbol{\beta}_j \in \mathbb{R}^{w_j}$ forms a group of w_j coefficients from $\boldsymbol{\beta}$ and $\sum_{j=1}^{n_g} w_j = n$. The solution of (5) is again a soft-thresholded version of $\mathbf{Y}$, although the threshold has the form $\lambda|Y_i|/\|\mathbf{Y}_j\|_2$ for observation i within group j. Hence without shrinkage, the best s_g group selection contains the values of $\mathbf{Y}$ from the s_g groups of observations whose ℓ_2 norms are the largest.

4 Mirror effect in group selection and discussion

In our simulation, 250 groups containing 20 coefficients $\boldsymbol{\beta}_j$ are generated so that $\boldsymbol{\beta} = (\boldsymbol{\beta}_j)_{j=1,\ldots,250}$ is a n-dimensional vector with $n = 5000$. Within group j, the $\boldsymbol{\beta}_j$ have the same probability p_j of being set to 0; for each j, a different value p_j is randomly drawn from the set $\mathbf{P} = (0.95, 0.80, 0.50, 0.05, 0.00)$ with respective probability $\mathbf{Q} = (0.02, 0.02, 0.01, 0.20, 0.75)$. The expected proportion of nonzeros is then $\langle \mathbf{P}, \mathbf{Q} \rangle = 1/20$ for the whole data $\boldsymbol{\beta}$. The nonzeros $\boldsymbol{\beta}$ are then distributed according to the zero inflated Laplace model $f_{\beta|\beta\neq 0}(\beta) = (a/2)\exp(-a|\beta|)$ where $a = 1/5$. The observations are $\mathbf{Y} = \boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is a n-vector of independent, standard normal errors. Estimates $\widehat{\boldsymbol{\beta}}$ are calculated considering four configurations: groups of size 20 (initial setting), 5 and 2 (subgroups built from the original groups) and 1 (unstructured selection).



The PE and Cp curves, solid and dashed lines respectively, are represented as functions of the selection size, for different sizes of group. The dotted line depicts the mirror curve estimated for unstrutured variables [5].

Figure 1: Mirror effect and group size impact.

Figure 1 plots the prediction error and Mallows's Cp as a function of the selection size for unstructured and 20-5-2-grouped variable selection. In each case, we observe that the PE and Cp curves are reflexion of each other with respect to a mirror curve. It is interesting to note that, for the signal-plus-noise model, the unstructured and group mirror curves coincide once the PE and

Cp curves are drifting apart. Also, it seems the bigger the group size gets, the closer the corresponding PE and Cp curves are. Hence when the group size grows, the mirror effect becomes smaller.

References

[1] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *JRSSB*, 57:289–300, 1995.

[2] G. Claeskens and N. Hjort. The focused information criterion. *JASA*, 98: 900–916, 2003.

[3] D. L. Donoho. For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829, 2006.

[4] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

[5] M. Jansen. Information criteria for variable selection under sparsity. *Biometrika*, 101(1):37–55, 2014.

[6] J. D. Lee, D. L. Sun, and J. E. Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

[7] H. Leeb and B.M. Pötscher. Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics*, 34(5):2554–2591, 2006.

[8] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

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

[10] M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007b.

[11] P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37:3468–3497, 2009.

Theoretical and simulation results on heavy-tailed fractional Pearson diffusions

Ivan Papić[*1], **Nikolai N. Leonenko**[2], **Alla Sikorskii**[3], **and Nenad Šuvak**[1]

[1]*Department of Mathematics, J.J. Strossmayer University of Osijek, Croatia*
[2]*School of Mathematics, Cardiff University, UK*
[3]*Department of Statistics and Probability, Michigan State University, USA*

We define heavy-tailed fractional reciprocal gamma and Fisher-Snedecor diffusions by a non-Markovian time change in the corresponding Pearson diffusions. We illustrate known theoretical results regarding these fractional diffusions via simulations.

Keywords: Fractional diffusion, Pearson diffusion, Stable subordinator, Transition density, Simulations

1 Introduction

Every continuous distribution with density satisfying the so called Pearson equation

$$\frac{\mathfrak{p}'(x)}{\mathfrak{p}(x)} = \frac{(a_1 - 2b_2)x + (a_0 - b_1)}{b_2x^2 + b_1x + b_0} \tag{1}$$

is called a Pearson distribution (see [8]). The family of Pearson distributions consists of six parametric subfamilies: normal, gamma, beta, Fisher-Snedecor, reciprocal gamma and Student distributions.

Strong solution of SDE

$$dX_t = \mu(X_t)dt + \sigma(X_t)dW_t,\ t \geq 0, \tag{2}$$

where

$$\mu(x) = a_0 + a_1x,\ \sigma(x) = \sqrt{2b(x)} = \sqrt{2(b_2x^2 + b_1x + b_0)}$$

is called the Pearson diffusion. They are called after Pearson since their stationary distributions belong to the Pearson family. Usually, it is convenient to re-parametrize drift and squared diffusion:

$$\mu(x) = -\theta(x - \mu),\ \sigma^2(x) = 2\theta k(B_2x^2 + B_1x + B_0),$$

*Corresponding author: ipapic@mathos.hr

where $\mu \in \mathbb{R}$ is the stationary mean depending on coefficients of the Pearson equation (1), $\theta > 0$ is the scaling of time determining the speed of the mean reversion, and k is a positive constant. Note that we need $\sigma^2(x) > 0$ on the diffusion state space (l, L).

Pearson diffusions could be categorized into six subfamilies, according to the degree of the polynomial $b(x)$ and, in the quadratic case $b(x) = b_2x^2 + b_1x + b_0$, according to the sign of its leading coefficient b_2 and the sign of its discriminant Δ:

- constant $b(x)$ - Ornstein-Uhlenbeck (OU) process with normal stationary distribution,
- linear $b(x)$ - Cox-Ingersol-Ross (CIR) process with gamma stationary distribution,
- quadratic $b(x)$ with $b_2 < 0$ - Jacobi diffusion with beta stationary distribution,
- quadratic $b(x)$ with $b_2 > 0$ and $\Delta > 0$ - Fisher-Snedecor (FS) diffusion with the Fisher-Snedecor stationary distribution,
- quadratic $b(x)$ with $b_2 > 0$ and $\Delta = 0$ - reciprocal gamma (RG) diffusion with reciprocal gamma stationary distribution,
- quadratic $b(x)$ with $b_2 > 0$ and $\Delta < 0$ - Student diffusion with the Student stationary distribution.

2 Fractional diffusions

The subject of our interest are fractional derivatives of order $0 < \alpha < 1$. We define Caputo fractional derivative of order $0 < \alpha < 1$ as

$$\frac{d^\alpha f(x)}{dx^\alpha} = \frac{1}{\Gamma(1-\alpha)} \int_0^\infty \frac{d}{dx} f(x-y) y^{-\alpha}\, dy,$$

or equivalently for absolutely continuous functions as

$$\frac{d^\alpha f(x)}{dx^\alpha} = \frac{1}{\Gamma(1-\alpha)} \int_0^x (x-y)^{-\alpha} f'(y)\, dy.$$

Interesting and detailed read regarding fractional derivatives one can find in [7, Chapter 2].

By $(X(t),\, t \geq 0)$ denote the Pearson diffusion solving (2). Introduce $(D_t,\, t \geq 0)$, the standard stable subordinator with index $0 < \alpha < 1$, which is independent of the process $(X(t),\, t \geq 0)$. D_t is a homogeneous Lèvy process with the Laplace transform

$$\mathbb{E}[e^{-sD_t}] = \exp\{-ts^\alpha\}.$$

Its inverse process

$$E_t = \inf\{x > 0 : D_x > t\}$$

is non-Markovian, non-decreasing, and for every t random variable E_t has a density, which will be denoted by $f_t(\cdot)$. The Laplace transform of this density is (see e.g., [9])

$$\mathbb{E}[e^{-sE_t}] = \int_0^\infty e^{-sx} f_t(x)dx = \mathcal{E}_\alpha(-st^\alpha), \tag{3}$$

where

$$\mathcal{E}_\alpha(z) := \sum_{j=0}^{\infty} \frac{(z)^j}{\Gamma(1+\alpha j)}$$

is the Mittag-Leffler function (see, for example [10]).
Notice that for $\alpha = 1$

$$\mathcal{E}_\alpha(z) = e^z,$$

i.e, Mittag-Leffler reduces to the exponential function.

Now, define the fractional Pearson diffusion $(X_\alpha(t),\, t \geq 0)$ as a composition of the Pearson diffusion and inverse of the stable subordinator, i.e.

$$X_\alpha(t) = X(E_t),\ t \geq 0. \tag{4}$$

We emphasize that $(X_\alpha(t),\, t \geq 0)$ is a non-Markovian process and define its transition density $p_\alpha(x,t;y)$ as

$$P(X_\alpha(t) \in B | X_\alpha(0) = y) = \int_B p_\alpha(x,t;y)dx \tag{5}$$

for any Borel subset B of (l, L).
Using results from [1] one can show that if the non-fractional Pearson diffusion satisfy SDE (2) with inital condition $X(0) = 0$, then the corresponding fractional Pearson diffusion defined with (4) satisfy SDE

$$dX_\alpha(t) = \mu(X_\alpha(t))dE_t + \sigma(X_\alpha(t))dB_{E_t} \tag{6}$$

with initial condition $X_\alpha(0) = 0$.
Integral form of this SDE is

$$X_\alpha(t) = X(E_t) = \int_0^{E_t} (a_0 + a_1 X(s))\, ds + \int_0^{E_t} \sqrt{2(b_0 + b_1 X(s) + b_2 (X(s))^2)}\, dB(s).$$

For details we refer to [1] and [4].
Non-heavy-tailed fractional Pearson diffusions (fractional Ornstein-Uhlenbeck (OU) , Cox-Ingersol-Ross (CIR) and Jacobi diffusion) are studied in detail in [3], while heavy-tailed fractional Pearson diffusions (fractional Fisher-Snedecor and reciprocal gamma diffusion) are studied in the recent paper [4]. Fractional Student diffusion have not yet been studied in detail since the nature behind the process (infinitesimal generator and spectrum) is much more complicated then in the other five cases. However, regarding non-fractional Student diffusion one can find some results in [5].
In this paper, we present simulation results regarding fractional Fisher-Snedecor and fractional reciprocal gamma diffusions, which illustrates theoretical results obtained in [4]. Therefore, we begin by stating the necessary theoretical results.

3 Fractional reciprocal gamma diffusion

The reciprocal gamma diffusion satisfies the SDE

$$dX_t = -\theta \left(X_t - \frac{\gamma}{\beta - 1} \right) dt + \sqrt{\frac{2\theta}{\beta - 1} X_t^2}\, dW_t, \quad t \geq 0,$$

with $\theta > 0$ and has invariant density

$$\mathfrak{rg}(x) = \frac{\gamma^\beta}{\Gamma(\beta)}\, x^{-\beta-1} e^{-\frac{\gamma}{x}} \mathrm{I}_{\langle 0,\infty \rangle}(x) \tag{7}$$

with parameters $\gamma > 0$ and $\beta > 1$, where the latter requirement ensures the existence of the stationary mean $\gamma/(\beta - 1)$.

Theorem 1. *The transition density of the fractional RG diffusion is given by*

$$\begin{aligned} p_\alpha(x, t; x_0) = & \sum_{n=0}^{\lfloor \frac{\beta}{2} \rfloor} \mathfrak{rg}(x)\, B_n(x)\, B_n(x_0)\, \mathcal{E}_\alpha(-\lambda_n t^\alpha) \\ & + \frac{\mathfrak{rg}(x)}{4\pi} \int_{\frac{\theta \beta^2}{4(\beta-1)}}^{\infty} \mathcal{E}_\alpha(-\lambda t^\alpha)\, b(\lambda)\, \psi(x, -\lambda)\, \psi(x_0, -\lambda)\, d\lambda, \end{aligned} \tag{8}$$

where B_n are normalized Bessel polynomials, λ_n are their eigenvalues, $b(\lambda)$ is a constant depending on λ and ψ is the solution of the corresponding Sturm-Liouville equation.

For proof and details see [4].

4 Fractional Fisher-Snedecor diffusion

The Fisher-Snedecor diffusion satisfies the SDE

$$dX_t = -\theta\left(X_t - \frac{\beta}{\beta-2}\right)dt + \sqrt{\frac{4\theta}{\gamma(\beta-2)}X_t(\gamma X_t+\beta)}\,dW_t, \quad t \geq 0$$

with $\theta > 0$ and has invariant density

$$\mathfrak{fs}(x) = \frac{\beta^{\frac{\beta}{2}}}{B\left(\frac{\gamma}{2},\frac{\beta}{2}\right)}\,\frac{(\gamma x)^{\frac{\gamma}{2}-1}}{(\gamma x+\beta)^{\frac{\gamma}{2}+\frac{\beta}{2}}}\,\gamma\,\mathrm{I}_{\langle 0,\infty\rangle}(x) \tag{9}$$

with parameters $\gamma > 0$ and $\beta > 2$, where the latter requirement ensures the existence of the stationary mean $\beta/(\beta-2)$.

Theorem 2. *The transition density of fractional FS diffusion is given by*

$$\begin{aligned} p_\alpha(x,t;x_0) = &\sum_{n=0}^{\lfloor\frac{\beta}{4}\rfloor} \mathfrak{fs}(x)\,F_n(x_0)\,F_n(x)\,\mathcal{E}_\alpha(-\lambda_n t^\alpha) \\ &+ \frac{\mathfrak{fs}(x)}{\pi}\int\limits_{\frac{\theta\beta^2}{8(\beta-2)}}^{\infty} \mathcal{E}_\alpha(-\lambda t^\alpha)\,a(\lambda)\,f_1(x_0,-\lambda)\,f_1(x,-\lambda)\,d\lambda, \end{aligned} \tag{10}$$

where F_n are normalized Fisher-Snedecor polynomials, λ_n are their eigenvalues, $a(\lambda)$ is a constant depending on λ and f_1 is the solution of the corresponding Sturm-Liouville equation.

For proof and details see [4].

5 Stationary distributions of the fractional reciprocal gamma and Fisher-Snedecor diffusions

By $p_\alpha(x,t)$ denote the density of $X_\alpha(t)$, by $p(x,t)$ the density of $X(t)$ and let f be the density of initial state $X_\alpha(0)$. Now, by the definition of transition

density it follows

$$p_\alpha(x,t) = \int_0^\infty p_\alpha(x,t;y)f(y)dy.$$

If we assume that the initial distribution is concentrated in one point, i.e. if $f(y) = \delta(x_0)$ we obtain

$$p_\alpha(x,t) = p_\alpha(x,t;x_0)$$

and since for fractional FS and RG diffusion, transition densities $p_\alpha(x,t;x_0)$ are given via (8) and (10), one can show that

$$p_\alpha(x,t) \to m(x) \; as \; t \to \infty, \tag{11}$$

where m is FS stationary distribution in fractional FS diffusion case, and RG stationary distribution in fractional RG case.
In fact, even without the assumption on the concentrated initial state, one can prove the statement, for details we refer to [4].
Also, obsverve that

$$p_\alpha(x,t) \to p(x,t) \; as \; \alpha \to 1. \tag{12}$$

6 Correlation structure of fractional Pearson diffusions

Stationary Pearson diffusion $X(t)$ such that the stationary distribution has finite second moment has the correlation function given by

$$\mathrm{Corr}\,[X(t),\, X(s)] = \exp(-\theta|t-s|), \tag{13}$$

where θ is the autocorrelation parameter. Since the autocorrelation function (13) falls off exponentially, Pearson diffusions exhibit short-range dependence.

We say that fractional Pearson diffusion $X_\alpha(t)$ defined by (4) is in the steady state if it starts from its invariant distribution with the density m. Then the autocorrelation function of $X_\alpha(t) = X(E_t)$ is given by

$$\mathrm{Corr}\,[X_\alpha(t),\, X_\alpha(s)] = \mathcal{E}_\alpha(-\theta t^\alpha) + \frac{\theta\alpha t^\alpha}{\Gamma(1+\alpha)} \int_0^{s/t} \frac{\mathcal{E}_\alpha(-\theta t^\alpha(1-z)^\alpha)}{z^{1-\alpha}} dz \tag{14}$$

for $t \geq s > 0$. The tehnique to prove this fact can be found in [2].
Observe that (14) implies the long-range dependence of the fractional diffusion $X_\alpha(t)$, since the autocorrelation function (14) falls off like power law with exponent $\alpha \in (0,1)$.

7 Simulation results

Simulation results are based on the algorithm introduced in [6]. Basically, idea is to seperately simulate trajectory of the inverse of the stable subordinator and trajectory of the non-fractional diffusion. Afterwards, by linear interpolation one gets trajectory of the fractional diffusion. This algorithm perfectly fits our setting, since we define fractional Pearson diffusion as a composition of the non-fractional Pearson diffusion and the inverse of the stable subordinator (which are assume to be independent).
Trajectories of such simulated fractional RG and FS diffusion are given in Figure 1, where the difference between non-fractional and fractional diffusions can be clearly seen. Unlike non-fractional diffusions, fractional diffusions have long resting periods of time due to change of time via inverse of the stable subordinator E_t.

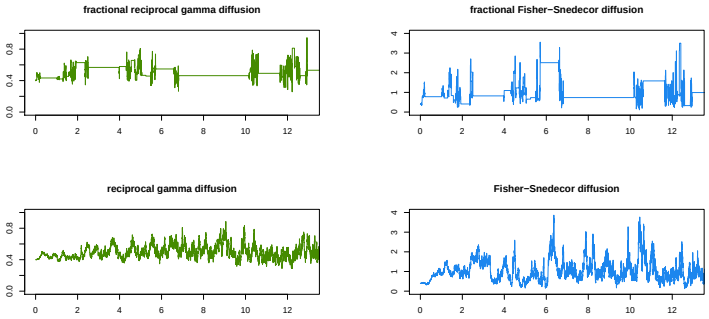


Figure 1: Sample paths of the fractional/non-fractional RG and FS diffusions with parameters $\gamma = 10$, $\beta = 20$, $\theta = 0.01$ and $\alpha = 0.7$, based on 10000 points with inital state $X_0 = 0.4$.

Next, we illustrate that density of fractional diffusion approach the stationary density as explained in Section 5. We simulated 1000 trajectories of the fractional RG diffusion and estimated densities at times $t = 0.02$, $t = 0.2$ and $t = 2$, see figure 2. Comparing densities $p_\alpha(x,t)$ and $p(x,t)$, we clearly observe slower approaching to the stationary density in fractional case. Autocorrelation function (14) of the fractional diffusion, in comparison with the autocorrelation function (13) of the non-fractional diffusion which decays exponentially fast, decays much slower, i.e. in polynomial rate. This is illustrated in Figure 3.

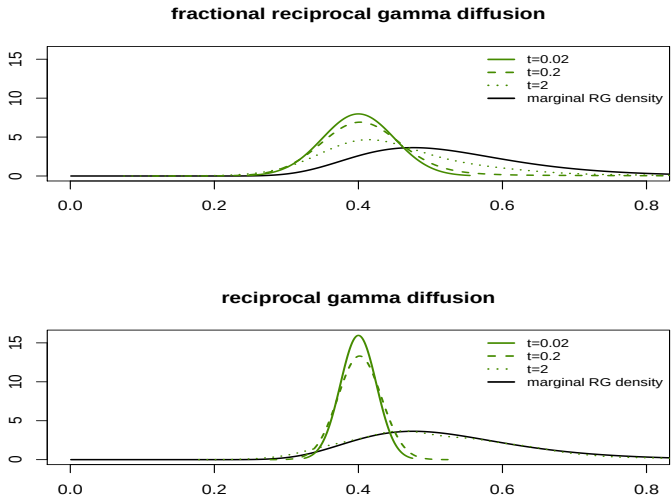


Figure 2: Estimated densities $p_\alpha(x,t)$ and $p(x,t)$ for reciprocal gamma diffusion with parameters $\gamma = 10$, $\beta = 20$, $\theta = 0.01$ and $\alpha = 0.7$, based on 1000 trajectories with inital state $X_0 = 0.4$.

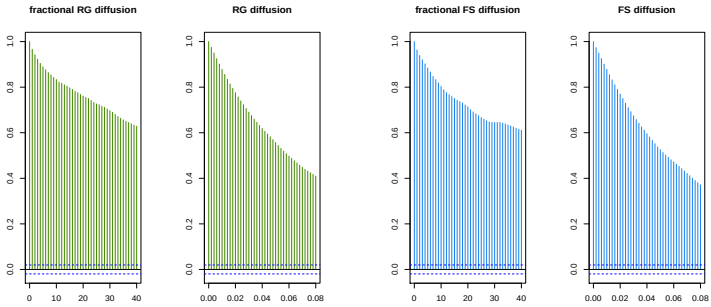


Figure 3: Estimated autocorrelation function of fractional/non fractional RG and FS diffusions with parameters $\gamma = 10$, $\beta = 20$, $\theta = 0.01$ and $\alpha = 0.7$, based on 10000 points with inital state $X_0 = 0.4$.

References

[1] K. Kobayashi. Stochastic calculus for a time-changed semimartingale and the associated stochastic differential equations. *Journal of Theoretical Probability*, 24(3):789–820, 2011.

[2] N. Leonenko, M. Meerschaert, and A. Sikorskii. Correlation structure of fractional Pearson diffusions. *Computers and Mathematics with Applications*, 66(5):737 – 745, 2013.

[3] N. Leonenko, M. Meerschaert, and A. Sikorskii. Fractional Pearson diffusions. *Journal of Mathematical Analysis and Applications*, 403(2):532 – 546, 2013.

[4] N. Leonenko, I. Papić, A. Sikorskii, and N. Šuvak. Heavy-tailed fractional Pearson diffusions. *Stochastic Processes and their Applications*, 2017.

[5] N. Leonenko and N. Šuvak. Statistical inference for Student diffusion process. *Stochastic Analysis and Applictions*, 28(6):972–1002, 2010.

[6] M. Magdziarz, A. Weron, and K. Weron. Fractional Fokker-Planck dynamics: Stochastic representation and computer simulation. *Physical Review E*, 75(1):016708, 2007.

[7] M. Meerschaert and A. Sikorskii. *Stochastic Models for Fractional Calculus*. De Gruyter, 2011.

[8] K. Pearson. *Tables for Statisticians and Biometricians, Part I*. Biometrics Laboratory, University College London, 1914.

[9] A. Piryatinska, A. Saichev, and W. Woyczynski. Models of anomalous diffusion: the subdiffusive case. *Physica A*, 349:375–420, 2005.

[10] T. Simon. Comparing Fréchet and positive stable laws. *Electronic Journal of Probability*, 19(16):1–25, 2014.

Copula based BINAR models with applications

Andrius Buteikis[*1]

[1]*Faculty of Mathematics and Informatics, Vilnius University, Naugarduko st. 24, LT-03225, Vilnius, Lithuania*

In this paper we study the problem of modelling the integer-valued vector observations. We consider the BINAR(1) models defined via copula-joint innovations. We review different parameter estimation methods and analyse estimation methods of the copula dependence parameter. We also examine the case where seasonality is present in integer-valued data and suggest a method of deseasonalizing them. Finally, an empirical application is carried out.

Keywords: Count data, BINAR, Poisson, Negative binomial distribution, Copula.

1 Introduction

Different financial institutions that issue loans do so following company-specific (and/or country-defined) rules which act as a safeguard so that loans are not issued to people who are known to be insolvent. The adequacy of a firms rules for issuing loans can be analysed by modelling the dependence between the number of loans which have defaulted and number of loans that have not defaulted via copulas.

The advantage of such approach is that copulas allow to model the marginal distributions (possibly from different distribution families) and their dependence structure (which is described via a copula) separately. Because of this feature, copulas were applied to many different fields (for some examples of copula applications see [2], [4], [5] and [6]). While these studies were carried out for continuous data, there is less developed literature on discrete models created with copulas: [7] discussed the differences and challenges of using copulas for discrete data compared to continuous data. By using bivariate integer-valued autoregressive models (BINAR) it is possible to account for both the discreteness and autocorrelation of the data. Furthermore, copulas can be used to model the dependence of innovations in the BINAR(1) models: [9] used the Frank copula and normal copula to model the dependence of the innovations of the BINAR(1) model.

*Corresponding author: andrius.buteikis@mif.vu.lt

In this short paper we analyse different BINAR(1) model with copula-joint innovations parameter estimation methods. We also discuss some issues concerning the seasonality in integer-valued data and suggest a method of deseasonalizing them. Finally, in order to analyse the presence of autocorrelation and copula dependence in loan data, an empirical application is carried out on weekly loan data. Estimation method comparisons and additional numerical results can be found in [3].

The paper is organized as follows. Section 2 presents the BINAR(1) process, Section 3 presents the definition of copulas. Section 4 compares different estimation methods for the BINAR(1) model. Seasonal adjustment of integer-valued data is presented in Section 5. In Section 6 an empirical application is carried out using different combinations of copula functions and marginal distribution functions. Conclusions are presented in Section 7.

2 The bivariate INAR(1) process

The BINAR(1) process was introduced in [11]. In this section we will provide the definition of the BINAR(1) model.

Definition 1. Let $\mathbf{R}_t = [R_{1,t}, R_{2,t}]'$, $t \in \mathbb{Z}$ be a sequence of independent identically distributed (i.i.d.) non-negative integer-valued bivariate random variables. A bivariate integer-valued autoregressive process of order 1 (BINAR(1)), $\mathbf{X}_t = [X_{1,t}, X_{2,t}]'$, $t \in \mathbb{Z}$, is defined as:

$$\mathbf{X}_t = \mathbf{A} \circ \mathbf{X}_{t-1} + \mathbf{R}_t = \begin{bmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{bmatrix} \circ \begin{bmatrix} X_{1,t-1} \\ X_{2,t-1} \end{bmatrix} + \begin{bmatrix} R_{1,t} \\ R_{2,t} \end{bmatrix}, \quad t \in \mathbb{Z}, \tag{1}$$

where $\alpha_j \in [0, 1)$, $j = 1, 2$, and the symbol '$\circ$' is the thinning operator which also acts as the matrix multiplication. We have that $\alpha_j \circ X_{j,t-1} := \sum_{i=1}^{X_{j,t-1}} Y_{j,t,i}$ and $Y_{j,t,1}, Y_{j,t,2}, \ldots$ is a sequence of i.i.d. Bernoulli random variables with $\mathbb{P}(Y_{j,t,i} = 1) = \alpha_j = 1 - \mathbb{P}(Y_{j,t,i} = 0)$, $\alpha_j \in [0, 1)$, such that these sequences are mutually independent and independent of the sequence $\mathbf{R}_t$, $t \in \mathbb{Z}$. For each t, $\mathbf{R}_t$ is independent of $\mathbf{X}_s$, $s < t$.

A number of thinning operator properties are provided in [12] and [13]. Properties of the BINAR(1) model can be easily derived and a number of these are provided in [12]. We will expand on the work by [9] and [11] by analysing additional copulas for the BINAR(1) model innovation distribution as well as estimation methods for the distribution parameters.

3 Copulas

Copulas are used for modelling the dependence between several random variables. The main advantage of using copulas is that they allow to model the marginal distributions separately from their joint distribution. More information about Copula theory, properties and applications can be found in [10] and [8].

Since innovations of a BINAR(1) model are non-negative integer-valued random variables, one needs to consider copulas linking discrete distributions. According to Sklar's theorem [14], if F_1 and F_2 are discrete marginals then a unique copula representation exists only for values in the range of $\mathrm{Ran}(F_1) \times \mathrm{Ran}(F_2)$. However, the lack of uniqueness does not pose a problem in empirical applications because it implies that there may exist more than one copula which describes the distribution of the empirical data. Bivariate copulas which will be used when constructing and evaluating the BINAR(1) model in this paper are:

- The Farlie-Gumbel-Morgenstern (FGM) copula with $\theta \in [-1, 1]$:

$$C(u_1, u_2; \theta) = u_1 u_2 (1 + \theta(1 - u_1)(1 - u_2)),$$

- The Frank copula with $\theta \in (-\infty, \infty) \setminus \{0\}$:

$$C(u_1, u_2; \theta) = -\frac{1}{\theta} \log \left(1 + \frac{(\exp(-\theta u_1) - 1)(\exp(-\theta u_2) - 1)}{\exp(-\theta) - 1} \right),$$

where $u_1 := F_1(x_1)$, $u_2 := F_2(x_2)$. Here θ is the dependence parameter and F_1, F_2 - marginal cdfs. See [10] for properties of these copulas.

4 Parameter estimation of the copula-based BINAR(1) model

In this section we examine different BINAR(1) model parameter estimation methods. Let $\mathbf{X}_t = (X_{1,t}, X_{2,t})'$ be a non-negative integer-valued time series given in Def. 1, where the joint distribution of $(R_{1,t}, R_{2,t})'$, with marginals F_1, F_2, is linked by a copula $C(\cdot, \cdot)$: $\mathbb{P}(R_{1,t} \leq x_1, R_{2,t} \leq x_2) = C(F_1(x_1), F_2(x_2))$ and let $C(F_1(x_1), F_2(x_2)) = C(F_1(x_1), F_2(x_2); \theta)$, where θ is a dependence parameter.

4.1 Conditional least squares (CLS) estimation

The Conditional Least Squares (CLS) estimator minimizes the squared distance between $\mathbf{X}_t$ and its conditional expectation. Similarly to the method in [13] for the INAR(1) model, we construct the CLS estimator in the case of the BINAR(1) model. The CLS estimators of $\alpha_j, \lambda_j,\ j = 1, 2$ are found by minimizing the sum

$$Q_j(\alpha_j, \lambda_j) \quad := \quad \sum_{t=2}^{N} (X_{j,t} - \alpha_j X_{j,t-1} - \lambda_j)^2 \ \longrightarrow \ \min_{\alpha_j, \lambda_j}, \quad j = 1, 2. \tag{2}$$

The asymptotic properties of the CLS estimators for the INAR(1) model case are provided in [13]. Assume now that the Poisson innovations $R_{1,t}$ and $R_{2,t}$ with parameters λ_1 and λ_2, respectively, are joint by a copula with dependence parameter θ. In order to estimate θ, [3] minimized the sum of squared differences

$$S^{(M_1,M_2)} \quad = \quad \sum_{t=2}^{N} \left(\widetilde{X}_{1,t}^{\text{CLS}} \widetilde{X}_{2,t}^{\text{CLS}} - \gamma^{(M_1,M_2)}(\widehat{\lambda}_1^{\text{CLS}}, \widehat{\lambda}_2^{\text{CLS}}; \theta) \right)^2, \tag{3}$$

where

$$\begin{aligned} \widetilde{X}_{j,t}^{\text{CLS}} \quad &:= \quad X_{j,t} - \widehat{\alpha}_j^{\text{CLS}} X_{j,t-1} - \widehat{\lambda}_j^{\text{CLS}}, \quad j = 1, 2, \\ \gamma^{(M_1,M_2)}(\lambda_1, \lambda_2; \theta) \quad &:= \quad \sum_{k=1}^{M_1} \sum_{l=1}^{M_2} kl\, c(F_1(k; \lambda_1), F_2(l; \lambda_2); \theta) - \lambda_1 \lambda_2, \end{aligned}$$

where $c(F_1(k; \lambda_1), F_2(s; \lambda_2); \theta)$ is the joint probability mass function and M_1 and M_2 are used to approximate the covariance $\gamma(\lambda_1, \lambda_2; \theta)$ as described in [3].

4.2 Conditional maximum likelihood (CML) estimation

BINAR(1) models can also be estimated via conditional maximum likelihood (CML) (see [11] and [9]). The log conditional likelihood function is:

$$\ell \quad = \quad \sum_{t=2}^{N} \log \mathbb{P}(X_{1,t} = x_{1,t}, X_{2,t} = x_{2,t} | X_{1,t-1} = x_{1,t-1}, X_{2,t-1} = x_{2,t-1})$$

for some initial values $x_{1,1}$ and $x_{2,1}$. In order to estimate the unknown parameters we maximize the log conditional likelihood:

$$\ell(\alpha_1, \alpha_2, \lambda_1, \lambda_2, \theta) \longrightarrow \max_{\alpha_1, \alpha_2, \lambda_1, \lambda_2, \theta}. \tag{4}$$

Numerical maximization is straightforward with the `optim` function from R statistical software.

For other marginal distribution cases where the marginal distribution has parameters other than λ_j, equation (4) would need to be minimized by those additional parameters.

4.3 Two-step estimation based on CLS and CML

Depending on the range of attainable values of the parameters and the sample size, CML maximization might take some time to compute. On the other hand, since CLS estimators of α_j and λ_j are easily derived, [3] proposed to substitute the parameters of the marginal distributions in eq. (4) with CLS estimates from eq. (2). Then we would only need to maximize ℓ with respect to a single dependence parameter θ.

4.4 Estimation method comparison via Monte Carlo simulation

A Monte Carlo simulation was carried out in [3] in order to compare the estimation methods. The estimates of the dependence parameter were similar in terms of MSE and bias for both CML and Two-step estimation method.

5 Seasonality

Assume now that the nonnegative integer-valued time series can be written in the following form $\mathbf{Z}_t = \mathbf{S}_t + \mathbf{X}_t$, where $\mathbf{X}_t$ is defined by equation (1) and $\mathbf{S}_t = (S_{1,t}, S_{2,t})'$ is the (deterministic) integer-valued seasonal component with period d, where $S_{j,t} = S_{j,t+d}$, $\forall t$ and $j = 1, 2$ and $\sum_{k=0}^{d-1} S_{j,t+k} = 0$.

In order to remove the seasonal effect but keep the nonnegative, integer-valued properties of the data, we defined the operator $s(L) = 1 + L + ... + L^{d-1}$, where $L^k Z_t = Z_{t-k}$, $k \geq 0$. By applying this operator, the seasonal component is removed and the sample size decreases by $d - 1$ observations. Alternatively, data can also be aggregated to a lower frequency (e.g. from daily to weekly data) in order to remove the seasonal effect at the cost of reducing the sample size d times. Finally, one can extend the seasonal INAR(1) model proposed in [1] to the BINAR(1) case.

Comparisons of these different seasonal adjustment methods is left for future research.

6 Application on default loan data

In this section we estimate a BINAR(1) model with the joint innovation distribution modelled by a copula cdf for empirical data. The dataset consists of weekly data on loans issued in Spain from October 21st, 2013, to January 1st, 2016 which includes loans that have defaulted and loans that were repaid without missing any payments. We will analyse and model the dependence between defaulted and non-defaulted loans as well as the presence of autocorrelation by considering a BINAR(1) model with different copulas for the innovations. For the marginal distributions of the innovations we considered Poisson as well as negative binomial distributions. We used the Two-step estimation method to estimate parameters. The dependence and variance parameter estimates when both marginals are negative binomial are provided in Table 1. Additional modelling results are provided in [3].

Table 1: Dependence and variance parameter estimates for BINAR(1) model via Two-step estimation method

Copula	$\hat{\theta}$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	AIC
FGM	0.8927	6.5581	45.3683	1466.1542
	(0.1867)	(1.2402)	(7.5522)	
Frank	2.3848	6.5875	45.426	1466.9795
	(0.5337)	(1.2613)	(7.5774)	

Overall, both Frank and FGM copulas provide similar fit in terms of AIC, regardless of the selected marginal distributions. The FGM copula is used to model weak dependence. Given a larger sample size, a Frank copula might be more appropriate because it can capture a stronger dependence than that of an FGM copula. Furthermore, the estimated dependence parameter is positive for the Frank and FGM copula cases, which indicates that there is a positive dependence between defaulted and non-defaulted loans.

7 Conclusions

In this short paper we have analysed different estimation methods for estimating parameters of a BINAR(1) model, including the dependence parameter of its innovations, which are linked via a copula. According to Monte Carlo simulations carried out in [3], BINAR(1) parameter estimates via CML had the smallest MSE and bias, however, estimates of the dependence parameter via CML and Two-step methods were similar. We also suggested a method to seasonally adjust the integer-valued data which exhibits a seasonal variation.

An empirical application on loan data was carried out and BINAR(1) models were estimated using different combinations of copula functions and marginal distribution functions. Additional estimation results are provided in [3]. The FGM copula provided the best model fit with Frank copula being very close in terms of AIC values. A larger sample size could help determine whether FGM or Frank copula is more appropriate to model the dependence between defaulted and non-defaulted loan amounts. Furthermore, the estimated copula dependence parameter indicates that the dependence between defaulted and non defaulted loans is positive.

Finally, one can apply different copula functions in order to analyse whether the loan data exhibits different forms of dependence. Lastly, the model can be extended by analysing the presence of structural changes within the data as well as extending the BINAR(1) model with copula joint innovations to account for the past values of other time series rather than only itself.

References

[1] M. Bourguignon, K.L.P. Vasconcellos, V.A. Reisen and M. Ispany. *A Poisson INAR(1) process with a seasonal structure.* Journal of Statistical Computation and Simulation, 86(2), 373–387, 2016.

[2] D. Brigo and A. Pallavicini and R. Torresetti. *Credit Models and the Crisis: A Journey into CDOs, Copulas, Correlations and Dynamic Models.* Wiley, United Kingdom, 2011.

[3] A. Buteikis and R. Leipus. *Application of copula-based BINAR models in loan modelling.* Preprint.

[4] U. Cherubini and S. Mulinacci and F. Gobbi and S. Romagnoli. *Dynamic Copula Methods in Finance.* Wiley, United Kingdom, 2011.

[5] J. Crook and F. Moreira. *Checking for Asymetric Default Dependence In a Credit Card Portfolio: A Copula Approach.* Journal of Empirical Finance, 18, 728–742, 2011.

[6] J. P. Fenech and H. Vosgha and S. Shafik. *Loan Default Correlation Using an Achimedean Copula Approach: A Case For Recalibration.* Economic Modelling, 47, 340–354, 2015.

[7] C. Genest and J. Neslehova. *A Primer on Copulas for Count Data.* Astin Bulletin, 37(2), 475–515, 2007.

[8] H. Joe. *Dependence Modeling with Copulas.* Chapman & Hall/CRC Monographs on Statistics and Applied probability, 134, 2015.

[9] D. Karlis and X. Pedeli. *Flexible Bivariate INAR(1) Processes Using Copulas.* Communications in Statistics - Theory and Methods, 42, 723–740, 2013

[10] R. Nelsen. *An Introduction to Copulas, 2nd edition.* Springer, New York, 2006.

[11] X. Pedeli and D. Karlis. *A bivariate INAR(1) process with application.* Statistical Modelling: An International Journal, 11(4), 325–349, 2011.

[12] X. Pedeli. *Modelling Multivariate Time Series for Count Data.* PhD thesis, Athens University of Economis And Business, 2011

[13] I. M. M. Silva. *Contributions to the analysis of discrete-valued time series.* PhD thesis, University of Porto, 2005

[14] M. Sklar. *Fonctions de Repartition a n Dimensions et Leurs Marges.* Publications de l'Institut de Statistique de L'Universite de Paris, 8, 229–231, 1959.

Efficient estimation for diffusions

Nina Munkholt Jakobsen[*1]

[1]*Department of Mathematical Sciences, University of Copenhagen*

We consider estimation of the diffusion parameter of a diffusion process observed over a fixed time interval. We present conditions on approximate martingale estimating functions under which estimators are rate optimal and efficient in the case of in-fill asymptotics. In this setup, limit distributions of the estimators are non-standard, in the sense that they are usually normal variance-mixtures. In particular, the mixing distribution depends on the full sample path of the diffusion process over the observation time interval. We also present the more applicable result that, after a suitable data-dependent normalisation, estimators converge in distribution to a standard Gaussian limit. The results presented here are joint work with Michael Sørensen, and published in [10].

Keywords: Stochastic differential equations, approximate martingale estimating functions, in-fill asymptotics, rate optimality, stable convergence

1 Introduction

Diffusion processes are used in a variety of fields to model continuous-time dynamics, for instance, in biology, finance, and neuroscience. However, the corresponding data are usually only observable at discrete time-points. Except in a few simple cases, the likelihood function based on the discrete-time observations is not known explicitly. Thus, for parameter estimation, alternatives to maximum likelihood estimation must be considered.

Here, we focus on a one-dimensional diffusion process $(X_t^\theta)_{t\geq 0}$, which solves a stochastic differential equation of the form

$$dX_t^\theta = a(X_t^\theta)\,dt + b(X_t^\theta, \theta)\,dW_t\,,$$

$\theta \in \Theta$, where $(W_t)_{t\geq 0}$ is a standard Wiener process. Let $\theta_0 \in \Theta$ denote the true, unknown parameter. We assume observations of $(X_t^{\theta_0})_{t\geq 0}$ over the fixed time-interval $[0,1]$ at times $t_i^n = i\Delta_n$, $i = 0, 1, \ldots, n$, with $\Delta_n = 1/n$. In the following, we put $X_t = X_t^{\theta_0}$ and $X_i^n = X_{t_i^n}^{\theta_0}$.

*Corresponding author: munkholt@math.ku.dk

For simplicity, we assume that $\Theta \subseteq \mathbb{R}$, but an extension of the following results to a multidimensional parameter would be straightforward. Similarly, the observation time interval $[0, 1]$ may be generalised to other compact time intervals by rescaling of the drift and diffusion coefficients a and b.

We consider estimators of the diffusion parameter θ, which are based on approximate martingale estimating functions. Many well-known estimators proposed in the literature may be formulated in terms of these estimating functions, see [12]. Our aim is to give a simple characterisation of the estimating functions that produce efficient estimators of the diffusion parameter when the sample size n increases to infinity.

Here, an approximate martingale estimating function $G_n(\theta)$ may be written on the form

$$G_n(\theta) = \sum_{i=1}^{n} g(\Delta_n, X_i^n, X_{i-1}^n, \theta)\,.$$

It is given by a real-valued function $g(t, y, x, \theta)$, which satisfies that for all $\theta \in \Theta$, the conditional expectation

$$\mathbb{E}\left(g\left(\Delta_n, X_{t_i^n}^\theta, X_{t_{i-1}^n}^\theta, \theta\right) \mid X_{t_{i-1}^n}^\theta\right)$$

is of order Δ_n^γ, for some constant $\gamma \geq 2$. A G_n-estimator solves the estimating equation $G_n(\theta) = 0$.

Under other asymptotic scenarios often considered for diffusion processes, limit distributions of estimators are typically Gaussian, with variances depending on θ_0, see e.g. [2, 4, 6, 11, 12]. Under the sampling scheme considered here, the limit distributions are usually normal variance-mixture distributions. In addition to depending on θ_0, these distributions may also depend on the full sample path of the diffusion process over the observation time interval. Estimation and asymptotics under the current observation scheme have previously been treated by, e.g., [1, 3, 5].

It was shown in [1, 5] that under suitable regularity conditions, the model and observation scheme considered here satisfy the local asymptotic mixed normality property with rate $\sqrt{n}$ and random asymptotic Fisher information

$$\mathcal{I}(\theta_0) = 2\int_0^1 \frac{\partial_\theta b(X_s, \theta_0)^2}{b^2(X_s, \theta_0)}\, ds\,.$$

Here, $\partial_\theta b(x, \theta)$ denotes the first partial derivative of b with respect to θ. This result is used to characterise a consistent estimator $\hat{\theta}_n$ as rate optimal and

efficient if

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{\mathcal{D}} L$$

as $n \to \infty$, where $L = \mathcal{I}(\theta_0)^{-1/2} Z$, with Z standard normal distributed and independent of $\mathcal{I}(\theta_0)$. We may interpret $\sqrt{n}$ as the fastest possible rate of convergence in distribution, and L as the limit distribution with the smallest possible variance, conditionally on $\mathcal{I}(\theta_0)$.

2 Main results

The work presented in [10] establishes existence, uniqueness, and asymptotic distribution results concerning consistent G_n-estimators, addressing the question of their rate optimality and efficiency. The essence of the main results of [10], Theorem 3.2 and Corollary 3.4, is summarized in the following Theorem 1, and Corollaries 1 and 2. Technicalities, as well as the existence and uniqueness results, are omitted here.

Theorem 1. *Assume suitable regularity assumptions. Suppose that*

$$\partial_y g(0, y, x, \theta)|_{y=x} = 0 \tag{1}$$

for all x and θ. Then, for any consistent G_n-estimator $\hat{\theta}_n$,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} W(\theta_0) Z$$

as $n \to \infty$, where Z is standard normal distributed and independent of

$$W(\theta_0) = \frac{\left(2 \int_0^1 b^4(X_s, \theta_0) \partial_y^2 g(0, X_s, X_s, \theta_0)^2 \, ds\right)^{1/2}}{\int_0^1 \partial_\theta b^2(X_s, \theta_0) \partial_y^2 g(0, X_s, X_s, \theta_0) \, ds}. \tag{2}$$

Here, $\partial_y^2 g$ denotes the second partial derivative of g with respect to y. Condition (1) ensures estimators that converge at the optimal rate $\sqrt{n}$. The proof of Theorem 1 relies on, among others, results from [8, 9], including a stable central limit theorem, Theorem IX.7.28, from [8]. The expression (2) reveals that $W(\theta_0)$ is usually random, and depends on the full sample path of the diffusion process over the observation time interval. For finite sample sizes, this sample path is only observed at discrete time-points. We use properties of stable convergence in distribution to deal with these complications. The result in Corollary 1 below shows that when suitably normalised, the estimators from Theorem 1 converge in distribution to a standard Gaussian limit.

Corollary 1. Assume suitable regularity assumptions, and suppose that (1) holds. Let $\hat{\theta}_n$ be any consistent G_n-estimator. Then

$$\widehat{W}_n = -\frac{\left(\frac{1}{\Delta_n}\sum_{i=1}^{n} g^2(\Delta_n, X_i^n, X_{i-1}^n, \hat{\theta}_n)\right)^{1/2}}{\sum_{i=1}^{n} \partial_\theta g(\Delta_n, X_i^n, X_{i-1}^n, \hat{\theta}_n)}$$

satisfies that $\widehat{W}_n \xrightarrow{\mathcal{P}} W(\theta_0)$, and it holds that

$$\sqrt{n}\,\widehat{W}_n^{-1}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)\,.$$

Finally, the additional condition (3) ensures efficiency of the estimators.

Corollary 2. Assume suitable regularity assumptions. Suppose that (1) and

$$\partial_y^2 g(0, y, x, \theta)\big|_{y=x} = C_\theta \frac{\partial_\theta b^2(x,\theta)}{b^4(x,\theta)} \tag{3}$$

hold for all x and θ, where C_θ is a non-zero constant. Then any consistent G_n-estimator is efficient.

For example, it may be verified that the estimating function given by

$$\tilde{g}(t, y, x, \theta) = \frac{\partial_\theta b^2(x,\theta)}{b^4(x,\theta)}\left((y-x)^2 - t b^2(x,\theta)\right)$$

satisfies (1) and (3), and corresponds to the efficient contrast function in [3], Theorem 5. It should be noted that conditions (1) and (3) also appear in [7, 12] under other sampling scenarios. Consequently, a number of approximate martingale estimating functions discussed in those papers satisfy our rate optimality and efficiency conditions.

3 Simulation study

The paper [10] also includes a simulation study. Visual comparisons are made of distributions pertaining to estimators based on two approximate martingale estimating functions, which are not covered by the theory of [3]. An excerpt from this simulation study is summarized here. Ten thousand sample paths of a diffusion process given by

$$dX_t^\theta = -2X_t^\theta\,dt + (\theta + (X_t^\theta)^2)^{-1/2}\,dW_t \tag{4}$$

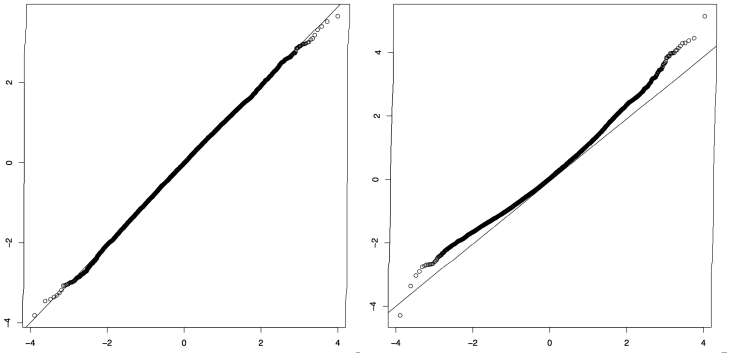


Figure 1: Q-Q plots comparing the distribution of $\sqrt{n}\,\widehat{W}_n^{-1}(\hat{\theta}_n - \theta_0)$ for the efficient (left) and inefficient (right) estimator, respectively, to the standard normal distribution, when $n = 1000$.

were simulated with $\theta_0 = 1$ and $X_0 = 0$, and parameter estimates were computed using the two estimating functions. These estimating functions were given by h and $\tilde{h}$, respectively:

$$h(t, y, x, \theta) = (y - (1 - 2t)x)^2 - (\theta + x^2)^{-1}t$$
$$\tilde{h}(t, y, x, \theta) = (\theta + x^2)^{10}h(t, y, x, \theta)$$

The functions h and $\tilde{h}$ both satisfy the rate-optimality condition (1). However, only h satisfies the efficiency condition (3) for the model (4). Figure 1 shows Q-Q plots comparing the distribution of $\sqrt{n}\,\widehat{W}_n^{-1}(\hat{\theta}_n - \theta_0)$ for the efficient (left) and inefficient (right) estimating function, respectively, to the standard normal distribution, when the sample size is $n = 1000$. In this example from [10], it seems that as the sample size increases, the standard normal distribution becomes a good approximation faster in the efficient case than in the inefficient case. This is an interesting observation, as the current theory does not speak about the speed of this convergence.

Acknowledgements: The work was partially supported by the Danish Council for Independent Research – Natural Science, through a grant to Susanne Ditlevsen. The research is also part of the Dynamical Systems Interdisciplinary Network funded by the University of Copenhagen Programme of Excellence.

References

[1] G. Dohnal. On estimating the diffusion coefficient. *Journal of Applied Probability*, 24(1):105–114, 1987.

[2] V. Genon-Catalot. Maximum contrast estimation for diffusion processes from discrete observations. *Statistics*, 21(1):99–116, 1990.

[3] V. Genon-Catalot and J. Jacod. On the estimation of the diffusion coefficient for multi-dimensional diffusion processes. *Annales de l'institut Henri Poincaré (B) Probabilités et Statistiques*, 29(1):119–151, 1993.

[4] A. Gloter and M. Sørensen. Estimation for stochastic differential equations with a small diffusion coefficient. *Stochastic Processes and their Applications*, 119(3):679–699, 2009.

[5] E. Gobet. Local asymptotic mixed normality property for elliptic diffusion: a Malliavin calculus approach. *Bernoulli*, 7(6):899–912, 2001.

[6] M. Jacobsen. Discretely observed diffusions: Classes of estimating functions and small Delta-optimality. *Scandinavian Journal of Statistics*, 28(1):123–149, 2001.

[7] M. Jacobsen. Optimality and small Delta-optimality of martingale estimating functions. *Bernoulli*, 8(5):643–668, 2002.

[8] J. Jacod and A. Shiryaev. *Limit Theorems for Stochastic Processes*. Number 288 in Grundlehren der mathematischen Wissenschaften. Springer-Verlag, Berlin Heidelberg New York, 2nd edition, 2003.

[9] J. Jacod and M. Sørensen. Aspects of asymptotic statistical theory for stochastic processes. Preprint. University of Copenhagen, 2012.

[10] N. M. Jakobsen and M. Sørensen. Efficient estimation for diffusions sampled at high frequency over a fixed time interval. *Bernoulli*, 23(3):1874–1910, 2017.

[11] M. Kessler. Estimation of an ergodic diffusion from discrete observations. *Scandinavian Journal of Statistics*, 24:211–229, 1997.

[12] M. Sørensen. Efficient estimation for ergodic diffusions sampled at high frequency. Preprint. University of Copenhagen, 2015.

Estimates for distributions of Hölder semi-norms of random processes from spaces $\mathbb{F}_\psi(\Omega)$

Dmytro Zatula[*1]

[1] *Taras Shevchenko National University of Kyiv*

We provide estimates for distributions of semi-norms of sample functions of random processes from spaces $\mathbb{F}_\psi(\Omega)$, defined on a compact space and on an infinite interval $[0, \infty)$, in Hölder spaces.

Keywords: random processes, $\mathbb{F}_\psi(\Omega)$ spaces of random variables, moduli of continuity, Hölder spaces, semi-norms

1 Introduction

$\mathbb{F}_\psi(\Omega)$ spaces of random variables and processes belonging to these spaces are investigated by Kozachenko and Mlavets' [5].

In the following we deal with estimates of distributions of Hölder semi-norms of sample functions of random processes from spaces $\mathbb{F}_\psi(\Omega)$, i.e. probabilities

$$\mathsf{P}\left\{\sup_{\substack{0<\rho(t,s)\leq\varepsilon\\ t,s\in\mathbb{T}}}\frac{|X(t)-X(s)|}{f(\rho(t,s))}>x\right\}.$$

Such estimates and assumptions under which semi-norms of sample functions of random processes from spaces $\mathbb{F}_\psi(\Omega)$, defined on a compact space, satisfy the Hölder condition were obtained by Zatula and Kozachenko [7]. Similar results were provided for Gaussian processes, defined on a compact space, by Dudley [3]. Buldygin and Kozachenko [2] generalized Dudley's results for random processes belonging to Orlicz spaces. Marcus and Rosen [4] obtained L^p moduli of continuity for a wide class of continuous Gaussian processes. Kozachenko et al. [6] studied the Lipschitz continuity of generalized sub-Gaussian processes and provided estimates for the distribution of Lipschitz norms of such processes. But all these problems were not considered yet for processes, defined on an infinite interval.

*Corresponding author: dm.zatula@gmail.com

2 Preliminaries

Definition 1. Let $\psi(u) > 0$, $u \geq 1$ be some increasing function such that $\psi(u) \to \infty$ as $u \to \infty$. We say that a random variable ξ belongs to the space $\mathbb{F}_\psi(\Omega)$ (see [5]) if

$$\sup_{u \geq 1} \frac{(\mathsf{E}|\xi|^u)^{1/u}}{\psi(u)} \leq \infty.$$

It is proved in the paper [5] that $\mathbb{F}_\psi(\Omega)$ is a Banach space with respect to the norm

$$\|\xi\|_\psi = \sup_{u \geq 1} \frac{(\mathsf{E}|\xi|^u)^{1/u}}{\psi(u)}.$$

Theorem 1 ([5]). *If a random variable ξ belongs to the space $\mathbb{F}_\psi(\Omega)$, then*

$$\mathsf{P}\{|\xi| > x\} \leq \inf_{u \geq 1} \frac{\|\xi\|_\psi^u \cdot (\psi(u))^u}{x^u}$$

for all $x > 0$.

Let $\xi_1, ..., \xi_n$ be random variables belonging to the space $\mathbb{F}_\psi(\Omega)$. Put $\eta_n = \max\limits_{1 \leq k \leq n} |\xi_k|$, $a_n = \max\limits_{1 \leq k \leq n} \|\xi_k\|_\psi$.

Definition 2. An $\mathbb{F}_\psi(\Omega)$ space has the property Z if there are monotonically non-decreasing function $z(x) > 0$, monotonically increasing function $U(n)$ and the real number $x_0 > 0$ such that for all sequence of random variables $(\xi_k,\ k = \overline{1, n})$ from the space $\mathbb{F}_\psi(\Omega)$, $\forall x > x_0$ and for all $n \geq 2$ the following holds

$$\mathsf{P}\{\eta_n > x \cdot a_n \cdot U(n)\} \leq \frac{1}{n} \exp\{-z(x)\}.$$

Definition 3 ([5]). We say that a random process $X = \{X(t),\ t \in \mathbb{T}\}$ belongs to the space $\mathbb{F}_\psi(\Omega)$ if random variables $X(t)$ belong to $\mathbb{F}_\psi(\Omega)$ for all $t \in \mathbb{T}$.

Definition 4 ([2]). Let $(\mathbb{T},\ \rho)$ be a metric space. The metric massiveness $N(u) := N_{(\mathbb{T},\ \rho)}(u)$ is the minimal number of closed balls (defined with respect to the metric ρ) that cover $\mathbb{T}$ and that have radiuses which do not exceed u.

Definition 5 ([2]). A function $q = \{q(t), t \in \mathbb{R}\}$ is called the modulus of continuity if $q(t) \geq 0$, $q(0) = 0$ and $q(t + s) \leq q(t) + q(s)$ for $t > 0$ and $s > 0$.

Definition 6 ([1]). A function $v(x)$ satisfy Hölder condition with exponent $\alpha \in (0, 1]$ if the following value is finite:

$$[v]_{\alpha, \mathbb{T}} = \sup_{\substack{t,s \in \mathbb{T} \\ t \neq s}} \frac{|v(t) - v(s)|}{|t - s|^\alpha}.$$

Hölder space $C^{0,\alpha}(\overline{\mathbb{T}})$ is a space of all continuous functions such that the Hölder condition is satisfied with exponent α in the space $\mathbb{T}$.

In the present we deal with a generalization of the semi-norm $[v]_{\alpha,\mathbb{T}}$ in the space $C^{0,\alpha}(\overline{\mathbb{T}})$. Let's consider the quantity

$$[v]_{q,\rho,\mathbb{T}} = \sup_{\substack{t,s\in\mathbb{T}\\ t\neq s}} \frac{|v(t)-v(s)|}{q(\rho(t,s))},$$

where ρ is a metric in the space $\mathbb{T}$, and $q = \{q(t), t \in \mathbb{T}\}$ is a modulus of continuity such that $\exists\, \alpha \in (0,1]\ \forall t,s \in \mathbb{T}, t \neq s : q(\rho(t,s)) \leq |t-s|^{\alpha}$.

3 Main results

In this section we formulate theorems on estimates for distributions of Hölder semi-norms and moduli of continuity of random processes from spaces $\mathbb{F}_{\psi}(\Omega)$, defined on a compact space and on infinite interval.

Theorem 2. *Let $(\mathbb{T}, \rho)$ be a metric compact space. Consider a separable random process $X = \{X(t),\, t \in \mathbb{T}\}$ belonging to the space $\mathbb{F}_{\psi}(\Omega)$ that has the property Z with functions $U(n)$, $z(x)$ and $x_0 > 0$.*

Suppose that there is a monotonically increasing continuous function $\sigma = \{\sigma(h),\, h \geq 0\}$ such that $\sigma(h) > 0$ as $h > 0$, $\sigma(0) = 0$ and the following inequality holds

$$\sup_{\rho(t,s)\leq h} \|X(t) - X(s)\|_{\psi} \leq \sigma(h). \tag{1}$$

Let $N(\varepsilon) = N_{\rho}(\mathbb{T}, \varepsilon)$ be a metric massiveness of the space $(\mathbb{T}, \rho)$. Consider $\varepsilon_0 = \sigma^{(-1)}\left(\sup\limits_{t,s\in\mathbb{T}} \rho(t,s)\right)$, where $\sigma^{(-1)}(h)$ is the inverse function of the function $\sigma(h)$, and

$$g_B(\varepsilon) = \int\limits_0^{\sigma(\varepsilon)} U(B^2 N^2(\sigma^{(-1)}(t)))dt < \infty, \qquad \varepsilon > 0.$$

Then for $x > x_0$, $\varepsilon \in (0, \varepsilon_0)$ and $B > 1$ the following inequality holds true

$$\mathsf{P}\left\{\sup_{0<\rho(t,s)\leq\varepsilon} \frac{|X(t)-X(s)|}{(6+4\sqrt{2})f_B(\rho(t,s)) + (5+2\sqrt{6})g_B(\rho(t,s))} > x\right\} \leq$$
$$\leq \frac{2B(2B+1)}{(B^2-1)N(\varepsilon)} \cdot \exp\{-z(x)\},$$

where $f_B(\varepsilon) = \int\limits_0^{\sigma(\varepsilon)} U(BN(\sigma^{(-1)}(t)))dt$, $\varepsilon > 0$.

Theorem 3. *Let the assumptions of Theorem 2 hold true. Then the following inequality holds*

$$\limsup_{\varepsilon \downarrow 0} \frac{\Delta(X;\varepsilon)}{x_0((6+4\sqrt{2})f_B(\varepsilon) + (5+2\sqrt{6})g_B(\varepsilon))} \le 1$$

with probability 1, where

$$\Delta(X;\varepsilon) = \sup_{\substack{t,s\in\mathbb{T} \\ 0<\rho(t,s)\le\varepsilon}} |X(t) - X(s)|,$$

$$f_B(\varepsilon) = \int\limits_0^{\sigma(\varepsilon)} U(BN(\sigma^{(-1)}(t)))dt,\ g_B(\varepsilon) = \int\limits_0^{\sigma(\varepsilon)} U(B^2N^2(\sigma^{(-1)}(t)))dt < \infty.$$

Now consider an infinite interval $[0,\infty)$. Let $[0,\infty) = \bigcup\limits_{i=0}^{\infty} A_i$, where $A_i = [a_i, a_{i+1}]$ and $\{a_i, i = 0,1,...,\infty\}$ is an increasing sequence, $a_0 = 0$. Denote $\alpha_i = a_{i+1} - a_i$ and $D_i = [a_i, a_{i+1} + \theta]$, where $\theta \in \left(0, \min\limits_{i\ge 0} \alpha_i\right)$. Let $N_i(\varepsilon)$ be metric massiveness for D_i, $i = 0,1,...$ with the metric $\rho(t,s) = |t-s|$, $t,s \in [0,\infty)$.

Theorem 4. *Consider a separable random process* $X = \{X(t), t \in [0,\infty)\}$ *belonging to the Banach space* $\mathbb{F}_\psi(\Omega)$ *that has the property Z with functions* $U(n)$, $z(x)$ *and* $x_0 > 0$. *Suppose that there are monotonically increasing continuous functions* $\sigma_i = \{\sigma_i(h),\ h \ge 0\}$ *such that* $\sigma_i(0) = 0, i = 0,1,...$ *and* $\forall i = 0,1,...$ *the following inequality holds*

$$\sup_{\substack{|t-s|\le h \\ t,s\in D_i}} \|X(t) - X(s)\|_\psi \le \sigma_i(h), \qquad 0 < h < \alpha_i + \theta. \tag{2}$$

Let also

$$\varepsilon_0 = \min_{i\ge 0}\left\{\sigma_i^{(-1)}\left(\sup_{t,s\in D_i}\rho(t,s)\right)\right\} = \min_{i\ge 0}\left\{\sigma_i^{(-1)}(\alpha_i+\theta)\right\},$$

where $\sigma_i^{(-1)}(h)$ *are inverse functions to functions* $\sigma_i(h)$, $i = 0,1,...$, *and* $\forall i = 0,1,...$:

$$g_{B,i}(\varepsilon) = \int\limits_0^{\sigma_i(\varepsilon)} U(B^2N_i^2(\sigma_i^{(-1)}(t)))dt < \infty;$$

$$f_{B,i}(\varepsilon) = \int\limits_0^{\sigma_i(\varepsilon)} U(BN_i(\sigma_i^{(-1)}(t)))dt,\ \varepsilon > 0.$$

Denoting $w_{B,i}(t,s) = (6+4\sqrt{2})f_{B,i}(|t-s|)+(5+2\sqrt{6})g_{B,i}(|t-s|)$, $t,s \in D_i$ *and* $w_B(t,s)$ *is such function that*

$$w_B(t,s) = \{w_{B,i}(t,s)\,|\,t,s \in A_i \ \textit{or}\ \min\{t,s\} \in A_i,\ \max\{t,s\} \in A_{i+1}\},$$

we obtain that for all $x > x_0$, $\varepsilon \in (0, \min\{\varepsilon_0, \theta\})$ *and* $\theta > \varepsilon$ *under the condition* $\sum\limits_{i=0}^{\infty} \frac{1}{\alpha_i} < \infty$ *the following inequality holds true:*

$$\mathsf{P}\left\{\sup_{\substack{0<|t-s|\le\varepsilon\\ t,s\in[0,\infty)}} \frac{|X(t)-X(s)|}{w_B(t,s)} > x\right\} \le \frac{4\varepsilon B(2B+1)}{B^2-1}\cdot\exp\{-z(x)\}\cdot\sum_{i=0}^{\infty}\frac{1}{\alpha_i+\varepsilon}.$$

Acknowledgements: I would like to thank Prof. Yurii Kozachenko for the support in the preparation of this material.

References

[1] A. Bressan. *Lecture Notes on Functional Analysis: With Applications to Linear Partial Differential Equations.* Graduate Studies in Mathematics 143. American Mathematical Society, Providence, RI, 2013.

[2] V. V. Buldygin, Yu. V. Kozachenko. *Metric Characterization of Random Variables and Random Processes.* American Mathematical Society, Providence, RI, 2000.

[3] R. M. Dudley. Sample functions of the Gaussian processes. *Ann. Probab.*, 1(1):3–68, 1973.

[4] M. B. Marcus and J. Rosen. L_p moduli of continuity of Gaussian processes and local times of symmetric Lévy processes. *Ann. Probab.*, 36(2):594–622, 2008.

[5] Yu. V. Kozachenko and Yu. Yu. Mlavets'. The Banach spaces $\mathbf{F}_{\psi}(\Omega)$ of random variables. *Theor. Probability and Math. Statist.*, 86:105–121, 2013.

[6] Yu. V. Kozachenko, T. Sottinen, O. Vasylyk. Lipschitz conditions for $\mathrm{Sub}_{\varphi}(\Omega)$-processes and applications to weakly self-similar processes with stationary increments. *Theor. Probability and Math. Statist.*, 82:57–73, 2011.

[7] D. V. Zatula and Yu. V. Kozachenko. Lipschitz conditions for stochastic processes in the Banach spaces $\mathbb{F}_{\psi}(\Omega)$ of random variables. *Theor. Probability and Math. Statist.*, 91:43–60, 2015.

Finite Mixture of C-vines for Complex Dependence

O. Evkaya *[1], C. Yozgatlıgil[2], and A. S. Kestel[2]

[1]*Atilim University*
[2]*Middle East Technical University*
[2]*Middle East Technical University*

Recently, there has been an increasing interest on the combination of copulas with a finite mixture model. Such a framework is useful to reveal the hidden dependence patterns observed for random variables flexibly in terms of statistical modeling. The combination of vine copulas incorporated into a finite mixture model is also beneficial for capturing hidden structures on a multivariate data set. In this respect, the main goal of this study is extending the study of Kim et al. (2013) with different scenarios. For this reason, finite mixture of C-vine is proposed for multivariate data with different dependence structures. The performance of the proposed model has been tested by different simulated data set including various tail dependence properties.

Keywords: copula; dependence; finite mixture; C-vine; tail dependence

1 Full Inference on C-vine Copula

This section is introduced to recall inference procedures of parameters in Vine copula, exemplified by C-vine copula. Generally, p-dimensional C-vine copula density can be written as in 1,

$$f(\boldsymbol{x};\boldsymbol{\phi_{cvine}}) = \prod_{k=1}^{p} f_k(x_k) \prod_{i=1}^{p-1}\prod_{j=1}^{p-i} c_{i,i+j|1:(j-1)} \left(F(x_i|x_1,...,x_{i-1}), F(x_{i+j}|x_1,...,x_{i-1}); \boldsymbol{\beta}_{i,i+j|(i+1):(i+j-1)}\right) \tag{1}$$

where $f_k(x_k)$ denotes the marginal densities, $c_{i,i+j|1:(j-1)}$ are the bivariate copula density functions with parameter(s) $\beta_{i,(i+j)|(i+1):(i+j-1)}$, and ϕ_{cvine} is the set of all parameters in p-dimensional C-vine density.

*Corresponding author: ozanevkaya@gmail.com

There exist one root node in the tree construction of C-vine model which results in following illustration in 4-dimension given by equation 2,

$$\begin{aligned} f(x_1, x_2, x_3, x_4; \boldsymbol{\phi}) = {} & c_{12}(F(x_1), F(x_2); \beta_{12}) c_{13}(F(x_1), F(x_3); \beta_{13}) \\ & c_{14}(F(x_1), F(x_4); \beta_{14}) \\ & c_{23|1}(F(x_2|x_1), F(x_3|x_1); \beta_{23|1}) c_{24|1}(F(x_2|x_1), F(x_4|x_1); \beta_{24|1}) \\ & c_{34|12}(F(x_3|x_1, x_2), F(x_4|x_1, x_2); \beta_{34|12}) \prod_{k=1}^{4} f_k(x_k) \end{aligned} \tag{2}$$

Under such multivariate framework, full inference on C-vine copula can be derived using the log-likelihood function presented in 3,

$$\begin{aligned} L(\boldsymbol{\phi}) = {} & \sum_{i=1}^{p-1} \sum_{j=1}^{p-i} \sum_{n=1}^{N} \log c_{i,i+j|(1):(j-1)} \\ & (F(x_{i,n}|x_{1,n}, ..., x_{i-1,n}), F(x_{i+j,n}|x_{1,n}, ..., x_{i-1,n}); \boldsymbol{\beta}_{i,i+j|(i+1):(i+j-1)}) \end{aligned} \tag{3}$$

and following three consecutive steps:

Step 1 Decide which variable is used as a root node in the first tree T_1 of a C-vine copula (i.e. joining the variables in which the root node variable is selected based on its significant relations with other variables)

Step 2 Specify the family and parametric shape of each pair-copula in an assumed C-vine copula

Step 3 Estimate all parameters of C-vine by maximizing the log-pseudo likelihood function given in 3

2 Finite Mixture of C-vines

The finite mixture model is introduced to connect m component C-vine densities to detect complex and hidden dependence structures in multivariate data with the related EM algorithm for estimating the parameters in the model.

Assume that a p-dimensional random vector X=(X_1, ..., X_p) is said to be generated from a mixture of M- component C-vine densities, where its density function is defined as in 4.

$$g(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{m=1}^{M} \pi_m f(\boldsymbol{x}, \boldsymbol{\phi_m}) \tag{4}$$

where π_m is the mixing proportion of the m-th component s.t. $0 < \pi_m < 1$ and $\sum_{m=1}^{M} \pi_m = 1$. Besides, ϕ_m is the m-th component-specific parameter vector for the C-vine density described in 4. Note that θ is the set of all parameters with dimension p and denoted by Θ, full product space (the simplex of π_m and the cross product space of ϕ_m). Here p is the total number of free parameters to be estimated and $p = (M-1) + \sum_{m=1}^{M} dim(\phi_m)$. For the estimation of equation 4, both the number of components M and the parameters θ are required to estimate, using the following EM-algorithm setup, proposed previously by Dempster et al. (1977).

Assume that N observations randomly drawn from a M component C-vine density given in 4, denoted as x_k=($x_{k,1}$, ..., $x_{k,p}$) where $k = 1, ..., p$. Then, log-likelihood of θ is described as given in 5

$$L(\boldsymbol{\theta}) = \log(\prod_{n=1}^{N} g(\boldsymbol{x}_n, \boldsymbol{\theta})) = \log(\prod_{n=1}^{N} \sum_{m=1}^{M} \pi_m f(\boldsymbol{x}_n, \boldsymbol{\phi}_m)) \tag{5}$$

Let z_n=($z_{n1}, ..., z_{nm}, ..., z_{nM}$) denotes latent variables, where $z_{nm} = 1$ if x_n drawn from the m-th component and $z_{nm} = 0$ otherwise. Here, z_n is i.i.d. from a multinomial distribution, i.e. z_n is Mult($M, \pi = (\pi_1, ..., \pi_m)$). Under this setup, the complete log-likelihood for the complete data set y_n=(x_n,z_n) is given by equation 6.

$$\begin{aligned} L(\boldsymbol{\theta})_c &= \log \prod_{n=1}^{N} \prod_{m=1}^{M} [\pi_m f(\boldsymbol{x}_n, \boldsymbol{\phi}_m)]^{z_{nm}} \\ &= \sum_{n=1}^{N} \sum_{m=1}^{M} z_{nm} \log(\pi_m) + \sum_{n=1}^{N} \sum_{m=1}^{M} z_{nm} \log(f(\boldsymbol{x}_n, \boldsymbol{\phi}_m)) \end{aligned} \tag{6}$$

Starting with initial values (initial guesses) for the parameters, θ^0, the repeated E-th and M-th step of the EM algorithm (to compute the successive estimates, θ^s) is described as follows:

E-step Calculates the conditional expectation of $L(\theta)_c$ given the observed data and current parameter estimates for θ. Such a computation is equivalent to the calculation of posterior probability that x_n belongs to the m-th

component, given the current values of the parameters formulated as in equation 7

$$\widehat{z_{nm}}^{(s)} = E[z_{nm}|\boldsymbol{x}, \theta^s] = P[z_{nm} = 1|\boldsymbol{x}, \theta^s] = \frac{\pi_m^{(s)} f(\boldsymbol{x}_n, \boldsymbol{\phi}_m^{(s)})}{\sum_{l=1}^{M} \pi_l^{(s)} f(\boldsymbol{x}_n, \boldsymbol{\phi}_l^{(s)})} \quad (7)$$

M-step Computes the parameter estimates for each component independently, $(\pi_1^{(s+1)}, ..., \pi_m^{(s+1)}, ..., \pi_M^{(s+1)})$ and $(\phi_1^{(s+1)}, ..., \phi_m^{(s+1)}, ..., \phi_M^{(s+1)})$ by maximizing the expected complete-data log-likelihood from E-step. It is also possible to obtain closed form solution for $\pi_m^{(s+1)} = \frac{\sum_{n=1}^{N} \widehat{z_{nm}}^{(s)}}{N}$. Afterwords, the estimation of $\phi_m^{(s+1)}$ in the m-th component C-vine or D-vine density function is equivalent to deriving the parameter estimates weighted by $\widehat{z_{nm}}^{(s)}$ for the parameters in a C-vine density in equation 4.

The E-step and the M-step are iterated until $L(\theta^{s+1}) - L(\theta^s)$ is smaller than a pre-specified tolerance value (ie. 10^{-6} or 10^{-8}), as a result of a nice property of EM algorithm that the log-likelihood is not decreased during the iteration. Based on the above setup, the given algorithm is run with multiple starting values randomly drawn from the parameter space and the best values is chosen from multiple local maximizer having the highest log-likelihood value. To accomplish the full inference on mixture of C-vines, three well known model selection criteria values are used:

- Akaike's Information Criterian (AIC) as defined $AIC = -2\log(L(\widehat{\theta})) + 2p$
- Bayesian Information Criterian (BIC) as defined $BIC = -2\log(L(\widehat{\theta})) + p\log(n)$
- Consistent AIC (CAIC) as defined $CAIC = -2\log(L(\widehat{\theta})) + p(\log(n) + 1)$

where $\widehat{\theta}$ is the estimate of p-dimensional θ defined in 4. In this study, the main objective is identifying the full inference on C-vine mixture model based on different scenarios. For this reason, the whole procedure for the full inference of C-vine copula can be summarized as follows:

Step 1 Derive the normalized ranks of d-dimensional observed data

Step 2 Decide the root node of each C-vine density by calculating all pairwise correlations.

Step 3 Consider different candidates of copulas for all pairs in an assumed mixture model

Step 4 Given a copula family, fit a mixture vine copula with M component and estimate the parametersin each model by employing the EM- algorithm

Step 5 Select the best fitted model by finding the model with smallest values of model selection criterians such as AIC, BIC and CAIC.

For the last step, naturally, even if the selection criteria measures give meaningful conclusions, they are not enough to decide the best fitted model. For this reason, available GOF tests are also required to improve the model selection part. Especially, Clarke and Vuong tests are widely used GOF test for comparing two different vine copulas might be considered in model selection.

3 Simulation Results

To test the performance of the mixture model, the base mixture model is constructed using Clayton and Gumbel pairs with parameters ($\beta_{12}^C = 8, \beta_{13}^C = 7, \beta_{23|1}^C = 6$) and ($\beta_{12}^G = 9, \beta_{13}^G = 6, \beta_{23|1}^G = 5$), respectively. As the number of parameters described, as a simple case, the mixture of C-vines are considered in 3-dimension with positive strong tail dependencies.

Here, as a simulated data, 2 component equally weighted mixture of C-vines is considered under the proposed mixture model to investigate the data generating process performs well or not. In this setup, the number of observations has been increased from $N = 50$ (small data set) to $N = 1000$ (large data set) to see the differences in parameter estimation process. For now, parameter estimation results of 2-component C-vine mixture with Clayton pairs are presented for illustration. Here, the estimated parameters for each pair of both components are obtained using the average value and the median values of 1000 different run given in () and [] table, respectively.

In table 1, as it is expected, parameter estimations of the first component are very close to true value since the correct copula family at each step is predefined as Clayton at the beginning for the simulated data. Besides, the number of observations has positive impact on closing the gap between parameter estimates and true values. Generally, the most suitable model will be determined by comparing the model comparison values among different scenarios like Clayton-Clayton, Clayton-Gumbel, Frank-Gumbel, assumed pair copulas in mixture model. As we expected, the most plausible result will be obtained from the Clayton-Gumbel pair families selection for the first-second component, same as the original simulated data.

Table 1: Parameter Estimation and Model Comparison Values

	Number of Observations				
	50	100	250	500	1000
$\widehat{\beta^C_{12}}$	(7.19)[7.3]	(7.95)[8.05]	(8.77)[8.82]	(8)[8]	(8)[8]
$\widehat{\beta^C_{13}}$	(6.49)[6.6]	(7.08)[7.15]	(7.76)[7.71]	(7)[7]	(7)[7]
$\widehat{\beta^C_{23\|1}}$	(4.29)[4.24]	(4.29)[4.25]	(4.81)[4.81]	(6)[6]	(6)[6]
$\widehat{\beta^C_{12}}$	(7.6)[7.66]	(7.29)[7.27]	(7.12)[7.22]	(7.07)[6.98]	(7.65)[7.39]
$\widehat{\beta^C_{13}}$	(6.3)[6.32]	(5.8)[5.79]	(5.55)[5.66]	(5.5)[5.43]	(6.03)[5.87]
$\widehat{\beta^C_{23\|1}}$	(3.32)[3.11]	(2.85)[2.53]	(2.53)[2.42]	(2.56)[2.54]	(2.14)[2.17]
AIC	-194	-363	-783	-1536	-3001
BIC	-188	-356	-773	-1524	-2986
CAIC	-185	-353	-770	-1521	-2983

References

[1] R.B Nelsen. *An Introduction to Copulas* 2nd edn. Springer Series in Statistics. Springer, New York, NY, 2006.

[2] K. Aas., C. Czado, A. Frigessi and H. Bakken. Pair-copula constructions of multiple dependence. *Insurance, Mathematics and Economics.* 44:182–198, 2009.

[3] T. Bedford, R. Cooke. Vines a new graphical model for dependent random variables. *Ann. Stat..* 30(4):1031–1068, 2002.

[4] D. Kim, J.M. Kim, S.M. Liao, Y.S. Jung. Mixture of D-vine Copulas for Modeling Dependence. *Computational Statistics and Data Analysis.* 64:1–19, 2013.

[5] E.C. Brechman, U. Schepsmeir. Modeling Dependence with C- and D-Vine Copulas: The R Package CDVine. *Journal of Statistical Software.* 52(3):1–27, 2013.

On Auto-Distance Correlation Matrix

Maria Pitsillou *[1] **and Konstantinos Fokianos**[1]

[1]*Department of Mathematics & Statistics, University of Cyprus*

We introduce the notions of auto-distance covariance and correlation matrices for multivariate time series and give their consistent estimators. In addition, a testing methodology for testing the i.i.d. hypothesis for multivariate time series data is developed. The resulting test statistic is compared with the related multivariate Ljung-Box statistic in a real data example.

Keywords: Characteristic function ; Correlation ; Stationarity ; U-statistic ; Wild Bootstrap

1 Introduction

There has been a considerable recent interest in measuring dependence by employing the concept of distance covariance function, a new measure of dependence for random variables, introduced by Székely et al. (2007). This tool has been recently defined to the context of multivariate time series by Zhou (2012), but without exploring the interrelationships between the various time series components. In this paper, we extend the notion of distance covariance to multivariate time series by defining its matrix version. Based on this new concept, we develop a multivariate testing methodology for testing independence.

2 Auto-distance covariance matrix

We denote by $\{\mathbf{X}_t : t = 0, \pm 1, \pm 2, \dots\}$ a d-dimensional time series process, with components $X_{t;r}$, $r = 1, \dots, d$. Suppose we have available a sample of size n, that is $\{\mathbf{X}_t, t = 1, \dots, n\}$. We define the pairwise auto-distance covariance function as a function of the joint and marginal characteristic functions of the pair $(X_{t;r}, X_{t+j;m})$, for $r, m = 1, \dots, d$. Denote by $\phi_j^{(r,m)}(u, v)$ the joint characteristic function of $X_{t;r}$ and $X_{t+j;m}$; that is

$$\phi_j^{(r,m)}(u, v) = E\left[\exp\left(i(uX_{t;r} + vX_{t+j;m})\right)\right], \quad j \in \mathbb{Z},$$

*Corresponding author: pitsillou.maria@ucy.ac.cy

and the marginal characteristic functions of $X_{t;r}$ and $X_{t+j;m}$ as $\phi^{(r)}(u) := \phi_j^{(r,m)}(u,0)$ and $\phi^{(m)}(v) := \phi_j^{(r,m)}(0,v)$ respectively, where $(u,v) \in \mathbb{R}^2$, and $i^2 = -1$. The *pairwise auto-distance covariance function (ADCV)* between $X_{t;r}$ and $X_{t+j;m}$, $V_{rm}(j)$, is defined as the positive square root of

$$V_{rm}^2(j) = \frac{1}{\pi^2} \int_{\mathbb{R}^2} \frac{\left|\phi_j^{(r,m)}(u,v) - \phi^{(r)}(u)\phi^{(m)}(v)\right|^2}{|u|^2 |v|^2} dudv, \quad j \in \mathbb{Z}.$$

The auto-distance covariance matrix, $V(j)$, is then defined by

$$V(j) = [V_{rm}(j)]_{r,m=1}^d, \quad j \in \mathbb{Z}.$$

The *pairwise auto-distance correlation function (ADCF)* between $X_{t;r}$ and $X_{t+j;m}$, $R_{rm}(j)$, is a coefficient that lies in the interval $[0,1]$ and also measures dependence and is defined as the positive square root of

$$R_{rm}^2(j) = \frac{V_{rm}^2(j)}{\sqrt{V_{rr}^2(0)}\sqrt{V_{mm}^2(0)}},$$

for $V_{rr}(0)V_{mm}(0) \neq 0$ and zero otherwise. The auto-distance correlation matrix of $\mathbf{X}_t$, is then defined as

$$R(j) = [R_{rm}(j)]_{r,m=1}^d, \quad j \in \mathbb{Z}.$$

When $j \neq 0$, $V_{rm}(j)$ measures the dependence of $X_{t;r}$ on $X_{t+j;m}$. In general, $V_{rm}(j) \neq V_{mr}(j)$ for $r \neq m$, since they measure different dependence structure between the series $\{X_{t;r}\}$ and $\{X_{t;m}\}$ for all $r, m = 1, 2, \ldots, d$. Thus, $V(j)$ and $R(j)$ are non-symmetric matrices, but $V(-j) = V'(j)$ and $R(-j) = R'(j)$. More properties can be found in Fokianos and Pitsillou (2017b). The empirical pairwise ADCV, $\widehat{V}_{rm}(j)$, for $j \geq 0$, is the non-negative square root of

$$\widehat{V}_{rm}^2(j) = \frac{1}{(n-j)^2} \sum_{t,s=1}^{n-j} A_{ts}^r B_{ts}^m,$$

where $A^r = A_{ts}$ and $B^m = B_{ts}$ are Euclidean distance matrices given by

$$A_{ts}^r = a_{ts}^r - \bar{a}_{t.}^r - \bar{a}_{.s}^r + \bar{a}_{..}^r,$$

with $a_{ts}^r = |X_{t;r} - X_{s;r}|$, $\bar{a}_{t.}^r = \left(\sum_{s=1}^{n-j} a_{ts}^r\right)/(n-j)$, $\bar{a}_{.s}^r = \left(\sum_{t=1}^{n-j} a_{ts}^r\right)/(n-j)$, $\bar{a}_{..}^r = \left(\sum_{t,s=1}^{n-j} a_{ts}^r\right)/(n-j)^2$. B_{ts}^m is defined analogously in terms of $b_{ts}^m = |X_{t+j;m} - X_{s+j;m}|$.

Fokianos and Pitsillou (2017b) showed that for a d-dimensional strictly stationary and ergodic process $\{\mathbf{X}_t\}$ with $E\left|X_{t;r}\right|^2 < \infty$, for $r = 1, \ldots, d$, then for all $j \in \mathbb{Z}$,

$$\widehat{V}(j) \to V(j),$$

almost surely, as $n \to \infty$. In addition, under pairwise independence it holds that

$$n\widehat{V}_{rm}^2(j) \to Z := \sum_k \lambda_k Z_k^2,$$

in distribution, as $n \to \infty$, where $\{Z_k\}$ is an i.i.d sequence of $N(0,1)$ random variables, and (λ_k) is a sequence of nonzero eigenvalues.

3 The testing problem

In this section, we develop a test statistic for testing the null hypothesis that $\{\mathbf{X}_t\}$ is an i.i.d. sequence. Following Hong's (1999) generalized spectral domain methodology, we first consider the generalized spectral density matrix

$$F(\omega, u, v) = \left[f^{(r,m)}(\omega, u, v)\right]_{r,m=1}^{d},$$

where

$$f^{(r,m)}(\omega, u, v) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \sigma_j^{(r,m)}(u, v) e^{-ij\omega}, \quad \omega \in [-\pi, \pi],$$

with p denoting the bandwidth parameter. Under the null hypothesis of independence, $F(\cdot, \cdot, \cdot)$ reduces to

$$F_0(\omega, u, v) = \frac{1}{2\pi}\left[\sigma_0^{(r,m)}(u, v)\right]_{r,m=1}^{d}.$$

Thus, comparing the Parzen's (1957) kernel-type estimators $\hat{F}(\omega, u, v)$ and $\hat{F}_0(\omega, u, v)$ via a Frobenious norm we result to a test statistic based on the ADCV matrix, given by

$$\widetilde{T}_n = \sum_{j=1}^{n-1} (n-j) k^2(j/p) \mathrm{tr}\{\widehat{V}^*(j)\widehat{V}(j)\}, \tag{1}$$

where $k(\cdot)$ is a univariate kernel function satisfying some standard properties. Moreover, $\widehat{V}^*(\cdot)$ denotes the complex conjugate matrix of $\widehat{V}(\cdot)$ and $\mathrm{tr}(A)$

denotes the trace of the matrix A. Fokianos and Pitsillou (2017b) also formed a similar test statistic in terms of the ADCF matrix, given by

$$\overline{T}_n = \sum_{j=1}^{n-1} (n-j)k^2(j/p)\mathrm{tr}\{\widehat{V}^*(j)\widehat{D}^{-1}\widehat{V}(j)\widehat{D}^{-1}\}. \tag{2}$$

Under the null hypothesis of independence and some further assumptions about the kernel function $k(\cdot)$, the standardized version of the test statistics $\widetilde{T}_n$ and $\overline{T}_n$ given in (1) and (2) were proved to follow $N(0,1)$ asymptotically and they are consistent. Fokianos and Pitsillou (2017a) developed a similar testing methodology based on ADCV/ADCF for testing serial dependence in a univariate strictly stationary time series setting.

4 Real data example

In this section we apply the proposed testing methodology to the monthly log returns of the stocks of IBM and the S&P 500 composite index starting from 29 May 1936 to 28 November 1975 for 474 observations. A larger data set and the aforementioned testing methodology are included in the R package **dCovTS** (Pitsillou and Fokianos, 2016). Assuming that the bivariate series follows a VAR model and employing the AIC to choose its best order, we obtain that a VAR(2) model fits well the data. Figure 1 shows the ADCF plot of the residuals after fitting a VAR(2) model to the original series. Based on this plot, the residuals of VAR(2) model do not have any strong dependence. The shown critical values (dotted horizontal line) are the 95% simultaneous critical values computed based on an algorithm suggested by Fokianos and Pitsillou (2017b) using the independent wild bootstrap approach (Dehling and Mikosch, 1994; Shao, 2010; Leucht and Neumann, 2013). To formally confirm the adequacy of this model fit, we perform tests of independence among the residuals for the following bandwidth values, $p = 6$, 11 and 20. The proposed statistic $\overline{T}_n$ and the related multivariate Ljung-Box statistic (Hosking, 1980) both give large p-values (0.254, 0.190, 0.098 and 0.958, 0.809, 0.811 respectively) suggesting the absence of any serial dependence among the residuals. The calculation of the statistic $\overline{T}_n$ is based on the Bartlett kernel. The computation of the p-values is based on 499 independent wild bootstrap realizations.

Acknowledgements: Financial support from a University of Cyprus research grant is greatly acknowledged.

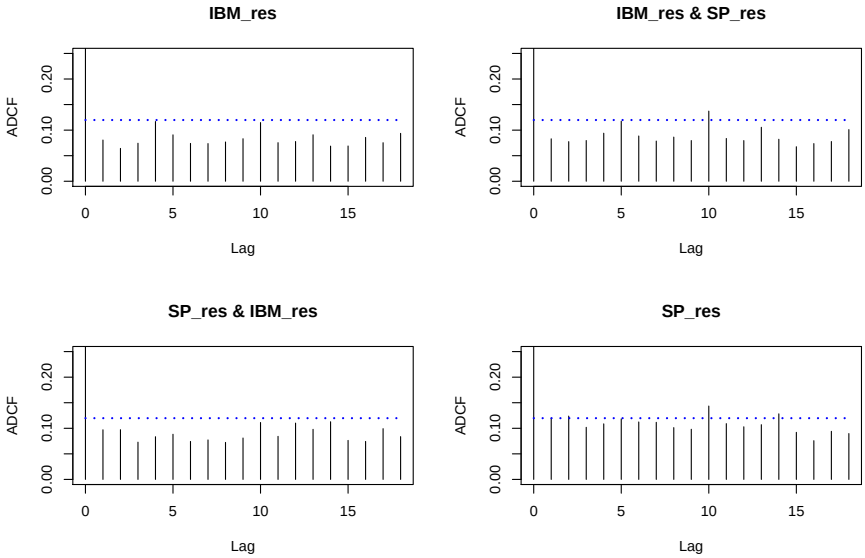


Figure 1: The sample ADCF of the residuals after fitting VAR(2) model to the bivariate series IBM and S&P500.

References

Dehling, H. and T. Mikosch (1994). Random quadratic forms and the boostrap for U-statistics. *Journal of Multivariate Analysis* **51**, 392–413.

Fokianos, K. and M. Pitsillou (2017a). Consistent testing for pairwise dependence in time series. *Technometrics* **59**, 262–270.

Fokianos, K. and M. Pitsillou (2017b). Testing pairwise independence for multivariate time series by the auto-distance correlation matrix. Under revision.

Hong, Y. (1999). Hypothesis testing in time series via the emprical characteristic function: A generalized spectral density approach. *Journal of the American Statistical Association* **94**, 1201–1220.

Hosking, J. R. M. (1980). Multivariate Portmanteau statistic. *Journal of the American Statistical Association* **75**, 349–386.

Leucht, A. and M. H. Neumann (2013). Dependent wild bootstrap for degenerate U- and V- statistics. *Journal of Multivariate Analysis* **117**, 257–280.

Parzen, E. (1957). On consistent estimates of the spectrum of a stationary time series. *Annals of Mathematical Statistics* **28**, 329–348.

Pitsillou, M. and K. Fokianos (2016). dCovTS: Distance covariance and correlation for time series analysis. *The R Journal* **8**, 324–340.

Shao, X. (2010). The dependent wild bootstrap. *Journal of the American Statistical Association* **105**, 218–235.

Székely, G. J., M. L. Rizzo, and N. K. Bakirov (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics* **35**, 2769–2794.

Zhou, Z. (2012). Measuring nonlinear dependence in time-series, a distance correlation approach. *Journal of Time Series Analysis* **33**, 438–457.

Stability of the Spectral EnKF under nested covariance estimators

Marie Turčičová*[1], Jan Mandel[2], and Kryštof Eben[3]

[1] *Charles University in Prague, Faculty of Mathematics and Physics, Sokolovská 83, Prague 8, 186 75, Czech Republic*
[2] *University of Colorado Denver, Denver, CO 80217-3364, USA*
[3] *Czech Academy of Sciences, Institute of Computer Science, Pod Vodárenskou věží 271/2, 182 07 Praha 8, Czech Republic*

In the case of traditional Ensemble Kalman Filter (EnKF), it is known that the filter error does not grow faster than exponentially for a fixed ensemble size [5]. The question posted in this contribution is whether the upper bound for the filter error can be improved by using an improved covariance estimator that comes from the right parameter subspace and has smaller asymptotic variance. Its effect on Spectral EnKF is explored by a simulation.

Keywords: nested covariance models, maximum likelihood, error of EnKF

1 Introduction

Estimating of large covariance matrices from small samples is an important problem in many fields, including spatial statistics, genomics, and ensemble filtering. One of the prominent applications is data assimilation, where a prior estimate of a random vector (usually representing a system state) is adjusted in order to be more consistent with current observations. The revised estimate is then plugged into a time-evolution model as an initial condition for the future time prediction. This approach, known as filtering, is used in many fields including meteorological predictions. A characteristic feature of this application is a large dimension of the system state (millions or larger), which results in high computational cost. One algorithm that deals with this problem is the Ensemble Kalman filter (EnKF), which approximates the mean and the covariance of the state vector from an ensemble. However, due to the high computational cost, this ensemble is always very small compared to the state dimension, and the approximation is very poor. In this contribution, we study improved estimation of the covariance matrix from a small ensemble, and its behaviour in high-dimensional EnKF.

*turcic@karlin.mff.cuni.cz

In particular, we consider a very special type of sparse approximation of a covariance matrix in spectral space, based on nested maximum likelihood models for diagonal matrices. The improved covariance estimator seems to have a positive effect in data assimilation, which is illustrated by a simulation.

2 Hierarchical maximum likelihood estimators

Suppose $\mathbb{X}_N = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N]$ is a random sample from a distribution on $\mathbb{R}^n$ with density $f(\boldsymbol{x}, \boldsymbol{\theta})$ with unknown parameter vector $\boldsymbol{\theta}$ in a parameter space $\Theta \subset \mathbb{R}^p$. The maximum likelihood estimator (MLE) $\hat{\boldsymbol{\theta}}_N$ of the true parameter $\boldsymbol{\theta}^0$ is defined by maximizing the log-likelihood $\ell(\boldsymbol{\theta}|\mathbb{X}_N) = \sum_{i=1}^{N} \log f(\boldsymbol{X}_i, \boldsymbol{\theta})$.

Further assume a hierarchical structure of the parameter space,

$$\boldsymbol{\theta}^0 \in \Psi \subset \Phi \subset \Theta,$$

where $\Psi \subset \mathbb{R}^m$, $\Phi \subset \mathbb{R}^k$, $m \le k \le p$. That is, $\boldsymbol{\theta}$ can be parametrized by a smaller number of parameters. We assume that the map $\boldsymbol{\varphi} \mapsto \boldsymbol{\theta}(\boldsymbol{\varphi})$ is one-to-one from Φ to Θ and continuously differentiable. Further assume that the associated Jacobi matrix $\nabla_{\boldsymbol{\varphi}} \boldsymbol{\theta}(\boldsymbol{\varphi}) = \left\{ \frac{\partial \theta_i}{\partial \varphi_j} \right\}$ has full rank for all $\boldsymbol{\varphi} \in \Phi$. We make analogous assumptions about the map $\boldsymbol{\psi} \mapsto \boldsymbol{\theta}(\boldsymbol{\psi})$ as well. Moreover, assume that $\boldsymbol{\theta}^0 = \boldsymbol{\theta}(\boldsymbol{\varphi}^0) = \boldsymbol{\theta}(\boldsymbol{\psi}^0)$ is an interior point of Ψ.

We will also adopt the usual assumptions in the maximum likelihood theory: (i) the density f determines the parameter $\boldsymbol{\theta}$ uniquely in the sense that $f(\boldsymbol{x}, \boldsymbol{\theta}_1) = f(\boldsymbol{x}, \boldsymbol{\theta}_2)$ a.e. if and only if $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$, and (ii) $f(\boldsymbol{x}, \boldsymbol{\theta})$ is a sufficiently smooth function of $\boldsymbol{x}$ and $\boldsymbol{\theta}$ (see [6] for details).

Under these assumptions, the error of the estimates is asymptotically normal

$$\sqrt{N}\left(\boldsymbol{\theta}\left(\hat{\boldsymbol{\varphi}}_N\right) - \boldsymbol{\theta}^0\right) \xrightarrow{d} \mathcal{N}_p\left(\mathbf{0}, Q_{\boldsymbol{\theta}(\boldsymbol{\varphi}^0)}\right) \text{ as } N \to \infty, \tag{1}$$

$$\sqrt{N}\left(\boldsymbol{\theta}\left(\hat{\boldsymbol{\psi}}_N\right) - \boldsymbol{\theta}^0\right) \xrightarrow{d} \mathcal{N}_p\left(\mathbf{0}, Q_{\boldsymbol{\theta}(\boldsymbol{\psi}^0)}\right) \text{ as } N \to \infty. \tag{2}$$

The matrices $Q_{\boldsymbol{\theta}(\boldsymbol{\varphi}^0)}$ and $Q_{\boldsymbol{\theta}(\boldsymbol{\psi}^0)}$ represent asymptotic variances of the parameters. These matrices are singular, but they can be understood as inverses of Fisher information matrices in a generalized sense. Their exact forms are given in [7].

The next theorem shows that for any two nested subspaces Φ and Ψ of the parameter space containing the true parameter, the asymptotic covariance matrices of the MLE are ordered in the same way. Hence, by confining the parameters to a smaller subspace, we can only improve the estimator.

Theorem 1 ([7]). *Under the assumptions listed previously, the matrix* $Q_{\boldsymbol{\theta}(\boldsymbol{\psi}^0)} - Q_{\boldsymbol{\theta}(\boldsymbol{\varphi}^0)}$ *is positive semidefinite (denoted as* $Q_{\boldsymbol{\theta}(\boldsymbol{\varphi}^0)} \leq Q_{\boldsymbol{\theta}(\boldsymbol{\psi}^0)}$*).*

In addition, if $U \sim \mathcal{N}_p\left(\mathbf{0}, Q_{\boldsymbol{\theta}(\boldsymbol{\varphi}^0)}\right)$ *and* $V \sim \mathcal{N}_p\left(\mathbf{0}, Q_{\boldsymbol{\theta}(\boldsymbol{\psi}^0)}\right)$ *are random vectors with the asymptotic distributions of the estimates* $\boldsymbol{\theta}\left(\hat{\boldsymbol{\varphi}}_N\right)$ *and* $\boldsymbol{\theta}(\hat{\boldsymbol{\psi}}_N)$*, then*

$$\mathrm{E}\,|U|^2 = \frac{1}{N}\,\mathrm{Tr}\,Q_{\boldsymbol{\theta}(\boldsymbol{\varphi}^0)} \leq \frac{1}{N}\,\mathrm{Tr}\,Q_{\boldsymbol{\theta}(\boldsymbol{\psi}^0)} = \mathrm{E}\,|V|^2\,, \tag{3}$$

where $|V| = \left(V^\top V\right)^{1/2}$ *is the standard Euclidean norm in* $\mathbb{R}^p$*.*

2.1 One specific hierarchical model for a covariance matrix

Consider three particular nested models for a diagonal covariance matrix. Such models appear to be useful in meteorological practice but more about our motivation will be said in the next section. The models have the form

- $D^{(n)} = \mathrm{diag}\{d_i, i = 1, \ldots, n\}$
- $D^{(3)} = \mathrm{diag}\{(c_1 - c_2\lambda_i)^{-1}(-\lambda_i)^{-\alpha}, i = 1, \ldots, n\}$
- $D^{(2)} = \mathrm{diag}\{c(-\lambda_i)^{-\alpha}, i = 1, \ldots, n\}$,

with $\{\lambda_i\}_{i=1}^n$ being the eigenvalues of a two-dimensional Laplace operator. The superscripts designate the number of parameters of each model. Under the normality assumption, all these parameters can be estimated from a random sample by the maximum likelihood method. Let $\hat{D}^{(n)}, \hat{D}^{(3)}$ and $\hat{D}^{(2)}$ be the resulting estimates. Notice that $\hat{D}^{(n)}$ is formed simply by the diagonal of sample covariance. The asymptotic hierarchical structure of $\mathrm{cov}(\hat{D}^{(n)})$, $\mathrm{cov}(\hat{D}^{(3)})$ and $\mathrm{cov}(\hat{D}^{(2)})$ is theoretically described in the previous section. The exact form of these estimators and their Fisher information matrices can be found in [7].

However, it is difficult to say something general about the MLEs based on small samples (although they are usually more of interest).

The simulations reported in [7] suggest that the hierarchical structure of the error (3) persists also for small samples. Here we use the hierarchical covariance models in data assimilation.

3 Covariance estimators in data assimilation

Our main objective is to demonstrate the positive effect of the improved covariance estimators $\hat{D}^{(3)}$ and $\hat{D}^{(2)}$ in data assimilation. First, let us briefly recall the Ensemble Kalman Filter (EnKF) [2], in the simple case when the whole state is observed. At the beginning, the distribution of the true state vector $\boldsymbol{X}_t$ is represented by a "forecast ensemble" $\boldsymbol{X}_f^1, \ldots, \boldsymbol{X}_f^N$. The

sample covariance of the forecast ensemble is denoted $\hat{C}_f$. Using the perturbed observations $\boldsymbol{y}^1, \ldots, \boldsymbol{y}^N$ (whose error has covariance R), the forecast ensemble is adjusted and results in an "analysis ensemble" $\boldsymbol{X}_a^1, \ldots, \boldsymbol{X}_a^N$, which is supposed to be "closer" to $\boldsymbol{X}_t$. Its sample covariance is denoted $\hat{C}_a$. The process is governed by the following equations:

$$\boldsymbol{X}_a^j = \boldsymbol{X}_f^j + \hat{C}_f \left(\hat{C}_f + R\right)^{-1} \left(\boldsymbol{y}^j - \boldsymbol{X}_f^j\right) \qquad j = 1, \ldots, N \tag{4}$$

$$\hat{C}_a = \left(I - \hat{C}_f \left(\hat{C}_f + R\right)^{-1}\right) \hat{C}_f. \tag{5}$$

Each member of the analysis ensemble is then pushed forward in time by a function $\eta(\cdot)$, which represents the evolution of the process $\boldsymbol{X}$ in time. This shifted ensemble becomes the forecast, and the whole cycle runs all over again.

It is possible to represent the covariance matrix $\hat{C}_f$ in spectral space [1]. Under the assumption of covariance stationarity, the spectral covariance matrix is diagonal with variances of the coefficients of the expansion of the state in the spectral basis. Filtering methods that take advantage of this result and perform the whole data assimilation process in spectral space, using only the diagonal of spectral sample covariance matrix for $\hat{C}_f$, were studied in [4]. Under the normality assumption, this corresponds to improving the spectral sample covariance by using the maximum likelihood estimator $\hat{D}^{(n)}$ from Subsection 2.1. The question is, whether the filter will perform better when using even more precise estimators like $\hat{D}^{(3)}$ and $\hat{D}^{(2)}$. The improvement can be achieved by searching for the MLE in a correct subspace (or close to it). However, based on climatological data, the power model $\hat{D}^{(2)}$ seems to be reasonable [3].

The critical point of every filtering method is its long-time behaviour and stability, especially for a small ensemble. In the case of traditional EnKF (given by equation (4) and (5)), the filter error does not grow faster than exponentially for a fixed ensemble size [5]. The question is, whether the upper bound for the filter error can be improved by using an improved covariance estimator that comes from the right subspace. This is the subject of our current research. The following simulation suggests that the answer may be positive.

The simulation setting was as follows. First, an initial forecast ensemble of size $N = 5$ and the initial true system state were generated from $\mathcal{N}_n(\mathbf{0}, C)$ with $n = 100$ and $C = FDF^\top$, where F is a Fourier transform and $D = \mathrm{diag}\{c(-\lambda_i)^{-\alpha}, i = 1, \ldots, n\}$ with $c = 50$ and $\alpha = 1.5$. In each cycle, the observations $\boldsymbol{y}^j = \boldsymbol{X}_t + \boldsymbol{\xi}^j$ were generated with $\boldsymbol{\xi}^j \sim \mathcal{N}_n(\mathbf{0}, R)$ and $R = 0.0064 \cdot \mathbb{I}$ and then assimilated with the forecast ensemble. The analysis part was done in the spectral space, following [4], where the theoretical covariance matrix D_f is assumed to be diagonal. After the assimilation part, the analysis ensemble

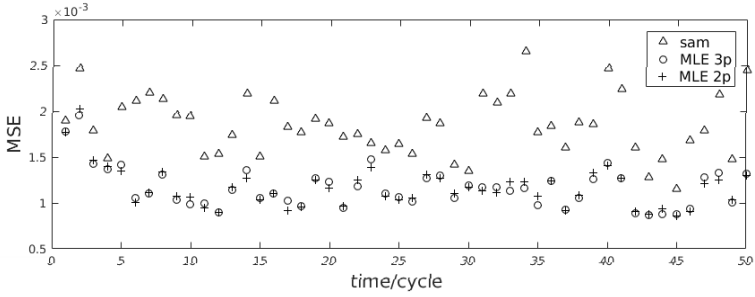


Figure 1: Mean square errors of the analysis ensemble mean.

was propagated in time by the model $\eta(\boldsymbol{X}_a) = A\boldsymbol{X}_a + \boldsymbol{b}$ with $A = 0.9 \cdot \mathbb{I}$ and $\boldsymbol{b} \sim \mathcal{N}_n(\boldsymbol{0}, C)$. The cycle consisting of analysis and propagation step then runs all over again. Three parallel filters were run with distinct estimators of D_f used in the analysis step. The estimators were $\hat{D}_f^{(n)}$ (denoted `sam`), $\hat{D}_f^{(3)}$ (`MLE 3p`) and $\hat{D}_f^{(2)}$ (`MLE 2p`).

In each of 50 cycles, the analysis ensemble was summarized into its mean $\bar{\boldsymbol{X}}_a = \frac{1}{N}\sum_{j=1}^{N} \boldsymbol{X}_a^j$ and the mean square error

$$\frac{1}{n}\sum_{i=1}^{n} \left(\bar{X}_a^j(i) - X_t(i)\right)^2$$

was plotted for every cycle. We denoted by $\bar{X}_a^j(i)$ the entries of $\bar{\boldsymbol{X}}_a^j$. As we can see at Fig. 1, the analysis that uses the more precise covariance estimator is closer to the true state vector (in terms of MSE). However, the performance of the analysis mean is not the only criterion. The stability of the analysis covariance C_a is also important. In Fig. 2, we can see a comparison of spectral representations of four matrices. The true filtering covariance descents from the original covariance C by propagation in time and by assimilation using the expression (5) (where $\hat{C}_f$ is substituted by the matrix C_f resulting from the time-propagation step). The other three matrices are distinct estimates of D_f based on the analysis ensemble after the last cycle. The estimate based on sample covariance is very rough. The MLEs follow the proper trend and provide stable estimates.

This short simulation indicates that the error of the EnKF is smaller when a better covariance estimate is used while the analysis covariance is stable. The theoretical background of this effect is a subject of further research.

Acknowledgements: This work was supported by the grant SVV 2017 No. 260454 and by the U.S. National Science Foundation under grant DMS-1216481.

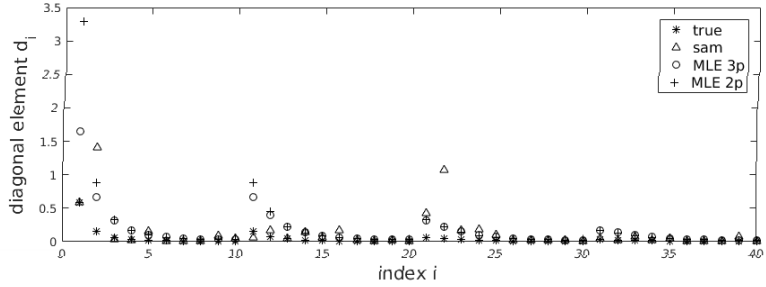


Figure 2: Spectral representations of the true filtering covariance and the analysis covariance matrices (the first 40 elements).

References

[1] P. Courtier, E. Andersson, W. Heckley, D. Vasiljevic, M. Hamrud, A. Hollingsworth, F. Rabier, M. Fisher, and J. Pailleux. The ECMWF implementation of three-dimensional variational assimilation (3D-Var). I: Formulation. *Quarterly Journal of the Royal Meteorological Society*, 124(550):1783–1807, 1998.

[2] G. Evensen. Sequential data assimilation with nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research*, 99 (C5)(10):143–162, 1994.

[3] G. Gaspari, S. E. Cohn, J. Guo, and S. Pawson. Construction and application of covariance functions with variable length-fields. *Quarterly Journal of the Royal Meteorological Society*, 132(619):1815–1838, 2006.

[4] I. Kasanický, J. Mandel, and M. Vejmelka. Spectral diagonal ensemble Kalman filters. *Nonlinear Processes in Geophysics*, 22(4):485 – 497, 2015.

[5] D. T. B. Kelly, K. J. H. Law, and A. M. Stuart. Well-posedness and accuracy of the ensemble Kalman filter in discrete and continuous time. *Nonlinearity*, 27(10):2579–2603, 2014.

[6] E. L. Lehmann and G. Casella. *Theory of point estimation*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 1998.

[7] M. Turčičová, J. Mandel, and K. Eben. Multilevel maximum likelihood estimation with application to covariance matrices. Technical report. Submitted in January 2017.

Methods for bandwidth detection in kernel conditional density estimations

Kateřina Konečná*[1] and Ivana Horová[2]

[1,2]*Department of Mathematics and Statistics, Masaryk University, Brno, Czech Republic.*

This contribution is focused on the kernel conditional density estimations (KCDE). The estimation depends on the smoothing parameters which influence the final density estimation significantly. This is the reason why a requirement of any data-driven method is needed for bandwidth estimation. In this contribution, the cross-validation method, the iterative method and the maximum likelihood approach are conducted for bandwidth selection of the estimator. An application on a real data set is included and the proposed methods are compared.

Keywords: kernel conditional density estimation, bandwidth detection, cross-validation method, iterative method, maximum likelihood method

Introduction

Kernel smoothing techniques belong to the most popular non-parametric techniques for data interpolation, especially for its simple usage and no strictly limiting requirements. Conditional density estimations offer the comprehensive information about the data structure – regression models only the conditional expectation while conditional density includes even the variability and the whole data distribution.

The estimator depends on the unknown parameters, called the smoothing parameters or bandwidths. They influence the quality of the estimation significantly, this is the reason why so much attention is given to the bandwidth determination. The optimal values of the smoothing parameters depend on the unknown conditional and marginal density, thus there is a necessity to develop an automatic data-driven bandwidth selectors. In this contribution, the widely used cross-validation method is supplemented with the iterative method and the leave-one-out maximum likelihood method.

*Corresponding author: xkonecn3@math.muni.cz

1 Statistical properties of the Nadaraya-Watson estimator of conditional density

The basic building block of kernel smoothing is a kernel function, which plays a role of weighting function. Let K be a real valued function satisfying

1. $K \in \mathrm{Lip}[-1,1]$, i. e. $|K(x) - K(y)| \leq L|x-y|$, $\forall x, y \in [-1,1]$, $L > 0$,
2. $\mathrm{supp}(K) = [-1,1]$,
3. moment conditions:

$$\int_{-1}^{1} K(x)\,\mathrm{d}x = 1, \int_{-1}^{1} xK(x)\,\mathrm{d}x = 0, \int_{-1}^{1} x^2K(x)\,\mathrm{d}x = \beta_2(K) \neq 0.$$

Such a function K is called a kernel of order 2.
Conditional density models the probability of a random variable Y given a fixed observation $X = x$. The Nadaraya-Watson estimator of conditional density takes the form

$$\hat{f}_{NW}(y|x) = \frac{1}{h_y} \sum_{i=1}^{n} w_i^{NW}(x) K\left(\frac{y - Y_i}{h_y}\right), \tag{1}$$

where $w_i^{NW}(x) = \dfrac{K\left(\frac{x-X_i}{h_x}\right)}{\sum_{j=1}^{n} K\left(\frac{x-X_j}{h_x}\right)}$ is a weight function in the point x, $h_x, h_y > 0$ are the smoothing parameters.

The statistical properties of the estimator are the rudiments for appraisal of suitability of the estimator and determination of the optimal values of bandwidths.
The Asymptotic Bias (AB) and the Asymptotic Variance (AV) of the Nadaraya-Watson estimator are given by Hyndman *et al.* ([4]) with the expressions

$$\mathrm{AB}\left\{\hat{f}_{NW}(y|x)\right\} = \frac{1}{2}h_x^2\beta_2(K)\left[2\frac{g'(x)}{g(x)} + \frac{\partial^2 f(y|x)}{\partial x^2}\right] + \frac{1}{2}h_y^2\beta_2(K)\frac{\partial^2 f(y|x)}{\partial y^2},$$

$$\mathrm{AV}\left\{\hat{f}_{NW}(y|x)\right\} = \frac{R^2(K) f(y|x)}{n h_x h_y g(x)},$$

where $R(K) = \int K^2(t)\,\mathrm{d}t$, $g(x)$ is a marginal density of a random variable X. The global quality of the estimate is measured by the Mean Integrated Squared Error (MISE) in the form

$$\mathrm{MISE}\left\{\hat{f}_{NW}(\cdot|\cdot)\right\} = \iint \mathrm{E}\left\{\left(\hat{f}_{NW}(y|x) - f(y|x)\right)^2\right\} g(x)\,\mathrm{d}x\,\mathrm{d}y.$$

The main term of MISE $\left\{\hat{f}_{NW}(\cdot|\cdot)\right\}$, the Asymptotic Mean Integrated Squared Error (AMISE), is of the form

$$\text{AMISE}\left\{\hat{f}_{NW}(\cdot|\cdot)\right\} = \frac{c_1}{nh_xh_y} + c_2h_x^4 + c_3h_y^4 + c_4h_x^2h_y^2,$$

where

$$\begin{aligned}
c_1 &= \int R^2(K)\,\mathrm{d}x,\\
c_2 &= \iint \frac{\beta_2^2(K)}{4}\left(2\frac{g'(x)}{g(x)}\frac{\partial f(y|x)}{\partial x} + \frac{\partial^2 f(y|x)}{\partial x^2}\right)^2 g(x)\,\mathrm{d}y\,\mathrm{d}x,\\
c_3 &= \iint \frac{\beta_2^2(K)}{4}\left(\frac{\partial^2 f(y|x)}{\partial y^2}\right)^2 g(x)\,\mathrm{d}y\,\mathrm{d}x,\\
c_4 &= \iint \frac{\beta_2^2(K)}{2}\left(2\frac{g'(x)}{g(x)}\frac{\partial f(y|x)}{\partial x} + \frac{\partial^2 f(y|x)}{\partial x^2}\right)\left(\frac{\partial^2 f(y|x)}{\partial y^2}\right) g(x)\,\mathrm{d}y\,\mathrm{d}x.
\end{aligned}$$

The optimal bandwidths $\left(h_x^*, h_y^*\right)$ minimize AMISE

$$(h_x^*, h_y^*) = \underset{(h_x,h_y)}{\arg\min}\,\text{AMISE}\left\{\hat{f}_{NW}(\cdot|\cdot)\right\},$$

where the nonequations $an^{-1/6} \le h_x \le bn^{-1/6}$ and $cn^{-1/6} \le h_y \le dn^{-1/6}$ are held for $0 < a < b < \infty$ and $0 < c < d < \infty$. The optimal values of smoothing parameters are derived by differentiating of AMISE, setting the derivatives to 0 and making several algebraic simplifications. They are given by Hyndman *et al.* in the paper [4] as follows

$$\begin{aligned}
h_x^* &= n^{-1/6}c_1^{1/6}\left[4\left(\frac{c_3^5}{c_4}\right)^{1/4} + 2c_5\left(\frac{c_3}{c_4}\right)^{3/4}\right]^{-1/6},\\
h_y^* &= h_x^*\left(\frac{c_3}{c_4}\right)^{1/4} = n^{-1/6}c_1^{1/6}\left[4\left(\frac{c_4^5}{c_3}\right)^{1/4} + 2c_5\left(\frac{c_4}{c_3}\right)^{3/4}\right]^{-1/6}.
\end{aligned}$$

2 Methods for bandwidth detection

The optimal values of the smoothing parameters depend on the unknown conditional and marginal density. This is the reason why any data-driven method for the estimation of them is needed.

One of the most common methods for choosing the bandwidths is the **cross-validation method** introduced by Fan and Yim [2] and Hansen [3]. The idea

of the method consists in minimization of the proper estimate of the Integrate Squared Error (ISE) represented by the cross-validation function

$$CV\left(h_x, h_y\right) = \frac{1}{n}\sum_{i=1}^{n}\int \hat{f}_{-i,NW}\left(y|X_i\right)^2 \,\mathrm{d}y - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{-i,NW}\left(Y_i|X_i\right),$$

where $\hat{f}_{-i,NW}\left(y|x\right)$ is the estimate in the pair of points (X_i, Y_i) using the points $\{(X_j, Y_j), j \neq i\}$. Thus, the estimates of bandwidths are given by

$$(\hat{h}_x^{CV}, \hat{h}_y^{CV}) = \underset{(h_x,h_y)}{\arg\min}\, CV(h_x, h_y).$$

The next proposed method is the **iterative method** suggested by Konečná and Horová ([5]). The method is based on a suitable estimation of AMISE which can be expressed by a sum of the Asymptotic Integrated Variance (AIV) and the Asymptotic Integrated Squared Bias (AISB). The relation (2) is derived by differentiating of AMISE, setting the derivatives to 0, and by replacing the terms by their estimations:

$$\mathrm{AIV}\left\{\hat{f}(\cdot|\cdot)\right\} - 2\widehat{\mathrm{ISB}}\left\{\hat{f}(\cdot|\cdot)\right\} = 0. \tag{2}$$

The term $\widehat{\mathrm{ISB}}\left\{\hat{f}(\cdot|\cdot)\right\}$ is an approximation of the $\mathrm{AISB}\left\{\hat{f}(\cdot|\cdot)\right\}$ term and it is of the form

$$\widehat{\mathrm{ISB}}\left\{\hat{f}(\cdot|\cdot)\right\} = \iint \left(\widehat{\mathrm{bias}}\left\{\hat{f}(y|x)\right\}\right)^2 g(x)\,\mathrm{d}x\,\mathrm{d}y$$
$$= \iint \left(\frac{\sum\limits_i K_{h_x\sqrt{2}}\left(x - X_i\right) K_{h_y\sqrt{2}}\left(y - Y_i\right)}{\sum\limits_i K_{h_x\sqrt{2}}\left(x - X_i\right)} - \hat{f}_{NW}(y|x)\right)^2 g(x)\,\mathrm{d}x\,\mathrm{d}y.$$

The supplemented equation $\hat{h}_y = \hat{c}\hat{h}_x$ to the equation (2) is represented by a relation $\hat{c}$ between the values of the smoothing parameters, $\hat{c}$ is given by the reference rule suggested by Bashtannyk and Hyndman in the paper [1]. The estimations of the smoothing parameters are derived as a solution of the system of two nonlinear equations (2) and the equation $\hat{h}_y = \hat{c}\hat{h}_x$.

The last suggested method is the **leave-one-out maximum likelihood method** which proceeds with the maximum likelihood method, a statistical standard procedure for estimating the unknown parameters. We consider a random vector $(\mathbf{X}, \mathbf{Y})$ with the independent and identically distributed

observations $(X_i, Y_i), i = 1, \ldots, n$ of the unknown conditional density. We define the modified likelihood function

$$\mathcal{L}(h_x, h_y \mid \mathbf{X}, \mathbf{Y}) = \prod_{j=1}^{n} \hat{f}_{-j,NW}(Y_j \mid X_j).$$

The modification of the classical likelihood approach consists in leaving one observation out. The estimations of the smoothing parameters are given by maximization of $\mathcal{L}(h_x, h_y \mid \mathbf{X}, \mathbf{Y})$, i.e.

$$(\hat{h}_x^{\mathcal{L}}, \hat{h}_y^{\mathcal{L}}) = \underset{(h_x, h_y)}{\arg\min}\, \mathcal{L}(h_x, h_y \mid \mathbf{X}, \mathbf{Y}).$$

3 Application on a real data

For comparison of the proposed methods, the `airquality` data from the `datasets` package in R ([6]) are concerned. The data describe daily air quality in New York, May to September 1973. The estimation of mean ozone concentration in parts per billion, given the maximum daily temperature in degrees Fahrenheit is focused on. There is 153 observations in total, in fact, we include only 116 observation because of some missing values.
The cross-validation method (CV), the iterative method (IT) and the leave-one-out maximum likelihood (ML) are used for bandwidth detection. The values of estimated bandwidths and the computational times are given in the Table 1.

Table 1: Estimates of the smoothing parameters and computational times for methods used for bandwidth determination.

method	$\hat{h}_x$	$\hat{h}_y$	computational time $[s]$
CV	1.845	7.638	182
IT	6.289	15.276	67.6
ML	2.517	10.017	31.3

As it can be seen, the CV method gives the most undersmoothed estimation due to small values of the smoothing parameters, whereas the IT method gives the most oversmoothed estimation. It seems that the ML gives the best results, supported by the shortest computational time. The IT method is the fastest – it takes about 30 seconds, the computational difficulty of the other two methods is evident. The IT method takes less than one third while the

ML method takes even one sixth of the CV's computational time.
It is important to emphasize that these results are valid for this real-data application. Several simulation studies should be executed for the proper assessment of the proposed methods, although this exceeds the extent of this contribution.

4 Conclusion

In this contribution, the methods for bandwidth determination were focused on. The classical approach for bandwidth detection, the cross-validation method, was supplemented with two suggested methods – the iterative and the leave-one-out maximum likelihood method.
These approaches could be extended to the other types of kernel conditional density estimations which have not been mentioned in this contribution. Future work could also involve variable bandwidths, on the other hand, their theoretical aspect as well as computational implementation would be quite difficult.

Acknowledgements: The research was supported by the Czech Science Foundation no. GA15-06991S.

References

[1] D. M. Bashtannyk and R. J. Hyndman. Bandwidth selection for kernel conditional density estimation. *Computational Statistics & Data Analysis*, 36(3):279 – 298, 2001.

[2] J. Fan and T. H. Yim. A crossvalidation method for estimating conditional densities. *Biometrika*, 91(4):819–834, 2004.

[3] B. E. Hansen. Nonparametric conditional density estimation. *Unpublished manuscript*, 2004.

[4] R. J. Hyndman, D. M. Bashtannyk, and G. K. Grunwald. Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statistics*, 5(4):315–336, 1996.

[5] K. Konečná and I. Horová. *Conditional Density Estimations*, pages 15–31. International Society for the Advancement of Science and Technology (ISAST), Athens, 2014.

[6] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

Inference on covariance matrices and operators using concentration inequalities

Adam B Kashlak*[1]

[1] *Cambridge Centre for Analysis, University of Cambridge, UK*

In the modern era of high and infinite dimensional data, classical statistical methodology is often rendered inefficient and ineffective when confronted with such big data problems as arise in genomics, medical imaging, speech analysis, and many other areas of research. Many problems manifest when the practitioner is required to take into account the covariance structure of the data during his or her analysis, which takes on the form of either a high dimensional low rank matrix or a finite dimensional representation of an infinite dimensional operator acting on some underlying function space. Thus, we propose using tools from the concentration of measure literature to construct rigorous descriptive and inferential statistical methodology for covariance matrices and operators. A variety of concentration inequalities are considered, which allow for the construction of nonasymptotic dimension-free confidence sets for the unknown matrices and operators. Given such confidence sets a wide range of estimation and inferential procedures can be and are subsequently developed.

Keywords: Sparse Matrix, Functional Data Analysis, Log Sobolev Inequality, Talagrand's Inequality, Confidence Sets

1 Overview

Concentration inequalities are a general category of results from geometry, functional analysis, and probability theory that control the tail behaviour of probability measures. In recent years, they have proved invaluable to statisticians due to their non-asymptotic dimension-free properties, which makes them particularly suitable for estimation and inference on finite samples of data living high or infinite dimensional space. Overviews of such results can be found in the monographs [3, 8, 11]. This manuscript introduces some of the author's doctoral research into using concentration inequalities for statistical estimation and inference on covariance matrices and operators.

*Corresponding author: ak852@cam.ac.uk or kashlak@ualberta.ca

1.1 Definitions and notation

Definition 1 (Empirical Covariance Matrix)**.** Let $X_1, \ldots, X_n \in \mathbb{R}^d$ be iid realizations of some random variable $X \in \mathbb{R}^d$ with unknown covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$. Then, the sample or empirical estimate for Σ is

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^{\mathrm{T}}$$

where $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$ is the sample mean of the data.

Definition 2 (Empirical Covariance Operator)**.** For $I \subseteq \mathbb{R}$, let $f_1, \ldots, f_n \in L^2(I)$ be iid realizations of some random function $f \in L^2(I)$ with unknown covariance operator $\Sigma \in Op(L^2)$. Then, the sample or empirical estimate for Σ is

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (f_i - \bar{f}) \otimes (f_i - \bar{f}) = \frac{1}{n} \sum_{i=1}^{n} (f_i - \bar{f})^{\otimes 2} = \frac{1}{n} \sum_{i=1}^{n} \big\langle (f_i - \bar{f}), \cdot \big\rangle \, (f_i - \bar{f})$$

where $\bar{f} = n^{-1} \sum_{i=1}^{n} f_i$ is the sample mean of the data.

Definition 3 (p-Schatten norm for matrices)**.** For an arbitrary matrix $\Sigma \in \mathbb{R}^{k \times l}$ and $p \in (1, \infty)$, the p-Schatten norm is

$$\|\Sigma\|_p^p = \mathrm{tr}\left((\Sigma^{\mathrm{T}} \Sigma)^{p/2} \right) = \|\boldsymbol{\nu}\|_{\ell^p}^p = \sum_{i=1}^{\min\{k,l\}} \nu_i^p$$

where $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_{\min\{k,l\}})$ is the vector of singular values of Σ and where $\|\cdot\|_{\ell^p}$ is the standard ℓ^p norm in $\mathbb{R}^d$. In the covariance matrix case where $\Sigma \in \mathbb{R}^{d \times d}$ is symmetric and positive definite, $\|\Sigma\|_p^p = \mathrm{tr}\,(\Sigma^p) = \|\boldsymbol{\lambda}\|_{\ell^p}^p$ where $\boldsymbol{\lambda}$ is the vector of eigenvalues of Σ.

When $p = \infty$, we have the standard operator norm on Euclidean space

$$\|\Sigma\|_\infty = \sup_{v \in \mathbb{R}^d, \|v\|_{\ell^2} = 1} \|\Sigma v\|_{\ell^2} = \sup_{v \in \mathbb{R}^d, \|v\|_{\ell^2} = 1} v^{\mathrm{T}} \Sigma v.$$

For covariance matrices, this coincides with the maximal eigenvalue of Σ.

Definition 4 (p-Schatten norm for operators)**.** Given two separable Hilbert spaces H_1 and H_2, a bounded linear operator $\Sigma : H_1 \to H_2$, and some $p \in [1, \infty)$, then the p-Schatten norm is $\|\Sigma\|_p^p = \mathrm{tr}\left((\Sigma^* \Sigma)^{p/2} \right)$. For $p = \infty$, the Schatten norm is the operator norm: $\|\Sigma\|_\infty = \sup_{f \in H_1} (\|\Sigma f\|_{H_2} / \|f\|_{H_1})$. In the case that Σ is compact, self-adjoint, and trace-class, then given the associated

eigenvalues $\{\lambda_i\}_{i=1}^{\infty}$, the p-Schatten norm coincides with the standard ℓ^p norm of the eigenvalues:

$$\|\Sigma\|_p^p = \begin{cases} \|\lambda\|_{\ell^p}^p = \sum_{i=1}^{\infty} |\lambda_i|^p, & p \in [1, \infty) \\ \max_{i \in \mathbb{N}} |\lambda_i|, & p = \infty \end{cases}$$

2 Covariance matrices

Given $X_1, \ldots, X_n \in \mathbb{R}^d$, past studies have shown that the empirical estimate for the covariance matrix, Definition 1, is a very poor estimator when the underlying true Σ is high dimensional, $d \gg n$, and sparse meaning that most of the off-diagonal entries are zero or negligible. Hence, much research has gone into better estimation techniques [1, 2, 5, 15, 14]. In [10], we propose using concentration inequalities to construct a non-asymptotic confidence set for the empirical estimate and then search the confidence set in order to find an improved estimator.

Let $d(\cdot, \cdot)$ be some metric measuring the distance between two covariance matrices, and let $\psi : \mathbb{R} \to \mathbb{R}$ be monotonically increasing. Then, the general form of the concentration inequalities is

$$\mathrm{P}\left(d(\Sigma_0, \hat{\Sigma}^{\mathrm{emp}}) \geq \mathrm{E}d(\Sigma_0, \hat{\Sigma}^{\mathrm{emp}}) + r\right) \leq \mathrm{e}^{-\psi(r)},$$

which is a bound on the tail of the distribution of $d(\Sigma_0, \hat{\Sigma}^{\mathrm{emp}})$ as it deviates above its mean. Thus, to construct a $(1-\alpha)$-confidence set, the variable $r = r_\alpha$ is chosen such that $\exp(-\psi(r_\alpha)) = \alpha$. Then, choose a $\hat{\Sigma}^{\mathrm{sp}}$ such that $d(\hat{\Sigma}^{\mathrm{sp}}, \hat{\Sigma}^{\mathrm{emp}}) \leq r_\alpha$.

The proposed search procedure is to sequentially set to zero the smallest entries in $\hat{\Sigma}^{\mathrm{emp}}$ while remaining inside the r_α-ball. The metric used is $d(\Sigma_0, \hat{\Sigma}^{\mathrm{emp}}) = \|\Sigma_0 - \hat{\Sigma}^{\mathrm{emp}}\|_p^{1/2}$ where $\|\cdot\|_p$ is the p-Schatten norm from Definition 3. This metric is shown to be Lipschitz $n^{-1/2}$ with respect to Euclidean distance in $\mathbb{R}^{d \times n}$.

In [10], three types of distributional assumptions are considered: log concave measures; sub-exponential measures; bounded random variables. In summary, applying our methodology to log concave measures, which include the multivariate Gaussian distribution, yielded excellent theoretical and experimental results. Our method is particularly good at support recovery or "sparsistency" in this case. For sub-exponential measures, the concentration inequalities do not yield nice theoretical results, but the methodology still gives good performance in simulation studies. This approach fails in the bounded random variable case as the resulting confidence sets are not dimension-free.

3 Covariance operators

In the functional data setting, $f_1, \ldots, f_n \in L^2(I)$ are iid random functions with $I \subseteq \mathbb{R}$. Similarly to the high dimensional case, covariance operators are of critical importance to inference and hypothesis testing. For example, the development of k-sample tests for the equality of covariance is a major area of research [4, 7, 12, 13].

In [9], we propose our own k-sample test for the equality of covariance by first using Talagrand's concentration inequality [16] in the Banach space setting to construct confidence sets for each of the covariance operator. For some desired p-Schatten norm, Definition 4, $\|\cdot\|_p$, with $p \in [1, \infty)$ and with conjugate $q = p/(p-1)$, we require the following terms, which correspond to the distance between the empirical covariance estimate and the true covariance operator and a weak variance term for this random variable:

$$Z = \left\| \frac{1}{n} \sum_{i=1}^n f_i \otimes f_i - \mathrm{E}\left(f_i \otimes f_i\right) \right\|_p = \left\| \hat{\Sigma} - \Sigma \right\|_p$$

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n \sup_{\|\Pi\|_q \leq 1} \mathrm{E} \left\langle f_i^{\otimes 2} - \mathrm{E} f_i^{\otimes 2}, \Pi \right\rangle^2.$$

In the above equation, the supremum is to be taken over a countably dense subset of the unit ball of $\Pi \in Op(L^2)$. For some $U \geq \|f_i^{\otimes 2}\|_{L^2}^2$ and $v_n = 2U\mathrm{E}Z + n\sigma^2$, the initial level $(1-\alpha)$ confidence set constructed is

$$C_{n,1-\alpha} = \left\{ \Sigma : Z \leq \mathrm{E}Z + \sqrt{-2v_n \log(2\alpha)/n} - U\log(2\alpha)/(3n) \right\}.$$

To make this confidence set usable on real data, the Rademacher average $R_n = n^{-1} \sum_{i=1}^n \varepsilon_i ((f_i - \bar{f})^{\otimes 2} - \hat{\Sigma})$, where $\mathrm{P}(\varepsilon_i = 1) = \mathrm{P}(\varepsilon_i = -1) = 0.5$ will be used as a proxy for the unknown $\mathrm{E}Z$.

In [9], this is not only applied to k-sample tests for equality of covariance, but also to the classification and clustering of functional data. This methodology is applied to a set of phoneme data detailed in [6], which is a collection of 400 log-periodograms for each of five different phonemes: /ɑ/ as in the vowel of "dark"; /ɔ/ as in the first vowel of "water"; /d/ as in the plosive of "dark"; /i/ as in the vowel of "she"; /ʃ/ as in the fricative of "she". Each curve contains the first 150 frequencies from a 32 ms sound clip sampled at a rate of 16-kHz. Comparisons of our concentration-based methodology with other methods of classification and clustering can be found in Tables 1 and 2, respectively.

	/ɑ/	/ɔ/	/d/	/i/	/ʃ/
CoM	76.9	76.8	96.6	**98.5**	99.4
KNN	72.4	79.1	**98.5**	97.4	**100.**
Kernel	72.0	**80.5**	98.4	97.2	99.9
GLM	**79.0**	72.3	98.2	95.9	99.2
Tree	70.8	69.4	95.6	87.8	92.6

Table 1: Percentage of correct classification of the five phonemes against the five methods: our concentration of measure approach (CoM); k-nearest-neighbours (KNN); kernel method (Kernel); generalized linear model (GLM); and regression trees (Tree).

	Concentration					**k-means**				
Cluster	A	B	C	D	E	A	B	C	D	E
/ɑ/	281	119	0	0	0	281	119	0	0	0
/ɔ/	125	273	1	1	0	126	272	1	1	0
/d/	0	0	384	15	1	0	2	386	10	2
/i/	1	0	1	393	5	1	3	2	381	13
/ʃ/	0	0	0	3	397	0	0	0	2	398

Table 2: Clustering 2000 phoneme curves into 5 clusters. Similar results achieved by both the concentration and k-means methods.

Acknowledgements: The author would like to acknowledge the support of his thesis advisers Professors John AD Aston and Richard Nickl from the University of Cambridge Statistical Laboratory. He would also like to thank Professor Linglong Kong at the University of Alberta for his collaboration on the research contained in Section 2.

References

[1] P. J. Bickel and E. Levina. Covariance regularization by thresholding. *The Annals of Statistics*, pages 2577–2604, 2008.

[2] J. Bien and R. J. Tibshirani. Sparse estimation of a covariance matrix. *Biometrika*, 98(4):807–820, 2011.

[3] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

[4] A. Cabassi, D. Pigoli, P. Secchi, and P. A. Carter. Permutation tests for the equality of covariance operators of functional data with applications to evolutionary biology. *arXiv preprint arXiv:1701.05870*, 2017.

[5] T. Cai and W. Liu. Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106(494):672–684, 2011.

[6] F. Ferraty and P. Vieu. Curves discrimination: a nonparametric functional approach. *Computational Statistics & Data Analysis*, 44(1):161–173, 2003.

[7] S. Fremdt, J. G. Steinebach, L. Horváth, and P. Kokoszka. Testing the equality of covariance operators in functional samples. *Scandinavian Journal of Statistics*, 40(1):138–152, 2013.

[8] E. Giné and R. Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press, 2016.

[9] A. B. Kashlak, J. A. D. Aston, and R. Nickl. Inference on covariance operators via concentration inequalities: k-sample tests, classification, and clustering via Rademacher complexities. *arXiv preprint arXiv:1604.06310*, 2016.

[10] A. B. Kashlak and L. Kong. A concentration inequality based methodology for sparse covariance estimation. *arXiv preprint arXiv:1705.02679*, 2017.

[11] M. Ledoux. *The concentration of measure phenomenon*, volume 89. American Mathematical Soc., 2001.

[12] V. M. Panaretos, D. Kraus, and J. H. Maddocks. Second-order comparison of Gaussian random functions and the geometry of DNA minicircles. *Journal of the American Statistical Association*, 105(490):670–682, 2010.

[13] D. Pigoli, J. A. Aston, I. L. Dryden, and P. Secchi. Distances and inference for covariance operators. *Biometrika*, page asu008, 2014.

[14] A. J. Rothman. Positive definite estimators of large covariance matrices. *Biometrika*, 99(3):733–740, 2012.

[15] A. J. Rothman, E. Levina, and J. Zhu. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186, 2009.

[16] M. Talagrand. New concentration inequalities in product spaces. *Inventiones mathematicae*, 126(3):505–563, 1996.

Abstracts of Remaining Talks

Simulating and Forecasting Human Population with General Branching Process

Plamen Trayanov*[1]

[1]*Sofia Univeristy "St. Kliment Ohridski"*

The branching process theory is widely used to describe a population dynamics in which particles live and produce other particles through their life, according to given stochastic birth and death laws. The theory of General Branching Processes (GBP) presents a continuous time model in which every woman has random life length and gives birth to children in random intervals of time. The flexibility of the GBP makes it very useful for modelling and forecasting human population. This paper is a continuation of previous developments in the theory, necessary to model the specifics of human population, and presents their application in forecasting the population age structure of Bulgaria. It also introduces confidence intervals of the forecasts, calculated by GBP simulations, which reflect both the stochastic nature of the birth and death laws and the branching process itself. The simulations are also used to determine the main sources of risk to the forecast.

*Corresponding author: plamentrayanov@gmail.com

Fréchet means and Procrustes analysis in Wasserstein space

Yoav Zemel*[1] and Victor Panaretos[1]

[1]*Ecole polytechnique fédérale de Lausanne*

We consider three interlinked problems in stochastic geometry: (1) constructing optimal multicouplings of random vectors; (2) determining the Fréchet mean of probability measures in Wasserstein space; and (3) registering collections of randomly deformed spatial point processes. We demonstrate how these problems are canonically interpreted through the prism of the theory of optimal transportation of measure on $\mathbb{R}^d$. We provide explicit solutions in the one dimensional case, consistently solve the registration problem and establish convergence rates and a (tangent space) central limit theorem for Cox processes. When $d > 1$, the solutions are no longer explicit and we propose a steepest descent algorithm for deducing the Fréchet mean in problem (2). Supplemented by uniform convergence results for the optimal maps, this furnishes a solution to the multicoupling problem (1). The latter is then utilised, as in the case $d = 1$, in order to construct consistent estimators for the registration problem (3). While the consistency results parallel their one-dimensional counterparts, their derivation requires more sophisticated techniques from convex analysis. This is joint work with Victor M. Panaretos

*Corresponding author: yoav.zemel@epfl.ch

Predict extreme influenza epidemics

Maud Thomas[*1] **and Holger Rootzén**[1]

[1] *Chalmers University of Technology*
[2] *Université Pierre et Marie Curie*

Influenza viruses are responsible for annual epidemics, causing more than 500,000 deaths per year worldwide. A crucial question for resource planning in public health is to predict the morbidity burden of extreme epidemics. We say that an epidemic is extreme whenever the influenza incidence rate exceeds a high threshold for at least one week. Our objective is to predict whether an extreme epidemic will occur in the near future, say the next couple of weeks.

The weekly numbers of influenza-like illness (ILI) incidence rates in France are available from the Sentinel network for the period 1991-2017. ILI incidence rates exhibit two different regimes, an epidemic regime during winter and a non-epidemic regime during the rest of the year. To identify epidemic periods, we use a two-state autoregressive hidden Markov model.

A main goal of Extreme Value Theory is to assess, from a series of observations, the probability of events that are more extreme than those previously recorded. Because of the autoregressive structure of the data, we choose to fit one of the mul- tivariate generalized Pareto distribution models proposed in Rootzén et al. (2016a) [Multivariate peaks over threshold models. arXiv:1603.06619v2]; see also Rootzén et al. (2016b) [Peaks over thresholds modeling with multivariate generalized Pareto distributions. arXiv:1612.01773v1]. For these models, explicit densities are given, and formulas for conditional probabilities can then be deduced, from which we can predict if an epidemic will be extreme, given the first weeks of observation.

[*]Corresponding author: maud.thomas@upmc.fr

Controlled branching processes in Biology: a model for cell proliferation

Carmen Minuesa Abril*[1], Miguel González Velasco[1], and Inés María del Puerto García[1]

[1] *University of Extremadura*

Branching processes are relevant models in the development of theoretical approaches to problems in applied fields such as, for instance, growth and extinction of populations, biology, epidemiology, cell proliferation kinetics, genetics and algorithm and data structures. The most basic model, the so-called Bienaymé-Galton-Watson process, consists of individuals that reproduce independently of the others following the same probability distribution, known as offspring distribution. A natural generalization is to incorporate a random control function which determines the number of progenitors in each generation. The resulting process is called controlled branching process.

In this talk, we deal with a problem arising in cell biology. More specifically, we focus our attention on experimental data generated by time-lapse video recording of cultured in vitro oligodendrocyte cells. In A.Y. Yakovlev et al. (2008) (Branching Processes as Models of Progenitor Cell Populations and Estimation of the Offspring Distributions, *Journal of the American Statistical Association*, 103(484):1357–1366), a two-type age dependent branching process with emigration is considered to describe the kinetics of cell populations. The two types of cells considered are referred as type T_1 (immediate precursors of oligodendrocytes) and type T_2 (terminally differentiated oligodendrocytes). The reproduction process of these cells is as follows: when stimulating to divide under in vitro conditions, the progenitor cells are capable of producing either their direct progeny (two daughter cells of the same type) or a single, terminally differentiated nondividing oligodendrocyte. Moreover, censoring effects as a consequence of the migration of progenitor cells out of the microscopic field of observation are modelled as the process of emigration of the type T_1 cells.

In this work, we propose a two-type controlled branching process to describe the embedded discrete branching structure of the age-dependent branching process aforementioned. We address the estimation of the offspring distribution of the cell population in a Bayesian outlook by making use of disparities. The importance of this problem yields in the fact that the behaviour of

*Corresponding author: cminuesaa@unex.es

these populations is strongly related to the main parameters of the offspring distribution and in practice, these values are unknown and their estimation is necessary. The proposed methodology introduced in M. González et al. (2017) (Robust estimation in controlled branching processes: Bayesian estimators via disparities. *Work in progress*), is illustrated with an application to the real data set given in A.Y. Yakovlev et al. (2008).

Best Unbiased Estimators for Doubly Multivariate Data

Arkadiusz Kozioł*[1], Roman Zmyślony[1], Ricardo Leiva[2], Miguel Fonseca[4], and Anuradha Roy[3]

[1]*Faculty of Mathematics, Computer Science and Econometrics University of Zielona Góra, Szafrana 4a, 65-516 Zielona Góra, Poland*
[2]*Departamento de Matemática F.C.E., Universidad Nacional de Cuyo, 5500 Mendoza, Argentina*
[3]*Department of Management Science and Statistics The University of Texas at San Antonio San Antonio, TX 78249, USA*
[4]*Centro de Matemática e Aplicações Universidade Nova de Lisboa Monte da Caparica, 2829-516 Caparica, Portugal*

The article addresses the best unbiased estimators of the block compound symmetric covariance structure for m-variate observations with equal mean vector over each level of factor or each time point (model with structured mean vector). Under multivariate normality, the free-coordinate approach is used to obtain unbiased linear and quadratic estimates for the model parameters. Optimality of these estimators follows from sufficiency and completeness of their distributions. Additionally, strong consistency is proven. The properties of the estimators in the proposed model are compared with the ones in the model with unstructured mean vector (the mean vector changes over levels of factor or time points).

*Corresponding author: a.koziol@wmie.uz.zgora.pl

Parameter Estimation for Discretely Observed Infinite-Server Queues with Markov-Modulated Input

Mathisca de Gunst[2], Bartek Knapik[2], Michel Mandjes[1], and Birgit Sollie*[1]

[1] *Universiteit van Amsterdam*
[2] *Vrije Universiteit Amsterdam*

The Markov-modulated infinite-server queue is a queueing system with infinitely many servers, where the arrivals follow a Markov-modulated Poisson process (MMPP), i.e. a Poisson process with rate modulating between several values. The modulation is driven by an underlying and unobserved continuous time Markov chain $\{X_t\}_{t\geq 0}$. The inhomogeneous rate of the Poisson process, $\lambda(t)$, stochastically alternates between d different rates, $\lambda_1, \ldots, \lambda_d$, in such a way that $\lambda(t) = \lambda_i$ if $X_t = i$, $i = 1, \ldots, d$.

We are interested in estimating the parameters of the arrival process for this queueing system based on observations of the queue length at discrete times only. We assume exponentially distributed service times with rate μ, where μ is time-independent and known. Estimation of the parameters of the arrival process has not yet been studied for this particular queueing system. Two types of missing data are intrinsic to the model, which complicates the estimation problem. First, the underlying continuous time Markov chain in the Markov-modulated arrival process is not observed. Second, the queue length is only observed at a finite number of discrete time points. As a result, it is not possible to distinguish the number of arrivals and the number of departures between two consecutive observations.

In this talk we show how we derive an explicit algorithm to find maximum likelihood estimates of the parameters of the arrival process, making use of the EM algorithm. Our approach extends the one used in Okamura et al. (2009), where the parameters of an MMPP are estimated based on observations of the process at discrete times. However, in contrast to our setting, Okamura et al. (2009) do not consider departures and therefore do not deal with the second type of missing data. We illustrate the accuracy of the proposed estimation algorithm with a simulation study.

*Corresponding author: b.sollie@vu.nl

Reference: Okamura H., Dohi T., Trivedi K.S. (2009). Markovian Arrival Process Parameter Estimation With Group Data. IEEE/ACM Transactions on Networking. Vol. 17, No. 4, pp. 1326–1339

Modeling of vertical and horizontal variation in multivariate functional data

Niels Olsen[*1]

[1]*Københvans Universitet*

We present a model for multivariate functional data that simultaneously model vertical and horisontal variation. Horisontal variation is modeled using warping functions represented by latent gaussian variables. Vertical variation is modeled using Gaussian processes using a generally applicable low-parametric covariance structure. We devise a method for maximum likelihood estimation using a Laplace approximation and apply it to three different data sets.

[*]Corresponding author: niels.olsen@math.ku.dk

Joint Bayesian nonparametric reconstruction of dynamical equations

Spyridon Hatjispyros[1] **and Christos Merkatas***[1]

[1] *Department of Mathematics, University of the Aegean, Greece*

We propose a Bayesian nonparametric mixture model for the joint full reconstruction of m dynamical equations, given m observed dynamically-noisy-corrupted chaotic time series. The method of reconstruction is based on the Pairwise Dependent Geometric Stick Breaking Processes mixture priors (PDGSBP) first proposed by Hatjispyros et al. (2017). We assume that each set of dynamical equations has a deterministic part with a known functional form i.e.

$$x_{ji} = g_j(\vartheta_j, x_{j,i-1}, \ldots, x_{j,i-l_j}) + \epsilon_{x_{ji}}, \;\; 1 \le j \le m, \;\; 1 \le i \le n_j.$$

under the assumption that the noise processes $(\epsilon_{x_{ji}})$ are independent and identically distributed for all j and i from some unknown zero mean process $f_j(\cdot)$. Additionally, we assume that a-priori we have the knowledge that the processes $(\epsilon_{x_{ji}})$ for $j = 1, \ldots, m$ have common characteristics, e.g. they may have common variances or even have similar tail behavior etc. For a full reconstruction, we would like to jointly estimate the following quantities

$$(\vartheta_j) \in \Theta \subseteq \mathcal{R}^{k_j}, \quad (x_{j,0}, \ldots, x_{j,l_j-1}) \in \mathcal{X}_j \subseteq \mathcal{R}^{l_j},$$

and perform density estimation to the m noise components (f_j).

Our contention is that whenever there is at least one sufficiently large data set, using carefully selected informative borrowing-of-strength-prior-specifications we are able to reconstruct those dynamical processes that are responsible for the generation of time series with small sample sizes; namely sample sizes that are inadequate for an independent reconstruction. We illustrate the joint estimation process for the case $m = 2$, when the two time series are coming from a quadratic and a cubic stochastic process of lag one and the noise processes are zero mean normal mixtures with common components.

*Corresponding author: cmerkatas@aegean.gr

Viterbi process for pairwise Markov models

Joonas Sova[*1]

[1] *University of Tartu*

My talk is based on ongoing joint work with my supervisor Jüri Lember.

We consider a Markov chain $Z = \{Z_k\}_{k\geq 1}$ with product state space $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y}$ is a finite set (state space) and $\mathcal{X}$ is an arbitrary separable metric space (observation space). Thus, the process Z decomposes as $Z = (X, Y)$, where $X = \{X_k\}_{k\geq 1}$ and $Y = \{Y_k\}_{k\geq 1}$ are random processes taking values in $\mathcal{X}$ and $\mathcal{Y}$, respectively. Following citepairwise,pairwise2,pairwise3, we call the process Z a *pairwise Markov model*. The process X is identified as an observation process and the process Y, sometimes called the *regime*, models the observations-driving hidden state sequence. Therefore our general model contains many well-known stochastic models as a special case: hidden Markov models, Markov switching models, hidden Markov models with dependent noise and many more. The *segmentation* or *path estimation* problem consists of estimating the realization of $(Y_1, \ldots, Y_n)$ given a realization $x_{1:n}$ of $(X_1, \ldots, X_n)$. A standard estimate is any path $v_{1:n} \in \mathcal{Y}^n$ having maximum posterior probability:

$$v_{1:n} = \underset{y_{1:n}}{\operatorname{argmax}}\, P(Y_{1:n} = y_{1:n} | X_{1:n} = x_{1:n}).$$

Any such path is called *Viterbi path* and we are interested in the behaviour of $v_{1:n}$ as n grows. The study of asymptotics of Viterbi path is complicated by the fact that adding one more observation, x_{n+1} can change the whole path, and so it is not clear, whether there exists a limiting infinite Viterbi path.

We show that under some conditions the infinite Viterbi path indeed exists for almost every realization $x_{1:\infty}$ of X, thereby defining an infinite Viterbi decoding of X, called the *Viterbi process.* This is done trough construction of *barriers*. A barrier is a fixed-sized block in the observations $x_{1:n}$ that fixes the Viterbi path up to itself: for every continuation of $x_{1:n}$, the Viterbi path up to the barrier remains unchanged. Therefore, if almost every realization of X-process contains infinitely many barriers, then the Viterbi process exists.

Having infinitely many barriers is not necessary for existence of infinite Viterbi path, but the barrier-construction has several advantages. One of them is that it allows to construct the infinite path *piecewise*, meaning that to

*Corresponding author: joonas.sova@ut.ee

determine the first k elements $v_{1:k}$ of the infinite path it suffices to observe $x_{1:n}$ for n big enough. Barrier construction has another great advantage: namely, the process $(Z, V) = \{(Z_k, V_k)\}_{k\geq 1}$, where $V = \{V_k\}_{k\geq 1}$ denotes the Viterbi process, is under certain conditions regenerative. This is can be proven by, roughly speaking, applying the Markov splitting method to construct regeneration times for Z which coincide with the occurrences of barriers. Regenerativity of (Z, V) allows to easily prove limit theorems to understand the asymptotic behaviour of inferences based on Viterbi paths. In fact, in a special case of hidden Markov model this regenerative property has already been known to hold and has found several applications citeAV,AVacta,Vsmoothing,Vrisk, iowa.