Austrian Journal of Statistics

AUSTRIAN STATISTICAL SOCIETY

Volume 46, Number 3–4, 2017

ISSN: 1026597X, Vienna, Austria

Österreichische Zeitschrift für Statistik

ÖSTERREICHISCHE STATISTISCHE GESELLSCHAFT

Austrian Journal of Statistics; Information and Instructions

GENERAL NOTES

The Austrian Journal of Statistics is an open-access journal with a long history and is published approximately quarterly by the Austrian Statistical Society. Its general objective is to promote and extend the use of statistical methods in all kind of theoretical and applied disciplines. Special emphasis is on methods and results in official statistics.

Original papers and review articles in English will be published in the Austrian Journal of Statistics if judged consistently with these general aims. All papers will be refereed. Special topics sections will appear from time to time. Each section will have as a theme a specialized area of statistical application, theory, or methodology. Technical notes or problems for considerations under Shorter Communications are also invited. A special section is reserved for book reviews.

All published manuscripts are available at

http://www.ajs.or.at

(old editions can be found at http://www.stat.tugraz.at/AJS/Editions.html)

Members of the Austrian Statistical Society receive a copy of the Journal free of charge. To apply for a membership, see the website of the Society. Articles will also be made available through the web.

PEER REVIEW PROCESS

All contributions will be anonymously refereed which is also for the authors in order to getting positive feedback and constructive suggestions from other qualified people. Editor and referees must trust that the contribution has not been submitted for publication at the same time at another place. It is fair that the submitting author notifies if an earlier version has already been submitted somewhere before. Manuscripts stay with the publisher and referees. The refereeing and publishing in the Austrian Journal of Statistics is free of charge. The publisher, the Austrian Statistical Society requires a grant of copyright from authors in order to effectively publish and distribute this journal worldwide.

OPEN ACCESS POLICY

This journal provides immediate open access to its content on the principle that making research freely available to the public supports a greater global exchange of knowledge.

ONLINE SUBMISSIONS

Already have a Username/Password for Austrian Journal of Statistics?
Go to http://www.ajs.or.at/index.php/ajs/login

Need a Username/Password?
Go to http://www.ajs.or.at/index.php/ajs/user/register

Registration and login are required to submit items and to check the status of current submissions.

AUTHOR GUIDELINES

The original LaTeX-file guidelinesAJS.zip (available online) should be used as a template for the setting up of a text to be submitted in computer readable form. Other formats are only accepted rarely.

SUBMISSION PREPARATION CHECKLIST

- The submission has not been previously published, nor is it before another journal for consideration (or an explanation has been provided in Comments to the Editor).
- The submission file is preferable in LaTeXfile format provided by the journal.
- All illustrations, figures, and tables are placed within the text at the appropriate points, rather than at the end.
- The text adheres to the stylistic and bibliographic requirements outlined in the Author Guidelines, which is found in About the Journal.

COPYRIGHT NOTICE

Manuscripts should be unpublished and not be under consideration for publication elsewhere. By submitting an article, the author(s) certify that the article is their original work, that they have the right to submit the article for publication, and that they can grant the above license.

Austrian Journal of Statistics

Volume 46, Number 3–4, 2017

Editor-in-chief: Matthias TEMPL

http://www.ajs.or.at

Published by the AUSTRIAN STATISTICAL SOCIETY

http://www.osg.or.at

Österreichische Zeitschrift für Statistik

Jahrgang 46, Heft 3–4, 2017

ÖSTERREICHISCHE STATISTISCHE GESELLSCHAFT

Impressum

Editor:	Matthias Templ, Statistics Austria & Vienna University of Technology
Editorial Board:	Peter Filzmoser, Vienna University of Technology Herwig Friedl, TU Graz Bernd Genser, University of Konstanz Peter Hackl, Vienna University of Economics, Austria Wolfgang Huf, Medical University of Vienna, Center for Medical Physics and Biomedical Engineering Alexander Kowarik, Statistics Austria, Austria Johannes Ledolter, Institute for Statistics and Mathematics, Wirtschaftsuniversität Wien & Management Sciences, University of Iowa Werner Mueller, Johannes Kepler University Linz, Austria Josef Richter, University of Innsbruck Milan Stehlik, Department of Applied Statistics, Johannes Kepler University, Linz, Austria Wolfgang Trutschnig, Department for Mathematics, University of Salzburg Regina Tüchler, Austrian Federal Economic Chamber, Austria Helga Wagner, Johannes Kepler University Walter Zwirner, University of Calgary, Canada
Book Reviews:	Ernst Stadlober, Graz University of Technology
Printed by	Statistics Austria, A-1110 Vienna

Published approximately quarterly by the Austrian Statistical Society, C/o Statistik Austria
Guglgasse 13, A–1110 Wien

Contents

Editorial

The Eleventh International Conference "Computer Data Analysis and Modeling: Theoretical and Applied Stochastics" (CDAM'2016) organized in Minsk by the Belarusian State University and Vienna University of Technology on September 6-10, 2016, was devoted to the topical problems in computer data analysis and modeling. There were 83 presentations by more than 120 participants from 22 countries.

The topics of the presentations corresponded to the following topical scientific problems: robust and nonparametric data analysis; multivariate data analysis; statistical analysis of time series and stochastic processes; probabilistic and statistical analysis of discrete data; statistical signal and image processing; econometric and financial analysis and modeling; survey analysis and official statistics; computer simulation of stochastic systems; computer intensive methods, algorithms and statistical software; computer data analysis and statistical modeling in applications.

This Special Issue contains 12 papers of the extended versions of the most significant presentations selected by the Organizing Committee after a refereeing process.

List of referees

K. Ducinskas, Klaipeda
G. Dzemyda, Vilnius
A. Egorov, Minsk
P. Filzmoser, Vienna
K. Fokianos, Nicosia
Yu. Kharin, Minsk
G. Medvedev, Minsk
Yu. Mishura, Kyiv
E. Orsingher, Rome
G. Shevlyakov, St. Petersburg
E. Zhuk, Minsk
A. Zubkov, Moscow

Both of us felt honored to be chosen as guest editors of this Special Issue of the Austrian Journal of Statistics. We are grateful to all the people who contributed to this Special Issue: to the authors for submitting such interesting articles, to the reviewers for their valuable comments, and in particular to the editor-in-chief Matthias Templ for giving us this opportunity and for his help. We also would like to mention that the CDAM'2016 Conference was supported by the CDAMCSS Project within the OeAD's programme IMPULSE.

Peter Filzmoser, Yuriy Kharin
(Guest Editors)

Yuriy Kharin
Research Institute for Applied
Problems of Mathematics and Informatics
Belarusian State University
Independence av. 4
220030 Minsk, Belarus
E-mail: Kharin@bsu.by

Peter Filzmoser
Institute of Statistics and
Mathematical Methods in Economics
Vienna University of Technology
Wiedner Hauptstr. 8–10
A-1040 Vienna, Austria
E-mail: p.filzmoser@tuwien.ac.at

Please note that all papers of this special issue are also available online at http://www.ajs.or.at

Austrian Journal of Statistics
April 2017, Volume 46, 3–11.
http://www.ajs.or.at/
doi:10.17713/ajs.v46i3-4.666

Robust Regression Analysis of Longitudinal Data under Censoring

Somnath Datta
University of Florida

Abstract

We consider regression analysis of longitudinal data when the temporal correlation is modeled by an autoregressive process. Robust R estimators of regression and autoregressive parameters are obtained. Our estimators are valid under censoring caused by detection limits. Efficient computation of the estimators is discussed. Theoretical and simulation studies of the estimators are presented. We analyze a real data set on air pollution using our methodology.

Keywords: rank estimators, left-censoring, censored rank, reweighting.

1. Introduction

We consider a time series $\{X_t : t \geq 1\}$ and an associated series of covariate vectors $\{Z_t : t \geq 1\}$, in $\Re^q$, for some $q \geq 1$. We postulate a linear model of the form $X_t = \beta_0 + \beta' Z_t + \alpha_t$, where the model errors α_t, is a stationary autoregressive time series of order p, for some $p \geq 1 : \alpha_t = \phi_1 \alpha_{t-1} + \cdots + \phi_p \alpha_{t-p} + \epsilon_t$, where $\{\epsilon_t\}$ are i.i.d. from a symmetric continuous distribution, and $\alpha_s = 0$, for $s \leq 0$. We assume that the coefficients ϕ satisfy the usual invertibility condition.

In some situations, the exact values of X_t may be unavailable due to censoring. In this paper, we develop our methodology for the situation when the censoring is to the left which may occur when the values of the time series X_t fall below a detection limit D_t. Thus, the observed data consists of $X_t^c = X_t \vee D_t$ and the censoring indicators $\delta_t = I(D_t \leq X_t)$. Our method can easily be adopted to the case of right censored data by simple changes in various formulas leading to our estimators. They can also be extended to the case when an observation is doubly censored but it requires more work.

The number of papers dealing with some form of censored time series data is limited (Vasudaven et al., 1996) although Zeger and Brookmeyer (1986) argue that censoring may occur naturally in longitudinal studies when there are detection limits on the observation that are being collected in time. They took a fully parametric approach to the above problem and fitted a Gaussian error model using the maximum likelihood approach via an EM type algorithm. In this paper, we take an estimating equation approach that is a robustified form of the least squares estimating function.

There is a sizable literature on R-estimators in the regression context (Hettmansperger and McKean, 2011) with i.i.d. errors but not for auto-regressive errors. Furthermore, an added complication arises due to censoring. As use the "approximate unbiasedness" principle of re-weighting data to construct our R-type estimating equation for the set of regression and the error autoregression parameters. Since this estimating equation involve ranks of quantities that are not computable due to censoring, the re-weighting principle is used again to compute the approximate ranks to be used in the estimating equation.

The rest of the paper is organized as follows. Section 2 describes our estimation method and discusses an efficient method of computing the estimator. We also present a model based resampling procedure for making inference using our estimators. Section 3 presents results from a number of simulation studies demonstrating the performance of our estimators. We illustrate our methodology on a real dataset dealing with air pollution in Section 4. The paper ends with a discussion section (Section 5).

2. The estimators

We develop two different estimators of the regression and the error autoregression parameters. In the first approach, we ignore the fact that the models errors are dependent and estimate the regression parameters first which are then used to estimate the autoregressive parameters. In the second approach, a joint objective function of both sets of parameters is formed.

2.1. The complete data case

First we consider the situation when there is no censoring so that we have fully observed the time series $X_t, 1 \leq t \leq n$. We form an estimating equation that is partly based on ranks of certain model residuals and is therefore yields more robust estimators than the corresponding least squares estimators. Define, for any vector $b \in \Re^q$, the residuals for the linear model part $a_t(b) \doteqdot X_t - b^T Z_t$, for $1 \leq t \leq n$, and $a_t(b) \doteqdot 0$, for $t \leq 0$. Note that even though the true errors α_t are not independent, they are still ergodic and thus we could use the same estimating equation of a traditional R-estimation in this context. Thus, we could obtain a "quick and dirty" consistent estimator of β by minimizing the objective function

$$D_{1,M}(b) = \sum_{t=1}^{n} \left\{ \phi_1 \left(\frac{R(a_t(b))}{n+1} \right) - \overline{\phi}_1 \right\} a_t(b), \tag{1}$$

where $R(a_t(b))$ is the rank of $a_t(b)$ amongst $a_1(b), \cdots, a_n(b)$. Here ϕ_1 is defined on $(0,1)$ such that ϕ_1 is monotonic and $\int \phi_1^2 < \infty$, and $\overline{\phi}_1 = n^{-1} \sum_{i=1}^{n} \phi_1 \left(i/(n+1) \right)$. After obtaining an estimate $\widehat{\beta}$, the intercept parameter can be (robustly) estimated as $\widehat{\beta}_0 = \text{med}(a_1(\widehat{\beta}), \cdots, a_n(\widehat{\beta}))$.

Having estimated the regression parameters, a similar objective function can now be formed to estimate the autoregressive part of the error time series:

$$D_{2,M}(h) = \sum_{t=1}^{n} \phi_2 \left(\frac{R(e_t(h))}{n+1} \right) \left\{ e_t(h) - \overline{e}(h) \right\}, \tag{2}$$

where $e_t(h) \doteqdot \{a_t(\widehat{\beta}) - \widehat{\beta}_0\} - \sum_{j=1}^{p} h_j \{a_{t-j}(\widehat{\beta}) - \widehat{\beta}_0\}$, $1 \leq t \leq n$, $R(e_t(h))$ is the rank of $e_t(h)$ amongst $e_1(h), \cdots, e_n(h)$, $\overline{e}(h) = n^{-1} \sum_{i=1}^{n} e_i(h)$; ϕ_2 is defined on $(0,1)$ such that ϕ_2 is monotonic and $\int \phi_2^2 < \infty$. In the rest of the paper, we refer to these estimators as "partial R estimators".

Finally, we consider a second approach where estimators are obtained by minimizing a joint objective function. For $b_0 \in \Re$, $b \in \Re^q$ and $h \in \Re^p$, define the model residuals (as a function of b_0, b and h), by $e_t(b_0, b, h) \doteqdot a_t(b_0, b) - \sum_{j=1}^{p} h_j a_{t-j}(b_0, b)$, where $a_t(b_0, b) = X_t - b_0 + b' Z_t$, $t \geq 1$, and $e_t(b_0, b, h) \doteqdot 0$, for $t \leq 0$. We then form a joint objective function

$$D_J(b_0, b, h) = \sum_{t=1}^{n} \phi_3 \left(\frac{R(e_t(b_0, b, h))}{n+1} \right) \Big\{ e_t(b_0, b, h) - \overline{e}(b_0, b, h) \Big\}, \tag{3}$$

where R and ϕ_3 are as before. In accordance with the earlier name, the resulting estimators obtained by minimizing D_J will be called the "full R-estimators".

2.2. Modification for censored data

Next, we will describe how to modify this estimating function in presence of left-censoring. Both the estimating function and the ranks have to be computed on the basis of observed data. However, in order to avoid any selection bias, the contribution of such a term has to be re-weighted by the corresponding inverse selection probability. These probabilities will have to be estimated from appropriate models fitted to the censoring distribution.

The censored data version of the objective function corresponding to (1) is of the form

$$D_{1,M}(b) = \sum_{t=1}^{n} \frac{\delta_t}{W_t} \Big\{ \phi_1 \left(\frac{R^c(a_t(b))}{n+1} \right) - \overline{\phi}_{1,c} \Big\} a_t(b), \tag{4}$$

with $\overline{\phi}_{1,c} = \left\{ \sum_{t=1}^{n} \frac{\delta_t}{W_t} \phi_1 \left(\frac{R^c(a_t(b))}{n+1} \right) / \sum_{t=1}^{n} \frac{\delta_t}{W_t} \right\}$, where the presence of δ_t indicates that the corresponding $a_t(b)$ is computable from the available data and W_t is the corresponding selection weight that is described later. The quantity R^c denotes a modified "rank" that accounts for censoring. To motivate the definition of rank for censored data, it will be useful to first consider the following sum representation for $R(a_t(b))$ in the full (uncensored) data situation:

$$R(a_t(b)) = \sum_{j=1}^{n} I[a_j(b) \leq a_t(b)].$$

Using the same re-weighting principle as before, we can define an "estimated rank" that is computable from left censored data by

$$R^c(a_t(b)) = \frac{\sum_{j=1}^{n} \frac{\delta_j}{W_j} I[a_j(b) \leq a_t(b)]}{\frac{1}{n} \sum_{j=1}^{n} \frac{\delta_j}{W_j}},$$

for any t with $\delta_t = 1$.

In the censored data case, a robust estimator of the intercept term can be obtained as

$$\widehat{\beta}_0 = \widehat{F}_a^{-1} \left(\frac{1}{2} \right), \tag{5}$$

where F_α is an estimator of the distribution function of the stationary distribution of the a_t based on the same re-weighting principle

$$\widehat{F}_\alpha(t) = \left(\sum_{t=1}^{n} \frac{\delta_t}{W_t} I[\, a_t(\widehat{b}\,) \leq t] \right) \Big/ \left(\sum_{t=1}^{n} \frac{\delta_t}{W_t} \right).$$

Next note that, in order to calculate the residual function $e_t(h)$ corresponding to time t, we need to have complete (i.e., uncensored) observations on X_j, $t-p \leq j \leq t$. Thus, the modified objective function for estimating ϕ will be of the form

$$D_{2,M}^c(h) = \sum_{t=1}^{n} \left(\prod_{j=t-p}^{t} \frac{\delta_j}{W_j} \right) \phi_2 \left(\frac{R^c(e_t(h))}{n+1} \right) \Big\{ e_t(h) - \overline{e}(h) \Big\}, \tag{6}$$

where the Ws are as before,

$$\overline{e}(h) = \left\{ \sum_{t=1}^{n} \left(\prod_{j=t-p}^{t} \frac{\delta_j}{W_j} \right) e_t(h) \right\} \Big/ \left\{ \sum_{t=1}^{n} \left(\prod_{j=t-p}^{t} \frac{\delta_j}{W_j} \right) \right\}$$

and

$$R^c(e_t(h)) = \frac{\sum_{j=1}^{n} \left(\prod_{k=j-p}^{j} \frac{\delta_k}{W_k} \right) I[e_j(h) \le e_t(h)]}{\frac{1}{n} \sum_{j=1}^{n} \left(\prod_{k=j-p}^{j} \frac{\delta_k}{W_k} \right)}, \tag{7}$$

for any t with $\prod_{j=t-p}^{t} \delta_j = 1$.
In the same way, the joint objective function can be modified to account for censored data as

$$D_J^c(b_0, b, h) = \sum_{t=1}^{n} \left(\prod_{j=t-p}^{t} \frac{\delta_j}{W_j} \right) \phi_3 \left(\frac{R^c(e_t(b_0, b, h))}{n+1} \right) \Big\{ e_t(b_0, b, h) - \overline{e}(b_0, b, h) \Big\}, \tag{8}$$

where W_j are the same as before and R^c is similarly defined as in (7).

2.3. Computation of the estimator

The estimating function can be optimized using a general purpose optimizer such as "optim" or "optimize" in R. For the $p = q = 1$ case, we can perform a grid search algorithm which we describe below.

Note that $D_{1,M}(b)$ is a linear function in b in regions where the ranks $R^c(a_t(b))$ do not change for t's with $\delta_t = 1$. For b to be such a change point, there will exist pairs of integers t and i such that $\delta_t = \delta_i = 1$ and $a_t(b) = a_i(b)$. Thus $X_t - bZ_t = X_i - bZ_i$ implying $b = (X_i - X_t)/(Z_i - Z_t)$, provided $Z_i \neq Z_t$. Let $\mathcal{B} = \{b_j : j = 1, \cdots, M\}$ be the sorted values of $\{(X_i - X_t)/(Z_i - Z_t) : 1 \le i \neq t \le n,\ \delta_i \delta_t = 1,\ Z_i \neq Z_t\}$. Then $D_{1,M}(b)$ is piecewise linear and continuous on $\mathcal{B}$. Hence $\widehat{\beta}$ can be obtained by maximizing $D_{1,M}$ on the grid of points $\mathcal{B}$. If $D_{1,M}(b)$ is constant on $[b_j, b_{j+1}]$, we will take $\widehat{\beta}$ to be the midpoint $(b_j + b_{j+1})/2$.
In the same way, $D_{2,M}^c$ can be maximized over the grid of points

$$\mathcal{H} = \left\{ \left(a_t(\widehat{\beta}) - a_i(\widehat{\beta}) \right) / \left(a_{t-1}(\widehat{\beta}) - a_{i-1}(\widehat{\beta}) \right) : 1 \le i \neq t \le n,\ \delta_i \delta_{i-1} \delta_t \delta_{t-1} = 1,\ a_{t-1}(\widehat{\beta}) \neq a_{i-1}(\widehat{\beta}) \right\}.$$

2.4. Computation of the weights

The weights W_j are estimates of the conditional (given X_j and Z_j) cumulative distribution function of the D_j, i.e., $W_j = \widehat{Pr}\{D_j \le X_j | X_j, Z_j\}$. The simplest way to estimate these will be to consider the corresponding (forward in time) hazard of $C_j = -D_j$, given X_j, Z_j,

$$\begin{aligned} &\lim_{dc \downarrow 0} \frac{\lambda_c(c \le C_j < c + dc | C_j \wedge (-X_j) \ge c, X_j, Z_j)}{dc} \\ &= \lim_{dc \downarrow 0} \frac{\lambda_c(c \le C_j < c + dc | C_j \wedge (-X_j) \ge c, Z_j)}{dc} \ \fallingdotseq \lambda_c(c | Z_j), \end{aligned}$$

where the equality is an independent censoring assumption that we impose throughout this paper. We now need a regression model on these hazard rates on the C. A flexible model is given by Aalen's linear hazards model that admits a closed form estimates of these quantities; see, e.g., Aalen (1989) or Datta and Satten (2002). A special case of these models, where we assume that $\lambda_c(c|Z_j)$ is free of the covariate Z_j, also yields the simplest choice of W_j obtained by the Kaplan-Meier estimator of the survival function based on the (right censored) C_j evaluated at $(-X_j)^-$.

2.5. Bootstrap inference

While it is possible to develop a large sample theory for our estimators by combining elements from Hettmansperger and McKean (2011), Datta and Satten (2002), and Datta and Beck (2014), we prefer to use model based resampling to perform statistical inference since it avoids the use of tuning parameter that is necessary for smoothing based asymptotic variance estimation.

Having fitted the regression model to the original data, we compute the model residuals $\widehat{\epsilon}_t = \widehat{\alpha}_t - \widehat{\phi}_1\widehat{\alpha}_{t-1} - \cdots - \widehat{\phi}_p\widehat{\alpha}_{t-p}$, with $\widehat{\alpha}_t = X_t - \widehat{\beta}_0 - \widehat{\beta}'Z_t$. Next, we resample the centered residuals to obtain ϵ_t^* which are used to compute $\alpha_t^* = \widehat{\phi}_1\alpha_{t-1}^* + \cdots + \widehat{\phi}_p\alpha_{t-p}^* + \epsilon_t^*$, and finally the bootstrapped complete data $X_t^* = \widehat{\beta}_0 + \widehat{\beta}'Z_t^* + \alpha_t^*$. The corresponding censoring times are independently generated from the fitted censoring hazard rate function based on the original data $D_t^* \sim \widehat{\lambda}_C$. Finally, we let $X_{t^c}^* = X_t^* \vee D_t^*$ and $\delta_t^* = I(D_t^* \leq X_t^*)$.

The rest of the bootstrap procedure is standard leading to either a percentile based confidence interval or a large sample normality based confidence interval where the asymptotic variance is replaced by the empirical variance of independent replicates of bootstrapped estimates of the parameter of interest.

3. Simulation studies

We consider a single continuous covariate Z that is generated from a $N(0, .64)$ distribution; we simulate the errors from an AR(1) model $\alpha_t = 0.5\alpha_{t-1} + \epsilon_t$. A number of distributions for the ϵ were investigated. The regression parameters used for the simulation were $\beta_0 = 2$ and $\beta_1 = 1$. The censoring times were generated as $D = 1/(E + 3) + m$, where E has a standard exponential distribution and $m \in \Re$ is chosen to control the censoring rate. Three choices of the censoring rates were used. The bias and variance of the estimators were empirically estimated based on $M = 1000$ Monte Carlo samples each.

Table 1 reports the results of the simulation. Some general trends are observed from this table. The joint estimators of the slope and autoregressive parameters have better performance that the corresponding partial estimators. The joint estimator of intercept parameter, on the other hand, exhibits substantial bias which worsens with the amount of censoring; however it is corrected by the modified estimator in all cases. The standard deviation, of the estimators of slope and autoregression parameters increases, albeit slightly, with the censoring level.

Next we compute the empirical coverage of the bootstrap based confidence intervals using the percentile methods, as well as the standardized statistic using bootstrap based variance estimate. The coverage appears to be very good when there is no censoring and is adequate even with 30% censoring (Table 2). Overall, confidence intervals using standardized statistics have better coverage as expected.

4. An application

We illustrate our methodology using monthly data on the chemical composition of atmospheric deposition of dry NH_4 collected by the Environmental Measurements Laboratory between 1977 and early 1980 at a number of sites in the United States (Toonkel 1981); the same data set was used by Zeger and Brookmeyer (1986) to illustrate their method. Since there are lower detection limits of the assays, the data is left-censored. Altogether, there were 43 data points out of which 6 were left-censored. In addition, there were three observations that were missing; in order to accommodate them into our framework, we treat them as left censored by an artificially set high value (larger than all the observed values in the data set). The data were log-transformed as in Zeger and Brookmeyer (1986). A plot of the log-transformed data is shown in Figure 1; where the incomplete observations are denoted by the symbol "+".

One of the main research question was to determine if the amount of deposit is increasing

with time. To that end we fit a regression model taking time as a covariate of the form $X_t = \beta_0 + \beta_1 t + \alpha_t$, where α_t was modeled by an AR(1) process. The resulting parameter estimates are given in Table 1; we include the parametric estimates by Zeger and Brookmeyer (ZB) for comparison.

While the two sets of parametric estimates are similar, the robust estimate of the intercept term is slightly smaller. More importantly, all confidence intervals for the slope term include 0 indicating that there is no significant change of the deposit levels with time. In a sense, the fact that the different analyses yielded the same scientific conclusion is reassuring.

5. Discussion

We introduce a robust and relatively model free technique of analyzing temporally correlated data that are subject to left-censoring. Although, our formulas are given here for left censored data, it is a mater of triviality to change them for right censored. With additional effort, it may be possible to extend the basic regression technique to other form of incomplete data. Another technical extension will be to consider other form of temporal correlation structures for the longitudinal responses.

This paper presents a number of novel components which may be useful for other incomplete data problems. In particular, the concept of an approximate or estimated "rank" may be applied to extend other rank based inference for censored data settings.

References

Aalen O.O. (1989). "A Linear Regression Model for the Analysis of Lifetimes." *Statistics in Medicine*, **8**, 907–925.

Datta S. and Beck J.D. (2014). "Robust Estimation of Marginal Regression Parameters in Clustered Data." *Statistical Modelling*, **14**, 489–501.

Datta S. and Satten G.A. (2002). "Estimation of Integrated Transition Hazards and Stage Occupation Probabilities for Non-Markov Systems under Stage Dependent Censoring." *Biometrics*, **58**, 792–802.

Hettmansperger T.P. and McKean J.W. (2011). *Robust Nonparametric Statistical Methods, 2nd ed.*. New York: Chapman & Hall.

Toonkel L.E. (1981). Appendix to Environmental Measurements Laboratory Environmental Report, New York: U.S. Department of Energy (available from the National Technical Information Service, U.S. Department of Commerce, Springfield, VA).

Vasudaven M. and Nair M.G. and Sithole M.M. (1996). "On Estimation for Censored Autoregressive Data." *Statistics & Probability Letters*, **31**, 97–105.

Zeger S.L. and Brookmeyer R. (1986). "Regression Analsis with Censored Autocorrelated Data." *Journal of the American Statistical Association*, **81**, 722–729.

Table 1: Performance of various estimators as measured by empirical bias and standard deviation in a simulation experiment.

			Partial R-estimators			Joint R-estimators			Modified
Sample size	Censoring %	Parameter	β_0	β_1	ϕ_1	β_0	β_1	ϕ_1	β_0
50	40%	Bias	0.002	-0.050	-0.086	-0.289	-0.032	-0.067	0.002
		SD	0.154	0.111	0.146	5.654	0.094	0.148	0.153
	30%	Bias	-0.015	-0.004	-0.070	-0.012	-0.001	-0.053	-0.015
		SD	0.157	0.107	0.138	1.429	0.087	0.139	0.157
	0%	Bias	-0.015	0.003	-0.067	-0.016	0.004	-0.051	-0.015
		SD	0.158	0.107	0.136	1.411	0.086	0.138	0.157
200	40%	Bias	0.009	-0.057	-0.040	0.042	-0.036	-0.034	0.009
		SD	0.072	0.057	0.069	1.084	0.048	0.069	0.072
	30%	Bias	-0.009	-0.011	-0.022	0.052	-0.007	-0.018	-0.009
		SD	0.073	0.053	0.063	1.075	0.043	0.064	0.073
	0%	Bias	-0.010	-0.004	-0.019	0.055	-0.002	-0.015	-0.010
		SD	0.073	0.053	0.063	1.072	0.043	0.063	0.073
500	40%	Bias	0.020	-0.050	-0.027	0.035	-0.031	-0.023	0.020
		SD	0.049	0.036	0.045	1.042	0.028	0.045	0.048
	30%	Bias	0.002	-0.005	-0.010	0.043	-0.003	-0.008	0.001
		SD	0.049	0.033	0.040	1.046	0.025	0.039	0.049
	0%	Bias	0.001	0.001	-0.008	0.044	0.001	-0.006	0.000
		SD	0.049	0.033	0.040	1.042	0.025	0.039	0.049

Table 2: Empirical coverages of bootstrap confidence intervals in simulated data

Censoring Rate		Parameter					
		Percentile Method			Normal Approximation		
	Nominal Coverage	β_0	β_1	ϕ_1	β_0	β_1	ϕ_1
0%	80%	0.772	0.766	0.794	0.800	0.772	0.802
	85%	0.826	0.810	0.826	0.850	0.814	0.852
	90%	0.880	0.882	0.888	0.896	0.880	0.902
	95%	0.926	0.940	0.938	0.940	0.952	0.950
	99%	0.964	0.990	0.984	0.974	0.988	0.984
30%	80%	0.732	0.730	0.800	0.742	0.732	0.818
	85%	0.776	0.778	0.848	0.802	0.792	0.870
	90%	0.836	0.832	0.898	0.846	0.840	0.904
	95%	0.882	0.924	0.938	0.912	0.922	0.942
	99%	0.964	0.978	0.984	0.980	0.984	0.988
40%	80%	0.714	0.426	0.662	0.756	0.644	0.748
	85%	0.766	0.472	0.720	0.806	0.702	0.808
	90%	0.828	0.560	0.780	0.854	0.760	0.874
	95%	0.892	0.678	0.860	0.928	0.842	0.936
	99%	0.954	0.846	0.958	0.974	0.946	0.974

Table 3: Parameter estimates for the Dry Deposition data

Parameter	β_0	β_1	ϕ_1	CI for ϕ_1
Our method	4.646	0.017	0.315	$(-0.018,\ 0.042)$ BS percentile method $(-0.011, 0.045)$ Using normal approximation
ZB	5.020	0.015	0.380	$(-0.042, 0.066)$

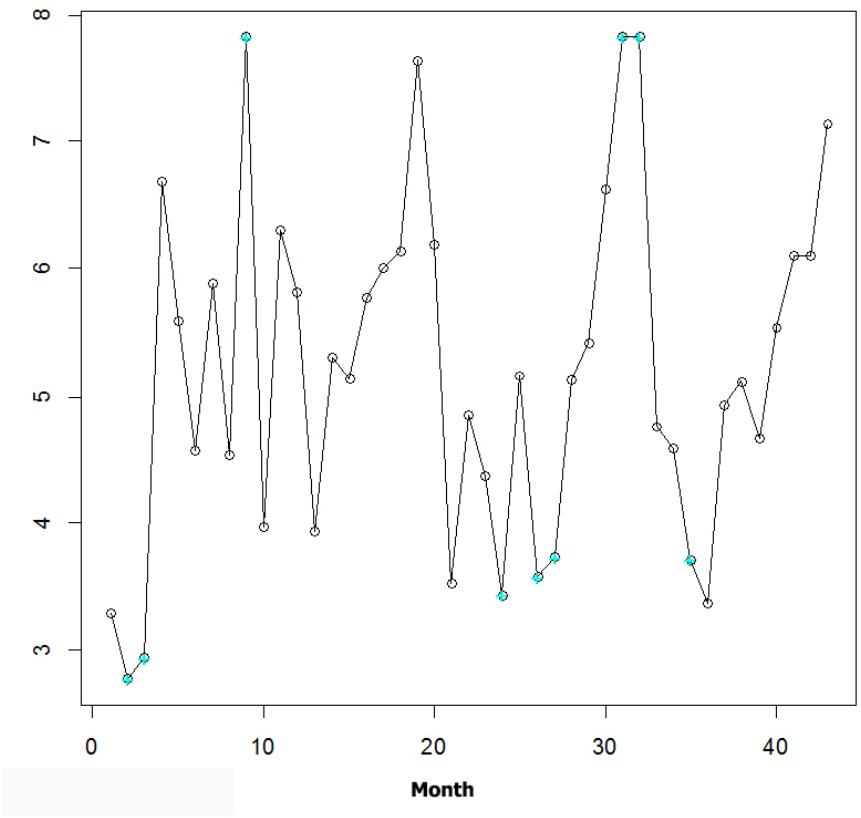


Figure 1: Log-transformed data of monthly deposition of dry NH_4; the incomplete data values are indicated by "+".

Affiliation:

Somnath Datta
Department of Biostatistics
University of Florida
Gainesville, FL 32610
E-mail: somnath.datta@ufl.edu
URL: http://www.somnathdatta.org/uf

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 46
April 2017

http://www.ajs.or.at/
http://www.osg.or.at/

Submitted: 2016-11-15
Accepted: 2017-02-02

Austrian Journal of Statistics
April 2017, Volume 46, 13–22.
http://www.ajs.or.at/
doi:10.17713/ajs.v46i3-4.667

The Spatial Sign Covariance Matrix and Its Application for Robust Correlation Estimation

Alexander Dürre
TU Dortmund

Roland Fried
TU Dortmund

Daniel Vogel
University of Aberdeen

Abstract

We summarize properties of the spatial sign covariance matrix and especially consider the relationship between its eigenvalues and those of the shape matrix of an elliptical distribution. The explicit relationship known in the bivariate case was used to construct the spatial sign correlation coefficient, which is a non-parametric and robust estimator for the correlation coefficient within the elliptical model. We consider a multivariate generalization, which we call the multivariate spatial sign correlation matrix. A small simulation study indicates that the new estimator is very efficient under various elliptical distributions if the dimension is large. We furthermore derive its influence function under certain conditions which indicates that the multivariate spatial sign correlation becomes more sensitive to outliers as the dimension increases.

Keywords: elliptical distribution, influence function, eigenvalues, fixed-point algorithm.

1. Introduction

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ denote a sample of independent p dimensional random variables from a distribution F and $s : \mathbb{R}^p \to \mathbb{R}^p$ with $s(\boldsymbol{x}) = \boldsymbol{x}/|\boldsymbol{x}|$ for $\boldsymbol{x} \neq 0$ and $s(0) = 0$ the spatial sign, then

$$S_n(\boldsymbol{t}_n, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = \frac{1}{n} \sum_{i=1}^{n} s(\boldsymbol{X}_i - \boldsymbol{t}_n) s(\boldsymbol{X}_i - \boldsymbol{t}_n)^T$$

denotes the empirical spatial sign covariance matrix (SSCM) with location $\boldsymbol{t}_n$. The canonical choice for the location estimator $\boldsymbol{t}_n$ is the spatial median

$$\boldsymbol{\mu}_n = \underset{\boldsymbol{\mu} \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} ||\boldsymbol{X}_i - \boldsymbol{\mu}||.$$

Besides its nice robustness properties like an asymptotic breakdown-point of 1/2, the spatial median has (under regularity conditions, see Kemperman 1987) the advantageous feature that it centres the spatial signs, i.e.,

$$\frac{1}{n} \sum_{i=1}^{n} s(\boldsymbol{X}_i - \boldsymbol{\mu}_n) = 0,$$

so that $S_n(\boldsymbol{\mu}_n, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ is indeed the empirical covariance matrix of the spatial signs of the data. If $\boldsymbol{t}_n$ is (strongly) consistent for a location $\boldsymbol{t} \in \mathbb{R}$, it was shown in Dürre, Vogel, and Tyler (2014) that under mild conditions on F the empirical SSCM is a (strongly) consistent estimator for its population counterpart $S(\boldsymbol{X}) = \mathsf{E}(s(\boldsymbol{X} - \boldsymbol{t})s(\boldsymbol{X} - \boldsymbol{t})^T)$, for $\boldsymbol{X} \sim F$. Results about $S(\boldsymbol{X})$ have been derived for continuous elliptical distributions F, i.e. if F possesses a density of the form

$$f(\boldsymbol{x}) = \det(V)^{-\frac{1}{2}} g((\boldsymbol{x} - \boldsymbol{\mu})^T V^{-1} (\boldsymbol{x} - \boldsymbol{\mu}))$$

for a location $\boldsymbol{\mu} \in \mathbb{R}^p$, a symmetric and positive definite shape matrix $V \in \mathbb{R}^{p \times p}$ and a function $g : [0, \infty) \to [0, \infty)$, which is often called the elliptical generator. Prominent members of the elliptical family are the multivariate normal distribution and elliptical t-distributions (e.g. Bilodeau and Brenner 1999, p. 208). If second moments exist, then $\boldsymbol{\mu}$ is the expectation of $\boldsymbol{X} \sim F$, and V a multiple of its covariance matrix. The shape matrix V is unique only up to a multiplicative constant. In the following, we consider the trace-normalized shape matrix $V_0 = V/\mathrm{tr}(V)$, which is convenient since $S(\boldsymbol{X})$ also has trace 1. If F is elliptical, then $S(\boldsymbol{X})$ and V share the same eigenvectors and the respective eigenvalues have the same ordering. For this reason, the SSCM has been proposed for robust principal component analysis (e.g. Locantore, Marron, Simpson, Tripoli, Zhang, and Cohen 1999; Marden 1999). In the present article, we study the eigenvalues of the SSCM.

In the following we discuss properties of the SSCM and extend it to correlation estimation. In Section 2 we summarize results about the eigenvalues of the SSCM and illustrate by means of two examples how the eigenvalues of the SSCM are connected with those of the shape matrix. In Section 3 a new estimator for the correlation matrix based on the SSCM is introduced, which we call the multivariate spatial sign correlation matrix. We describe a fixed-point algorithm to calculate the estimator numerically. Furthermore we investigate the efficiency of the spatial sign correlation matrix in a small simulation under different elliptical distributions and derive its influence function under specific assumptions.

2. Eigenvalues of the SSCM

Let $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$ denote the eigenvalues of V_0 and $\delta_1 \geq \ldots \geq \delta_p \geq 0$ those of $S(\boldsymbol{X})$. Explicit formulae that relate the δ_i to the λ_i are only known for $p = 2$ (see Vogel, Köllmann, and Fried 2008; Croux, Dehon, and Yadine 2010), namely

$$\delta_i = \frac{\sqrt{\lambda_i}}{\sqrt{\lambda_1} + \sqrt{\lambda_2}}, \quad i = 1, 2. \tag{1}$$

Assuming $\lambda_2 > 0$, we have $\delta_1/\delta_2 = \sqrt{\lambda_1/\lambda_2} \leq \lambda_1/\lambda_2$, thus the eigenvalues of the SSCM are closer together than those of the corresponding shape matrix. It is shown in Dürre, Tyler, and Vogel (2016) that this holds true for arbitrary $p > 2$,

$$\lambda_i/\lambda_j \geq \delta_i/\delta_j \quad \text{for } 1 \leq i < j \leq p \tag{2}$$

as long as $\lambda_j > 0$. There is no explicit map between the eigenvalues known for $p > 2$. Dürre *et al.* (2016) give a representation of δ_i as one-dimensional integral, which permits fast and accurate numerical evaluations for arbitrary p,

$$\delta_i = \frac{\lambda_i}{2} \int_0^\infty \frac{1}{(1 + \lambda_i x) \prod_{j=1}^p (1 + \lambda_j x)^{\frac{1}{2}}} dx, \quad i = 1, \ldots, p. \tag{3}$$

We use this formula, which is implemented in R (R Core Team 2016) in the package sscor (Dürre and Vogel 2016b), to get an impression how the eigenvalues of $S(\boldsymbol{X})$ look like in comparison to those of V_0. We first look at equidistantly spaced eigenvalues

$$\lambda_i = \frac{2(p + 1 - i)}{p(p + 1)}, \quad i = 1, \ldots, p,$$

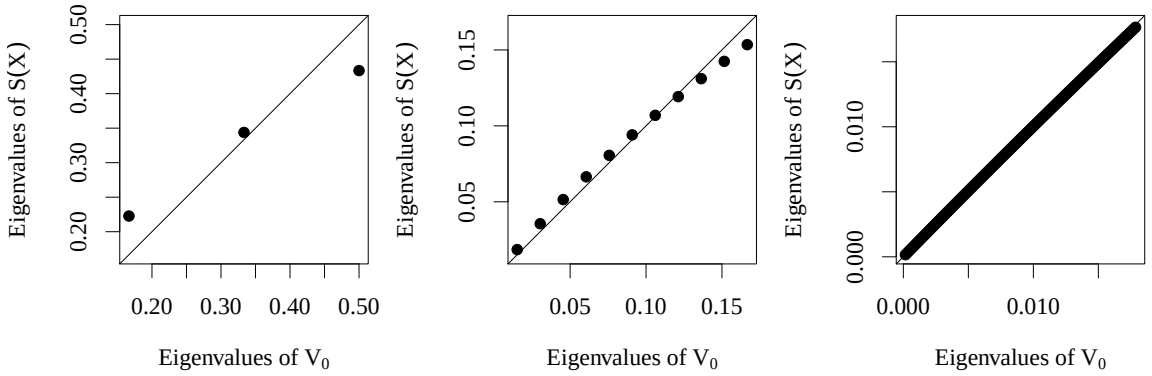


Figure 1: Eigenvalues of the SSCM w.r.t. the corresponding eigenvalues of the shape matrix in the equidistant setting $p = 3$ (left), $p = 11$ (centre) and $p = 101$ (right).

for different $p = 3,\ 11,\ 101$. The magnitude of the eigenvalues necessarily decreases as p increases, since $\sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} \delta_i = 1$ per definition of V_0 and $S(\boldsymbol{X})$. As one can see in Figure 1, the eigenvalues of $S(\boldsymbol{X})$ and V_0 approach each other for increasing p. In fact the maximal absolute difference for $p = 101$ is roughly $2 \cdot 10^{-4}$. In the second scenario, we take $p - 1$ equidistantly spaced eigenvalues and one eigenvalue 5 times larger than the rest, i.e.,

$$\lambda_i = \begin{cases} \frac{5(p-1)}{p((p+1)/2+5)-5} & i = 1, \\ \frac{p-i}{p((p+1)/2+5)-5} & i = 1, \ldots, p-1. \end{cases}$$

This models the case where the dependence is mainly driven by one principal component. As

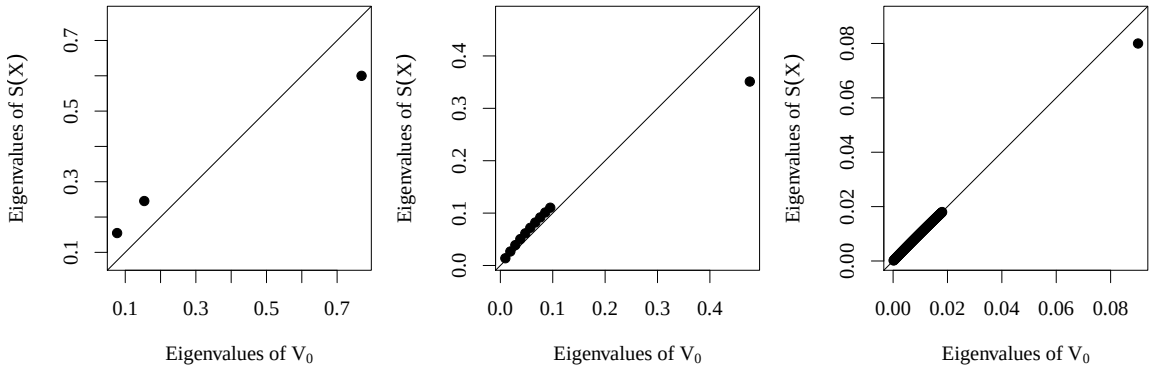


Figure 2: Eigenvalues of the SSCM wrt the corresponding eigenvalues of shape matrix in the setting of one large eigenvalue for $p = 3$ (left), $p = 11$ (centre) and $p = 101$ (right).

one can see in Figure 2, the distance between the two largest eigenvalues is smaller for $S(\boldsymbol{X})$ than for V_0. This is not surprising in the light of (2). Thus in general, the eigenvalues of the SSCM are less separated than those of V_0, which is one reason why the use of the SSCM for robust principal component analysis has been questioned (e.g. Bali, Boente, Tyler, and Wang 2011; Magyar and Tyler 2014). However, the differences appear to be generally small in higher dimensions.

3. Estimation of the correlation matrix

In the bivariate case, a robust estimator for the correlation coefficient based on the SSCM

can be obtained by inverting (1). Let

$$\rho_n = \hat{v}_{12}/\sqrt{\hat{v}_{11}\hat{v}_{22}} \quad \text{where} \quad V_n = (\hat{v}_{ij})_{i,j=1,2} = S_n^2,$$

and S_n is the bivariate SSCM. We call this estimator the spatial sign correlation coefficient. For more information, see [6]. Under mild regularity assumptions this estimator is consistent under elliptical distributions and asymptotically normal with variance

$$\mathrm{ASV}(\rho_n) = (1-\rho^2)^2 + \frac{1}{2}(a + a^{-1})(1-\rho^2)^{3/2}, \tag{4}$$

where $a = \sqrt{v_{11}/v_{22}}$ is the ratio of the marginal scales and $\rho = v_{12}/\sqrt{v_{11}v_{22}}$ is the generalized correlation coefficient, which coincides with the usual moment correlation coefficient if second moments exists. Equation (4) indicates, that for fixed ρ, the variance of ρ_n is minimal for $a = 1$, but can get arbitrarily large if a tends to infinity or 0.

Therefore a two-step procedure has been proposed, the *two-stage spatial sign correlation* $\rho_{\sigma,n}$, which first margin-wise standardizes the data by a robust scale estimator, e.g., the median absolute deviation (MAD), and then computes the spatial sign correlation of the standardized data. Under mild conditions (see Dürre and Vogel 2016a), this two-step procedure yields an asymptotic variance of

$$\mathrm{ASV}(\rho_{\sigma,n}) = (1-\rho^2)^2 + (1-\rho^2)^{3/2}, \tag{5}$$

which equals that of ρ_n for the most favourable case of $a = 1$. Since (5) only depends on the parameter ρ, the two-stage spatial sign correlation coefficient is very suitable to construct robust and non-parametric confidence intervals for the correlation coefficient under ellipticity. It turns out that these intervals are quite accurate even for rather small sample sizes of $n = 10$ and in fact more accurate than those based on the sample moment correlation coefficient (Dürre and Vogel 2016a).

3.1. The multivariate spatial sign correlation matrix

One can construct an estimator of the correlation matrix R by filling the off-diagonal positions of the matrix estimate with the bivariate spatial sign correlation coefficients of all pairs of variables. This was proposed in Dürre, Vogel, and Fried (2015). Equation (3) allows an alternative approach: First standardize the data marginally by a robust scale estimator and compute the SSCM of the transformed data. Then apply a singular value decomposition

$$S_n(\boldsymbol{t}_n, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = \hat{U}\hat{\Delta}\hat{U}^T,$$

where $\hat{\Delta}$ contains the ordered eigenvalues $\hat{\delta}_1 \geq \ldots \geq \hat{\delta}_p$. One obtains estimates $\hat{\lambda}_1, \ldots, \hat{\lambda}_p$ by inverting (3). Although theoretical results are yet to be established, we found in our simulations that the following fix point algorithm

$$\begin{aligned}
\hat{\lambda}_i^{(0)} &= \hat{\delta}_i, && i = 1, \ldots, p, \\
\tilde{\lambda}_i^{(k+1)} &= 2\hat{\delta}_i \left(\int_0^\infty \frac{1}{(1+\hat{\lambda}_i^{(k)}x)\prod_{j=1}^p (1+\hat{\lambda}_j^{(k)}x)^{\frac{1}{2}}} dx, \right)^{-1}, && i = 1, \ldots, p,\ k = 1, 2, \ldots \\
\hat{\lambda}_i^{(k+1)} &= \tilde{\lambda}_i^{(k+1)} \left(\sum_{j=1}^p \tilde{\lambda}_j^{(k+1)} \right)^{-1}, && i = 1, \ldots, p,\ k = 1, 2, \ldots
\end{aligned}$$

works reliably and converges fast, converging usually within 5 iterations if p is large. Let $\hat{\Lambda}$ denote the diagonal matrix containing $\hat{\lambda}_1, \ldots, \hat{\lambda}_p$, then $\hat{V} = \hat{U}\hat{\Lambda}\hat{U}^T$ is a suitable estimator for the shape of the standardized data and $\hat{R}$ with $\hat{\rho}_{ij} = \hat{v}_{ij}/\sqrt{\hat{v}_{ii}\hat{v}_{jj}}$ an estimator for the correlation matrix, which we call the *multivariate spatial sign correlation matrix*. As opposed

to the pairwise approach, the multivariate spatial sign correlation matrix is positive semi-definite by construction.

3.2. Simulation under elliptical distributions

By a small simulation study we want to obtain an impression of the efficiency of the multivariate spatial sign correlation matrix. We compare the variances of the moment correlation, the pairwise as well as the multivariate spatial sign correlation under several elliptical distributions: normal, Laplace and t distributions with 5 and 10 degrees of freedom. The latter three generate heavier tails than the normal distribution. The Laplace distribution is obtained by the elliptical generator $g(x) = c_p \exp(-\sqrt{|x|}/2)$, where c_p is the appropriate integration constant depending on p (e.g. Bilodeau and Brenner 1999, p. 209).

First we take the identity matrix as shape matrix and compare the variances of an off-diagonal element of the matrix estimates for different dimensions $p = 2,\ 3,\ 5,\ 10,\ 50$ and sample sizes $n = 100,\ 1000$. We use the R packages mvtnorm (Genz, Bretz, Miwa, Mi, Leisch, Scheipl, and Hothorn 2015) and MNM (Nordhausen and Oja 2011) for the data generation. The results based on 10000 runs are summarized in Table 1.

Table 1: Simulated variances (multiplied by n) of one off-diagonal element of the correlation matrix estimate based on the moment correlation (cor), the pairwise spatial sign correlation (sscor pairwise) and the multivariate spatial sign correlation matrix (sscor multivariate) for spherical normal (N), t_5, t_{10}, and Laplace (L) distribution, several dimensions p and sample sizes $n = 100, 1000$.

	n	100					1000				
	p	2	3	5	10	50	2	3	5	10	50
N	cor	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	sscor pairwise	1.9	1.9	1.9	1.9	1.9	2.0	2.0	2.0	2.0	2.0
	sscor multivariate	1.9	1.6	1.4	1.2	1.0	2.0	1.7	1.4	1.2	1.0
t_{10}	cor	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.4	1.3
	sscor pairwise	2.0	1.9	1.9	2.0	1.9	2.0	2.0	2.0	2.0	2.0
	sscor multivariate	2.0	1.7	1.3	1.2	1.0	2.0	1.7	1.4	1.2	1.0
t_5	cor	2.0	2.1	2.1	2.1	2.1	2.6	2.6	2.6	2.6	2.6
	sscor pairwise	2.0	2.0	1.9	2.0	1.9	2.1	2.0	2.0	2.0	2.0
	sscor multivariate	2.0	1.7	1.4	1.2	1.1	2.1	1.7	1.4	1.2	1.0
L	cor	1.6	1.5	1.3	1.2	1.1	1.6	1.5	1.3	1.2	1.1
	sscor pairwise	1.9	1.9	1.9	2.0	2.0	2.0	2.0	2.0	2.0	2.0
	sscor multivariate	1.9	1.6	1.4	1.2	1.1	2.0	1.7	1.4	1.2	1.1

Except for the moment correlation at the t_5 distribution, the results for $n = 100$ and $n = 1000$ are very similar. Note that the variance of the moment correlation decreases at the Laplace distribution as the dimension p increases, but not so for the other distributions considered. The lower dimensional marginals of the Laplace distribution are, contrary to the normal and the t-distributions, not within the same distributional class (see Kano 1994), and the kurtosis of the one-dimensional marginals of the Laplace distribution in fact decreases as p increases.

Equation (5) yields an asymptotic variance of 2 for the pairwise spatial sign correlation matrix elements regardless of the specific elliptical generator. This can also be observed in the simulation results. The moment correlation is twice as efficient under normality, but it has a higher variance at heavy tailed distributions. For uncorrelated t_5 distributed random variables, the spatial sign correlation outperforms the moment correlation. Looking at the multivariate spatial sign correlation, we see a strong increase of efficiency for larger p. For $p = 50$ the variance is comparable to that of the moment correlation. Since the asymptotic variance of the SSCM does not depend on the elliptical generator, this is expected to apply also for the

multivariate spatial sign correlation, and this claim is confirmed by the simulations. The multivariate spatial sign correlation is more efficient than the moment correlation even under slightly heavier tails for moderately large p.

Following a referee's suggestion, we simulate also from other shape matrices, e.g., the equi-correlation matrix

$$V = \begin{pmatrix} 1 & 0.5 & \dots & 0.5 \\ 0.5 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0.5 \\ 0.5 & \dots & 0.5 & 1 \end{pmatrix}.$$

The results can be found in Table 2. Except for the general smaller asymptotic variances we get the same picture. The asymptotic variance of the multivariate spatial sign correlation matrix is shrinking with growing dimension and approaches that of the sample correlation under normality, albeit more slowly than in the uncorrelated case.

Table 2: Simulated variances (multiplied by n) of one off-diagonal element of the correlation matrix estimate based on the moment correlation (cor), the pairwise spatial sign correlation (sscor pairwise) and the multivariate spatial sign correlation matrix (sscor multivariate) for equi-correlated normal (N), t_5, t_{10}, and Laplace (L) distribution, several dimensions p and sample sizes $n = 100, 1000$.

	n	100					1000				
	p	2	3	5	10	50	2	3	5	10	50
N	cor	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6
	sscor pairwise	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2
	sscor multivariate	1.2	1.0	1.0	0.8	0.8	1.2	1.0	0.9	0.8	0.7
t_{10}	cor	0.8	0.7	0.7	0.8	0.8	0.8	0.7	0.8	0.7	0.8
	sscor pairwise	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2
	sscor multivariate	1.2	1.1	0.9	0.8	0.8	1.2	1.0	0.9	0.8	0.7
t_5	cor	1.2	1.2	1.2	1.2	1.2	1.5	1.5	1.5	1.5	1.5
	sscor pairwise	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2
	sscor multivariate	1.2	1.0	0.9	0.8	0.7	1.2	1.0	0.9	0.8	0.8
L	cor	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1
	sscor pairwise	1.2	1.2	1.2	1.3	1.2	1.2	1.2	1.2	1.2	1.2
	sscor multivariate	1.2	1.1	0.9	0.9	0.7	1.2	1.0	0.9	0.8	0.7

An increase of efficiency for larger p is not uncommon for robust scatter estimators. It can be observed amongst others for M-estimators, the Tyler shape matrix, the MCD, and S-estimators (see e.g. Croux and Haesbroeck 1999; Taskinen, Croux, Kankainen, Ollila, and Oja 2006). All of these are affine equivariant estimators, requiring $n > p$. This restriction is not necessary for the spatial sign correlation matrix.

3.3. Sensitivity to outliers

One may expect that the efficiency gain for large p is at the expense of robustness. We therefore investigate the influence function of one off-diagonal element of the multivariate spatial sign correlation. The influence function is based on the concept that estimators are functionals working on distributions. In this setting the specific estimate based on a given dataset equals the functional evaluated at the corresponding empirical distribution. Denote $\check{\rho}$ the functional representation of the multivariate spatial sign correlation with matrix-elements $\check{\rho}_{i,j},\ 1 \le i < j \le p$. Then the influence function $IF(\boldsymbol{x}, \check{\rho}_{i,j}, F)$ is defined as

$$IF(\boldsymbol{x}, \check{\rho}_{i,j}, F) = \lim_{\epsilon \to 0} \frac{\check{\rho}_{i,j}((1-\epsilon)F + \epsilon\Delta_{\boldsymbol{x}}) - \check{\rho}_{i,j}(F)}{\epsilon}$$

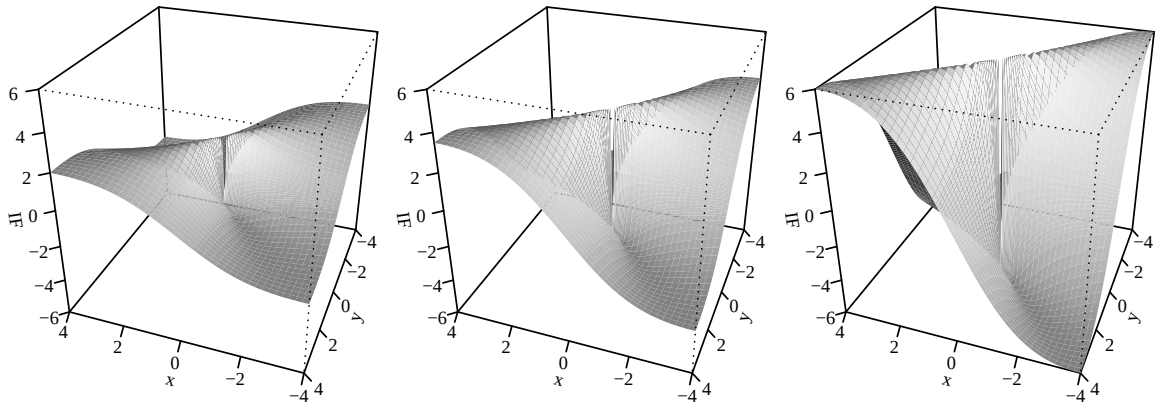


Figure 3: Partial influence functions of the off-diagonal element of multivariate spatial sign correlation $\check{\rho}_{12}$ for $\boldsymbol{x} = (x, y, 0, \ldots, 0)$ under spherical distribution for $p = 2$ (left), $p = 5$ (centre) and $p = 10$ (right).

where $\Delta_{\boldsymbol{x}}$ denotes the Dirac measure putting its mass at $\boldsymbol{x}$. For further explanations and details about the influence function, see Huber and Ronchetti (2009).
Since we do not have an explicit representation for the estimated eigenvalues $\hat{\lambda}_1, \ldots, \hat{\lambda}_p$, it seems to be challenging to calculate the influence function for arbitrary F and $\boldsymbol{x}$. Nevertheless, we can get results if we restrict ourselves to the case where F is elliptical with shape $V = I_p$ and $\boldsymbol{x}$ lies in a special hyperplane of $\mathbb{R}^p$. Furthermore we look at the case where the proportions of the marginal scales are known, respectively the data is not standardized prior to the computation of the SSCM. The following proposition, the proof of which can be found in the appendix, states the influence function in the outlined situation.

Proposition 1. *Let F be elliptical with shape $V = I_p$ and $\boldsymbol{\mu} = 0$. If we let $\check{\rho}_{i,j}$ denote the functional representation of the off-diagonal element of the multivariate spatial sign correlation without pre-standardization and let $\boldsymbol{x} = (x, y, 0, \ldots, 0)^T$ with $x, y \in \mathbb{R}$, then*

$$IF(\boldsymbol{x}, \check{\rho}_{1,2}, F) = (p+2)\frac{xy}{x^2+y^2}. \tag{6}$$

For $p = 2$, Proposition 1 is a special case of Proposition 4 in Dürre *et al.* (2015) which gives the influence function for arbitrary V. Although Proposition 1 is restricted to the situation where there is only contamination in the first two components, it provides evidence that the sensitivity of the multivariate spatial sign correlation increases with increasing dimension. One can see in Figure 3 respectively formula (6) that the influence functions are proportional to each other and that $|IF(\boldsymbol{x}, \check{\rho}_{1,2}, F)|$ increases linearly in p for fixed $\boldsymbol{x} = (x, y, 0 \ldots, 0)$. This result indicates that the multivariate spatial sign correlation is more effected by outliers if p is large.

4. Conclusion

We have discussed properties of the spatial sign covariance matrix, in particular those concerning its eigenvalues under elliptical distributions. We expand on the eigenvalue representation as one-dimensional integrals given in Dürre *et al.* (2016). First we use it to investigate the function mapping the eigenvalues of the shape matrix onto the ones of the spatial sign covariance. The eigenvalues of the spatial sign covariance matrix are closer together than the ones of the shape matrix on a logarithmic scale, see Dürre *et al.* (2016). Two examples suggest

that this behaviour diminishes as the dimension increases. One may suspect that the map between the eigenvalues of the spatial sign covariance matrix and the shape matrix converges towards the identity modulo a multiplicative constant as the dimension tends to infinity.
Our second application of the integral representation is the construction of the multivariate spatial sign correlation matrix. By a fixed-point algorithm one can invert the map between the eigenvalues of the shape and the spatial sign covariance matrix and, based on this, estimate the correlation matrix of an elliptical distributed random vector. We found the fixed-point algorithm to work reliably and fast for various shape matrices and dimensions. Simulations show that the resulting estimator is highly efficient in larger dimensions. Its asymptotic variance appears to approach that of the sample correlation under normality as the dimension is growing. Asymptotics confirming the simulation results are of great interest. The calculated partial influence function indicates that the efficiency gain of the spatial sign correlation matrix is at the cost of robustness. So the estimator does not seem to be very robust in the case of very high dimensions, but is nevertheless very efficient under heavy-tailed distributions.

Acknowledgements

Alexander Dürre and Roland Fried were supported in part by the Collaborative Research Grant 823 of the German Research Foundation. The authors are grateful to the referee for the constructive comments, which helped to improve the presentation of the article.

Appendix

Proof of Proposition 1: Let denote $\check{S}$ the functional representation of the spatial sign covariance matrix $\check{S}(F) = \mathsf{E}_F\left(\frac{\boldsymbol{X}\boldsymbol{X}^T}{\boldsymbol{X}^T\boldsymbol{X}}\right)$ where $\boldsymbol{X}$ has distribution F. Since $\check{S}((1-\epsilon)F + \epsilon\Delta_{\boldsymbol{x}}) = (1-\epsilon)I_p + \epsilon\boldsymbol{x}\boldsymbol{x}^T$ is a block diagonal matrix, we get the following eigenvalue decomposition $\check{S}((1-\epsilon)F + \epsilon\boldsymbol{x}\boldsymbol{x}^T) = U_{xy}\Delta_\epsilon U_{xy}^T$ where

$$U_{x,y} = \begin{pmatrix} \frac{x}{\sqrt{x^2+y^2}} & \frac{y}{\sqrt{x^2+y^2}} & 0 & \dots & 0 \\ \frac{y}{\sqrt{x^2+y^2}} & \frac{-x}{\sqrt{x^2+y^2}} & 0 & \dots & 0 \\ 0 & 0 & 1 & & 0 \\ & & & \ddots & \\ 0 & 0 & 0 & & 1 \end{pmatrix} \quad \text{and} \quad \Delta_\epsilon = \begin{pmatrix} \frac{1+(p-1)\epsilon}{p} & 0 & 0 & \dots & 0 \\ 0 & \frac{1-\epsilon}{p} & 0 & \dots & 0 \\ 0 & 0 & \frac{1-\epsilon}{p} & & 0 \\ & & & \ddots & \\ 0 & 0 & 0 & & \frac{1-\epsilon}{p} \end{pmatrix}.$$

We need to know how the perturbation of the eigenvalues of the SSCM translates into the eigenvalues of the shape matrix. The function $\Phi : \mathbb{R}^p \to \mathbb{R}^p$ which maps the eigenvalues of the shape to the eigenvalues of the SSCM is injective (see Proposition 1 in Dürre *et al.* 2016). Therefore the shape matrix related to Δ_ϵ contains only two distinct eigenvalues: λ_1 and $\lambda_2 = \ldots = \lambda_p$. We can simplify the situation even further since the eigenvalues are not uniquely defined and standardize them such that $\lambda_2 = \ldots, \lambda_p = 1$. On the other hand we have $\sum_{i=1}^p \delta_i = 1$ and therefore $\delta_i = \frac{1-\delta_1}{p-1}$, $i = 2, \ldots, p$. Consequently in this case the connection between the eigenvalues can be expressed by the one-dimensional function $f : [0,1] \to [0,\infty)$ which maps the first eigenvalue of Δ_ϵ to the first of the shape matrix.
Let $\gamma : \mathbb{R}^{p\times p} \to [-1,1]$ denote the function which computes the correlation coefficient between the first and second component given the shape matrix: $\gamma(A) = \frac{a_{12}}{\sqrt{a_{11}a_{22}}}$ and denote further $k(\epsilon) = \frac{1+(p-1)\epsilon}{p}$, then straightforward calculations yields,

$$\begin{aligned} \lim_{\epsilon\to 0} \frac{\check{\rho}_{i,j}((1-\epsilon)F + \epsilon\Delta_{\boldsymbol{x}}) - \check{\rho}_{i,j}(F)}{\epsilon} &= \lim_{\epsilon\to 0} \frac{\gamma\left(U_{xy} f\left(\frac{1+(p-1)\epsilon}{p}\right) U_{xy}^T\right)}{\epsilon} \\ &= \lim_{\epsilon\to 0} \frac{1}{\epsilon} \frac{(f[k(\epsilon)]-1)xy}{\sqrt{y^2 + f[k(\epsilon)]x^2}\sqrt{x^2 + f[k(\epsilon)]y^2}} =: \frac{\partial}{\partial\epsilon} h(f[k(\epsilon)])\Big|_{\epsilon=0}. \end{aligned}$$

By the chain rule we get:

$$\frac{\partial}{\partial \epsilon} h(f[k(\epsilon)])\Big|_{\epsilon=0} = \frac{\partial}{\partial \epsilon} h(x)\Big|_{x=1} \cdot \frac{\partial}{\partial y} f(y)\Big|_{y=1/p} \cdot \frac{\partial}{\partial \epsilon} k(\epsilon)\Big|_{\epsilon=0}.$$

Whereas differentiation of h and k is straightforward, we do not have an explicit representation of f. Since we only need its derivative, we can apply the inverse function theorem. Using (3) and Leibniz's rule we arrive at

$$\begin{aligned}\frac{\partial}{\partial x} f(x)\Big|_{x=1/p} &= \frac{1}{\frac{\partial}{\partial x} f^{-1}(x)|_{x=1}} \\ &= 1/\left(\frac{1}{2}\int_0^\infty \frac{1}{(1+z)^{\frac{p}{2}+1}} dz - \frac{3}{4}\int_0^\infty \frac{z}{(1+z)^{\frac{p}{2}+2}} dz\right) =: \frac{1}{A_1 + A_2}.\end{aligned}$$

For A_1 and A_2 we can apply formula 3.193-3 in Gradshteyn and Ryzhik (2000):

$$\int_0^\infty \frac{x^{\mu-1} dx}{(1+\beta x)^\nu}\, dx = B(\mu, \nu-\mu) \quad \text{for} \quad \nu > \mu > 0$$

where B denotes the beta function. Setting $\beta = 1$, $\mu = 1$ and $\nu = p/2+1$ for A_1 respectively $\mu = 2$ and $\nu = p/2+2$ for A_2 and using the relationship between beta and gamma function we arrive at $A_1 = \frac{1}{p}$ and $A_2 = \frac{3}{2p(p/2+1)}$. Straightforward term manipulations yield the stated formula (6). □

References

Bali JL, Boente G, Tyler DE, Wang JL (2011). "Robust Functional Principal Components: A Projection-pursuit Approach." *The Annals of Statistics*, **39**(6), 2852–2882.

Bilodeau M, Brenner D (1999). *Theory of Multivariate Statistics*. Springer Science & Business Media.

Croux C, Dehon C, Yadine A (2010). "The k-step Spatial Sign Covariance Matrix." *Advances in data analysis and classification*, **4**(2-3), 137–150.

Croux C, Haesbroeck G (1999). "Influence Function and Efficiency of the Minimum Covariance Determinant Scatter Matrix Estimator." *Journal of Multivariate Analysis*, **71**(2), 161–190.

Dürre A, Tyler DE, Vogel D (2016). "On the Eigenvalues of the Spatial Sign Covariance Matrix in More than Two Dimensions." *Statistics & Probability Letters*, **111**, 80–85.

Dürre A, Vogel D (2016a). "Asymptotics of the Two-stage Spatial Sign Correlation." *Journal of Multivariate Analysis*, **144**, 54–67.

Dürre A, Vogel D (2016b). *sscor: Spatial Sign Correlation*. R package version 0.2, URL http://CRAN.R-project.org/package=sscor.

Dürre A, Vogel D, Fried R (2015). "Spatial Sign Correlation." *Journal of Multivariate Analysis*, **135**, 89–105.

Dürre A, Vogel D, Tyler DE (2014). "The Spatial Sign Covariance Matrix with Unknown Location." *Journal of Multivariate Analysis*, **130**, 107–117.

Genz A, Bretz F, Miwa T, Mi X, Leisch F, Scheipl F, Hothorn T (2015). *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.0-3, URL http://CRAN.R-project.org/package=mvtnorm.

Gradshteyn I, Ryzhik I (2000). *Table of Integrals, Series, and Products. Translation edited and with a preface by Alan Jeffrey and Daniel Zwillinger.* 6th edition. Amsterdam: Elsevier/Academic Press.

Huber PJ, Ronchetti E (2009). *Robust Statistics.* Wiley.

Kano Y (1994). "Consistency Property of Elliptic Probability Density Functions." *Journal of Multivariate Analysis*, **51**(1), 139–147.

Kemperman JHB (1987). "The Median of a Finite Measure on a Banach Space." In Y Dodge (ed.), *Statistical Data Analysis Based on the L_1-Norm and Related Methods*, pp. 217–230. Amsterdam: North-Holland.

Locantore N, Marron J, Simpson D, Tripoli N, Zhang J, Cohen K (1999). "Robust Principal Component Analysis for Functional Data." *Test*, **8**(1), 1–28.

Magyar AF, Tyler DE (2014). "The Asymptotic Inadmissibility of the Spatial Sign Covariance Matrix for Elliptically Symmetric Distributions." *Biometrika*, **101**(3), 673–688.

Marden JI (1999). "Some Robust Estimates of Principal Components." *Statistics & Probability Letters*, **43**(4), 349–359.

Nordhausen K, Oja H (2011). "Multivariate L_1 Methods: The Package MNM." *Journal of Statistical Software*, **43**(5), 1–28. URL http://www.jstatsoft.org/v43/i05/.

R Core Team (2016). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Taskinen S, Croux C, Kankainen A, Ollila E, Oja H (2006). "Influence Functions and Efficiencies of the Canonical Correlation and Vector Estimates Based on Scatter and Shape Matrices." *Journal of Multivariate Analysis*, **97**(2), 359–384.

Vogel D, Köllmann C, Fried R (2008). "Partial Correlation Estimates Based on Signs." In *Proceedings of the 1st Workshop on Information Theoretic Methods in Science and Engineering. TICSP series.*

Affiliation:

Alexander Dürre
Department of Statistics
Technische Universität Dortmund
44221 Dortmund, Germany
E-mail: alexander.duerre@udo.edu
URL: https://www.statistik.tu-dortmund.de/duerre-en.html

Austrian Journal of Statistics
April 2017, Volume 46, 23–36.
http://www.ajs.or.at/
doi:10.17713/ajs.v46i3-4.668

Performance and Robustness Analysis of Sequential Hypotheses Testing for Time Series with Trend

Alexey Kharin
Belarusian State Universitiy

Ton That Tu
Belarusian State Universitiy,
Da Nang University of Education

Abstract

The problem of sequential testing of simple hypotheses for time series with a trend is considered. Analytic expressions and asymptotic expansions for error probabilities and expected numbers of observations are obtained. Robustness analysis is performed. Numerical results are given.

Keywords: sequential test, time series with trend, error probabilities, expected sample sizes, asymptotic expansions.

1. Introduction

The sequential approach to test parametric hypotheses proposed by Wald (see Wald (1947)) has been applied in many practical problems of computer data analysis. The sequential probability ratio test (SPRT) is proved to be optimal in terms of minimizing expected sample size under the assumption that type I and type II error probabilities do not exceed preassigned values (see Wald and Wolfowitz (1948)). The problem of sequential test performance characteristics (error probabilities and expected number of observations) evaluation is well studied for the case of identical distribution of observations (see Govindarajulu (2004), Kharin (2013), Kharin (2016)). In this paper, the model of non-identical distribution is considered for the problem of two simple hypotheses testing (see Kharin and Ton (2016)).

In practice, data does not often follow the hypothetical model exactly (see Huber (1981), Kharin (2005)), and the problem of robustness under distortions (see Kharin (1997), Maevskii and Kharin (2002)) is important for sequential testing (see Kharin (2011), Kharin and Kishylau (2015)). Here we consider the problem of robustness of sequential tests for time series with trend.

2. Mathematical model

Let $x_1, x_2, ...$ be time series with a trend:

$$x_t = \theta^T \psi(t) + \xi_t,\ t = 1, 2, 3, ..., \tag{1}$$

where $\psi(t) = (\psi_1(t), \psi_2(t), ..., \psi_m(t))^T$, $t \geq 1$, are the vectors of basic functions of trend, $\theta = (\theta_1, \theta_2, ..., \theta_m)^T \in \mathbb{R}^m$ is an unknown vector of coefficients, and $\{\xi_t, t \geq 1\}$ is the sequence of independent identically distributed random variables, $\xi_t \sim N(0, \sigma^2)$.

Consider two simple hypotheses ($\theta^0, \theta^1 \in \mathbb{R}^m$ are known vectors):

$$H_0 : \theta = \theta^0, H_1 : \theta = \theta^1. \tag{2}$$

Denote the accumulated log-likelihood ratio statistic:

$$\Lambda_n = \Lambda_n(x_1, x_2, ..., x_n) = \sum\nolimits_{t=1}^{n} \lambda_t, \tag{3}$$

where $\lambda_t = \ln\left(\dfrac{p_t(x_t, \theta^1)}{p_t(x_t, \theta^0)}\right)$ is the log-likelihood ratio calculated on observation x_t, and $p_t(x, \theta)$ is the probability density function of x_t provided the true parameter value is θ. To test hypotheses (2), after n observations one makes the decision:

$$d = \mathbf{1}_{[C_+, +\infty)}(\Lambda_n) + 2 \cdot \mathbf{1}_{(C_-, C_+)}(\Lambda_n). \tag{4}$$

Thresholds C_- and C_+ are the parameters of the test. Decisions $d = 0$ and $d = 1$ mean stopping of the observation process and acceptance of H_0 or H_1 correspondently. According to Wald (1947),

$$C_+ = \ln\left((1 - \beta_0)/\alpha_0\right), C_- = \ln\left(\beta_0/(1 - \alpha_0)\right),$$

where α_0, β_0 are given values for error probabilities of types I and II respectively.

3. Some auxiliary results

Let $\zeta_n, n \geq 1$ be a sequence of random variables satisfying the following conditions:

$$\text{i) } \zeta_n \in T = \{0, 1, 2, ..., K, K+1\}, n \in \mathbb{N}; \tag{5}$$

$$\text{ii) } P(\zeta_n = i_1 | \zeta_k = i_1) = 1, i_1 \in \{0, 1\}, n > k; \tag{6}$$

$$\text{iii) } P(\zeta_n = i_1 | \zeta_k = i_2, \zeta_l = i_3) = P(\zeta_n = i_1 | \zeta_k = i_2), n > k > l \geq 1, i_1, i_2, i_3 \in T. \tag{7}$$

Remark 1. *A Markov chain with a finite state space T, in which the states 0, 1 are absorbing, satisfies conditions (5)-(7).*

Introduce the notation:

$$T_1 = \{0, 1\}, T_2 = \{2, 3, ..., K, K+1\}, T = T_1 \cup T_2,$$
$$P(k) = \{p_{ij}(k)\}_{(K+2)\times(K+2)}, P(k, l) = \{p_{ij}(k, l)\}_{(K+2)\times(K+2)},$$
$$p_{ij}(k) = P(\zeta_k = j | \zeta_{k-1} = i), p_{ij}(k, l) = P(\zeta_{k+l} = j | \zeta_k = i).$$

Since (6), matrices $P(k)$ and $P(1, k)$ can be expressed as follows:

$$P(k) = \left(\begin{array}{c|c} I_2 & O_{2\times K} \\ \hline R_k & Q_k \end{array}\right), k \geq 2; \quad P(1, k) = \left(\begin{array}{c|c} I_2 & O_{2\times K} \\ \hline \bar{R}_k & \bar{Q}_k \end{array}\right), k \geq 1, \tag{8}$$

where $R_k, \bar{R}_k$ are some matrices of size $K \times 2$, I_k is the identity matrix of size k, $O_{2\times K}$ is the $2 \times K$-matrix with all elements equal to 0, and $Q_k, \bar{Q}_k$ are some matrices of size $K \times K$. For $k < n < k + l$ we have:

$$p_{ij}(k, l) = \sum_{t \in T} P(\zeta_n = t | \zeta_k = i) P(\zeta_{k+l} = j | \zeta_n = t) = \sum_{t \in T} p_{it}(k, n-k) p_{tj}(n, k+l-n),$$

which implies that

$$P(k, l) = P(k, n-k) P(n, k+l-n), \ k < n < k+l. \tag{9}$$

Therefore,

$$P(1,k) = P(2)P(3)...P(k+1),\ k \geq 1. \tag{10}$$

From (8) and (10), we get ($Q_1 = I_K$):

$$\bar{Q}_k = Q_1 Q_2 ... Q_k Q_{k+1}. \tag{11}$$

Let t be the total number of time moments for which process ζ_n belongs to T_2; $n_j (j \in T_2)$ be the number of time moments for which $\zeta_n = j$; u_j^k be the function that is 1 if the process $\zeta_n = j$ after k steps, and is 0 otherwise; $E_i(\cdot)$ be the conditional expected value given $\zeta_1 = i$. Denote: $N = \sum_{k=0}^{+\infty} \bar{Q}_k$; $\tau = N\mathbf{1}_K$; $\mathbf{1}_K$ is the vector of size K with all elements equal to 1.

Theorem 1. *In the above notation, for sequence (5)-(7) the following equations are satisfied:*

$$\{E_i(n_j)\}_{K\times K} = N,\ i,j \in T_2, \{E_i(t)\}_{K\times 1} = \tau,\ i \in T_2. \tag{12}$$

Proof. Consider representation $n_j = \sum_{k=0}^{+\infty} u_j^k$. Therefore,

$$\{E_i(n_j)\} = \{\sum_{k=0}^{+\infty} E_i(u_j^k)\} = \sum_{k=0}^{+\infty} \{p_{ij}(1,k)\} = \sum_{k=0}^{+\infty} \bar{Q}_k;$$
$$\{E_i(t)\} = \{\sum_{j\in T_2} E_i(n_j)\} = N\mathbf{1}_K = \tau.$$

□

Let B be a matrix of size $K \times 2$, $B = \{b_{ij}\}_{K\times 2}$, where b_{ij} is the probability that the sequence ζ_n started in i is absorbed in j, $i \in T_2, j \in T_1$.

Theorem 2. *If conditions (5)-(7) hold for ζ_n, then*

$$B = \sum_{k=1}^{+\infty} \bar{Q}_{k-1} R_{k+1}. \tag{13}$$

Proof. Let $B(k), k \geq 1$, be matrix of size $K \times 2$, $B(k) = \{b_{ij}(k)\}_{K\times 2}$, where $b_{ij}(k)$ is the probability that the process ζ_n starting in i is absorbed in j after exactly k steps, $i \in T_2, j \in T_1$. We obtain (13) from the facts that $B(k) = \bar{Q}_{k-1} R_{k+1}$ and $B = \sum_{k=1}^{+\infty} B(k)$. □

Corollary 1. *If $\pi = (\pi_0, \pi_1, ..., \pi_K, \pi_{K+1}), \pi_i = P(\zeta_1 = i), i \in T$ and $\pi' = (\pi_2, ..., \pi_K, \pi_{K+1})$, then the total expected value $E(t)$ equals:*

$$E(t) = \pi'\tau. \tag{14}$$

Let $U, V, T \in \mathbb{R}$ be three independent random variables, $U \sim N(\mu_u, \sigma_u^2), V \sim N(\mu_v, \sigma_v^2), T \sim N(\mu_t, \sigma_t^2)$. From the properties of multivariate normal distributions (see Bilodeau and Brenner (1999)), we have:

$$f_{U+V|V}(x|y) = \frac{f_{U+V,V}(x,y)}{f_V(y)} = n_1(x; y+\mu_u, \sigma_u^2), \tag{15}$$

$$f_{U+V+T|V+T,T}(x|y,z) = f_{U+V+T|V+T}(x|y) = n_1(x; y+\mu_u, \sigma_u^2), \tag{16}$$

$$|f_U^{'}(x)| \leq \frac{e^{-1/2}}{\sigma_u^2\sqrt{2\pi}}, \quad \forall x, \tag{17}$$

$$|f_{U+V|V}^{'}(x|y)| \leq \frac{e^{-1/2}}{\sigma_u^2\sqrt{2\pi}}, \quad \forall x, y. \tag{18}$$

Lemma 1. *(Gut 2005) If X is a non-negative, integer valued random variable, then*

$$E(X) = \sum_{n=1}^{+\infty} P(X \geq n).$$

Lemma 2. *(Gut 2005) Let $r > 0$, and suppose that X is a non-negative random variable. Then the following inequalities hold:*

$$\sum_{n=1}^{+\infty} n^{r-1} P(X \geq n) \leq E(X^r) \leq 1 + \sum_{n=1}^{+\infty} n^{r-1} P(X \geq n),$$

and

$$E(|X|^r) < \infty \text{ if and only if } \sum_{n=1}^{+\infty} n^{r-1} P(X \geq n) < \infty.$$

Lemma 3. *(Coope 1996) For positive semidefinite matrices A, B of the same order*

$$0 \leq tr(AB) \leq tr(A)tr(B).$$

4. Performance and robustness analysis

4.1. Performance analysis for the hypothetical model

For model (1), (3) we have ($t \geq 1$):

$$x_t \sim N(\theta^T \psi(t); \sigma^2),\ p_t(x, \theta) = \frac{1}{\sigma\sqrt{2\pi}} exp\left\{-\frac{1}{2\sigma^2}(x - \theta^T\psi(t))^2\right\};$$

$$\lambda_t = \lambda_t(x_t) = -\frac{1}{2\sigma^2}\{2x_t(\theta^0 - \theta^1)^T\psi(t) + (\theta^1)^T\psi(t)\psi^T(t)\theta^1 - (\theta^0)^T\psi(t)\psi^T(t)\theta^0\}.$$

Due to the properties of the normal distribution, λ_t and Λ_n have also normal distributions with the following parameters:

$$E(\lambda_t) = -\frac{1}{2\sigma^2}\{2(\theta^0 - \theta^1)^T\psi(t)\psi^T(t)\theta + (\theta^1)^T\psi(t)\psi^T(t)\theta^1 - (\theta^0)^T\psi(t)\psi^T(t)\theta^0\},$$

$$D(\lambda_t) = \frac{(\theta^0 - \theta^1)^T\psi(t)\psi^T(t)(\theta^0 - \theta^1)}{\sigma^2};$$

$$E(\Lambda_n) = -\frac{1}{2\sigma^2}\{2(\theta^0 - \theta^1)^T H_n\theta + (\theta^1)^T H_n\theta^1 - (\theta^0)^T H_n\theta^0\},$$

$$D(\Lambda_n) = \frac{(\theta^0 - \theta^1)^T H_n(\theta^0 - \theta^1)}{\sigma^2},$$

where $H_n = \sum_{t=1}^{n} \psi(t)\psi^T(t)$.

Introduce the notation: $E^{(k)}(\cdot), D^{(k)}(\cdot)$ are conditional expected value and variance provided the hypothesis H_k is true ($k = 0, 1$), $\Phi(\cdot)$ is the cumulative distribution function of the standard normal law,

$$\sigma_n^2 = D^{(0)}(\lambda_n) = D^{(1)}(\lambda_n) = \frac{(\theta^0 - \theta^1)^T\psi(n)\psi^T(n)(\theta^0 - \theta^1)}{\sigma^2},$$

$$\mu_n^{(k)} = E^{(k)}(\lambda_n) = \frac{(-1)^{k+1}\sigma_n^2}{2}, \quad s_n^2 = \sum_{t=1}^{n} \sigma_t^2, \quad m_n^{(k)} = \sum_{t=1}^{n} \mu_t^{(k)} = \frac{(-1)^{k+1}s_n^2}{2};$$

$$A_n = \{a_{ij}\}_{n\times n}, \quad a_{ij} = \begin{cases} 1, & i \geq j, \\ 0, & \text{otherwise}, \end{cases}$$

$$X_n = (\lambda_1, \lambda_2, ..., \lambda_n)^T, \quad T_n = (\Lambda_1, \Lambda_2, ..., \Lambda_n)^T = A_n X_n;$$

$$\mu_{T_n}^{(k)} = E^{(k)}(T_n) = A_n E^{(k)}(X_n), \quad \Sigma_{T_n} = Cov(T_n, T_n) = A_n Cov(X_n, X_n)A_n^T;$$

$$N = \inf\{n \in \mathbb{N} : \Lambda_n \notin (C_-, C_+)\}, \quad \Gamma = (\theta^0 - \theta^1)(\theta^0 - \theta^1)^T.$$

Theorem 3. *If $tr(\Gamma H_n) \to +\infty$ as $n \to +\infty$, then test (3)-(4) terminates finitely with probability 1.*

Proof. We have:

$$s_n^2 = \sum_{t=1}^{n} \sigma_t^2 = \frac{1}{\sigma^2} \sum_{t=1}^{n} (\theta^0 - \theta^1)^T \psi(t) \psi^T(t) (\theta^0 - \theta^1)$$
$$= \frac{1}{\sigma^2} \sum_{t=1}^{n} tr\{(\theta^0 - \theta^1)^T \psi(t) \psi^T(t) (\theta^0 - \theta^1)\} = \frac{1}{\sigma^2} tr(\Gamma H_n).$$

Under the theorem condition, we get $s_n^2 \to +\infty$ as $n \to +\infty$. Furthermore, we also have $P_k(N > n) = P_k(\Lambda_i \in (C_-, C_+), i = \overline{1,n}), k = 0, 1$, and

$$P_k(\Lambda_i \in (C_-, C_+), i = \overline{1,n}) \le P_k(\Lambda_n \in (C_-, C_+)) =$$
$$= \Phi\left(\frac{C_+ - \mu_n^{(k)}}{s_n}\right) - \Phi\left(\frac{C_- - \mu_n^{(k)}}{s_n}\right)$$
$$= \Phi\left(\frac{2C_+ - (-1)^{k+1} s_n^2}{2s_n}\right) - \Phi\left(\frac{2C_- - (-1)^{k+1} s_n^2}{2s_n}\right),$$

which implies that $\lim\limits_{n \to +\infty} P_k(\Lambda_i \in (C_-, C_+), i = \overline{1,n}) = 0$ or $\lim\limits_{n \to +\infty} P_k(N > n) = 0$. Therefore, $P_k(N < +\infty) = 1 - P_k(N = +\infty) = 1 - \lim\limits_{n \to +\infty} P_k(N > n) = 1$. □

Corollary 2. *Under the conditions of Theorem 3 we get:*

$$\sum_{i=1}^{n} \sum_{j=1}^{m} \psi_j^2(i) \to +\infty \text{ as } n \to +\infty. \tag{19}$$

Additionally, if $(\theta_i^1 - \theta_i^0)^T \psi_i(t), i = \overline{1,m}$, are simultaneously nonnegative (or nonpositive) functions at t, then the result of Theorem 3 still holds if $\sum_{i=1}^{n} \psi_k^2(i) \to +\infty$, where k is such an index that $\theta_k^1 \neq \theta_k^0$.

Proof. Note that Γ and H_n are positive semi-definite matrices. The proof is derived directly from Lemma 3 and the facts that:

$$tr(H_n) = \sum_{i=1}^{n} \sum_{j=1}^{m} \psi_j^2(i), tr(\Gamma H_n) = \sum_{i=1}^{n} (\sum_{k=1}^{m} (\theta_k^0 - \theta_k^1) \psi_k(i))^2.$$

□

Remark 2. *If $\sum_{i=1}^{n} \sum_{j=1}^{m} \psi_j^2(i)$ is bounded, then there exists a positive constant L such that $s_n^2 \to L$ as $n \to +\infty$. In this case, we get $\sigma_n \to 0$ and $\mu_n^{(k)} \to 0, k = 0, 1$. It means $\lambda_n \xrightarrow{P} 0$ as $n \to +\infty$ under hypothesis H_0 or H_1. In addition, we also have:*

$$\lim_{n \to +\infty} P_k(\Lambda_n \in (C_-, C_+)) = \Phi\left(\frac{2C_+ - (-1)^{k+1} L}{2\sqrt{L}}\right) - \Phi\left(\frac{2C_- - (-1)^{k+1} L}{2\sqrt{L}}\right) > 0.$$

Theorem 4. *Under the conditions of Theorem 3 the following expressions are valid for the characteristics of test (3), (4):*

$$E^{(k)}(N) = 1 + \sum_{i=1}^{+\infty} \int_{C_-}^{C_+} ds_i \int_{C_-}^{C_+} ds_{i-1} \ldots \int_{C_-}^{C_+} n_i(s; \mu_{T_i}^{(i)}, \Sigma_{T_i}) ds_1, \; k = 0, 1, \tag{20}$$

$$\alpha = \int_{C_+}^{+\infty} n_1(s_1; \mu_1^{(0)}, \sigma_1^2) ds_1 + \sum_{i=2}^{+\infty} \int_{C_+}^{+\infty} ds_i \int_{C_-}^{C_+} ds_{i-1} \ldots \int_{C_-}^{C_+} n_i(s; \mu_{T_i}^{(0)}, \Sigma_{T_i}) ds_1, \tag{21}$$

$$\beta = \int_{-\infty}^{C_-} n_1(s_1; \mu_1^{(1)}, \sigma_1^2) ds_1 + \sum_{i=2}^{+\infty} \int_{-\infty}^{C_-} ds_i \int_{C_-}^{C_+} ds_{i-1} \ldots \int_{C_-}^{C_+} n_i(s; \mu_{T_i}^{(1)}, \Sigma_{T_i}) ds_1. \tag{22}$$

Proof. Under the condition of Theorem 4 the test terminates finitely with probability 1. From Lemma 1, we have:

$$E^{(k)}(N) = 1 + \sum_{i=1}^{+\infty} P_k(N > i) = 1 + \sum_{i=1}^{+\infty} \int_{C_-}^{C_+} ds_i \int_{C_-}^{C_+} ds_{i-1} ... \int_{C_-}^{C_+} n_i(s; \mu_{T_i}^{(k)}, \Sigma_{T_i}) ds_1.$$

For the error type I probability we get:

$$\begin{aligned} \alpha &= P_0(\Lambda_N \geq C_+) = \sum_{i=1}^{+\infty} P_0(N = i, \Lambda_N \geq C_+) \\ &= P_0(\Lambda_1 \geq C_+) + \sum_{i=2}^{+\infty} P_0(\Lambda_i \geq C_+, \Lambda_j \in (C_-, C_+), j = \overline{1, i-1}). \end{aligned} \tag{23}$$

From (23) we get (21). The expression of β in (22) is proved analogously. □

In practice, it is difficult to use formulae (20)-(22) for computing the characteristics of the test: using numerical methods for approximating the multiple integration in the right hand sides of these equalities is unfeasible. To get upper bounds for these test characteristics, we can use the following estimate:

$$P(\Lambda_1 \in (a_1, b_1), ..., \Lambda_n \in (a_n, b_n)) \leq P(\Lambda_i \in (a_i, b_i), ..., \Lambda_n \in (a_n, b_n)), \tag{24}$$

where i is a fixed value in $\{1, 2, ..., n\}$.

It is obvious that the smaller value i, the stricter the inequality (24). In particular, when $tr(\Gamma H_n)$ tends to $+\infty$ slowly, we should select the value i smaller to get better estimates.

Corollary 3. *Under the Theorem 4 condition, the following inequalities hold:*

$$\begin{aligned} E^{(k)}(N) &\leq \Phi\left(\frac{C_+ - \mu_1^{(k)}}{\sigma_1}\right) + \Phi\left(\frac{\mu_1^{(k)} - C_-}{\sigma_1}\right) \\ &\quad + \sum_{i=2}^{+\infty} \int_{C_-}^{C_+} \int_{C_-}^{C_+} n_1(x; m_{i-1}^{(k)}, s_{i-1}^2) n_1(y; x + \mu_i^{(k)}, \sigma_i^2) dx dy, k \in \{0, 1\}, \\ \alpha &\leq 1 - \Phi\left(\frac{C_+ - \mu_1^{(0)}}{\sigma_1}\right) + \sum_{i=2}^{+\infty} \int_{C_+}^{+\infty} \int_{C_-}^{C_+} n_1(x; m_{i-1}^{(0)}, s_{i-1}^2) n_1(y; x + \mu_i^{(0)}, \sigma_i^2) dx dy, \\ \beta &\leq \Phi\left(\frac{C_- - \mu_1^{(1)}}{\sigma_1}\right) + \sum_{i=2}^{+\infty} \int_{-\infty}^{C_-} \int_{C_-}^{C_+} n_1(x; m_{i-1}^{(1)}, s_{i-1}^2) n_1(y; x + \mu_i^{(1)}, \sigma_i^2) dx dy. \end{aligned}$$

Proof. The inequalities are resulting from (24), (15) with $i = n - 1$. □

Time series data is usually collected at certain intervals. Sometimes, there are patterns that can repeat over fixed periods of time within the data set. Such patterns are known as periodic fluctuations or seasonality. In this case, the function of trend $g(t) = \theta^T \psi(t)$ will be periodic with some period $T > 0$. In particular, function $h(t) = (\theta^0 - \theta^1)^T \psi(t)$ will also be periodic.

Theorem 5. *If there exists an integer $T \geq 1$ such that $h(t + T) = h(t), \forall t = 1, 2, ...$, then the stopping time N has finite moments of any order.*

Proof. Without loss of generality, assume that hypothesis H_0 is true. Due to the Theorem conditions, λ_t and λ_{t+kT} have the same distribution for all $t, k \in \mathbb{N}$.

On event $\{N > kT\}, k \geq 1$, $\Lambda_{kT} = \lambda_1 + ... + \lambda_{kT} \in (C_-, C_+)$, which implies that $\Lambda_{kT}^2 \leq (C_+ - C_-)^2$. Since $D(\Lambda_{kT}) = \sum_{i=1}^{kT} D(\lambda_i) = kD(\Lambda_T) = \frac{k}{\sigma^2} tr(\Gamma H_T) \to +\infty$ as $k \to +\infty$, there exists $k_0 \geq 1$ such that on $\{N > k_0 T\}$:

$$P_0(\Lambda_{k_0 T}^2 < (C_+ - C_-)^2) = q \in (0, 1).$$

Put $n_0 = hk_0T, h \geq 1$. Note that from the fact $N > n_0$, we have

$$|\Lambda_{rk_0T} - \Lambda_{(r-1)k_0T}| < C_+ - C_-, \forall r = \overline{1,h}.$$

Because $\{\Lambda_{rk_0T} - \Lambda_{(r-1)k_0T}, r = \overline{1,h}\}$ is a sequence of independent identically distributed random variables and $\{N > n_0\} \subset \{|\Lambda_{rk_0T} - \Lambda_{(r-1)k_0T}| < C_+ - C_-, \forall r = \overline{1,h}\}$, we have

$$P_0(N > n_0) \leq P_0(|\Lambda_{rk_0T} - \Lambda_{(r-1)k_0T}| < C_+ - C_-, \forall r = \overline{1,h}) = q^h = q^{\frac{n_0}{k_0T}}.$$

Therefore, for any $n > k_0T$

$$P_0(N > n) \leq q^{\left[\frac{n}{k_0T}\right]} \leq q^{\frac{n}{k_0T}-1},$$

which implies $P_0(N > n) = o(n^{-r})$ as $n \to +\infty$ for any finite r.
The rest part of proof is derived from Lemma 2. □

Corollary 4. *If the basic vector function of trend $\psi(t)$ is periodic on the set $\mathbb{N}$, then the stopping time N has finite moments of any order.*

4.2. Special case

Assume that there exists a constant $a \neq 0$ such that $h(t) = a, \forall t \geq 1$, and H_0 is the true hypothesis.

In this case, $\{\lambda_t, t \geq 1\}$ becomes the sequence of independent and identically distributed random variables from $N(\mu, \sigma_0^2)$, where

$$\mu = \frac{-a^2}{2\sigma^2}, \quad \sigma_0^2 = \frac{a^2}{\sigma^2}.$$

Let x be a fixed value and put $\Lambda_n^x = x + \Lambda_n$. The new test based on Λ_n^x is equivalent to the original SPRT whose region of indifference is the interval $(C_- - x, C_+ - x)$. Let $\beta_\theta(x)$ and $N_\theta(x)$ be the operating characteristic and average sample size functions of this SPRT respectively. From the Markov property of log-likelihood ratio statistic Λ_n and when H_0 is true, $\beta_{\theta^0}(x)$ and $N_{\theta^0}(x)$ are known to satisfy the Fredholm integral equations (see Basseville and Nikiforov (1993), Cox and Miller (1965)):

$$\beta_{\theta^0}(x) = F_{\theta^0}(C_- - x) + \int_{C_-}^{C_+} K_{\theta^0}(x,y)\beta_{\theta^0}(y)dy,$$

$$N_{\theta^0}(x) = 1 + \int_{C_-}^{C_+} K_{\theta^0}(x,y)N_{\theta^0}(y)dy,$$

where $F_{\theta^0}(x) = P_{\theta_0}(\lambda_1 < x) = \Phi\left(\frac{x-\mu}{\sigma_0}\right)$ and $K_{\theta^0}(x,y) = \frac{\partial}{\partial y}F_{\theta^0}(y-x) = \frac{1}{\sigma_0}\varphi\left(\frac{y-x-\mu}{\sigma_0}\right); \varphi(z) = n_1(z;0,1), z \in \mathbb{R}$.

A numerical method is used for solving these equations. Let m_0 be a positive integer and $\{y_i, i = \overline{1,m_0}\}, C_- = y_1 < y_2 < ... < y_{m_0} = C_+$, be a partition of the interval $[C_-, C_+]$, in which y_i having the smallest absolute value is set to be 0. Note that $\alpha = 1 - \beta_{\theta^0}(0)$ and $E(N|H_0) = N_{\theta^0}(0)$. Let $\tilde{\beta}_{\theta^0}(y_i)$ be approximations of $\beta_{\theta^0}(y)$ at $y = y_i, i = \overline{1,m_0}$. Using the trapezoid formula (see Hoffman (2001)) we have:

$$\begin{aligned}\int_{C_-}^{C_+} K_{\theta^0}(x,y)\beta_{\theta^0}(y)dy &\approx \sum_{i=1}^{m_0-1} \frac{y_{i+1}-y_i}{2\sigma_0}\left(\varphi\left(\frac{y_i-x-\mu}{\sigma_0}\right)\tilde{\beta}_{\theta^0}(y_i) + \right.\\ &\quad \left. + \varphi\left(\frac{y_{i+1}-x-\mu}{\sigma_0}\right)\tilde{\beta}_{\theta^0}(y_{i+1})\right)\\ &\approx \sum_{i=1}^{m_0} \rho_i\varphi\left(\frac{y_i-x-\mu}{\sigma_0}\right)\tilde{\beta}_{\theta^0}(y_i),\end{aligned}$$

where

$$\rho_i = \begin{cases} \frac{y_2-y_1}{2\sigma_0}, & i=1, \\ \frac{y_{m_0}-y_{m_0-1}}{2\sigma_0}, & i=m_0, \\ \frac{y_{i+1}-y_{i-1}}{2\sigma_0}, & i \in \{2,3,...,m_0-1\}. \end{cases}$$

In the case of uniform partition, the error of this approximation is $O(h^2)$, where $h = y_{i+1} - y_i$ (see Hoffman (2001)). From the approximation above we get the following system of linear equations:

$$\begin{cases} \tilde{\beta}_{\theta^0}(y_1) & = \Phi\left(\frac{C_- - y_1 - \mu}{\sigma_0}\right) + \sum_{i=1}^{m} \rho_i \varphi\left(\frac{y_i - y_1 - \mu}{\sigma_0}\right) \tilde{\beta}_{\theta^0}(y_i), \\ \quad \vdots & \quad \vdots \qquad\qquad \vdots \qquad\qquad\qquad \vdots \\ \tilde{\beta}_{\theta^0}(y_{m_0}) & = \Phi\left(\frac{C_- - y_{m_0} - \mu}{\sigma_0}\right) + \sum_{i=1}^{m_0} \rho_i \varphi\left(\frac{y_i - y_{m_0} - \mu}{\sigma_0}\right) \tilde{\beta}_{\theta^0}(y_i). \end{cases} \tag{25}$$

This system can be represented in the matrix form:

$$(I_{m_0} - A) \cdot \tilde{\beta}_{\theta^0} = B,$$

where

$$A = \{a_{ij}\}_{m_0 \times m_0}, a_{ij} = \rho_j \varphi\left(\frac{y_j - y_i - \mu}{\sigma_0}\right),$$
$$\tilde{\beta}_{\theta^0} = (\tilde{\beta}_{\theta^0}(y_1), ..., \tilde{\beta}_{\theta^0}(y_{m_0}))^T,$$
$$B = \left(\Phi\left(\frac{C_- - y_1 - \mu}{\sigma_0}\right), ..., \Phi\left(\frac{C_- - y_{m_0} - \mu}{\sigma_0}\right)\right).$$

Similarly, we also have the system of linear equations for the approximations $\tilde{N}_{\theta^0}(y_i)$ of $N_{\theta^0}(y)$ at $y = y_i, i = \overline{1, m_0}$:

$$\begin{cases} \tilde{N}_{\theta^0}(y_1) & = 1 + \sum_{i=1}^{m_0} \rho_i \varphi\left(\frac{y_i - y_1 - \mu}{\sigma_0}\right) \tilde{N}_{\theta^0}(y_i), \\ \quad \vdots & \quad \vdots \qquad\qquad \vdots \qquad\qquad\qquad \vdots \\ \tilde{N}_{\theta^0}(y_{m_0}) & = 1 + \sum_{i=1}^{m_0} \rho_i \varphi\left(\frac{y_i - y_{m_0} - \mu}{\sigma_0}\right) \tilde{N}_{\theta^0}(y_i). \end{cases} \tag{26}$$

This system is also represented in the matrix form:

$$(I_{m_0} - A) \cdot \tilde{N}_{\theta^0} = \bar{B},$$

where $\tilde{N}_{\theta^0} = (\tilde{N}_{\theta^0}(y_1), ..., \tilde{N}_{\theta^0}(y_{m_0}))^T$ and $\bar{B}$ is a column vector of size m_0 whose all elements are equal to one.

4.3. Approximation of the random sequence Λ_n

Let us split the state space of Λ_n on $K+2$ cells:

$$A_0 = (-\infty, C_-),\ A_i = [C_{i-1}, C_i),\ i = \overline{1, K},\ A_{K+1} = [C_+, +\infty),$$

$$C_- = C_0 < C_1 < C_2 < ... < C_K = C_+,\ C_i = C_- + ih,\ h = \frac{C_+ - C_-}{K}, i = \overline{1, K}.$$

Denote $S_K = \{A_1, A_2, ..., A_K\}$.

Lemma 4. *If* $\inf_n tr(\Gamma\psi(n)\psi^T(n)) \geq C, C = const > 0$, *then for all* $i, j, k, i < j < k$ *the following asymptotic expansions are valid:*

$$P(\Lambda_k \in A^1 | \Lambda_j \in A^2, \Lambda_i \in A^3) = P(\Lambda_k \in A^1 | \Lambda_j \in A^2) + R(h), \tag{27}$$

where $A^1, A^2, A^3 \in \{A_0, A_1, ..., A_K, A_{K+1}\}$, *and* $R(h) = \begin{cases} O(h), & A^1 \in \{A_0, A_{K+1}\}, \\ O(h^2), & A^1 \in S_K. \end{cases}$

Proof. From the Lemma condition and (17)-(18), there exist two positive constants C_1, C_2 such that:

$$|f'_{\Lambda_n}(x)| \le C_1, \forall n \ge 1, |f'_{\Lambda_n|\Lambda_k}(x|y)| \le C_2, \forall k < n, \forall x, y \in \mathbb{R}.$$

The rest part of proof is similar to the proof of Theorem 1 in Kharin (2008). □

Denote $f_{C_-}^{C_+}(x) = \left(\left[\frac{x-C_-}{h}\right] + 1\right) \cdot \mathbf{1}_{(C_-,C_+)}(x) + (K+1) \cdot \mathbf{1}_{[C_+,+\infty)}(x)$.
For the random sequence Λ_n introduce the discrete random sequence Z_n with the finite state space $V = \{0, 1, ..., K+1\}$. Put $Z_1 = f_{C_-}^{C_+}(\Lambda_1)$. For $n \ge 2$:

$$Z_n = \begin{cases} 0, & \text{if } Z_{n-1} = 0, \\ K+1, & \text{if } Z_{n-1} = K+1, \\ f_{C_-}^{C_+}(\Lambda_n), & \text{otherwise.} \end{cases}$$

To simplify the notation, let us renumerate the states space of Z_n:

$$V = \{\{0\}, \{K+1\}, \{1\}, ..., \{K\}\}.$$

Denote:

$$P^{(n)}(\theta^k) = \left(\begin{array}{c|c} I_2 & \mathbf{O}_{2\times K} \\ \hline R_n(\theta^k) & Q_n(\theta^k) \end{array}\right), k \in \{0,1\},$$

$$Q_n(\theta^k) = \{q_{ij}^{(n)}(\theta^k)\}_{K\times K}, R_n(\theta^k) = \{r_{ij}^{(n)}(\theta^k)\}_{K\times 2},$$

$$q_{ij}^{(n)}(\theta^k) = \frac{\int_{A_i} n_1(y; m_{n-1}^{(i)}, s_{n-1}^2) \int_{A_j} n_1(x; y + \mu_n^{(k)}, \sigma_n^2) dxdy}{\int_{A_i} n_1(y; m_{n-1}^{(k)}, s_{n-1}^2) dy}, 1 \le i, j \le K,$$

$$r_{i1}^{(n)}(\theta^k) = \frac{\int_{A_i} n_1(y; m_{n-1}^{(k)}, s_{n-1}^2) \int_{A_0} n_1(x; y + \mu_n^{(k)}, \sigma_n^2) dxdy}{\int_{A_i} n_1(y; m_{n-1}^{(k)}, s_{n-1}^2) dy}, 1 \le i \le K,$$

$$r_{i2}^{(n)}(\theta^k) = \frac{\int_{A_i} n_1(y; m_{n-1}^{(k)}, s_{n-1}^2) \int_{A_{K+1}} n_1(x; y + \mu_n^{(k)}, \sigma_n^2) dxdy}{\int_{A_i} n_1(y; m_{n-1}^{(k)}, s_{n-1}^2) dy}, 1 \le i \le K,$$

$$S(\theta^k) = I_K + \sum_{i=1}^{+\infty} \prod_{j=1}^{i+1} Q_j(\theta^k), \quad B(\theta^k) = R_2(\theta^k) + \sum_{i=2}^{+\infty} \prod_{j=1}^{i} Q_j(\theta^k) R_{i+1}(\theta^k),$$

$B_{(j)}(\cdot)$ is the j^{th} column of matrix $B(\cdot), \pi(\theta^k)$ is the probability distribution of $Z_1, k \in \{0,1\}$.

Theorem 6. *If* $\exists C = const > 0, \inf_n tr(\Gamma\psi(n)\psi^T(n)) \ge C$, *then the characteristics of the test* (3),(4) *satisfy the following expansions:*

$$E^{(k)}(N) = 1 + (\pi(\theta^k))' S(\theta^k) \cdot \mathbf{1}_K + O(h), k \in \{0,1\}, \tag{28}$$

$$\alpha = (\pi(\theta^0))' B_{(2)}(\theta^0) + \pi_{K+1}(\theta^0) + O(h), \tag{29}$$

$$\beta = (\pi(\theta^1))' B_{(1)}(\theta^1) + \pi_0(\theta^1) + O(h), \tag{30}$$

where $(\pi(\theta^k))' = (\pi_1(\theta^k), ..., \pi_K(\theta^k))$.

Proof. Under the conditions of Theorem 6 sequence Z_n satisfies the conditions (5)-(7) asymptotically as $h \to 0$. The results of this theorem are derived from the Lemma 4, Theorems 1 and 2. □

4.4. Robustness evaluation

In practice the observed data can often come from more complicated sources than hypothetical ones because of the distortion. Some contamination models (see Huber (1981)) can be used

to analyze the robustness of statistical procedures. However, in this paper we consider only the case where the noise components ξ_t in model (1) are distorted by another noises in the Gaussian distribution family.

Suppose that the observed data come from the following mixed model:

$$\bar{x}_t = \theta^T \psi(t) + (1-\varepsilon)\xi_t + \varepsilon\tilde{\xi}_t, \tag{31}$$

where $\{\tilde{\xi}_t, t \geq 1\}$ is a sequence of independent identically distributed random variables, $\tilde{\xi}_t \sim N(0, \tilde{\sigma}^2)$, $\tilde{\sigma}$ is a given positive constant, ξ_t and $\tilde{\xi}_t$ are independent for all t, and $\varepsilon \in [0, 1/2)$ is a level of distortion.

Introduce the notation: $\bar{\lambda}_t = \lambda_t(\bar{x}_t), \bar{\mu}_t^{(k)} = E^{(k)}(\bar{\lambda}_t), \bar{\sigma}_t^2 = D(\bar{\lambda}_t), \bar{m}_n^{(k)} = \sum_{i=1}^n \bar{\mu}_i^{(k)}, \bar{s}_n^2 = \sum_{i=1}^n \bar{\sigma}_i^2$, and $\bar{P}^{(n)}(\theta^k), \bar{R}_n(\theta^k), \bar{Q}_n(\theta^k), \bar{\pi}(\theta^k), \bar{\alpha}, \bar{\beta}, \bar{t}(\theta^k)$ are calculated analogously by replacing x_t with $\bar{x}_t$.

For $t, n \geq 1, k \in \{0, 1\}$, we get:

$$\bar{\mu}_t^{(k)} = \mu_t^{(k)}, \bar{m}_n^{(k)} = m_n^{(k)}, \bar{\sigma}_t^2 = \frac{\sigma_t^2}{\sigma^2}[(1-\varepsilon)^2\sigma^2 + \varepsilon^2\tilde{\sigma}^2], \bar{s}_n^2 = \frac{s_n^2}{\sigma^2}[(1-\varepsilon)^2\sigma^2 + \varepsilon^2\tilde{\sigma}^2]. \tag{32}$$

Theorem 7. *For the contaminated model (31), under the conditions of Theorem 6 the following asymptotic expansions hold as $\varepsilon \to 0$ and $h \to 0$:*

$$\bar{\alpha} = \alpha + O(\varepsilon) + O(h), \tag{33}$$

$$\bar{\beta} = \beta + O(\varepsilon) + O(h), \tag{34}$$

$$\bar{t}(\theta^k) = t(\theta^k) + O(\varepsilon) + O(h), k \in \{0, 1\}. \tag{35}$$

Proof. From (32), as $\varepsilon \to 0$ we have:

$$\frac{1}{\bar{\sigma}_t^2} = \frac{1}{\sigma_t^2}(1 + O(\varepsilon)), \quad \frac{1}{\bar{s}_n^2} = \frac{1}{s_n^2}(1 + O(\varepsilon)),$$

$$n_1(x; \bar{\mu}_1^{(k)}, \bar{\sigma}_1^2) = \frac{1 + O(\varepsilon)}{\sigma_1\sqrt{2\pi}} \exp\left[-\frac{(x - \mu_1^{(k)})^2}{2\sigma_1^2}(1 + O(\varepsilon))\right]$$
$$= (1 + O(\varepsilon))n_1(x; \mu_1^{(k)}, \sigma_1^2), \forall x \in \mathbb{R},$$

which implies that

$$\bar{\pi}_i(\theta^k) = \pi_i(\theta^k) + O(\varepsilon), i = \overline{0, K+1}, k \in \{0, 1\}. \tag{36}$$

Similarly, we also have as $\varepsilon \to 0$

$$n_1(y; \bar{m}_{n-1}^{(k)}, \bar{s}_{n-1}^2)n_1(x; y + \bar{\mu}_n^{(k)}, \bar{\sigma}_n^2) = (1 + O(\varepsilon))^2 n_1(y; m_{n-1}^{(k)}, s_{n-1}^2)n_1(x; y + \mu_n^{(k)}, \sigma_n^2),$$

and obtain

$$\bar{r}_{ij}^{(n)}(\theta^k) = r_{ij}^{(n)}(\theta^k) + O(\varepsilon), 1 \leq i \leq K, j = 1, 2, \tag{37}$$

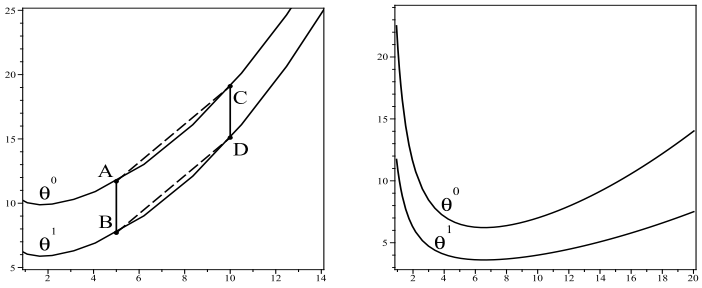
$$\bar{q}_{ij}^{(n)}(\theta^k) = q_{ij}^{(n)}(\theta^k) + O(\varepsilon), 1 \leq i, j \leq K. \tag{38}$$

Combining (36)-(38) and using Theorem 6 we get (33)-(35). □

5. Numerical examples

First consider the probability model (1) for the special case from Section 4.2 with the following parameters values: $m = 4, \sigma = 10, \psi(t) = (1 + 1/t, 4 - t/10, t/10, t^2/10)^T, \theta^0 =$

$(1,2,2,1)^T, \theta^1 = (1,1,1,1)^T$; hypotheses (2) were tested. Denote the sample estimate of a characteristic γ with Monte-Carlo method by $\hat{\gamma}$. The number of simulation runs used in this method was 100 000. Denote $t_k = E^{(k)}(N), k = \overline{0,1}$.



(a) (b)

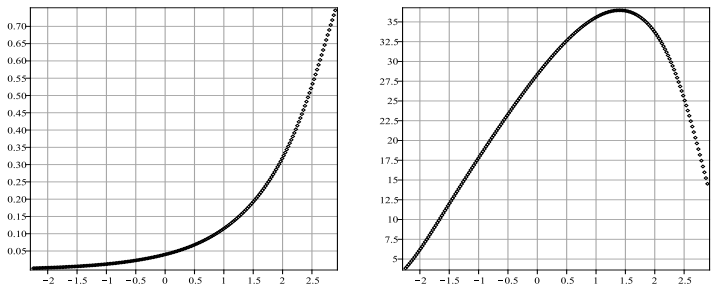
Figure 1: Trend functions

Suppose that hypothesis H_0 is true. Then we have $a = (\theta^0 - \theta^1)^T \psi(t) = 4, \forall t$. This means that the lengths of segments AB and CD are always the same in all positions such that they are parallel to the vertical axis (figure 1a). In addition, $\lambda_t, t \geq 1$, are independent identically distributed random variables, $\lambda_t \sim N(\mu, \sigma_0^2)$, where $\mu = -50, \sigma_0 = 0.4$. Monte-Carlo estimates ($\hat{\alpha}$ and $\hat{t}_0$) and approximate values ($\tilde{\alpha} = 1 - \beta_{\theta^0}(0)$ and $\tilde{t}_0 = \tilde{N}_{\theta^0}(0)$) calculated according to Section 4.2 for Type I error probability α and conditional average number of observations t_0 respectively are presented in Table 1. When the value of m_0 increases, the approximate values of test characteristics tend to their corresponding Monte-Carlo estimates.

Table 1: Performance characteristics estimates

α_0	β_0	$\hat{\alpha}$	$\hat{t}_0$	m_0	$\tilde{\alpha}$	$\tilde{t}_0$
0.1	0.1	0.08034	25.73666	200	0.08236	25.70478
				500	0.08076	25.72103
0.05	0.1	0.04072	28.46724	200	0.04128	28.46644
				500	0.04080	28.47073
0.05	0.05	0.03990	36.83060	200	0.03816	36.83469
				500	0.03974	36.80202

The dependence of the operating characteristic and the average sample size functions on the initial value x in the modified test is presented in figure 2 for the case of $m_0 = 200, \alpha_0 = 0.05, \beta_0 = 0.1$. Under hypothesis H_0, $\beta_{\theta^0}(x)$ is a decreasing function with respect to x (figure 2a). This fact is easily understood because when x increases, the probability that $x + \Lambda_n$ comes out of the interval $(C_- - x, C_+ - x)$ through the upper boundary $C_+ - x$ also increases. However, function $N_{\theta^0}(x)$ increases to the maximum value in the interval (C_-, C_+) before dropping (figure 2b).



(a) (b)

Figure 2: Plots of functions $1 - \beta_{\theta^0}(x)$ and $N_{\theta^0}(x)$

In the next examples, the following values of parameters are used for calculating: $m = 4, \sigma = 10, \psi(t) = (1, t/10, t^2/100, 10/t)^T, \theta^0 = (1,2,2,2)^T, \theta^1 = (1,1,1,1)^T$. In this case, $P_k(\Lambda_{50} \in (C_-, C_+)) \leq 10^{-6}, k \in \{0,1\}$. All infinite sums were replaced by the sums of the first 50 summands, this provides the accuracy of the order 0.00001. Figure 1b shows the plots of trend functions. The upper bounds for the test performance characteristics constructed in Corollary 3 are given in Tables 2 and 3.

Table 2: The upper bounds for error probabilities

α_0	β_0	$\hat{\alpha}$	$\alpha \leq$	$\hat{\beta}$	$\beta \leq$
0.1	0.1	0.07166	0.12155	0.07244	0.12155
0.05	0.1	0.03392	0.05307	0.07006	0.11402
0.01	0.05	0.00534	0.00761	0.03216	0.04734

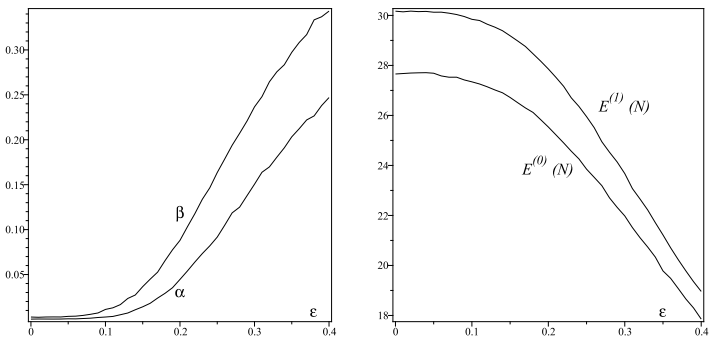
Table 3: The upper bounds for the average number of observations

α_0	β_0	$\hat{t}_0$	$E^{(0)}(N) \leq$	$\hat{t}_1$	$E^{(1)}(N) \leq$
0.1	0.1	16.07760	18.09357	16.03320	18.09357
0.05	0.1	17.03316	18.68805	19.70408	21.30107
0.01	0.05	21.13740	22.13213	25.61644	26.41215

Denote the main terms of asymptotic expansions of $\alpha, \beta, E^{(k)}(N)$ respectively by $\bar{\alpha}_{ASYM}$, $\bar{\beta}_{ASYM}$, $\bar{E}^{(k)}_{ASYM}(N)$. The numerical results for these main terms are presented in Table 4.

Table 4: The main terms of asymptotic expansions of the test characteristics

α_0	β_0	K	$\bar{\alpha}_{ASYM}$	$\bar{\beta}_{ASYM}$	$\bar{E}^{(0)}_{ASYM}(N)$	$\bar{E}^{(1)}_{ASYM}(N)$
0.1	0.1	10	0.07449	0.07449	15.85942	15.85942
		20	0.07273	0.07273	16.03470	16.03470
		40	0.07225	0.07225	16.08029	16.08029
0.05	0.1	10	0.03501	0.07367	16.76240	19.49727
		20	0.03433	0.07104	16.97260	19.65483
		40	0.03414	0.07033	17.02777	19.69613
0.01	0.05	10	0.00586	0.03466	20.83295	25.57559
		20	0.00579	0.03291	21.06344	25.65018
		40	0.00576	0.0324	21.12768	25.67067



(a) (b)

Figure 3: Dependence of performance characteristics on probability of contamination ε

Due to symmetric property, we have $\bar{\alpha}_{ASYM} = \bar{\beta}_{ASYM}, \bar{E}^{(0)}_{ASYM}(N) = \bar{E}^{(1)}_{ASYM}(N)$ provided $\alpha_0 = \beta_0$. When the value K increases, the main terms of asymptotic expansions of the test characteristics become closer to their corresponding Monte-Carlo estimates (see Tables 2 and 3). The orders of approximation in (28)-(30) are only $O(h)$. Therefore, if we want to make the main terms of asymptotic expansions better, the value K must be larger. However, with the large value K the computation on infinite sums (in practice, they can be reasonably replaced with finite ones because of the termination of the test) of matrices with high dimensions $S(\theta^i)$ and $B(\theta^i)$ will cost much time.

Figure 3 shows the dependence of the error probabilities and average number of observations on the probability of contamination ε in the model (31), when $\tilde{\sigma}^2 = 50\sigma^2, \alpha_0 = 0.001, \beta_0 = 0.005$. When the contamination probability ε increases, both error probabilities increase. For both conditional average numbers of observations, there are opposite pictures.

6. Conclusion

The problem of sequential testing for time series with trend is studied. The sufficient condition of termination of the test is given. Beside the explicit (but not useful for further analysis) formulae for the test characteristics, an approach to approximate test characteristics is also constructed. This approach allows us not only to estimate the test characteristics, but also to analyze the robustness of the test.

References

Basseville M, Nikiforov I (1993). *Detection of Abrupt Changes: Theory and Application*. Prentice Hall.

Bilodeau M, Brenner D (1999). *The Theory of Multivariate Statistics*. Springer-Verlag.

Coope I (1996). "On Matrix Trace Inequalities and Related Topics for Products of Hermitian Matrices." *Journal of Mathematical Analysis and Applications*, **188**, 999–1001.

Cox D, Miller H (1965). *The Theory of Stochastic Processes*. John Wiley and Sons.

Govindarajulu Z (2004). *Sequential Statistics*. World Scientifc.

Gut A (2005). *Probability: A Graduate Course*. Springer Science-Business Media Inc.

Hoffman J (2001). *Numerical Methods for Engineers and Scientists*. Marcel Dekker Inc.

Huber P (1981). *Robust Statistics*. John Wiley and Sons.

Kharin A (2005). "Robust Bayesian Prediction under Distortions of Prior and Conditional Distributions." *Journal of Mathematical Sciences*, **126**(1), 992–997.

Kharin A (2008). "Robustness Evaluation in Sequential Testing of Composite Hypotheses." *Austrian Journal of Statistics*, **37**(1), 51–60.

Kharin A (2011). "Robustness Analysis for Bayesian Sequential Testing of Composite Hypotheses under Simultaneous Distortion of Priors and Likelihoods." *Austrian Journal of Statistics*, **40**(1&2), 65–73.

Kharin A (2013). "Robustness of Sequential Testing of Hypotheses on Parameters of M-valued Random Sequences." *Journal of Mathematical Sciences*, **189**(6), 924–931.

Kharin A (2016). "Performance and Robustness Evaluation in Sequential Hypotheses Testing." *Communications in Statistics - Theory and Methods*, **45**(6), 1693–1709.

Kharin A, Kishylau D (2015). "Robust Sequential Test for Hypotheses about Discrete Distributions in the Presence of "Outliers"." *Journal of Mathematical Sciences*, **205**(1), 68–73.

Kharin A, Ton T (2016). "Sequential Statistical Hypotheses Testing on Parameters of Time Series with a Trend under Missing Values." *Proceedings of the National Academy of Sciences of Belarus. Series of Physical-Mathematical Sciences*, (3), 38–46.

Kharin Y (1997). "Robustness of Clustering under Outliers." *Lecture Notes in Computer Science*, **1280**, 501–512.

Maevskii V, Kharin Y (2002). "Robust Regressive Forecasting under Functional Distortions in a Model." *Automation and Remote Control*, **63**(11), 1803–1820.

Wald A (1947). *Sequential Analysis*. John Wiley and Sons.

Wald A, Wolfowitz J (1948). "Optimum Character of the Sequential Probability Ratio Test." *The Annals of Mathematical Statistics*, **19**(3), 326–339.

Affiliation:

Alexey Kharin, Ton That Tu
Department of Probability Theory and Mathematical Statistics
Faculty of Applied Mathematics and Informatics
Belarusian State University
Independence avenue, 4
Minsk 220030
Belarus
E-mail: KharinAY@bsu.by

Ton That Tu
Faculty of Mathematics
Da Nang University of Education
Ton Duc Thang street, 459
Da Nang 555726
Vietnam
tthattu@gmail.com

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/

Volume 46 *Submitted:* 2016-11-15
April 2017 *Accepted:* 2017-02-02

Austrian Journal of Statistics
April 2017, Volume 46, 37–45.
http://www.ajs.or.at/
doi:10.17713/ajs.v46i3-4.669

High-order Vector Markov Chain with Partial Connections in Data Analysis

Yuriy Kharin
Belarusian State University,
Department of Mathematical
Modelling and Data Analysis

Michail Maltsew
Belarusian State University,
Research Institute
for Applied Problems
of Mathematics and Informatics

Abstract

A new mathematical model for discrete time series is proposed: homogenous vector Markov chain of the order s with partial connections. Conditional probability distribution for this model is determined only by a few components of previous vector states. Probabilistic properties of the model are given: ergodicity conditions and conditions under which the stationary probability distribution is uniform. Consistent statistical estimators for model parameters are constructed.

Keywords: vector Markov chain with partial connections, ergodicity conditions, statistical estimation of parameters.

1. Introduction

Markov chain is a wide used mathematical model for discrete time series. It is applied in economics (Kemeny and Snell 1963), biology (Waterman 1999), sociology (Bonacich 2003) and in other fields. Markov chain of the order s (Doob 1953) is an adequate model for description of high-depth dependences in data. In practice data is often represented in time-indexed blocks, and it is reasonable to use vector Markov chains. The state space for such models consists of m-vectors for some finite value m. Unfortunately, it is difficult to use s-order Markov chain in practice, because the number of parameters D for this model increases exponentially when s grows. That is why small-parametric (parsimonious) models are used in applications (Kharin 2012). For such models D depends polynomially on s. Markov chain of order s with r partial connections (MC(s, r)) is an example of a parsimonious model for the univariate case $m = 1$. It was developed in Belarusian state university (Kharin and Petlitskii 2007). Conditional distribution for this model doesn't depend on all s previous states but only on r selected states. In this paper we propose a generalization of the MC(s, r) for high-order vector Markov chain. Note in addition that parsimonious models are also useful in robust statistical analysis (Kharin 2016; Kharin and Kishylau 2015; Kharin and Shlyk 2009; Kharin and Zhuk 1998; Kharin 1997).

2. Vector Markov chain with partial connections

2.1. Mathematical model

Introduce the notation: $\mathbf{N}$ is the set of positive integers; $A = \{0, 1, \dots, N-1\}$ is the discrete set with N elements, $2 \le N < \infty$; $m \in \mathbf{N}$, $J_i = (j_{i1}, \dots, j_{im}) \in A^m$, $i = 1, 2, \dots$, is an m-dimensional vector; $J_a^b = (J_a, \dots, J_b)$, $a, b \in \mathbf{N}$, $a \le b$, is a sequence of $b - a + 1$ ordered m-dimensional vectors;

$$x_t = (x_{t1}, \dots, x_{tm}) \in A^m, t \in \mathbf{N}$$

is a homogeneous vector Markov chain of the order s ($2 \le s < \infty$) with the state space A^m, with some initial probability distribution

$$\pi^{(0)}_{J_1,\dots,J_s} = \mathrm{P}\{x_1 = J_1, \dots, x_s = J_s\}, \tag{1}$$

and some $(s+1)$-dimensional matrix of transition probabilities:

$$\begin{gathered} P = (p_{J_1^s, J_{s+1}}), \\ p_{J_1^s, J_{s+1}} = \mathrm{P}\{x_t = J_{s+1} | x_{t-1} = J_s, \dots, x_{t-s} = J_1\}, t = s+1, s+2, \dots. \end{gathered} \tag{2}$$

We will denote this Markov chain VMC(s) (Vector Markov Chain of the order s).

The number of independent parameters for the VMC(s) is determined by formula:

$$D_s = N^{ms}(N^m - 1). \tag{3}$$

In Table 1 we present the number of parameters for the binary VMC(s) when $m = 8$ for different values of s .

Table 1: The number of parameters for the binary VMC(s)

s	1	2	4	8	16
D_s	65 280	16 711 680	$\approx 1,095 \cdot 10^{12}$	$\approx 4,704 \cdot 10^{21}$	$\approx 8,677 \cdot 10^{40}$

Table 1 illustrates the "curse of dimensionality" for the s-order Markov chain. To overcome this difficulty we construct a modification of the VMC(s) similarly to the paper (Kharin and Petlitskii 2007).

We will use the notation:

$$M_r = \{(k_1, l_1), (k_2, l_2), \dots, (k_r, l_r)\} \subseteq M_* = \{(k, l) : 1 \le k \le s, 1 \le l \le m\} \tag{4}$$

is an ordered set of $1 \le r \le sm$ pairs of indices, $k_1 = 1$, which we will call a template-set; $\mathbf{M}_r$ is a set of all possible template-sets;

$$S_{M_r}(J_t, \dots, J_{t+s-1}) = (j_{t+k_1-1,l_1}, \dots, j_{t+k_r-1,l_r}),\ t = 1, 2, \dots$$

is a selector function, that associates s vectors with their r components: $S_{M_r} : A^{ms} \to A^r$; $Q = (q_{(i_1,\dots,i_r),I_{r+1}})$ is a stochastic $N^r \times N^m$ matrix, $i_1, \dots, i_r \in A$, $I_{r+1} \in A^m$.

The Markov chain $\{x_t \in A^m : t \in \mathbf{N}\}$ is called the vector Markov chain of the order s with r partial connections if its transition probabilities have the following form:

$$p_{J_1^s, J_{s+1}} = q_{S_{M_r}(J_1,\dots,J_s),J_{s+1}} = q_{(j_{k_1,l_1},\dots,j_{k_r,l_r}),J_{s+1}},\ J_1, \dots, J_s, J_{s+1} \in A^m. \tag{5}$$

We will denote this model VMC(s, r). The m-dimensional VMC(s, r) is determined by the following parameters:

- $s \ge 1$ is the order of the Markov chain;

- $r \in \{1, \ldots, sm\}$ is the number of connections;
- M_r is the template-set of connections;
- $Q = (q_{(i_1,\ldots,i_r),I_{r+1}})$ is a stochastic $N^r \times N^m$-matrix, $(i_1, \ldots, i_r) \in A^r$, $I_{r+1} \in A^m$.

The definition (5) of the VMC(s, r) means that the probability distribution of time series x_t at time point t depends not on all ms components of s previous states, but it depends only on r selected components determined by the template-set M_r. If $r = sm$, then $M_r = M_*$ and we have fully-connected s-order Markov chain: VMC$(s, ms) \equiv$ VMC(s). If $m = 1$, then the VMC(s, r) transforms into the Markov chain with partial connections (Kharin and Petlitskii 2007).

The number of independent parameters for the VMC(s, r) is determined by formula:

$$d = N^r(N^m - 1) + 2r - 1. \tag{6}$$

In Table 2 we present the number of parameters for the binary VMC(s, r) when $N = 2$, $m = 8$ for different values of s and r .

Table 2: The number of parameters for the binary VMC(s, r)

(s, r)	(1, 2)	(2, 4)	(4, 6)	(8, 8)	(16, 10)	(32, 16)
d	1 023	4 087	16 331	65 295	261 139	16 711 711

From comparison of Table 1 and Table 2 we can see sufficient gain in the number of parameters of the VMC(s, r)-model comparing to the VMC(s)-model.

2.2. Ergodicity conditions

Let us give now ergodicity conditions for the Markov chain of conditional order.

Theorem 1. *Homogenous vector Markov chain with partial connections VMC(s, r) is ergodic if and only if there exists a number $c \in \mathbf{N}$, such that the following inequality holds:*

$$\min_{J_1^s, J_{c+1}^{c+s} \in A^{ms}} \sum_{J_{s+1}^c \in A^{m(c-s)}} \prod_{t=1}^{c} q_{S_{M_r}}(J_t, \ldots, J_{t+s-1}), J_{t+s} > 0. \tag{7}$$

Proof. Consider the first-order Markov chain with extended state space, which is equivalent to x_t:

$$\bar{x}_t = (x_{t,1}, \ldots, x_{t,m}, x_{t+1,1}, \ldots, x_{t+1,m}, \ldots, x_{t+s-1,1}, \ldots, x_{t+1,m}) \in A^{ms},$$

Transition matrix for $\bar{x}_t$ has the following form:

$$\bar{P} = (\bar{p}_{I_1^s, J_1^s}),\ \bar{p}_{I_1^s, J_1^s} = \mathbf{I}\{I_2^s = J_1^{s-1}\} p_{I_1^s, J_s}, \tag{8}$$

where $\mathbf{I}\{C\}$ is the indicator function of the event C, $p_{I_1^s, J_s}$ is determined by (5). Ergodicity of x_t is equivalent to ergodicity of $\bar{x}_t$. According to (Kemeny and Snell 1963) $\bar{x}_t$ is ergodic if and only if there exists a number $c \in \mathbf{N}$, such that the following inequality holds:

$$\min_{J_1^s, J_{c+1}^{c+s} \in A^{ms}} \bar{p}^{(c)}_{J_1^s, J_{c+1}^{c+s}} > 0,$$

where $\bar{p}^{(c)}_{J_1^s, J_{c+1}^{c+s}}$ is the c-step transition probability from J_1^s to J_{c+1}^{c+s} for the Markov chain $\bar{x}_t$. In (Kharin and Maltsew 2011) the following representation for this probability was obtained:

$$\bar{p}^{(c)}_{J_1^s, J_{c+1}^{c+s}} = \sum_{J_{s+1}^c \in A^{m(c-s)}} \prod_{t=1}^{c} p_{J_t^{t+s-1}, J_{t+s}}.$$

Using this result and equation (5) we come to the criterion (7). Theorem is proved.

Corollary 1. If all elements of matrix Q are positive, then the VMC(s,r) is ergodic.

If the VMC(s,r) is ergodic one, then the stationary probability distribution exists (Doob 1953). We will denote it $(\pi_{J_1^s})$, $J_1^s \in A^{ms}$, and its marginal distributions will be denoted as

$$\pi_s^{M_r}(i_1,\ldots,i_r) = \mathrm{P}\{S_{M_r}(x_t,\ldots,x_{t+s-1}) = (i_1,\ldots,i_r)\},$$

$$\pi_{s+1}^{M_r}(i_1,\ldots,i_r,I_{r+1}) = \mathrm{P}\{S_{M_r}(x_t,\ldots,x_{t+s-1}) = (i_1,\ldots,i_r), x_{t+s} = I_{r+1}\} =$$

$$= \pi_s^{M_r}(i_1,\ldots,i_r)q_{(i_1,\ldots,i_r),I_{r+1}}.$$

3. Statistical estimators for VMC(s,r) parameters

3.1. Likelihood function

Let us construct now statistical estimators for VMC(s,r) parameters. At first, we need to construct the likelihood function.

Introduce the notation: $X^{(n)} \in A^{mn}$ is the observed vector time series of length n;

$$\nu_{s+1}^{M_r}(i_1,\ldots,i_r,I_{r+1}) = \sum_{t=1}^{n-s} \mathbf{I}\{S_{M_r}(x_t,\ldots,x_{t+s-1}) = (i_1,\ldots,i_r), x_{t+s} = I_{r+1}\}, \tag{9}$$

$$\nu_s^{M_r}(i_1,\ldots,i_r) = \sum_{I_{r+1}\in A^m} \nu_{s+1}^{M_r}(i_1,\ldots,i_r,I_{r+1}),\ (i_1,\ldots,i_r) \in A^r,\ I_{r+1} \in A^m, \tag{10}$$

are frequency statistics for the VMC(s,r) based on $X^{(n)}$.

Lemma 1. If the true values s, r and M_r, are known, then the likelihood function for the VMC(s,r) has the following form:

$$L_n(X^{(n)},Q) = \pi^{(0)}_{x_1,\ldots,x_s} \prod_{t=s}^{n-1} q_{S_{M_r}(x_{t-s+1},\ldots,x_t),x_{t+1}}. \tag{11}$$

Proof. Equation (11) follows from the expression for the n-dimensional probability distribution that we get using theorem on compound probabilities, the Markov property and definition of the VMC(s,r):

$$\mathrm{P}\{x_1 = j_1, x_2 = j_2, \ldots, x_n = j_n\} =$$

$$= \mathrm{P}\{X_1^s = J_1^s\} \prod_{t=s}^{n-1} \mathrm{P}\{x_{t+1} = J_{t+1} | X_1^t = J_1^t\} = \pi^{(0)}_{J_1,\ldots,J_s} \prod_{t=s}^{n-1} q_{S_{M_r}(J_{t-s+1},\ldots,J_t),J_{t+1}}.$$

Lemma is proved.

Corollary 2. The loglikelihood function for the VMC(s,r) has the following form:

$$l_n(X^{(n)},Q) = \ln \pi^{(0)}_{J_1,\ldots,J_s} + \sum_{i_1,\ldots,i_r\in A} \sum_{I_{r+1}\in A^m} \nu_{s+1}^{M_r}(i_1,\ldots,i_r,I_{r+1}) \ln q_{(i_1,\ldots,i_r),I_{r+1}}. \tag{12}$$

3.2. Estimators for transition probabilities

Construct now maximum likelihood estimators (MLE) for the matrix Q of one-step transition probabilities.

Theorem 2. If the true values s, r and the template-set M_r are known, then the maximum likelihood estimators for the one-step transition probabilities (5) are

$$\hat{q}_{(i_1,\ldots,i_r),I_{r+1}} = \begin{cases} \dfrac{\nu^{M_r}_{s+1}(i_1,\ldots,i_r,I_{r+1})}{\nu^{M_r}_{s}(i_1,\ldots,i_r)}, \ if\ \nu^{M_r}_{s}(i_1,\ldots,i_r) > 0, \\ \\ \dfrac{1}{N^m}, \ if\ \nu^{M_r}_{s}(i_1,\ldots,i_r) = 0. \end{cases} \tag{13}$$

Proof. In order to construct the MLE we need to solve the following conditional extremum problem:

$$\begin{cases} l_n(X^{(n)},Q) \to \max\limits_{Q}, \\ \sum\limits_{I_{r+1}\in A^m} q_{(i_1,\ldots,i_r),I_{r+1}} = 1, \ i_1,\ldots,i_r \in A. \end{cases}$$

This problem splits into N^r subproblems for each set $(i_1,\ldots,i_r)$:

$$\begin{cases} \sum\limits_{I_{r+1}\in A^m} \nu^{M_r}_{s+1}(i_1,\ldots,i_r,I_{r+1}) \ln q_{(i_1,\ldots,i_r),I_{r+1}} \to \max\limits_{q_{(i_1,\ldots,i_r),I_{r+1}}}, \\ \sum\limits_{I_{r+1}\in A^m} q_{(i_1,\ldots,i_r),I_{r+1}} = 1. \end{cases}$$

Using the Lagrange multipliers method for solving these subproblems we get estimators (13). Theorem is proved.

Consistency of estimators (13) follows from the next theorem.

Theorem 3. If the VMC(s,r) is stationary Markov chain, then MLE (13) are consistent estimators as $n \to \infty$:

$$\hat{q}_{(i_1,\ldots,i_r),I_{r+1}} \xrightarrow{\mathrm{P}} q_{(i_1,\ldots,i_r),I_{r+1}}, \ i_1,\ldots,i_r \in A, \ I_{r+1} \in A^m. \tag{14}$$

Proof. Normalized frequencies of the states for the s-order Markov chain tend to the stationary probability distribution as $n \to \infty$ (Kharin and Maltsew 2011):

$$\hat{\pi}_{J_1^{s+1}} = \frac{1}{n-s}\sum_{t=1}^{n-s} \mathrm{I}\{x_t = J_1,\ldots,x_{t+s} = J_{s+1}\} \xrightarrow{\mathrm{P}} \pi_{J_1^{s+1}} = \pi_{J_1^s} p_{J_1^s,J_{s+1}}.$$

Since frequencies $\nu^{M_r}_{s+1}(i_1,\ldots,i_r,I_{r+1})$, $\nu^{M_r}_{s}(i_1,\ldots,i_r)$ are sums of the frequencies of the $(s+1)$-tuples, we have convergence property for $n \to \infty$:

$$\hat{\pi}^{M_r}_{s}(i_1,\ldots,i_r) = \frac{1}{n-s}\nu^{M_r}_{s}(i_1,\ldots,i_r) \xrightarrow{\mathrm{P}} \pi^{M_r}_{s}(i_1,\ldots,i_r), \tag{15}$$

$$\hat{\pi}^{M_r}_{s+1}(i_1,\ldots,i_r,I_{r+1}) = \frac{1}{n-s}\nu^{M_r}_{s+1}(i_1,\ldots,i_r,I_{r+1}) \xrightarrow{\mathrm{P}}$$

$$\xrightarrow{\mathrm{P}} \pi^{M_r}_{s+1}(i_1,\ldots,i_r,I_{r+1}) = \pi^{M_r}_{s}(i_1,\ldots,i_r) q_{(i_1,\ldots,i_r),I_{r+1}}. \tag{16}$$

Using (15), (16) and theorem on functional transformations of convergent random sequences from (Borovkov 1998), we come to (14). Theorem is proved.

3.3. Estimators for the template-set

To construct estimators for the template-set M_r we will also use the maximum likelihood method. Let

$$H(M_r) = \sum_{i_1,\ldots,i_r\in A}\ \sum_{I_{r+1}\in A^m} \nu^{M_r}_{s+1}(i_1,\ldots,i_r,I_{r+1}) \ln \frac{\nu^{M_r}_{s+1}(i_1,\ldots,i_r,I_{r+1})}{\nu^{M_r}_{s}(i_1,\ldots,i_r)} \tag{17}$$

be the plug-in statistical estimator for the Shannon conditional information on the future vector x_{t+1}.

Theorem 4. If the order s and the number of connections r are a priori known, then the MLE for the template-set M_r is

$$\hat{M}_r = arg \max_{M_r \in \mathbf{M}_r} H(M_r). \tag{18}$$

Proof. Formula (18) follows from (17) and from the representation of the loglikelihood function (12) for the VMC(s, r).

Computational complexity of the exhaustive search in the formula (18) is $\mathcal{O}(nmN^{r+m}(sm)^{r-1})$, that is why we can use it only for small values of m and r. Therefore we developed a special algorithm for calculation of estimators (18) to reduce computational complexity. This algorithm is based on step-by-step extension of the initial template-set.

Let $\mathbf{M}_r^+(M_{r-1})$ be the set of templates built by extension of the template M_{r-1} by one element from the set $M_* \setminus M_{r-1}$, $r = 2, 3, \ldots$. At the first step of the algorithm we find the initial template $\hat{M}_{r_-}$, $r_- \geq 1$, using the exhaustive search in (18). Then we find sequentially the estimators:

$$\hat{M}_{r_-+1}, \hat{M}_{r_-+2}, \ldots, \hat{M}_r. \tag{19}$$

Estimator $\hat{M}_{r'}$, $r' = r_- + 1, r_- + 2, \ldots, r$, is constructed as follows:

$$\hat{M}_{r'} = arg \max_{M_{r'} \in \mathbf{M}_{r'}(M_{r'-1})} H(M_{r'}), \tag{20}$$

i. e. we extend the template-set $\hat{M}_{r'-1}$ by one additional element.

3.4. Estimators for the order and the number of connections

In order to estimate the order s and the number of connections r we use the Bayesian information criterion (BIC) (Csiszar and Shields 1999), that has the following expression for our model:

$$(\hat{s}, \hat{r}) = arg \min_{2 \leq s' \leq s_+,\ 1 \leq r' \leq r_+} BIC(s', r'),$$

$$BIC(s', r') = -l_n(X^{(n)}, Q, M_r) + 2d \ln(n - s') =$$

$$= - \sum_{i_1, \ldots, i_{r'} \in A} \sum_{I_{r'+1} \in A^m} \nu_{s'+1}^{M_{r'}}(i_1, \ldots, i_{r'}, I_{r'+1}) \ln \frac{\nu_{s'+1}^{M_{r'}}(i_1, \ldots, i_{r'}, I_{r'+1})}{\nu_{s'}^{M_{r'}}(i_1, \ldots, i_{r'})} + 2d \ln(n - s'),$$

where $s_+ \geq 1$, $1 \leq r_+ \leq ms_+$ are maximal admissible values of s and r respectively, d is the number of independent parameters of the VMC(s, r) determined by formula (6).

4. Results of computer experiments

4.1. Simulated data

Estimation of the matrix Q

Let us illustrate the properties of the constructed statistical estimators by computer simulation. Binary vector Markov chain with the following values of parameters:

$$m = 4,\ s = 4,\ r = 6,\ M_6 = \{(1,1), (2,2), (2,4), (3,1), (3,2), (4,3)\} \tag{21}$$

was simulated. Matrix Q has dimension $2^6 \times 2^4$, its elements were generated as random variables with the uniform probability distribution in the interval [0, 1].

We generated 100 independent realizations of VMC(s, r), each realization consisted of n binary m-vectors, $n \in \{10^5, 2 \cdot 10^5, \ldots, 100 \cdot 10^5\}$. Statistical estimators for transition probabilities were constructed according to formula (13), and the mean square estimation error for estimation of Q was calculated:

$$\Delta_n = \sum_{i_1,\ldots,i_6 \in \{0,1\}} \sum_{I \in \{0,1\}^4} \left(q_{(i_1,\ldots,i_6),I} - \hat{q}_{(i_1,\ldots,i_6),I} \right)^2 .$$

Figure 1 represents dependence of Δ_n on the time series length n and illustrates the consistency property of statistical estimators (13).

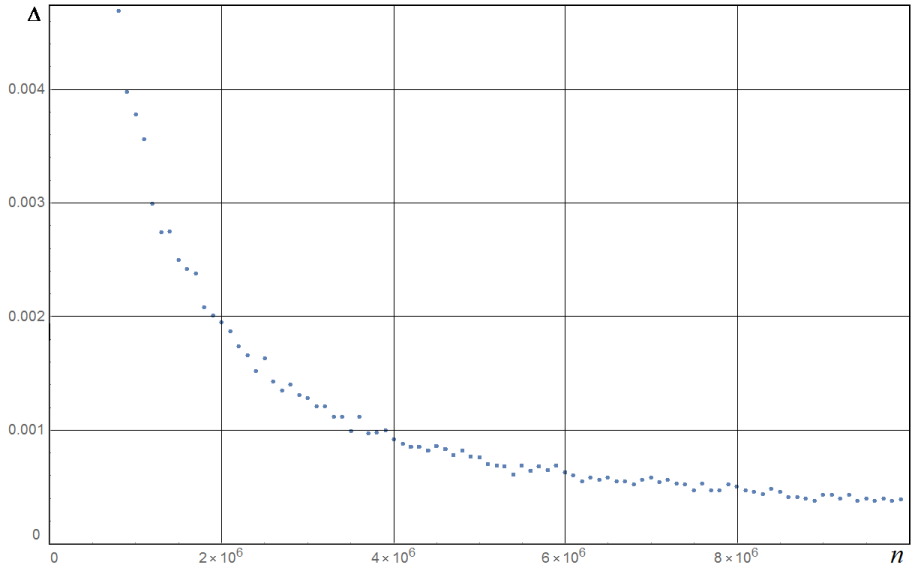


Figure 1: Estimation error of Q

Estimation of the template-set M_r

Let us analyze now properties of the estimators $\hat{M}_r$ built by the extension algorithm presented in the subsection 3.3. In computer experiments we simulated VMC(s, r) with parameters (21) for the time series length $n = 1000, 10000, 20000, \ldots, 150000$. For each n we simulated $U = 1000$ independent realizations. For each realization the template-set estimator $\hat{M}_r$ was computed with the extension algorithm. Then we calculate the frequency estimate of true decision:

$$\delta_n = \frac{1}{U} \sum_{u=1}^{U} \mathbf{I}\{\hat{M}_r^{(u)} = M_r\}, \tag{22}$$

where $\hat{M}_r^{(u)}$ is the estimator for the template-set obtained for the u-th realization.

Results presented in Table 3 illustrate the consistency property of the estimator $\hat{M}_r$.

Table 3: Error of estimating M_r

n	1000	10000	20000	30000	50000	60000	80000	100000	120000	150000
δ_n	0.001	0.341	0.609	0.762	0.914	0.943	0.979	0.993	0.998	1

4.2. Real data

In conclusion let us present results of experiments on real genetic data. We tested genetic DNA sequence LN589993 (https://www.ncbi.nlm.nih.gov/nuccore) which consists of four nucleotides (A, C, G, T). Detection of dependences in DNA sequences is an important problem

in bioinformatics (Waterman 1999; Voloshko, Medved, and Kharin 2016), and we constructed stochastic model for the observed sequence using VMC(s, r).

We recoded the sequence to binary form: "A" corresponds to 00, "C" corresponds to 10, "G" corresponds to 01, "T" corresponds to 11. The sequence was splitted on $n = 5708$ nonoverlapping triplets and represented as 6-variate ($m = 6$) time series. Using step-by-step extension algorithm we estimated the template-set, and using (13) we estimated the transition probabilities and the mean square deviation of matrix Q from uniform distribution

$$\Delta = \max_{i_1,\ldots,i_r \in \{0,1\},\ I \in \{0,1\}^m} |\hat{q}_{(i_1,\ldots,i_r),I} - 1/64|. \tag{23}$$

For $s = 3$, $r = 7$ we get the following results:

$$\hat{M} = \{(1,3),(1,4),(2,1),(2,2),(3,2),(3,3),(3,4)\},\ \Delta = 0.175.$$

As we can see deviation of matrix $\hat{Q}$ from the case of "pure randomness" is quite significant, and constructed VMC(3, 7) model detects stochastic dependences in the analyzed genetic sequence.

References

Bonacich P (2003). "Asymptotics of a Matrix Valued Markov Chain Arising in Sociology." *Stochastic Processes and their Applications*, **104**(1), 155–171.

Borovkov A (1998). *Mathematical Statistics*. Gordon and Breach, New York.

Csiszar I, Shields P (1999). "Consistency of the BIC Order Estimator." *Electronic research announcements of the American mathematical society*, **5**, 123–127.

Doob J (1953). *Stochastic Processes*. Wiley, New York.

Kemeny J, Snell J (1963). *Finite Markov Chains*. D. Van Nostrand Company, Princeton NJ.

Kharin A (2016). "Performance and Robustness Evaluation in Sequential Hypotheses Testing." *Communications in Statistics - Theory and Methods*, **45**(6), 1693–1709.

Kharin A, Kishylau D (2015). "Robust Sequential Test for Hypotheses about Discrete Distributions in the Presence of "Outliers"." *Journal of Mathematical Sciences*, **205**(1), 68–73.

Kharin A, Shlyk P (2009). "Robust Multivariate Bayesian Forecasting under Functional Distortions in the ξ^2-metric." *Journal of Statistical Planning and Inference*, **139**(11), 3842–3846.

Kharin Y (1997). "Robustness of Clustering under Outliers." *Lecture Notes in Computer Science*, **1280**, 501–512.

Kharin Y (2012). "Parsimonious Models for High-order Markov Chains and Their Statistical Analysis." *VIII World Congress on Probability and Statistics*, pp. 168–169.

Kharin Y, Maltsew M (2011). "Algorithms for Statistical Analysis of Markov Chain with Conditional Memory Depth." *Informatics*, **1**, 34–43(in Russian).

Kharin Y, Petlitskii A (2007). "A Markov Chain of Order s with r Partial Connections and Statistical Inference on Its Parameters." *Discrete Mathematics and Applications*, **17**(3), 295–317.

Kharin Y, Zhuk E (1998). "Filtering of Multivariate Samples Containing "Outliers" for Clustering." *Pattern Recognition Letters*, **19**(11), 1077–1085.

Voloshko V, Medved E, Kharin Y (2016). "Multiresolution Statistical Analysis of DNA." *Proceedings of the 13th international conference*, pp. 178–181.

Waterman M (1999). *Mathematical Methods for DNA Sequences*. Chapman and Hall/CRC, Boca Raton, Florida.

Affiliation:

Yuriy Kharin
Department of Mathematical Modelling and Data Analysis
Belarusian State University
Independence av. 4
220030 Minsk, Belarus
E-mail: kharin@bsu.by

Michail Maltsew
Research Institute for Applied Problems of Mathematics and Informatics
Belarusian State University
Independence av. 4
220030 Minsk, Belarus
E-mail: maltsew@bsu.by

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 46 *Submitted:* 2016-11-15
April 2017 *Accepted:* 2017-02-02

Austrian Journal of Statistics
April 2017, Volume 46, 47–56.
http://www.ajs.or.at/
doi:10.17713/ajs.v46i3-4.670

Statistical Estimation and Classification Algorithms for Regime-Switching VAR Model with Exogenous Variables

Vladimir Malugin
Belarusian State University

Alexander Novopoltsev
Belarusian State University

Abstract

We consider a vector autoregression model with exogenous variables and Markov-switching regimes to describe complex systems with cyclic changes of states. To estimate and forecast the states, we propose EM and discriminant analysis algorithms based on non-classified and classified data samples accordingly. The accuracy of the algorithms is examined by means of computer simulation experiments.

Keywords: regime-switching models, vector autoregression models with exogenous variables, EM algorithm, discriminant analysis algorithm, dynamic programming approach, probability of misclassification.

1. Introduction

Regime-switching models are a convenient tool for the analysis of complex systems with cyclic changes of states (Hamilton 2008). Most studies are devoted to Markov-switching vector autoregression model (MS-VAR) (Krolzig 1997). If the regimes are independent or there is a high uncertainty regarding the classes of states, then the models with independent-switching regimes may be more preferable. The autoregression and regression models of such type were entirely studied in Malugin and Kharin (1986) and Malugin (2014). The object of the study is a vector autoregression model with Markov-switching states including exogenous variables (MS-VARX), thus allowing a multivariate linear regression ones (Malugin 2014).

2. Models and tasks of research

Let a complex system at time t be characterized by a random observation vector defined on the probability space $(\Omega, \mathbf{F}, \mathbf{P})$, where Ω is a space of elementary objects $\omega \in \Omega$; $\mathbf{P}$ is a probability measure: $\mathbf{P}(A) = \mathbf{P}\{\omega \in A\}$, $A \in \mathbf{F}$. Let $\{\Omega_0, \ldots, \Omega_{L-1}\}$ be a decomposition of Ω into a finite number of non-empty disjoint subsets, such that: $\Omega_l \in \mathbf{F}, \mathbf{P}\{\Omega_l\} = \mathbf{P}\{\omega \in \Omega_l\} > 0, \bigcup_{l \in S(L)} \Omega_l = \Omega, S(L) = \{0, \ldots, L-1\}$. These subsets are the classes of states of a complex system, and L is the number of classes.

A random vector $y_t = (x'_t, z'_t)' \in R^n$ can be partitioned into subvectors of endogenous

variables $x_t = (x_{tj}) \in \Re^N$ and deterministic exogenous variables (regressors) $z_t = (z_{tk}) \in \mathfrak{Z} \subset \Re^M$. It is assumed that, in general, the time series is described by a model RS-VARX(p)($p \geq 1$):

$$x_t = \sum_{i=1}^{p} A_{d(t),i} x_{t-i} + B_{d(t)} z_t + \eta_{d(t),t}, \quad t = 1, \ldots, T, \tag{1}$$

where $x_{1-p}, \ldots, x_0 \in \Re^N$ are a set of the given initial values; $\eta_{d(t),t} \in \Re^N$ is a random disturbances or innovation process; and $d(t) \in S(L) = \{0, \ldots, L-1\}$ is a state of a system at time t.

Model (1) satisfies the following assumptions:

M.1. Segmented-stationary condition: for each class of states $l \in S(L)$ matrices of autoregression coefficients $\{A_{l,i}\}(i = 1, ..., p)$ satisfy the stationarity condition for VAR(p) model (Lutkepohl 2005);

M.2. Disturbance assumptions: disturbances $\{\eta_{l,r}\}\,(t, s = 1, ..., T,\ l \in S(L))$ are independent Gaussian random vectors with parameters: $\mathbf{E}\{\eta_{l,t}\} = 0_N \in \Re^N, \mathbf{E}\{\eta_{l,t}\eta'_{l,s}\} = \delta_{t,s}\Sigma_l$, where $\delta_{r,s}$ — the Kronecker delta.

M.3. Structural heterogeneity conditions: for matrices of autoregression and regression coefficients: $A_l \neq A_k$ and (or) $B_l \neq B_k \forall k \neq l,\ k, l \in S(L)$.

We consider a model with L ($2 \leq L < s + 1$) classes of states: where $s \geq 1$ — number of state switching points $1 < \tau_1 < \ldots < \tau_s < T$. Concerning the sequence of states $d(t) \equiv d_t \in S(L)(t = 1, ..., T)$ there are two types of assumptions:

d1. $d_t\,(t = 1, ..., T)$ — independent identically distributed random variables with probability distribution $\mathbf{P}\{d_t = l\} = \pi_l > 0\,(l \in S(L))\,, \sum_{l \in S(L)} \pi_l = 1;\ \mathbf{P}\{d_t = l\} = \pi_l > 0\,(l \in S(L)) \sum_{l \in S(L)} \pi_l = 1$;

d2. $d_t\,(t = 1, ..., T)$ — homogeneous ergodic Markov chain (GCM) with the distribution determined by the vector of probability of the initial state π and matrix one-step transition probabilities P:

$$\pi = (\pi_l),\ \pi_l = \mathbf{P}\{d_1 = l\} > 0\ \ (l \in S(L))\,, \sum_{l \in S(L)} \pi_l = 1;$$
$$P = (p_{kl}), p_{kl} = P\{d_{t+1} = l | d_t = k\} \geq 0 (k, l \in S(L)), \sum_{l \in S(L)} p_{kl} = 1, k \in S(L). \tag{2}$$

Under the conditions of d1 and d2, we deal with the models IS-VARX and MS-VARX respectively. Model (1) includes a number of special cases: model of multivariate linear regression RS-MLR, if $p = 0$, $M \geq 1$ (Malugin 2014); model RS-VAR without exogenous variables, if $p > 0$, $M = 0$ (Krolzig 1997).

The true values of model parameters $\{A_l, B_l, \Sigma_l\,(l \in S(L)), \pi\, P$ and the moments of switching state $\{\tau_i\}(i = 1, ..., s)$ are unknown. There is either classified or a non-classified sample of observations $(\bar{X}, \bar{Z})$ ($\bar{X} = (x_t) \in \Re^{NT}$, $\bar{Z} = (z_t) \in \mathfrak{Z}^T \subseteq \Re^{T}$), so that the vector of states $\bar{d} = (d_t) \in S^T(L)$ is either known or unknown. We presented two statistical classification algorithms for MS-VARX model in these cases: EM algorithm for joint parameters and vector of states estimation for non-classified sample and discriminant analysis algorithm in the case of classified sample for classification of out-of-sample observations. For IS-MLR and IS-VARX models the listed tasks are solved in Malugin (2014).

3. Splitting of mixtures described by MS-VARX

Representations for the model parameters. Model (1) under the assumptions M.1-M.3, d.1 and d.2 can be represented in the regression form

$$x_t = \Pi_{d(t)} u_t + \eta_{d(t),t}, \tag{3}$$

where $\Pi_{d(t)} = (A_{d(t),1}, \ldots, A_{d(t),p}, B_{d(t)})$ is the block $N \times (pN + M)$ — matrix of parameters; $u_t = (x'_{t-1}, \ldots, x'_{t-p}, z_t)' \in \Re^{Np+M}$ — the stacked vector of predetermined variables formed from lagged endogenous and exogenous variables with values known at time t.

In this case we use a sample of observations $(\bar{X}, \bar{U})$, where $\bar{X} = (x'_1, \ldots, x'_T)' \in \Re^{NT}$ — the values of the endogenous variables, which correspond to the values $\bar{U} = (u'_1, \ldots, u'_T)' \in \Re^{NpT} \times \mathfrak{Z}^T \subseteq \Re^{(Np+M)T}$ of predefined variables. For the model (3) we will also denote:

$\theta_l \in \Re^m (m = N \times (pN + M) + N(N + 1)/2)$ — stacked vector of parameters for the class $l \in S(L)$ consisting of independent elements of matrices $\{\Pi_l, \Sigma_l\}$ $(l \in S(L))$;

$\phi \in \Re^q (q = Lm + (L-1)(L+1))$ — parameters of a mixture model, including $\{\theta_l\}$ and π, P, $\hat{\phi} \in \Re^q$ — statistical estimate of $\phi \in \Re^q$;

$D = (d_1, \ldots, d_T)' \in S^T(L)$ — the state vector for the period under observation;

$\tilde{\gamma}_{l,t} = \mathbf{P}\{d_t = l | \bar{X}, \bar{U}; \tilde{\phi}\}$ — posterior probabilities of the class $l \in S(L)$ at the moment t;

$\tilde{\xi}_{kl,t} = \mathbf{P}\{d_{t+1} = l | d_t = k; \bar{X}, \bar{U}; \tilde{\phi}\}$ — posterior probability of transition from class $k \in S(L)$ to class $l \in S(L)$ at the moment $t (t = 1, \ldots, T-1)$.

If the model (3) satisfies the assumptions M.1-M.3, then the random vector $x_t \in \Re^N$ under the given values $u_t \in \Re^{Np+M}$ and $d_t = l (l \in S(L))$ has conditional normal distribution with density

$$p_X(x, u, \theta_l) = (2\pi)^{-\frac{N}{2}} |\Sigma_l|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x - \Pi_l u)' \Sigma_l^{-1} (x - \Pi_l u)\right\}, \; x \in \Re^N, u \in \Re^{Np+M}. \quad (4)$$

The likelihood function for parameters ϕ under the fixed state vector $D \in S^T(L)$ and assumptions (4) and d.2 is of a form:

$$L(\phi; \bar{X}, \bar{U}, D) = \pi_{d_1} p_X(x_1; u_1, \theta_{d_1}) \prod_{t=2}^{T} p_{d_{t-1}, d_t} p_X(x_t; u_t, \theta_{d_t}). \quad (5)$$

Let $\Lambda(\phi, \tilde{\phi})$ be the conditional expectation of the log-likelihood function $l(\phi; \bar{X}, \bar{U}, D) = \ln L(\phi; \bar{X}, \bar{U}, D)$ induced by the distribution $P\{D | \bar{X}, \bar{Z}; \tilde{\phi}\}$ of the random vector D given the fixed sample $(\bar{X}, \bar{U})$ and initial value $\tilde{\phi}$ of the parameter vector, i.e.

$$\begin{aligned} \Lambda(\phi, \tilde{\phi}) &= E_{\tilde{\phi}}\{l(\phi; \bar{X}, \bar{U}, D) | \bar{X}, \bar{U}; \tilde{\phi}\} = \\ &= \sum_{l \in S(L)} \tilde{\gamma}_{l,1} \ln \pi_l + \sum_{t=2}^{T} \sum_{k \in S(L)} \sum_{l \in S(L)} \tilde{\xi}_{kl,t} \ln p_{kl} + \sum_{t=1}^{T} \sum_{l \in S(L)} \tilde{\gamma}_{l,t} \ln p_X(x_t; u_t, \tilde{\theta}_l) = \\ &= Q_1(\{\pi_l\}) + Q_2(\{p_{kl}\}) + Q_3(\{\theta_l\}). \end{aligned} \quad (6)$$

In accordance with a general approach (Malugin 2014; Bilmes 1998) we obtain an analytical representation for the unknown characteristics. In the considered case we have conditional normal distribution for vector of endogenous variables with the density $p_X(x; u, \theta_l)$ for the given vector of predetermined (lagged or exogenous) variables $u_t = (x'_{t-1}, \ldots, x'_{t-p}, z_t)' \in \Re^{Np+M}$. Formulas for the posterior probabilities $\{\tilde{\gamma}_{l,t}\}, \{\tilde{\xi}_{l,t}\}$ are based on the density $p_X(x; u, \theta_l)$ and followed from the Lemma 1.

Lemma 1. *For fixed values of the parameters $\{\tilde{\theta}_l\}$, $\tilde{\pi}$, $\tilde{P}$ of the model (3) posterior probabilities $\tilde{\gamma}_{t,l}$, $\tilde{\xi}_{kl,t}$ for the sample (X, Z) are of a form:*

$$\tilde{\gamma}_{l,t} = \frac{\tilde{\alpha}_{l,t} \tilde{\beta}_{l,t}}{\sum\limits_{k \in S(L)} \tilde{\alpha}_{k,t} \tilde{\beta}_{k,t}}, \quad l \in S(L), \; t = 1, \ldots, T; \quad (7)$$

$$\tilde{\xi}_{kl,t} = \frac{\tilde{\alpha}_{k,t} \tilde{p}_{kl} p_X(x_{t+1}; u_{t+1}, \tilde{\theta}_l) \tilde{\beta}_{l,t+1}}{\sum\limits_{r \in S(L)} \sum\limits_{s \in S(L)} \tilde{\alpha}_{r,t} \tilde{p}_{rs} p_X(x_{t+1}; u_{t+1}, \tilde{\theta}_s) \tilde{\beta}_{s,t+1}}, \quad k, l \in S(L), \; t = 1, \ldots, T-1; \quad (8)$$

$$\tilde{\alpha}_{l,1} = \tilde{\pi}_l p_X(x_1; u_1, \tilde{\theta}_l),\ \tilde{\alpha}_{l,t} = \left(\sum_{k \in S(L)} \tilde{\alpha}_{k,t-1} \tilde{p}_{kl} \right) p_X(x_t; u_t, \tilde{\theta}_l), \quad t = 2, \ldots, T; \tag{9}$$

$$\tilde{\beta}_{l,T} \equiv 1,\ \tilde{\beta}_{l,t} = \sum_{k \in S(L)} \tilde{p}_{lk} p_X(x_{t+1}; u_{t+1}, \tilde{\theta}_k) \tilde{\beta}_{k,t+1}, \quad t = T-1, T-2, \ldots, 1. \tag{10}$$

The proof of the Lemma 1 is based on the method from Bilmes (1998) for Gaussian Mixture with Markov regime switching.

The representation for estimate $\hat{\phi} \in \Re^q$ is obtained by maximization of the conditional expectation of the log-likelihood function (6) for some given initial value $\tilde{\phi} \in \Re^q$, that is:

$$\hat{\phi} = \arg\max_{\phi \in \Re^q} \Lambda(\phi, \tilde{\phi}) = \arg\max_{\phi \in \Re^q} E_{\tilde{\phi}} \{ l(\phi; \bar{X}, \bar{U}, D) | \bar{X}, \bar{U}; \tilde{\phi} \}, \tag{11}$$

Theorem 1. *If model MS-VARX (3) satisfies the assumptions M.1-M.3, d.2, the estimates* $\{\hat{\Pi}_l, \hat{\Sigma}_l\}$ $(l \in S(L))$, $\hat{\pi}$, $\hat{P}$ *on a sample* $(\bar{X}, \bar{U})$ *are the solution of equation (11) for a given* $\tilde{\phi} \in \Re^q$*:*

$$\hat{\pi}_l = \tilde{\gamma}_{l,1}, \quad \hat{p}_{kl} = \sum_{t=2}^{T} \tilde{\xi}_{kl,t} \left(\sum_{t=2}^{T} \tilde{\gamma}_{k,t-1} \right)^{-1}, \hat{\Pi}_l = \sum_{t=1}^{T} \tilde{\gamma}_{l,t} x_t u'_t \left(\sum_{t=1}^{T} \tilde{\gamma}_{l,t} u_t u'_t \right)^{-1}, \tag{12}$$

$$\hat{\Sigma}_l = \sum_{t=1}^{T} \tilde{\gamma}_{l,t} (x_t - \hat{\Pi}_l z_t)(x_t - \hat{\Pi}_l z_t)' \left(\sum_{t=1}^{T} \tilde{\gamma}_{l,t} \right)^{-1}, \tag{13}$$

where posterior probabilities $\{\tilde{\gamma}_{l,t}\}, \{\tilde{\xi}_{kl,t}\}$ *are described by the formulas (7)–(10).*

Proof. Three terms Q_1, Q_2 and Q_3 in the formula (6) depend on the various parameter sets. Therefore, the optimization problem for $\Lambda(\phi, \tilde{\phi})$ can be partitioned into three independent optimization problem for continuous in the parameters functions where a posterior probabilities $\{\tilde{\gamma}_{l,t}\}, \{\tilde{\xi}_{kl,t}\}$ are given. To maximize the functions Q_1, Q_2 with equality constrained we use the method of Lagrange multipliers. Maximizing the function Q_3 of the form

$$Q_3(\{\theta_l\}) = \sum_{l \in S(L)} \sum_{t=1}^{T} \tilde{\gamma}_{l,t} \left(-\tfrac{N}{2} \ln(2\pi) - \tfrac{1}{2} \ln |\Sigma_l| - \tfrac{1}{2} (x_t - \Pi_l u_t)' \Sigma_l^{-1} (x_t - \Pi_l u_t) \right)$$

is carried out separately on matrices Π_l and Σ_l by calculating the derivatives and using properties of matrices operations (Anderson 1984). □

Corollary. Using the known block structure for matrices $\hat{\Pi}_l$, we can get the estimates $\{\hat{A}_{l,1}, \ldots, \hat{A}_{l,p}, \hat{B}_l\}$ $(l \in S(L))$.

EM-algorithm for MS-VARX. For joint estimation of all parameters $\phi \in \Re^q$ and state vector $D \in S^T(L)$ the EM MS-VARX-algorithm (*Expectation-Maximization algorithm for MS-VARX*) is addressed. EM MS-VARX-algorithm belongs to the family of Baum – Welch algorithms of splitting of a mixture of multivariate distributions, controlled by a hidden Markov chain (Bilmes 1998).

The algorithm includes the following steps (superscript k in brackets indicates the iteration number).

Preliminary step, that includes: 1) setting the initial values of the parameters: $\phi^{(0)} \equiv \tilde{\phi}$; or $D^{(0)} = (d_t^{(0)})(t = 1, ..., T)$; 2) setting the parameters defining the accuracy rate of the objective function calculation ε $(0 < \varepsilon << 1)$ and the maximum number of iterations $\bar{k}$.

For iteration k $(k = 1, 2, \ldots)$:

Step E. Calculation of $\{\tilde{\gamma}_{l,t},\ \tilde{\xi}_{kl,t}\}$ by the formulas (7)–(10) assuming $\tilde{\phi} \equiv \phi^{(k-1)}$. Estimation of $D^{(k)} = (d_t^{(k)}) \in S^T(L)$ $(t = 1, \ldots, T\}$ by the decision rule of the maximum a posteriori probability of the class:

$$d_t^{(k)} = \arg\max_{l \in S(L)} \left\{ \gamma_{l,t}^{(k)} \right\}, \quad t = 1, \ldots, T. \tag{14}$$

Step M. Computation of the parameter estimates $\{\Pi_l^{(k)}, \Sigma_l^{(k)}\}$ $(l \in S(L)), \pi^{(k)}, P^{(k)}$by the formulas (12), (13) with using the probability $\gamma_{l,t}^{(k-1)}$ and $\xi_{ij,t}^{(k-1)}$ calculated in the Step E.

Checking two stop conditions (Bilmes 1998): 1) $k = \bar{k}$; 2) $l_X^{(k)} \geq l_X^{(k-1)}$ and $\left(l_X^{(k)} - l_X^{(2)}\right) < (1+\varepsilon)\left(l_X^{(k-1)} - l_X^{(2)}\right)$, where $l_X^{(k)} = \ln P\{\bar{X}, \bar{U}|\phi^{(k)}\} = \ln\left(\sum_{l \in S(L)} \alpha_{l,T}^{(k)}\right)$. If one of the conditions is satisfied, we set: $\hat{\phi} = \phi^{(k)}$, $\hat{D} = D^{(k)}$,$\hat{l}_X = l_X^{(k)}$, $\hat{\gamma}_{l,t} = \gamma_{l,t}^{(k)}$, $(l \in S(L),\ t = 1, \ldots, T)$. In this case, the algorithm terminates, otherwise the algorithm proceeds to Step E.

Convergence problems for this type of algorithms are investigated in numerous studies, particularly in Krolzig (1997); Malugin (2014). The convergence of the algorithm ensures the consistence of the resulting parameters estimates $\hat{\phi}$, $\hat{\pi}$, $\hat{P}$ as well as the consistence of the classification rule (13).

4. Discriminant analysis of the MS-VARX

The decision classification rule of multivariate autoregression observations $(\bar{X}, \bar{U})$ described by the MS-VARX model in general case can be defined as:$\hat{D} = (\hat{d}_t) = D(\bar{X}, \bar{U})$, $\hat{d}_t = \hat{d}_t(\bar{X}, \bar{U}) \in S(L)$, $t = 1, \ldots, T$. The accuracy of classification for this rule is characterized by the probability of misclassification:

$$r = r(D(\bar{X}, \bar{U})) = P\{||\hat{D} - D^0|| \neq 0\}, \quad ||D - D|| = \sum_{t=1}^{T} (1 - \delta_{\hat{d}t, d_t^0}), \tag{15}$$

where $D^0 = (d_t^0)$ and $\hat{D} = (\hat{d}_t)$ are the true state vector and its estimate respectively.

Assume first all parameters of the MS-VARX (3) to be known. Describe an optimal classification rule, called *Bayesian decision rules* (BDR) (Malugin 2014; Kharin 1996), which minimizes the probability of misclassification (15). Bayesian decision rules of pointwise and groupwise classification of multivariate observations described by IS-VARX and IS-MLR models, have been proposed and studied in Malugin (2014). In the considered case of MS-VARX model we addressing the groupwise classification decision rule. A similar problem in the case of a parametric family of continuous probability distributions was considered in Kharin (1996). To formulate the decision rule we will use the log-likelihood function, which for some fixed vector D according to (5) simplifies to:

$$l(\phi; \bar{X}, \bar{U}, D) = \ln(L(\phi; \bar{X}, \bar{U}, D)) = \ln \pi_{d_1} + \sum_{t=2}^{T} \ln p_{d_{t-1}, d_t} + \sum_{t=1}^{T} \ln p_X(x_t; u_t, \theta_{d_t}). \tag{16}$$

Lemma 2. *If model MS-VARX (3) satisfies the assumptions of M.1-M.3, d.2 and the staked vector of parameters $\phi \in \Re^q$ is known, BDR of groupwise classification is determined by the condition*

$$\hat{D} \equiv \hat{D}(\bar{X}_1^T, \bar{U}_1^T) = \arg\max_{D \in S^T(L)} l(\phi; \bar{X}_1^T, \bar{U}_1^T, D), \tag{17}$$

where $(\bar{X}_1^T, \bar{U}_1^T)$ $(\bar{X}_1^T = (x'_1, ..., x'_T)' \in \Re^{NT}$, $\bar{U}_1^T = (u'_1, \ldots, u'_T{}' \in \Re^{NpT} \times \mathfrak{Z}^T \subseteq \Re^{(Np+M)T})$ is a sample of observations to be classified.

Proof. It is known (Kharin 1996) that the decision rule of the form (17) for arbitrary family of parametric continuous distributions minimizes a probability of error classification. Such decision rules are known as Bayesian decision rules. Under the conditions of the Lemma 2 the vector of endogenous variables $x_t \in \Re^N$ corresponding to fixed values $u_t \in \Re^{Np+M}$ and

$d_t = l(l \in S(L))$ has conditional Gaussian distribution with density (4) which belong to the mentioned above family of parametric continuous distributions. □

To solve the integer optimization task (17) for some fixed continuous vector $\phi \in \Re^q (q = Lm + (L-1)(L+1))$ we will use the dynamic programming method (Kharin 1996; Bellman and Dreyfus 1962). Its implementation requires a special representation of the log-likelihood function $l(\phi; \bar{X}, \bar{U}, D)$ through the so-called Bellman functions.

Theorem 2. *Under the conditions of Lemma 2, the BDR of groupwise classification of sample* $(\bar{X}_1^T, \bar{U}_1^T)$ *is implemented using the dynamic programming method in accordance with the following relationships:*

$$\hat{d}_T = \arg\max_{k \in S(L)} F_T(k),\ \hat{d}_t = \arg\max_{k \in S(L)} \left(f_t(k, \hat{d}_{t+1}) + F_t(k) \right), \quad t = T-1, T-2, \ldots, 1, \tag{18}$$

$$F_1(l) \equiv 0,\ F_{t+1}(l) = \max_{k \in S(L)} \left(f_t(k,l) + F_t(k) \right),\ l \in S(L), \quad t = 1, \ldots, T-1, \tag{19}$$

where $\{F_t(k)\}$ *are Bellman functions and* $\{f_t(k,l)\}$ *are described by formulas*

$$f_t(k,l) = \delta_{t,1}(\ln \pi_k + \ln p_X(x_1; u_1, \theta_k)) + \ln p_{kl} + \ln p_X(x_{t+1}; u_{t+1}, \theta_l), k, l \in S(L), \tag{20}$$

δ_{t1} — *Kronecker symbol,* $t = 1, \ldots, T-1$.

Proof. In conditions of Lemma 2 the formulas (18)–(20) are obtained by means of equivalent transformation of function $l(\phi; \bar{X}, \bar{U}, D)$. Indeed, on the basis of (16), (17) and (20) we obtain:

$$\hat{D} = \arg\max_{D \in S^T(L)} \sum_{t=1}^{T-1} f_t(d_t, d_{t+1}). \tag{21}$$

It is known that a dynamic programming procedure includes the following two stages which use formulas (19) and (18) respectively:

1) recursive calculation of Bellman functions $\{F_t(l)\}$ $(l \in S(L), t = 1, ..., T-1)$ by the formulas

$$F_{t+1}(l) = \max_{k \in S(L)} (f_t(k,l) + F_t(k)),\ F_1(l) \equiv 0;$$

2) calculation of vector $\hat{D}$ components in the reverse order:

$$\hat{d}_t = \arg\max_{k \in S(L)} \left(f_t(k, \hat{d}_{t+1}) + F_t(k) \right)\ (t = T-1, T-2, ..., 1),$$
$$\hat{d}_T = \arg\max_{k \in S(L)} F_T(k).$$

□

Since parameters $\{\theta_l\}$ $(l \in S(L))$, π, P are unknown, we need to use their estimates obtained from some sample of classified observations. To get such a sample as to find the estimates $\{\hat{\theta}_l\}$ $(l \in S(L))$, $\hat{\pi}$, $\hat{P}$ it is suggested to apply the proposed above EM MS-VARX algorithm. Thus, the following statement is true.

Corollary. *If* $\{\hat{\theta}_l\}$ $(l \in S(L)), \hat{\pi}, \hat{P}$ *are consistent estimates of parameter for model (3), then using them in (15)–(17) instead of unknown values of parameters we obtain a consistent "plug-in" Bayesian decision rule.*

The "plug-in" BDR of group classification can be used to forecast future states of complex system for a given horizon $h \geq 1$ using new out-of-sample observations $(\bar{X}_{T+1}^{T+h}, \bar{U}_{T+1}^{T+h})$, where $\bar{X}_{T+1}^{T+h} = (x'_{T+1}, \ldots, x'_{T+h})' \in \Re^{Nh}$, $\bar{U}_{T+1}^{T+h} = (u'_{T+1}, \ldots, u'_{T+h})' \subseteq \Re^{(Np+M)h}$.

5. Performance evaluations

Description of test models and examples. We consider the model MS-VARX in the form (1) or (3) under the assumptions M.1–M.3, d.2 with cyclic changes in the matrix of

regression coefficients. The aim of experiments is to evaluate the accuracy of classification and prediction for the proposed decision rules. We use the following notation for the proposed classification algorithms for MS-VARX: BDR — Bayesian decision rule of groupwise classification algorithms; EBDR — estimated ("plug-in") BDR algorithms; EM — EM MS-VARX algorithms.

General description of the test models: $L = 2$, $N = 2$, $M = 3$; $A_1 = A_0$, $\Sigma_0 = \Sigma_1 = \Sigma$; $T \in \{100, 200, 500, 1000, 2000\}$, $h = 100$. The exogenous vector $z_t = (z_{tj}) \in \Re^M$ has a uniform distribution in $\mathfrak{Z} = a^M$, $a = [1, 10]$ with mean value $\tilde{z} = \mathbf{E}\{z\} = (5.5, 5.5, 5.5)'$. Interclass Mahalanobis distance defined for the mean value of the vector of exogenous variables is denoted by $\Delta(\tilde{z})$.

Parameter values for various experiments

$$\Sigma_0 = \Sigma_1 = \Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}; \quad B_0 = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 0 & 3 \end{pmatrix}, \quad B_1 = B_0 + H;$$

$$B.1.\ H = \begin{pmatrix} 0 & 0 & 0 \\ -0.5 & 0 & 0 \end{pmatrix}, \quad B.2.\ H = \begin{pmatrix} 0 & 0 & 0 \\ -1 & 1 & 1 \end{pmatrix}, \quad B.3.\ H = \begin{pmatrix} 0 & 0 & 0 \\ -1 & 0 & -1 \end{pmatrix};$$

$$\pi_0 = \pi_1 = 0.5, \quad P = \begin{pmatrix} 1-\omega & \omega \\ \omega & 1-\omega \end{pmatrix} \quad (0 < \omega < 0.5).$$

Characteristics of classification and estimation accuracy. The matrix $H = B_1 - B_0$ in the case $A_1 = A_0$, $\Sigma_0 = \Sigma_1 = \Sigma$ determines the degree of distinctiveness of classes, caused by structural changes in the matrix of regression coefficients. The probability of misclassification under the model assumptions is calculated according to the formulas (Malugin 2014; Kharin 1996):

$$r(\tilde{z}) = \pi_0 r_0(\tilde{z}) + \pi_1 r_1(\tilde{z}), \quad r_l(\tilde{z}) = \Phi\left(-\frac{\Delta(\tilde{z})}{2} - (-1)^l \frac{h}{\Delta(\tilde{z})}\right), \quad h = \ln\frac{\pi_0}{\pi_1}\ (l \in \{0, 1\}),$$

where $\Phi(\cdot)$ — the function of standard normal distribution, $\Delta(\tilde{z})$ — interclass Mahalanobis distance at point $\tilde{z}$.

The probability of misclassification is calculated by averaging the classification results of $K = 100$ random samples for each set of parameters using the formulas $\hat{r} = K^{-1}\sum_{i=1}^{K} \hat{r}_i$, $\hat{r}_i = 1 - T^{-1}\sum_{t=1}^{T} \delta_{\hat{d}_t^i, d_t^0}$ where $D^0 = (d_t^0)$, $\hat{D}^i = (\hat{d}_t^i)$ — true state vector and its estimate respectively for the i-th sample.

The accuracy of the parameter estimates is determined by the characteristics $\delta_\theta = ||\hat{\theta} - \theta||$, $\delta_P = ||\hat{P} - P||$, where $||\cdot||$ is the Euclidean norm of the matrix and vector.

Analysis of the results of experiments.

Case 1. The impact of differences in matrix of regression and autoregression coefficients for different classes. Parameters value (set 1): variants B.1–B.3 for the matrix of regression coefficients, $A_1 = A_2 = \mathrm{O}_{N\times N}$, $\omega = 0.2$. The estimates of accuracy measures for these experiments are presented in Table 1.

Table 1: The impact of structural changes in regression coefficients.

Variants of matrix B	Accuracy of classification and estimation algorithms					
	$\Delta(\tilde{z})$	$\hat{r}_{BDR}$	$\hat{r}_{EM}$	$\hat{r}^h_{EBDR}$	δ_θ	δ_P
B.1	1.23	0.198	0.294	0.34	0.265	0.28
B.2	2.46	0.097	0.1	0.109	0.191	0.073
B.3	4.919	0.017	0.02	0.018	0.166	0.059

Parameters values (set 2): variant B.3 for matrix of regression coefficients, $\omega = 0.2$,

$$A.1.\ A_1 = A_0 = \begin{pmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{pmatrix}; \quad A.2.\ A_0 = A_1 = \begin{pmatrix} 0.6 & 0 \\ 0 & 0.6 \end{pmatrix} \quad (\textit{the same matrices});$$

$$A.3.\ A_0 = -A_1 = \begin{pmatrix} 0.6 & 0 \\ 0 & 0.6 \end{pmatrix}; \quad A.4.\ A_0 = -A_1 = \begin{pmatrix} 0.9 & 0 \\ 0 & 0.9 \end{pmatrix} \quad (\textit{different matrices}).$$

The estimates of accuracy measures for these experiments are presented in Table 2.

Table 2: The impact of structural changes in autoregression coefficients.

T	A.1		A.2		A.3		A.4	
	$\hat{r}_{BDA}$	$\hat{r}_{EM}$	$\hat{r}_{BDA}$	$\hat{r}_{EM}$	$\hat{r}_{BDA}$	$\hat{r}_{EM}$	$\hat{r}_{BDA}$	$\hat{r}_{EM}$
100	0.0077	0.0787	0.0077	0.0588	0.0013	0.0015	0.0001	0.0049
200	0.0074	0.0128	0.0074	0.0082	0.0012	0.0013	0.0002	0.0002

Conclusion 1. The accuracy of classification depends on the number of parameters, subject to structural changes and the severity of structural changes; the presence of structural changes in the matrices of autoregression coefficients leads to a decrease in the probability of misclassification (compare the values of $\hat{r}_{BDA}$ and $\hat{r}_{EM}$ for the cases A.2. ($A_0 = A_1$) and A.4. ($A_0 = -A_1$) in Table 2).

Case 2. The impact of training sample size T on the accuracy of the algorithms. Parameters value: variant B.2 for the matrix of regression coefficients; $T \in \{100, 200, 500, 1000, 2000\}$, forecast horizon $h = 100$. The dependence of the accuracy of classification and prediction on the size of training sample is illustrated in Figure 1.

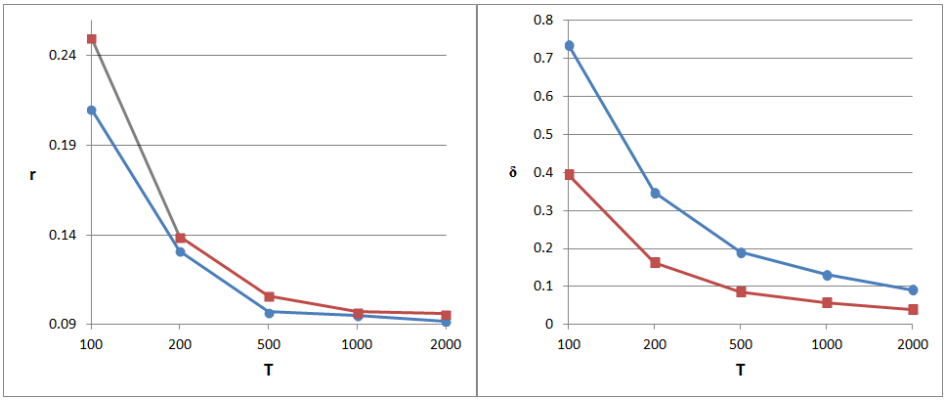


Figure 1: The dependence of the accuracy of algorithms on the size of training sample: left — $\hat{r}_{EM}$ (circles) and $\hat{r}^{h}_{EDA}$ (squares); right — δ_θ (circles) and δ_P (squares)

Conclusion 2. There is observed an expected rise in the accuracy of the classification and estimation algorithms with increasing interclass distance and volume of observations (Figures 1, 2, Table 1);

Case 3. The effect of uncertainty regarding the class of state on the efficiency of the EM MS-VARX algorithm. Parameters value: under conditions of Case 2 the uncertainty of state is described by the parameters $\omega \in \{0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.5\}$, $T = 100$. The value $\omega = 0.1$ corresponds to the high degree of certainty, the value $\omega = 0.5$ corresponds to the highest degree of uncertainty. The dependence of the accuracy of classification and estimation on the parameter ω is illustrated in Figure 2.

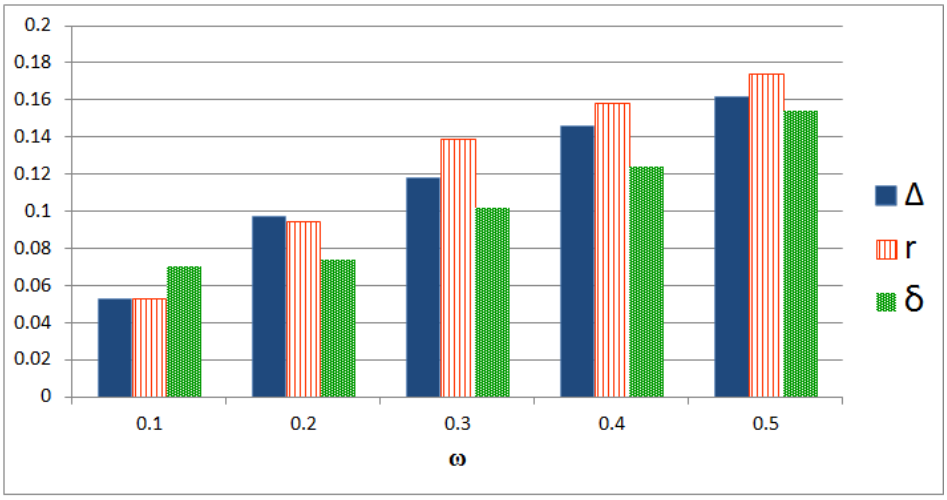


Figure 2: The effect of uncertainty regarding the class of the EM MS-VARX algorithm (columns from left to the right): interclass distance $\Delta(\tilde{z})$; estimate of the probability of misclassification $\hat{r}_{EM}$; characteristics of parameters estimation accuracy δ_θ

Conclusion 3. The increasing degree of uncertainty regarding the state of the system have the following effects for the EM MS-VARX and EDBR algorithms: interclass distance decreases and the probability of misclassification falls significantly (compare the values of $\hat{r}_{EM}$ for the cases $\omega = 0.1$ and $\omega = 0.5$ in Figure 2). This indicates the feasibility of using in these cases the IS MS-VARX algorithm (Malugin 2014) for independent classes of states.

References

Anderson T (1984). *An Introduction to Multivariate Statistical Analysis.* Wiley.

Bellman R, Dreyfus S (1962). *Applied Dynamic Programming.* Princeton University Press Princeton, New Jersey.

Bilmes J (1998). *A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models.* Int. Computer Science Institute, Berkeley CA.

Hamilton J (2008). "Regime-switching Models." *New Palgrave Dictionary of Economics*, pp. 1755–1804.

Kharin Y (1996). *Robustness in Statistical Pattern Recognition.* Dordrecht, Boston, London Kluwer Academic Publishers.

Krolzig H (1997). *Markov Switching Vector Autoregressions. Modelling Statistical Inference and Application to Business Cycle Analysis.* Berlin, Springer-Verlag.

Lutkepohl H (2005). *New Introduction to Multiple Time Series Analysis.* Berlin, Springer-Verlag.

Malugin V (2014). *Methods of Analysis of Multivariate Econometric Models with Heterogeneous Structure.* Minsk, Belarusian State University.

Malugin V, Kharin Y (1986). "On Optimal Classification or Random Observations Different in Regression Equations." *Automation and Remote Control*, (7), 61–69.

Affiliation:

Vladimir Malugin
Department of Mathematical Modeling and Data Analysis
Belarusian State University
220030 Minsk, Belarus
E-mail: malugin@bsu.by
URL: http://fpmi.bsu.by/en/main.aspx?guid=24341

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 46
April 2017

http://www.ajs.or.at/
http://www.osg.or.at/

Submitted: 2016-11-15
Accepted: 2017-02-02

Austrian Journal of Statistics
April 2017, Volume 46, 57–66.
http://www.ajs.or.at/
doi:10.17713/ajs.v46i3-4.671

On Independent Component Analysis with Stochastic Volatility Models

Markus Matilainen
University of Turku

Jari Miettinen
University of Jyvaskyla

Klaus Nordhausen
University of Turku

Hannu Oja
University of Turku

Sara Taskinen
University of Jyvaskyla

Abstract

Consider a multivariate time series where each component series is assumed to be a linear mixture of latent mutually independent stationary time series. Classical independent component analysis (ICA) tools, such as fastICA, are often used to extract latent series, but they don't utilize any information on temporal dependence. Also financial time series often have periods of low and high volatility. In such settings second order source separation methods, such as SOBI, fail. We review here some classical methods used for time series with stochastic volatility, and suggest modifications of them by proposing a family of vSOBI estimators. These estimators use different nonlinearity functions to capture nonlinear autocorrelation of the time series and extract the independent components. Simulation study shows that the proposed method outperforms the existing methods when latent components follow GARCH and SV models. This paper is an invited extended version of the paper presented at the CDAM 2016 conference.

Keywords: blind source separation, GARCH model, nonlinear autocorrelation, multivariate time series.

1. Introduction

In this paper we assume that the observed p-variate time series $\boldsymbol{x} = (\boldsymbol{x}_t)_{t=0,\pm1,\pm2,\ldots}$ follows the basic *independent component (IC) model*

$$\boldsymbol{x}_t = \boldsymbol{\mu} + \boldsymbol{\Omega}\boldsymbol{z}_t, \quad t = 0, \pm1, \pm2, \ldots,$$

where $\boldsymbol{\mu}$ is a p-variate location vector, $\boldsymbol{\Omega}$ is a full-rank $p \times p$ mixing matrix and $\boldsymbol{z} = (\boldsymbol{z}_t)_{t=0,\pm1,\pm2,\ldots}$ is an unobservable p-variate stationary time series such that

$$(i) \quad \mathsf{E}(\boldsymbol{z}_t) = \boldsymbol{0}, \quad (ii) \quad \mathrm{COV}(\boldsymbol{z}_t) = \boldsymbol{I}_p \text{ and}$$
$$(iii) \text{ the component series of } \boldsymbol{z} \text{ are independent.}$$

Then $\boldsymbol{x}$ is also stationary with $\mathsf{E}(\boldsymbol{x}_t) = \boldsymbol{\mu}$ and $\mathrm{COV}(\boldsymbol{x}_t) = \boldsymbol{\Sigma} = \boldsymbol{\Omega}\boldsymbol{\Omega}'$. In *independent component analysis (ICA)* the goal is to find, using the observed time series $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$, an

estimate of an unmixing matrix $\boldsymbol{W}$ such that $\boldsymbol{Wx} = (\boldsymbol{Wx}_t)_{t=0,\pm1,\pm2,\ldots}$ has independent component series.

The IC model has recently received a lot of attention in financial time series analysis as complicated p-variate time series models can then be replaced by p simple univariate (e.g. ARMA or GARCH) models in parameter estimation and prediction problems. The model also serves as a dimension reduction tool as often only few component series in $\boldsymbol{z}$ are relevant while the rest just present noise. For some recent contributions, see Broda and Paolella (2009); Chen, Härdle, and Spokoiny (2007); García-Ferrer, González-Prieto, and Peña (2012); Lu, Wu, and Lee (2009); Oja, Kiviluoto, and Malaroiu (2000).

In the literature standard ICA methods, such as fastICA, are often used to estimate an unmixing matrix $\boldsymbol{W}$ in a time series context although such methods only use the marginal distribution of $\boldsymbol{x}_t$ and make no use of the information on temporal dependence. On the other hand, there exist second order source separation methods, like SOBI (Belouchrani, Abed Meraim, Cardoso, and Moulines 1997), which are particularly popular for analyzing biomedical data. Such methods use autocovariances and cross-autocovariances for the estimation. They are capable of separating time series with nonzero linear autocorrelations, but they do not utilize nonlinear autocorrelations.

Volatility clustering is a common feature in economic and financial time series, i.e. there are periods of lower and higher volatility. As the transitions between such periods do not typically have any clear pattern, they are treated as random occurrences. There are a vast amount of different models that have been invented for such situations. In our simulations we consider two popular choices, the GARCH model (Bollerslev 1986) and the SV model (Taylor 1982). For further information on stochastic volatility and a recent overview of stochastic volatility models, see for example Matteson and Ruppert (2011).

In this paper we review various independent component estimators that use nonlinear autocorrelations, and compare their performance to that of fastICA in a simulation study where independent time series components follow GARCH and SV models. The paper has the following structure. First, in Section 2 we define the aforementioned univariate stochastic volatility models. In Section 3 we discuss the ICA methods which are considered in this paper and suggest our extension. In Section 4 we show that this extension has the important affine equivariance property. Section 5 consists of the simulation study.

2. Stochastic volatility models for univariate series

There are several different stochastic volatility models. Here we concentrate on two widely used ones. The first one is GARCH (Generalized Autoregressive Conditional Heteroscedasticity) process (Bollerslev 1986) defined as follows. A univariate GARCH(p, q) process is given by

$$x_t = \sigma_t \epsilon_t,$$

where ϵ_t is an independent white noise process and σ_t^2 is a conditional variance process

$$\sigma_t^2 = Var(x_t | x_u, u < t) = \omega + \sum_{i=1}^{p} \alpha_i x_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2,$$

with $\omega > 0$ and $\alpha_i, \beta_j \geq 0\ \forall i, j$. For (second order) stationarity, we require that $\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1$.

Another popular model is the SV (Stochastic Volatility) model (Taylor 1982), defined as

$$\begin{aligned} x_t &= e^{h_t/2} \epsilon_t, \\ h_t &= \mu + \phi(h_{t-1} - \mu) + \sigma \eta_t, \end{aligned}$$

where ϵ_t and η_t are two independent white noise innovation processes. The parameter μ is the level, ϕ is the persistence and $\sigma\eta_t$ is the volatility of the log-variance. The process h_t is called the volatility process and it is strongly stationary with $N(0,1)$ innovations and initial state $h_0 \sim N(\mu, \sigma^2/(1-\phi^2))$. For stationarity, we require $|\phi| < 1$ and $\mu \in \mathbb{R}$.

3. Source separation for multivariate time series

Under our model assumption, the standardized multivariate series of $\boldsymbol{x}_t$ is given by $\boldsymbol{x}_t^{st} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x}_t - \boldsymbol{\mu})$. One of the key results in ICA states that there exists an orthogonal matrix $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p)'$ such that $\boldsymbol{z}_t = \boldsymbol{U}\boldsymbol{x}_t^{st}$ (up to signs and order of the components) (Miettinen, Taskinen, Nordhausen, and Oja 2015). Here $\boldsymbol{z}_t$ denotes the vector of independent series. The final unmixing matrix functional is then given by $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Sigma}^{-1/2}$. The estimate of $\boldsymbol{W}$ is then obtained by replacing $\boldsymbol{\Sigma}$ and $\boldsymbol{U}$ by their sample counterparts. For finding $\boldsymbol{U}$, we next list the criterion functions in different approaches.

In the symmetric *fastICA* (Hyvärinen and Oja 1997) approach $\boldsymbol{U}$ maximizes

$$\sum_{i=1}^{p} \left| \mathsf{E}\left[G(\boldsymbol{u}_i'\boldsymbol{x}_t^{st})\right] \right|,$$

and in the symmetric squared fastICA (Miettinen, Nordhausen, Oja, Taskinen, and Virta 2017a) $\boldsymbol{U}$ maximizes

$$\sum_{i=1}^{p} \left(\mathsf{E}\left[G(\boldsymbol{u}_i'\boldsymbol{x}_t^{st})\right] \right)^2.$$

Here a twice continuously differentiable, nonlinear and nonquadratic function G is chosen so that $E[G(y)] = 0$ if $y \sim N(0,1)$. Two common choices for G are $G(z) = z^4 - 3$ and $G(z) = \log(\cosh(z)) - E[\log(\cosh(y))]$, where $y \sim N(0,1)$. Notice that both utilize only the stationary (marginal) distribution of $\boldsymbol{x}_t$.

The estimators presented below make use of the joint distributions of $(\boldsymbol{x}_t, \boldsymbol{x}_{t+k})$, $k = 1, 2, \ldots$. The classical *SOBI* uses only second moments and it was originally defined as a method which jointly diagonalizes several autocovariance matrices. However, SOBI can be reformulated so that $\boldsymbol{U}$ maximizes

$$\sum_{i=1}^{p} \sum_{k=1}^{K} \left(\mathsf{E}\left[(\boldsymbol{u}_i'\boldsymbol{x}_t^{st})(\boldsymbol{u}_i'\boldsymbol{x}_{t+k}^{st})\right] \right)^2.$$

The solution is unique if, for all pairs $i \neq j$ there exists a k, $1 \leq k \leq K$, such that $\mathsf{E}(z_{t,i}z_{t+k,i}) \neq \mathsf{E}(z_{t,j}z_{t+k,j})$. SOBI fails to separate GARCH and SV time series as all lagged autocovariances are in such cases zero.

The *gFOBI* procedure proposed in Matilainen, Nordhausen, and Oja (2015) maximizes a sum of fourth moments

$$\sum_{i=1}^{p} \sum_{k=0}^{K} \left(\mathsf{E}\left[(\boldsymbol{u}_i'\boldsymbol{x}_{t+k}^{st})||\boldsymbol{x}_t^{st}||^2\right] \right)^2,$$

where $||\cdot||$ is the Frobenius (matrix) norm. For $K = 0$, the regular ICA method FOBI (Cardoso 1989) is obtained.

The *gJADE* procedure (Matilainen *et al.* 2015), in turn, uses a much richer sum of fourth cumulants and maximizes

$$\sum_{i=1}^{p} \sum_{r=1}^{p} \sum_{s=1}^{p} \sum_{k=0}^{K} \left(\kappa(\boldsymbol{u}_i'\boldsymbol{x}_{t+k}^{st}, \boldsymbol{u}_i'\boldsymbol{x}_{t+k}^{st}, \boldsymbol{x}_{t,r}^{st}, \boldsymbol{x}_{t,s}^{st}) \right)^2,$$

where

$$\kappa(z_1, z_2, z_3, z_4) = \mathsf{E}(z_1z_2z_3z_4) - \mathsf{E}(z_1z_2)\mathsf{E}(z_3z_4) - \mathsf{E}(z_1z_3)\mathsf{E}(z_2z_4) - \mathsf{E}(z_1z_4)\mathsf{E}(z_2z_3).$$

Again, for $K = 0$, the regular ICA method JADE (Cardoso and Souloumiac 1993) is obtained. Both, gFOBI and gJADE, were created having stochastic volatility models in mind.

FastICA does not use any knowledge of temporal dependence, but there exists some fixed-point algorithms aimed for a time series context. The *FixNA* (Fixed-point algorithm for maximizing the nonlinear autocorrelation) method was introduced in Shi, Jiang, and Zhou (2009), and its criterion function to be maximized is

$$D_1(\boldsymbol{U}) = \sum_{i=1}^{p} \sum_{k=1}^{K} \mathsf{E}\left[G(\boldsymbol{u}_i' \boldsymbol{x}_t^{st}) G(\boldsymbol{u}_i' \boldsymbol{x}_{t+k}^{st})\right],$$

where G is a twice continuously differentiable function. The G-functions suggested in Shi *et al.* (2009) are $G(z) = \log(\cosh(z))$ and $G(z) = z^2$.

A similar function to be maximized is of the form

$$D_2(\boldsymbol{U}) = \sum_{i=1}^{p} \sum_{k=1}^{K} \left| \mathsf{E}\left[G(\boldsymbol{u}_i' \boldsymbol{x}_t^{st}) G(\boldsymbol{u}_i' \boldsymbol{x}_{t+k}^{st})\right] - \mathsf{E}\left[G(\boldsymbol{u}_i' \boldsymbol{x}_t^{st})\right]^2 \right|,$$

and we will denote it as *FixNA2*. It was first proposed in Hyvärinen (2001), however only with $G(z) = z^2$ and $K = 1$. We further similarly suggest a natural extension of SOBI with the criterion function

$$D_3(\boldsymbol{U}) = \sum_{i=1}^{p} \sum_{k=1}^{K} \left(\mathsf{E}\left[G(\boldsymbol{u}_i' \boldsymbol{x}_t^{st}) G(\boldsymbol{u}_i' \boldsymbol{x}_{t+k}^{st})\right] - \mathsf{E}\left[G(\boldsymbol{u}_i' \boldsymbol{x}_t^{st})\right]^2 \right)^2.$$

As a variant of SOBI, we call this estimator *vSOBI*.

The term $\mathsf{E}\left[G(\boldsymbol{u}_i' \boldsymbol{x}_t^{st})\right]^2$ in $D_2(\boldsymbol{U})$ and $D_3(\boldsymbol{U})$ is used to normalize the summands. Notice that in SOBI, $G(z) = z$, and hence the aforementioned term equals to 0. When $G(z) = z^2$, the term equals to 1.

Remark 1. Instead of using lags $k = 1, 2, \ldots, K$, also more flexible lag combinations could be used (see Taskinen, Miettinen, and Nordhausen 2016, for example).

To obtain the estimating equations for matrix $\boldsymbol{U}$, the Lagrangian multiplier technique can be used as in Miettinen, Illner, Nordhausen, Oja, Taskinen, and Theis (2016). The Lagrangian function to be optimized is

$$L(\boldsymbol{U}, \boldsymbol{\Lambda}) = D_r(\boldsymbol{U}) - \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} \lambda_{ij} \boldsymbol{u}_i' \boldsymbol{u}_j - \sum_{i=1}^{p} \lambda_{ii} (\boldsymbol{u}_i' \boldsymbol{u}_i - 1), \text{ for } r = 1, 2, 3,$$

where $\boldsymbol{\Lambda} = (\lambda_{ij})$ is a symmetric matrix that contains the $p(p+1)/2$ Lagrangian multipliers. Write next

$$\boldsymbol{T}_{r,i} = \boldsymbol{T}_{r,i}(\boldsymbol{U}) = \frac{\partial}{\partial \boldsymbol{u}_i} D_r(\boldsymbol{U}), \ i = 1, \ldots, p, \ r = 1, 2, 3,$$

and $\boldsymbol{T}_r = \boldsymbol{T}_r(\boldsymbol{U}) = (\boldsymbol{T}_{r,1}, \ldots, \boldsymbol{T}_{r,p})'$. Now

$$\frac{\partial}{\partial \boldsymbol{u}_i} L(\boldsymbol{U}, \boldsymbol{\Lambda}) = \boldsymbol{T}_{r,i} - \sum_{j<i} \lambda_{ji} \boldsymbol{u}_j - \sum_{j>i} \lambda_{ij} \boldsymbol{u}_j - 2\lambda_{ii} \boldsymbol{u}_i = 0.$$

Then multiplying this by $\boldsymbol{u}_j'$ from the left side gives

$$\begin{cases} j > i: & \boldsymbol{u}_j' \boldsymbol{T}_{r,i} - \lambda_{ij} = 0 \\ j < i: & \boldsymbol{u}_j' \boldsymbol{T}_{r,i} - \lambda_{ji} = 0 \end{cases}$$

Now notice that $\boldsymbol{u}_j' \boldsymbol{T}_{r,i} - \lambda_{ji} = \boldsymbol{u}_i' \boldsymbol{T}_{r,j} - \lambda_{ij}$. Then

$$\boldsymbol{u}_j' \boldsymbol{T}_{r,i} = \boldsymbol{u}_i' \boldsymbol{T}_{r,j}, \ i \neq j, \ i, j = 1, \ldots, p.$$

From here the estimating equations for an orthogonal $\boldsymbol{U}$ can be obtained.

Result 2. The estimating equations for an orthogonal matrix $\boldsymbol{U}$ are

$$\boldsymbol{U}\boldsymbol{T}_r' = \boldsymbol{T}_r\boldsymbol{U}' \text{ and } \boldsymbol{U}\boldsymbol{U}' = \boldsymbol{I}_p,$$

or, equivalently,

$$\boldsymbol{U} = (\boldsymbol{T}_r\boldsymbol{T}_r')^{-1/2}\boldsymbol{T}_r.$$

For vSOBI for example, to maximize $D_3(\boldsymbol{U})$, we need to calculate $\boldsymbol{T}_3 = (\boldsymbol{T}_{3,1}, \ldots, \boldsymbol{T}_{3,p})'$, where

$$\begin{aligned}\boldsymbol{T}_{3,i} = \sum_{i=1}^{p}\sum_{k=1}^{K} &\left(\mathsf{E}\left[G(\boldsymbol{u}_i'\boldsymbol{x}_t^{st})G(\boldsymbol{u}_i'\boldsymbol{x}_{t+k}^{st})\right] - \mathsf{E}\left[G(\boldsymbol{u}_i'\boldsymbol{x}_t^{st})\right]^2\right)\\ &\cdot\left(\mathsf{E}\left[G'(\boldsymbol{u}_i'\boldsymbol{x}_t^{st})G(\boldsymbol{u}_i'\boldsymbol{x}_{t+k}^{st})\boldsymbol{x}_t^{st}\right] + \mathsf{E}\left[G(\boldsymbol{u}_i'\boldsymbol{x}_t^{st})G'(\boldsymbol{u}_i'\boldsymbol{x}_{t+k}^{st})\boldsymbol{x}_{t+k}^{st}\right]\right.\\ &\left.-2\mathsf{E}\left[G(\boldsymbol{u}_i'\boldsymbol{x}_t^{st})\right]\mathsf{E}\left[G'(\boldsymbol{u}_i'\boldsymbol{x}_t^{st})\boldsymbol{x}_t^{st}\right]\right),\end{aligned}$$

for $i = 1, \ldots, p$.

Therefore in all three cases $r = 1, 2, 3$ the computation of $\boldsymbol{U}$ which maximizes $D_r(\boldsymbol{U})$ can be done iteratively given some initial orthogonal matrix $\boldsymbol{U}_0$ and some tolerance limit ε as described in the algorithm below.

Data: Standardized time series $\boldsymbol{x}_t^{st} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x}_t - \boldsymbol{\mu})$
Result: $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Sigma}^{-1/2}$
$\boldsymbol{U}_{old} = \boldsymbol{U}_0$;
$\Delta = \infty$;
while $\Delta > \varepsilon$ **do**
  $\boldsymbol{T}_r = \boldsymbol{T}_r(\boldsymbol{U}_{old})$;
  $\boldsymbol{U}_{new} = (\boldsymbol{T}_r\boldsymbol{T}_r')^{-1/2}\boldsymbol{T}_r$;
  $\Delta = ||\boldsymbol{U}_{new} - \boldsymbol{U}_{old}||$;
  $\boldsymbol{U}_{old} = \boldsymbol{U}_{new}$;
end
$\boldsymbol{U} = \boldsymbol{U}_{new}$;

4. Affine equivariance

In blind source separation it is desirable that an unmixing matrix estimator is affine equivariant, which means that the separation performance does not depend on the actual value of the mixing matrix $\boldsymbol{\Omega}$. Let $\boldsymbol{x}_t^* = \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{b}$ be an affine transformation of $\boldsymbol{x}_t$, where $\boldsymbol{A}$ is a non-singular $p \times p$ matrix and $\boldsymbol{b}$ is a p-vector. Then a method is called affine equivariant if $\boldsymbol{W}^* = \boldsymbol{W}\boldsymbol{A}^{-1}$ and $\boldsymbol{W}\boldsymbol{x}_t = \boldsymbol{W}^*\boldsymbol{x}_t^*$, up to location shifts, sign changes and the order of the components.

Result 3. FixNA, FixNA2 and vSOBI algorithms are affine equivariant.

As an example, consider the affine equivariance of vSOBI algorithm: As $\mathrm{COV}(\boldsymbol{x}_t^*) = \boldsymbol{\Sigma}^* = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}'$, then by Theorem 2.1 in Ilmonen, Oja, and Serfling (2012), $(\boldsymbol{\Sigma}^*)^{-1/2} = \boldsymbol{V}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{A}^{-1}$ and therefore $\boldsymbol{x}_t^{*st} = \boldsymbol{V}\boldsymbol{x}_t^{st}$ for some orthogonal $p \times p$ matrix $\boldsymbol{V}$. Let $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p)'$ be the orthogonal matrix which maximizes

$$D_3(\boldsymbol{U}) = \sum_{i=1}^{p}\sum_{k=1}^{K}\left(\mathsf{E}\left[G(\boldsymbol{u}_i'\boldsymbol{x}_t^{st})G(\boldsymbol{u}_i'\boldsymbol{x}_{t+k}^{st})\right] - \mathsf{E}\left[G(\boldsymbol{u}_i'\boldsymbol{x}_t^{st})\right]^2\right)^2.$$

Next write $\boldsymbol{U}^* = \boldsymbol{U}\boldsymbol{V}' = (\boldsymbol{u}_1^*, \ldots, \boldsymbol{u}_p^*)'$, where $\boldsymbol{u}_i^* = \boldsymbol{V}\boldsymbol{u}_i$, for $i = 1, \ldots, p$. Now $\boldsymbol{U}^*$ is an orthogonal matrix and

$$G(\boldsymbol{u}_i^{*\prime}\boldsymbol{x}_t^{*st}) = G((\boldsymbol{V}\boldsymbol{u}_i)'(\boldsymbol{V}\boldsymbol{x}_t^{st})) = G(\boldsymbol{u}_i'\boldsymbol{x}_t^{st}).$$

Clearly also $G(\boldsymbol{u}_i^{*\prime}\boldsymbol{x}_{t+k}^{*st}) = G(\boldsymbol{u}_i{}'\boldsymbol{x}_{t+k}^{st})$, where $k = 1, \ldots, K$. Then $D_3(\boldsymbol{U}^*)$ is the maximum of the criterion function for $\boldsymbol{x}_t^*$, and

$$\boldsymbol{W}^* = \boldsymbol{U}^*(\boldsymbol{\Sigma}^*)^{-1/2} = \boldsymbol{U}\boldsymbol{V}'\boldsymbol{V}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{A}^{-1} = \boldsymbol{W}\boldsymbol{A}^{-1}.$$

Using the same arguments it can be shown that also FixNA and FixNA2 algorithms are affine equivariant. Notice also that the fastICA versions discussed here, SOBI, gFOBI and gJADE are all affine equivariant.

5. Simulation study

The following simulations are conducted using R 3.2.2 (R Core Team 2016) with the packages fGarch (Wuertz and Rmetrics Core Team 2013), fICA (Miettinen, Nordhausen, Oja, and Taskinen 2014), JADE (Miettinen, Nordhausen, and Taskinen 2017b), stochvol (Kastner 2016) and tsBSS (Matilainen, Miettinen, Nordhausen, Oja, and Taskinen 2016). In our simulation studies we compare the following methods:

- FixNA, FixNA2 and vSOBI with both $G(z) = z^2$ and $G(z) = \log(\cosh(z))$ and lags $k = 1, \ldots, 12$
- symmetric fastICA and symmetric squared fastICA with both $G(z) = z^4 - 3$ and $G(z) = \log(\cosh(z)) - E[\log(\cosh(y))]$, where $y \sim N(0, 1)$
- gFOBI, gJADE with lags $k = 0, 1, \ldots, 12$ and SOBI with lags $k = 1, \ldots, 12$

In the following we refer to G-functions as pow and lcosh. Hence the methods are denoted as vSOBI(pow) or FixNA2(lcosh), for example.

As a performance measure we use the Minimum Distance Index (Ilmonen, Nordhausen, Oja, and Ollila 2010), which is defined as

$$\hat{D} = \hat{D}(\hat{\boldsymbol{W}}) = \frac{1}{\sqrt{p-1}} \inf_{\boldsymbol{C}\in\boldsymbol{\mathcal{C}}} ||\boldsymbol{C}\hat{\boldsymbol{W}}\boldsymbol{\Omega} - \boldsymbol{I}_p||,$$

where $\boldsymbol{\mathcal{C}}$ is the set of all matrices with exactly one non-zero element in each row and column, and $||\cdot||$ is the Frobenius (matrix) norm. The index has the range $0 \leq \hat{D} \leq 1$, where zero indicates perfect separation.

For time series of lengths $T = 100, 200, 400, \ldots, 25600$ we report the averages $T(p-1)\hat{D}^2$ based on 2000 repetitions. Such an average represents a global measure of variation of an unmixing matrix, see Ilmonen *et al.* (2010) for details. As all the methods are affine equivariant, we can choose $\boldsymbol{\Omega} = \boldsymbol{I}_p$ without loss of generality and consider the following two 4-variate settings:

- **GARCH setting:** The sources are four GARCH(1, 1) processes with normal innovations. The parameters (α_1, β_1) are chosen so that the first eight moments are finite, and are: (i) (0.05, 0.9), (ii) (0.1, 0.7), (iii) (0.1, 0.8) and (iv) (0.2, 0.5).
- **SV setting:** In the second setup the four sources are SV processes with normal innovations and (μ, ϕ, σ)-parameter vectors $(-10, 0.8, 0.1)$, $(-10, 0.9, 0.2)$, $(-10, 0.9, 0.3)$ and $(-10, 0.95, 0.4)$. Again, all the first eight moments exist.

Performance Figure 1 summarizes the results for both settings. Notice that for clarity we omitted the results for SOBI as it, as expected, failed to find any latent sources. From the four different fastICA versions only the best one is presented, and for FixNA only the pow version is shown as FixNA(lcosh) was much worse than any other FixNA method. FixNA(lcosh) also exhibited a strange convergence pattern as it was deteriorating when the sample sizes increased.

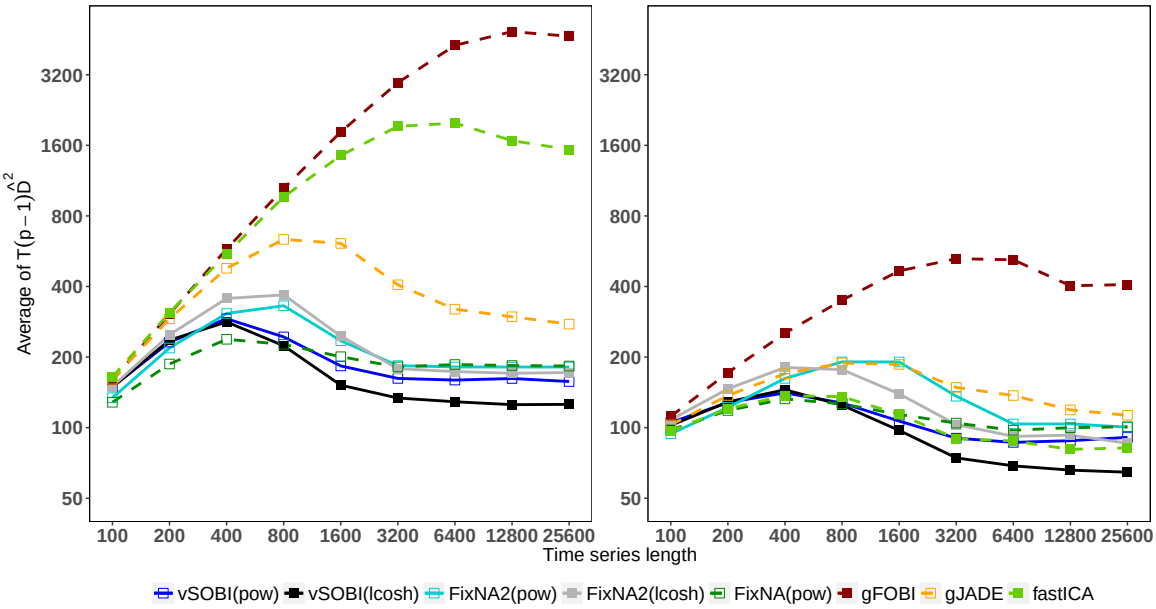


Figure 1: Comparison of performance of algorithms in the GARCH setting (left panel) and in the SV setting (right panel).

It is seen in Figure 1 that the proposed vSOBI estimator works very well both in GARCH and in SV settings and, when using $G(z) = \log(\cosh(z))$ as the nonlinearity function, it clearly outperforms all the other estimators. Squared fastICA algorithms produce slightly better results than the original fastICA algorithms. In GARCH setting the symmetric squared fastICA with $G(z) = z^4 - 3$ produces better results that other fastICA algorithms, but compared to other algorithms its performance is still not good, as only gFOBI and SOBI have poorer performance. In the SV setting however, while regular fastICA and squared fastICA algorithms with $G(z) = z^4 - 3$ are not among the best, both versions with $\log(\cosh(z))$ type nonlinearities give good results. The squared version is the better one and, when compared to other algorithms, we can see that only vSOBI with $G(z) = \log(\cosh(z))$ is better. Notice that unlike the other estimators, fastICA algorithms do not utilize any information on temporal dependence.

In general it seems better to use FixNA2 instead of FixNA, perhaps due to FixNA2 having the more natural centering part in the objective function. The performance of gFOBI is poor in both settings, but gJADE seems to work decently in SV setting.

Convergence The convergence percentages of the proposed vSOBI algorithms are good, and in time series of length 800 onwards very close to 100 %. SOBI has no converge issues, but as stated before, it has very poor performance. Also gFOBI and gJADE have very few convergence issues, and only when the time series is very short.

On the other hand, FixNA2 algorithms have lots of convergence problems in short time series, as they converged in less than 25 % of repetitions in time series of length 100. As the time series length increases, the convergence rate approaches 100 %. FastICA algorithms have more issues in the GARCH setting (less than 50 % of repetitions converging in the beginning) than in the SV setting.

FixNA(pow) has only some convergence issues in short time series. On the other hand, convergence of FixNA algorithm when using $G(z) = \log(\cosh(z))$ as the nonlinearity function is surprising in SV setting, as there are more convergence issues when time series length increases, contrary to what is expected (not shown in figures). For time series length 100 its convergence percentage is 98.5, while with time series length 25600 it is only 76.2. The reason

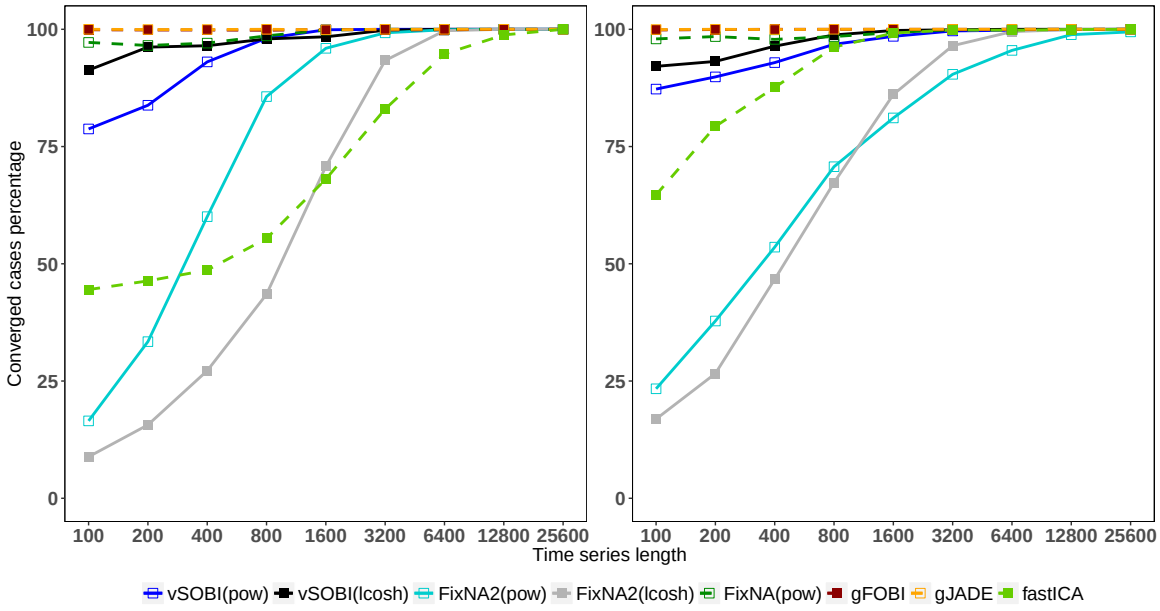


Figure 2: Comparison of convergence percentages of algorithms in the GARCH setting (left panel) and in the SV setting (right panel)

behind this behaviour is not yet known. In GARCH setting this behaviour is not seen after time series length of 800.

Converge results are summarized in Figure 2. Except for FixNA(lcosh) in the SV setting, convergence percentages of all algorithms are larger than 99.4 % in time series of length 25600.

6. Discussion

In this paper we surveyed different blind source separation methods suitable for multivariate time series with stochastic volatility features. Such methods were earlier quite scattered in the literature. We suggested a small modification to existing methods yielding the family of vSOBI estimators. The simulations were used to compare the vSOBI estimators with previously proposed methods using stochastic volatility models. The proposed vSOBI estimator with $G(z) = \log(\cosh(z))$ as the nonlinearity function showed the best performance among its competitors.

SOBI as a second order method was designed to function on time series with nonzero linear autocovariances, such as ARMA processes. In stochastic volatility models linear autocovariances are zero, and therefore SOBI is useless. On the other hand, the methods which exploit nonlinear autocorrelations, considered in this paper, are far from optimal in separating ARMA processes. In practical situations one can easily imagine a time series independent component model where some components are ARMA processes and others exhibit stochastic volatility features. Therefore it is desirable to derive methods which work in such cases.

In our future research we will consider a weighted combination of SOBI and vSOBI yielding for example the objective function

$$\sum_{i=1}^{p}\sum_{k=1}^{K}\left(a\left(\mathsf{E}\left[(\boldsymbol{u}_i'\boldsymbol{x}_t^{st})(\boldsymbol{u}_i'\boldsymbol{x}_{t+k}^{st})\right]\right)^2+(1-a)\left(\mathsf{E}\left[G(\boldsymbol{u}_i'\boldsymbol{x}_t^{st})G(\boldsymbol{u}_i'\boldsymbol{x}_{t+k}^{st})\right]-\mathsf{E}\left[G(\boldsymbol{u}_i'\boldsymbol{x}_t^{st})\right]^2\right)^2\right).$$

with some suitable weight $a \in [0,1]$ and function G. This could still be generalized to case where SOBI and vSOBI parts use different combinations of lags.

Acknowledgements

We thank the reviewer for careful reading of the paper and helpful comments. This work was supported by the Academy of Finland (grants 251965, 256291 and 268703).

References

Belouchrani A, Abed Meraim K, Cardoso JF, Moulines E (1997). "A Blind Source Separation Technique Based on Second Order Statistics." *IEEE Transactions on Signal Processing*, **45**, 434–444.

Bollerslev T (1986). "Generalized Autoregressive Conditional Heteroskedasticity." *Journal of Econometrics*, **31**(3), 307–327.

Broda SA, Paolella MS (2009). "CHICAGO: A Fast and Accurate Method for Portfolio Risk Calculation." *Journal of Financial Econometrics*, **7**(4), 412–436.

Cardoso JF (1989). "Source Separation Using Higher Order Moments." In *International Conference on Acoustics, Speech, and Signal Processing*, pp. 2109–2112.

Cardoso JF, Souloumiac A (1993). "Blind Beamforming for Non-Gaussian Signals." In *IEE Proceedings F*, volume 140, pp. 362–370.

Chen Y, Härdle W, Spokoiny V (2007). "Portfolio Value at Risk Based on Independent Component Analysis." *Journal of Computational and Applied Mathematics*, **205**, 594–607.

García-Ferrer A, González-Prieto E, Peña D (2012). "A Conditionally Heteroskedastic Independent Factor Model With an Application to Financial Stock Returns." *International Journal of Forecasting*, **28**(1), 70 – 93.

Hyvärinen A (2001). "Blind Source Separation by Nonstationarity of Variance: A Cumulant-based Approach." *IEEE Transactions on Neural Networks*, **12**(6), 1471–1474.

Hyvärinen A, Oja E (1997). "A Fast Fixed-Point Algorithm for Independent Component Analysis." *Neural Computation*, **9**, 1483–1492.

Ilmonen P, Nordhausen K, Oja H, Ollila E (2010). "A New Performance Index for ICA: Properties Computation and Asymptotic Analysis." In V Vigneron, V Zarzoso, E Moreau, R Gribonval, E Vincent (eds.), *"Latent Variable Analysis and Signal Separation", LNCS*, volume 6365, pp. 229–236. Springer, Heidelberg.

Ilmonen P, Oja H, Serfling R (2012). "On Invariant Coordinate System (ICS) Functionals." *International Statistical Review*, **80**, 93–110.

Kastner G (2016). "Dealing with Stochastic Volatility in Time Series Using the R Package stochvol." *Journal of Statistical software*, **69**(5), 1–30.

Lu CJ, Wu JY, Lee TS (2009). "Application of Independent Component Analysis Preprocessing and Support Vector Regression in Time Series Prediction." In *International Joint Conference on Computational Sciences and Optimization*, volume 1, pp. 468–471.

Matilainen M, Miettinen J, Nordhausen K, Oja H, Taskinen S (2016). *tsBSS: Tools for Blind Source Separation for Time Series*. R package version 0.2, URL https://CRAN.R-project.org/package=tsBSS.

Matilainen M, Nordhausen K, Oja H (2015). "New Independent Component Analysis Tools for Time Series." *Statistics & Probability Letters*, **105**, 80–87.

Matteson D, Ruppert D (2011). "Time-Series Models of Dynamic Volatility and Correlation." *IEEE Signal Processing Magazine*, **28**(5), 72–82.

Miettinen J, Illner K, Nordhausen K, Oja H, Taskinen S, Theis F (2016). "Separation of Uncorrelated Stationary Time Series Using Autocovariance Matrices." *Journal of Time Series Analysis*, **37**(3), 337–354.

Miettinen J, Nordhausen K, Oja H, Taskinen S (2014). *fICA: Classical, Reloaded and Adaptive FastICA Algorithms*. R package version 1.0-2, URL http://CRAN.R-project.org/package=fICA.

Miettinen J, Nordhausen K, Oja H, Taskinen S, Virta J (2017a). "The Squared Symmetric FastICA Estimator." *Signal Processing*, **131**, 402–411.

Miettinen J, Nordhausen K, Taskinen S (2017b). "Blind Source Separation Based on Joint Diagonalization in R: The Packages JADE and BSSasymp." *Journal of Statistical Software*, **76**(2). URL http://dx.doi.org/10.18637/jss.v076.i02.

Miettinen J, Taskinen S, Nordhausen K, Oja H (2015). "Fourth Moments and Independent Component Analysis." *Statistical Science*, **30**, 372–390.

Oja E, Kiviluoto K, Malaroiu S (2000). "Independent Component Analysis for Financial Time Series." In *Adaptive Systems for Signal Processing, Communications, and Control Symposium*, pp. 111–116.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. R version 3.2.4, URL http://www.R-project.org/.

Shi Z, Jiang Z, Zhou F (2009). "Blind Source Separation with Nonlinear Autocorrelation and Non-Gaussianity." *Journal of Computational and Applied Mathematics*, **223**(1), 908–915.

Taskinen S, Miettinen J, Nordhausen K (2016). "A More Efficient Second Order Blind Identification Method for Separation of Uncorrelated Stationary Time Series." *Statistics & Probability Letters*, **116**, 21–26.

Taylor SJ (1982). "Financial Returns Modelled by the Product of Two Stochastic Processes – A Study of Daily Sugar Prices 1961–79." In OD Anderson (ed.), *Time Series Analysis: Theory and Practice 1*, pp. 203–216. Springer, North-Holland, Amsterdam.

Wuertz W, Rmetrics Core Team (2013). *fGarch: Rmetrics - Autoregressive Conditional Heteroskedastic Modelling*. R package version 3010.82, URL http://CRAN.R-project.org/package=fGarch.

Affiliation:

Markus Matilainen
Department of Mathematics and Statistics
FI-20014 University of Turku, Finland
E-mail: markus.matilainen@utu.fi
URL: http://users.utu.fi/manmat

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 46
April 2017

http://www.ajs.or.at/
http://www.osg.or.at/

Submitted: 2016-11-15
Accepted: 2017-02-02

Austrian Journal of Statistics
April 2017, Volume 46, 67–78.
http://www.ajs.or.at/
doi:10.17713/ajs.v46i3-4.672

Maximum Likelihood Drift Estimation for Gaussian Process with Stationary Increments

Yuliya Mishura
Taras Shevchenko
National University of Kyiv

Kostiantyn Ralchenko
Taras Shevchenko
National University of Kyiv

Sergiy Shklyar
Taras Shevchenko
National University of Kyiv

Abstract

The paper deals with the regression model $X_t = \theta t + B_t$, $t \in [0, T]$, where $B = \{B_t, t \geq 0\}$ is a centered Gaussian process with stationary increments. We study the estimation of the unknown parameter θ and establish the formula for the likelihood function in terms of a solution to an integral equation. Then we find the maximum likelihood estimator and prove its strong consistency. The results obtained generalize the known results for fractional and mixed fractional Brownian motion.

Keywords: Gaussian process, stationary increments, discrete observations, continuous observations; maximum likelihood estimator, strong consistency, fractional Brownian motion, mixed fractional Brownian motion.

1. Introduction

We study the problem of the drift parameter estimation for the stochastic process

$$X_t = \theta t + B_t, \tag{1}$$

where $\theta \in \mathbb{R}$ is an unknown parameter, and $B = \{B_t, t \geq 0\}$ is a centered Gaussian process with stationary increments, $B_0 = 0$. In the particular case when $B = B^H$ is a fractional Brownian motion, this model has been studied by many authors. Mention the paper by Norros, Valkeila, and Virtamo (1999) that treats the maximum likelihood estimation by continuous observations of the trajectory of X on the interval $[0, T]$ (see also Le Breton (1998)). Further, Hu, Nualart, Xiao, and Zhang (2011) investigate the exact maximum likelihood estimator by discrete observations at the points $tk = kh$, $k = 1, 2, \ldots, N$; Bertin, Torres, and Tudor (2011) consider the maximum likelihood estimation in the discrete scheme of observations, where the trajectory of X is observed at the points $t_k = \frac{k}{N}$, $k = 1, 2, \ldots, N^\alpha$, $\alpha > 1$. For hypothesis testing of the drift parameter sign in the model (1) driven by a fractional Brownian motion, see Stiburek (2017). The paper Cai, Chigansky, and Kleptsyna (2016) treats the likelihood function for Gaussian processes not necessarily having stationary increments. However, on the one hand, our approach is different from their one, it cannot be deduced from their general formulas and on the other hand, gives rather elegant representations. The construction of the maximum likelihood estimator in the case where B is the sum of two fractional Brow-

nian motions was studied in Mishura (2016) and Mishura and Voronov (2015). A similar non-Gaussian model driven by the Rosenblatt process was considered in Bertin *et al.* (2011).

As already mentioned, we consider the case when B is a centered Gaussian processes with stationary increments. We construct the maximum likelihood estimators for both discrete and continuous schemes of observations. The assumptions on the process in the continuous case are formulated in terms of the second derivative of its covariance function, see Assumptions 1 and 2. The exact formula for the maximum likelihood estimator contains a solution of an integral equation with the kernel obtained after the differentiation. We give the sufficient conditions for the strong consistency of the estimators. Several examples of the process B are considered.

The paper is organized as follows. Section 2 is devoted to the case of the discrete observations. The maximum likelihood estimation for continuous time is studied in Section 3.

2. Maximum likelihood estimation by discrete observations

We start with the construction of the likelihood function and the maximum likelihood estimator in the case of discrete observations. In the next section these results will be used for the derivation of the likelihood function in the continuous-time case, see the proof of Theorem 3.3.

Let the process X defined by formula (1) be observed at the points $t_k, k = 0, 1, \dots, N$,

$$0 = t_0 < t_1 < \ldots < t_N \le T. \tag{2}$$

The problem is to estimate the parameter θ by the observations X_{t_k}, $k = 0, 1, \dots, N$ of the process X_t.

2.1. Likelihood function and construction of the estimator

Denote

$$\Delta X^{(N)} = \left(X_{t_k} - X_{t_{k-1}}\right)_{k=1}^{N}, \quad \Delta B^{(N)} = \left(B_{t_k} - B_{t_{k-1}}\right)_{k=1}^{N}.$$

Note that in our model $X_{t_0} = X_0 = 0$, and the N-dimensional vector $\Delta X^{(N)}$ is a one-to-one function of the observations. The vectors $\Delta B^{(N)}$ and $\Delta X^{(N)}$ are Gaussian with different means (except the case $\theta = 0$) and the same covariance matrix. We denote this covariance matrix by $\Gamma^{(N)}$. The next maximum likelihood estimator coincides with the least square estimator considered in Rao (2002, eq. (4a.1.5)).

Lemma 2.1. *Assume that the Gaussian distribution of the vector $(B_{t_k})_{k=1}^N$ is nonsingular. Then one can take the function*

$$L^{(N)}_{\Delta X^{(N)}=x}(\theta) = \frac{f_\theta(x)}{f_0(x)} = \exp\left\{\theta z^\top \left(\Gamma^{(N)}\right)^{-1} x - \frac{\theta^2}{2} z^\top \left(\Gamma^{(N)}\right)^{-1} z\right\}, \tag{3}$$

where $z = (t_k - t_{k-1})_{k=1}^N$, as a likelihood function in the discrete-time model. MLE is linear with respect to the observations and equals

$$\hat\theta^{(N)} = \frac{z^\top \left(\Gamma^{(N)}\right)^{-1} \Delta X^{(N)}}{z^\top \left(\Gamma^{(N)}\right)^{-1} z}. \tag{4}$$

Proof. The pdf of $\Delta B^{(N)}$ with respect to the Lebesgue measure equals

$$f_\theta(x) = \frac{1}{(2\pi)^{N/2}\sqrt{\det \Gamma^{(N)}}} \exp\left\{-\frac{1}{2}(x - \theta z)^\top \left(\Gamma^{(N)}\right)^{-1}(x - \theta z)\right\}.$$

The density of the observations for given θ with respect to the distribution of the observations for $\theta = 0$ is taken as a likelihood function. □

Remark 2.2. Let the process X be observed on a regular grid, i.e., at the points $t_k = kh$, $k = 1, \dots, N$, where $h > 0$. Then $\Gamma^{(N)}$ is a Toeplitz matrix, that is

$$\Gamma^{(N)}_{k+l,l} = \Gamma^{(N)}_{l,k+l} = \mathsf{E}\left(B_{(k+l)h} - B_{(k+l-1)h}\right)\left(B_{lh} - B_{(l-1)h}\right) = \mathsf{E}\, B_{(k+1)h}B_h - \mathsf{E}\, B_{kh}B_h$$

does not depend on l due to the stationarity of increments. This simplifies the numerical computation of MLE.

2.2. Properties of the estimator

Since $\Delta X^{(N)} = \Delta B^{(N)} + \theta z$, the maximum likelihood estimator (4) equals

$$\hat\theta^{(N)} = \theta + \frac{z^\top \left(\Gamma^{(N)}\right)^{-1} \Delta B^{(N)}}{z^\top \left(\Gamma^{(N)}\right)^{-1} z}\,.$$

Lemma 2.3. *Under the assumptions of Lemma 2.1, the estimator $\hat\theta^{(N)}$ is unbiased and normally distributed. Its variance equals*

$$\operatorname{var}\hat\theta^{(N)} = \frac{1}{z^\top \left(\Gamma^{(N)}\right)^{-1} z}\,.$$

Proof. The estimator $\hat\theta^{(N)}$ is unbiased and normally distributed because $\hat\theta^{(N)} - \theta$ is linear and centered Gaussian vector $\Delta B^{(N)}$. The variance of the estimator is equal to

$$\begin{aligned}\operatorname{var}\hat\theta^{(N)} &= \frac{\operatorname{var}\left(z^\top \left(\Gamma^{(N)}\right)^{-1} \Delta B^{(N)}\right)}{\left(z^\top \left(\Gamma^{(N)}\right)^{-1} z\right)^2} = \frac{z^\top \left(\Gamma^{(N)}\right)^{-1} \operatorname{var}\left(\Delta B^{(N)}\right) \left(\Gamma^{(N)}\right)^{-1} z}{\left(z^\top \left(\Gamma^{(N)}\right)^{-1} z\right)^2} \\ &= \frac{z^\top \left(\Gamma^{(N)}\right)^{-1} \Gamma^{(N)} \left(\Gamma^{(N)}\right)^{-1} z}{\left(z^\top \left(\Gamma^{(N)}\right)^{-1} z\right)^2} = \frac{1}{z^\top \left(\Gamma^{(N)}\right)^{-1} z}\,.\end{aligned}$$

□

To prove the consistency of the estimator, we need the following technical result.

Lemma 2.4. *If $A \in \mathbb{R}^{N\times N}$ is a positive definite matrix, $x \in \mathbb{R}^N$, $x \neq 0$ is a non-zero vector, then*

$$x^\top A^{-1} x \geq \frac{\|x\|^4}{x^\top A x}\,.$$

Proof. As the matrix A is positive definite, $x^\top A x > 0$ and there exists a positive definite matrix $A^{1/2}$ (and so the matrix $A^{1/2}$ is symmetric and nonsingular) such that $(A^{1/2})^2 = A$. By the Cauchy-Schwarz inequality,

$$\|x\|^4 = \left(x^\top A^{1/2} A^{-1/2} x\right)^2 \leq \left\|A^{1/2}x\right\|^2 \left\|A^{-1/2}x\right\|^2 = \left(x^\top A x\right)\left(x^\top A^{-1} x\right),$$

whence the desired inequality follows. □

In the rest of this section we assume that the process X is observed on a regular grid, at the points $t_k = kh$, $k = 1, \dots, N$, for some $h > 0$. We also assume that for any N the Gaussian distribution of the vector $(B_{kh})_{k=1}^N$ is nonsingular.

Theorem 2.5. *Let $h > 0$, and*

$$\mathsf{E}\left(B_{(k+1)h} - B_{kh}\right) B_h \to 0 \quad \textit{as} \quad k \to \infty.$$

Let $\hat\theta^{(N)}$ be the ML estimator of parameter θ of the model (1) by the observations X_{kh}, $k = 1, \dots, N$. Then the estimator $\hat\theta^{(N)}$ is mean-square consistent, i.e.,

$$\mathsf{E}\left(\hat\theta^{(N)} - \theta\right)^2 \to 0 \quad \textit{as} \quad N \to \infty.$$

Proof. By Remark 2.2, the matrix $\Gamma^{(N)}$ has Toeplitz structure,

$$\Gamma^{(N)}_{l,m} = \Gamma^{(N)}_{l,m} = \mathsf{E}\left(B_{(|l-m|+1)h} - B_{|l-m|\,h}\right) B_h.$$

Moreover, $\Gamma^{(N)}_{k,l}$ does not depend on N as soon as $N \geq \max(k,l)$. By Toeplitz theorem,

$$\frac{1}{N^2}\sum_{l=1}^{N}\sum_{m=1}^{N}\Gamma^{(N)}_{l,m} = \frac{1}{N}\,\mathsf{E}(B_h)^2 - \sum_{k=2}^{N}\frac{2(N+1-k)}{N^2}\,\mathsf{E}\left(B_{kh} - B_{(k-1)h}\right) B_h \to$$
$$\to \lim_{k\to\infty}\mathsf{E}\left(B_{kh} - B_{(k-1)h}\right) B_h = 0 \quad \text{as} \quad N \to \infty.$$

For a regular grid we have that $z = (h,\dots,h)^\top$. Hence, in this case,

$$z^\top \Gamma^{(N)} z = h^2 \sum_{l=1}^{N}\sum_{m=1}^{N}\Gamma^{(N)}_{l,m}, \qquad \|z\| = h\sqrt{N}.$$

Finally, with the use of Lemma 2.3,

$$\mathsf{E}\left(\hat\theta^{(N)} - \theta\right)^2 = \frac{1}{z^\top\left(\Gamma^{(N)}\right)^{-1} z} \leq \frac{z^\top\Gamma^{(N)}z}{\|z\|^4} = \frac{1}{h^2N^2}\sum_{l=1}^{N}\sum_{m=1}^{N}\Gamma^{(N)}_{l,m} \to 0 \quad \text{as} \quad N\to\infty. \quad \square$$

To prove the strong consistency, we need the following auxiliary statement.

Lemma 2.6. *Let $h > 0$, and $\hat\theta^{(N)}$ be the ML estimator of parameter θ of the model (1) by the observations X_{kh}, $k = 1,\dots,N$. Then the random process $\hat\theta^{(N)}$ has independent increments.*

Proof. In the next paragraph, $N_2 \leq N_3$ are positive integers, $(I,0) = (I_{N_2}, 0_{N_2\times(N_3-N_2)})$ is $N_2 \times N_3$ diagonal matrix with ones on the diagonal, and its transpose is denoted by $\left(\begin{smallmatrix} I \\ 0 \end{smallmatrix}\right)$. The vector $\Delta B^{(N_2)} = (B_h,\dots,B_{(N_2-1)h} - B_{N_2h})^\top$ is the beginning of vector $\Delta B^{(N_3)} = (B_h,\dots,B_{(N_3-1)h} - B_{N_3h})^\top$; the vector $z_{N_2} = (h,\dots,h)^\top \in \mathbb{R}^{N_2}$ is the beginning of vector $z_{N_3} = (h,\dots,h)^\top \in \mathbb{R}^{N_3}$, so

$$\Delta B^{(N_2)} = (I,0)\,\Delta B^{(N_3)}, \qquad z_{N_2} = (I,0)\,z_{N_3}.$$

Then

$$\mathsf{E}\,\Delta B^{(N_3)}\left(\Delta B^{(N_2)}\right)^\top = \mathsf{E}\,\Delta N^{(N_3)}\left(\Delta B^{(N_3)}\right)^\top \left(\begin{smallmatrix} I \\ 0 \end{smallmatrix}\right) = \Gamma^{(N_3)}\left(\begin{smallmatrix} I \\ 0 \end{smallmatrix}\right),$$
$$\mathsf{E}\,\hat\theta^{(N_3)}\hat\theta^{(N_2)} = \frac{\mathsf{E}\,z_{N_3}^\top\left(\Gamma^{(N_3)}\right)^{-1}\Delta B^{(N_3)}\left(\Delta B^{(N_2)}\right)^\top\left(\Gamma^{(N_2)}\right)^{-1}z_{N_2}}{z_{N_3}^\top\left(\Gamma^{(N_3)}\right)^{-1}z_{N_3}\,z_{N_2}^\top\left(\Gamma^{(N_2)}\right)^{-1}z_{N_2}} =$$
$$= \frac{z_{N_3}^\top\left(\Gamma^{(N_3)}\right)^{-1}\Gamma^{(N_3)}\left(\begin{smallmatrix} I \\ 0 \end{smallmatrix}\right)\left(\Gamma^{(N_2)}\right)^{-1}z_{N_2}}{z_{N_3}^\top\left(\Gamma^{(N_3)}\right)^{-1}z_{N_3}\,z_{N_2}^\top\left(\Gamma^{(N_2)}\right)^{-1}z_{N_2}} =$$
$$= \frac{z_{N_3}^\top\left(\begin{smallmatrix} I \\ 0 \end{smallmatrix}\right)\left(\Gamma^{(N_2)}\right)^{-1}z_{N_2}}{z_{N_3}^\top\left(\Gamma^{(N_3)}\right)^{-1}z_{N_3}\,z_{N_2}^\top\left(\Gamma^{(N_2)}\right)^{-1}z_{N_2}} =$$
$$= \frac{z_{N_2}^\top\left(\Gamma^{(N_2)}\right)^{-1}z_{N_2}}{z_{N_3}^\top\left(\Gamma^{(N_3)}\right)^{-1}z_{N_3}\,z_{N_2}^\top\left(\Gamma^{(N_2)}\right)^{-1}z_{N_2}} = \frac{1}{z_{N_3}^\top\left(\Gamma^{(N_3)}\right)^{-1}z_{N_3}}.$$

For $N_1 \leq N_2 \leq N_3$

$$\mathsf{E}\,\hat\theta^{(N_3)}\left(\hat\theta^{(N_2)} - \hat\theta^{(N_1)}\right) = \mathsf{E}\,\hat\theta^{(N_3)}\hat\theta^{(N_2)} - \mathsf{E}\,\hat\theta^{(N_3)}\hat\theta^{(N_1)}$$
$$= \frac{1}{z_{N_3}^\top\left(\Gamma^{(N_3)}\right)^{-1}z_{N_3}} - \frac{1}{z_{N_3}^\top\left(\Gamma^{(N_3)}\right)^{-1}z_{N_3}} = 0;$$

therefore for $N_1 \le N_2 \le N_3 \le N_4$

$$\mathsf{E}\left(\hat\theta^{(N_4)} - \hat\theta^{(N_3)}\right)\left(\hat\theta^{(N_2)} - \hat\theta^{(N_1)}\right) = 0.$$

Thus, the Gaussian process $\{\hat\theta^{(N)},\ N = 1, 2, \ldots\}$ is proved to have uncorrelated increments. Hence its increments are independent. □

Theorem 2.7. *Under the assumptions of Theorem 2.5, the estimator $\theta^{(N)}$ is strongly consistent, i.e. $\hat\theta^{(N)} \to \theta$ as $N \to \infty$ almost surely.*

Proof. By Theorem 2.5 $\operatorname{var}\hat\theta^{(N)} \to 0$ as $N \to \infty$, so

$$\operatorname{var}\left(\hat\theta^{(N)} - \hat\theta^{(N_0)}\right) = \operatorname{var}\hat\theta^{(N)} + \operatorname{var}\hat\theta^{(N_0)} - 2\sqrt{\operatorname{var}\hat\theta^{(N)}\operatorname{var}\hat\theta^{(N_0)}}\operatorname{corr}\left(\hat\theta^{(N)}, \hat\theta^{(N_0)}\right) \to \operatorname{var}\hat\theta^{(N_0)}$$

as $N \to \infty$. The process $\hat\theta^{(N)}$ has independent increments. Therefore by Kolmogorov's inequality, for $\epsilon > 0$ and $N \in \mathbb{N}$

$$\mathsf{P}\left(\sup_{N \ge N_0}\left|\hat\theta^{(N)} - \hat\theta^{(N_0)}\right| > \frac{\epsilon}{2}\right) \le \frac{4}{\epsilon^2}\lim_{N\to\infty}\operatorname{var}\left(\hat\theta^{(N)} - \hat\theta^{(N_0)}\right) = \frac{4}{\epsilon^2}\operatorname{var}\hat\theta^{(N_0)}.$$

Then, using the unbiasedness of the estimator, we get

$$\mathsf{P}\left(\sup_{N \ge N_0}\left|\hat\theta^{(N)} - \theta\right| \ge \epsilon\right) \le \mathsf{P}\left(\left|\hat\theta^{(N_0)} - \theta\right| \ge \frac{\epsilon}{2}\right) + \mathsf{P}\left(\sup_{N \ge N_0}\left|\hat\theta^{(N)} - \hat\theta^{(N_0)}\right| \ge \frac{\epsilon}{2}\right) \le$$
$$\le \frac{4}{\epsilon^2}\operatorname{var}\hat\theta^{(N_0)} + \frac{4}{\epsilon^2}\operatorname{var}\hat\theta^{(N_0)} = \frac{8}{\epsilon^2}\operatorname{var}\hat\theta^{(N_0)} \to 0 \quad \text{as} \quad N_0 \to \infty,$$

whence $|\hat\theta^{(N)} - \theta| \to 0$ as $N \to \infty$ almost surely. □

Example 2.8. Let us consider the model (1) with $B_t = B_t^{H_1} + B_t^{H_2}$, where $B_t^{H_1}$ and $B_t^{H_2}$ are two independent fractional Brownian motions with Hurst indices $H_1, H_2 \in (0, 1)$, i. e. centered Gaussian processes with covariance functions

$$\mathsf{E}\, B_t^{H_i} B_s^{H_i} = \frac{1}{2}\left(t^{2H_i} + s^{2H_i} - |t-s|^{2H_i}\right), \quad t \ge 0,\ s \ge 0,\ i = 1, 2.$$

These processes have stationary increments, and

$$\mathsf{E}\left(B^{H_i}_{(k+1)h} - B^{H_i}_{kh}\right)B^{H_i}_h \sim h^{2H_i} H_i\,(2H_i - 1)\,k^{2H_i - 2} \to 0, \quad \text{as } k \to \infty,$$

see e. g. (Mishura 2008, Sec. 1.2). Taking into account the independence of centered processes $B_t^{H_1}$ and $B_t^{H_2}$, we obtain that

$$\mathsf{E}(B_{(k+1)h} - B_{kh})B_h = \mathsf{E}\left(B^{H_1}_{(k+1)h} - B^{H_1}_{kh}\right)B^{H_1}_h + \mathsf{E}\left(B^{H_2}_{(k+1)h} - B^{H_2}_{kh}\right)B^{H_2}_h \to 0 \quad \text{as} \quad k \to \infty.$$

Thus, the assumptions of Theorem 2.5 are satisfied.

3. Maximum likelihood estimation by continuous observations

Let the process X be observed on the whole interval $[0, T]$. It is required to estimate the unknown parameter θ by these observations.

3.1. Likelihood function and construction of the estimator

In this section we construct a formula for continuous-time MLE, similar to the formula (4) for the discrete case.

Assumption 1. The covariance function of B_t has a mixed derivative

$$\frac{\partial^2}{\partial s\,\partial t}\,(\mathsf{E}\,B_tB_s) = K(t-s),$$

where $K(t)$ is an even function, $K \in L_1[-T,T]$.

Lemma 3.1. *Under Assumption 1, the integral $\int_0^T f(t)\,dB_t$ exists as the mean square limit of the corresponding Riemann sums for any $f \in L_2[0,T]$. Moreover,*

$$\mathsf{E}\left[\int_0^T f(t)\,dB_t\,\int_0^T g(s)\,dB_s\right] = \int_0^T f(t)\int_0^T K(t-s)g(s)\,ds\,dt \tag{5}$$

for any $f, g \in L_2[0,T]$.

Proof. According to Huang and Cambanis (1978) (see also Cramér and Leadbetter 2004, Sec. 5.3), the integral $\int_0^T f(t)\,dB_t$ exists if and only if the double Riemann integral $\int_0^T\int_0^T f(t)f(s)K(t-s)\,ds\,dt$ exists. Moreover, if the both integrals $\int_0^T f(t)\,dB_t$ and $\int_0^T g(s)\,dB_s$ exist, then formula (5) holds. However, using the properties of a convolution, one can prove that

$$\int_0^T\int_0^T f(t)f(s)K(t-s)\,ds\,dt \le \|K\|_{L_1[-T,T]}\,\|f\|^2_{L_2[0,T]} < \infty. \qquad \square$$

Define a linear operator $\Gamma_T\colon L_2[0,T] \to L_2[0,T]$ by

$$\Gamma_T f(t) = \int_0^T K(t-s)f(s)\,ds. \tag{6}$$

It follows from (6) that

$$\mathsf{E}\left[\int_0^T f(t)\,dB_t\,\int_0^T g(s)\,dB_s\right] = \int_0^T \Gamma_T f(t)g(t)\,dt. \tag{7}$$

For a fixed set of points $t_1,\dots,t_N$ which satisfy (2) define mutually adjoint linear operators $M : L_2[0,T] \to \mathbb{R}^N$ and $M^* : \mathbb{R}^N \to L_2[0,T]$. If $f \in \mathbb{R}^N$, then let $Mf \in \mathbb{R}^N$ be a vector whose k-th element is equal to $\int_{t_{k-1}}^{t_k} f(s)\,ds$. The adjoint operator M^* is

$$M^*x = \sum_{k=1}^N x_k\,\mathbf{1}_{[t_{k-1},t_k]}, \qquad x \in \mathbb{R}^N.$$

The basic properties of the operator Γ_T are collected in the following evident lemma.

Lemma 3.2. *Let Assumption 1 hold. Then*

(i) The operator Γ_T is bounded ($\|\Gamma_T\| \le \|K\|_{L_1[-T,T]}$) and self-adjoint;

(ii) The following relation between the operator Γ_T and the covariance matrix $\Gamma^{(N)}$ from Proposition 2.1 holds:

$$M\Gamma_T M^* = \Gamma^{(N)},$$

i. e., the matrix of the operator $M\Gamma_T M^*\colon \mathbb{R}^N \to \mathbb{R}^N$ equals $\Gamma^{(N)}$.

Now we are ready to formulate our key assumption on the kernel K (in terms of the operator Γ_T).

Assumption 2. For all $T > 0$, the constant function $\mathbf{1}_{[0,T]}(t) = 1$, $t \in [0,T]$, belongs to the range of the operator Γ_T, i. e. there exists a function $h_T \in L_2[0,T]$ such that

$$\Gamma_T h_T = \mathbf{1}_{[0,T]}\,.$$

Theorem 3.3. *If all finite-dimensional distributions of the process* $\{B_t,\ t \in (0,T]\}$*, are nonsingular and Assumptions 1 and 2 hold, then*

$$L(\theta) = \exp\left\{\theta \int_0^T h_T(s)\, dB_s - \frac{\theta^2}{2} \int_0^T h_T(s)\, ds\right\} \tag{8}$$

is a likelihood function.

Proof. Let us show that the function $L(\theta)$ defined in (8) is a density function for the distribution of the process X_t for a given θ with respect to the density function of a distribution of the process B_t (it coincides with X_t when $\theta = 0$). In other words, we need to prove that

$$dP_\theta = L(\theta)\, dP_0,$$

where P_θ is the probability measure that corresponds to the value of the parameter θ. It suffices to show that for all partitions $0 = t_0 < t_1 < \ldots < t_N \le T$ of the interval $[0,T]$ and for all cylinder sets $A \in \mathcal{F}_N$ the following equality holds:

$$\int_A dP_\theta = \int_A L(\theta)\, dP_0, \tag{9}$$

where $\mathcal{F}_N$ is the σ-algebra, generated by the values B_{t_k} of the process B_t at the points t_k, $k = 1, \ldots, N$. We have

$$\int_A dP_\theta = \int_A L^{(N)}(\theta)\, P_0,$$
$$\int_A L(\theta)\, dP_0 = \int_A \mathsf{E}_0[L(\theta) \mid \mathcal{F}_N]\, dP_0,$$

where $L^{(N)}$ is the likelihood function (3) for the discrete-time model, and $\mathsf{E}_0[\,\cdot \mid \mathcal{F}_N]$ is the conditional expectation corresponding to the probability measure P_0. To prove (9), it suffices to show that

$$L^{(N)}(\theta) = \mathsf{E}_0[L(\theta) \mid \mathcal{F}_N].$$

If $\theta = 0$, then $X_t = B_t$,

$$\mathsf{E}_0[L(\theta) \mid \mathcal{F}_N] = \mathsf{E} \exp\left\{\theta \int_0^T h_T(s)\, dB_s - \frac{\theta^2}{2} \int_0^T h_T(s) ds\right\}.$$

Due to joint normality of $\int_0^T h_T(s)\, dB_s$ and $\Delta B^{(N)}$, the conditional distribution of $\int_0^T h_T(s)\, dB_s$ with respect to $\mathcal{F}_N$ is Gaussian (Anderson 2003, Theorem 2.5.1); its conditional variance is nonrandom. Let us find its parameters. By the least squares method,

$$\mathsf{E}[B_t \mid \mathcal{F}_N] = \mathsf{E}\left[B_t \mid \Delta B^{(N)}\right] = \operatorname{cov}\left(B_t, \Delta B^{(N)}\right)\left(\operatorname{cov}\left(\Delta B^{(N)}, \Delta B^{(N)}\right)\right)^{-1} \Delta B^{(N)}.$$

We have $\operatorname{cov}\left(\Delta B^{(N)}, \Delta B^{(N)}\right) = \Gamma^{(N)}$. Calculate $\operatorname{cov}\left(B_t, \Delta B^{(N)}\right)$:

$$\operatorname{cov}\left(B_t,\ B_{t_k} - B_{t_{k-1}}\right) = \int_{s=0}^{t} \int_{u=t_{k-1}}^{t_k} K(s-u)\, du\, ds,$$

and for any vector $x = (x_k)_{k=1}^N \in \mathbb{R}^N$

$$\begin{aligned}
\operatorname{cov}\left(B_t, \Delta B^{(N)}\right) x &= \sum_{k=1}^{N} \operatorname{cov}\left(B_t,\ B_{t_k} - B_{t_{k-1}}\right) x_k = \int_{s=0}^{t} \sum_{k=1}^{N} \int_{u=t_{k-1}}^{t_k} K(s-u) x_k\, du\, ds = \\
&= \int_0^t \int_0^T K(s-u) M^* x(u)\, du\, ds = \int_0^T 1_{[0,t]}(s)\, \Gamma_T M^* x(s)\, ds = \\
&= \int_0^T \Gamma_T\, 1_{[0,t]}(s)\, M^* x(s)\, ds = \left(M \Gamma_T\, 1_{[0,t]}\right)^\top x,
\end{aligned}$$

whence we get

$$\operatorname{cov}\left(B_t, \Delta B^{(N)}\right) = \left(M\Gamma_T \mathbf{1}_{[0,t]}\right)^\top.$$

Therefore

$$\mathsf{E}[B_t \mid \mathcal{F}_N] = \left(M\Gamma_T \mathbf{1}_{[0,t]}\right)^\top \left(\Gamma^{(N)}\right)^{-1} \Delta B^{(N)} = \int_0^t \left(\Gamma_T M^* \left(\Gamma^{(N)}\right)^{-1} \Delta B^{(N)}\right)(s)\, ds.$$

Then

$$\begin{aligned}\mathsf{E}\left[\int_0^T h_T(t)\, dB_t \Big| \mathcal{F}_N\right] &= \int_0^T h_T(s) \left(\Gamma_T M^* \left(\Gamma^{(N)}\right)^{-1} \Delta B^{(N)}\right)(s)\, ds \\ &= \int_0^T (\Gamma_T h_T)(s) \left(M^* \left(\Gamma^{(N)}\right)^{-1} \Delta B^{(N)}\right)(s)\, ds \\ &= (M\Gamma_T h_T)^\top \left(\Gamma^{(N)}\right)^{-1} \Delta B^{(N)},\end{aligned}$$

where we have used that the operator Γ_T is self-adjoint.
Further, $M\Gamma_T h_T = M\,\mathbf{1}_{[0,\mathrm{T}]} = z$, where the vector z is defined after (3). Hence,

$$\mathsf{E}\left[\int_0^T h_T(t)\, dB_t \Big| \mathcal{F}_N\right] = z^\top \left(\Gamma^{(N)}\right)^{-1} \Delta B^{(N)}.$$

In order to calculate the variance we apply the partition-of-variance equality

$$\operatorname{var}\left[\int_0^T h_T(t)\, dB_t\right] = \operatorname{var}\left(\mathsf{E}\left[\int_0^T h_T(t)\, dB_t \Big| \mathcal{F}_N\right]\right) + \operatorname{var}\left[\int_0^T h_T(t)\, dB_t \Big| \mathcal{F}_N\right].$$

We have

$$\operatorname{var}\left[\int_0^T h_T(t)\, dB_t\right] = \int_0^T (\Gamma_T h_T)(t)\, h_T(t)\, dt = \int_0^T \mathbf{1}_{[0,\mathrm{T}]}(t) h_T(t)\, dt = \int_0^T h_T(t)\, dt,$$

and

$$\operatorname{var}\left(\mathsf{E}\left[\int_0^T h_T(t)\, dB_t \Big| \mathcal{F}_N\right]\right) = \operatorname{var}\left(z^\top \left(\Gamma^{(N)}\right)^{-1} \Delta B^{(N)}\right) = z^\top \left(\Gamma^{(N)}\right)^{-1} z.$$

Hence,

$$\operatorname{var}\left[\int_0^T h_T(t)\, dB_t \Big| \mathcal{F}_N\right] = \int_0^T h_T(t)\, dt - z^\top \left(\Gamma^{(N)}\right)^{-1} z. \tag{10}$$

Applying the formula for the mean of the log-normal distribution, we obtain

$$\begin{aligned}\mathsf{E}_0[L(\theta) \mid \mathcal{F}_N] = \mathsf{E} \exp\Bigg\{&\theta z^\top \left(\Gamma^{(N)}\right)^{-1} \Delta B^{(N)} \\ &+ \frac{\theta^2}{2}\left(\int_0^T h_T(t)\, dt - z^\top \left(\Gamma^{(N)}\right)^{-1} z\right) - \frac{\theta^2}{2} \int_0^T h_T(s) ds\Bigg\} = L^{(N)}(\theta).\end{aligned}$$

Thus, (9) is proved. □

Corollary 3.4. *The maximum likelihood estimator of θ by continuous observations is given by*

$$\hat{\theta}_T = \frac{\int_0^T h_T(t)\, dX_t}{\int_0^T h_T(t)\, dt}. \tag{11}$$

3.2. Properties of the estimator

It follows immediately from (11) that the maximum likelihood estimator $\hat{\theta}_T$ is equal to

$$\hat{\theta}_T = \theta + \frac{\int_0^T h_T(t)\, dB_t}{\int_0^T h_T(t)\, dt}. \tag{12}$$

Proposition 3.5. *The estimator $\hat{\theta}_T$ is unbiased and normally distributed. Its variance is equal to*

$$\operatorname{var}\hat{\theta}_T = \mathsf{E}\left(\hat{\theta}_T - \theta\right)^2 = \frac{1}{\int_0^T h_T(t)\,dt}. \tag{13}$$

Proof. Unbiasedness and normality follows from the fact that $\hat{\theta} - \theta$ is a linear functional of centered Gaussian process B. By (7),

$$\operatorname{var}\left(\int_0^T h_T(t)\,dB_t\right) = \int_0^T \Gamma_T h(t)\,h_T(t)\,dt = \int_0^T 1_{[0,\mathrm{T}]}(t)\,h_T(t)\,dt = \int_0^T h_T(t)\,dt.$$

Thus, equation (13) immediately follows from (12). □

Corollary 3.6. *Let the process $B = \{B_t,\ t \geq 0\}$ satisfy Assumptions 1 and 2. If*

$$\int_0^T h_T(t)\,dt \to \infty, \quad \text{as} \quad T \to +\infty, \tag{14}$$

then the maximum likelihood estimator $\hat{\theta}_T$ is mean-square consistent, i.e., $\mathsf{E}(\hat{\theta}_T - \theta)^2 \to 0$, as $T \to +\infty$.

It can be hard to verify the condition (14). The following result gives sufficient conditions for the consistency in terms of the autocovariance function of B.

Theorem 3.7. *Let the process $B = \{B_t,\ t \geq 0\}$ satisfy Assumptions 1 and 2. If the covariance function of the increment process $B_N - B_{N-1}$ tends to 0:*

$$\mathsf{E}\,(B_{N+1} - B_N)\,B_1 \to 0 \quad \text{as} \quad N \to \infty,$$

then the maximum likelihood estimator $\hat{\theta}_T$ is mean-square consistent.

Proof. The estimator $\hat{\theta}^{(N)}$ from the discrete sample $\{X_1, \ldots, X_N\}$ is mean-square consistent by Theorem 2.5. The estimator from the continuous-time sample $\{X_t,\ t \in [0, T]\}$ is unbiased. Now compare the variances of the discrete and continuous-time estimators.

The desired inequalities are got from the proof of Theorem 2.5. Suppose that $T \geq 1$, N is an integer such that $N \leq T < N + 1$. By equation (10) we have

$$\int_0^T h_T(t)\,dt \geq z^\top \left(\Gamma^{(N)}\right)^{-1} z. \tag{15}$$

As $\operatorname{var}\hat{\theta}_T = \frac{1}{\int_0^T h_T(t)\,dt}$, $\operatorname{var}\hat{\theta}^{(N)} = \left(z^\top \left(\Gamma^{(N)}\right)^{-1} z\right)^{-1}$, we have $\operatorname{var}\hat{\theta}_T \leq \operatorname{var}\hat{\theta}^{(N)}$, and

$$\lim_{T\to+\infty} \mathsf{E}\left(\hat{\theta}_T - \theta\right)^2 = \lim_{N\to\infty} \mathsf{E}\left(\hat{\theta}^{(N)} - \theta\right)^2 = 0.$$ □

To prove the strong consistency of $\hat{\theta}_T$, we need the following auxiliary result.

Lemma 3.8. *Let the process B satisfy the conditions of Theorem 3.3. Then the estimator process $\hat{\theta} = \{\hat{\theta}_T, T \geq 0\}$ has independent increments.*

Proof. Let $T_2 \leq T_3$. Then

$$\begin{aligned}\mathsf{E}\left[\int_0^{T_3} h_{T_3}(t)\,dB_t \int_0^{T_2} h_{T_2}(s)\,dB_s\right] &= \int_0^{T_2} \Gamma_{T_3} h_{T_3}(t)\,h_{T_2}(t)\,dt \\ &= \int_0^{T_2} 1_{[0,T_3]}(t)\,h_{T_2}(t)\,dt = \int_0^{T_2} h_{T_2}(t)\,dt.\end{aligned}$$

Thus, if $0 < T_1 \le T_2 \le T_3 \le T_4$, then

$$\begin{aligned}
&\mathsf{E}(\hat\theta_{T_4} - \hat\theta_{T_3})(\hat\theta_{T_2} - \hat\theta_{T_1}) \\
&\quad = \mathsf{E}\left(\frac{\int_0^{T_4} h_{T_4}(t)\,dB_t}{\int_0^{T_4} h_{T_4}(t)\,dt} - \frac{\int_0^{T_3} h_{T_3}(t)\,dB_t}{\int_0^{T_3} h_{T_3}(t)\,dt}\right)\left(\frac{\int_0^{T_2} h_{T_2}(t)\,dB_t}{\int_0^{T_2} h_{T_2}(t)\,dt} - \frac{\int_0^{T_1} h_{T_1}(t)\,dB_t}{\int_0^{T_1} h_{T_1}(t)\,dt}\right) = \\
&\quad = \frac{\mathsf{E}\left[\int_0^{T_4} h_{T_4}(t)\,dB_t \int_0^{T_2} h_{T_2}(t)\,dB_t\right]}{\int_0^{T_4} h_{T_4}(t)\,dt \int_0^{T_2} h_{T_2}(t)\,dt} - \frac{\mathsf{E}\left[\int_0^{T_3} h_{T_3}(t)\,dB_t \int_0^{T_2} h_{T_2}(t)\,dB_t\right]}{\int_0^{T_3} h_{T_3}(t)\,dt \int_0^{T_2} h_{T_2}(t)\,dt} - \\
&\qquad - \frac{\mathsf{E}\left[\int_0^{T_4} h_{T_4}(t)\,dB_t \int_0^{T_1} h_{T_1}(t)\,dB_t\right]}{\int_0^{T_4} h_{T_4}(t)\,dt \int_0^{T_1} h_{T_1}(t)\,dt} + \frac{\mathsf{E}\left[\int_0^{T_3} h_{T_3}(t)\,dB_t \int_0^{T_1} h_{T_1}(t)\,dB_t\right]}{\int_0^{T_3} h_{T_3}(t)\,dt \int_0^{T_1} h_{T_1}(t)\,dt} = \\
&\quad = \frac{1}{\int_0^{T_4} h_{T_4}(t)\,dt} - \frac{1}{\int_0^{T_3} h_{T_3}(t)\,dt} - \frac{1}{\int_0^{T_4} h_{T_4}(t)\,dt} + \frac{1}{\int_0^{T_3} h_{T_3}(t)\,dt} = 0.
\end{aligned}$$

Similarly to the proof of Lemma 2.6, the random process $\hat\theta_T$ is Gaussian and its increments are proved to be uncorrelated so they are independent. □

Theorem 3.9. *Under conditions of Theorem 3.7 the estimator $\hat\theta_T$ is strongly consistent.*

Proof. By Kolmogorov's inequality, for any $\epsilon > 0$ and $t_0 > 0$

$$\begin{aligned}
\mathsf{P}\left(\sup_{T>t_0} |\hat\theta_T - \theta| > \epsilon\right) &\le \mathsf{P}\left(|\hat\theta_{t_0} - \theta| > \frac{\epsilon}{2}\right) + \mathsf{P}\left(\sup_{T>t_0} |\hat\theta_T - \theta_{t_0}| > \frac{\epsilon}{2}\right) \le \\
&\le \frac{4}{\epsilon^2}\operatorname{var}\hat\theta_{t_0} + \frac{4}{\epsilon^2}\lim_{T\to+\infty}\operatorname{var}\left(\hat\theta_T - \hat\theta_{t_0}\right) = \frac{8}{\epsilon^2}\operatorname{var}\hat\theta_{t_0}.
\end{aligned}$$

By Theorem 3.7,

$$\lim_{t_0\to+\infty} \mathsf{P}\left(\sup_{T>t_0} |\hat\theta_T - \theta| > \epsilon\right) = 0 \quad \text{for all} \quad \epsilon > 0,$$

whence the strong consistency follows. □

Remark 3.10. The Brownian motion does not satisfy Assumption 1 (as for covariance function $\max(s,t)$ of Wiener process, $\frac{\partial \max(s,t)}{\partial t}$ is not continuous in s). So we extend our model such that it can handle Wiener process. Let the process B be a sum of two independent random processes,

$$B_t = B_t^{\mathrm{C}} + W_t, \tag{16}$$

where B^{C} satisfies Assumption 1, and W is a standard Wiener process. Let us look at the changes of the statements if the process B admits representation (16) (Assumption 1 for B is dropped). Lemma 3.1 changes as follows:

$$\mathsf{E}\left[\int_0^T f(t)\,dB_t \int_0^T g(s)\,dB_s\right] = \int_0^T f(t)\int_0^T K(t-s)g(s)\,ds\,dt + \int_0^T f(t)g(t)\,dt.$$

Equation (7) will stand true, if we set

$$\Gamma_T f(t) = f(t) + \Gamma_T^{\mathrm{C}} f(t) = f(t) + \int_0^T K(t-s)f(s)\,ds.$$

Lemma 3.2 stands true (with $\|\Gamma_T\| \le \|K\|_{L_1[-T,T]} + 1$). Theorem 3.3 holds true; minor changes in the proof are required.

Table 1: The means and variances of $\hat{\theta}_T$

H	T	Sample Mean	Sample Variance	Theoretical Variance
0.6	1	2.0344	1.9829	1.8292
	10	2.0047	0.2584	0.2356
0.7	1	1.9995	1.9956	1.9692
	10	2.0296	0.3484	0.3270
0.8	1	2.0165	2.0435	1.9930
	10	2.0071	0.5155	0.4392
0.9	1	2.0117	2.0376	1.9984
	10	2.0099	0.7710	0.5867

3.3. Examples

Example 3.11. Let B be a fractional Brownian motion with the Hurst index $H \in (1/2, 1)$. Then $K(t) = \frac{H(2H-1)}{|t|^{2-2H}}$. We denote by Γ_T^H the corresponding operator Γ_T. Then for the function

$$h_T(s) = C_H s^{1/2-H}(T-s)^{1/2-H},$$

$C_H = \left(H(2H-1)\mathrm{B}\left(H-\frac{1}{2}, \frac{3}{2}-H\right)\right)^{-1}$, we have that $\Gamma_T^H h_T = 1_{[0,\mathrm{T}]}$, see Norros *et al.* (1999). The maximum likelihood estimator is given by

$$\hat{\theta}_T = \frac{T^{2H-2}}{\mathrm{B}(3/2-H,\, 3/2-H)} \int_0^T s^{1/2-H}(T-s)^{1/2-H}\, dX_s.$$

Example 3.12. Consider the following model:

$$X_t = \theta t + W_t + B_t^H, \tag{17}$$

where W is a standard Wiener process, B^H is a fractional Brownian motion with Hurst index H, and random processes W_t and B_t^H are independent. The process $W_t + B_t^H$ admits representation (16) with $B^{\mathrm{C}} = B^H$. Corresponding operator Γ_T is $\Gamma_T = I + \Gamma_T^H$ (see Example 3.11 for the definition of Γ_T^H). The operator Γ_T^H is self-adjoint and positive semi-definite. Hence, the operator Γ_T is invertible. Thus Assumption 2 holds true.

The function $h_T = \Gamma_T^{-1} 1_{[0,\mathrm{T}]}$ can be evaluated iteratively

$$h_T = \sum_{k=0}^{\infty} \frac{\left(\frac{1}{2}\left\|\Gamma_T^H\right\| I - \Gamma_T^H\right)^k 1_{[0,\mathrm{T}]}}{\left(1+\frac{1}{2}\left\|\Gamma_T^H\right\|\right)^{k+1}}. \tag{18}$$

3.4. Simulations

We illustrate the behavior of the maximum likelihood estimator for Example 3.12 by means of simulation experiments. For $T = 1$ and $T = 10$ and various values of H we find h_T iteratively by (18). Then for $\theta = 2$ we simulate 1000 realizations of the process (17) for each H and compute the estimates by (11). The means and variances of these estimates are reported in Table 1. The theoretical variances calculated by (13) are also presented. We see that these simulation studies confirm the theoretical properties of $\widehat{\theta}_T$, especially unbiasedness and consistency.

References

Anderson TV (2003). *An Introduction to Multivariate Statistical Analysis*. Wiley, Hoboken NJ.

Bertin K, Torres S, Tudor CA (2011). "Maximum-likelihood Estimators and Random Walks in Long Memory Models." *Statistics*, **45**(4), 361–374.

Cai C, Chigansky P, Kleptsyna M (2016). "Mixed Gaussian Processes: A Filtering Approach." *Ann. Probab.*, **44**(4), 3032–3075.

Cramér H, Leadbetter MR (2004). *Stationary and Related Stochastic Processes*. Dover Publications, Inc., Mineola, NY. Sample function properties and their applications, Reprint of the 1967 original.

Hu Y, Nualart D, Xiao W, Zhang W (2011). "Exact Maximum Likelihood Estimator for Drift Fractional Brownian Motion at Discrete Observation." *Acta Math. Sci. Ser. B Engl. Ed.*, **31**(5), 1851–1859.

Huang ST, Cambanis S (1978). "Stochastic and Multiple Wiener Integrals for Gaussian Processes." *Ann. Probab.*, **6**(4), 585–614.

Le Breton A (1998). "Filtering and Parameter Estimation in a Simple Linear System Driven by a Fractional Brownian Motion." *Statist. Probab. Lett.*, **38**(3), 263–274.

Mishura Y (2008). *Stochastic Calculus for Fractional Brownian Motion and Related Processes*, volume 1929. Springer Science & Business Media.

Mishura Y (2016). "Maximum Likelihood Drift Estimation for the Mixing of Two Fractional Brownian Motions." In *Stochastic and Infinite Dimensional Analysis*, pp. 263–280. Springer.

Mishura Y, Voronov I (2015). "Construction of Maximum Likelihood Estimator in the Mixed Fractional–Fractional Brownian Motion Model with Double Long-range Dependence." *Mod. Stoch. Theory Appl.*, **2**(2), 147–164.

Norros I, Valkeila E, Virtamo J (1999). "An Elementary Approach to a Girsanov Formula and Other Analytical Results on Fractional Brownian motions." *Bernoulli*, **5(4)**, 571–587.

Rao CR (2002). *Linear Statistical Inference and its Applications*. Second edition. Wiley, New York. Wiley Series in Probability and Statistics.

Stiburek D (2017). "Statistical Inference on the Drift Parameter in Fractional Brownian Motion with a Deterministic Drift." *Comm. Statist. Theory Methods*, **46**(2), 892–905.

Affiliation:

Yuliya Mishura, Kostiantyn Ralchenko, Sergiy Shklyar
Department of Probability Theory, Statistics and Actuarial Mathematics
Taras Shevchenko National University of Kyiv
64 Volodymyrska
01601 Kyiv, Ukraine
E-mail: myus@univ.kiev.ua, k.ralchenko@gmail.com, shklyar@univ.kiev.ua

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/

Volume 46 *Submitted:* 2016-11-15
April 2017 *Accepted:* 2017-02-02

Austrian Journal of Statistics
April 2017, Volume 46, 79–87.
http://www.ajs.or.at/
doi:10.17713/ajs.v46i3-4.673

Extracting Information from Interval Data Using Symbolic Principal Component Analysis

M. R. Oliveira, M. Vilela, A. Pacheco
CEMAT and Instituto Superior Técnico,
Universidade de Lisboa, Portugal

Rui Valadas
IT and Instituto Superior Técnico
Universidade de Lisboa, Portugal

Paulo Salvador
IT and Universidade de Aveiro, Portugal

Abstract

We introduce generic definitions of symbolic variance and covariance for random interval-valued variables, that lead to a unified and insightful interpretation of four known symbolic principal component estimation methods: CPCA, VPCA, CIPCA, and SymCov-PCA. Moreover, we propose the use of truncated versions of symbolic principal components, that use a strict subset of the original symbolic variables, as a way to improve the interpretation of symbolic principal components. Furthermore, the analysis of a real dataset leads to a meaningful characterization of Internet traffic applications, while highligting similarities between the symbolic principal component estimation methods considered in the paper.

Keywords: interval data, symbolic principal component analysis, Internet data.

1. Introduction

The low cost of information storage combined with recent advances in search and retrieval technologies has made available huge amounts of data, the so-called *big data* explosion. New statistical analysis techniques are now required to deal with the volume and complexity of this data. One promising technique is Symbolic Data Analysis (SDA), introduced by E. Diday (1987).

In conventional data analysis, the variables that characterize an object can only take single values. SDA introduces symbolic random variables which can take values over complex data structures like lists, intervals, histograms or even distributions (Billard and Diday 2006). Symbolic data may exist on their own right or may result from the aggregation of a base dataset according to the researchers interest. For example, suppose that our goal is to characterize the ages of university teachers. The variable that records the teachers' age will have as many observations as teachers, and these can differ among universities. Let us assume that a given university has 1000 teachers, and the values $\omega_1, \ldots, \omega_{1000}$ are the teachers' ages. SDA

calls these values *micro-data*. In conventional statistical analysis, the universities would have to be characterized by single-valued variables, e.g. the mean teachers' age. SDA can deal with more complex data structures, called *macro-data*. For example, the teachers' age can be aggregated into one or multiple intervals. Our main interest in this paper is on interval-valued data, where *macro-data* corresponds to the interval between minimum and maximum of *micro-data* values: $[a, b] = [\min\{\omega_1, \ldots, \omega_{1000}\}, \max\{\omega_1, \ldots, \omega_{1000}\}]$.

The paper is organized as follows. Section 2 presents basic descriptive statistics, including symbolic variances and covariances, for interval-valued data. Section 3 introduces Symbolic Principal Component Analysis (SPCA) for interval-valued data. Section 4 uses SPCA on the analysis of Internet data produced by six different Internet applications. Finally, some conclusions are drawn in Section 5.

2. Basic descriptive statistics

There have been several proposals for definitions of symbolic versions of sample mean, variance, covariance, and correlation, according to various types of symbolic data, including interval-valued data (Billard and Diday 2006).

We assume that the collected interval-valued data are realizations of random vectors. As such, we consider a random interval-valued vector $\boldsymbol{X} = (X_1, \ldots, X_p)^t$, where $X_j = [A_j, B_j]$, with A_j and B_j being random variables verifying $P(A_j \leq B_j) = 1$, denotes the j-th random interval-valued variable of $\boldsymbol{X}$. Even though this is the common representation of random interval-valued variables, we follow the approach of Vilela (2015) and write the intervals X_j in terms of their centers, $C_j = (A_j + B_j)/2$, and their ranges, $R_j = B_j - A_j$. Equivalently, $[A_j, B_j] = [C_j - R_j/2\,,\, C_j + R_j/2]$. This choice leads to a clear interpretation of an interval in terms of its "location" on the real line along with its length; moreover it enables for the unification of several results in the literature (cf. Vilela (2015) and references therein). Likewise, the random vector $\boldsymbol{X}$ is equivalently represented by the random vector of centers, $\boldsymbol{C} = (C_1, \ldots, C_p)^t$, and the random vector of ranges, $\boldsymbol{R} = (R_1, \ldots, R_p)^t$.

Let $(\boldsymbol{C}_1, \ldots, \boldsymbol{C}_n)^t$ and $(\boldsymbol{R}_1, \ldots, \boldsymbol{R}_n)^t$ denote the vectors of centers and ranges obtained from a random sample of size n from $\boldsymbol{X}$, where $\boldsymbol{C}_i = (C_{i1}, \ldots, C_{ip})^t$ and $\boldsymbol{R}_i = (R_{i1}, \ldots, R_{ip})^t$ characterize the i-th entity or object of the sample. In this setting, a natural proposal for sample symbolic mean of the interval-valued variable X_j is to use the traditional sample mean of the centers, $\overline{X}_j = \overline{C}_j$ with $\overline{C}_j = \sum_{i=1}^n C_{ij}/n$ (Billard and Diday 2006).

As concerns the sample symbolic variance of the interval-valued variable X_j, we express the proposals available in the literature as the sum of two components, the first accounting for the variability of the associated centers and the second for the size of the associated ranges, in the form

$$S_{jj}^{(\alpha)} = \sum_{i=1}^{n} \frac{\left(C_{ij} - \overline{C}_j\right)^2}{n} + \alpha \sum_{i=1}^{n} \frac{R_{ij}^2}{n}, \tag{1}$$

with the nonnegative weight α accounting for the relevance given to the sizes of the ranges. In particular, we address three cases, with respective values 0, 1/4, 1/12, for the weight α. The first case ($\alpha = 0$) ignores the contribution of the ranges, simply turning the symbolic variance into the variance of the centers (Billard and Diday 2006). Concerning the second case ($\alpha = 1/4$), we note that as $R_{ij}/2$ represents the radius of the j-th random interval-valued variable measured at the i-th entity, thus $\sum_{i=1}^n R_{ij}^2/(4n)$ may be interpreted as the sample second order moment of the radius of the j-th random interval-valued variable. This was originally proposed by De Carvalho, Brito, and Bock (2006). The third case, presented in Bertrand and Goupil (2000), corresponds to choosing the weight $\alpha = 1/12$, which is derived assuming that *micro-data* are uniformly distributed on the associated *macro-data* interval.

In the same manner, we consider proposals for the sample symbolic covariance between two interval-valued variables X_j and X_l that express it as the sum of two components, the first

accounting for the sample covariance of the associated centers and the second for the size of the associated ranges, in the form

$$S_{jl}^{(\beta)} = \sum_{i=1}^{n} \frac{(C_{ij} - \overline{C}_j)(C_{il} - \overline{C}_l)}{n} + \beta \sum_{i=1}^{n} \frac{R_{ij} R_{il}}{n}, \tag{2}$$

with the nonnegative weight β accounting for the relevance given to the sizes of the ranges associated to the interval-valued variables X_j and X_l. The case $\beta = 0$ was introduced by Billard and Diday (2003), $\beta = 1/12$ by Billard (2008), and finally $\beta = 1/4$ by Vilela (2015).

In sequence, we may use (1)-(2) to construct a sample symbolic covariance matrix $\boldsymbol{S}^{(\alpha,\beta)}$ having on the diagonal the sample symbolic variances $S_{jj}^{(\alpha)}$, given in (1), and outside the diagonal the sample symbolic covariances $S_{jl}^{(\beta)}$, $j \neq l$, given in (2), leading to

$$\boldsymbol{S}^{(\alpha,\beta)} = \boldsymbol{S}_{CC} + (\alpha - \beta)\, \mathrm{D}iag\left(\frac{\boldsymbol{\mathcal{R}}\boldsymbol{R}^t}{n}\right) + \beta\, \frac{\boldsymbol{\mathcal{R}}\boldsymbol{R}^t}{n}, \tag{3}$$

with $\boldsymbol{S}_{CC}$ denoting the sample covariance matrix of the centers and $\boldsymbol{\mathcal{R}} = [R_{ij}]$ the $(n \times p)$ matrix of random sample ranges. Particular cases of sample symbolic covariance matrices, $\boldsymbol{S}^{(\alpha,\beta)}$, with $\alpha \in \{0, 1/12, 1/4\}$ and $\beta = \alpha$ or $\beta = 0$ have been introduced in the literature (vide Vilela (2015) and references therein). Details about the links between these sample symbolic covariance matrices and SPCA for interval-valued data are discussed in the next section.

3. Symbolic principal component analysis

Principal component analysis (PCA) is one of the most popular statistical methods to analyse real data. There have been several proposals to extend this methodology to the symbolic data analysis framework, in particular to interval-valued data. The majority of the available methods rely on a strategy called symbolic-conventional-symbolic, meaning that: (i) input data is symbolic (interval-valued, in here), (ii) the data is converted into conventional, to which the conventional PCA method is applied, and (iii) at the end, the PCA results are turned into symbolic, usually by a method called Maximum Covering Area Rectangle (MCAR).

We study four SPCA methods: centers (CPCA) and vertices (VPCA) methods, presented in Cazes, Chouakria, Diday, and Schektman (1997); the Complete Information PCA method (CIPCA), introduced by Wang, Guan, and Wu (2012); and Symbolic Covariance PCA (SymCovPCA), proposed by Le-Rademacher and Billard (2012). CPCA and VPCA corresponds to the first SPCA methods proposed in the literature and the last two are among the most recent alternatives. All these methods rely on the symbolic-conventional-symbolic strategy, which can be specified as follows: (i) compute the associated $(p \times p)$ sample symbolic covariance matrix $\boldsymbol{S}^{(\alpha,\beta)}$ (vide Table 1); (ii) obtain the spectral decomposition of $\boldsymbol{S}^{(\alpha,\beta)}$, as in the conventional PCA, and (iii) transform the conventional scores into symbolic scores, e.g. using MCAR.

Note that $\boldsymbol{S}^{(\frac{1}{4},0)}$ and $\boldsymbol{S}^{(\frac{1}{12},0)}$ (vide Table 1) are covariance matrices that use a definition of symbolic variance of an interval-valued variable that does not coincide with the definition of symbolic covariance between the same interval-valued variable and itself. On one hand, this violates a basic rule in the conventional framework, namely that the variance of a variable equals the covariance of the variable with itself. On the other hand, $\boldsymbol{S}^{(\frac{1}{4},0)}$ and $\boldsymbol{S}^{(\frac{1}{12},0)}$ can be seen as the sum of two symmetric semi-positive definite matrices, which guarantees that they are themselves symmetric semi-positive definite matrices and, therefore, its eigenvalues and eigenvectors verify the usual properties associated with conventional principal components. Overall, Wang *et al.* (2012), who proposed $\boldsymbol{S}^{(\frac{1}{12},0)}$, argue that the variance of a variable being defined differently from the covariance of the variable with itself turns into an advantage of their method.

Table 1: Sample symbolic covariance matrices, $\boldsymbol{S}^{(\alpha,\beta)}$, defined by the combination of several proposals for symbolic variances and covariances along with the corresponding SPCA method.

(α,β)	$\boldsymbol{S}^{(\alpha,\beta)}$	SPCA Method
(0,0)	$\boldsymbol{S}_{CC}$	CPCA
$(\frac{1}{4},\frac{1}{4})$	$\boldsymbol{S}_{CC}+\frac{1}{4}\frac{\boldsymbol{\mathcal{R}}\boldsymbol{R}^t}{n}$	—
$(\frac{1}{12},\frac{1}{12})$	$\boldsymbol{S}_{CC}+\frac{1}{12}\frac{\boldsymbol{\mathcal{R}}\boldsymbol{R}^t}{n}$	SymCovPCA
$(\frac{1}{4},0)$	$\boldsymbol{S}_{CC}+\frac{1}{4}\mathrm{D}iag\left(\frac{\boldsymbol{\mathcal{R}}\boldsymbol{R}^t}{n}\right)$	VPCA
$(\frac{1}{12},0)$	$\boldsymbol{S}_{CC}+\frac{1}{12}\mathrm{D}iag\left(\frac{\boldsymbol{\mathcal{R}}\boldsymbol{R}^t}{n}\right)$	CIPCA

Similarly to the conventional PCA, it may be interesting to define the SPCA based on standardized interval-valued variables, and to do so we introduce the sample correlation matrix as: $\boldsymbol{P}^{(\alpha,\beta)}=\boldsymbol{U}_{(\alpha)}^{-1}\boldsymbol{S}^{(\alpha,\beta)}\boldsymbol{U}_{(\alpha)}^{-1}$, where $\boldsymbol{U}_{(\alpha)}=\mathrm{D}iag\left(S_{11}^{(\alpha)},\ldots,S_{pp}^{(\alpha)}\right)^{1/2}$, for $\boldsymbol{S}^{(\alpha,\beta)}=[S_{jl}^{(\alpha,\beta)}]$, where $S_{jj}^{(\alpha,\beta)}=S_{jj}^{(\alpha)}$ and $S_{jl}^{(\alpha,\beta)}=S_{jl}^{(\beta)}$, for $j\neq l$. Equivalently, $\boldsymbol{S}^{(\alpha,\beta)}=\boldsymbol{U}_{(\alpha)}\boldsymbol{P}^{(\alpha,\beta)}\boldsymbol{U}_{(\alpha)}$. Thus, SPCA methods based on standardized interval-valued variables just have to use $\boldsymbol{P}^{(\alpha,\beta)}$ instead of $\boldsymbol{S}^{(\alpha,\beta)}$.

The most common way to transform conventional objects into symbolic ones for methods following the symbolic-conventional-symbolic strategy is the MCAR representation. This representation was introduced by Chouakria (1998) to obtain the symbolic scores of the CPCA and VPCA methods, but can be used with any other method following the symbolic-conventional-symbolic strategy. Accordingly with this proposal, the sample interval-value score of the i-th object on the j-th symbolic principal component (SPC) is:

$$\widehat{\mathrm{SPC}}_{ij}=\Big[\mathrm{PC}_j(\min i),\mathrm{PC}_j(\max i)\Big], \tag{4}$$

where $j=1,\ldots,p,\ i=1,\ldots,n$ and $\hat{\boldsymbol{\gamma}}_j$ is the j-th eigenvector of $\boldsymbol{S}^{(\alpha,\beta)}$, the sample symbolic covariance matrix under consideration. Moreover, the lower bound is formed by the linear combinations of the lower bounds of the original intervals in case of positive weights, $\hat{\gamma}_{kj}>0$, plus the combination of the upper bounds if the weights are negative, $\hat{\gamma}_{kj}<0$, leading to

$$\mathrm{PC}_j(\min i)=\sum_{l:\,\hat{\gamma}_{lj}>0}(a_{il}-\bar{c}_l)\,\hat{\gamma}_{lj}\ +\sum_{l:\,\hat{\gamma}_{lj}<0}(b_{il}-\bar{c}_l)\,\hat{\gamma}_{lj}, \tag{5}$$

where $\bar{c}_l=\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\frac{a_{il}+b_{il}}{2}$. Likewise,

$$\mathrm{PC}_j(\max i)=\sum_{l:\,\hat{\gamma}_{lj}>0}(b_{il}-\bar{c}_l)\,\hat{\gamma}_{lj}\ +\sum_{l:\,\hat{\gamma}_{lj}<0}(a_{il}-\bar{c}_l)\,\hat{\gamma}_{lj}. \tag{6}$$

Moreover, the hyper-rectangle formed by the first k SPC, $(\widehat{\mathrm{SPC}}_{i1},\ldots,\widehat{\mathrm{SPC}}_{ik})^t$ is the MCAR k-dimensional representation of the i-th object.

4. Analysis of Internet data

In this section we illustrate the use of SPCA through a dataset of Internet traffic, typically observed in backbone networks, and measured during July 2014. Specifically, the dataset

contains traffic produced by six different Internet applications, namely Web browsing (produced by HTTP), file sharing (produced by Torrent), streaming, video (YouTube), port scans (produced by NMAP), and snapshots. The first four applications correspond to regular traffic and the last two to Internet attacks. The analysis usually aims at detecting the various Internet applications within a traffic aggregate and/or the separation between regular and illicit traffic.

The dataset comprises 917 traffic objects, corresponding to packet flows of specific applications, which we call *datastreams*. For each datastream, we registered five different traffic characteristics observed in 0.1 seconds intervals, during 5 minutes. The traffic characteristics registered were the following: number of upstream packets (PUp), number of downstream packets (PDw), number of upstream bytes (BUp), number of downstream bytes (BDw), and number of active TCP sessions (Ses). Thus, each object is characterized by a total of 3000 observations per traffic characteristic, which constitutes our *micro-data*.

The conventional approach to analyse this data is based on summary statistics of each traffic characteristic. In particular, (Pascoal, Oliveira, Valadas, Filzmoser, Salvador, and Pacheco 2012; Pascoal 2014) used 8 summary statistics (minimum, 1^{st} quartile, median, mean, 3^{rd} quartile, maximum, standard deviation, and median absolute deviation) for the above five traffic characteristics, giving a total of 40 variables to describe the datastreams. This approach usually requires a pre-processing step to remove irrelevant and redundant variables; Pascoal (2014) used a robust feature selection method based on mutual information for that purpose.

This dataset is naturally symbolic, since each traffic characteristic is multi-valued. SDA takes into consideration the complex structure of these data, and may lead to clearer interpretation and new insights. In our case, we will use interval-valued variables for each traffic characteristic (our *macro-data*), instead of the 8 summary statistics listed above.

Given the nature of the data and the existence of potential atypical observations among the *micro-data*, we decided to trim 1% of the lower and 1% of the higher values. This was only done for the regular applications given that illicit ones have few datastreams and small variability and would be completely eliminated from the dataset, even for such small trimming percentiles. Apart from that, and following the recommendations in (Pascoal *et al.* 2012; Pascoal 2014), data was smoothed using a logarithm transformation ($\ln(x+1)$, to overcome the existence of zeros). SPCA, estimated according with the four methods under study, was applied to this dataset and percentages of explained variance from the conventional and symbolic approach are summarized in Table 2.

Table 2: Eigenvalues of the sample symbolic covariance matrices for each estimation method, and associated cumulative percentage of total conventional variance.

	CPCA		SymCovPCA		VPCA		CIPCA	
	$\hat{\lambda}_j$	Cumul.	$\hat{\lambda}_j$	Cumul.	$\hat{\lambda}_j$	Cumul.	$\hat{\lambda}_j$	Cumul.
PC_1	4.292	65.4%	18.731	87.5%	25.939	50.8%	10.999	51.4%
PC_2	1.625	90.1%	1.736	95.6%	16.743	83.6%	6.178	80.3%
PC_3	0.540	98.4%	0.740	99.1%	3.584	90.6%	1.900	89.1%
PC_4	0.098	99.9%	0.181	100.0%	2.734	95.9%	1.401	95.7%
PC_5	0.009	100.0%	0.011	100.0%	2.072	100.0%	0.922	100.0%

The conventional analysis of Table 2 suggests to retain 2 principal components, explaining between 80.3% (CIPCA) and 95.6% (SymCovPCA) of the total sample variance associated with $\boldsymbol{S}^{(\alpha,\beta)}$, meaning that e.g. for CIPCA, $(\hat{\lambda}_1+\hat{\lambda}_2)/\sum_{j=1}^{5}\hat{\lambda}_j = 0.803$.

The results obtained with CPCA and SymCovPCA are similar, and so are the results obtained with VPCA and CIPCA. Moreover, these similarities are easily explained by the expressions of Table 1.

Table 3 shows the loadings of the first and second SPC, obtained with the four methods. In the case of SymCovPCA, the number of upstream and downstream bytes (BUp, BDw) have the highest loading (on absolute value) in the definition of the first SPC. Thus, the center

Table 3: Eigenvectors of the sample symbolic covariance matrices for each estimation method, called loadings.

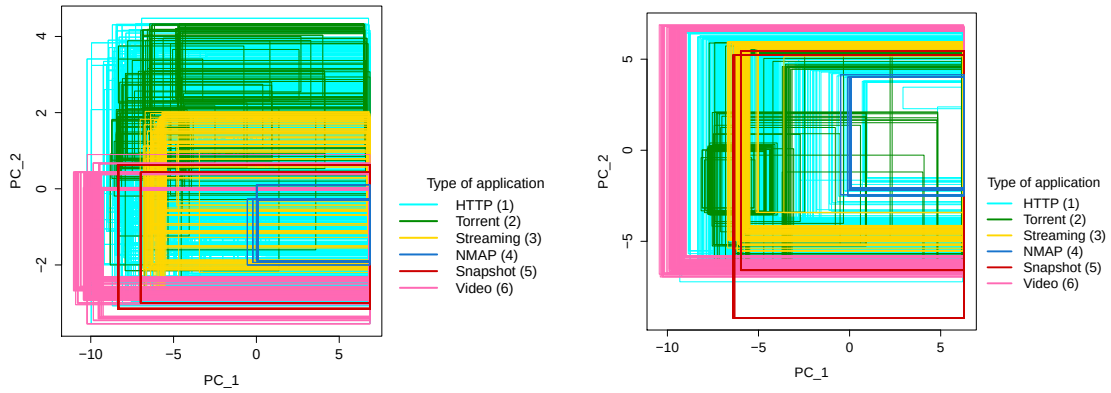
	CPCA		SymCovPCA		VPCA		CIPCA	
	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_1$	$\hat{\gamma}_2$
$\ln(\mathrm{P}Dw+1)$	-0.316	-0.152	-0.264	-0.171	-0.049	-0.036	-0.125	-0.059
$\ln(\mathrm{B}Dw+1)$	**-0.752**	-0.016	**-0.730**	-0.043	**-0.986**	0.157	**-0.932**	0.337
$\ln(\mathrm{P}Up+1)$	-0.295	-0.151	-0.255	-0.168	-0.044	-0.038	-0.113	-0.070
$\ln(\mathrm{B}Up+1)$	**-0.491**	0.050	**-0.571**	0.075	-0.153	**-0.986**	-0.318	**-0.937**
$\ln(\mathrm{S}es+1)$	-0.082	**0.976**	-0.079	**0.967**	-0.011	-0.012	-0.029	-0.027

and range of the first SPC can be interpreted as a weighted sum of the number of upstream and downstream bytes. The number of bytes is sometimes referred to as the traffic volume. For the center, the negative coefficients indicate that datastreams with high (low) number of bytes in both directions have low (high) center values on the first SPC. For the range, the coefficients are taken in absolute value, so datastreams with high (low) number of bytes in both directions have high (low) range values on the first SPC. Recall that the range expresses the inner variability of *micro-data*. As for the second SPC, the loading associated with number of sessions stands out. Thus, datastreams characterized by an high (low) number of sessions have high (low) center and range values on the second SPC.

The SymCovPCA scores are shown in Figure 1(a). Each datastream is represented by a rectangle, defined by the centers and ranges of the first two SPC. It can be said that the various Internet applications are, in general, well identified, since the datastreams show similar patterns for the same application. Most datastreams have a small minimum traffic volume (number of bytes), with the corresponding rectangles leaning to the right side. HTTP shows no distinctive characteristic, since the datastreams spread over all score ranges. This can be explained by the heterogeneity of user behaviours and accessed Web pages, typical of Web browsing. Torrent is concentrated on the upper part of the graph, due to its high number of sessions. The high number of sessions and large variability of the traffic volume is mostly explained by the variation on the number of available peers during traffic sharing sessions. The graph also suggests the existence of several Torrent groups, but this pattern will become clearer with the CIPCA method. The behaviour of video related with the second SPC contrasts with that of Torrent: it is concentrated in the lower part of the graph, due to its low number of sessions. Moreover, video is the application with the highest traffic volume. We may say that video datastreams are characterized by a low number of high volume sessions, and Torrent by a high number of high volume sessions. Streaming has a behaviour similar to video, but with higher number of sessions and lower traffic volume. NMAP is the application with smallest volume and variability, and has also a relatively low number of sessions. Finally, the behaviour of snapshot is in-between video and streaming, both in terms of volume and number of sessions. Snapshot has two clear groups, that differ on the peak traffic volume, and correspond to full desktop and partial desktop uploads, respectively.

Table 3 shows that the loadings obtained with CIPCA are much higher (in absolute value) for BDw (first component) and BUp (second component). Thus, the first SPC can be interpreted as the number of bytes down (BDw) and the second one as the number of bytes up (BUp). The CIPCA scores are shown in Figure 1(b). Snapshot has the highest upstream peak traffic volume, and is now better separated from video and streaming. NMAP is again the application with smaller rectangles. However, it is now better separated from HTTP, since most HTTP datastreams have higher traffic volume range simultaneously in the upstream and downstream directions. Video and streaming are also well separated, since video datastreams have consistently higher traffic volume ranges simultaneously in both directions. Regarding Torrent, it is now possible to distinguish among three groups: the group centers occur at approximately the same upstream traffic volume; one group has small traffic range in both directions (small rectangles) and high downstream volume, another has high traffic ranges in the downstream direction but small in the upstream direction, and a third one has small

downstream volumes but high upstream traffic ranges. These groups emerge from differences on the relative location of peers and the quality/stability of links. The first group corresponds to closer peers from which it is possible to download at higher speeds, the third to farther peers for which the links are less stable and unable to download at high speeds, and the third group is a mixture of the two previous ones.



(a) SymCovPCA. (b) CIPCA.

Figure 1: Symbolic scores, estimated by MCAR method.

To validate these interpretations, we follow an approach similar to Cadima and Jolliffe (2001) where we consider the truncated SPC with respective loadings, $\hat{\gamma}_j^{tr}$, equal to $\hat{\gamma}_j$ for the input variables considered to be relevant for interpreting $\hat{\gamma}_j$ (highlighted in boldface in Table 3) and zero otherwise. With these truncated weights, we compute the corresponding centers and ranges of the truncated SPC. Finally, we calculate the correlations between the correspondent truncated and complete (i.e., non-truncated) ranges and centers. If these correlations are close to 1, then the selection of variables performed to give a meaning to the j-th SPC by considering its truncated version gets validated. In our case, for CPCA and SymCovPCA we made the following selection: $\hat{\gamma}_1^{tr} = (0, \hat{\gamma}_{21}, 0, \hat{\gamma}_{41}, 0)^t$ and $\hat{\gamma}_2^{tr} = (0, 0, 0, 0, \hat{\gamma}_{51})^t$ and for VPCA and CIPCA we choose $\hat{\gamma}_1^{tr} = (0, \hat{\gamma}_{21}, 0, 0, 0)^t$ and $\hat{\gamma}_2^{tr} = (0, 0, 0, \hat{\gamma}_{42}, 0)^t$. In sequence, the first two truncated SPC of the centers and ranges are calculated as well as their sample correlation with the corresponding non-truncated SPC of centers and ranges, with the resulting values being summarized in Table 4. Note that all these correlations are close to 1, with the exception of the correlation between the truncated and complete versions of $SPC_2^{SymCovPCA}$ ranges. We note that if the inclusion of an additional input variable for the interpretation of $SPC_2^{SymCovPCA}$ ranges has the merit of increasing the correlation between the truncated and complete versions from 0.774 to 0.902. Nevertheless, it has the drawback of making the interpretation of $SPC_2^{SymCovPCA}$ ranges more complex and difficult.

Table 4: Sample Correlation between non-truncated, SPC_j^M, and truncated, $[SPC_j^M]^{tr}$, symbolic principal components centers, $C[\cdot]$, and ranges, $R[\cdot]$, according with the selection made for each estimation method, M.

	M			
	CPCA	SymCovPCA	VPCA	CIPCA
$\hat{Cor}(C[SPC_1^M], C[SPC_1^M]^{tr})$	0.997	0.998	0.997	0.986
$\hat{Cor}(R[SPC_1^M], R[SPC_1^M]^{tr})$	0.989	0.997	0.996	0.984
$\hat{Cor}(C[SPC_2^M], C[SPC_2^M]^{tr})$	0.995	0.984	0.988	0.915
$\hat{Cor}(R[SPC_2^M], R[SPC_2^M]^{tr})$	0.856	0.774	0.988	0.960

5. Conclusion

Starting from a generic definition of symbolic variance and covariance for random interval-valued variables, we have used a common insightful framework to present four symbolic principal component estimation methods that rely on a symbolic-conventional-symbolic strategy: CPCA, VPCA, CIPCA, and SymCovPCA. This framework highlighted similarities and differences between these methods.

Aiming at improving the interpretation of symbolic principal components, we proposed the use of truncated versions of symbolic principal components, which are obtained from the (complete) symbolic principal components by relying only on a strict subset of the original symbolic variables.

The analysis of a symbolic dataset containing Internet traffic lead to a clear interpretation of the underlying Internet applications (Web browsing, file sharing, streaming, video, port scans, and snapshots). The analysis pointed out the difficulties in separating illicit traffic from regular one, suggesting the need to develop outlier detection methods for symbolic data. Furthermore, the analysis highlighted similarities between the symbolic principal component estimation methods considered in the paper.

Acknowledgements

This work has been supported by Fundação para a Ciência e Tecnologia (FCT), Portugal, through the projects UID/Multi/04621/2013 and PTDC/EEI-TEL/5708/2014.

References

Bertrand P, Goupil F (2000). "Descriptive Statistics for Symbolic Data." In HH Bock, E Diday (eds.), *Analysis of Symbolic Data*, Studies in Classification, Data Analysis, and Knowledge Organization, pp. 106–124. Springer Berlin Heidelberg.

Billard L (2008). "Sample Covariance Functions for Complex Quantitative Data." In *Proceedings of World IASC Conference, Yokohama, Japan,*, pp. 157–163.

Billard L, Diday E (2003). "From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis." *Journal of the American Statistical Association*, **98**, 470–487.

Billard L, Diday E (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. John Wiley & Sons.

Cadima JFCL, Jolliffe IT (2001). "Variable Selection and the Interpretation of Principal Subspaces." *Journal of Agricultural, Biological, and Environmental Statistics*, **6**(1), 62–79.

Cazes P, Chouakria A, Diday E, Schektman Y (1997). "Extension de l'Analyse en Composantes Principales à des Données de Type Intervalle." *Revue de Statistique Appliquée*, **45**(3), 5–24.

Chouakria A (1998). *Extension des Méthodes d'Analyse Factorielle à des Données de Type Intervalle*. Ph.D. thesis, Université Paris-Dauphine.

De Carvalho FdA, Brito P, Bock HH (2006). "Dynamic Clustering for Interval Data Based on L_2 Distance." *Computational Statistics*, **21**(2), 231–250.

Diday E (1987). "The Symbolic Approach in Clustering and Related Methods of Data Analysis." In *Proceedings of First conference IFCS,Aachen, Germany*. H. Bock ed.North-Holland.

Le-Rademacher J, Billard L (2012). "Symbolic Covariance Principal Component Analysis and Visualization for Interval-Valued Data." *Computational and Graphical Statistics*, **21**(2), 413–432.

Pascoal C (2014). *Contributions to Variable Selection and Robust Anomaly Detection in Telecommunications*. Ph.D. thesis, Instituto Superior Técnico, Universidade de Lisboa, Portugal.

Pascoal C, Oliveira M, Valadas R, Filzmoser P, Salvador P, Pacheco A (2012). "Robust Feature Selection and Robust PCA for Internet Traffic Anomaly Detection." In *INFOCOM, 2012 Proceedings IEEE*, pp. 1755–1763. ISSN 0743-166X.

Vilela M (2015). *Classical and Robust Symbolic Principal Component Analysis for Interval Data*. Master's thesis, Instituto Superior Técnico, Universidade de Lisboa, Portugal.

Wang H, Guan R, Wu J (2012). "CIPCA: Complete-Information-based Principal Component Analysis for Interval-valued Data." *Neurocomputing*, **86**, 158–169.

Affiliation:

M. Rosário Oliveira
CEMAT and Dep. Matemática
Instituto Superior Técnico
Universidade de Lisboa, Portugal
E-mail: rosario.oliveira@tecnico.ulisboa.pt

Austrian Journal of Statistics
April 2017, Volume 46, 89–98.
http://www.ajs.or.at/
doi:10.17713/ajs.v46i3-4.674

Random Graphs' Robustness in Random Environment

Marina Leri
Institute of Applied Mathematical Research, Karelian Research Centre, Russian Academy of Sciences

Yury Pavlov
Institute of Applied Mathematical Research, Karelian Research Centre, Russian Academy of Sciences

Abstract

We consider configuration graphs the vertex degrees of which are independent and follow the power-law distribution. Random graphs dynamics takes place in a random environment with the parameter of vertex degree distribution following uniform distributions on finite fixed intervals. As the number of vertices tends to infinity the limit distributions of the maximum vertex degree and the number of vertices with a given degree were obtained. By computer simulations we study the robustness of those graphs from the viewpoints of link saving and node survival in the two cases of the destruction process: the "targeted attack" and the "random breakdown". We obtained and compared the results under the conditions that the vertex degree distribution was averaged with respect to the distribution of the power-law parameter or that the values of the parameter were drawn from the uniform distribution separately for each vertex.

Keywords: random graphs, random environment, power-law distribution, robustness, simulation modeling, forest fire model.

1. Introduction

Recently, random graphs have been widely used for modeling complex networks such as the Internet, social, transport or telecommunication networks (see, e.g., Durrett 2007). One of the random graph models showing its best fit is a so called configuration graph (Bollobas 1980). Real data observations showed (Faloutsos, Faloutsos, and Faloutsos 1999; Durrett 2007) that their topology can be adequately represented by configuration graphs with vertex degrees being independent identically distributed (i.i.d.) random variables possessing natural values. Let ξ be a random variable equal to the degree of any vertex. In (Reittu and Norros 2004) the authors showed that in modeling of huge networks it is more appropriate to use the following vertex degree distribution of the random variable ξ:

$$\mathbf{P}\{\xi = k\} = k^{-\tau} - (k+1)^{-\tau}, \tag{1}$$

where $k = 1, 2, \ldots$, $\tau > 0$. This distribution is known as a power-law vertex degree distribution, and, moreover, many authors note that in real networks the value of the parameter

τ usually lies in the interval $(1, 2)$ (see, e.g., Faloutsos *et al.* 1999; Reittu and Norros 2004; Durrett 2007). However other values of the parameter τ are not without interest. Thus, for example, it turned out that configuration graphs with vertex degree distribution (1) could be used for modeling forest fires, where $\tau > 2$ is even more important (see, Leri and Pavlov 2014, 2016).

Each vertex degree equals the number of stubs or semiedges coming from it, i.e. the number of edges for which connected vertices are not yet found. All stubs are numbered in an arbitrary order. The sum of vertex degrees has to be even, so if it turns out to be odd, one extra stub is added to an equiprobably chosen vertex. The graph is constructed by joining all stubs one to another pairwise and equiprobably to form edges. Obviously, such graph may contain loops, cycles and multiple edges.

In (Leri and Pavlov 2014, 2016) we studied the robustness of configuration graphs to targeted and random destruction influences. In the case of a random breakdown equiprobably chosen vertices are removed from the graph sequentially with all the incident edges. In the case of a targeted destruction vertices with the highest degrees are removed. We proposed a criterion of graph destruction and found estimates of graph breakdown probability depending on the fraction of removed vertices.

Recently, there appeared some works where the authors note that with growing network size vertex degree distributions of corresponding random graphs may change and even become random variables (see, e.g., Bianconi and Barabasi 2001). Therefore, it was natural to start studying random graphs in random environment, where the vertex degree distribution is not fixed, as in Equation (1), but has a random behaviour.

In (Pavlov 2016) we considered configuration graphs where vertex degrees follow the distribution (1) under the condition that the parameter τ is a random variable uniformly distributed on the interval $[a, b]$, where $0 < a < b < \infty$. Then, the random variable ξ has the following distribution:

$$\begin{gathered} p_1 = \mathbf{P}\{\xi = 1\} = 1 - \frac{1}{(b-a)\ln 2}\left(\frac{1}{2^a} - \frac{1}{2^b}\right), \\ p_k = \mathbf{P}\{\xi = k\} = \frac{1}{(b-a)\ln k}\left(\frac{1}{k^a} - \frac{1}{k^b}\right) - \frac{1}{(b-a)\ln(k+1)}\left(\frac{1}{(k+1)^a} - \frac{1}{(k+1)^b}\right), \end{gathered} \tag{2}$$

where $k = 2, 3, \ldots$. For such graphs we studied the limit distribution of degree structure for all zones where the number of vertices and the number of edges tends to infinity. New results on the asymptotical behaviour of the vertex degrees are given below.

In this paper we consider the robustness of configuration graphs with random vertex degree distribution. Distribution (2), as well as (1), is the same for all vertices, thus it is not yet a random environment. It is more natural to suppose that vertex degrees are defined by Equation (1), where the value of τ is chosen from the interval $[a, b]$ equiprobably for each vertex (see, Leri 2016). This work includes the results on robustness of both models. Comparison of the results showed their similarity. It means that the study of the graphs' behaviour in the considered random environment can be replaced by the study of the model with an averaged vertex degree distribution (2).

In the next section limit theorems of the maximum vertex degree and the number of vertices of the given degree are proved. Section 3 contains the description of the process of modeling graphs with vertex degree distribution (2). In section 4 we consider graphs in random environment. In section 5 we discuss the results of comparing the negative outcomes of destruction on these two models.

2. Degree structure

Let N be the number of vertices. We denote by $\xi_{(N)}$ and μ_r the random variables equal to the maximum vertex degree and the number of vertices with degree r, respectively. In (Reittu

and Norros 2004) the authors showed that if vertex degrees follow the distribution (1), $\xi_{(N)}$ is proportional to $N^{1/\tau}$ as $N \to \infty$, and μ_r for large r is proportional to $Nr^{-(\tau+1)}$. Let us consider the limit behaviour of these characteristics in the case of vertex degree distribution (2).

Theorem 1. Let $N \to \infty$. Then for any fixed x

$$\mathbf{P}\{a \ln \xi_{(N)} - \ln N + \ln \ln N + \ln(b-a) - \ln a \le x\} \to e^{-e^{-x}}.$$

Proof. Let us denote by $\xi_1, \ldots, \xi_N$ the degrees of vertices $1, \ldots, N$. It is clear that

$$\mathbf{P}\{\xi_{(N)} \le y\} = \mathbf{P}\{\xi_1 \le y, \ldots, \xi_N \le y\} = \mathbf{P}^N\{\xi_1 \le y\}.$$

From this and (2) we obtain that

$$\mathbf{P}\{\xi_{(N)} \le y\} = \left(1 - \frac{1}{(b-a)\ln([y]+1)}\left(\frac{1}{([y]+1)^a} - \frac{1}{([y]+1)^b}\right)\right)^N, \tag{3}$$

where $[y]$ is the integer part of y. It is easy to see that as $y \to \infty$

$$\frac{1}{(b-a)\ln([y]+1)}\left(\frac{1}{([y]+1)^a} - \frac{1}{([y]+1)^b}\right) \sim \frac{1}{(b-a)y^a \ln y}. \tag{4}$$

The theorem assertion follows from (3) and (4) if

$$y = \left(\frac{a}{(b-a)} e^x \frac{N}{\ln N}\right)^{1/a}. \tag{5}$$

Remark 1. From Theorem 1 and (5) we can see that the maximal vertex degree is proportional to $(N/\ln N)^{1/a}$.

Theorem 2. Let $N \to \infty$ and k is a natural number. The following assertions are true.

1. If $Np_r \to \infty$ then uniformly in k such that $u_r = (k - Np_r)/\sqrt{Np_r(1-p_r)}$ lies in any fixed finite interval

$$\mathbf{P}\{\mu_r = k\} = (2\pi N p_r(1-p_r))^{-1/2} e^{-u_r^2/2}(1 + o(1)).$$

2. If $r \to \infty$ then uniformly in k such that $(k - Np_r)/\sqrt{Np_r}$ lies in any fixed finite interval

$$\mathbf{P}\{\mu_r = k\} = \frac{(Np_r)^k}{k!} e^{-Np_r}(1 + o(1)).$$

Proof. Random variables $\xi_1, \ldots, \xi_N$ are independent, therefore

$$\mathbf{P}\{\mu_r = k\} = \binom{N}{k} p_r^k (1-p_r)^{N-k}. \tag{6}$$

Under the first condition of the theorem $Np_r(1-p_r) \to \infty$ and in (6) we can use the normal approximation of binomial probabilities. Under the second condition $p_r \to 0$ and these probabilities allow Poisson approximation. Straight from here follow assertions of Theorem 2.

Remark 2. Let $r \to \infty$. From (2) we can find that

$$p_r \sim \frac{a}{(b-a)r^{a+1}\ln r}.$$

From here and Theorem 2 it follows that μ_r is proportional to $N/(r^{a+1}\ln r)$.

Remark 3. Let $p_r \to 0$. By Theorem 2 normal and Poisson approximations of μ_r distribution are possible at the same time.

3. Robustness of graphs with given degree distribution

The problem of robustness and vulnerability of present-day huge complex networks to various types of breakdowns remains rather pressing (see, e.g., Bollobas and Riordan 2004; Durrett 2007; Norros and Reittu 2008). Therefore, along with the studies of random graph's structure, we consider the process of its destruction aiming to look at how the main structural characteristics would change with the removal of graph vertices. In this Section we study configuration random graphs with vertex degrees following the distribution (2) on a predefined interval $[a, b]$. We consider two types of the destruction process. During the first one (link saving) we remove graph vertices sequentially with all the incident edges. The aim is to consider changes of the graph structure with vertex removal. The second process (node survival = forest fire model) takes the similarity with the process of fire spreading (it could also be any other destructive influence (Arinaminparthy, Kapadia, and May 2012; Bertoin 2011, 2012; Drossel and Schwabl 1992)), where the destruction starts from a chosen vertex and then spreads over the graph through its edges in some certain way. For each type of destruction process we consider two cases: "targeted attack" means the removal of vertices with the highest degrees (link saving) or "targeted lightning-up" of a vertex with the highest degree (node survival) on the one hand, and "random breakdown" – the removal of equiprobably chosen vertices (link saving) or "random ignition" of an equiprobably chosen vertex (node survival) on the other. The main method of studies described below is simulation modeling followed by statistical analysis of the obtained data.

3.1. Link saving: preserving graph connectivity

The first considered graph destruction process – link saving – was done on simulation models of graphs sized from 1000 to 10000 vertices and three intervals $[a, b]$: $(1, 2)$, $(1, 3]$ and $[2, 3]$. Power-law configuration graphs with the parameter $\tau \in (1, 2)$ are known to be a good representation of the AS-level topology, where AS means autonomous systems (see, e.g., Faloutsos *et al.* 1999; Mahadevan, Krioukov, Fomenkov, Huffaker, Dimitropoulos, Claffy, and Vahdat 2006; Reittu and Norros 2004). Power-law graphs with the parameter $\tau \in [2, 3]$ are useful for the studies of forest fire models (Leri and Pavlov 2014, 2016; Leri 2016). The interval $(1, 3]$ is chosen as a generalization. The purpose is to estimate the graph breakdown probability depending on the percentage of removed vertices.

Let random variables $\eta_1, \eta_2, \ldots, \eta_s$ be equal to the percentages of vertices in the graph components in decreasing order. Thus η_1 is the percentage of vertices in the largest component, η_2 – the percentage of vertices in the second-sized component, etc., where s is the number of graph components. In (Leri and Pavlov 2014) we proposed the criterion of graph destruction to be the occurrence of the following event $\mathrm{A} : \{\eta_1 \leq 2\eta_2\}$, which means that the percentage of vertices in the largest component becomes less than or equal to the two values of the percentage of vertices in the second largest component. This criterion is suggested because the largest components of the considered graphs are far larger than their second components. The average initial ratios of the sizes of the first two largest components of the considered graphs depend on the graph size N as follows: $\eta_1/\eta_2 = 0.15N + 81$ for $(1, 2)$; $\eta_1/\eta_2 = 0.002N + 7.7$ for $[2, 3]$; $\eta_1/\eta_2 = 0.08N + 50$ for $(1, 3]$, where determination coefficients R^2 of all these regression models are equal to 0.99. This means that, for example, when $[a, b] = (1, 2)$ the ratio η_1/η_2 is more than 200 for $N = 1000$ and more than 1600 for $N = 10000$.

In Figures 1 and 2 we plot the following regression relations between the probabilities $\mathbf{P}\{\mathrm{A}\}$ of graph destruction, the percentage of vertices removed from the graph r and the initial graph size N. For the process of targeted attack on vertices with the highest degrees we obtained

the following regression models:

$$\begin{aligned} \mathbf{P}\{A\} &= -0.47 + 0.054r^{1.9} + 0.046\ln N, \qquad (1,2), \qquad (R^2 = 0.96) \\ \mathbf{P}\{A\} &= 0.26 + 0.31\ln r + 0.1\ln N, \qquad [2,3], \qquad (R^2 = 0.99) \\ \mathbf{P}\{A\} &= -0.63 + 0.27r^{1.3} + 0.06\ln N, \qquad (1,3], \qquad (R^2 = 0.95) \end{aligned}$$

where the value of r is limited as follows: $4.21 - 0.354\ln N \le r \le 5.75 - 0.108\ln N$ for $(1,2)$, $0.115 - 0.01\ln N \le r \le 2.89 - 0.256\ln N$ for $[2,3]$ and $2.24 - 0.2\ln N \le r \le 4.03 - 0.122\ln N$ for $(1,3]$. In practice, it is clear that for the out-of-limits values of r the probability $P\{A\} = 0$ when r is less than the lower limit and $P\{A\} = 1$ when r is larger than the upper limit. The same is true for the models given below. In the case of random breakdowns the relations are as follows:

$$\begin{aligned} \mathbf{P}\{A\} &= -0.4 + 0.00035r^2 - 0.04\ln N, \qquad (1,2), \qquad (R^2 = 0.92) \\ \mathbf{P}\{A\} &= 0.51 + 0.02r - 0.07\ln N, \qquad [2,3], \qquad (R^2 = 0.99) \\ \mathbf{P}\{A\} &= -0.45 + 0.00041r^2 - 0.04\ln N, \qquad (1,3], \qquad (R^2 = 0.92) \end{aligned}$$

with the following limits for r: $35.32 + 1.25\ln N \le r \le 63.59 + 0.81\ln N$ for $(1,2)$, $0.056 + 0.00074N \le r \le 24.5 + 3.5\ln N$ for $[2,3]$ and $34.36 + 1.12\ln N \le r \le 59.77 + 0.74\ln N$ for $(1,3]$.

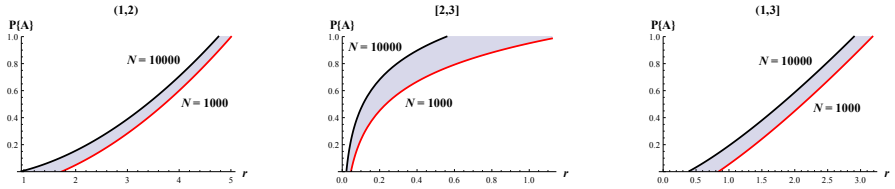


Figure 1: Probability of graph destruction in the case of targeted attacks on graphs with $[a,b] = (1,2)$, $[2,3]$ and $(1,3]$, respectively.

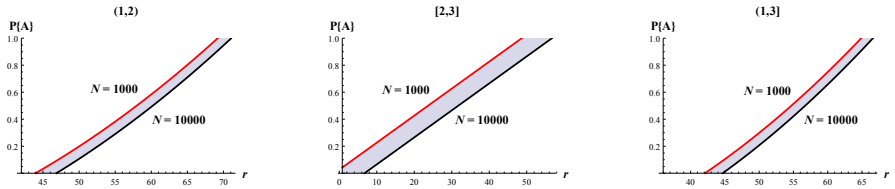


Figure 2: Probability of graph destruction in the case of random breakdowns of graphs with $[a,b] = (1,2)$, $[2,3]$ and $(1,3]$, respectively.

The dependencies for $1000 < N < 10000$ fall within shaded regions. Simulation results show that configuration graphs with vertex degree distribution (2) are more robust to random breakdowns than to targeted attacks on the vertices with the highest degrees. To destroy such a graph by removing vertices with high degrees it is enough to take away $1 - 5\%$ of them. However, in the case of random vertex removal, the graph will be ruined by the destruction of $50 - 70\%$ of its vertices. Thus, in the case when $[a,b] = [2,3]$ graphs are more vulnerable to both targeted and random breakdowns than in the cases when $[a,b] = (1,2)$ and $[a,b] = (1,3]$.

3.2. Node survival: forest fire model

The study of a destruction process which is called a forest fire model imposes some constraints on the graphs being considered. Since we view graph vertices as trees growing in a limited area of a real forest, their number as well as the number of vertices in a corresponding graph has to

be limited. Thus, we propose to use an auxiliary square lattice graph of the size 100×100 (Leri 2016; Leri and Pavlov 2016). Two vertices of the graph are connected if on the corresponding tree topology a fire can move from one tree to another. We consider 15 different relative allocations of edges and vertices on our lattice graphs. Let m be an averaged inner vertex degree. For a fully packed square lattice $m = 8$. Then for each lattice graph topology we calculated corresponding values of m and N. Let the size of our power-law configuration graph be equal to the size of the auxiliary lattice graph. Knowing that $m = \zeta(\tau)$ (where $\zeta(x)$ is the Riemann zeta function), we obtained a regression relation between the power-law configuration graph size $N \leq 10000$ and the parameter τ of vertex degree distribution (1) (see Figure 3):

$$N = [9256\tau^{-1.05}], \qquad R^2 = 0.97. \tag{7}$$

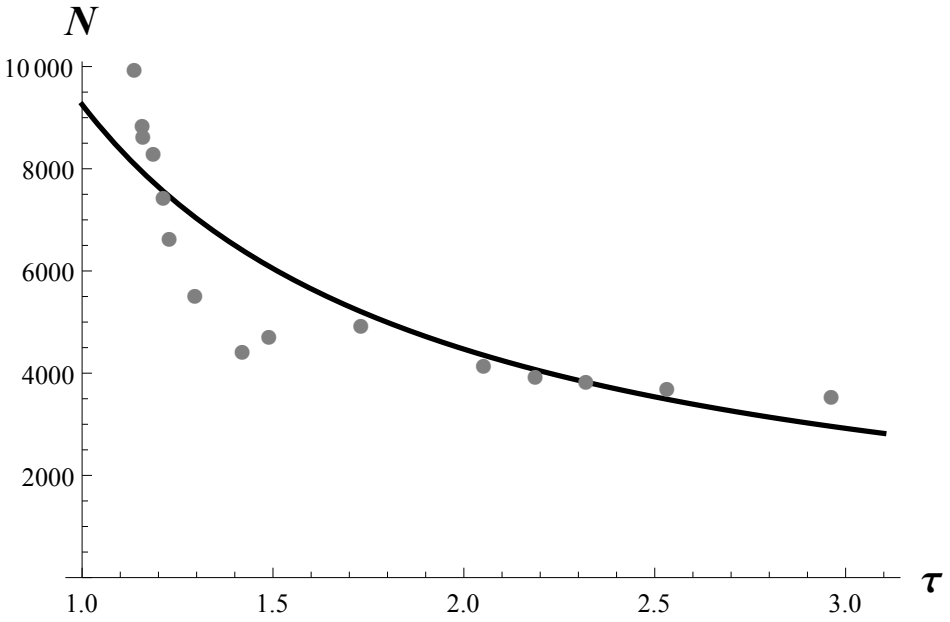


Figure 3: Regression relation between N and τ.

Let $\overline{\tau} = (a+b)/2$, then the relation (7) confines the number of vertices N in a corresponding power-law graph. Here, as in Subsection 3.1, we consider the same three intervals $[a, b]$: $(1, 2)$, $[2, 3]$, $(1, 3]$ with the N values obtained from (7) being equal to 6046, 3536 and 4470, respectively. As it was mentioned above, we analyse two cases of starting a fire propagation process: targeted lightning-up and random ignition. When the fire starts it spreads through the incident edges to connected vertices with the probability of fire transition p. This probability could be either a predefined value $p \in (0, 1]$ fixed for all the graph edges or a random variable following the standard uniform distribution. Here we study both cases. The purpose is to find the optimal interval $[a, b]$ of the distribution (2) that would ensure maximum survival of graph vertices in case of a fire. For all the three intervals we found relations between the average number of vertices surviving in a fire n and the probability of fire transition p. Plots on Figure 4 show how the number of remaining vertices n depends on the probability p in the two fire-start cases: targeted lightning-up (left plot) and random ignition (right plot).

It is clear that with the increase of the probability p the number of remaining vertices n will be decreasing. Furthermore, at lower values of p graphs with $[a, b] = (1, 2)$ prove to be more resilient to the fire destruction process in both cases of fire start. But as the value of p increases the topology with $[a, b] = [2, 3]$ will ensure a better survival of graphs vertices. As for graphs with $[a, b] = (1, 3]$, they are the most vulnerable to this kind of destruction in both fire-start cases.

Further we consider the probability of fire transition p to be a random variable drawn from the standard uniform distribution. Table 1 shows an average number of vertices having remained

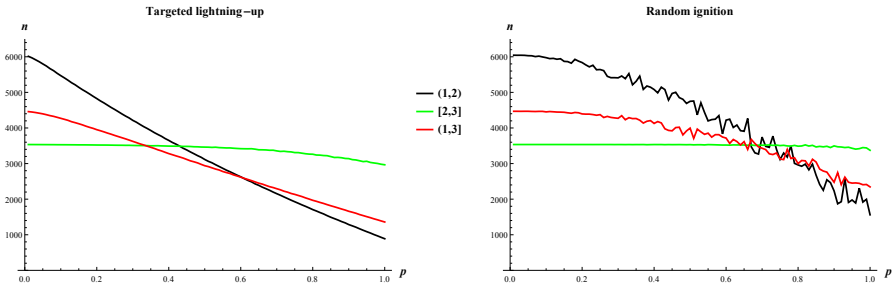


Figure 4: Relation between the number of surviving vertices n and the probability of fire transition.

in the fire for the three considered intervals and the two fire-start cases.

Table 1: Average number of nodes surviving in a fire $\overline{n}$.

$[a,b]$	targeted lightning-up	random ignition
$(1,2)$	3122	4605
$[2,3]$	3471	3531
$(1,3]$	2950	3959

Thus, when the fire starts from a vertex with the highest degree more vertices will survive on the graph topology with the vertex degree distribution (2) with $[a,b]=[2,3]$. But in the case of random ignition graphs with $[a,b]=(1,2)$ will be more resilient to the fire.

4. Graphs' robustness in random environment

In this Section, as distinct from Section 3, vertex degrees of the considered configuration graphs follow the distribution (1), where $\tau \sim \mathbf{U}[a,b]$, which means that the parameter τ is uniformly distributed on a predefined interval $[a,b]$ and the values of τ are chosen separately for each vertex. Here we follow the same study scheme as in Section 3, considering the two types of the destruction process: link saving and node survival with the two cases of destruction spreading: targeted and random.

4.1. Link saving: preserving graph connectivity

As before, we consider graphs of the sizes $N \in [1000, 10000]$ and the three intervals $[a,b]$: $(1,2)$, $(1,3]$ and $[2,3]$. Having obtained statistical data from computer simulations we derived regression relations between the probabilities $\mathbf{P}\{\mathrm{A}\}$ of graph destruction, the percentage of vertices removed from the graph r and the initial graph size N. The following regression models were obtained in the case of a targeted attack (see Figure 5):

$$\begin{aligned}
\mathbf{P}\{A\} &= -0.44 + 0.04r^{2.1} + 0.05\ln N, && (1,2), && (R^2 = 0.96)\\
\mathbf{P}\{A\} &= 0.72 + 0.3\ln r + 0.04\ln N, && [2,3], && (R^2 = 0.99)\\
\mathbf{P}\{A\} &= -0.65 + 0.3r^{1.2} + 0.06\ln N, && (1,3], && (R^2 = 0.95)
\end{aligned}$$

and in the case of random breakdowns (see Figure 6):

$$\begin{aligned}
\mathbf{P}\{A\} &= -0.53 + 0.00033r^2 - 0.016\ln N, && (1,2), && (R^2 = 0.93)\\
\mathbf{P}\{A\} &= 0.42 + 0.019r - 0.056\ln N, && [2,3], && (R^2 = 0.99)\\
\mathbf{P}\{A\} &= -0.17 + 0.0004r^2 - 0.064\ln N, && (1,3], && (R^2 = 0.93)
\end{aligned}$$

with the following limits for r in the case of a targeted attack: $6.91 - 0.745\ln N \le r \le 5.59 - 0.108\ln N$ for $(1,2)$, $0.064 - 0.00408\ln N \le r \le 1.79 - 0.114\ln N$ for $[2,3]$ and $2.11 -$

$0.186 \ln N \leq r \leq 4.17 - 0.133 \ln N$ for $(1, 3]$. and in the case of random breakdowns: $40.32 + 0.54 \ln N \leq r \leq 68.15 + 0.34 \ln N$ for $(1, 2)$, $-0.28 + 0.0006N \leq r \leq 30.5 + 3 \ln N$ for $[2, 3]$ and $25.88 + 1.92 \ln N \leq r \leq 55 + 1.23 \ln N$ for $(1, 3]$. It is easy to see that the obtained models are similar in function forms and only slightly differ in regression models' coefficients.

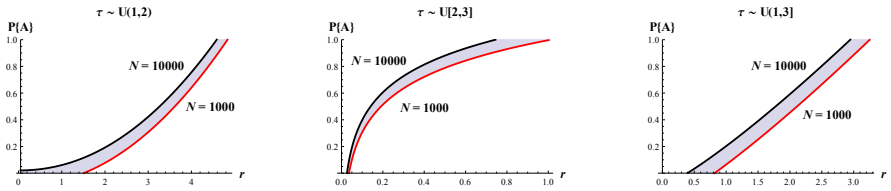


Figure 5: Probability of graph destruction in the case of targeted attacks on graphs with $\tau \sim \mathbf{U}(1, 2)$, $\tau \sim \mathbf{U}[2, 3]$ and $\tau \sim \mathbf{U}(1, 3]$, respectively.

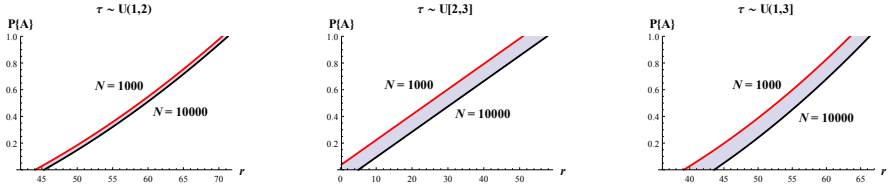


Figure 6: Probability of graph destruction in the case of random breakdowns of graphs with $\tau \sim \mathbf{U}(1, 2)$, $\tau \sim \mathbf{U}[2, 3]$ and $\tau \sim \mathbf{U}(1, 3]$, respectively.

Again, similarly to the graphs with the distribution (2), average initial ratios between the sizes of the first two largest components depend on the graph size N in the following way: $\eta_1/\eta_2 = 0.14N + 100$ for $(1, 2)$; $\eta_1/\eta_2 = 0.002N + 7.1$ for $[2, 3]$; $\eta_1/\eta_2 = 0.08N + 68$ for $(1, 3]$, where $R^2 = 0.99$ for all the three models.

Thus, the considered graphs also proved to be more robust to random breakdowns than to targeted attacks on vertices with the highest degrees. In the case of targeted attack it is enough to remove $1-5\%$ of vertices to destroy a graph, and in the case of random breakdowns it takes $50 - 70\%$. Here, like in Section 3, the same notes concerning the smallest values of the removed vertices are true (see Figures 5 and 6). Similarly to the results described in Section 3, in the case when $\tau \sim \mathbf{U}[2, 3]$ graphs are more vulnerable to both targeted and random breakdowns than in the cases where $\tau \sim \mathbf{U}(1, 2)$ and $\tau \sim \mathbf{U}(1, 3]$.

4.2. Node survival: forest fire model

Here we discuss the results of studying forest fire modeling on configuration graphs with the vertex degree distribution (1) and $\tau \sim \mathbf{U}[a, b]$. As in Section 3, we use an auxiliary square lattice graph of the size 100×100 to confine the number of vertices N through the relation (7) with $\overline{\tau} = (a+b)/2$. We consider the same three intervals $[a, b]$: $(1, 2)$, $[2, 3]$, $(1, 3]$ on which the parameter τ is uniformly distributed. The values of N obtained from (7) were the same as in Section 3.2. Again we consider the two fire-start cases: targeted lightning-up and random ignition. The probability of fire transition p is either a predefined value $p \in (0, 1]$ fixed for all graph edges or a random variable following the standard uniform distribution. The aim is to find the optimal interval of the parameter τ that would ensure maximum survival of graph vertices in case of a fire. Plots in Figure 7 show relations between the average number of vertices surviving in a fire n and the probability of fire transition p for the two fire-start cases.

It is quite clear that the obtained results are rather similar to those in Section 3. So are the results for the case where the probability of fire transition p is a random variable uniformly distributed on $(0, 1]$ (see Table 2).

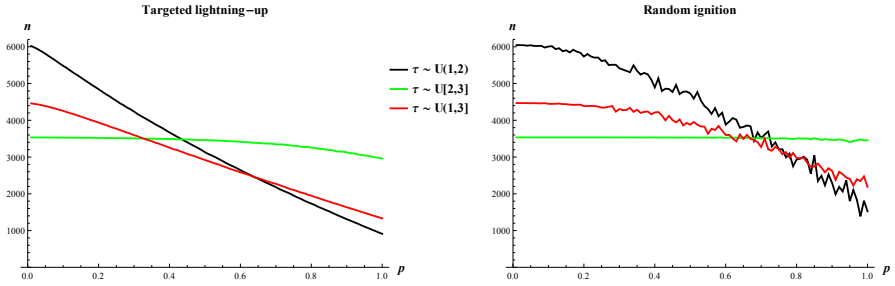


Figure 7: Relation between the number of surviving vertices n and the probability of fire transition.

Table 2: An average number of nodes surviving in a fire $\overline{n}$.

$\tau \sim \mathbf{U}[a, b]$	targeted lightning-up	random ignition
$(1, 2)$	3146	4933
$[2, 3]$	3464	3533
$(1, 3]$	2917	3927

5. Conclusions

The simulation model of the graph destruction process in random environment used in Section 4 is more complicated and requires far more computations than the model with fixed vertex degree distribution (2) in Section 3. At the same time, it is easy to see that the results obtained in Sections 3 and 4 are quite close to each other. This means that we can study the dynamics of random graphs in random environment using averaged degree distributions. But the problem of finding the conditions when such an interchange is incorrect is still open.

6. Acknowledgements

The study is supported by the Russian Foundation for Basic Research, grant 16-01-00005. The authors would like to thank professor A.M. Zubkov (Steklov Mathematical Institute of RAS) for constructive discussion of the problem.

References

Arinaminparthy N, Kapadia S, May R (2012). "Size and Complexity in Model Financial Systems." *Proceedings of the National Academy of Sciences of the USA*, **109**, 18338–18343.

Bertoin J (2011). "Burning Cars in a Parking Lot." *Commun. Math. Phys.*, **306**, 261–290.

Bertoin J (2012). "Fires on Trees." *Annales de l'Institut Henri Poincare Probabilites et Statistiques*, **48**(4), 909–921.

Bianconi G, Barabasi AL (2001). "Bose-Einstein Condensation in Complex Networks." *Physical Review Letters*, **86**, 5632–5635.

Bollobas B (1980). "A Probabilistic Proof of an Asymptotic Formula for the Number of Labelled Regular Graphs." *Eur. J. Comb.*, **1**, 311–316.

Bollobas B, Riordan O (2004). "Robustness and Vulnerability of Scale-free Random Graphs." *Internet Mathematics*, **1**(1), 1–35.

Drossel B, Schwabl F (1992). "Self-organized Critical Forest-fire Model." *Phys. Rev. Lett.*, **69**, 1629–1632.

Durrett R (2007). *Random Graph Dynamics*. Cambridge Univ. Press, Cambridge.

Faloutsos C, Faloutsos P, Faloutsos M (1999). "On Power-law Relationships of the Internet Topology." *Computer Communications Rev.*, **29**, 251–262.

Leri M (2016). "Forest Fire Model on Configuration Graphs with Random Node Degree Distribution." In *XVII-th International Summer Conference on Probability and Statistics: Conference Proceedings and Abstracts*, pp. 29–32.

Leri M, Pavlov Y (2014). "Power-law Random Graphs' Robustness: Link Saving and Forest Fire Model." *Austrian Journal of Statistics*, **43**(4), 229–236.

Leri M, Pavlov Y (2016). "Forest Fire Models on Configuration Random Graphs." *Fundamenta Informaticae*, **145**(3), 313–322.

Mahadevan P, Krioukov D, Fomenkov M, Huffaker B, Dimitropoulos X, Claffy K, Vahdat A (2006). "The Internet AS-Level Topology: Three Data Sources and One Definitive Metric." *ACM SIGCOMM Computer Communication Review (CCR)*, **36**(1), 17–26.

Norros I, Reittu H (2008). "Attack Resistance of Power-law Random Graphs in the Finite Mean, Infinite Variance Region." *Internet Mathematics*, **5**(3), 251–266.

Pavlov Y (2016). "On Conditional Configuration Graphs with Random Distribution of Vertex Degrees." *Transactions of Karelian Research Centre of Russian Academy of Science: Mathematical Modeling and Information Technologies*, **8**, 62–72. In Russian.

Reittu H, Norros I (2004). "On the Power-law Random Graph Model of Massive Data Networks." *Performance Evaluation*, **55**, 3–23.

Affiliation:

Yury Pavlov
Institute of Applied Mathematical Research
Karelian Research Centre
Russian Academy of Sciences
Pushkinskaja,11, Petrozavodsk 185910, Russia
E-mail: pavlov@krc.karelia.ru
URL: http://mathem.krc.karelia.ru/member.php?id=9&plang=e

Austrian Journal of Statistics
April 2017, Volume 46, 99–105.
http://www.ajs.or.at/
doi:10.17713/ajs.v46i3-4.675

A Modification of Linfoot's Informational Correlation Coefficient

Georgy Shevlyakov
Peter the Great St.Petersburg Polytechnic

Nikita Vasilevskiy
Itivity Group AB

Abstract

Performance of the Linfoot's informational correlation coefficient is experimentally studied at the bivariate normal distribution. It is satisfactory in the case of a strong correlation and on large samples. To reduce the bias of estimation, a symmetric version of this correlation measure is proposed. On small and large samples, this modified informational correlation coefficient outperforms Linfoot's correlation measure at the bivariate normal distribution in the wide range of the correlation coefficient.

Keywords: informational measures, correlation, bivariate normal distribution.

1. Introduction

Pearson's correlation coefficient is a well-defined measure of the linear dependence between continuous random variables X and Y, as well as the closely related to it rank measures: namely, the quadrant, Spearman and Kendall correlation coefficients. However, if one is interested either in processing discrete data or in revealing the possible nonlinear relationship between random variables, then difficulties may arise both in the implementation of those classical measures as well as in their interpretation.

In the literature, several proposals have been made to solve these problems, for instance, Gebelein's and Sarmanov's correlation coefficients (Gebelein 1941), (Sarmanov 1958). Recently, the distance correlation coefficient was proposed (see (Székely, Rizzo, and Bakirov 2007)).

In what follows, we focus on the informational measures of association between random variables (Shannon 1948). The dependence measure by Joe (1989) exploits the concept of the relative entropy that measures the similarity of two random variables with the distributions $p(x)$ and $q(x)$ in the discrete case

$$D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}\,.$$

Silvey (1964) uses the measure of dependence between two random variables defined by the ratio of their joint density and the product of their marginal densities

$$\varphi(x, y) = p(x, y)/[p(x)p(y)].$$

The introduced measure is defined as $\Delta = E[d(x)]$, where

$$d(x) = \int_{y:\,\varphi(x,y)>1} [p(y|x) - p(y)]\, dy.$$

Thus, it can be rewritten as

$$\Delta = \int\int_{(x,y):\,\varphi(x,y)>1} [p(x,y) - p(x)p(y)]\, dxdy.$$

Granger, Maasoumi, and Racine (2004) introduce another measure of dependence

$$S_p = \frac{1}{2}\int\int \left[p(x,y)^{1/2} - [p(x)p(y)]^{1/2}\right]^2 dxdy.$$

However, Joe's measure of dependence is not symmetric, whereas Silvey's and Granger's measures are hard to compute.

Mutual information ($I(X,Y)$) for any pair of discrete and continuous random variables X and Y is defined as follows

$$I(X,Y) = \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}\,, \quad I(X,Y) = \int\int f(x,y)\log\frac{f(x,y)}{f(x)f(y)}\,dxdy\,.$$

The informational correlation coefficient (ICC), firstly introduced by Linfoot (1957), is defined as follows

$$\rho_{ICC}(X,Y) = \sqrt{1 - e^{-2I(X,Y)}}\,. \tag{1}$$

Note that ICC is equal to the classical Pearson's correlation coefficient at the bivariate normal distribution: $\rho_{ICC}(X,Y) = \rho$.

2. Problem setting

Although the Linfoot's correlation coefficient ICC was introduced more than 60 years ago, its properties as a statistical measure of correlation have not yet been studied; it was not checked how well this measure estimates the correlation coefficient based on a sample of a given size.

Below, using Monte Carlo method on small and large samples, we experimentally examine the unbiasedness of ICC at the bivariate normal distribution over the wide range of its correlation coefficient.

Moreover, in order to improve the performance of ICC, namely to reduce its bias, we propose and study its modified symmetric version denoted as $MICC$.

3. Monte Carlo experiment

3.1. Description of the computational algorithm

All numerical experiments are performed using R language, especially its "entropy" library. The first problem is how to compute mutual information, which is used in (1). This is solved by applying a shrink-algorithm (Hausser and Strimmer 2010).

There exist several different algorithms of computing $I(X,Y)$; in our work, we choose the most precise one, not the fastest (for comparative analysis, see (Hausser and Strimmer 2010)). All experiments are performed at the standard bivariate normal distribution with density $f(x,y) = N(x,y;0,0,1,1,\rho)$.

The general algorithm can be described as follows:

1. Generate a sample of a fixed size: $N = 20, 60, 100, 400, 1000, 10000$.
2. Extract x- and y-components from the sample, which are dependent random variables with the correlation coefficient ρ.
3. Construct the table of frequencies, the discrete analog of the joint distribution: we take a rectangle $[x_{min}, x_{max}] \times [y_{min}, y_{max}]$ on plane and divide it into $n_x \times n_y$ "bins" of equal size. Thus, the table of dimension n_x, n_y is built, each element of which is equal to the number of points in the corresponding bin.
4. Mutual information $I(X, Y)$ and ICC are computed using this table of frequencies.

This sequence is repeated 1000 times, allowing us to compute Monte Carlo estimates of the mean and variance of ICC: computations are performed for $\rho = 0, 0.1, 0.2, \ldots, 0.9, 1$; the number of bins is taken equal to 400. Typical results are exhibited in Figure 1.

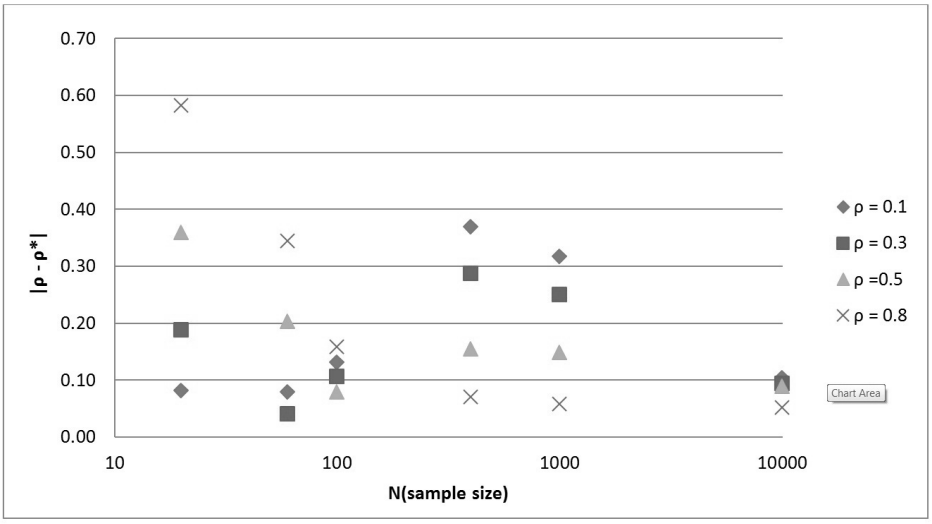


Figure 1: Monte Carlo biases of ICC

3.2. Monte Carlo results for ICC

From Figure 1 it follows that estimation biases are considerably large (on small samples, they can even be greater than 0.5). Relatively small biases are observed only on large samples $N = 1000$ and $N = 10\,000$.

Satisfactory performance is observed in the case of a strong correlation—the ICC biases decrease with the growth of the sample size.

We may also add that the coefficient of variance is less than 0.2 for all examined combinations of (ρ, N).

A remark on the choice of the number of bins. The shrink-algorithm takes the table of frequencies as an input. It appeared that the algorithm performance depends on the relation N/K^2, where K is the linear dimension of the table. We observed that results are almost independent of the changes of K, as they depend only on the coefficient $B = N/K^2$. For $\rho = 0.5$, the value $B = 7$ is optimal. Given a data sample, we can choose an appropriate value of K, which is a tuning parameter of our algorithm.

4. A symmetric version of Linfoot's correlation coefficient

4.1. Main result

Mutual entropy, also known as the Kullback-Leibler distance, has a serious disadvantage – it is not symmetric, i.e., $D_K L(p||q) \neq D_K L(q||p)$. Thus, the Kullback-Leibler divergence is

used (Kullback 1959)

$$Div(p||q) = D_{KL}(p||q) + D_{KL}(q||p). \tag{2}$$

Analogously, a symmetric version of mutual information can be introduced as it is natural to use a symmetric measure of correlation for estimation of the Pearson correlation coefficient, a symmetric measure of interdependence between random variables—in this case we expect lesser biases of estimation.

$$J(X,Y) = I(X,Y) + I^*(X,Y)$$
$$= \int\int f(x,y)\log\frac{f(x,y)}{f(x)f(y)}\,dxdy + \int\int f(x)f(y)\log\frac{f(x)f(y)}{f(x,y)}\,dxdy\,.$$

Our idea is to repeat Linfoot's derivation of Equation (1), replacing the mutual information $I(X,Y)$ with its symmetric version $J(X,Y)$. In this case, the following result holds.

Theorem *A symmetric analog of the Linfoot's informational correlation coefficient* (1) *called as the modified informational correlation coefficient* ($MICC$) *is given by*

$$\rho_{MICC} = \sqrt{1 - \frac{2}{W(2e^{2(J+1)})}} \tag{3}$$

with the particular case $\rho_{MICC} = \rho$ *at the bivariate normal distribution.*

Here W is the Lambert's function—the inverse function for xe^x; it cannot be expressed in terms of elementary functions. Its properties are well developed, and there are special methods to compute it (see (Corless, Gonnet, Hare, Jeffrey, and Knuth 1996)).

Proof The first step is to express the mutual information via the correlation coefficient similarly to that obtained by Linfoot (1957)

$$I(X,Y) = -\frac{1}{2}\log(1-\rho^2).$$

Compute $I^*(X,Y)$ at the bivariate normal with density $f(x,y) = N(x,y;0,0,\sigma_x^2,\sigma_y^2,\rho)$

$$I^*(X,Y) = \int\int f(x)f(y)\log\frac{f(x)f(y)}{f(x,y)}\,dxdy$$
$$= \frac{1}{2\pi\sigma_x\sigma_y}\int\int \exp\left(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}\right)\left(\frac{\rho^2}{1-\rho^2}\frac{x^2}{2\sigma_x^2} + \frac{\rho^2}{1-\rho^2}\frac{y^2}{2\sigma_y^2} - \frac{\rho xy}{\sigma_x\sigma_y}\right)dxdy.$$

Consider the following three integrals:

$$I_1(X,Y) = \frac{1}{2\pi\sigma_x\sigma_y}\int\int \exp\left(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}\right)\frac{\rho^2}{1-\rho^2}\frac{x^2}{2\sigma_x^2}\,dxdy$$

$$= \frac{1}{2\pi\sigma_x\sigma_y}\frac{\rho^2}{1-\rho^2}\int \frac{x^2}{2\sigma_x^2}e^{-\frac{x^2}{2\sigma_x^2}}\,dx\int e^{-\frac{y^2}{2\sigma_y^2}}\,dy$$
$$= \frac{1}{2\pi\sigma_x\sigma_y}\frac{\rho^2}{1-\rho^2}\frac{1}{2\sigma_x^2}\frac{\sqrt{\pi}}{2}2^{\frac{3}{2}}\sigma_x^3\sqrt{2\pi}\sigma_y$$
$$= \frac{\rho^2}{2(1-\rho^2)}.$$

Analogously, we have

$$I_2(X,Y) = \frac{1}{2\pi\sigma_x\sigma_y}\int\int \exp\left(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}\right)\frac{\rho^2}{1-\rho^2}\frac{y^2}{2\sigma_y^2}\,dxdy$$

$$= \frac{\rho^2}{2(1-\rho^2)}.$$

Next,

$$I_3(X,Y) = -\frac{\rho}{\sigma_x\sigma_y}\int x\exp\left(-\frac{x^2}{2\sigma_x^2}\right)dx\int y\exp\left(-\frac{y^2}{2\sigma_y^2}\right)dy = 0.$$

Thus, we get

$$I^*(X,Y) = I_1 + I_2 + I_3 = \frac{\rho^2}{1-\rho^2},$$

$$J(X,Y) = -\frac{\log(1-\rho^2)}{2} + \frac{\rho^2}{1-\rho^2}.$$

Now we express the correlation coefficient ρ via the symmetrized mutual information $J(X,Y)$ inversing the above Equation:

$$J(X,Y) = J = -\frac{\log(1-\rho^2)}{2} + \frac{\rho^2}{1-\rho^2}.$$

Set $1-\rho^2 = t$. Then we get

$$J = -\frac{\log t}{2} + \frac{1-t}{t} = -\frac{\log t}{2} + \frac{1}{t} - 1.$$

Further

$$-2(J+1) = \log t - \frac{2}{t} \Rightarrow \frac{\exp(-2(J+1))}{2} = \frac{t}{2}\exp(-2/t),$$

and setting $2/t = p$, we finalize the derivation

$$\frac{1}{2e^{2(J+1)}} = \frac{1}{p}e^{-p} \Rightarrow 2e^{2(J+1)} = pe^p \Rightarrow p = W(2e^{2(J+1)}),$$

where W is the Lambert function, namely the inverse function for pe^p. Finally, we arrive at Equation (3).

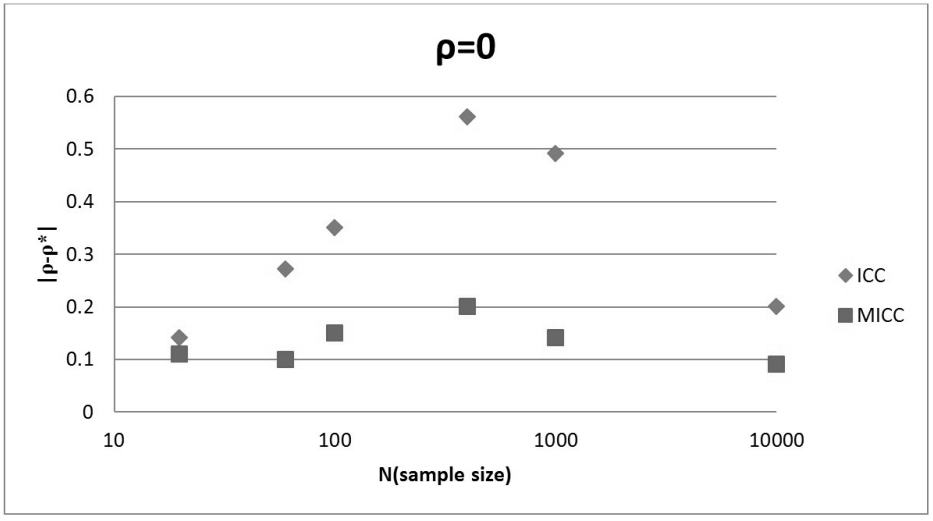


Figure 2: Monte Carlo biases of ICC and $MICC$ at $\rho = 0$

4.2. Monte Carlo results for $MICC$

The results of comparison of these two correlation measures are exhibited in Figures 2–4: from them it follows that the modified informational correlation coefficient $MICC$ outperforms its classical Linfoot's analog ICC in almost all of the combinations of sample sizes and correlation

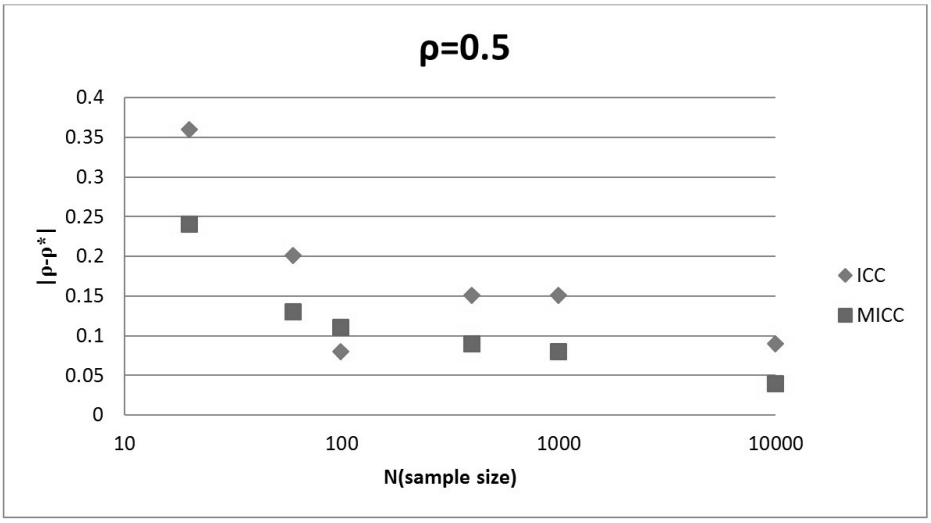


Figure 3: Monte Carlo biases of *ICC* and *MICC* at $\rho = 0.5$

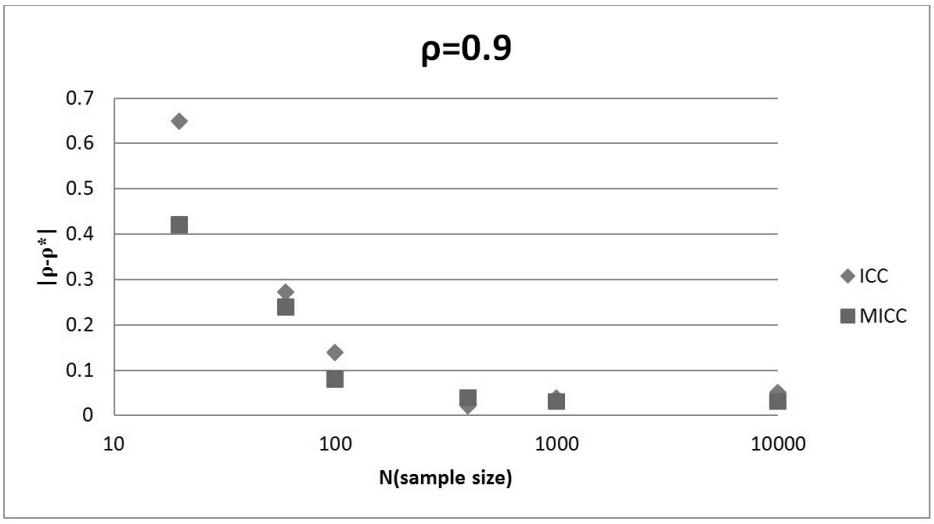


Figure 4: Monte Carlo biases of *ICC* and *MICC* at $\rho = 0.9$

coefficients. The observed improvement is more considerable on small samples and low values of the correlation coefficient—just in the most difficult cases for *ICC*.

5. Concluding remarks

- The statistical performance of the Linfoot's informational correlation coefficient is studied: considerable biases of estimation are observed, especially on small samples.
- To reduce the biases of *ICC*, a modified symmetric version of it, namely *MICC*, is proposed, which proved to provide much lesser estimation biases as compared to its prototype.
- The proposed modified informational correlation coefficient MICC is recommended for processing Big Data, as the obtained results show that its best performance is achieved on large samples.
- Work in process: since the problem of estimation of nonlinear dependencies between random variables still remains important, it seems advantageous to use informational measures of correlation in this case, and the comparative study of the performance of those and other measures of nonlinear correlation, such as the coefficient of determination, Sarmanov's correlation coefficient (Sarmanov 1958) and distance correlation

coefficient of (Székely *et al.* 2007), is in process.

References

Corless R, Gonnet G, Hare D, Jeffrey D, Knuth D (1996). "On the Lambert W Function." *Advances in Computational Mathematics*, **11**, 329–359.

Gebelein H (1941). "Das Statistische Problem der Korrelation als Variations und Eigenwerproblem und Sein Zuzammenhang mit der Ausgleichsrechnung." *Z. Angewandte Math. Mech.*, **21**, 364–379.

Granger C, Maasoumi E, Racine J (2004). "A Dependence Metric for Possibly Nonlinear Processes." *Journal of Time Series Analysis*, **25**, 649–669.

Hausser J, Strimmer K (2010). "Entropy Inference and the James-Stein Estimator with Application to Nonlinear Gene Association Networks." *Journal of Machine Learning Research*, **10**, 1469–1484.

Joe H (1989). "Relative Entropy Measures of Multivariate Dependence." *Journal of the American Statistical Association*, **84**, 157–164.

Kullback S (1959). *Information Theory and Statistics*. Wiley.

Linfoot E (1957). "An Informational Measure of Correlation." *Information and Control*, **1**, 85–89.

Sarmanov O (1958). "Maximum Correlation Coefficient (Symmetric Case)." *Doklady Akad. Nauk SSSR*, **120**, 715–718.

Shannon C (1948). "A Mathematical Theory of Communication." *Bell System Technical Journal*, **27**, 379–423.

Silvey S (1964). "A Measure of Association." *The Annals of Mathematical Statistics*, **35**, 1157–1166.

Székely G, Rizzo M, Bakirov N (2007). "Measuring and Testing Dependence by Correlation of Distances." *Ann. Statist.*, **35**, 2769–2794.

Affiliation:

Georgy Shevlyakov
Peter the Great St. Petersburg Polytechnic University
Polytechnicheskaya, 29, 195251 St. Petersburg, Russia
Institute for Problems in Mechanical Engineering, Russian Academy of Sciences
V.O., Bolshoy pr., 61, 199178 St. Petersburg, Russia
E-mail: Georgy.Shevlyakov@phmf.spbstu.ru

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 46 *Submitted:* 2016-11-15
April 2017 *Accepted:* 2017-02-02

Austrian Journal of Statistics
April 2017, Volume 46, 107–115.
http://www.ajs.or.at/
doi:10.17713/ajs.v46i3-4.676

Comparison of Partially Ranked Lists

Eugenia Stoimenova
Institute of Information and Communication Technologies,
Institute of Mathematics and Informatics,
Bulgarian Academy of Sciences

Abstract

In this paper we introduce a measure of closeness of partial rankings based on a metric on permutations, and we analyze some of its properties. We consider two types of partial rankings: ranking the k favorite items out of n and classification into several ordered categories.

Keywords: partial ranking, top-k list, metric on permutations.

1. Introduction

In many situations, there are different methods for analyzing the same data. For example, several methods exist for finding differentially expressed genes using RNA-seq data. They tend to produce similar, but not identical significant genes and rankings of the gene list. When comparing different methods applied to the same data, we are interested in how close are their outputs. The main idea is to define appropriate distance on the sample space. Further, the interpretation of the rough distance between two rankings should be made on the basis of its statistical significance. That means we need to know the distribution of the distance under some common hypotheses about a sample of rankings. In recent years, many new applications appear in different areas including bioinformatics pattern recognition, information retrivial Jurman, Merler, Barla, Paoli, Galea, and Furlanello (2007), Jurman, Riccadonna, Visintainer, and Furlanello (2009), Chan, Yan, Kittler, and Mikolajczyk (2015), Fagin, Kumar, Mahdian, Sivakumar, and Vee (2006), Fagin, Kumar, and Sivakumar (2003), etc.

In this paper we define an appropriate mathematical framework that include special cases of partially ranked lists of items. Any ranked list can be complete, which means all n items are ranked, or incomplete, which means some items are not ranked. The incomplete ranking include the case where the most significant k items are ranked, with group $k+1$ consisting of the remaining items. Any ranking of n items corresponds a permutation $\langle \alpha(1), \ldots, \alpha(n) \rangle$ from the set of all permutations S_n. We define appropriate distance measures on S_n in order to compare full or incomplete rankings or rankings of different types. The distance can be thought of as a measure of the similarity of the two rankings.

Let α and β be two permutations from S_n corresponding to two rankings and let d be a metric on the permutation group S_n. Then $d : S_n \times S_n \to [0, \infty)$ satisfies the usual axioms: $d(\alpha, \beta) \geq 0 \quad \forall \alpha, \beta \in S_n$, $d(\alpha, \beta) = 0 \Leftrightarrow \alpha = \beta$; $d(\alpha, \beta) = d(\beta, \alpha) \quad \forall \alpha, \beta \in S_n$; and the triangle inequality $d(\alpha, \beta) \leq d(\alpha, \gamma) + d(\gamma, \beta) \quad \forall \alpha, \beta, \gamma \in S_n$.

Invariance is natural in many problems. Right-invariance means that the distance does not depend on arbitrary labeling or reordering of the data:

$$d(\alpha, \beta) = d(\alpha\tau, \beta\tau).$$

Here $\alpha\tau$ is the product of two permutations α and τ and defined by $\alpha\tau(i) = \alpha(\tau(i))$. Right-invariant property allows to compute the distance between two permutations α and β through the the distance of $\alpha\beta^{-1}$ to the identity permutation.

For α and $\beta \in S_n$ the following functions are commonly used as statistical measures of association:

$$\begin{aligned}
F(\alpha,\beta) &= \sum_i \mid \alpha(i) - \beta(i) \mid && \text{Spearman's footrule}\\
R^2(\alpha,\beta) &= \sum_i (\alpha(i) - \beta(i))^2 && \text{Spearman's rho}\\
T(\alpha,\beta) &= \textit{number of pairs } (i,j) \textit{ such that} && \\
&\qquad \alpha(i) < \alpha(j) \textit{ and } \beta(i) > \beta(j) && \text{Kendall's tau}\\
H(\alpha,\beta) &= \#\{i = 1,\ldots,n : \alpha(i) \neq \beta(i)\} && \text{Hamming distance}\\
L(\alpha,\beta) &= \sum_i \min(\mid \alpha(i) - \beta(i) \mid, n - \mid \alpha(i) - \beta(i) \mid) && \text{Lee distance}\\
M(\alpha,\beta) &= \max_{1\le i\le n} \mid \alpha(i) - \beta(i) \mid && \text{Chebyshev's distance}
\end{aligned}$$

All these measures are right-invariant metrics on S_n. By right-invariance of a distance it is sufficient to study its statistical properties when one of the rankings is the identity permutation.

2. Complete or incomplete ranking

A ranking of n items is represented by an ordered n-tuple, which simply lists the items in their ranked order. The most preferred item is listed first, and the least preferred item appears in the n-th position. Any ranking corresponds to a permutation which is an element of the set S_n of permutations. Given a set of rankings, the problem of their comparison reduced to a problem of choosing appropriate measure of association on the set of all rankings. There are several usefull distance measures on S_n thoroughly discussed in statistical literature like Kendall's τ, Spearman's ρ, Spearman's footrule. Therefore, for two permutations $\alpha, \beta \in S_n$ the distance $d(\alpha, \beta)$ can be thought of as a measure of similarity of the two rankings. Excellent references on statistical analysis of rankings are the monographs by Diaconis 1988, Critchlow 1985, and Marden 1995.

There are many situations, in which complete ranking of all n items is not compulsory. The goal might be to rank only their favorite k out of n items or just to choose their k favorite items. In other cases it is important to classify items into groups or categories according to some criterion of "goodness". Further, we need appropriate distances to measure closeness of such rankings.

The general partitioning problem can be described as follows. Let $\{1,\ldots,n\}$ be n given items. We wish to partition them into a fixed number of disjoint categories, such that each category contains a certain preassigned number of items. The first category contains n_1 favorite items, the second category contains the n_2 next preferred items, and so on; the final category contains the n_r least favorite items, where $\sum n_i = n,\ n_i \geq 1$. We do not state any preferences among members of the same category.

If we assign values to r and n_i we obtain several special cases of interest.

(1) To choose the best single item ($r = 2,\ n_1 = 1,\ n_2 = n - 1$);

(2) To choose the best k items without regard to order ($r = 2, n_1 = k, n_2 = n - k$);

(3) To choose the best k items with regard to order ($r = k+1, n_1 = \ldots = n_k = 1,\ n_{k+1} = n-k$);

(4) To order all items ($r = n, n_1 = \ldots = n_r = 1$);

(5) To partition the items into a fixed number of categories.

Many of the decision procedures that one might use within the scope of these ranking problems have a corresponding structure which is invariant under a group of transformations. We consider suitable models for analysis of such partially ranked data thoroughly described by Critchlow 1985.

The full ranking (goal 4) of n items is viewed as an element of the permutation group S_n. The corresponding permutation $\alpha \in S_n$ is a bijection function from $\alpha : \{1, \ldots, n\} \to \{1, \ldots, n\}$ onto itself, where $\alpha(i)$ denotes the rank given to item i and $\alpha^{-1}(i)$ denotes the item assigned to rank i.

With composition of permutations defined by $(\alpha\beta)(i) = \alpha(\beta(i))$, the set S_n constitutes the permutation group. The inverse permutation α^{-1} satisfies $\alpha\alpha^{-1} = \alpha^{-1}\alpha = e$, where e is the identity permutation $e(i) = i$.

2.1. The ranking of the k favorite items out of n

This is probably the most popular goal in ranking problems. Any such partial ranking is identified with permutation from the subgroup $S_{n-k} \subset S_n$ which leaves the first k integers fixed and permutes the remaining $n - k$ integers between themselves:

$$S_{n-k} = \{\alpha \in S_n : \quad \alpha(i) = i \quad for\ all \quad i = 1, \ldots, k\}.$$

Define an equivalence relation on S_n as follows: two permutations α and β are equivalent if and only if there exists $\gamma \in S_{n-k}$ so that $\alpha = \gamma\beta$. For any $\alpha \in S_n$, the equivalence class $S_{n-k}\alpha$ induced by α consists of all permutations equivalent to α. Hence, each partial ranking of k out of n items can be identified with the set of all full permutations which induce it. The set of all such partial rankings can be identified with the set of all such right cosets. Clearly, there is a one-to-one correspondence between the partial rankings of type "k out of n" and right cosets of S_{n-k}. This coset space is denoted by S_n/S_{n-k}.

2.2. Classification into r ordered categories

Let $n_1, \ldots, n_r$ be an ordered sequence of r strictly positive numbers summing to n. Such an ordered partition corresponds to a partial ranking with n_1 items in the first group, n_2 items in the second group and so on. No further information is conveyed about orderings within each group. The special case of ranking the top k items corresponds to $n_1 = \cdots = n_k = 1,\ n_{k+1} = k + 1$.

Formally, denote $N_1, \ldots, N_r$ are the following partition of $\{1, \ldots, n\}$:

$$\begin{aligned} N_1 &= \{1, \ldots, n_1\} \\ N_2 &= \{n_1 + 1, \ldots, n_1 + n_2\} \\ . \quad &. \quad . \\ N_r &= \{n_1 + \cdots + n_{r-1} + 1, \ldots, n\}. \end{aligned} \tag{1}$$

Let S denote the subgroup of all rankings which permute the first n_1 items among the first n_1 ranks, and which permute the next n_2 items among the next n_2 ranks, and so on. The equivalence class $[\alpha]$, that assigns the same set of ranks to the items from the each category as α, is the right coset $S\alpha$. There is a one-to-one correspondence between the partitioning "of type $n_1, \ldots, n_r$" and the right cosets of S.

3. Distances on partial rankings

In the above algebraic structure the problem of comparing of partial rankings is reduced to a problem of extending the metrics on the permutation group S_n to metrics on the corresponding coset space. We discuss an extension of the above metrics for the cases of partial rankings. One natural way of extending it is to construct the induced Hausdorff metrics. Its particular benefit is that it keeps the metric properties of the original distance.

Proposition 1. *Let G be an arbitrary finite group, K be any subgroup of G, and d be a right-invariant metric on G. Then d induces a right-invariant metric on the coset space G/K defined by*

$$d^*(K\alpha, K\beta) = \max\left\{\max_{\sigma\in K\beta}\min_{\pi\in K\alpha} d(\pi,\sigma), \max_{\pi\in K\alpha}\min_{\sigma\in K\beta} d(\pi,\sigma)\right\}.$$

In this formula, the quantity $\min_{\pi\in K\alpha} d(\pi,\sigma)$ is the distance between σ and the set $K\alpha$. Therefore, the quantity $\max_{\sigma\in K\beta}\min_{\pi\in K\alpha} d(\pi,\sigma)$ is the maximal distance of a member of $K\beta$ to the set $K\alpha$. Similarly, $\max_{\pi\in K\alpha}\min_{\sigma\in K\beta} d(\pi,\sigma)$ is the maximal distance of a member of $K\alpha$ to the set $K\beta$.

We focus on Chebyshev's metric between partial rankings. These ere obtained by suitable generalization the M distance on S_n to coset spaces of S_n. The Hausdorff versions of five other metrics are due to Critchlow 1985.

4. Chebyshev's metric for partial rankings

In this section we derive an extension of Chebyshev's metric for partial rankings of type (3) and (5). Theorems 1 and 2 below state the extensions of Chebyshev's metric to the metric on the coset spaces S_n/S_{n-k} and S_n/S. The extensions preserve the invariant properties of the metric. The construction is based on the Hausdorff distance between cosets.

The Hausdorff metrics on S_n/S_{n-k} induced by Chebyshev's metric is defined by taking $G = S_n$ and $K = S_{n-k}$ in Proposition 1.

Theorem 1. *Let A, B, D, E be the following partition of $\{1,\ldots,n\}$:*

$$\begin{aligned} A &= \{i = 1,\ldots,n;\ \alpha(i) \le k,\ \beta(i) \le k\} \\ B &= \{i = 1,\ldots,n;\ \alpha(i) \le k,\ \beta(i) > k\} \\ D &= \{i = 1,\ldots,n;\ \alpha(i) > k,\ \beta(i) \le k\} \\ E &= \{i = 1,\ldots,n;\ \alpha(i) > k,\ \beta(i) > k\}. \end{aligned}$$

Then the Hausdorff metrics metrics on S_n/S_{n-k} induced by Chebyshev metric are

$$\begin{aligned} M^*(S_{n-k}\alpha, S_{n-k}\beta) &= \max\Big\{\delta(k-h)\max_{m\in A}|\alpha(m)-\beta(m)|, \\ &\qquad \delta(h)(n-p_1), \delta(h)(n-s_1), \delta(n-k-h)h\Big\}. \end{aligned}$$

Here $p_1 < \ldots < p_h$ is an ordering of the set $\cup_{i\in B}\{\alpha(i)\}$, $s_1 < \ldots < s_h$ is an ordering of the set $\cup_{i\in D}\{\beta(i)\}$ and h is the number of elements in B (or D), and $\delta(x) = 1$ for $x > 0$ and 0 for $x \le 0$.

The proof is in the Appendix.

The Hausdorff metric on S_n/S induced by Chebyshev's metric is defined by taking $G = S_n$ and $K = S$, in the Proposition 1.

Theorem 2. *Let n_{ij} be the number of elements in the set $\{\alpha^{-1}(N_i) \cap \beta^{-1}(N_j)\}$. Then*

$$\begin{aligned} M^*(\alpha,\beta) = \max_{1<i,j<r}\Big[\delta(n_{ij})\max\Big\{ &|\sum_{k=1}^{i-1} n_k + \sum_{k=1}^{j-1} n_{ik} - \sum_{k=1}^{i-1} n_k - \sum_{k=i+1}^{r} n_{kj}|, \\ &|\sum_{k=1}^{i-1} n_k + \sum_{k=j+1}^{r} n_{ik} - \sum_{k=1}^{j-1} n_k - \sum_{k=1}^{i-1} n_{kj}|\Big\}\Big] \end{aligned}$$

is right-invariant metric on S_n/S.

The proofs of these theorems are based of the fact that Chebyshev's metric on S_n satisfies *the transposition property*.

Definition 1. ***(Transposition property)*** *Let $\alpha, \beta, \gamma \in S_n$ be permutations such that α and β differ by a single transposition; that is there exist integers $p, q \in \{1, \ldots, n\}$ such that*

$$\begin{aligned} \alpha(p) &= \beta(q) \\ \alpha(q) &= \beta(p) \\ \alpha(i) &= \beta(i) \quad \forall i \neq p, q. \end{aligned}$$

If $\alpha(p) \leq \alpha(q)$ and $\gamma(p) \leq \gamma(q)$, then $M(\alpha, \gamma) \leq M(\beta, \gamma)$.

Proposition 2. *Chebyshev's metric satisfies the transposition property.*

The proof can be found in Stoimenova (2000). For metrics possessing the transposition property, the permutations $\beta_{\max}$, $\alpha_{\min}(\beta_{\max})$, $\alpha_{\max}$ and $\beta_{\min}(\alpha_{\max})$ have a simple special form.

5. Distributional properties of the metrics

Suppose that two partial rankings α^* and β^* are generated independently from a uniform distribution on all possible partial rankings and calculate the distance $d^*(\alpha^*, \beta^*)$. Thus the distance is a random variable and one might study its distribution on the set of permutations. Figure 1 (left) shows the distribution of Chebyshev's metric for full rankings based on 10000 choices of σ from a uniform distribution. Since the metric is right invariant we calculate the distribution of the distance from the identity permutation.

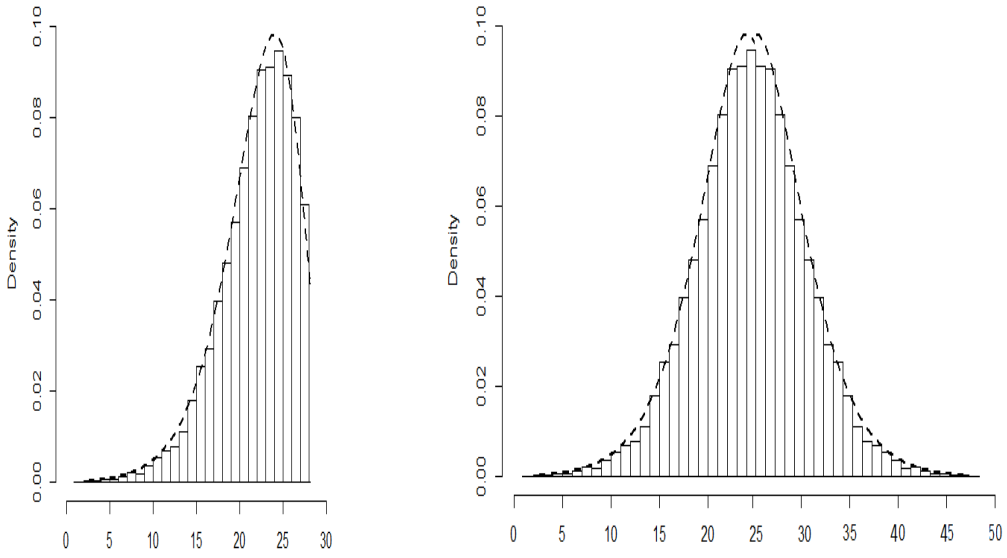


Figure 1: Distribution of Chebyshev distance and Spearman's rho between 2 random permutations

It is evident in the figure that the distribution of the Chebyshev's metric is left skewed and it has similar form on partial rankings as well. We suggest a chi-square approximation of the distribution. The other metrics discussed in Section 1 have symmetrical distributions and for large n they exhibit normality. The distribution of Spearman's rho for full rankings is presented on Figure 1 (right). Normal approximation is in the following sense.

Definition 2. *The metric d^* on S_n/S is asymptotically normally distributed if for partial rankings α^* and β^* the following limit distribution is valid*

$$\lim_{n\to\infty} P\left(\frac{d^*(\alpha^*, \beta^*) - E\, d^*(\alpha^*, \beta^*)}{\sqrt{var(d^*(\alpha^*, \beta^*))}} \leq x\right) = \Phi(x)$$

for all real numbers x, where Φ, is the standard normal cumulative distribution function.

By using the exact or the approximate distribution of a distance on permutations, one can calculate the probability that d^* is less than or equal to the observed value $d^*(\alpha^*, \beta^*)$. This probability is the p-value for α^* and β^*. Smaller values of p indicate stronger evidence that α^* and β^* are "similar".

To compute the p-value, Critchlow finds the probability distribution of some popular metrics on permutations under the appropriate uniformity assumption. The critical values of the distribution of Chebyshev's metric under uniformity assumptions can be calculated for different choices of the sizes of the categories. The R code for Chebyshev's metric is available by the author on request. Therefore, the significance of the distance can be used to estimate the similarity between the two partial rankings. The interpretation is very much like as the significance of a correlation coefficient.

Chan *et al.* 2015 computes the distributions of several metrics between ranking descriptors in a texture image from a real dataset and applies them to image intensities and to filter responses. The distributions of Chebyshev's metric and Spearman's rho have the same features we see on Figure 1.

Acknowledgments. The author acknowledge funding by the Bulgarian fund for scientific investigations Project I02/19.

References

Chan CH, Yan F, Kittler J, Mikolajczyk K (2015). "Full Ranking as Local Descriptor for Visual Recognition: A Comparison of Distance Metrics on S_n." *Pattern Recognition*, **48**(4), 1328–1336.

Critchlow DE (1985). *Metric Methods for Analyzing Partially Ranked Data.* Lecture Notes in Statistics, 34. Berlin etc.: Springer-Verlag.

Diaconis P (1988). *Group Representations in Probability and Statistics.* IMS Lecture Notes-Monograph Series, 11. Hayward, CA: Institute of Mathematical Statistics.

Fagin R, Kumar R, Mahdian M, Sivakumar D, Vee E (2006). "Comparing Partial Rankings." *SIAM J. Discrete Math.*, **20**(3), 628–648.

Fagin R, Kumar R, Sivakumar D (2003). "Comparing Top k Lists." *SIAM J. Discrete Math.*, **17**(1), 134–160.

Jurman G, Merler S, Barla A, Paoli S, Galea A, Furlanello C (2007). "Algebraic Stability Indicators for Ranked Lists in Molecular Profiling ." *Bioinformatics*, **24**(2), 258–264.

Jurman G, Riccadonna S, Visintainer R, Furlanello C (2009). "Canberra Distance on Ranked Lists." In KC Agrawal C Burges (ed.), *In Proceedings of Advances in Ranking NIPS 09 Workshop*, pp. 22–27.

Marden JI (1995). *Analyzing and Modeling Rank Data.* Monographs on Statistics and Applied Probability. 64. London: Chapman.

Stoimenova E (2000). "Rank Tests Based on Exceeding Observations." *Ann. Inst. Stat. Math.*, **52**(2), 255–266.

Appendix: Proofs of Theorems

Proof of Theorem 1. The result for S_n/S_{n-k} is a special case of Theorem 2 with $r = k + 1,\ n_1 = \ldots = n_k = 1$. We derive it separately because of its simpler form.

Since the Chebyshev's distance is right invariant the induced Hausdorff metric on S_n/S_{n-k} is represented as:

$$\begin{aligned} M^*(S_{n-k}\alpha, S_{n-k}\beta) &= \max\Big[\max_{\sigma\in S_{n-k}\beta}\min_{\pi\in S_{n-k}\alpha} d(\pi,\sigma), \max_{\pi\in S_{n-k}\alpha}\min_{\sigma\in S_{n-k}\beta} d(\pi,\sigma)\Big] \\ &= \max\Big[\max_{\sigma\in S_{n-k}\beta} d(\alpha_{\min}(\sigma),\sigma), \max_{\pi\in S_{n-k}\alpha} d(\pi,\beta_{min}(\pi))\Big] \\ &= \max\Big[M(\alpha_{\min}(\beta_{\max}),\beta_{\max}), M(\beta_{\min}(\alpha_{\max}),\alpha_{\max})\Big], \end{aligned}$$

where for any $\beta \in S_n$, $\alpha_{\min}(\beta) = \min M(.,\beta)$, $\beta_{\max} = \max M(\alpha_{\min}(.),.)$, and for any $\alpha \in S_n$, $\beta_{\min}(\alpha) = \min M(\alpha,.)$, $\alpha_{\max} = \max M(.,\beta_{\min}(.))$.

Since the Chebyshev's distance satisfies the Transposition property 1 then $\beta_{\max}$, $\alpha_{\min}(\beta_{\max})$, $\alpha_{\max}$ and $\beta_{\min}(\alpha_{\max})$ have the forms:

$$\begin{aligned} \beta_{\max} &= <\beta^{-1}(1),\ldots,\beta^{-1}(k), N_{k+1,k+1}, \alpha^{-1}(p_h),\ldots,\alpha^{-1}(p_1)> \\ \alpha_{\min}(\beta_{\max}) &= <\alpha^{-1}(1),\ldots,\alpha^{-1}(k), \beta^{-1}(s_1),\ldots,\beta^{-1}(s_h), N_{k+1,k+1}> \\ \alpha_{\max} &= <\alpha^{-1}(1),\ldots,\alpha^{-1}(k), N_{k+1,k+1}, \beta^{-1}(s_h),\ldots,\beta^{-1}(s_1)> \\ \beta_{\min}(\alpha_{\max}) &= <\beta^{-1}(1),\ldots,\beta^{-1}(k), \alpha^{-1}(p_1),\ldots,\alpha^{-1}(p_h), N_{k+1,k+1}>, \end{aligned}$$

where $N_{k+1,k+1}$ are the elements of the set E in ascending order (or arbitrary fixed order.)

The sets A,B and D are as follows:

$$\begin{aligned} A &= \{\alpha^{-1}(1),\ldots,\alpha^{-1}(k)\} \cap \{\beta^{-1}(1),\ldots,\beta^{-1}(k)\} \\ B &= \{\alpha^{-1}(p_h),\ldots,\alpha^{-1}(p_1)\} \\ D &= \{\beta^{-1}(s_h),\ldots,\beta^{-1}(s_1)\}. \end{aligned}$$

Thus

$$\begin{aligned} \max_{m\in A} \mid \alpha_{\min}(\beta_{\max})(m) - \beta_{\max}(m) \mid &= \max_{m\in A} \mid \alpha(m) - \beta(m) \mid; \\ \max_{m\in B} \mid \alpha_{\min}(\beta_{\max})(m) - \beta_{\max}(m) \mid &= \max_{1\le m\le h} \mid n+1-m-p_m \mid \\ &= \max\{n-p_1, p_h+h-n-1\} = n-p_1; \\ \max_{m\in D} \mid \alpha_{\min}(\beta_{\max})(m) - \beta_{\max}(m) \mid &= \max_{1\le m\le h} \mid k+m-s_m \mid \\ &= \max\{k+1-s_1, s_h-h-k\} = k+1-s_1; \\ \max_{m\in E} \mid \alpha_{\min}(\beta_{\max})(m) - \beta_{\max}(m) \mid &= h. \end{aligned}$$

Hence

$$\begin{aligned} M(\alpha_{\min}(\beta_{\max}),\beta_{\max}) &= \max_{1\le m\le n} \mid \beta_{\min}(\alpha_{\max})(m) - \alpha_{\max}(m) \mid \\ &= \max\Big[\delta(k-h)\max_{m\in A} \mid \alpha(m)-\beta(m) \mid, \delta(h)(n-p_1), \delta(h)(k+1-s_1), \\ &\qquad \delta(h)(s_h-h-k), \delta(n-k-h)h\Big]. \end{aligned}$$

Similarly

$$\begin{aligned} \max_{m\in A} \mid \beta_{min}(\alpha_{max})(m) - \alpha_{max}(m) \mid &= \max_{m\in A} \mid \alpha(m) - \beta(m) \mid; \\ \max_{m\in B} \mid \beta_{\min}(\alpha_{\max})(m) - \alpha_{\max}(m) \mid &= \max\{k+1-p_1, p_h-h-k\} = k+1-p_1 \\ \max_{m\in D} \mid \beta_{\min}(\alpha_{\max})(m) - \alpha_{\max}(m) \mid &= \max\{n-s_1, s_h+h-n-1\} = n-s_1 \\ \max_{m\in E} \mid \beta_{\min}(\alpha_{\max})(m) - \alpha_{\max}(m) \mid &= h. \end{aligned}$$

and hence

$$M(\beta_{\min}(\alpha_{\max}),\alpha_{\max}) = \max_{1\le m\le n} \mid \beta_{\min}(\alpha_{\max})(m) - \alpha_{\max}(m) \mid$$
$$= \max\Big[\delta(k-h)\max_{m\in A} \mid \alpha(m)-\beta(m) \mid, \delta(h)(k+1-p_1), \delta(h)(p_h-h-k), \delta(h)h\Big].$$

Proof of Theorem 2. Note that n_{ij} is the number of items placed in the i-th category by the first judge and in the j-th category by the second. From Lemma 3 (Critchlow (1985), p.53) and Proposition 2 it follows that $\beta_{\max}$, $\alpha_{\min}(\beta_{\max})$, $\alpha_{\max}$ and $\beta_{\min}(\alpha_{\max})$ have the forms:

$$\begin{aligned}
\beta_{\max} &= < N_{r1},\ldots,N_{11},N_{r2},..,N_{12},\ldots,N_{rj},\ldots,N_{1j},\ldots,N_{rr},\ldots,N_{1r} > \\
\alpha_{\min}(\beta_{\max}) &= < N_{11},\ldots,N_{1r},\ldots,N_{i1},\ldots,N_{ir},\ldots,N_{r1},\ldots,N_{rr} > \\
\alpha_{\max} &= < N_{1r},\ldots,N_{11},..,N_{ir},\ldots,N_{i1},\ldots,N_{rr},\ldots,N_{r1} > \\
\beta_{\min}(\alpha_{\max}) &= < N_{11},\ldots,N_{r1},..,N_{1j},\ldots,N_{rj},\ldots,N_{1r},\ldots,N_{rr} > .
\end{aligned}$$

where $N_{i,j}$ are the elements of the set $\{\alpha^{-1}(N_i)\cap\beta^{-1}(N_j)\}$.
Let m_{ij} be the smallest number in N_{ij}. There are

$$\sum_{k=1}^{i-1} n_k + \sum_{k=1}^{j-1} n_{ik}$$

numbers occurring to the left of m_{ij} in the bracket representation of $\alpha_{min}(\beta_{\max})$, and

$$\sum_{k=1}^{i-1} n_k + \sum_{k=i+1}^{r} n_{kj}$$

numbers occurring to the left of m_{ij} in the bracket representation of $\beta_{\max}$. Hence

$$\alpha_{\min}(\beta_{\max})(m_{ij}) - \beta_{\max}(m_{ij}) = \sum_{k=1}^{i-1} n_k + \sum_{k=1}^{j-1} n_{ik} - \sum_{k=1}^{i-1} n_k - \sum_{k=i+1}^{r} n_{kj}.$$

Therefore

$$\begin{aligned}
M^*(\alpha_{\min}(\beta_{\max}),\beta_{\max}) &= \max_{1\le m\le n} |\alpha_{\min}(\beta_{\max})(m) - \beta_{\max}(m)| \\
&= \max_{1\le i,j\le r}\Big[\delta(n_{ij})\max_{m\in N_{ij}} |\alpha_{\min}(\beta_{\max})(m) - \beta_{\max}(m)|\Big] \\
&= \max_{1\le i,j\le r}\Big[\delta(n_{ij}) \mid \sum_{k=1}^{i-1} n_k + \sum_{k=1}^{j-1} n_{ik} - \sum_{k=1}^{j-1} n_k - \sum_{k=i+1}^{r} n_{kj} \mid \Big].
\end{aligned}$$

Similarly there are

$$\sum_{k=1}^{i-1} n_k + \sum_{k=j+1}^{r} n_{ik}$$

numbers occurring to the left of m_{ij} in the bracket representation of $\alpha_{\max}$, and

$$\sum_{k=1}^{i-1} n_k + \sum_{k=1}^{i-1} n_{kj}$$

numbers occurring to the left of m_{ij} in the bracket representation of $\beta_{\min}(\alpha_{\max})$. Hence

$$\beta_{\min}(\alpha_{\max})(m_{ij}) - \alpha_{\max}(m_{ij}) = \sum_{k=1}^{i-1} n_k + \sum_{k=j+1}^{r} n_{ik} - \sum_{k=1}^{j-1} n_k - \sum_{k=1}^{i-1} n_{kj}.$$

Note $\beta_{\min}(\alpha_{\max})(m) - \alpha_{\max}(m)$ is a constant for $m \in N_{ij}$, so

$$\begin{aligned} M(\beta_{\min}(\alpha_{\max}), \alpha_{\max}) &= \max_{1\le i,j\le r} \left[\delta(n_{ij})_m \max_{m\in N_{ij}} \mid \beta_{\min}(\alpha_{\max})(m) - \alpha_{\max}(m)) \mid \right] \\ &= \max_{1\le i,j\le r} \left[\delta(n_{ij})_m \mid \sum_{k=1}^{1-1} n_k + \sum_{k=j+1}^{r} n_{ik} - \sum_{k=1}^{j-1} n_k - \sum_{k=1}^{i-1} n_{kj} \mid \right]. \end{aligned}$$

Affiliation:

Eugenia Stoimenova
Institute of Information and Communication Technologies
& Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Acad. G.Bontchev str., block 25A
1113 Sofia
Bulgaria
E-mail: jeni@parallel.bas.bg

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 46
April 2017

http://www.ajs.or.at/
http://www.osg.or.at/

Submitted: 2016-11-15
Accepted: 2017-02-02

Austrian Journal of Statistics
April 2017, Volume 46, 117–128.
http://www.ajs.or.at/
doi:10.17713/ajs.v46i3-4.677

Depth-based Classification for Multivariate Data

Ondřej Vencálek
Palacký University in Olomouc

Abstract

Concept of data depth provides one possible approach to the analysis of multivariate data. Among other it can be also used for classification purposes. The present paper is an overview of the research in the field of depth-based classification for multivariate data. It provides a short summary of current state of knowledge in the field of depth-based classification followed by detailed discussion of four main directions in the depth-based classification, namely semiparametric depth-based classifiers, maximal depth classifier, (maximal depth) classifiers which use local depth functions and finally advanced depth-based classifiers. We do not restrict our attention only on proposed classifiers. The paper rather aims to overview the ideas connected with depth-based classification and problems that were discussed in this context.

Keywords: data depth, classification, Bayes optimal, overview.

1. Introduction

Depth function is basically any function which provides ordering (or quasi-ordering) of points in multidimensional space $\mathbb{R}^d$ with respect to some probability measure P defined on this space. Existence of ordering enables generalization of quantiles (median, in special case) and related nonparametric techniques proposed for univariate variables. Thus the notion of depth creates one possible basis of nonparametric multivariate data analysis.

Data depth has been also applied in classification. We use the term classification, where sometimes term discriminant analysis or supervised learning is used. The aim of classification is to create a rule for allocation of new observations into one of two (or more) groups. Formally, we consider two unknown absolutely continuous probability distributions P_1 and P_2 on $\mathbb{R}^d$. Independent random samples from these distributions are available. Together they constitute so called training set. Empirical distributions based on the training set are denoted $\widehat{P}_1$ and $\widehat{P}_2$. A classifier (rule for classification) is thus a function $c : \mathbb{R}^d \to \{1, 2\}$. The theory can be generalized for more than two groups.

Possibility to use data depth for classification was firstly mentioned by Liu already in 1990, Liu (1990). She suggested that the new observation "should be assigned to the population whose training sample leads to a smaller relative rank for it." However, after this first reference it lasted more than ten years until the depth-based classifiers started to be studied systematically. Thus we can say that the depth-based classification has been developed since 2000.

The aim of the present paper is to summarize main ideas how the depth can be used in classification. The paper renders readers interested in classification insight to the current state of knowledge in the field of depth-based classification. Also it provides a comprehensive overview of the research made in the area of depth-based classification in the last 15 years.

2. Short guide on depth-based classification

The reader not familiar with the depth-based classification might ask in which situation it was useful to use this concept. In order to facilitate the basic orientation in this area we shortly summarize current state of knowledge in the field of depth-based classification. We do it even before dealing with particular methods of depth-based classification since it might be useful for the reader to know what to expect from the depth-based classification as soon as possible.

- The depth-based classifiers are applicable when one want to avoid strict parametric assumptions (like normality) on the considered distributions. They can utilize possible global properties of the considered distributions (like their symmetry), but can be applied even if there are no such properties, e.g. for non-symmetric distributions. They work well for unimodal distributions since the depth is global property which characterizes location of a point w.r.t. the whole distribution. If the considered distributions might be multimodal or could have nonconvex levelsets of density a local depth should be used to overcome this problem.
- Simple advice how to decide weather it is reasonable to use depth-based classification is to answer the question weather it is reasonable to classify by median and other quantiles. The median which is the point with highest depth might lie in the area where the density is low. In such a case depth-based classifiers would have problems. Only few of them are able to overcome these problems.
- Depth-based classifiers were primarily constructed for continuous distributions, little attention was paid for categorical "explanatory" variables.
- One should be aware that the computation of depth is not easy task and for most of the depth functions might be very slow in dimensions higher than five when the number of points in training set is high. The depth-based procedures are advisable in dimensions from 2 to 5, they are not advisable in dimensions higher than 20. However, the latest depth-based classifiers, as the one proposed by Dutta and Ghosh (2015), were shown to perform well even in dimension 100. For really high-dimensional data the depth-based classifiers usually need to be preceded by reduction of dimensionality.
- When more than two classes are considered, the depth-based classifiers usually rely on majority voting principle.
- The main advantage of the depth-based classifiers is their affine invariance. Most of the depth functions are affine invariant and thus the classification procedures do not change e.g. with the change of units (scales). The depth-based classifiers also usually have good robust properties.
- The simplest classifiers, so called maximal depth classifiers, which will be described in section 4.2 are not satisfactory in most practical situations. More complicated classifiers described in section 4.4 need to be employed. Probably the most universal depth-based classifier is that by Paindaveine and Van Bever (2015) (see end of the section 4.4) since it is shown to be nonparametrically consistent under very mild conditions.
- For practitioners who need reliable implemented classifiers an R-package ddalpha is advisable, see Pokotylo, Mozharovskyi, and Dyckerhoff (2016). In this package the DD-plot classifier by Li, Cuesta-Albertos, and Liu (2012) and the DDα-classifier by Lange,

Mosler, and Mozharovskyi (2014b) are implemented. Both these classifiers (mentioned in section 4.4) belong to the best depth-based classifiers that are currently available.

- There are also depth-based methods for classification of functional data. They utilize so called functional data depth. The current paper deals with classification for multivariate data so the classification of functional data is not included.

3. Concept of data depth

There are several depth functions commonly used for classification – halfspace depth, projection depth, spatial depth, Mahalanobis depth, zonoid depth, and some others. Let us recall here the first four of the depth functions listed above:

- The *halfspace depth* of a point $\boldsymbol{x}$ in $\mathbb{R}^d$ with respect to a probability measure P is defined as the minimum probability mass carried by any closed halfspace containing $\boldsymbol{x}$, that is

$$D(\boldsymbol{x}, P) = \inf\left\{\mathrm{P}(\mathbb{H}) : \mathbb{H} \text{ a closed halfspace in } \mathbb{R}^d : \boldsymbol{x} \in \mathbb{H}\right\}.$$

- The *projection depth* of a point $\boldsymbol{x}$ in $\mathbb{R}^d$ with respect to a probability measure P is defined as

$$D(\boldsymbol{x}, P) = \frac{1}{1 + O(\boldsymbol{x}, P)}, \quad \text{where} \quad O(\boldsymbol{x}, P) = \sup_{\|\boldsymbol{u}\|=1} \frac{|\boldsymbol{u}'\boldsymbol{x} - \mu_{P_{\boldsymbol{u}}}|}{\sigma_{P_{\boldsymbol{u}}}},$$

  where $\mu_{P_{\boldsymbol{u}}}$ is some location and $\sigma_{P_{\boldsymbol{u}}}$ is some scale measure of distribution of random variable $\boldsymbol{u}'\boldsymbol{X}$ ($\boldsymbol{X} \sim P$), usually $\mu_{P_{\boldsymbol{u}}} = \mathrm{median}(\boldsymbol{u}'\boldsymbol{X})$ and $\sigma_{P_{\boldsymbol{u}}} = \mathrm{MAD}(\boldsymbol{u}'\boldsymbol{X})$, where MAD stands for median absolute deviation.

- The *Mahalanobis depth* of a point $\boldsymbol{x}$ in $\mathbb{R}^d$ with respect to a probability measure P with mean $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{\Sigma}$ is defined as

$$D(\boldsymbol{x}, P) = \frac{1}{1 + O(\boldsymbol{x}, P)}, \quad \text{where} \quad O(\boldsymbol{x}, P) = (\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}).$$

- The *spatial depth* (also called L_1*-depth*) of a point $\boldsymbol{x}$ in $\mathbb{R}^d$ with respect to a probability measure P with variance matrix $\boldsymbol{\Sigma}$ is defined as

$$D(\boldsymbol{x}, P) = 1 - \mathrm{E}_P \left\| \frac{\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{X})}{\left\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{X})\right\|} \right\|.$$

All the depth functions listed above have desirable properties like affine invariance, maximality at a point of symmetry (if the distribution is symmetric in some sense, e.g. angularly), monotonicity on rays from the point with the maximal depth – so called deepest point, which can be considered as multivariate analogy to median.

The very important difference among the considered depth functions consists in different behaviour of their empirical versions. While the empirical halfspace depth is equal to zero for any point which lies outside of the convex hull of the data, the empirical versions of the latter three depth functions are nonzero everywhere.

The depth of a point is a characteristic of the point specifying its centrality or outlyingness with respect to the considered distribution. Since the whole distribution is considered, the depth is said to be a "global" characteristic of the point. However, in recent years there have been attempts to "localize" depth. Later we will discuss importance of these attempts for classification purposes.

4. Depth-based classifiers

We can distinguished four main groups of depth-based classifiers for multidimensional data: semiparametric classifiers which use data depth, maximal depth classifier which use global depth functions, maximal depth classifiers which use local depth functions and "advanced" depth-based classifiers.

4.1. Semiparametric depth-based classifiers

- The first thorough study devoted to possible use of data depth in context of classification entitled "Measuring overlap in binary regression" was done by Christmann and Rousseeuw (2001). The authors considered classical logistic regression model in which the 0-1 response variable coded group membership in two classes classification problem. They realized that the regression depth can be used to measure amount of separation between the two groups. Enumeration of the regression depth of one particular 'fit' is accompanied by finding the hyperplane which minimizes number of misclassified points from the training set. Similar ideas can be found also in Christmann, Fischer, and Joachims (2002) and Christmann (2006).

- Ghosh and Chaudhuri extended substantially the above mentioned ideas, see Ghosh and Chaudhuri (2005a). They came with the following extensions: (1) They found connection of the linear classification with the halfspace depth ("the estimated linear projection is orthogonal to the hyperplane, which defines the halfspace depth of the origin w.r.t. the data cloud formed by the differences $\boldsymbol{x}_{1i} - \boldsymbol{x}_{2j}$ in the d-dimensional space", where $\boldsymbol{x}_{1i}$ are observations from one group and $\boldsymbol{x}_{2j}$ from the other). (2) They suggested use of weighed regression depth in linear classification based on regression depth to deal with the situation in which prior probabilities are not proportional to their training sample sizes. (3) They generalized the classifiers for nonlinear separating surfaces. To construct such surfaces they projected the original d-dimensional observations $\boldsymbol{x}_i$ into a higher-dimensional space of features $\boldsymbol{z}_i = (f_1(\boldsymbol{x}_i), \ldots, f_h(\boldsymbol{x}_i))$, where $f_1(\cdot), \ldots, f_h(\cdot)$ are some given functions, e.g. powers, and performed linear classification on that h-dimensional space. In this way they could obtained e.g. quadratic classification. Similar idea, but in slightly different context can be found in Lange *et al.* (2014b). (4) They discussed the problem of more than two groups. They proposed majority voting or pairwise coupling.

4.2. Maximal depth classifier

- The simplest depth-based classifier is so called maximal depth classifier. The points close to the centre (multivariate median) of some distribution have high depth with respect to this distribution and it seems to be natural to classify them to this distribution. The idea of the maximal depth classifier – to classify a new observation to the group where its depth is maximal – is thus in accordance to common sense:

$$c(\boldsymbol{x}) = \arg\max_{i=1,2} D(\boldsymbol{x}; \widehat{P}_i) \tag{1}$$

  Different authors advocated use of different depth functions in this classifier. As far as we know this classifier was first studied by Jörnsten (2004). She used spatial depth, similarly as Hartikainen and Oja (2006). Jörnsten also came with the idea of relative depth which measures uncertainty of classification of a given point. The relative depth is defined as difference between maximal and second highest depth, i.e. $\max_i D(\boldsymbol{x}; \widehat{P}_i) - \max_{i \neq c(\boldsymbol{x})} D(\boldsymbol{x}; \widehat{P}_i)$. If the difference is high we are pretty sure in classification. Small relative depth indicates possible problems in the class assignment. Jörnsten suggested deletion of points with small relative depth from the training sample

which might lead to improvement in classification. The classifier proposed by Jörnsten was used in comparative study focused on classifiers for high-dimensional data presented in Hall, Titterington, and Xue (2009). However, this study highlighted rather componentwise median-based classifier and its truncated form.

Mosler and Hoberg were first who pointed out so called "outsider" problem in Mosler and Hoberg (2006). The problem consists in zero empirical depth of points that are outside of the convex hull of points in the training set when using some depth functions like zonoid depth or halfspace depth. To overcome the outsider problem they suggested to combine zonoid depth with Mahalanobis depth which does not suffer from this problem.

Use of projection depth in the maximal depth classifier was advocated by Kosiorowski (2008) who emphasized good robust properties of such a classifier. Also the classifier proposed by Hubert and Van der Veeken (2010b) can be understand as the maximal depth classifier using projection depth where the outlyingness adjusted for skewness of the distribution is used instead of commonly used outlyingness. Similarly as Jörnsten they consider removal of the possible problematic points from the training set before the construction of the final classifier. Their contribution can be viewed also in (rather exceptional) discussion of depth-based classification for high-dimensional data. They advocated use of robust SIMCA (Soft Independent Modelling by Class Analogy) method which lies in application of (robust) PCA method in each group. An extended projection depth which takes into account possible difference in dispersion of considered distributions was suggested by Cui, Lin, and Yang (2008).

An interesting technical application of maximal depth classifier for real time sensor node tracking and location was recently proposed in Kumar, Kumar, Kumar, and Hegde (2015). Specificity of this classifier lies in the fact that not one point, but the group of multidimensional points is classified at once.

- Although many authors suggesting maximal depth classifier refered to the original paper by Liu (1990), the idea presented in that paper was to classify rather according to the relative rank (based on depth), not the depth itself (see section 1). This idea was rediscovered much later by Billor, Abebe, Turkmen, and Nudurupati (2008). Although they call their classifier as "depth transvariation classifier", it can be called more directly maximal rank classifier since it has the following form:

$$c(\boldsymbol{x}) = \arg\max_{i=1,2} \operatorname{rank}(\boldsymbol{x}; \widehat{P}_i), \tag{2}$$

where $\operatorname{rank}(\boldsymbol{x}; \widehat{P}_i)$ denotes percentage of points from the training set of the i-th group which have smaller depth w.r.t. $\widehat{P}_i$. The same idea was also considered by Hubert and Van der Veeken (2010a).

The maximal depth classifiers mentioned in this section are already outdated and were replaced by maximal depth classifiers that use some local depth described in section 4.3 or by more advanced classifiers described in section 4.4. Insufficiency of the maximal depth classifiers was revealed already by Ghosh and Chaudhuri (2005b). They proved that the maximal depth classifier attains minimal possible probability of misclassification (known as Bayes risk) only in very special cases – they showed optimality when the considered distributions are elliptically symmetric with the density decreasing from the centre, differing only in location and having equal prior probabilities. The optimality is lost even if only one of these assumptions is not fulfilled. The following simple example illustrates these problems. Let us consider two one-dimensional normal distributions with the same mean but different variances. Then all points different from the mean will be classified to the distribution with smaller variance. This is due to the affine invariance of the depth.

4.3. Maximal depth classifiers with local depth

The depth of a given point characterizes its location w.r.t. the whole distribution. Thus the classifiers which use any "global" depth function perform well only if the considered distributions have some global properties like symmetry or unimodality. To obtain good performance also in more general settings, for example in the case of multimodality or nonconvexity of levelsets of density, use of some local depth started to be promoted recently. A new problem that emerged and need to be handle is choice of localization level.

The first classifiers which employed local depth appeared only in 2013. Hlubinka and Vencalek (2013) used weighted halfspace depth. Paindaveine and Van Bever (2013) developed more complex approach which enables localization of any "global" depth function which can be subsequently used in the maximal depth classifier. The proposed local depth is defined as global depth conditional on some neighborhood of the point of interest. The neighborhood itself is defined in terms of data depth. It is worthwhile to recall the main ideas leading to the localization of depth here:

Let $P^{\boldsymbol{X}}$ be a probability distribution of a d-dimensional random variable $\boldsymbol{X}$, let $D(\cdot, P^{\boldsymbol{X}})$ be any depth function and let $\beta \in (0, 1]$. *Depth regions* of a distribution $P^{\boldsymbol{X}}$ are sets of the following form: $\left\{\boldsymbol{x} \in \mathbb{R}^d : D(\boldsymbol{x}, P^{\boldsymbol{X}}) \geq \alpha\right\}$. The *symmetrized version of a distribution* $P^{\boldsymbol{X}}$ with the centre in a given point $\boldsymbol{x} \in \mathbb{R}^d$ is defined as a mixture $P_{\boldsymbol{x}} = \frac{1}{2}P^{\boldsymbol{X}} + \frac{1}{2}P^{2\boldsymbol{x}-\boldsymbol{X}}$. *The probability β neighborhood of a point $\boldsymbol{x} \in \mathbb{R}^d$ w.r.t. $P^{\boldsymbol{X}}$* is defined as the smallest depth region of $P_{\boldsymbol{x}}$ with $P_{\boldsymbol{x}}$ probability larger than or equal to β. It is denoted by $R^{\beta}(P_{\boldsymbol{x}})$. The *$\beta$-local depth of a point $\boldsymbol{x} \in \mathbb{R}^d$ w.r.t. $P^{\boldsymbol{X}}$* is defined as $D(\boldsymbol{x}, P^{\beta}_{\boldsymbol{x}})$, where $P^{\beta}_{\boldsymbol{x}}$ is conditional distribution of $P^{\boldsymbol{X}}$, conditional on depth neighborhood of $\boldsymbol{x}$ $R^{\beta}(P_{\boldsymbol{x}})$.

An interesting economical application of maximal depth classifier which uses local L^p depth for so called algorithmic trading was presented by Kosiorowski, Bocian, and Bujak (2014). The classes considered in the paper characterize different states of market.

4.4. Advanced depth-based classifiers

The paper by Ghosh and Chaudhuri (2005b) uncovered insufficiency of the maximal depth classifier and started the search for depth-based classifiers which would be applicable in a broad class of distributional settings. Typical depth-based classifier can be described as a two-steps procedure:

1. The first step consists in computation of depths of the new observation $\boldsymbol{x}$ with respect to both parts of the training set. Each point is characterized by a pair of depths, these pairs lies in so called DD-space (depth-versus-depth space). Typically the DD-space is subset of $[0, 1] \times [0, 1] \subset \mathbb{R}^2$ and thus the first step can be usually (but not necessarily) considered as reduction of dimensionality – from $\mathbb{R}^d$ to the compact subset of $\mathbb{R}^2$. This step is connected with the question "Which depth function should be used?"

2. The second step consists in application of some classification procedure in the DD-space. This step is connected with the question "Which classification procedure should be applied in the DD-space?" This "new" question is in the center of current research in the field of depth-based classification.

The scheme of typical depth-based classification procedure is shown in Figure 1.

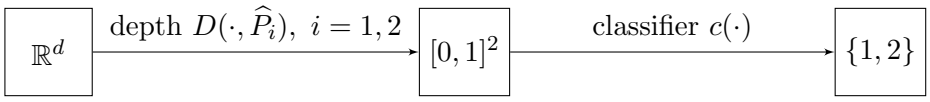


Figure 1: Scheme of typical advanced depth-based classifier.

The difference among the classifiers proposed in literature consists mainly in different answers to the two questions connected with the two steps – different depth functions can be applied in

the first step and different classification procedures can be applied in the second step. Depth function used in the first step can be either global or local. Also procedures that are applied in the DD-space can be also either "global" or "local" in nature. As the global we denote the procedures that take into account all points of the training set when constructing the classifier while the local procedures take into account only points of the training set close to the point which is classified.

The list of advanced depth-based classifiers follows:

- The first classifier which overcame insufficiency of the maximal depth classifier was suggested already by Ghosh and Chaudhuri (2005b). They discovered that in the case of elliptically symmetric distribution the classifier which minimizes probability of misclassification can be expressed as

$$c(\boldsymbol{x}) = \arg\max_{i=1,2} \pi_i \theta_i(D(\boldsymbol{x}, \widehat{P}_i)),$$

  where π_i are prior probabilities and $\theta_i, i = 1, 2$ are some unknown real functions. This holds due to the correspondence between depth and density in case of elliptically symmetric distributions. They found explicit formula for the $\theta_i(D(\boldsymbol{x}, \widehat{P}_i))$, when the halspace depth is used. Later Dutta and Ghosh (2012) found similar formula also for the projection depth. The classifiers which utilize these relations have the following form:

$$\begin{aligned} c(\boldsymbol{x}) &= \arg\max_i k_i \rho_i(\gamma_i(HD(\boldsymbol{x}, \widehat{P}_i)))/\gamma_i(HD(\boldsymbol{x}, \widehat{P}_i))^{d-1}, \\ c(\boldsymbol{x}) &= \arg\max_i k_i^* \rho_i^*(PD(\boldsymbol{x}, \widehat{P}_i)) \cdot PD(\boldsymbol{x}, \widehat{P}_i)^{d-3}/(1 - PD(\boldsymbol{x}, \widehat{P}_i))^{d-1}, \end{aligned}$$

  where the first classifier uses halfspace depth HD while the second classifier uses projection depth PD, k_i (k_i^* respectively) are unknown constants estimated by minimizing misclassification rate, $\gamma_i(HD(\boldsymbol{x}, \widehat{P}_i))$ denotes Mahalanobis distance whose relation to the halfspace depth is known, and finally ρ_i (ρ_i^* respectively) are unknown functions that need to be estimated by kernel density estimation technique. Here it is useful to note that the density that need to be estimated is always one-dimensional.

- The previously mentioned classifiers are optimal if the considered distributions are l_2-symmetric (after standardization). However, their authors have shown that for any l_p-symmetric distribution with $p \neq 2$ the density cannot be a function of halfspace depth. Dutta and Ghosh (2016) suggested to use L_p depth ($D(\boldsymbol{x}, P) = 1/(1 + \left\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu})\right\|_p$), where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are location and scale parameters of the distribution P) to overcome this problem and obtain classifier optimal for a broader class of l_p-symmetric distributions. They used relationship between depth and density in this case to estimate the density from the depth by one-dimensional kernel density estimation. The choice of p had to be discussed. The authors proposed restricted maximal likelihood estimate and discussed further how the restriction should be made.

- Another interesting classifier was proposed by Dutta and Ghosh (2015). They first transform the original data to the DD-space using spatial or localized spatial depth. Subsequently they use these depths as explanatory variables in multinomial additive logistic regression model (a spacial case of generalized additive models, called GAMs) where the response variable indicates group membership (class labels). Degree of localization determined by a single parameter h of locality is object of interest. The authors suggest "multiscale approach" in which the final classifier is based on weighted average of posterior probabilities estimated with different values of the parameter h. The weights are based on estimated misclassification rates. As mentioned already in section 2, this classifier was tested for high-dimensional data and demonstrated its good properties in this settings.

- The *DD-plot classifier* proposed by Li *et al.* (2012) works also in two steps. The first is transformation of the data into DD-space (depth versus depth). Visualization of the DD-space is by means of so called DD-plot. This step is followed by separating the transformed data points by a curve from a given family, for example by a polynom. The separation is done in a way which minimizes number of errors when classifying points from the training set. The classifier thus has the following form:

$$c(\boldsymbol{x}) = \arg\max_i r(D(\boldsymbol{x}, \widehat{P}_i)),$$

  where r is a function from a given class of functions (e.g. polynoms). In the simplest case the points might be separated by the straight line going through origin (which represents points with zero empirical depth to each group, so called outsiders). The only parameter that need to be estimated here (by minimizing average error rate on training set) is the slope of the line. In a special case we obtain maximal depth classifier which corresponds to the 45 degree line.

  The idea to separate classes by straight line going from the origin in the DD-space was proposed already by Jin and Cui (2010). Unfortunately, their paper remains unnoticed so far. They suggested (and subsequently used) new notion of depth, which is defined as $D_c(\boldsymbol{x}) = D(c\boldsymbol{x} + (1-c)\boldsymbol{\mu}, P)$, where D is some well known depth function and $c > 0$ is a tuning parameter. Apart from the slope of the line this parameter need to be estimated. For a given tuning parameter c the slope b is estimated to make probability of misclassification rate in the first group smaller than a given small $\alpha \in (0,1)$, i.e. $P_1(D(\boldsymbol{x}, \widehat{P}_1) > bD(\boldsymbol{x}, \widehat{P}_2)) \geq 1 - \alpha$. The parameter c is then tuned to make the misclassification rate in the second group as small as possible. Different notions of depth can be used to further minimize this misclassification rate.

- The *DD-alpha* procedure proposed by Lange *et al.* (2014b) belongs to the best currently available depth-based classifiers. Instead of the pair $[D(\boldsymbol{x}, \widehat{P}_1), D(\boldsymbol{x}, \widehat{P}_2)]$, it works with a higher-dimensional vector of "features". Such a vector might be like this:

$$\boldsymbol{z} := [D(\boldsymbol{x}, \widehat{P}_1), D(\boldsymbol{x}, \widehat{P}_2), D(\boldsymbol{x}, \widehat{P}_1) \cdot D(\boldsymbol{x}, \widehat{P}_2), D(\boldsymbol{x}, \widehat{P}_1)^2, D(\boldsymbol{x}, \widehat{P}_2)^2].$$

  Note that the the particular features always have the form of product $D(\boldsymbol{x}, \widehat{P}_1)^{k_1} \cdot D(\boldsymbol{x}, \widehat{P}_2)^{k_2}$, where in the pravious example $k_1 + k_2 \leq 2$, but in general higher powers can be used as well. Linear separation (by a hyperplane) in the feature space leads in general to nonlinear separation in the DD-space. Lange et al. proposed a heuristic for finding proper parameters which specify separating hyperplane given by the equation $aD(\boldsymbol{x}, \widehat{P}_1) + bD(\boldsymbol{x}, \widehat{P}_2) + cD(\boldsymbol{x}, \widehat{P}_1)D(\boldsymbol{x}, \widehat{P}_2) + dD(\boldsymbol{x}, \widehat{P}_1)^2 + eD(\boldsymbol{x}, \widehat{P}_2)^2 = 0$. The procedure was successfully tested on many real datasets leading usually to low misclassification rates, see Mozharovskyi, Mosler, and Lange (2015). The procedure which is very fast and robust was implemented in the R-package `ddalpha`, see Pokotylo *et al.* (2016). Several depth functions are implemented in this package. The choice of proper depth function was discussed in Lange, Mosler, and Mozharovskyi (2014a). The research on DD-alpha procedure is nicely summarized in Mozharovskyi (2015).

- The *k-depth-nearest neighbour classifier* highlighted by Vencalek (2013) is quite simple – it uses the well known k-nearest neighbour procedure in the DD-space. The question is which metric should be used to measure distances between distinct points.

- A classifier which uses a specific local depth was suggested by Pokotylo and Mosler (2016). Instead of term "DD-plot" used by Li *et al.* (2012) they use term "pot-pot plot", where pot-pot is a shortcut for potential versus potential. The potential of a class in a given point is defined as a kernel density estimate in this point multiplied by the class's prior probability. Proper choice of kernel can make the potential be affine invariant and thus it can be viewed as a local depth. Any of previously mentioned classifiers can be applied in the pot-pot plot.

- The transformation used in the first step of advanced depth-based classifier does not have to be directly transformation to the DD-space. Hubert, Rousseeuw, and Segaert (2015) suggested transformation to the distance-distance space. The considered distance is so called bagdistance which is based on (halfspace) depth. The bagdistance of a point $\boldsymbol{x} \in \mathbb{R}^d$ w.r.t. distribution P is given by the ratio of the Euclidean distance of $\boldsymbol{x}$ to the multivariate median $\boldsymbol{\theta}$ and the Euclidean distance of a point $c(\boldsymbol{x})$ to $\boldsymbol{\theta}$, where $c(\boldsymbol{x})$ is defined as the intersection of the boundary of so call "bag" and a ray from the $\boldsymbol{\theta}$ through $\boldsymbol{x}$. The bag is the smallest depth region (for definition see section 4.3) with at least 50% probability mass. The ratio is not defined in $\boldsymbol{\theta}$, where the distance is defined additionally to be zero. In the distance-distance space they proposed to use k-nearest neighbour method, however any other classifier mentioned in this section can be used as well.

- There are two depth-based advanced classifiers that does not follow the scheme presented in Figure 1. Although they are different they both can be called k-depth nearest neighbours (k-depthNN) method.

  The first one was proposed by Vencalek (2011). The classifier was based on assumption that there exists a function which relates depth and density function. This assumption holds true for elliptically symmetric distributions. In other cases localization of depth might be used to bring the depth closer to density. Vencalek defined distributional neighbourhood of a given point $\boldsymbol{x} \in \mathbb{R}^d$ as a set of points whose depth w.r.t. a given distribution P does not differ from $D(\boldsymbol{x}, P)$ of more that a given $\epsilon > 0$. In analogy to classical kNN he considered points in neigbouhoods of $\boldsymbol{x}$ that contain a fixed number (k) of points. Since there are two distributions, there are also two such distributional neighbourhoods. Vencalek suggested to classify a new observation to the group with the smallest (in the sense of Lebesgue measure) distributional neighbourhood. The main practical problem of the procedure is computation (estimation) of Lebesgue measures of the distributional neighbourhoods.

  Maybe the most promising depth-based classifier is that by Paindaveine and Van Bever (2015). They used the idea of symmetrization: any given point $\boldsymbol{x} \in \mathbb{R}^d$ is in the centre of the equal mixture of original distribution $P^{\boldsymbol{X}}$ and its reflection $P^{2\boldsymbol{x}-\boldsymbol{X}}$ (see also section 4.3) and thus it is the deepest point w.r.t. this mixture. The points from the original training set can be ordered according to their depth w.r.t. this symmetrized distribution. Then k points with the highest depth form closest distributional neighbours of the point $\boldsymbol{x}$. The point is assigned to the group with the highest number of representatives among the k nearest neighbours. The procedure was shown to be "nonparametrically consistent" (which is just a little bit weaker property than universal consistency). The main practical disadvantage might be viewed in large number of computation needed to classify a single point.

5. Conclusion

The data depth provides basis for nonparametric inference on multidimensional data. Possibility of its use in classification has been investigated for more than 15 years. Although one can expect broad applicability of the nonparametric depth-based classifiers the optimality of many proposed classifiers can be guaranteed only under some restrictive assumptions. Global depth functions and global classification techniques applied on the DD-space lead to good results only if the considered distributions have some global properties like unimodality. In more general settings localization is needed – one can use local depth functions or local classifiers used in the DD-space. Simple classifiers like maximal depth classifier have been already overcome. The DD-plot classifier by Li *et al.* (2012), DD-alpha classifier by Lange *et al.* (2014b) (both implemented in the R-package ddalpha) and classifiers based on symmetrization proposed in Paindaveine and Van Bever (2015) and Paindaveine and Van Bever

(2013) belong to the currently top depth-based classifiers. Interesting new ways how to use depth in the context of classification like the one proposed in Gilad-Bachrach and Burges (2013) continue to appear.

Acknowledgement

The research was supported by the grant of Czech Science Foundation GA15-06991S.

References

Billor N, Abebe A, Turkmen A, Nudurupati SV (2008). "Classification Based on Depth Transvariations." *Journal of classification*, **25**(2), 249–260.

Christmann A (2006). "Regression Depth and Support Vector Machine." *DIMACS series in discrete mathematics and theoretical computer science*, **72**, 71–86.

Christmann A, Fischer P, Joachims T (2002). "Comparison between Various Regression Depth Methods and the Support Vector Machine to Approximate the Minimum Number of Missclassifications." *Computational Statistics*, **17**(2), 273–287.

Christmann A, Rousseeuw PJ (2001). "Measuring Overlap in Binary Regression." *Computational Statistics & Data Analysis*, **37**(1), 65–75.

Cui X, Lin L, Yang G (2008). "An Extended Projection Data Depth and Its Applications to Discrimination." *Communications in Statistics – Theory and Methods*, **37**(14), 2276–2290.

Dutta S, Ghosh AK (2012). "On Robust Classification Using Projection Depth." *Annals of the Institute of Statistical Mathematics*, **64**(3), 657–676.

Dutta S, Ghosh AK (2015). "Multi-scale Classification Using Localized Spatial Depth." *arXiv preprint arXiv:1504.03804*.

Dutta S, Ghosh AK (2016). "On Affine Invariant *L_p* Depth Classifiers based on an Adaptive Choice of *p*." *arXiv preprint arXiv:1611.05668*.

Ghosh AK, Chaudhuri P (2005a). "On Data Depth and Distribution-free Discriminant Analysis Using Separating Surfaces." *Bernoulli*, pp. 1–27.

Ghosh AK, Chaudhuri P (2005b). "On Maximum Depth and Related Classifiers." *Scandinavian Journal of Statistics*, **32**(2), 327–350.

Gilad-Bachrach R, Burges CJ (2013). "Classifier Selection Using the Predicate Depth." *Journal of Machine Learning Research*, **14**(1), 3591–3618.

Hall P, Titterington D, Xue JH (2009). "Median-Based Classifiers for High-Dimensional Data." *Journal of the American Statistical Association*, **104**(488), 1597–1608.

Hartikainen A, Oja H (2006). "On Some Parametric, Nonparametric and Semiparametric Discrimination Rules." *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, **72**, 61–70.

Hlubinka D, Vencalek O (2013). "Depth-based Classification for Distributions with Nonconvex Support." *Journal of Probability and Statistics*, **2013**.

Hubert M, Rousseeuw PJ, Segaert P (2015). "Multivariate and Functional Classification Using Depth and Distance." *arXiv preprint arXiv:1504.01128*.

Hubert M, Van der Veeken S (2010a). "Fast and Robust Classifiers Adjusted for Skewness." In *Lechevallier, Y., Saporta, G. (Eds.), Proceedings of COMPSTAT 2010*, pp. 1135–1142. Physica-Verlag.

Hubert M, Van der Veeken S (2010b). "Robust Classification for Skewed Data." *Advances in Data Analysis and Classification*, **4**(4), 239–254.

Jin J, Cui H (2010). "Discriminant Analysis Based on Statistical Depth." *Journal of Systems Science and Complexity*, **23**(2), 362–371.

Jörnsten R (2004). "Clustering and Classification Based on the L_1 Data Depth." *Journal of Multivariate Analysis*, **90**(1), 67–89.

Kosiorowski D (2008). "Robust Classification and Clustering Based on the Projection Depth Function." In *Brito, P. (Eds.), Proceedings of COMPSTAT 2008*, volume II, pp. 209–216. Physica-Verlag.

Kosiorowski D, Bocian M, Bujak A (2014). "A Combination of Localdepth and svm Algorithms in Automatic Identification and Prediction of a Market State." In *Knowledge-Economy-Society: Contemporary Tools of Organizational Resources Management*, pp. 225–237. Cracow University of Economics.

Kumar S, Kumar A, Kumar A, Hegde RM (2015). "Hybrid Maximum Depth-kNN Method for Real Time Node Tracking Using Multi-sensor Data." In *2015 IEEE International Conference on Communications (ICC)*, pp. 6652–6657. IEEE.

Lange T, Mosler K, Mozharovskyi P (2014a). "DDα-classification of Asymmetric and Fat-tailed Data." In *Data Analysis, Machine Learning and Knowledge Discovery*, pp. 71–78. Springer.

Lange T, Mosler K, Mozharovskyi P (2014b). "Fast Nonparametric Classification Based on Data Depth." *Statistical Papers*, **55**(1), 49–69.

Li J, Cuesta-Albertos JA, Liu RY (2012). "DD-classifier: Nonparametric Classification Procedure Based on DD-plot." *Journal of the American Statistical Association*, **107**(498), 737–753.

Liu RY (1990). "On a Notion of Data Depth Based on Random Simplices." *The Annals of Statistics*, **18**(1), 405–414.

Mosler K, Hoberg R (2006). "Data Analysis and Classification with the Zonoid Depth." *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, **72**, 49–59.

Mozharovskyi P (2015). *Contributions to Depth-based Classification and Computation of the Tukey Depth*. Ph.D. thesis.

Mozharovskyi P, Mosler K, Lange T (2015). "Classifying Real-world Data with the {DD}\ Alpha-procedure." *Advances in Data Analysis and Classification*, **9**(3), 287–314.

Paindaveine D, Van Bever G (2013). "From Depth to Local Depth: A Focus on Centrality." *Journal of the American Statistical Association*, **108**(503), 1105–1119.

Paindaveine D, Van Bever G (2015). "Nonparametrically Consistent Depth-based Classifiers." *Bernoulli*, **21**(1), 62–82.

Pokotylo O, Mosler K (2016). "Classification with the Pot-pot Plot." *arXiv preprint arXiv:1608.02861*.

Pokotylo O, Mozharovskyi P, Dyckerhoff R (2016). "Depth and Depth-based Classification with R-package ddalpha." *arXiv preprint arXiv:1608.04109*.

Vencalek O (2011). *Weighted Data Depth and Depth Based Discrimination*. Ph.D. thesis, Doctoral Thesis. Charles University. Prague. URL http://artax.karlin.mff.cuni.cz/~venco2am/DataDepth.html.

Vencalek O (2013). "New Depth-based Modification of the k-nearest Neighbour Method." *SOP Transactions on Statistics and Analysis*, **1**(2), 131–138.

Affiliation:

Ondřej Vencálek
Department of Mathematical Analysis and Applications of Mathematics
Faculty of Science, Palacký University in Olomouc
17. listopadu 12, 771 46 Olomouc, Czech Republic
E-mail: ondrej.vencalek@upol.cz

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 46
April 2017

http://www.ajs.or.at/
http://www.osg.or.at/

Submitted: 2016-11-15
Accepted: 2017-02-02

ISSN 1026–597X

Contents