# Austrian Journal of Statistics

AUSTRIAN STATISTICAL SOCIETY

**Volume 43, Number 3-4, 2014**

Special Issue

**10th International Conference**
**COMPUTER DATA ANALYSIS & MODELING 2013, Minsk, Belarus**

# 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

The author(s) retain any copyright on the submitted material. The contributors grant the journal the right to publish, distribute, index, archive and publicly display the article (and the abstract) in printed, electronic or any other form.

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 43, Number 3-4, 2014

Editor: Matthias TEMPL

http://www.ajs.or.at

Published by the AUSTRIAN STATISTICAL SOCIETY

http://www.osg.or.at

## Österreichische Zeitschrift für Statistik

### Jahrgang 43, Heft 3-4, 2014

ÖSTERREICHISCHE STATISTISCHE GESELLSCHAFT

## Impressum


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

# Preface

The Tenth International Conference "Computer Data Analysis and Modeling: Complex Stochastic Data and Systems" (CDAM'2013) organized in Minsk by the Belarusian State University and Vienna University of Technology on September 10-14, 2013, was devoted to the topical problems in computer data analysis and modeling. There were 99 presentations by more than 130 participants from 19 countries.

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

This Special Issue contains 11 papers of the extended versions of the most significant presentations selected by the Organizing Committee after a refereeing process.

List of referees

W. Charemza, Leicester
G. Dzemyda, Vilnius
K. Ducinskas, Klaipeda
P. Filzmoser, Vienna
Yu. Kharin, Minsk
V. Malugin, Minsk
G. Medvedev, Minsk
E. Stoimenova, Sofia
E. Zhuk, Minsk

Peter Filzmoser, Yuriy Kharin
(Guest Editors)

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

Peter Filzmoser
Department of Statistics and
and Probability Theory
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
*June 2014, Volume 43/3-4, 167–179.*
*http://www.ajs.or.at/*

# Sensitivity Analysis for the Decomposition of Mixed Partitioned Multivariate Models into Two Seemingly Unrelated Submodels

**Eva Fišerová**
Palacký University Olomouc

**Lubomír Kubáček**
Palacký University Olomouc

**Abstract**

The paper is focused on the decomposition of mixed partitioned multivariate models into two seemingly unrelated submodels in order to obtain more efficient estimators. The multiresponses are independently normally distributed with the same covariance matrix. The partitioned multivariate model is considered either with, or without an intercept. The elimination transformation of the intercept that preserves the BLUEs of parameter matrices and the MINQUE of the variance components in multivariate models with and without an intercept is stated. Procedures on testing the decomposition of the partitioned model are presented. The properties of plug-in test statistics as functions of variance components are investigated by sensitivity analysis and insensitivity regions for the significance level are proposed. The insensitivity region is a safe region in the parameter space of the variance components where the approximation of the variance components can be used without any essential deterioration of the significance level of the plug-in test statistic. The behavior of plug-in test statistics and insensitivity regions is studied by simulations.

*Keywords*: multivariate model, decomposition, plug-in statistic, joint test, variance components, insensitivity region.

## 1. Introduction

A multivariate approach to modeling (see, e.g., Anderson 1958; Kshirsagar 1972; Kubáček 2008; Seber 2004) has several advantages in comparison with a series of univariate models. Specifically, multivariate models respect the association between outcomes, and thus, in general, procedures are more efficient. Further, they can evaluate the joint influence of predictors on all outcomes and avoid the issue of multiple testing. On the other hand, there are situations when the multivariate model can be decomposed to a series of simpler models, univariate or multivariate, depending on the issue. Moreover, from a practical point of view, collecting data is usually easier in decomposed models.

The paper deals with a special case of a decomposition of a partitioned multivariate model with independent multiresponses with the same covariance matrix into two seemingly unrelated multivariate submodels (Zellner 1962) in order to obtain more efficient estimators. Namely, the multiresponse variables in the model are partitioned into two sets $\underline{\mathbf{Y}}^1$ and $\underline{\mathbf{Y}}^2$. Similarly,

the set of predictors is partitioned into two sets $\mathbf{X}_1$ and $\mathbf{X}_2$. As an example, let us consider the nutrigenomic study in the mouse. The response variable might be expressions of chosen genes ($\underline{\boldsymbol{Y}}_{i\cdot}^1$) and concentrations of hepatic fatty acids ($\underline{\boldsymbol{Y}}_{i\cdot}^2$) measured on subjects. The predictors might be genotype ($\mathbf{X}_1$) and type of diet ($\mathbf{X}_2$). The problem is to decide, roughly speaking, if it is possible to explain separately expressions of genes by genotype and hepatic fatty acids concentrations by diet or not. Fišerová and Kubáček (2012) proposed tests for the verification of the significance of a model decomposition under normality of random errors in the case when the covariance matrix is known or completely unknown. Further, Fišerová and Kubáček (2013) shown that the proposed tests may be used in models without an intercept, as well as in models with an intercept. These tasks are summarized in Section 2. Here, we will focused on the situation when the covariance matrix includes unknown variance components.

If variance components can be estimated via the maximum likelihood method, the technique of Kenward and Roger (1996) is useful for testing hypotheses about the decomposition of the model. Nevertheless, the maximum likelihood approach is suitable for replicated models or models with large number of observations. In the paper we consider a model without replications when the minimum norm quadratic unbiased estimators (MINQUE) based on Rao's procedure (Rao and Kleffe 1988) are used instead. This approach is valid even for models with small number of observations. The MINQU estimators are derived in Section 3. Estimated values of variance components can be plugged into the test statistic for a known covariance matrix. The investigation of statistical properties of a plug-in test statistic is rather difficult and therefore we will study the quality of a plug-in test statistic as a function of the variance components by sensitivity analysis. The sensitivity approach provides the so-called insensitivity regions (Kubáček 1996) in the space of variance components where the approximation of variance components do not cause any essential damage of the chosen statistical characteristic. Namely, we propose the insensitivity region for the significance level (Kubáček 2007b) as it is shown in Section 4. If we know that the true value of the variance components is with sufficiently high probability within the insensitivity region for the significance level, then the significance level of the plug-in test statistic does not exceed the chosen tolerable value. The sensitivity approach is investigated mostly in univariate models, e.g., Kubáček (1996); Fišerová and Kubáček (2003, 2004, 2006); Kubáček and Fišerová (2003); Lešanská (2002a,b). Some results for multivariate models are presented in Kubáček (2006, 2007a,b) and Fišerová and Kubáček (2009). The behavior of plug-in statistics and insensitivity regions is studied by simulations in Section 5.

## 2. Tests of the decomposition in case of a known covariance matrix

Let us consider the multivariate model in a partitioned form

$$\underset{(n\times p_1)}{\left(\underline{\mathbf{Y}}^1\right.}\,,\,\underset{(n\times p_2)}{\left.\underline{\mathbf{Y}}^2\right)} = \underset{(n\times k_1)}{\left(\mathbf{X}_1\right.}\,,\,\underset{(n\times k_2)}{\left.\mathbf{X}_2\right)} \begin{pmatrix} \underset{(k_1\times p_1)}{\mathbf{B}_{11}} & \underset{(k_1\times p_2)}{\mathbf{B}_{12}} \\ \underset{(k_2\times p_1)}{\mathbf{B}_{21}} & \underset{(k_2\times p_2)}{\mathbf{B}_{22}} \end{pmatrix} + \underset{(n\times p_1)}{\left(\underline{\boldsymbol{\varepsilon}}_1\right.}\,,\,\underset{(n\times p_2)}{\left.\underline{\boldsymbol{\varepsilon}}_2\right)}. \tag{1}$$

Here $\underline{\mathbf{Y}} = (\underline{\mathbf{Y}}^1, \underline{\mathbf{Y}}^2)$ is a random matrix (observation matrix), $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ is a known design matrix, $\mathbf{B}_{11}$, $\mathbf{B}_{12}$, $\mathbf{B}_{21}$ and $\mathbf{B}_{22}$ are matrices of unknown parameters and $\underline{\boldsymbol{\varepsilon}} = (\underline{\boldsymbol{\varepsilon}}_1, \underline{\boldsymbol{\varepsilon}}_2)$ is a random error matrix. We will assume that the matrix $\mathbf{X}$ is of full column rank, the multiresponses are independent with the same positive definite covariance matrix $\boldsymbol{\Sigma}$ and the random errors are normally distributed. The covariance matrix $\boldsymbol{\Sigma}$ of the multiresponse $\underline{\boldsymbol{Y}}_{i\cdot} = (Y_{i1}, Y_{i2}, \ldots, Y_{ip})'$ is partitioned in the same way, i.e.,

$$\operatorname{var}\begin{pmatrix} \underline{\boldsymbol{Y}}_{i\cdot}^1 \\ \underline{\boldsymbol{Y}}_{i\cdot}^2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{11}, & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21}, & \boldsymbol{\Sigma}_{22} \end{pmatrix}, \ \mathrm{i} = 1, 2, \ldots, \mathrm{n}. \tag{2}$$

Further, let us consider a system of two seemingly unrelated (Zellner 1962) multivariate submodels

$$\underset{(n\times p_1)}{\underline{\mathbf{Y}}^1} = \underset{(n\times k_1)}{\mathbf{X}_1}\,\underset{(k_1\times p_1)}{\mathbf{B}_1} + \underset{(n\times p_1)}{\underline{\boldsymbol{\varepsilon}}_1}, \qquad \underset{(n\times p_2)}{\underline{\mathbf{Y}}^2} = \underset{(n\times k_2)}{\mathbf{X}_2}\,\underset{(k_2\times p_2)}{\mathbf{B}_2} + \underset{(n\times p_2)}{\underline{\boldsymbol{\varepsilon}}_2} \tag{3}$$

with the covariance matrix $\boldsymbol{\Sigma}$ of the multiresponse $\underline{\boldsymbol{Y}}_{i\cdot}$ in the form (2). Note that models in (3) are seemingly unrelated because there is a link between them described by $\mathrm{cov}(\underline{\boldsymbol{Y}}^1_{\mathrm{i}\cdot}, \underline{\boldsymbol{Y}}^2_{\mathrm{i}\cdot}) = \boldsymbol{\Sigma}_{12}$. If $\boldsymbol{\Sigma}_{12} = \mathbf{0}$, the models in (3) are independent. The problem is to decide which of the models (1) and (3) should be chosen for modeling in order to obtain more efficient estimators.

The issue with a decomposition of model (1) into (3) leads to testing the hypothesis that "the system of two seemingly unrelated multivariate submodels (3) is a true model", i.e., to test $\mathbf{B}_{12} = \mathbf{0}$ and $\mathbf{B}_{21} = \mathbf{0}$ simultaneously. If the covariance matrix $\boldsymbol{\Sigma}$ is known, Fišerová and Kubáček (2012) proposed the test statistics

$$T_{21} = \mathrm{Tr}\big[(\underline{\mathbf{Y}}^1)'\mathbf{M}_{\mathrm{X}_1}\mathbf{X}_2(\mathbf{X}_2'\mathbf{M}_{\mathrm{X}_1}\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_{\mathrm{X}_1}\underline{\mathbf{Y}}^1\boldsymbol{\Sigma}_{11}^{-1}\big] \sim \chi^2_{\mathrm{p_1k_2}} \text{ under } \mathbf{B}_{21} = \mathbf{0}, \tag{4}$$

$$T_{12} = \mathrm{Tr}\big[(\underline{\mathbf{Y}}^2)'\mathbf{M}_{\mathrm{X}_2}\mathbf{X}_1(\mathbf{X}_1'\mathbf{M}_{\mathrm{X}_2}\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{M}_{\mathrm{X}_2}\underline{\mathbf{Y}}^2\boldsymbol{\Sigma}_{22}^{-1}\big] \sim \chi^2_{\mathrm{p_2k_1}} \text{ under } \mathbf{B}_{12} = \mathbf{0}. \tag{5}$$

The symbol $\mathrm{Tr}(\boldsymbol{\Sigma})$ denotes trace of the matrix $\boldsymbol{\Sigma}$ and $\mathbf{M}_{X_i} = \mathbf{I}_n - \mathbf{X}_i(\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'$, $i = 1, 2$. To test the hypotheses $\mathbf{B}_{21} = \mathbf{0}$ and $\mathbf{B}_{12} = \mathbf{0}$ simultaneously, one can use, e.g., the Bonferroni correction in order to preserve the type I error rate $\alpha$. More precisely, if $T_{21} \leq \chi^2_{p_1k_2}(1-\alpha/2)$ and $T_{12} \leq \chi^2_{p_2k_1}(1-\alpha/2)$, where $\chi^2_{p_1k_2}(1-\alpha/2)$ denotes the $(1-\alpha/2)$-quantile of a $\chi^2_{p_1k_2}$ distribution, neither of the hypotheses $\mathbf{B}_{21} = \mathbf{0}$, $\mathbf{B}_{12} = \mathbf{0}$ can be rejected on the significance level $\alpha$.

Note that the decomposition of model (1) leads to two seemingly unrelated submodels. If the decomposition is significant, the prediction of $\underline{\mathbf{Y}}^1$ conditional on $\mathbf{X}_1$ is not improved also by regressing on $\mathbf{X}_2$. However the predictors $\mathbf{X}_2$ are necessary for the calculation of the prediction of $\underline{\mathbf{Y}}^1$. Analogous conclusions hold for the prediction of $\underline{\mathbf{Y}}^2$.

Until now we have considered only the model without an intercept. A partitioned form of the model with the intercept can be written as

$$\left(\underset{(n\times p_1)}{\underline{\mathbf{Y}}^1}, \underset{(n\times p_2)}{\underline{\mathbf{Y}}^2}\right) = \left(\underset{(n\times 1)}{\mathbf{1}}, \underset{(n\times k_1)}{\mathbf{X}_1}, \underset{(n\times k_2)}{\mathbf{X}_2}\right)\begin{pmatrix} \underset{(1\times p_1)}{\mathbf{b}_1}, & \underset{(1\times p_2)}{\mathbf{b}_2} \\ \underset{(k_1\times p_1)}{\mathbf{B}_{11}}, & \underset{(k_1\times p_2)}{\mathbf{B}_{12}} \\ \underset{(k_2\times p_1)}{\mathbf{B}_{21}}, & \underset{(k_2\times p_2)}{\mathbf{B}_{22}} \end{pmatrix} + \left(\underset{(n\times p_1)}{\underline{\boldsymbol{\varepsilon}}_1}, \underset{(n\times p_2)}{\underline{\boldsymbol{\varepsilon}}_2}\right), \tag{6}$$

where $\mathbf{1}$ is a vector of ones. We will assume that the design matrix $(\mathbf{1}, \mathbf{X}_1, \mathbf{X}_2)$ is of full column rank, and therefore all regression parameters are unbiasedly estimable. If the model includes also an intercept, then the question is, where the intercept should go in the decomposed model, in $\mathbf{X}_1$, in $\mathbf{X}_2$, or in both $\mathbf{X}_1$ and $\mathbf{X}_2$. Naturally, all cases are possible and results depend on particular tasks. To avoid this situation, Fišerová and Kubáček (2013) proposed the transformation for an elimination of the intercept that leads to the identical BLUEs of parameter matrices $\mathbf{B}_{11}$, $\mathbf{B}_{12}$, $\mathbf{B}_{21}$ and $\mathbf{B}_{22}$ in model (6) with the intercept and model without the intercept in the form

$$\left(\mathbf{M}_1\underline{\mathbf{Y}}^1, \mathbf{M}_1\underline{\mathbf{Y}}^2\right) = \left(\mathbf{M}_1\mathbf{X}_1, \mathbf{M}_1\mathbf{X}_2\right)\begin{pmatrix} \mathbf{B}_{11}, & \mathbf{B}_{12} \\ \mathbf{B}_{21}, & \mathbf{B}_{22} \end{pmatrix} + \left(\mathbf{M}_1\underline{\boldsymbol{\varepsilon}}_1, \mathbf{M}_1\underline{\boldsymbol{\varepsilon}}_2\right). \tag{7}$$

Here $\mathbf{M}_1 = \mathbf{I} - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'$. Therefore testing the decomposition of the partitioned model with the intercept can be done similarly as for the model without the intercept. The process is the following. First we transform model (6) to (7). Next we test hypotheses $\mathbf{B}_{12} = \mathbf{0}$ and $\mathbf{B}_{21} = \mathbf{0}$ simultaneously via test statistics $T_{12}$, $T_{21}$ using the substitution

$$\underline{\mathbf{Y}}^j \to \mathbf{M}_1\underline{\mathbf{Y}}^j,\ \mathbf{X}_j \to \mathbf{M}_1\mathbf{X}_j,\ \boldsymbol{\Sigma} \to \mathbf{M}_1\boldsymbol{\Sigma}\mathbf{M}_1,\ j = 1, 2.$$

Obviously, test statistics $T_{12}$ and $T_{21}$ have the same degrees of freedom in the case of the model without the intercept since the assumptions on full column rank of the design matrices imply that the ranks of the transformed design matrices $\mathbf{M}_1\mathbf{X}_j$ are equal to $k_j$, $j = 1, 2$, as well.

## 3. Tests of the decomposition in case of a covariance matrix with unknown variance components

Now we will consider the covariance matrix $\boldsymbol{\Sigma}$ of the structure $\boldsymbol{\Sigma} = \sum_{i=1}^{s} \vartheta_i \mathbf{V}_i$, where $\mathbf{V}_1, \ldots, \mathbf{V}_s$ are known $(p_1 + p_2) \times (p_1 + p_2)$ symmetric and positive semidefinite matrices and $\vartheta_i > 0, i = 1, \ldots, s$, are unknown parameters (variance components). Denote $\boldsymbol{\Sigma}_0 = \sum_{i=1}^{s} \vartheta_{0,i} \mathbf{V}_i$, where $\boldsymbol{\vartheta}_0 = (\vartheta_{0,1}, \ldots, \vartheta_{0,s})'$ is an approximate value of the vector parameter $\boldsymbol{\vartheta}$. The $\boldsymbol{\vartheta}_0$-locally minimum norm quadratic unbiased estimator ($\boldsymbol{\vartheta}_0$-LMINQUE) of $\boldsymbol{\vartheta}$ based on Rao's procedure (Rao and Kleffe 1988) is stated in the following lemma.

**Lemma 1** The $\boldsymbol{\vartheta}_0$-locally MINQUE of $\boldsymbol{\vartheta}$ in the model (1) is

$$\widehat{\boldsymbol{\vartheta}} = \frac{1}{n - k_1 - k_2} \mathbf{S}^{-1}_{\Sigma_0^{-1}} \widehat{\boldsymbol{\gamma}},$$

where the $i$th component of the vector $\widehat{\boldsymbol{\gamma}} = (\widehat{\gamma}_1, \ldots, \widehat{\gamma}_s)'$ is

$$\widehat{\gamma}_i = \mathrm{Tr}\left[ \begin{pmatrix} (\underline{\mathbf{Y}}^1)' \\ (\underline{\mathbf{Y}}^2)' \end{pmatrix} \mathbf{M}_{(X_1, X_2)} (\underline{\mathbf{Y}}^1, \underline{\mathbf{Y}}^2) \boldsymbol{\Sigma}_0^{-1} \mathbf{V}_i \boldsymbol{\Sigma}_0^{-1} \right]$$

and the $(i, j)$th element of $(s \times s)$ matrix $\mathbf{S}_{\Sigma_0^{-1}}$ is

$$\left\{ \mathbf{S}_{\Sigma_0^{-1}} \right\}_{i,j} = \mathrm{Tr}(\boldsymbol{\Sigma}_0^{-1} \mathbf{V}_i \boldsymbol{\Sigma}_0^{-1} \mathbf{V}_j).$$

Under normality, the covariance matrix of the estimator $\widehat{\boldsymbol{\vartheta}}$ at the point $\boldsymbol{\vartheta}_0$ is

$$\mathrm{var}_{\vartheta_0}(\widehat{\boldsymbol{\vartheta}}) = \frac{2}{n - k_1 - k_2} \mathbf{S}^{-1}_{\Sigma_0^{-1}}. \tag{8}$$

**Proof.** For simplicity, the proof proceeds for the univariate form of model (1) which can be expressed as

$$\mathrm{vec}(\underline{\mathbf{Y}}^1, \underline{\mathbf{Y}}^2) \sim N_{n(p_1+p_2)} \left\{ \left[ \mathbf{I}_{p_1+p_2} \otimes (\mathbf{X}_1, \mathbf{X}_2) \right] \mathrm{vec} \begin{pmatrix} \mathbf{B}_{11}, & \mathbf{B}_{12} \\ \mathbf{B}_{21}, & \mathbf{B}_{22} \end{pmatrix}, \sum_{i=1}^{s} \vartheta_i \mathbf{V}_i \otimes \mathbf{I}_n \right\}. \tag{9}$$

Here, the symbol $\mathrm{vec}(\underline{\mathbf{Y}}^1)$ denotes the column vector composed of the columns of $\underline{\mathbf{Y}}^1$. The notation $\otimes$ means the Kronecker multiplication of matrices (Rao and Mitra 1988). Then, according to Rao and Kleffe (1988), the $\boldsymbol{\vartheta}_0$-locally MINQUE of the vector parameter $\boldsymbol{\vartheta}$ in model (9) is

$$\widehat{\boldsymbol{\vartheta}} = \mathbf{S}^{-1}_{\mathbf{D}} \widehat{\boldsymbol{\gamma}}, \quad \mathbf{D} = [\mathbf{M}_{[I \otimes (X_1, X_2)]} (\boldsymbol{\Sigma}_0 \otimes \mathbf{I}) \mathbf{M}_{[I \otimes (X_1, X_2)]}]^{+},$$

where the $i$th component of the vector $\widehat{\boldsymbol{\gamma}}$ is given as

$$\widehat{\gamma}_i = \left[ \mathrm{vec}(\underline{\mathbf{Y}}^1, \underline{\mathbf{Y}}^2) \right]' \mathbf{D} (\mathbf{V}_i \otimes \mathbf{I}) \mathbf{D} \mathrm{vec}(\underline{\mathbf{Y}}^1, \underline{\mathbf{Y}}^2), \ i = 1, \ldots, s.$$

Now it is sufficient to take into account the equality $\mathbf{D} = \boldsymbol{\Sigma}_0^{-1} \otimes \mathbf{M}_{(X_1, X_2)}$, which is substituted into the previous formulas, and to simplify each of the expressions. □

The estimator of the variance components and its covariance matrix depends on approximate values $\boldsymbol{\vartheta}_0$. To eliminate this dependency, it is necessary to use an iterative procedure. The calculated estimated values of the variance components are used in the next iteration as

approximate ones. The iterative procedure is very robust. It usually stops after two iterations for any initial value of the variance components even for distributions different from the Gaussian (Bognárová, Kubáček, and Volaufová 1996).

Note that the iterated MINQUE is practically the same as the maximum likelihood estimator in the case of Gaussian distribution. Moreover, the MINQUE procedure can be used even for negative variance components and symmetric matrices $\mathbf{V}_i$ for errors with normal distribution. The formulas for the estimators are the same as under the assumption of positive variance components and p.s.d. matrices $\mathbf{V}_i$.

If model (1) can be decomposed into (3), the variance components can be estimated on the basis of either $\underline{\mathbf{Y}}^1$ or $\underline{\mathbf{Y}}^2$ in model (3). The explicit formulas for $\boldsymbol{\vartheta}_0$-LMINQUE in the submodels follows directly from Lemma 1. Particularly, using the notation $\boldsymbol{\Sigma}_{jj,0}$ for an approximate value of matrix $\boldsymbol{\Sigma}_{jj}$, $j = 1, 2$, the expressions for $\boldsymbol{\vartheta}_0$-LMINQUE of $\boldsymbol{\vartheta}$ in submodels (3) are equal to

$$\widehat{\boldsymbol{\vartheta}} = \frac{1}{n-k_j}\mathbf{S}^{-1}_{\Sigma^{-1}_{jj,0}}\widehat{\boldsymbol{\gamma}}, \quad \mathrm{var}_{\vartheta_0}(\widehat{\boldsymbol{\vartheta}}) = \frac{2}{n-k_j}\mathbf{S}^{-1}_{\Sigma^{-1}_{jj,0}}, \quad j = 1, 2.$$

The $i$th component of the vector $\widehat{\boldsymbol{\gamma}}$ and the $(p, q)$th element of the matrix $\mathbf{S}_{\Sigma^{-1}_{jj,0}}$ are given as

$$\widehat{\gamma}_i = \mathrm{Tr}\left[(\underline{\mathbf{Y}}^j)'\mathbf{M}_{X_j}\underline{\mathbf{Y}}^j\boldsymbol{\Sigma}^{-1}_{jj,0}\mathbf{V}_i\boldsymbol{\Sigma}^{-1}_{jj,0}\right], \quad \left\{\mathbf{S}_{\Sigma^{-1}_{jj,0}}\right\}_{p,q} = \mathrm{Tr}(\boldsymbol{\Sigma}^{-1}_{jj,0}\mathbf{V}_p\boldsymbol{\Sigma}^{-1}_{jj,0}\mathbf{V}_q).$$

If the mixed partitioned model includes also an intercept, the elimination transformation (7) can be used. This transformation preserves not only the BLUEs of the regression parameters matrices, but also the estimates of the variance components, as it will be shown in the following theorem.

**Theorem 1.** The $\boldsymbol{\vartheta}_0$-locally MINQUE of $\boldsymbol{\vartheta}$ in models (6) and (7) are the same.

**Proof.** For the sake of simplicity, the proof proceeds for the univariate form of model (6). Let us denote

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \mathrm{vec}(\underline{\boldsymbol{\varepsilon}}_1) \\ \mathrm{vec}(\underline{\boldsymbol{\varepsilon}}_2) \end{pmatrix}, \mathbf{Y} = \begin{pmatrix} \mathrm{vec}(\underline{\mathbf{Y}}^1) \\ \mathrm{vec}(\underline{\mathbf{Y}}^2) \end{pmatrix}, \quad \boldsymbol{\Sigma}\otimes\mathbf{I} = \begin{pmatrix} \boldsymbol{\Sigma}_{11}\otimes\mathbf{I}, & \boldsymbol{\Sigma}_{12}\otimes\mathbf{I} \\ \boldsymbol{\Sigma}_{21}\otimes\mathbf{I}, & \boldsymbol{\Sigma}_{22}\otimes\mathbf{I} \end{pmatrix},$$

$$\mathbf{A}_1 = \begin{pmatrix} \mathbf{I}\otimes\mathbf{1}, & \mathbf{0} \\ \mathbf{0}, & \mathbf{I}\otimes\mathbf{1} \end{pmatrix}, \quad \mathbf{A}_2 = \begin{pmatrix} \mathbf{I}\otimes\mathbf{X}_1, & \mathbf{0}, & \mathbf{I}\otimes\mathbf{X}_2, & \mathbf{0} \\ \mathbf{0}, & \mathbf{I}\otimes\mathbf{X}_2, & \mathbf{0}, & \mathbf{I}\otimes\mathbf{X}_1 \end{pmatrix}$$

$$\boldsymbol{\beta}_1 = (\mathbf{b}_1, \mathbf{b}_2)', \quad \boldsymbol{\beta}_2 = \left(\mathrm{vec}(\mathbf{B}_{11})', \mathrm{vec}(\mathbf{B}_{22})', \mathrm{vec}(\mathbf{B}_{21})', \mathrm{vec}(\mathbf{B}_{12})'\right)'.$$

Then the model (6) can be rewritten as

$$\mathbf{Y} = (\mathbf{A}_1, \mathbf{A}_2)\begin{pmatrix}\boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2\end{pmatrix} + \boldsymbol{\varepsilon}. \tag{10}$$

By Rao and Kleffe (1988), the $\boldsymbol{\vartheta}_0$-locally MINQUE of $\boldsymbol{\vartheta}$ in model (10) is given as

$$\widehat{\boldsymbol{\vartheta}} = \mathbf{S}^{-1}_G\widehat{\boldsymbol{\gamma}}, \quad \mathbf{G} = [\mathbf{M}_{(A_1,A_2)}(\boldsymbol{\Sigma}_0\otimes\mathbf{I})\mathbf{M}_{(A_1,A_2)}]^+,$$

where $\widehat{\gamma}_i = \mathbf{Y}'\mathbf{G}(\mathbf{V}_i\otimes\mathbf{I})\mathbf{G}\mathbf{Y}$, $i = 1, \ldots, s$. Further, using the relationships

$$\mathbf{P}_{A_1} = \mathbf{A}_1(\mathbf{A}_1'\mathbf{A}_1)^{-1}\mathbf{A}_1', \mathbf{P}_{(A_1,A_2)} = \mathbf{P}_{A_1} + \mathbf{P}_{M_{A_1}A_2},$$
$$\mathbf{P}_{(A_1,A_2)}\mathbf{P}_{A_1} = \mathbf{P}_{A_1}, \mathbf{M}_{(A_1,A_2)} = \mathbf{M}_{A_1} - \mathbf{P}_{M_{A_1}A_2},$$

we obtain

$$\mathbf{M}_{(A_1,A_2)}\mathbf{M}_{A_1} = \mathbf{M}_{(A_1,A_2)}, \quad \mathbf{M}_{(A_1,A_2)}\mathbf{M}_{M_{A_1}A_2} = \mathbf{M}_{A_1}\mathbf{M}_{M_{A_1}A_2}.$$

Thus, the matrix $\mathbf{G}$ can be rewritten as

$$\mathbf{G} = [\mathbf{M}_{(A_1,A_2)}(\boldsymbol{\Sigma}_0\otimes\mathbf{I})\mathbf{M}_{(A_1,A_2)}]^+ = [\mathbf{M}_{M_{A_1}A_2}(\mathbf{M}_{A_1}(\boldsymbol{\Sigma}_0\otimes\mathbf{I})\mathbf{M}_{A_1})\mathbf{M}_{M_{A_1}A_2}]^+ = \widetilde{\mathbf{G}}.$$

Therefore, the equalities

$$\mathbf{G}(\mathbf{V}_i \otimes \mathbf{I})\mathbf{G} = \widetilde{\mathbf{G}}\mathbf{M}_{A_1}(\mathbf{V}_i \otimes \mathbf{I})\mathbf{M}_{A_1}\widetilde{\mathbf{G}},$$

$$\widehat{\gamma}_i = \mathbf{Y}'\mathbf{G}(\mathbf{V}_i \otimes \mathbf{I})\mathbf{G}\mathbf{Y} = \mathbf{Y}'\mathbf{M}_{A_1}\widetilde{\mathbf{G}}\mathbf{M}_{A_1}(\mathbf{V}_i \otimes \mathbf{I})\mathbf{M}_{A_1}\widetilde{\mathbf{G}}\mathbf{M}_{A_1}\mathbf{Y} = \widetilde{\gamma}_i,$$

also holds. Summarizing the above results we obtain $\widehat{\boldsymbol{\vartheta}} = \mathbf{S}_G^{-1}\widehat{\boldsymbol{\gamma}} = \mathbf{S}_{\widetilde{G}}^{-1}\widetilde{\boldsymbol{\gamma}}$, i.e., the $\boldsymbol{\vartheta}_0$-LMINQUE of $\boldsymbol{\vartheta}$ in model (10) and in the model $\mathbf{M}_{A_1}\mathbf{Y} = \mathbf{M}_{A_1}\mathbf{A}_2\boldsymbol{\beta}_2 + \mathbf{M}_{A_1}\boldsymbol{\varepsilon}$ are the same. Using the relationship $\mathbf{M}_{A_1} = \mathrm{diag}\{\mathbf{I} \otimes \mathbf{M}_1, \mathbf{I} \otimes \mathbf{M}_1\}$ (diagonal matrix), the last model can be rewritten into multivariate form (7), and thus the proof is finished. □

If the variance components are estimated, the hypothesis about the decomposition of model (1) can be tested by the plug-in statistics $T_{21}$ and $T_{12}$, when the matrices $\widehat{\boldsymbol{\Sigma}}_{jj} = \sum_{i=1}^{s} \widehat{\vartheta}_i \mathbf{V}_{i,(jj)}$, where $\mathbf{V}_{i,(jj)}$ is the corresponding part of $\mathbf{V}_i$, $j = 1, 2$, are plugged into formulas (4). Testing the decomposition of the model with the intercept (6) can be done by the plug-in statistics $T_{21}$ and $T_{12}$ for the transformed model (7), since the transformation affects neither the BLUEs of the regression parameter matrices nor the MINQUE of the variance components.

Obviously, the substitution of the true values of the variance components by their estimated values influences the optimum quality of the estimators $\widehat{\mathbf{B}}_{21}$, $\widehat{\mathbf{B}}_{12}$ and, consequently, the significance level and the power of the test. The investigation of statistical properties of the plug-in test statistics $T_{21}$ and $T_{12}$ is rather difficult and therefore we will study the quality of the plug-in test statistic as a function of the variance components by sensitivity analysis as it is shown in the next section.

## 4. Sensitivity analysis for the significance level

The main idea of the sensitivity approach (Kubáček 1996) is to consider the plug-in statistic as a function of the variance components and to find a safe region in the parameter space of the variance components where the approximation of the variance components does not cause any essential damage of the significance level of the plug-in test statistic (Kubáček 2007b). The plug-in test statistic can have a higher significance level. Let $\varepsilon > 0$ be the maximum admissible increase of the significance level. The goal is to find a region in the parameter space of the variance components such that shifts $\delta\boldsymbol{\vartheta}$ around the true value $\boldsymbol{\vartheta}^*$ within this region cause the significance level of the plug-in test statistic $T_{21}$ to be not greater than $\alpha/2 + \varepsilon/2$. (We consider the significance level $\alpha/2+\varepsilon/2$ since the Bonferroni correction for multiple tests on $\mathbf{B}_{21} = \mathbf{0}$ and $\mathbf{B}_{12} = \mathbf{0}$ is used.) Such a region is called an *insensitivity region for the significance level* and will be denoted by $\mathcal{N}_{\varepsilon,T_{21}}$. More precisely, $\mathcal{N}_{\varepsilon,T_{21}}$ is a neighborhood of the vector $\boldsymbol{\vartheta}^*$ with the property

$$\boldsymbol{\vartheta} \in \mathcal{N}_{\varepsilon,T_{21}} \Rightarrow P\Big\{T_{21}(\boldsymbol{\vartheta}) \le \chi^2_{k_2 p_1}(1-\alpha/2)\Big\} \ge 1 - \alpha/2 - \varepsilon/2.$$

The derivation of the insensitivity region for the significance level is based on an approximation of the plug-in test statistic $T_{21}$ by $T_{12}(\boldsymbol{\vartheta}) = T_{21}(\boldsymbol{\vartheta}^*) + \delta T_{21}$. The variable $\delta T_{21} = \delta\boldsymbol{\vartheta}' \frac{\partial T_{21}}{\partial \boldsymbol{\vartheta}}$ characterizes the change of the statistic $T_{21}(\boldsymbol{\vartheta}^*)$ caused by the shift $\delta\boldsymbol{\vartheta}$ around $\boldsymbol{\vartheta}^*$. Obviously, the significance level of $T_{21}$ increases with increasing $\delta T_{21}$ and vice versa. Hence the problem is to find the upper limit for $\delta T_{21}$ so that the significance level increased by a maximum tolerated value. Using the Chebyshev inequality it holds that

$$P\left\{|\delta T_{21} - E(\delta T_{21})| \ge t\sqrt{\mathrm{var}(\delta T_{21})}\right\} \le \frac{1}{t^2},\ t > 0. \tag{11}$$

The inequality (11) together with the probability statement for the tolerated significance level

$$P\left\{T_{21} + \delta T_{21} \ge \chi^2_{k_2 p_1}(1-\alpha/2)\right\} \le \alpha/2 + \varepsilon/2,$$

implies that for a sufficiently large $t > 0$ such that

$$P\{\delta T_{21} < E(\delta T_{21}) + t\sqrt{\mathrm{var}(\delta T_{21})}\} \approx 1, \tag{12}$$

the inequality $E(\delta T_{21}) + t\sqrt{\mathrm{var}(\delta T_{21})} \leq \delta_{\varepsilon,T_{21}}$, where

$$\delta_{\varepsilon,T_{21}} = \chi^2_{k_2 p_1}(1-\alpha/2) - \chi^2_{k_2 p_1}(1-\alpha/2-\varepsilon/2), \tag{13}$$

is a sufficient condition for the upper limit for $\delta T_{21}$. The explicit form of the insensitivity region $\mathcal{N}_{\varepsilon,T_{21}}$ is stated in Theorem 2.

**Theorem 2** The insensitivity region $\mathcal{N}_{\varepsilon,T_{21}}$ for the significance level of the statistic $T_{21}$ is

$$\mathcal{N}_{\varepsilon,T_{21}} = \left\{\boldsymbol{\vartheta}^* + \delta\boldsymbol{\vartheta} : \left(\delta\boldsymbol{\vartheta} - \delta_{\varepsilon,T_{21}}\mathbf{A}^{-1}\mathbf{a}\right)' \mathbf{A} \left(\delta\boldsymbol{\vartheta} - \delta_{\varepsilon,T_{21}}\mathbf{A}^{-1}\mathbf{a}\right) \leq \delta^2_{\varepsilon,T_{21}}\left(1 + \mathbf{a}'\mathbf{A}^{-1}\mathbf{a}\right)\right\},$$

where $\delta_{\varepsilon,T_{21}}$ is given by (13),

$$\mathbf{a} = k_2\left[\mathrm{Tr}(\boldsymbol{\Sigma}_{11}^{-1}\mathbf{V}_{1,(11)}), \ldots, \mathrm{Tr}(\boldsymbol{\Sigma}_{11}^{-1}\mathbf{V}_{s,(11)})\right]', \quad \mathbf{A} = 2t^2 k_2 \mathbf{S}_{\Sigma_{11}^{-1}} - \mathbf{a}\mathbf{a}',$$

and $t > 0$ is a sufficiently large number such that the probability statement (12) holds.

**Proof.** It is necessary to determine the mean value and the variance of variable $\delta T_{21}$ which characterizes a change of the statistic $T_{21}(\boldsymbol{\vartheta}^*)$ caused by the shift $\delta\boldsymbol{\vartheta}$. Let

$$\xi_i = \frac{\partial T_{21}(\boldsymbol{\vartheta})}{\partial \vartheta_i}\Big|_{\vartheta=\vartheta^*} = -\mathrm{Tr}\left[\underline{\mathbf{Y}}_1'\mathbf{P}_{M_{X_1}X_2}\underline{\mathbf{Y}}_1\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{\vartheta}^*)\mathbf{V}_{i,(11)}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{\vartheta}^*)\right],\ i = 1, \ldots, s.$$

The mean value of the variable $\xi_i$ is

$$\begin{aligned} E_{\vartheta^*}(\xi_i) &= -\mathrm{Tr}\left(\left\{\left[\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{\vartheta}^*)\mathbf{V}_{i,(11)}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{\vartheta}^*)\right] \otimes \mathbf{P}_{M_{X_1}X_2}\right\}\left[\boldsymbol{\Sigma}_{11}(\boldsymbol{\vartheta}^*) \otimes \mathbf{I}\right]\right) \\ &= -\mathrm{Tr}(\mathbf{P}_{M_{X_1}X_2})\mathrm{Tr}\left[\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{\vartheta}^*)\mathbf{V}_{i,(11)}\right] = -k_2\mathrm{Tr}\left[\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{\vartheta}^*)\mathbf{V}_{i,(11)}\right]. \end{aligned}$$

Further we calculate the covariance between the variables $\xi_i$ and $\xi_j$:

$$\begin{aligned} \mathrm{cov}_{\vartheta^*}(\xi_i, \xi_j) &= 2\mathrm{Tr}\left(\left\{\left[\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{\vartheta}^*)\mathbf{V}_{i,(11)}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{\vartheta}^*)\right] \otimes \mathbf{P}_{M_{X_1}X_2}\right\}\left[\boldsymbol{\Sigma}_{11}(\boldsymbol{\vartheta}^*) \otimes \mathbf{I}\right]\right. \\ &\quad \left.\times \left\{\left[\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{\vartheta}^*)\mathbf{V}_{j,(11)}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{\vartheta}^*)\right] \otimes \mathbf{P}_{M_{X_1}X_2}\right\}\left[\boldsymbol{\Sigma}_{11}(\boldsymbol{\vartheta}^*) \otimes \mathbf{I}\right]\right) \\ &= 2\mathrm{Tr}(\mathbf{P}_{M_{X_1}X_2})\mathrm{Tr}\left[\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{\vartheta}^*)\mathbf{V}_{i,(11)}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{\vartheta}^*)\mathbf{V}_{j,(11)}\right] = 2k_2\left\{\mathbf{S}_{\Sigma_{11}^{-1}(\vartheta^*)}\right\}_{i,j}. \end{aligned}$$

Now we are able to determine the upper limit for $\delta T_{21} = \delta\boldsymbol{\vartheta}'\boldsymbol{\xi}$. Let $t > 0$ be a sufficiently large number such that (12) holds, i.e., with probability sufficiently near to one it is true that

$$\delta\boldsymbol{\vartheta}'\boldsymbol{\xi} < E_{\vartheta^*}(\boldsymbol{\xi}')\delta\boldsymbol{\vartheta} + t\sqrt{\delta\boldsymbol{\vartheta}'\mathrm{var}_{\vartheta^*}(\boldsymbol{\xi})\delta\boldsymbol{\vartheta}} \leq \delta_{\varepsilon,T_{21}}.$$

Substituting the mean value and the covariance matrix of the vector $\boldsymbol{\xi}$, and by a simple calculation we obtain the inequality

$$\delta\boldsymbol{\vartheta}'(2t^2k_2\mathbf{S}_{\Sigma_{11}^{-1}(\vartheta^*)} - \mathbf{a}\mathbf{a}')\delta\boldsymbol{\vartheta} - 2\mathbf{a}'\delta\boldsymbol{\vartheta} \leq \delta^2_{\varepsilon,T_{21}},$$

which is equivalent with

$$\begin{aligned} &\left[\delta\boldsymbol{\vartheta} - \delta_{\varepsilon,T_{21}}(2t^2k_2\mathbf{S}_{\Sigma_{11}^{-1}(\vartheta^*)} - \mathbf{a}\mathbf{a}')^{-1}\mathbf{a}\right]'(2t^2k_2\mathbf{S}_{\Sigma_{11}^{-1}(\vartheta^*)} - \mathbf{a}\mathbf{a}') \\ &\quad \times \left[\delta\boldsymbol{\vartheta} - \delta_{\varepsilon,T_{21}}(2t^2k_2\mathbf{S}_{\Sigma_{11}^{-1}(\vartheta^*)} - \mathbf{a}\mathbf{a}')^{-1}\mathbf{a}\right] \leq \delta^2_{\varepsilon,T_{21}} + \delta^2_{\varepsilon,T_{21}}\mathbf{a}'(2t^2k_2\mathbf{S}_{\Sigma_{11}^{-1}(\vartheta^*)} - \mathbf{a}\mathbf{a}')^{-1}\mathbf{a}, \end{aligned}$$

thereby the statement is proved. □

Analogous considerations can be made for the plug-in test statistic $T_{12}$. In this case the explicit formula for the insensitivity region for the significance level results in

$$\mathcal{N}_{\varepsilon,T_{12}} = \left\{ \boldsymbol{\vartheta}^* + \delta\boldsymbol{\vartheta} : \left(\delta\boldsymbol{\vartheta} - \delta_{\varepsilon,T_{12}}\mathbf{B}^{-1}\mathbf{b}\right)' \mathbf{B} \left(\delta\boldsymbol{\vartheta} - \delta_{\varepsilon,T_{12}}\mathbf{B}^{-1}\mathbf{b}\right) \leq \delta^2_{\varepsilon,T_{12}} \left(1 + \mathbf{b}'\mathbf{B}^{-1}\mathbf{b}\right) \right\},$$

where

$$\delta_{\varepsilon,T_{12}} = \chi^2_{k_1p_2}(1-\alpha/2) - \chi^2_{k_1p_2}(1-\alpha/2-\varepsilon/2),$$
$$\mathbf{b} = k_1\left[\mathrm{Tr}(\boldsymbol{\Sigma}_{22}^{-1}\mathbf{V}_{1,(22)}), \ldots, \mathrm{Tr}(\boldsymbol{\Sigma}_{22}^{-1}\mathbf{V}_{p,(22)})\right]', \quad \mathbf{B} = 2t^2k_1\mathbf{S}_{\Sigma_{22}^{-1}} - \mathbf{b}\mathbf{b}'.$$

The size of the insensitivity region depends on the parameters $\varepsilon$ and $t$ chosen by the user. The parameter $\varepsilon$ is related to the user's opinion that $\varepsilon$ causes a tolerable increase of the significance level. The larger $\varepsilon$, the larger the insensitivity region, but also a higher significance level follows. The parameter $t$ corresponds to the approximation of the plug-in test statistic, namely with the upper limit for the variable $\delta T_{21}$ that describes the change of the statistic $T_{21}(\boldsymbol{\vartheta}^*)$ caused by the shift $\delta\boldsymbol{\vartheta}$. For $t = 5$, from the Chebyshev inequality (11) it follows that at least 96% of the data values of $\delta T_{21}$ must be within 5 standard deviations of the mean or, equivalently, no more than 4% of the data values can be more than 5 standard deviations away from the mean. If $\delta T_{21}$ is approximately normally distributed, at least 99.7% of the data values of $\delta T_{21}$ must be within 3 standard deviations of the mean. Hence it is reasonable to choose the parameter $t$ in the interval $\langle 3, 5\rangle$. The smaller $t$, the larger the insensitivity region but also cases a higher tail probability. The procedure for the optimal choice of the parameter $t$ that maximizes the size of the insensitivity region is derived in Lešanská (2002a).

Both insensitivity regions $\mathcal{N}_{\varepsilon,T_{12}}$ and $\mathcal{N}_{\varepsilon,T_{21}}$ are suitable for a justification of the utilization of plug-in joint tests $T_{12}$, $T_{21}$ for a decomposition of model (1) into two seemingly unrelated submodels (3). The process is as follows. First, we determine estimates of the variance components. Then we compute the insensitivity regions $\mathcal{N}_{\varepsilon,T_{12}}$ and $\mathcal{N}_{\varepsilon,T_{21}}$ for the estimated values of the variance components and chosen values $\varepsilon$ and $t$. Finally, we set the confidence domain for the variance components for a sufficiently high confidence level and check whether this confidence domain is embedded into the insensitivity regions. If this confidence domain is included into both insensitivity regions, plug-in joint tests are admissible and, moreover, the significance level of plug-in joint tests does not exceed the value of $\alpha+\varepsilon$. If the confidence domain is not embedded into both insensitivity regions, the experiment requires better design, other measurement devices, or more observations to be sure that the approximation of the variance components by their estimates do not cause an increase in significance level by more than a tolerable $\varepsilon$. The criterion is very demanding, in some cases the confidence domain is not embedded into the insensitivity regions, however, the estimated values of the variance components lie in the insensitivity regions what implies that the increase of the significance level is almost a tolerable one (see Section 5).

The determination of an exact confidence domain for the variance components is difficult since the distribution of the estimator $\widehat{\boldsymbol{\vartheta}}$ is unknown even for a normally distributed vector $\mathrm{vec}(\mathbf{Y})$. Some approximation can be derived using the Bonferroni inequality and the Chebyshev inequality which imply that

$$P\left\{\forall i = 1, \ldots, s: \ |\widehat{\vartheta}_i - E(\widehat{\vartheta}_i)| \leq \sqrt{\frac{s}{\alpha}}\sqrt{\mathrm{var}(\widehat{\vartheta}_i)}\right\} \geq 1-\alpha.$$

Hence at least a $(1-\alpha)100\%$-confidence domain for the variance components is a set given by the Cartesian product

$$I_{1-\alpha}(\boldsymbol{\vartheta}) = \mathrm{X}_{i=1}^{s}\left\{u : |u - \widehat{\vartheta}_i| \leq \sqrt{\frac{s}{\alpha}}\sqrt{\mathrm{var}(\widehat{\vartheta}_i)}\right\}.$$

Recall that the covariance matrix of the variance components estimator is given by (8).

It can be easily proved, similarly as in Lešanská (2002b), that a $k^2$-multiple of $\boldsymbol{\vartheta}^*$ makes a homothetic change of the boundary of the insensitivity region for the significance level with the coefficient $k^2$ and the centre at the point 0.

## 5. Simulation study

By simulations we will study the behavior of the plug-in test statistics $T_{12}$, $T_{21}$ for the decomposition of model (1) and the insensitivity regions for the significance level. We will consider different choices of the covariance matrix, parameter matrices, number of observations and the true model (model (1) or the system of seemingly unrelated submodels (3)).

We considered $n = 40$ and $n = 400$ observations, a multiresponse $\underline{\mathbf{Y}}_{i.}^{j}$, $j = 1, 2$, with dimensions $p_1 = 3$ and $p_2 = 4$, and the number of regressors equal to $k_1 = 2$ and $k_2 = 2$. The parameter matrices were chosen as

$$\mathbf{B}_1 = \mathbf{B}_{11} = \begin{pmatrix} 3, & 2, & 2 \\ 2, & 3, & 3 \end{pmatrix}, \quad \mathbf{B}_2 = \mathbf{B}_{22} = \begin{pmatrix} 2, & 4, & 4, & 1.5 \\ 4, & 2, & 4, & 4 \end{pmatrix}, \tag{14}$$

$$\mathbf{B}_{12} = \begin{pmatrix} 1, & 3, & 1, & 15 \\ 2, & 7, & 8, & 3 \end{pmatrix}, \quad \mathbf{B}_{21} = \begin{pmatrix} 4, & 4, & 1 \\ 2, & 8, & 3 \end{pmatrix}. \tag{15}$$

The design matrices were considered in the form of $\mathbf{X} = \mathbf{1}_{10} \otimes \mathbf{T}$ and $\mathbf{X} = \mathbf{1}_{100} \otimes \mathbf{T}$, with the matrix $\mathbf{T}$

$$\mathbf{T} = \begin{pmatrix} 1, & 0, & 0, & 0 \\ 1, & 1, & 0, & 0 \\ 1, & 1, & 1, & 0 \\ 1, & 0, & 1, & 1 \end{pmatrix}.$$

The symbol $\mathbf{1}_{10}$ denotes the vector of 10 ones. Similar designs of experiments were used in the simulation study in Fišerová and Kubáček (2012).

The observation matrices $\underline{\mathbf{Y}}^1$ and $\underline{\mathbf{Y}}^2$ were generated in a natural way, a normally distributed error term was added to the true mean. The multiresponses were considered to be independent with the same covariance matrix $\boldsymbol{\Sigma}$ chosen in the following two forms: either $\mathbf{V}_1 = \mathrm{diag}\{1, 1, 1, 0, 0, 0, 0\}$ and $\mathbf{V}_2 = \mathrm{diag}\{0, 0, 0, 1, 1, 1, 1\}$ (the corresponding covariance matrix is denoted by $\boldsymbol{\Sigma}_1$), or $\mathbf{V}_1 = \mathrm{diag}\{1, 1, 0, 0, 0, 0, 1\}$ and $\mathbf{V}_2 = \mathrm{diag}\{0, 0, 1, 1, 1, 1, 0\}$ (covariance matrix $\boldsymbol{\Sigma}_2$). The variance components were chosen either $\vartheta_1 = 5$ and $\vartheta_2 = 3$, or $\vartheta_1 = 0.05$ and $\vartheta_2 = 0.03$.

Table 1: Empirical probabilities (in %) of rejecting the hypothesis "the true model is the system of two seemingly unrelated models (3)" on the significance level $\alpha$. Data are simulated from model (3).

| | | $\vartheta_1 = 5$ and $\vartheta_2 = 3$ | | | | $\vartheta_1 = 0.05$ and $\vartheta_2 = 0.03$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $n = 400$ | | $n = 40$ | | $n = 400$ | | $n = 40$ | |
| parameter matrices | $\alpha$ | $\boldsymbol{\Sigma}_1$ | $\boldsymbol{\Sigma}_2$ | $\boldsymbol{\Sigma}_1$ | $\boldsymbol{\Sigma}_2$ | $\boldsymbol{\Sigma}_1$ | $\boldsymbol{\Sigma}_2$ | $\boldsymbol{\Sigma}_1$ | $\boldsymbol{\Sigma}_2$ |
| (14) | 5 | 5.2 | 4.9 | 6.1 | 5.9 | 5.2 | 5.1 | 5.7 | 5.6 |
| (14) | 1 | 0.9 | 1.1 | 1.5 | 1.2 | 0.9 | 1.1 | 1.3 | 1.3 |
| 100*(14) | 5 | 5.1 | 5.1 | 5.6 | 5.6 | 4.6 | 5.0 | 6.2 | 5.6 |
| 100*(14) | 1 | 1.1 | 1.1 | 1.5 | 1.1 | 1.0 | 1.2 | 1.3 | 1.3 |

10 000 simulations were done for all cases. First, the data were simulated from the system of two seemingly unrelated submodels (3), i.e., for matrices $\mathbf{B}_1$, $\mathbf{B}_2$ given by (14) and for zero matrices $\mathbf{B}_{12}$, $\mathbf{B}_{21}$. The empirical probabilities of rejecting the hypothesis $\mathbf{B}_{12} = \mathbf{0}$ and $\mathbf{B}_{21} = \mathbf{0}$ simultaneously are presented in Table 1. We can see that the obtained empirical significance levels for plug-in test statistics $T_{12}$ and $T_{21}$ are essentially equal to the nominal

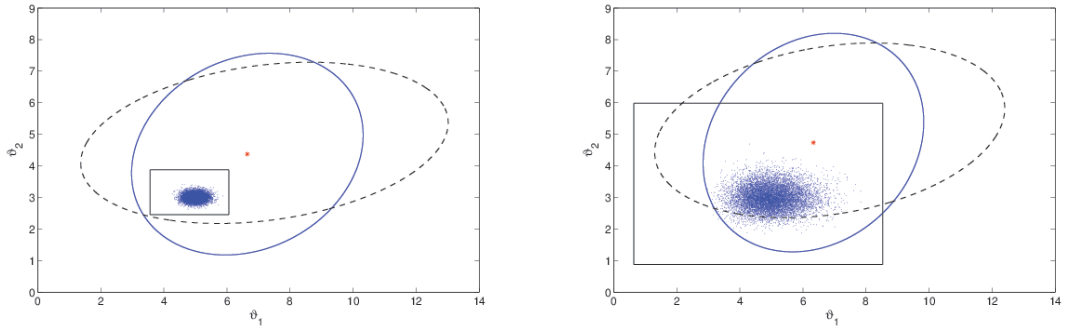

Figure 1: Insensitivity regions for the significance level $\alpha = 1\%$ together with 95%-confidence domain for the variance components for different sample sizes (left: $n = 400$; right: $n = 40$). $\mathcal{N}_{5,T_{21}}$ by solid line, $\mathcal{N}_{5,T_{12}}$ by dashed line. Data are simulated from model (3) for $\mathbf{\Sigma}_2$, $\vartheta_1 = 5$, $\vartheta_2 = 3$.

levels. Fišerová and Kubáček (2012) have shown in a simulation study that the joint test by $T_{12}$ and $T_{21}$ is conservative in case of a known covariance matrix, the obtained empirical significance level was equal to half of the nominal one. For data simulated from the model (1), i.e., for matrices $\mathbf{B}_{11}$, $\mathbf{B}_{22}$ and $\mathbf{B}_{12}$, $\mathbf{B}_{21}$ given by (14) and (15), respectively, the plug-in test statistics $T_{21}$ and $T_{12}$ rejected the decomposition of model (1) in all cases.

Finally we will investigate the insensitivity regions for the significance level. Let us assume a tolerable increase of the significance level $\alpha = 1\%$ or $\alpha = 5\%$ equal to $\varepsilon = 5\%$. It means, we are satisfied if the true type I error rate is $\alpha/2 + \varepsilon/2 = 5\%$ (3%) for a nominal significance level $\alpha = 5\%$ ($\alpha = 1\%$) for each of the plug-in test statistic $T_{21}$ and $T_{12}$. Further, we assume that the parameter $t$ equals 3, i.e., at least 89% of data values of $\delta T_{21}$ and $\delta T_{12}$ must be within 3 standard errors of the mean.

For $\alpha = 1\%$, the resulting insensitivity regions are displayed together with 95%-confidence domain for the variance components in Figure 1. The data were simulated from model (3) for the covariance matrix $\mathbf{\Sigma}_2$ and $\vartheta_1 = 5$, $\vartheta_2 = 3$. We can see that the insensitivity regions $\mathcal{N}_{5,T_{21}}$ and $\mathcal{N}_{5,T_{12}}$ are large enough. The statistic $T_{12}$ allows greater shifts in direction of $\vartheta_1$, and $T_{21}$ in direction of $\vartheta_2$. It means, the statistic $T_{12}$ is more sensitive to changes in $\vartheta_2$, and $T_{21}$ in $\vartheta_1$. Furthermore, we can notice, that the confidence domain for the variance components is embedded in both insensitivity regions for sample size $n = 400$ (left figure). For smaller sample size, the confidence domain increases more than the insensitivity ones (enlarged only slightly) and thus the confidence domain is not embedded into the insensitivity regions. Nevertheless, almost all variance components estimates lie within the insensitivity regions. Unfortunately, this is not generally true. The observed relative frequencies of the variance components estimates within the insensitivity regions are indicated in Table 2. The results are averages of a hundred times repeated 10 000 simulations. Obviously, the confidence region for the variance components is smaller for greater sample size and thus the relative frequencies are essentially 100%. However, for smaller sample size and significance level $\alpha = 5\%$, the relative frequencies are only 55-70%. Nevertheless, in this case the plug-in test is sufficiently good as it is shown in Table 1. Large differences between the relative frequencies are due to large differences in size of the insensitivity regions for significance levels $\alpha = 5\%$ and $\alpha = 1\%$. For example, in the case $\vartheta_1 = 5$ and $\vartheta_2 = 3$, the semiaxes of insensitivity region $\mathcal{N}_{5,T_{21}}$ are 1.75 and 1.28 for $\alpha = 5\%$, and 4.35 and 2.98 for $\alpha = 1\%$. This effect is related to the construction of the insensitivity regions, namely with the fact that a tolerable increase $\varepsilon$ of the significance level $\alpha$ leads to larger $\delta_{\varepsilon,T_{21}}$, $\delta_{\varepsilon,T_{12}}$ for smaller $\alpha$.

Note that the insensitivity regions for the significance level are shown as ellipses in Figure 1, although, in general, they should be open sets to the right hand corner. This relates to the fact that the p-values becomes less extreme with increasing variances. However, the construction

Table 2: Relative frequency (in %) of the variance components estimates within insensitivity regions for the significance level. Data are simulated from model (3) for $\mathbf{\Sigma}_2$.

| | | $\vartheta_1 = 5$ and $\vartheta_2 = 3$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | $n = 400$ | | | $n = 40$ | | |
| parameters | $\alpha$ | $\mathcal{N}_{5,T_{21}} \cap \mathcal{N}_{5,T_{12}}$ | $\mathcal{N}_{5,T_{21}}$ | $\mathcal{N}_{5,T_{12}}$ | $\mathcal{N}_{5,T_{21}} \cap \mathcal{N}_{5,T_{12}}$ | $\mathcal{N}_{5,T_{21}}$ | $\mathcal{N}_{5,T_{12}}$ |
| (14) | 5 | 98.9 | 99.8 | 99.6 | 56.7 | 66.7 | 67.6 |
| (14) | 1 | 100 | 100 | 100 | 96.7 | 98.3 | 97.9 |
| 100*(14) | 5 | 98.8 | 99.8 | 99.7 | 55.1 | 67.3 | 64.5 |
| 100*(14) | 1 | 100 | 100 | 100 | 96.9 | 98.5 | 98.0 |

| | | $\vartheta_1 = 0.05$ and $\vartheta_2 = 0.03$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | $n = 400$ | | | $n = 40$ | | |
| parameters | $\alpha$ | $\mathcal{N}_{5,T_{21}} \cap \mathcal{N}_{5,T_{12}}$ | $\mathcal{N}_{5,T_{21}}$ | $\mathcal{N}_{5,T_{12}}$ | $\mathcal{N}_{5,T_{21}} \cap \mathcal{N}_{5,T_{12}}$ | $\mathcal{N}_{5,T_{21}}$ | $\mathcal{N}_{5,T_{12}}$ |
| (14) | 5 | 98.68 | 99.35 | 99.22 | 54.79 | 65.61 | 65.54 |
| (14) | 1 | 100 | 100 | 100 | 97.04 | 98.47 | 98.17 |
| 100*(14) | 5 | 98.91 | 99.84 | 99.78 | 58.35 | 68.85 | 68.30 |
| 100*(14) | 1 | 100 | 100 | 100 | 95.94 | 98.22 | 97.21 |

of the insensitivity regions for estimators or the confidence level is different, and thus we proposed closed regions due to a uniform methodology and for easier handling.

The insensitivity regions for the significance level and the confidence domain for variance components result in intervals for the covariance matrix $\mathbf{\Sigma}_1$ since in this case the statistic $T_{21}$ is a function of the parameter $\vartheta_1$ only, and $T_{12}$ is a function of $\vartheta_2$. The obtained results are similar as in the case of the covariance matrix $\mathbf{\Sigma}_2$ and therefore they are omitted.

## 6. Conclusion

The proposed plug-in joint test seems to be a proper method for a decomposition of a mixed multivariate model (with independent responses with the same covariance matrix) into two seemingly unrelated submodels. The decomposition is advantageous at least from two viewpoints. The estimators of the regression parameters are more efficient, and data collection can be easier. The sensitivity approach is an appropriate technique for a justification that the estimated values of the variance components can be plugged in without any essential deterioration of the regression coefficients estimates and the inference. This is based on identifying safe regions in the space of the variance components where plug-in estimators cause only negligible changes of the optimum quality of estimators and test statistics. In particular, the proposed insensitivity region for the significance level guarantees that the true type I error rate does not exceed the chosen tolerable value.

The used methodology is general and suitable for more complex (mixed) models, e.g. models with restrictions on the regression parameters, singular models, and other statistical inference such as a confidence level or power of a test as well.

## Acknowledgements

The authors thankfully acknowledge the support by the Operational Program Education for Competitiveness - European Social Fund (project CZ.1.07/2.3.00/20.0170 of the Ministry of Education, Youth and Sports of the Czech Republic) and the Grant No. PrF-2013-013 of the Internal Grant Agency of the Palacký University in Olomouc.

## References

Anderson T (1958). *An Introduction to Multivariate Statistical Analysis.* J. Wiley, New York.

Bognárová M, Kubáček L, Volaufová J (1996). "Comparison of MINQUE and LMVQUIE by Simulation." *Acta Universitatis Palackianae Olomucensis. Facultas Rerum Naturalium. Mathematica*, **35**, 25–38.

Fišerová E, Kubáček L (2003). "Sensitivity Analysis in Singular Mixed Linear Models with Constraints." *Kybernetika*, **39**, 317–332.

Fišerová E, Kubáček L (2004). "Statistical Problems of Measurement in Triangle." *Folia Facultatis Scientiarium Naturalium Universitatis Masarykiana Brunesis, Mathematica*, **15**, 77–94.

Fišerová E, Kubáček L (2006). "Insensitivity Regions and Outliers in Mixed Models with Constraints." *Austrian Journal of Statistics*, **35**, 245–252.

Fišerová E, Kubáček L (2009). "Insensitivity Regions for Deformation Measurement on a Dam." *Environmetrics*, **20**, 776–789.

Fišerová E, Kubáček L (2012). "Decomposition of Multivariate Statistical Models." *Tatra Mountains Mathematical Publications*, **51**, 33–43.

Fišerová E, Kubáček L (2013). "Some remarks on decomposition of partitoned multivariate models into two seemingly unrelated submodels." In S Aivazian, P Filzmoser, Y Kharin (eds.), *Computer Data Analysis and Modeling: Theoretical and Applied Stochastics. Vol.1.*, pp. 36–41. Publishing center of BSU, Minsk.

Kenward M, Roger J (1996). "Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood." *Biometrics*, **53**, 983–997.

Kshirsagar A (1972). *Multivariate Analysis.* Marcel Dekker, Inc., New York.

Kubáček L (1996). "Linear Model with Inaccurate Variance Components." *Applications of Mathematics*, **41**, 433–445.

Kubáček L (2006). "Statistical methods in environmetrics." In L Dušek, J Hřebíček, J Žižka (eds.), *Proceedings of the 2nd International Summer School on Computational Biology*, pp. 68–83. Masaryk University, Brno.

Kubáček L (2007a). "Multivariate Regression Model with Constraints." *Mathematica Slovaca*, **57**, 271–296.

Kubáček L (2007b). "Test of Linear Hypothesis in Multivariate Models." *Kybernetika*, **43**, 463–470.

Kubáček L (2008). *Multivariate Statistical Models Revisited.* Palacký University, Olomouc.

Kubáček L, Fišerová E (2003). "Problems of Sensitiveness and Linearization in a Determination of Isobestic Points." *Mathematica Slovaca*, **53**, 407–426.

Lešanská E (2002a). "Optimization of the Size of Nonsensitiveness Regions." *Applications of Mathematics*, **47**, 9–23.

Lešanská E (2002b). "Insensitivity Regions and their Properties." *Journal of Electrical Engineering*, **53**, 68–71.

Rao C, Kleffe J (1988). *Estimation of Variance Components and Applications.* North–Holland, Amsterdam, New York, Oxfords, Tokyo.

Rao C, Mitra S (1988). *Generalized Inverse of Matrices and Its Applications.* J. Wiley, New York.

Seber G (2004). *Multivariate Observations*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Zellner A (1962). "An Efficient Method of Estimating Seemingly Unrelated Regression Equations and Tests for Aggregation Bias." *Journal of the American Statistical Association*, **57**, 348–368.

**Affiliation:**

Eva Fišerová
Mathematical Analysis and Applications of Mathematics
Faculty of Science
Palacký University
771 46 Olomouc, Czech Republic
E-mail: eva.fiserova@upol.cz
URL: http://mant.upol.cz/en/clen.asp?id=7

Lubomír Kubáček
Mathematical Analysis and Applications of Mathematics
Faculty of Science
Palacký University
771 46 Olomouc, Czech Republic
E-mail: lubomir.kubacek@upol.cz
URL: http://mant.upol.cz/en/clen.asp?id=3

Austrian Journal of Statistics
*June 2014, Volume 43/3-4, 181–193.*
*http://www.ajs.or.at/*

# On Outliers and Interventions in Count Time Series following GLMs

**Roland Fried** TU Dortmund
**Tobias Liboschik** TU Dortmund
**Hanan Elsaied** Suez Canal Univ.
**S. Kitromilidou** Univ. of Cyprus
**K. Fokianos** Univ. of Cyprus

**Abstract**

We discuss the analysis of count time series following generalised linear models in the presence of outliers and intervention effects. Different modifications of such models are formulated which allow to incorporate, detect and to a certain degree distinguish extraordinary events (interventions) of different types in count time series retrospectively. An outlook on extensions to the problem of robust parameter estimation, identification of the model orders by robust estimation of autocorrelations and partial autocorrelations, and online surveillance by sequential testing for outlyingness is provided.

*Keywords*: discrete data, model identification, robustness, (partial) autocorrelations, surveillance.

## 1. Introduction

Time series of counts are measured in various disciplines whenever a number of events is counted during certain time periods. Examples are the monthly number of car accidents in a region, the weekly number of new cases in epidemiology, the number of transactions at a stock market per minute in finance, or the number of photon arrivals per microsecond in a biological experiment. A natural modification of the popular autoregressive moving average (ARMA) models for continuous variables is based on the assumption that the observation $Y_t$ at time $t$ is generated by a generalised linear model (GLM) conditionally on the past, choosing an adequate distribution for count data like the Poisson and a link function $\eta(\cdot)$. This approach of time series following a GLM is pursued e.g. by Kedem and Fokianos (2002). Focusing on first order models, we consider time series $(Y_t : t \in \mathbb{N}_0)$ following a Poisson model

$$\begin{aligned} Y_t|\mathcal{F}_{t-1}^Y &\sim Pois(\lambda_t), \\ \eta(\lambda_t) &= \beta_0 + \beta_1\eta(Y_{t-1}+c) + \gamma_1\eta(\lambda_{t-1}),\ t \geq 1, \end{aligned} \tag{1}$$

where $\mathcal{F}_{t-1}^Y$ stands for the $\sigma$-algebra created by $\{Y_{t-1}, \ldots, Y_0, \lambda_0\}$, while $\beta_0, \beta_1, \gamma_1$ are unknown parameters, and $c$ is a known constant. Models employing other distributions like the negative binomial could be treated similarly.

The natural choice for $\eta$ is the logarithm, and this is the reason for adding the constant $c$ to $Y_{t-1}$ in the term $\eta(Y_{t-1}+c)$, since we need to avoid difficulties arising from observations which are equal to 0. Following Fokianos and Tjøstheim (2011), who develop ergodicity conditions

for a subclass of the arising log–linear models, we set $c = 1$. Another choice for $\eta$ which has received some attention is the identity, $\eta = id$, see e.g. Ferland, Latour and Oraichi (2006). In this case we can set $c$ to 0. For ergodicity conditions for this model class see Fokianos, Rahbek and Tjøstheim (2009).

We briefly discuss possible interpretations of models like those given in (1) in the context of epidemiology, with $Y_t$ denoting the number of new cases observed at time $t$. For a fixed population size, the conditional mean $\lambda_t$ measures the risk of a person to fall ill at time $t$ then. Our model assumes that all effects on $\lambda_t$ are linear after transformation to a suitable scale by $\eta$. The term $\eta(Y_{t-1} + c)$ in the second equation models the dependence of the transformed conditional mean $\eta(\lambda_t)$ and thus of the observation $Y_t$ on the previous value $Y_{t-1}$, with $\beta_1$ measuring the strength of this dependence. A large number of cases $Y_{t-1}$ at time $t-1$ can cause a large number of cases $Y_t$ at time $t$ because the risk of infection increases. The term $\eta(\lambda_{t-1})$ additionally describes that there can be periods of increased risk also because of certain weather conditions or expositions, for instance, and $\gamma_1$ measures the size of such dependencies.

Given a model as formulated in (1), a basic question is whether it properly describes all the observations of a given time series, or whether some observations have been influenced by extraordinary effects, which are called interventions in what follows. Outlier and intervention analysis for ARMA processes of continuous variables has been developed by Fox (1972), Box and Tiao (1975), Tsay (1986), Chang, Tiao and Chen (1988) and Chen and Liu (1993), among others. However, counts are positive and typically right-skewed, causing a need for especially designed models and procedures.

The remainder of the paper is organised as follows. Section 2 generalises the intervention models proposed by Fokianos and Fried (2010, 2012) for time series which are Poisson conditionally on the past, with $\eta$ being the identity and the log-link, respectively. Section 3 reviews first attempts of robust fitting of models with known link function and model orders. Section 4 reports a first study of model identification for the linear model applying the identity link, using robust estimators of the autocorrelations and partial autocorrelations. Section 5 provides an outlook to surveillance, that is online monitoring by sequential outlier detection.

## 2. Models for Intervention Analysis

A possibility to introduce an extraordinary effect on a time series $(Y_t)$ generated by (1) is the assumption that from a time point $\tau$ on the underlying conditional mean process is changed by adding terms $\omega\delta^{t-\tau}I(t \geq \tau)$ to $\eta(\lambda_t)$, so that instead of $(Y_t)$ we observe a contaminated process $(Z_t)$ generated from a model with contamination,

$$\begin{aligned} Z_t|\mathcal{F}_{t-1}^Z &\sim Pois(\lambda_t^c), \\ \eta(\lambda_t^c) &= \beta_0 + \beta_1\eta(Z_{t-1} + c) + \gamma_1\eta(\lambda_{t-1}^c) + \omega\delta^{t-\tau}I(t \geq \tau),\ t \geq 1. \end{aligned} \tag{2}$$

In obvious notation, $(\lambda_t^c)$ is the contaminated process of conditional means, which coincides with $(\lambda_t)$ until time $\tau - 1$ and then becomes affected, while $\mathcal{F}_{t-1}^Z$ denotes the $\sigma$-algebra representing the information on the past of the contaminated process and the initial values, analogous to $\mathcal{F}_{t-1}^Y$. The new parameter $\omega$ determines the size of the effect, $I(t \geq \tau)$ indicates whether $t \geq \tau$ or not, and $\delta \in [0, 1]$ determines whether the effect is concentrated on time $\tau$ (in case of $\delta = 0$), causing a spiky outlier, whether the whole level is shifted from time $\tau$ on ($\delta = 1$), or whether a geometrically decaying transient shift with rate $\delta \in (0, 1)$ occurs. Note that even in case of $\delta = 0$ the whole future of the process is affected by an intervention, since its effect enters the dynamics both via $Z_t$ and $\eta(\lambda_t^c)$, $t \geq \tau$. Continuing the explanations given above in the context of epidemiology, an intervention according to (2) can be interpreted as an internal change of the data generating process. For some reason, e.g. due to particular weather conditions or other expositions, the conditional mean of the process (the risk) changes in an unpredictable manner at time $\tau$, and this changes the observation for that time point,

and also the observations thereafter.

Liboschik et al. (2013) explore another intervention model in case of the identity link. In their approach, an intervention affects the observation at time $\tau$, but not the underlying conditional mean. This can be understood as an external change, as the contaminated observation $Z_\tau$ equals the sum of the uncontaminated value $Y_\tau$ plus a random number $C_\tau$, which arises because of extraordinary reasons and enters the dynamics of the process in the same way as $Y_\tau$, while the underlying risk $\lambda_\tau$ initially is not affected. An example might be people being infected due to external reasons, e.g. on a journey. The modified intervention model with a general link function $\eta$ reads

$$\begin{aligned} Z_t|\mathcal{F}_{t-1}^Z &\sim Pois(\lambda_t^c), \\ \eta(\lambda_t^c) &= \eta(\lambda_t) + \omega\delta^{t-\tau} I(t \geq \tau), \\ \eta(\lambda_t) &= \beta_0 + \beta_1 \eta(Z_{t-1} + c) + \gamma_1 \eta(\lambda_{t-1}),\ t \geq 1. \end{aligned} \tag{3}$$

The last two equations describing the conditional mean process can be summarised as

$$\eta(\lambda_t^c) = \beta_0 + \beta_1 \eta(Z_{t-1} + c) + \gamma_1 \left(\eta(\lambda_{t-1}^c) - \omega\delta^{t-1-\tau} I(t-1 \geq \tau)\right) + \omega\delta^{t-\tau} I(t \geq \tau)\ .$$

This shows the difference to model (2) more clearly.

If the time point $\tau$ and the type of an intervention, i.e. the value of $\delta$, both are known, an intervention model as formulated in (2) or (3) can be fitted by maximising the conditional likelihood iteratively, starting from suitable initial values. The existence of such a known intervention can be confirmed by comparing the test statistics of the corresponding score test to the upper percentiles of its asymptotical $\chi_1^2$-distribution, as described in the papers mentioned above. If only the time point $\tau$ is unknown, but the type is known, simulation experiments indicate that parametric bootstrap procedures work rather well: fit the model without intervention effects and calculate the score test statistics for all time points. Use the maximum of all score test statistics for all time points as the final test statistic. Then generate artificial time series without interventions from the fitted model and calculate the corresponding maximum score test statistic as well. Opt for an intervention at that time point which maximises the score test statistic for the real data, if it is among the largest $100\alpha$-percent of all maximum score test statistics. If the type of the intervention is unknown as well, the maximum score test statistics can be calculated for each type given either model (2) or (3). The simulations suggest that preference should be given to level shifts ($\delta = 1$) if they turn out to be significant, since a level shift usually causes the test statistics for the other types of intervention effects also to become large, while the reverse effect is much less pronounced. Multiple interventions can be dealt with by estimating the effect of a detected intervention and subtracting it from the time series, before the cleaned data are analysed with respect to further interventions.

Note that the above intervention models are not able to describe so called additive outliers representing e.g. pure measurement or reporting errors, i.e. the case where a single observation is changed without any effects on the future of the process. Actually, such additive outliers are difficult to deal with by a frequentist approach, since we would need to condition on the unobserved value $Y_\tau$ instead of the contaminated $Z_\tau$. Fried et al. (2013) develop a Bayesian approach for additive outliers, applying Markov Chain Monte Carlo techniques. Their simulation results provide evidence that in this way it is possible to deal with additive outliers if there are several of them. A single or very few additive outliers pose difficulties to a Bayesian approach based on little informative prior distributions, since they do not provide enough information on that component of the underlying mixture distribution which causes the outliers.

Furthermore it should be noted that we implicitly assume intervention effects to be additive when using the identity link, and multiplicative on the original scale when using the log-link, since for simplicity we introduce the intervention effects in the same way as the dependencies on the past. Another assumption underlying the intervention models formulated above, and

also the common outlier and intervention models which have been proposed for ARMA processes in the literature, is that the dynamics of the process does not change and follows the same model after an intervention as before it.

For an illustration we analyse an artificial time series of length $n = 200$ generated from model (2) with $\eta = id$, $\beta_0 = 3$, $\beta_1 = 0.4$, $\gamma_1 = 0.3$, an internal level shift of size $\omega_1 = 4$ at time $\tau_1 = 100$ and an internal spike of size $\omega_2 = 30$ at time $\tau_2 = 150$.

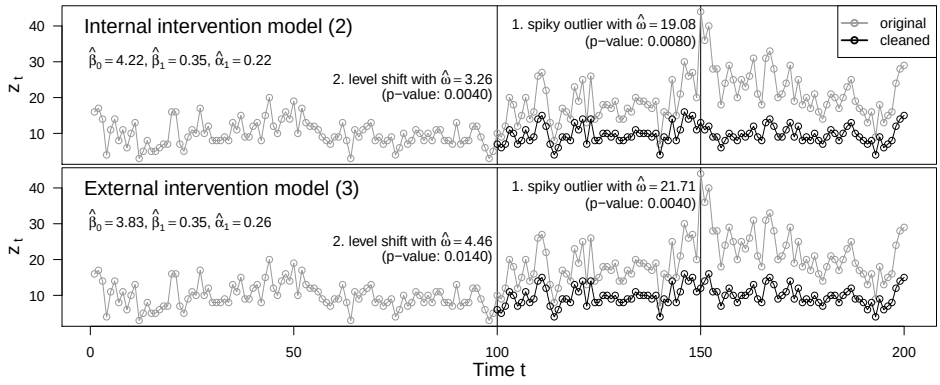


Figure 1: Results obtained from fitting both intervention models to a time series with an internal level shift at time 100 and an internal spike at time 150.

The results obtained from fitting both intervention models to these data are illustrated in Figure 1. The spike and the level shift are detected when using either of these two models, albeit with some differences between the estimated parameter values and outlier sizes, according to the different influences of such patterns on the time series. These findings confirm those of Liboschik et al. (2013): interventions can be detected successfully even if the wrong model is used. This is good news and also bad news: it is good news since it implies a certain robustness against model misspecification, but it makes a statement about the cause of an intervention effect and about its mechanism (internal / external) difficult. More work on model selection is needed for this.

## 3. Robust estimation

First attempts are available concerning the robust estimation of the model parameters in the presence of outliers and intervention effects. This is even more important because of the difficulties in specifying intervention effects correctly and because of the remaining difficulties in dealing with a single or a few additive outliers outlined above.

M-estimators are a popular generalisation of (conditional) maximum likelihood estimators which provide some robustness against outliers by replacing the log-likelihood or the score function by more robust alternatives. An M-estimator of a parameter $\theta$ can be defined as the solution of a score equation

$$\sum_{t=1}^{n} \psi(y_t, \hat{\theta}) = 0 \ . \tag{4}$$

Maximum likelihood estimation is derived by choosing $\psi(y, \theta)$ as the derivative of the log-density $\ln f_\theta(y)$ with respect to $\theta$, i.e. as the usual score function, while $\psi(y, \theta) = y - \theta$ corresponds to least squares and $\psi(y, \theta) = \text{sign}(y - \theta)$ to least absolute deviation estimation of location. The popular Huber M-estimator of the location parameter $\theta$ in a location-scale model with known (or preliminarily estimated) scale $\sigma$ is derived from

$$\psi(y, \theta) = \frac{y - \theta}{\sigma} I(-k\sigma \leq y - \theta \leq k\sigma) + k \ \text{sign}(y - \theta) I(|y - \theta| > k\sigma),$$

where $k$ is a tuning constant which determines the efficiency and the robustness of the resulting estimator. For $k = 0$ we get least absolute deviations and for $k \to \infty$ we get least squares. The score function of the Huber M-estimator is monotone. This guarantees a unique solution which can easily be determined iteratively starting from any initial value. The score function of the Tukey M-estimator,

$$\psi(y, \theta) = \frac{y - \theta}{\sigma} \left( k^2 - \frac{(y - \theta)^2}{\sigma^2} \right)^2 I(-k\sigma \leq y - \theta \leq k\sigma),$$

however, is redescending to 0 as $y-\theta$ approaches $\pm k\sigma$. This leads to the possibility of multiple solutions of the defining score equations (4).

M-estimation of generalised linear models using the Huber $\psi$-function has been treated by Cantoni and Ronchetti (2001). However, in our basic model (1) we regress on previous observations and previous conditional means, and it is well known that monotone M-estimators like those based on the Huber function need further modifications to become robust against outlying regressors. Cantoni and Ronchetti (2001) consider covariates following an elliptical distribution and use weights based on robustly estimated Mahalanobis distances to down-weight observations with outlying regressors. This approach is not natural in our context, since we regress on previous observations, which are conditionally Poisson, or some transformation of them. Empirical work on model (2) with the log-link and $\gamma_1 = 0$, that is a model without feedback, indicates that in the cases of level shift and transient shift there are no significant differences between the classical maximum likelihood estimation and the approach based on Cantoni and Ronchetti (2001). This agrees with findings for Gaussian ARMA models, that maximum likelihood and least squares work rather well in case of outliers which conform to the dynamics of the process. In the case of additive outliers, the weighted approach through robust Mahalanobis distances was found to perform much better than the classical maximum likelihood estimation, especially as the number of outliers increases. In fact, some further empirical work on the feedback case ($\gamma_1 \neq 0$) indicates that the Cantoni and Ronchetti (2001) estimation approach performs better with weights (Kitromilidou and Fokianos, 2014).

Maronna, Martin and Yohai (2006) recommend Tukey's $\psi$-function since its redescending behavior completely eliminates the influence of huge outliers and provides some robustness even in the case of outlying regressors. However, we need to use highly robust initial parameter estimates then, in order not to get trapped in a wrong solution when trying to solve (4) iteratively. This and the discreteness and strong asymmetries of Poisson models pose further problems which are not encountered in ordinary symmetric location-scale models. This will briefly be illustrated in the context of independent Poisson data in the following.

Cadigan and Chen (2001) investigate a modification of the Huber score function for the Poisson distribution. Under Poisson assumptions, the variance $\sigma^2$ equals the mean $\theta$, so that we can replace $\sigma$ by $\sqrt{\theta}$ in the above score functions, see also Elsaied (2012). Furthermore, the expectation of $\psi(Y, \theta)$ has to be zero for getting asymptotically unbiased estimates. This can be accomplished by introducing a bias correction $a$ and replacing $(y-\theta)/\sigma$ by $(y-\theta)/\sqrt{\theta} - a$ in the above formulae. Given the need for a highly robust initial estimate when using the Tukey $\psi$-function, we might want to apply the median of the data, but this only works if it is not zero because of our scaling by $\sqrt{\hat{\theta}}$, and it provides only a very rough estimate if the sample median is small. Elsaied (2012) proposes an adaptive estimate instead, combining the sample median with an estimate derived from the frequency of zero observations.

The asymptotical distribution of an M-estimator under suitable regularity conditions is $N(\theta, V_\psi(\theta))$, with the asymptotical variance $V_\psi(\theta) = E(\psi(Y, \theta)/B_\theta)^2$, where $B_\theta = \partial E\psi(Y, \theta)/\partial\theta$, see e.g. Maronna, Martin and Yohai (2006). The relative efficiency of an M-estimator as compared to the maximum likelihood estimator, which is the sample mean, under these conditions thus becomes $\theta/V_\psi(\theta)$, and is illustrated in Figure 2. Note that an estimator with a fixed tuning constant $k$ does not achieve a desirable high level of efficiency

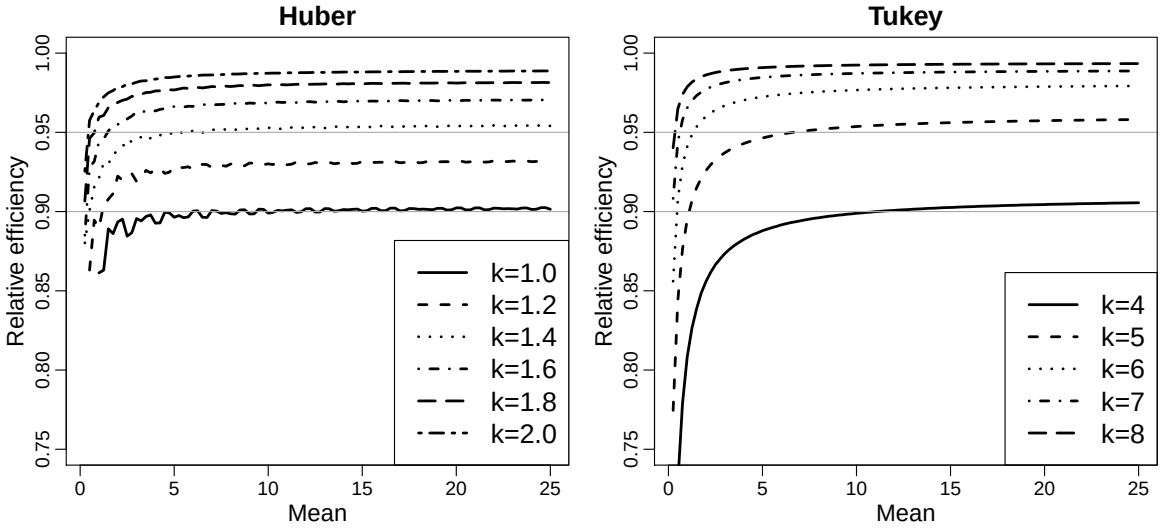


Figure 2: Asymptotical efficiencies of the Huber and the Tukey M-estimator with different tuning constants $k$ for different values of the mean $\theta$.

for all possible values of $\theta$. For further investigations in this respect and a first approach to robust M-estimation for model (1) with the identity link see Elsaied (2012).

## 4. Robust model identification

Besides the robust estimation of the parameters of a specific model, the proper identification of the link function and the model orders gets more complicated in the presence of outliers. In the following we provide a first robustness study for the identification of the model orders in case of a linear model with the identity link.

Two common tools for the choice of the model orders of linear time series models are the sample autocorrelation function (SACF) and the sample partial autocorrelation function (SPACF). However, these are strongly affected by outlying observations so that there is a need for robust and efficient alternatives. Let $\boldsymbol{y} = (y_1, \ldots, y_n)'$ be an observed time series. We consider estimation of the autocorrelation at lag $h$ by a robust bivariate correlation estimator applied to the vector $\boldsymbol{y}_t^h = (y_{1+h}, \ldots, y_n)'$ and the vector of lagged observations $\boldsymbol{y}_{t-h}^h = (y_1, \ldots, y_{n-h})'$. We consider the rank-based correlation estimators Spearman's $\rho$, Kendall's $\tau$ and Gaussian rank (for a comparison in the bivariate context see Boudt et al., 2012). Another class of autocorrelation estimators, which is based on an idea of Gnanadesikan and Kettenring (1972), employs any robust univariate scale estimator $\widehat{\text{var}}(\cdot)$. We use a variant bounded between -1 to 1 inclusive, which at lag $h$ is given by

$$\widehat{\text{acf}}_{GK}(\boldsymbol{y}; h) = \frac{\widehat{\text{var}}(\boldsymbol{y}_t^h + \boldsymbol{y}_{t-h}^h) - \widehat{\text{var}}(\boldsymbol{y}_t^h - \boldsymbol{y}_{t-h}^h)}{\widehat{\text{var}}(\boldsymbol{y}_t^h + \boldsymbol{y}_{t-h}^h) + \widehat{\text{var}}(\boldsymbol{y}_t^h - \boldsymbol{y}_{t-h}^h)}.$$

Ma and Genton (2000) study this Gnanadesikan-Kettenring (GK) approach in the Gaussian framework, using the highly robust $Q_n$ estimator of scale proposed by Croux and Rousseeuw (1992). We additionally consider the median absolute deviation from the median (MAD), the 10% and 20% winsorised variance, the interquartile range (IQR), as well as the highly robust $S_n$ (Croux and Rousseeuw, 1992) and $\tau$ (Maronna and Zamar, 2002) estimators of scale. Apart from the winsorised variance, these estimators are on the scale of the original data and need to be squared.

We compare estimators which are corrected such that they achieve consistency at the normal distribution. Note that the normal distribution is a limiting case of a Poisson distribution

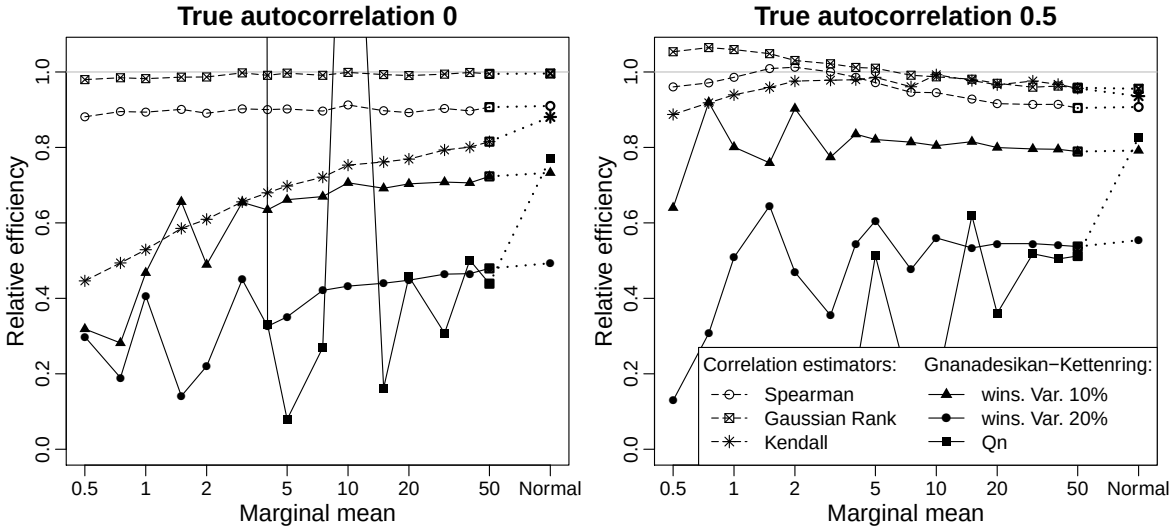


Figure 3: Efficiency of autocorrelation estimators at lag $h = 1$ relatively to the SACF. Time series of length 100 are simulated from model (1) with the marginal mean given on the horizontal axis and from a $N(\lambda_t, \lambda_t)$ model with a marginal mean of 50 (points on the very right of each plot).

with mean tending to infinity. However, we cannot expect this Fisher-consistency correction to hold true, especially in the case of a clearly skewed Poisson distribution with a small mean. Moreover, the marginal distribution of a time series from model (1) is strictly speaking only Poisson under the null hypothesis of independence.

In our simulation study we generate time series with 100 observations from the first order linear Poisson model (1) with $\eta = id$, $c = 0$ and $\gamma_1 = 0$. We consider scenarios with a true autocorrelation at lag $h = 1$ of zero ($\beta_1 = 0$) and of 0.5 ($\beta_1 = 0.5$). The results are averaged over 10 000 repetitions for each scenario and reported as a function of the marginal mean $\mu = \beta_0/(1 - \beta_1)$. The shown relative efficiencies are the ratio of the mean square errors of the SACF and the respective estimator.

The GK autocorrelation estimators based on $Q_n$ (see Figure 3), $S_n$, MAD and IQR are unsuitable for small counts, as these estimators are unstable due to the high proportion of ties in such data. It frequently happens that the scale estimations $\widehat{\text{var}}(\boldsymbol{y}_t^h + \boldsymbol{y}_{t-h}^h)$ and $\widehat{\text{var}}(\boldsymbol{y}_t^h - \boldsymbol{y}_{t-h}^h)$ coincide, resulting in an autocorrelation estimate of zero, or that one or both of them collapse to zero, resulting in an estimate of $\pm 1$ or a non-computable autocorrelation estimation, respectively. Particularly for small marginal means, we get zero estimates with high probability, causing a super-efficient performance if the true autocorrelation is zero. Implosion, that is breakdown to zero, is a known problem of many robust scale estimators. But not even the $Q_n$ estimator, which showed the best performance with respect to implosion among many other alternatives in a study of Gather and Fried (2003), does perform acceptably in the case of small counts. We also tried variants of the $Q_n$ using the 50%- and 75%-quantile of the pairwise distances, instead of the 25%-quantile as it is usually employed. Yet, for counts with low means none of these alternatives perform well. The $\tau$ estimator of scale as implemented by Maronna and Zamar (2002) is based on the variance estimation of the MAD and hence also performs poorly. We conclude that none of these popular highly robust scale estimators seems to be appropriate for small counts. Particularly for a low winsoring proportion, the winsorised variance estimator results in smaller problems with stability than the estimators mentioned before and will be considered further.

Figure 3 reconfirms the result that the efficiency of the estimators relatively to the SACF tends to its value achieved under a normal distribution. The Gaussian rank estimator has a very high relative efficiency both for uncorrelated and autocorrelated data, which does not

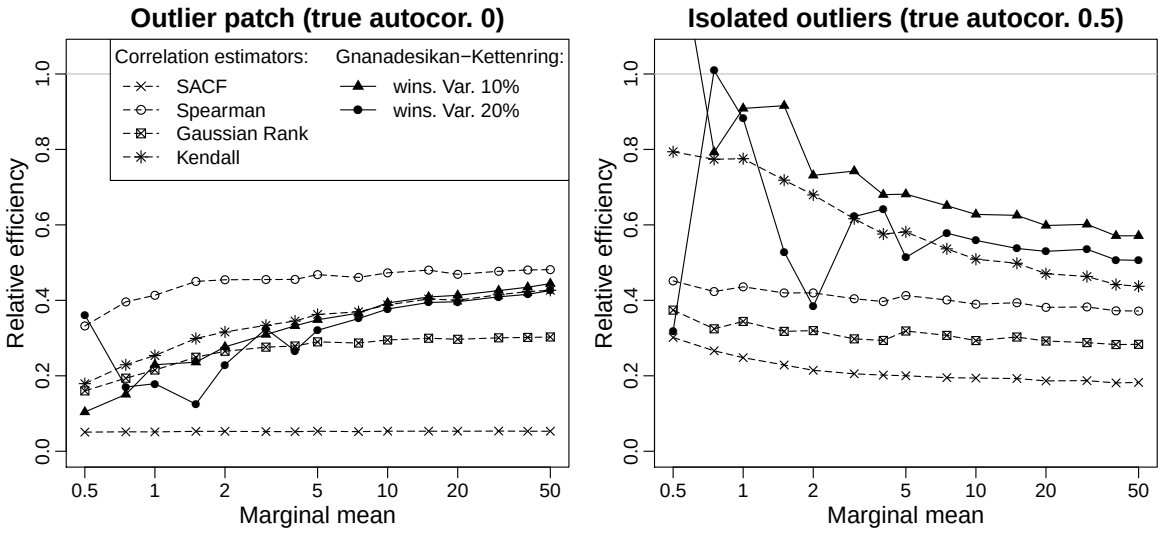


Figure 4: Efficiency of autocorrelation estimators at lag $h = 1$ for contaminated Poisson data relatively to the SACF for uncontaminated Poisson data. We contaminated 5% of the 100 observations with additive outliers of size five times the marginal standard deviation. Left: Patchy outliers in the centre. Right: Isolated outliers at arbitrarily chosen positions 17, 40, 55, 72 and 92.

depend a lot on the marginal mean. Spearman's $\rho$ correlation estimator behaves in a similar fashion, but has a lower relative efficiency of about 90% on uncorrelated data. In contrast, the relative efficiency of Kendall's $\tau$ depends very much on the marginal mean. In case of uncorrelated data its relative efficiency is below 50% for small means and even for large means slightly below Spearman's $\rho$.

To study the robustness properties of the estimators, we contaminate the time series of independent data, that is $\beta_1 = 0$, with a patch of 5% additive outliers in the centre and the autocorrelated ones with 5% of isolated additive outliers. The first outlier scenario is known to bias the estimation towards one and the latter one biases towards zero, which is away from the true values of zero and 0.5, respectively. For autocorrelation estimation when $\beta_1 = 0$, outlier patches are the worst case, whereas for time series with $\beta_1 > 0$ they can even compensate for an existing downward bias in finite samples. The simulation results in Figure 4 can be interpreted as the loss of efficiency compared to the SACF for uncontaminated data from the same model.

The outlier patch has a strong effect on the efficiency of the autocorrelation estimators for uncorrelated data (see Figure 4 left). The ordinary SACF is not robust and drops down to a relative efficiency of around 5%. The rank-based autocorrelation estimators show qualitatively the same pattern of increasing relative efficiency for increasing marginal mean. The Gaussian rank correlation, which has been the most efficient rank-based estimator for clean uncorrelated data, is the least robust one, because it gives more influence to the largest and the smallest observations. The 10%-winsorised variance has an efficiency of around 10% relatively to the SACF for clean data, which also increases with the marginal mean to about 40%. The 20%-winsorised variance is in principle slightly less efficient and shows a similar behaviour but is, as for uncontaminated data, quite unstable for low means.

The same number of isolated outliers for moderately correlated data has a weaker effect on the efficiency of the autocorrelation estimators than the outlier patch for uncorrelated data (see Figure 4 right). Unlike in the latter situation, we observe a decreasing relative efficiency for an increasing marginal mean for all estimators, except for the instability of the GK estimation based on the 20%-winsorised variance, which has been discussed before. Again, the Gaussian rank based estimator is the least efficient among the rank-based estimators, but this time

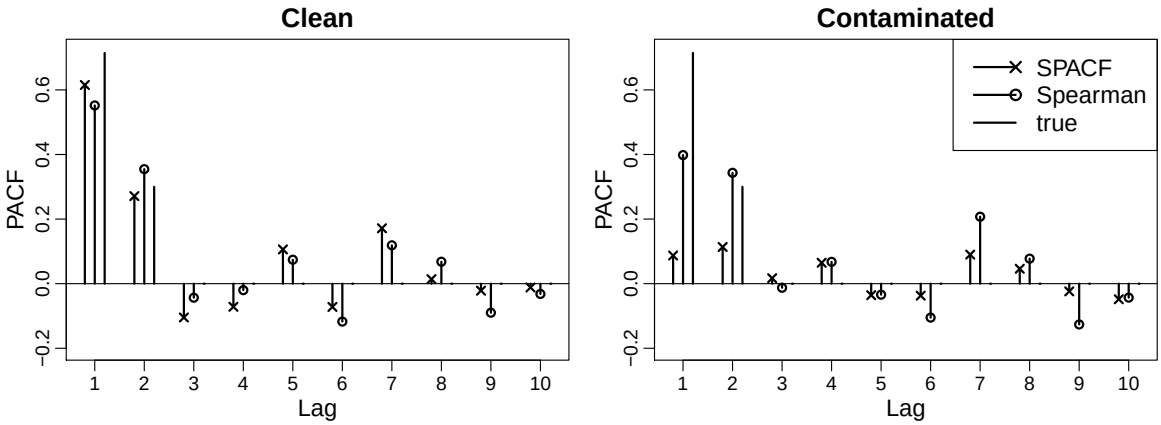


Figure 5: Estimated PACF of a simulated INARCH(2) time series of length 100 with parameters $\beta_0 = 0.4$, $\beta_1 = 0.5$ and $\beta_2 = 0.3$. Left: Clean data. Right: Contaminated with five additive outliers of size five times the marginal standard deviation at arbitrarily chosen positions 17, 40, 55, 72 and 92.

Kendall's $\tau$ is much more efficient than Spearman's $\rho$, particularly for low marginal means.

Because of the instability of most of the other estimators we recommend to use one of the rank-based autocorrelation estimators for count time series with small counts. When choosing an autocorrelation estimator one should take into account both, the desired efficiency at clean data and the desired robustness properties.

We illustrate the usefulness of robust autocorrelation estimation for identification of the model order with a simulated example. Consider a time series $(Y_t : t \in \mathbb{N}_0)$ from an integer-valued ARCH model of unknown order $p \in \mathbb{N}_0$, called INARCH($p$), with $Y_t|\mathcal{F}_{t-1}^Y \sim Pois(\lambda_t)$ and conditional mean equation $\lambda_t = \beta_0 + \beta_1 Y_{t-1} + \cdots + \beta_p Y_{t-p}$ for $t \geq 1$. We want to determine the model order $p$. The time series $(Y_t : t \in \mathbb{N}_0)$ has the same second-order properties as an AR($p$) model (cf. Ferland et al., 2006). Hence, it is known that the partial autocorrelation function (PACF) is non-zero for lags up to $p$ and zero for larger lags. We obtain the estimated partial autocorrelation function from the estimated autocorrelation function by applying the Durbin-Levinson algorithm (see for example Morettin, 1984).

Looking at Figure 5, we see that one can correctly identify the model order of an INARCH(2) model by looking at the SPACF or at the estimated PACF derived from the ACF estimation based on Spearman's $\rho$: both estimations are clearly larger than zero for the first two lags and close to zero for all other lags. In case of a contamination with isolated outliers the non-robust estimation with the SPACF is pushed towards zero, such that one might falsely identify a model of order $p = 0$. As opposed to this, the robust estimation of the PACF with Spearman's $\rho$ is not so strongly affected by the outliers and would still allow a correct model specification.

Since the Spearman correlation coefficient measures monotone, but not necessarily linear dependence, one might speculate about its possible value for the identification of the model orders in case of models applying (monotone) link functions different from the identity. However, a thorough examination of this is beyond the scope of this work.

## 5. Surveillance

The methods for detection of intervention effects in count time series described above can be applied retrospectively, i.e. when we observe the whole time series before it is analysed. An open problem so far is how these models can be used for surveillance, i.e. online detection of changes. This is an interesting problem for example in epidemiology, where we want to detect

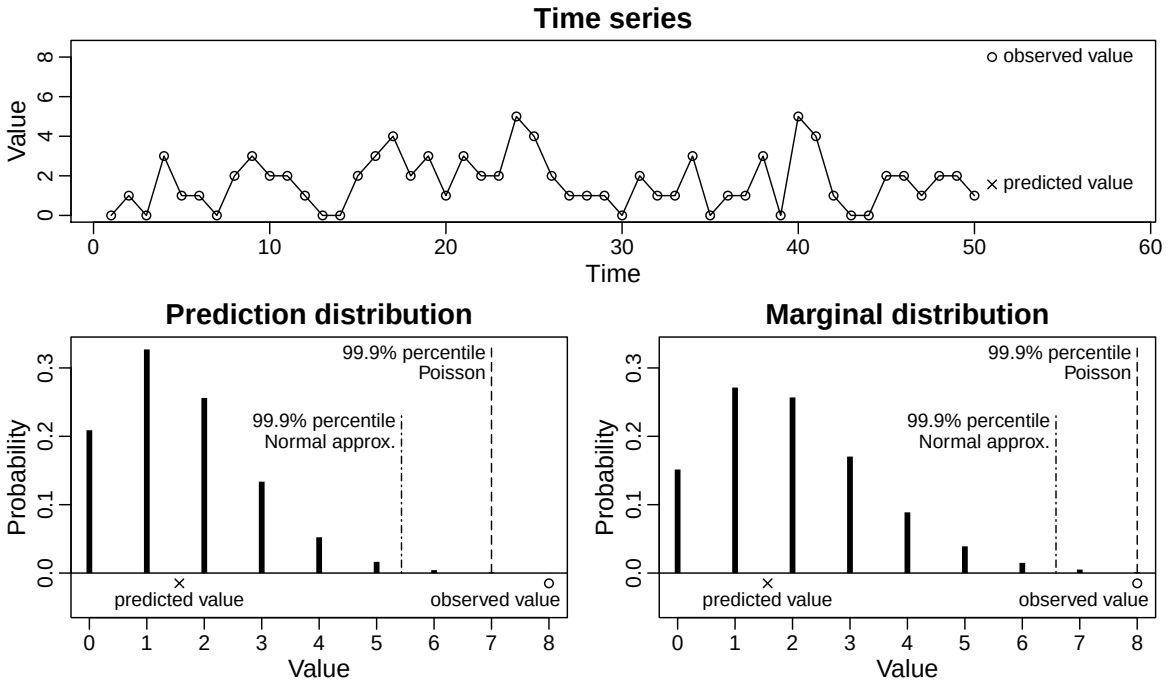


Figure 6: Simulated example for the proposed monitoring procedure. The value observed at time 51 is beyond the 99.9% percentile of the prediction, but not the marginal distribution, and would thus be identified as an outlier. A normal approximation would provide somewhat different critical values but the same conclusion in this case.

the outbreak of an epidemic with only short time delays.

An intuitive approach is to compare an incoming observation $y_{n+1}$ to its 1-step prediction $\hat{\lambda}_{n+1}$, obtained by fitting model (1) from the data observed until time point $n$, plugging in the estimated parameters into the formula for $\eta(\lambda_{n+1})$ and applying the inverse transform $\eta^{-1}$. Given such a model, there is evidence of an extraordinary effect at time $n+1$ if $y_{n+1}$ is larger than the upper $1-\alpha_N$ percentile of a Poisson distribution with mean $\hat{\lambda}_{n+1}$. Assuming the model and its parameters to be known exactly, choosing $\alpha_N = 1-(1-\alpha)^{1/N}$ ensures that we do not falsely detect any outlier with probability $1-\alpha$ when applying this rule to $N$ subsequent predictions. This follows along the same lines as in Davies and Gather (1993), who treat the independent case, since we control the probability of detecting an outlier conditionally on the past $\mathcal{F}_n^Y$. As an example, for $N = 50$ predictions an individual level of $\alpha_N = 0.1\%$ yields a global level of $\alpha = 4.9\%$. Other error probabilities $\alpha_N$ can be chosen for tuning the sensitivity and the specificity of the sequential detection procedure. For large means $\hat{\lambda}_{n+1}$ of the prediction distribution one would also need to consider downward outliers. In this case one defines, in the terminology of Davies and Gather (1993), an outlier identifier by a lower and an upper bound both depending on $\alpha_N$.

We illustrate the approach outlined above with a simulated example. We generate a time series from the first order linear Poisson model (1) with $\eta = id$, $c = 0$, $\beta_0 = 1$, $\beta_1 = 0.3$ and $\gamma_1 = 0.2$ (see Figure 6 top). In order to assess whether observation $y_{51}$ is notably large, we fit the model on the previous observations $y_1, \ldots, y_{50}$ and, based on this, compute its 1-step ahead prediction $\hat{\lambda}_{51}$. Compared with the $99,9\%$ percentile of the 1-step prediction distribution for $y_{51}$, a Poisson with mean $\hat{\lambda}_{51}$, the observed value $y_{51}$ is large and therefore identified as a potential outlier (see Figure 6 bottom left). In this case, one would have come to the same decision if we compare $y_{51}$ with the 99.9% percentile of a $N(\hat{\lambda}_{51}, \hat{\lambda}_{51})$, a normal approximation of the 1-step prediction distribution. Note that we would not have identified this observation as a potential outlier if we compare it with the 99.9% percentile of the marginal distribution of the process (see Figure 6 bottom right). Since no analytical

formula for this percentile is available, we approximated it by simulation of a time series with 100 000 observations.

An analysis of a single observation cannot tell us which type of intervention occurs, e.g. whether there is a spiky outlier or a level shift. For this we need to wait some more time points until further values $y_{n+2}, y_{n+3}, \ldots, y_{n+m}$ are observed, with a suitably chosen delay $m \in \mathbb{N}$. Instead of its 1-step ahead prediction, a comparison of $y_{n+h}$ to its $h$-step ahead prediction might be advantageous then, since the 1-step ahead prediction will strongly be affected by a level shift at time $n+1$ due to its use of $y_{n+1}, \ldots, y_{n+h-1}$. To the best of our knowledge, so far there are no simple formulae available for the conditional expectation of $Y_{n+h}$ given $\mathcal{F}_n^Y$ if $h \geq 2$, which is the natural candidate for $h$-step ahead prediction, so that we would need to rely on simulating the future given the fitted model, or use simple linear predictions instead, sticking the previous predictions $\hat{y}_{t+h-1} = \hat{\lambda}_{t+h-1}$ into the formula for $\eta(\hat{\lambda}_{t+h})$ for $h = 2, 3, \ldots, m$. However, note that the conditional distribution of $Y_{n+h}$ given $\mathcal{F}_n^Y$ is not Poisson for $h \geq 2$, so that there is need for more research on these models.

**Acknowledgements**
The work of K. Fokianos and S. Kitromilidou is supported by Cyprus Research Promotion Foundation TEXNOLOGIA/THEPIS/0609(BE)/02. The work of R. Fried and T. Liboschik is supported by the German Research Foundation (DFG, SFB 823 "Statistical modelling of nonlinear dynamic processes").

## References

Boudt K., Cornelissen J., Croux C. (2012). The Gaussian Rank Correlation Estimator: Robustness Properties. *Statistics and Computing*. Vol. **22**, pp. 471–483.

Box G.E.P., Tiao G.C. (1975). Intervention Analysis With Applications to Economics and Environmental Problems. *Journal of the American Statistical Association*. Vol. **70**, pp. 70–79.

Cadigan N.G., Chen J. (2001). Properties of Robust M-estimators for Poisson and Negative Binomial Data. *Journal of Statistical Computation and Simulation*. Vol. **70**, pp. 273–288.

Cantoni E., Ronchetti E. (2001). Robust Inference for Generalized Linear Models. *Journal of the American Statistical Association*. Vol. **96**, pp. 1022–1030.

Chang I., Tiao G.C., Chen C. (1988). Estimation of Time Series Parameters in the Presence of Outliers. *Technometrics*. Vol. **30**, pp. 193–204.

Chen C., Liu L.-M. (1993). Joint Estimation of Model Parameters and Outlier Effects in Time Series. *Journal of the American Statistical Association*. Vol. **88**, pp. 284–297.

Croux C., Rousseeuw P.J. (1993). Time-Efficient Algorithms for two Highly Robust Estimators of Scale. In: Dodge Y., Whittaker J.*Computational Statistics Volume 1*, pp. 284–297. Physika, Heidelberg.

Davies L., Gather U. (1993). The Identification of Multiple Outliers. *Journal of the American Statistical Association*. Vol. **88**, pp. 782–792.

Elsaied H. (2012). Robust Modelling of Count Data. *Unpublished PhD thesis. Department of Statistics, TU Dortmund University, Germany*. http://hdl.handle.net/2003/29404.

Ferland R., Latour A., Oraichi D. (2006). Integer–valued GARCH Processes. *Journal of Time Series Analysis*. Vol. **27**, pp. 923–942.

Fokianos K., Fried R. (2010). Interventions in INGARCH Processes. *Journal of Time Series Analysis*. Vol. **31**, pp. 210–225.

Fokianos K., Fried R. (2012). Interventions in Log-linear Poisson Autoregression. *Statistical Modelling*. Vol. **12**, pp. 299–322.

Fokianos K., Tjøstheim D. (2011). Log–linear Poisson Autoregression. *Journal of Multivariate Analysis*. Vol. **102**, pp. 563–578.

Fokianos K., Rahbek A., Tjøstheim D. (2009). Poisson Autoregression. *Journal of the American Statistical Association*. Vol. **104**, pp. 1430–1439.

Fox A.J. (1972). Outliers in Time Series. *Journal of the Royal Statistical Society. Series B*. Vol. **34**, pp. 350–363.

Fried R., Agueusop I., Bornkamp B., Fokianos K., Fruth J., Ickstadt K. (2013). Bayesian Outlier Detection in INGARCH Time Series. *Statistics and Computing*. To appear.

Gather U., Fried R. (2003). Robust Estimation of Scale for Local Linear Temporal Models. *Tatra Mathematical Publications*. Vol. **26**, pp. 87–101.

Gnanadesikan R., Kettenring J.R. (1972). Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data. *Biometrics*. Vol. **28**, pp. 81–124.

Kedem B., Fokianos K. (2002). *Regression Models for Time Series Analysis*. Wiley, Hoboken, NJ.

Kitromilidou S., Fokianos K. (2014). *Robust Estimation Methods for a Class of Count Time Series Log-Linear Models*. Submitted for publication.

Liboschik T., Kerschke P., Fokianos K., Fried R. (2013). Modelling Interventions in INGARCH Processes. *SFB 823 Discussion Paper 03/13, TU Dortmund University, Germany*. http://hdl.handle.net/2003/29878.

Ma Y., Genton M.G. (2000). Highly Robust Estimation of the Autocovariance Function. *Journal of Time Series Analysis*. Vol. **21** pp. 663–684.

Maronna R.A., Martin R.D., Yohai V.J. (2006). *Robust Statistics*. Wiley, New York.

Maronna R.A., Zamar R.H. (2002). Robust Estimates of Location and Dispersion of High-dimensional Datasets. *Technometrics*. Vol. **44**, pp. 307–317.

Morettin P.A. (1984). The Levinson Algorithm and its Applications in Time Series Analysis. *International Statistical Review*. Vol. **52**, pp. 83–92.

Tsay R.S. (1986). Time Series Model Specification in the Presence of Outliers. *Journal of the American Statistical Association*. Vol. **81**, pp. 132–141.

**Affiliation:**

Roland Fried
Statistics in Biosciences
Department of Statistics
TU Dortmund University
44221 Dortmund, Germany
E-mail: fried@statistik.tu-dortmund.de
URL: http://www.statistik.tu-dortmund.de/fried.html

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 43/3-4
June 2014

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

*Submitted:* 2013-11-25
*Accepted:* 2014-03-10

Austrian Journal of Statistics
*June 2014, Volume 43/3-4, 195–203.*
*http://www.ajs.or.at/*

# An Approach to Robustness Evaluation for Sequential Testing under Functional Distortions in $L_1$-metric

**Alexey Kharin**
Belarusian State University

**Sergey Chernov**
Belarusian State University

**Abstract**

The problem of sensitivity analysis for the sequential probability ratio test under functional distortions of the observation probability distribution is considered. For the situation where distorted densities of the log likelihood ratio statistic belong to $\varepsilon$-neighborhoods of hypothetical centers in the $L_1$-metric the least favorable distributions that maximize the conditional error probabilities are constructed. The instability coefficient is obtained to enable robustness evaluation for the sequential probability ratio test and its modification – trimmed sequential probability ratio test.

*Keywords*: sequential probability ratio test, error probability, distortion, $L_1$-metric, least favorable distribution, instability coefficient, robustness.

## 1. Introduction

The sequential approach to hypothesis testing (Wald 1947) is applied in various practical problems of statistical data analysis (Mukhopadhyay and de Silva 2009). If hypothetical suppositions are fulfilled, sequential tests require less observations at average in comparison with classical analogues based on the fixed number of observations, to provide the fixed small levels of error probabilities. However, in practice there are distortions in statistical data, i.e. the factual probability distribution of observations deviate from the hypothetical model (Kharin and Voloshko 2011). Therefore it is important to characterize the influence of the distortions on the error probabilities.

Similar problems of robustness analysis were investigated in Kharin (2002), Kharin and Kishylau (2005), Kharin (2013a) for discrete data under "contamination" (Huber and Ronchetti 2009). The problems of robustness analysis and of robust decision rules construction for case of composite hypotheses are investigated in Kharin (2008), Kharin (2011a) using the methodology of the asymptotic expansion construction for the characteristics w.r.t. the small parameter of distortion developed in Kharin and Shlyk (2009), Kharin (2005).

In Chernov and Kharin (2013) error probabilities of the sequential probability ratio test (SPRT) under functional distortions described by neighborhoods in the $L_2$-metric were studied.

In this paper we consider the case of continuos probability distribution of observations and analyze the influence of the distortions in the $L_1$-metric on the error probabilities of the SPRT. For a given maximal possible distance between the factual and the hypothetical probability distributions of the log likelihood ratio statistic the least favorable distributions (LFD) that maximize the conditional error probability of the SPRT are constructed. This maximal value of the error probability is required for the quantitative robustness analysis of sequential tests.

## 2. Mathematical Model

Consider the mathematical model from Kharin and Chernov (2011). Let $x_1, x_2, \ldots \in \mathbf{R}$ be independent and identically distributed random observations on a probability space $(\Omega, \mathcal{F}, \mathrm{P})$. Let $f(x, \theta)$ be the probability density function (p.d.f.) of $x_i$, $i \in \mathbf{N} = \{1, 2, \ldots\}$, with a parameter $\theta \in \Theta = \{\theta_0, \theta_1\}$; $F(x, \theta)$ be the cumulative distribution function that corresponds to $f(x, \theta)$.

There are two simple hypotheses concerning the unknown value of the parameter $\theta$:

$$\mathcal{H}_0 : \ \theta = \theta_0, \quad \mathcal{H}_1 : \ \theta = \theta_1. \tag{1}$$

Denote the accumulated log likelihood ratio test statistic:

$$\Lambda_n = \Lambda_n(x_1, \ldots, x_n) = \sum_{k=1}^{n} \lambda_k, \tag{2}$$

where

$$\lambda_k = \lambda(x_k) = \ln \frac{f(x_k, \theta_1)}{f(x_k, \theta_0)} \tag{3}$$

is the logarithm of the likelihood ratio statistic calculated for the observation $x_k$, $k \in \mathbf{N}$.

To test hypotheses (1) by observations $x_1, x_2, \ldots$ the SPRT (Wald 1947) can be used:

$$N = \min\{n \in \mathbf{N} : \ \Lambda_n \notin (C_-, C_+)\}, \tag{4}$$

$$d = \begin{cases} 0, & \Lambda_N \le C_- \\ 1, & \Lambda_N \ge C_+ \end{cases}, \tag{5}$$

where $N$ is the random stopping time; at this time point the decision $d$ is made according to (5). In (4) the parameters $C_-, C_+ \in \mathbf{R}$ are the test thresholds defined according to Wald (1947):

$$C_- = \ln \frac{\beta_0}{1 - \alpha_0}, \ C_+ = \ln \frac{1 - \beta_0}{\alpha_0}, \tag{6}$$

where $\alpha_0, \beta_0 \in (0, \frac{1}{2})$ are given maximal admissible values of probabilities of type I (to accept $\mathcal{H}_1$ provided $\mathcal{H}_0$ is true) and II (acceptance of $\mathcal{H}_0$ provided the true hypothesis is $\mathcal{H}_1$) errors respectively.

Let $\alpha(f)$ and $\beta(f)$ be the error probabilities of the test (4), (5) for the case where observations $x_1, x_2, \ldots$ have the probability density function $f(\cdot)$.

It is known that $\alpha_0$ and $\beta_0$ are only approximate values of the factual error probabilities $\alpha(f)$ and $\beta(f)$ of types I and II for the SPRT (4) – (6) (see Wald 1947) and can deviate from $\alpha(f)$ and $\beta(f)$ significantly (Kharin 2013a).

Without loss of generality, suppose that the hypothesis $\mathcal{H}_0$ is true, so the value of the type I error probability $\alpha$ is considered. To make formulation shorter, introduce the simplified notation:

$$F(x) = F(x, \theta_0), \ f(x) = f(x, \theta_0), \ F_\lambda(x) = \mathrm{P}_{\mathcal{H}_0}\{\lambda_1 \le x\},$$

where $\mathrm{P}_{\mathcal{H}_0}\{\cdot\}$ means the probability under the hypothesis $\mathcal{H}_0$. Let the probability density function $p_\lambda(x)$ corresponds to the cumulative distribution function $F_\lambda(x)$.

## 3. Inequalities for Error Probabilities of the SPRT

Let $x(\omega)$ and $y(\omega)$ be random variables on some probability space $(\Omega, \mathcal{F}, \mathrm{P})$ with some probability density functions $a(x)$ and $b(y)$ respectively; let also $1_A(\cdot)$ be the indicator function of the set $A$.

**Lemma 1** *If the inequality $\lambda(x(\omega)) \geq \lambda(y(\omega))$ is satisfied for every $\omega \in \Omega$, then the inequality*

$$\alpha(a) \geq \alpha(b)$$

*takes place.*

***Proof.*** It follows from the Lemma condition that

$$\Lambda_n(a) = \sum_{k=1}^{n} \lambda(x_k) \geq \sum_{k=1}^{n} \lambda(y_k) = \Lambda_n(b). \tag{7}$$

From (5) we have

$$\alpha(a) = \mathrm{P}_{\mathcal{H}_0}\{\Lambda_N(a) \geq C_+\},$$

where $N$ is the random stopping time. Because of (7) we get the relation between the random events:

$$\{\Lambda_N(a) \geq C_+\} \supseteq \{\Lambda_N(b) \geq C_+\},$$

therefore, $\alpha(a) \geq \alpha(b)$. ■

**Lemma 2** *If the inequality $\lambda(x) \geq \lambda(y)$ is satisfied for*

$$x \in M_{a>b} = \{z : \ a(z) > b(z)\}, \ y \in M_{b\geq a} = \mathbf{R} \setminus M_{a>b},$$

*then the inequality $\alpha(a) \geq \alpha(b)$ holds.*

***Proof.*** From the norm conditions for $a(\cdot)$, $b(\cdot)$ we have:

$$\int_{M_{a>b}} a(x)dx + \int_{M_{b\geq a}} a(x)dx \equiv 1 \equiv \int_{M_{b\geq a}} b(x)dx + \int_{M_{a>b}} b(x)dx.$$

Using these equations denote

$$p = \int_{M_{a>b}} (a(x) - b(x))dx = \int_{M_{b\geq a}} (b(x) - a(x))dx \in [0, 1].$$

Note that if $p = 0$, then $a(\cdot)$ and $b(\cdot)$ coincide, if $p = 1$, they are orthogonal in the sense that $a(x)b(x) = 0$, $\forall x$.

Let $\eta = \eta(\omega)$ be the Bernoulli random variable with the parameter value $p$:

$$\mathrm{P}\{\eta = 1\} = p, \ \mathrm{P}\{\eta = 0\} = 1 - p;$$

$\xi = \xi(\omega)$, $\xi^+ = \xi^+(\omega)$ and $\xi^- = \xi^-(\omega)$ be random variables with the p.d.f.s

$$p_\xi(x) = \frac{\min\{a(x), b(x)\}}{1 - p},$$

$$p_{\xi^+}(x) = \frac{1_{M_{a>b}}(x)(a(x) - b(x))}{p}, \tag{8}$$

$$p_{\xi^-}(x) = \frac{1_{M_{b\geq a}}(x)(b(x) - a(x))}{p},$$

respectively, and $\eta$, $\xi$, $\xi^+$, $\xi^-$ be independent.

The norm condition is satisfied for the functions determined by (8):

$$\int_{-\infty}^{+\infty} p_\xi(x)dx = \frac{1}{1-p}\left(\int_{M_{a>b}} b(x)dx + \int_{M_{b\geq a}} a(x)dx\right) =$$

$$\frac{1}{1-p}\left(1 - \int_{M_{b\geq a}} b(x)dx + \int_{M_{b\geq a}} a(x)dx\right) =$$

$$\frac{1}{1-p}\left(1 - \int_{M_{b\geq a}} (b(x)-a(x))\,dx\right) = \frac{1}{1-p}(1-p) \equiv 1;$$

$$\int_{-\infty}^{+\infty} p_{\xi^+}(x)dx = \frac{1}{p}\int_{M_{a>b}} (a(x)-b(x))\,dx \equiv 1;$$

$$\int_{-\infty}^{+\infty} p_{\xi^-}(x)dx = \frac{1}{p}\int_{M_{b\geq a}} (b(x)-a(x))\,dx \equiv 1.$$

The p.d.f.s $p_{\xi^+}(\cdot)$ and $p_{\xi^-}(\cdot)$ are orthogonal, and $\xi^-(\omega) \geq \xi^+(\omega)$, $\omega \in \Omega$.
Construct random variables $\xi_a = \xi_a(\omega)$, $\xi_b = \xi_b(\omega)$ on $(\Omega, \mathcal{F}, \mathrm{P})$:

$$\xi_a(\omega) = (1-\eta(\omega))\xi(\omega) + \eta(\omega)\xi^+(\omega),\ \xi_b(\omega) = (1-\eta(\omega))\xi(\omega) + \eta(\omega)\xi^-(\omega). \tag{9}$$

The p.d.f.s of random variables (9) can be found by (8):

$$p_{\xi_a}(x) = p \cdot p_{\xi^+}(x) + (1-p)\cdot p_\xi(x) =$$

$$\frac{1-p}{1-p}\cdot \min\{a(x), b(x)\} + \frac{p}{p}\cdot 1_{M_{a>b}}(x)\cdot(a(x)-b(x)) =$$

$$\begin{cases} b(x)+a(x)-b(x), & \text{if } a(x) > b(x), \\ a(x)+0, & \text{if } a(x) \leq b(x), \end{cases} \equiv a(x). \tag{10}$$

Analogously we get

$$p_{\xi_b}(x) = p \cdot p_{\xi^-}(x) + (1-p)\cdot p_\xi(x) \equiv b(x). \tag{11}$$

From the construction of $\xi^-$, $\xi^+$ and the condition of this Lemma it follows that $\lambda(\xi^+) \geq \lambda(\xi^-)$.
Analyze now the two available cases using (9).
1. If $\omega$: $\eta(\omega) = 1$, then $\xi_a(\omega) = \xi^+$, $\xi_b(\omega) = \xi^-$.
2. If $\omega$: $\eta(\omega) = 0$, then $\xi_a(\omega) = \xi_b(\omega) = \xi(\omega)$.
Combining these two results, we have $\lambda(\xi_a) \geq \lambda(\xi_b)$, $\forall\omega \in \Omega$.
Finally, using Lemma 1 we get

$$\alpha(p_{\xi_a}) \geq \alpha(p_{\xi_b}),$$

that is equivalent to $\alpha(a) \geq \alpha(b)$ because of (10), (11). ■

## 4. Robustness Evaluation for SPRT

Let the hypothetical model described in Section 1 be not satisfied, so the log likelihoods $\lambda_n = \lambda(x_n)$, $n \in \mathbf{N}$, are independent and identically distributed random variables with some p.d.f. $\tilde{p}_\lambda(x)$, that may deviate from the hypothetical p.d.f. $p_\lambda(x)$, but the distance between $\tilde{p}_\lambda(x)$ and $p_\lambda(x)$ in the $L_1$-metric does not exceed $\varepsilon$:

$$\rho_{L_1}(\tilde{p}_\lambda(\cdot), p_\lambda(\cdot)) = \int_{\mathbf{R}} |\tilde{p}_\lambda(x) - p_\lambda(x)|dx \leq \varepsilon, \tag{12}$$

where $0 \leq \varepsilon \leq \varepsilon_0$, and the maximal admissible deviation $\varepsilon_0$ is a priori known.

Denote by $L_1(p_\lambda, \varepsilon)$ the family of probability density functions $\tilde{p}_\lambda(x)$ that satisfy the inequality (12) for the fixed value of $\varepsilon$. Let the cumulative probability distribution function $\tilde{F}_\lambda(x)$ corresponds to the p.d.f. $\tilde{p}_\lambda(x)$. Let $\alpha(\tilde{p}_\lambda, \varepsilon)$ be the type I error probability for the SPRT (4), (5), when the log likelihood (3) has the p.d.f. $\tilde{p}_\lambda(\cdot) \in L_1(p_\lambda, \varepsilon)$.

Let us construct the least favorable probability distribution of $\lambda_n$, i.e. the p.d.f. that maximizes the value of $\alpha(\cdot, \varepsilon)$ within the set $L_1(p_\lambda, \varepsilon)$.

Consider the p.d.f.

$$\bar{p}_\lambda(x) = 1_{(g^-, +\infty)}(x) p_\lambda(x) + \frac{\varepsilon}{2}\delta(x - g^+), \tag{13}$$

where $\delta(\cdot)$ is the Dirac $\delta$-function,

$$g^+ = C_+ - C_-, \; F_\lambda(g^-) = \frac{\varepsilon}{2}.$$

**Lemma 3** *The function $\bar{p}_\lambda(\cdot)$ belongs to $L_1(p_\lambda, \varepsilon)$.*

***Proof.*** Find $\rho_{L_1}(\bar{p}_\lambda(\cdot), p_\lambda(\cdot))$ using (13):

$$\int_{\mathbf{R}} |\bar{p}_\lambda(x) - p_\lambda(x)|dx = \int_{(-\infty, g^-)} p_\lambda(x)dx + \int_{(g^-, +\infty)} |\bar{p}_\lambda(x) - p_\lambda(x)|dx =$$

$$\frac{\varepsilon}{2} + \frac{\varepsilon}{2} \cdot \int_{(g^-, +\infty)} \delta(x - g^+)dx = \varepsilon. \tag{14}$$

Lemma is proved. ■

Now let us prove that if the random variables $\{\lambda_n\}$ have the p.d.f. $\bar{p}_\lambda(x)$, then the type I error probability $\alpha(\bar{p}_\lambda)$ is the highest value within the neighborhood $L_1(p_\lambda, \varepsilon)$.

**Theorem 1** *If the p.d.f. $\tilde{p}_\lambda(\cdot)$ belongs to $L_1(p_\lambda, \varepsilon)$, then the following inequality holds:*

$$\alpha(\tilde{p}_\lambda, \varepsilon) \le \alpha(\bar{p}_\lambda, \varepsilon). \tag{15}$$

***Proof.*** Take any p.d.f. $\tilde{p}_\lambda(\cdot) \in L_1(p_\lambda, \varepsilon)$. Denote as in Lemma 2:

$$p = \int_{(\tilde{p}_\lambda > p_\lambda)} (\tilde{p}_\lambda(x) - p_\lambda(x))\, dx = \int_{(\tilde{p}_\lambda \le p_\lambda)} (p_\lambda(x) - \tilde{p}_\lambda(x))\, dx;$$

$$\varepsilon^-(q) = \int_{(-\infty, g^-)} q(y)dy,$$

where $q(\cdot)$ is some arbitrary p.d.f.

Note that $p \le \varepsilon/2$ and construct the auxiliary p.d.f.s $q_1(\cdot)$ and $q_2(\cdot)$:

$$q_1(x) = 1_{(\tilde{p}_\lambda \le p_\lambda)}(x)\tilde{p}_\lambda(x) + 1_{(\tilde{p}_\lambda > p_\lambda)}(x)p_\lambda(x) + p \cdot \delta(x - g^+) =$$

$$1_{(\tilde{p}_\lambda < p_\lambda)}(x)\tilde{p}_\lambda(x) + 1_{(\tilde{p}_\lambda \ge p_\lambda)}(x)p_\lambda(x) + p \cdot \delta(x - g^+),$$

$$q_2(x) = 1_{(g^-, +\infty) \cap (\tilde{p}_\lambda < p_\lambda)}(x)\tilde{p}_\lambda(x) + 1_{(g^-, +\infty) \cap (\tilde{p}_\lambda \ge p_\lambda)}(x)p_\lambda(x) +$$

$$\varepsilon^-(q_1)\delta(x - g^-) + p \cdot \delta(x - g^+). \tag{16}$$

The p.d.f. $q_1(x)$ is constructed from $\tilde{p}_\lambda(x)$ by "transferring" of the probability measure equals to $p$ from the set $\{\tilde{p}_\lambda > p_\lambda\}$ to the point $\{g^+\}$. The p.d.f. $q_2(x)$ is constructed from $q_1(x)$ by "transferring" of the probability measure (equaled to $\varepsilon^-(q_1) = \int_{(-\infty, g^-)} q_1(y)dy$) from the sets $\{\tilde{p}_\lambda < p_\lambda\} \cap (-\infty, g^-)$ and $\{\tilde{p}_\lambda \ge p_\lambda\} \cap (-\infty, g^-)$ to the point $\{g^-\}$.

Compare the four error probabilities $\alpha(\tilde{p}_\lambda)$, $\alpha(q_1)$, $\alpha(q_2)$ and $\alpha(\bar{p}_\lambda)$ using (16). Consider the sets, where the mentioned p.d.f.s differ from each other:

$$\{x:\ \tilde{p}_\lambda(x) < q_1(x)\} \subseteq \{g^+\},\ \{x:\ \tilde{p}_\lambda(x) > q_1(x)\} \subseteq \{x:\ \tilde{p}_\lambda(x) > p_\lambda(x)\} \setminus \{g^+\},$$

$$\{x:\ q_1(x) < q_2(x)\} \subseteq \{g^-\},\ \{x:\ q_1(x) > q_2(x)\} \subseteq (g^-, +\infty),$$
$$\{x:\ q_2(x) < \bar{p}_\lambda(x)\} \subseteq ((g^-, +\infty) \cap \{x:\ \tilde{p}_\lambda(x) < p_\lambda(x)\}) \cup \{g^+\},$$
$$\{x:\ q_2(x) > \bar{p}_\lambda(x)\} \subseteq \{g^-\}.$$

According to Lemma 2 we have inequalities

$$\alpha(\tilde{p}_\lambda, \varepsilon) \le \alpha(q_1, \varepsilon) \le \alpha(q_2, \varepsilon) \le \alpha(\bar{p}_\lambda, \varepsilon).$$

Therefore, the inequality (15) holds. ■

**Corollary 1** *The error probability $\alpha(\bar{p}_\lambda, \varepsilon)$ is a monotone function w.r.t. the neighborhood size $\varepsilon$, and $\forall \varepsilon \in [0, \varepsilon_0]$ the following inequality holds:*

$$\alpha(\bar{p}_\lambda, \varepsilon) \le \alpha(\bar{p}_\lambda, \varepsilon_0).$$

***Proof*** follows from the result of Lemma 2. ■

Calculate now the instability coefficient $\kappa$ (Kharin 2013b) that characterizes the relative increment of the type I error probability for the SPRT under distortion (12) from the hypothetical version:

$$\kappa = \frac{\alpha^+ - \alpha^0}{\alpha^0} \ge 0,$$

where

$$\alpha^0 = \alpha(p_\lambda),\ \alpha^+ = \sup_{\tilde{p}_\lambda \in L_1(p_\lambda, \varepsilon),\ \varepsilon \in [0, \varepsilon_0]} \alpha(\tilde{p}_\lambda, \varepsilon).$$

**Corollary 2** *The instability coefficient for the error type I probability of the SPRT is equal to*

$$\kappa = \frac{\alpha(\bar{p}_\lambda, \varepsilon_0) - \alpha(p_\lambda)}{\alpha(p_\lambda)} \ge 0.$$

***Proof.*** The result follows from Lemma 3, Theorem 1 and Corollary 1. ■

## 5. Robustness Evaluation for Trimmed SPRT

To decrease the influence of distortions on the error probabilities of the test (4), (5) we construct the trimmed probability density function $p_\lambda(x)$ for the log likelihood (3) following the idea of Kharin (2002):

$$p^g_\lambda(x) = 1_{(g^-, g^+)} p_\lambda(x) + \varepsilon^- \delta(x - g^-) + \varepsilon^+ \delta(x - g^+), \tag{17}$$

where $g^-, g^+ \in \mathbf{R}$, $g^- < g^+$, are some trimming parameters for $\lambda_n$;

$$\varepsilon^- = \varepsilon^-(p_\lambda) = F_\lambda(g^-),\ \varepsilon^+ = \varepsilon^+(p_\lambda) = 1 - F_\lambda(g^+). \tag{18}$$

Note that the function $p^g_\lambda(x)$ defined by (17) is some probability density function as it is nonnegative and the norm condition holds:

$$\int_{\mathbf{R}} p^g_\lambda(y)dy = \int_{\mathbf{R}} 1_{(g^-, g^+)} p_\lambda(y)dy + \int_{\mathbf{R}} \varepsilon^- \delta(y - g^-)dy + \int_{\mathbf{R}} \varepsilon^+ \delta(y - g^+)dy =$$
$$\int_{(g^-, g^+)} p_\lambda(y)dy + \varepsilon^- + \varepsilon^+ = (F_\lambda(g^+) - F_\lambda(g^-)) + F_\lambda(g^-) + (1 - F_\lambda(g^+)) = 1.$$

The sequential test (4) – (6) constructed using the test statistic with the trimmed probability density function (17) instead of $\lambda(\cdot)$ will be called the trimmed SPRT. If $g^- = -\infty$ and $g^+ = +\infty$, then the trimmed p.d.f. $p^g_\lambda(\cdot)$ coincides with $p_\lambda(\cdot)$, i.e. we have no trimming.

Prove now that if the p.d.f. $\tilde{p}_\lambda(\cdot)$ belongs to the $\varepsilon$-neighborhood in the $L_1$-metric of the function $p_\lambda(\cdot)$, then the trimmed p.d.f. $\tilde{p}^g_\lambda(x)$ belongs to the $\varepsilon$-neighborhood of the function $p^g_\lambda(\cdot)$ in the same metric.

**Lemma 4** *If $\tilde{p}_\lambda \in L_1(p_\lambda, \varepsilon)$, then $\tilde{p}^g_\lambda \in L_1(p^g_\lambda, \varepsilon)$.*

***Proof.*** Using (17), (18) evaluate the distance:

$$\int_{\mathbf{R}} |\tilde{p}^g_\lambda(x) - p^g_\lambda(x)|dx = \int_{(g^-,g^+)} |\tilde{p}_\lambda(x) - p_\lambda(x)|dx +$$

$$|\varepsilon^-(\tilde{s}) - \varepsilon^-(s)| \cdot \int_{\mathbf{R}} \delta(x - g^-)dx + |\varepsilon^+(\tilde{s}) - \varepsilon^+(s)| \cdot \int_{\mathbf{R}} \delta(x - g^+)dx =$$

$$\int_{(g^-,g^+)} |\tilde{p}_\lambda(x) - p_\lambda(x)|dx + |\varepsilon^-(\tilde{s}) - \varepsilon^-(s)| + |\varepsilon^+(\tilde{s}) - \varepsilon^+(s)| =$$

$$\int_{(g^-,g^+)} |\tilde{p}_\lambda(x) - p_\lambda(x)|dx + \left| \int_{(-\infty,g^-)} (\tilde{p}_\lambda(x) - p_\lambda(x))\, dx \right| +$$

$$\left| \int_{(g^+,+\infty)} (\tilde{p}_\lambda(x) - p_\lambda(x))\, dx \right| \le \int_{\mathbf{R}} |\tilde{p}_\lambda(x) - p_\lambda(x)|dx \le \varepsilon,$$

that proves the statement of the Lemma. ■

Let us find now the least favorable probability distribution for the fixed parameters of trimming $g^-$ and $g^+$, that maximizes the value of $\alpha(\cdot, \varepsilon)$ within $L_1(p^g_\lambda, \varepsilon)$. In other words, let us prove that if $\bar{p}_\lambda(\cdot)$ corresponds to the LFD in $L_1(p_\lambda, \varepsilon)$, then $\bar{p}^g_\lambda(\cdot)$ corresponds to the LFD in $L_1(p^g_\lambda, \varepsilon)$.

If $\bar{p}_\lambda(\cdot)$ satisfies (13), then $\bar{p}^g_\lambda(\cdot)$, constructed according to (17), is determined by the equation

$$\bar{p}^g_\lambda(x) = 1_{(g^-,g^+)}(x)\, p_\lambda(x) + \left(\varepsilon^- - \frac{\varepsilon}{2}\right) \delta(x - g^-) + \left(\varepsilon^+ + \frac{\varepsilon}{2}\right) \delta(x - g^+). \tag{19}$$

**Theorem 2** *If the probability density function $\tilde{p}_\lambda(\cdot)$ belongs to the set $L_1(p_\lambda, \varepsilon)$, then the following inequality holds:*

$$\alpha(\tilde{p}^g_\lambda, \varepsilon) \le \alpha(\bar{p}^g_\lambda, \varepsilon).$$

***Proof.*** The Theorem statement follows from Lemma 4 and Theorem 1. ■

**Corollary 3** *The error probability $\alpha(\bar{p}^g_\lambda, \varepsilon)$ is a monotone function w.r.t. the variable $\varepsilon$, and for every $\varepsilon$, $0 \le \varepsilon \le \varepsilon_0$, the following inequality takes place:*

$$\alpha(\bar{p}^g_\lambda, \varepsilon) \le \alpha(\bar{p}^g_\lambda, \varepsilon_0).$$

***Proof.*** The Corollary statement follows from Lemma 4 and Theorem 1. ■

Now calculate the instability coefficient (Kharin 2011b) for the type I error probability of the SPRT under distortion (12).

**Corollary 4** *The instability coefficient for the error type I probability of the trimmed SPRT is equal to*

$$\kappa = \frac{\alpha(\bar{p}^g_\lambda, \varepsilon_0) - \alpha(p^g_\lambda)}{\alpha(p^g_\lambda)} \ge 0.$$

***Proof*** follows from Lemma 4, Theorem 2 and Corollary 3. ■

## 6. Conclusions

The least favorable probability distributions of the log likelihood ratio statistic are constructed in the paper for the distortions in the $L_1$-metric. The obtained results are useful for evaluation of the difference between hypothetical and actual error probabilities under functional distortions in observation distributions, adjusted in the mentioned metric.

The results for the error type II probabilities are obtained in the same way.

The instability coefficient characterizes robustness of the SPRT and of the trimmed SPRT quantitatively.

The research is partially supported by the ISTC Project B-1910.

## References

Chernov S, Kharin A (2013). "Error Probabilities for Sequential Testing of Simple Hypotheses Under Functional Distortions in the $L_2$-Metric." *Statistical Methods of Estimation and Hypotheses Testing (in Russian)*, **25**, 64–72.

Huber P, Ronchetti E (2009). *Robust Statistics*. Wiley, New York.

Kharin A (2002). "On Robustifying of the Sequential Probability Ratio Test for a Discrete Model Under "Contaminations"." *Austrian Journal of Statistics*, **31**(4), 267–277.

Kharin A (2005). "Robust Bayesian Prediction Under Distrtions 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 (2011a). "Robustness Analysis for Bayesian Sequential Testing of Composite Hypotheses Under Simultaneous Distortions of Priors and Likelihoods." *Austrian Journal of Statistics*, **40**(1), 65–73.

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

Kharin A, Chernov S (2011). "Evaluation of the Error Probabilities for the Sequential Probability Ratio Test." *Proc. of the Belarusian State University (in Russian)*, (1), 96–100.

Kharin A, Kishylau D (2005). "Robust Sequential Testing of Hypotheses on Discrete Probability Distributions." *Austrian Journal of Statistics*, **34**(2), 153–162.

Kharin A, Shlyk P (2009). "Robust Multivariate Bayesian Forecasting Under Functional Distortions in the Chi-square Metric." *Journal of Statistical Planning and Inference*, **139**, 3842–3846.

Kharin Y (2011b). "Robustness of the Mean Square Risk in Forecasting of Regression Time Series." *Communications in Statistics – Theory and Methods*, **40**(16), 2893–2906.

Kharin Y (2013b). "Robustness in Statistical Forecasting." In C Becker et al (ed.), *Robustness and Complex Data Sructures*, pp. 225–242. Springer, Berlin.

Kharin Y, Voloshko V (2011). "Robust Estimation of AR Coefficients Under Simultaneously Influencing Outliers and Missing Values." *Journal of Statistical Planning and Inference*, **141**(9), 3276–3288.

Mukhopadhyay N, de Silva B (2009). *Sequential Methods and Their Applications*. Chapman and Hall / CRC, Boca Raton.

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

**Affiliation:**

Alexey Kharin
Department of Probability Theory and Mathematical Statistics
Belarusian State University
Independence av. 4
220030 Minsk, Belarus
E-mail: KharinAY@bsu.by
Telephone: +375 17 2095129
Fax: +375 17 2095054

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 43/3-4
June 2014

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

*Submitted:* 2013-11-25
*Accepted:* 2014-03-10

Austrian Journal of Statistics
*June 2014, Volume 43/3-4, 205–216.*
*http://www.ajs.or.at/*

# Markov Chain of Conditional Order: Properties and Statistical Analysis

**Yuriy Kharin**
Belarusian State University

**Mikhail Maltsau**
Belarusian State University

**Abstract**

The paper deals with finite Markov chain of conditional order, that is a special case of high-order Markov chain with a small number of parameters. Statistical estimators for parameters and statistical tests for parametric hypotheses are constructed and their properties are analyzed. Results of computer experiments on simulated and real data are presented.

*Keywords*: markov chain, conditional order, ergodicity, statistical estimator, hypothesis testing.

## 1. Introduction

Finite Markov chain of the order $s$ ($1 \leq s < \infty$) described by Doob (1953) is a well-known universal mathematical model to analyze long memory discrete-valued time series in many applied fields. It is used for statistical data analysis in genetics (see Waterman 1999), economics (see Ching 2004), signal processing (see Li, Dong, Zhang, Zhao, Shi, and Zhao 2010) and other areas.

Unfortunately, there is a significant disadvantage of this model. It has exponential complexity since the number of independent parameters $D(s)$ of the $N$-state Markov chain of the order $s$ increases exponentially w.r.t. $s$:

$$D(s) = (N-1)N^s = O(N^{s+1}).$$

Because of the "curse of dimensionality" to identify this model one needs time series of big size (length of time series) $n \geq D(s)$ not available in practice Kharin (2013), Kharin (2005), Kharin and Shlyk (2009). Therefore, small-parametric or parsimonious models are developed to overcome this difficulty. These models are special cases of the $s$-order Markov chain, but the number of parameters required to determine the one-step transition probability matrix is much less than $D(s)$. Let us give some examples of such parsimonious models: the Markov chain of the order $s$ with $r$ partial connections (see Kharin and Petlitskii 2007), Raftery model (see Raftery 1985), variable length Markov chain (see Buhlmann and Wyner 1999). For example, the conditional probability distribution of the current state of the Markov chain of the order $s$ with $r$ partial connections depends not on all $s$ previous states, but only on $r$

selected states. This paper is devoted to a new parsimonious model called Markov chain of conditional order proposed by authors in Kharin and Maltsew (2012).

## 2. Mathematical model

At first let us introduce the notation: $\mathbb{N}$ is the set of positive integers, $N \in \mathbb{N}$, $2 \le N < \infty$, $A = \{0, 1, \dots, N-1\}$ is the finite state space with $N$ elements; $J_n^m = (j_n, \dots, j_m) \in A^{m-n+1}$, $m \ge n$, is the multiindex (subsequence of indices from a sequence $j_1, j_2, \dots$); $\{x_t \in A : t \in \mathbb{N}\}$ is a homogeneous Markov chain of the order $s$, $(2 \le s < \infty)$ with $(s+1)$-dimensional matrix of transition probabilities $P = (p_{J_1^{s+1}})$:

$$p_{J_1^{s+1}} = \mathrm{P}\{x_{t+s} = j_{s+1} | x_{t+s-1} = j_s, \dots, x_t = j_1\},\ J_1^{s+1} \in A^{s+1},\ t \in \mathbb{N};$$

$L \in \{1, 2, \dots, s-1\}$, $K = N^L - 1$ are some positive integers; $Q^{(1)}, \dots, Q^{(M)}$ are $M$ $(1 \le M \le K+1)$ different square stochastic matrices of the order $N$:

$$Q^{(m)} = (q_{i,j}^{(m)}),\ 0 \le q_{i,j}^{(m)} \le 1,\ \sum_{j \in A} q_{i,j}^{(m)} \equiv 1,\ i, j \in A,\ 1 \le m \le M;$$

$< J_n^m >= \sum_{k=n}^{m} N^{k-n} j_k \in \{0, 1, \dots, N^{m-n+1} - 1\}$ is the numeric representation of the multiindex $J_n^m \in A^{m-n+1}$; $\mathrm{I}\{C\}$ is the indicator function of event $C$.

The Markov chain $\{x_t \in A : t \in \mathbb{N}\}$ is called the Markov chain of conditional order (see Kharin and Maltsew 2012), if its one-step transition probabilities have the following parsimonious form:

$$p_{J_1^{s+1}} = \sum_{k=0}^{K} \mathrm{I}\{< J_{s-L+1}^s >= k\} q_{j_{b_k}, j_{s+1}}^{(m_k)}, \tag{1}$$

where $1 \le m_k \le M$, $1 \le b_k \le s - L$, $0 \le k \le K$, $\min\limits_{0 \le k \le K} b_k = 1$; it is assumed that all elements of the set $\{1, 2, \dots, M\}$ occur in the sequence $m_0, \dots, m_K$. The sequence of elements $J_{s-L+1}^s$ is called the base memory fragment (BMF) of the random sequence, $L$ is the length of BMF; the value $s_k = s - b_k + 1$ is called the conditional order. Thus the conditional probability distribution of the state $x_t$ at time $t$ depends not on all $s$ previous states, but it depends only on $L+1$ selected states $(j_{b_k}, J_{s-L+1}^s)$. Note that if $L = s-1$, $s_0 = s_1 = \dots = s_K = s$, we have the fully-connected Markov chain of the order $s$. If $M = K+1$, then each transition matrix corresponds to only one value of the BMF, otherwise there exists a common matrix which corresponds to several values of BMF.

Therefore the Markov chain of conditional order is determined by the following parameters:

- unconditional order $s$ of the Markov chain;
- the length of BMF $L$;
- $K+1$ conditional orders $\{s_k : 0 \le k \le K\}$;
- $K+1$ parameters $\{m_k : 0 \le k \le K\}$ which determine the transition matrices;
- $M$ stochastic matrices of the order $N$ which are described by $MN(N-1)$ independent parameters.

Hence the transition matrix $P = (p_{J_1^{s+1}})$, $J_1^{s+1} \in A^{s+1}$, of the Markov chain of conditional order is determined by

$$d = 2(N^L + 1) + MN(N-1) \tag{2}$$

independent parameters. For example, we need no more than 66 parameters for the Markov chain of conditional order if $s = 10$, $L = 2$, whereas the fully-connected Markov chain of this order requires $D(s) = 1024$ parameters.

## 3. Statistical estimators for parameters

In this section we present statistical estimators for parameters of the Markov chain of conditional order. Introduce the notation: $X_1^n \in A^n$ is the observed time series of length $n$, $\pi^0_{J_1^s} = \mathrm{P}\{x_1 = j_1, \ldots, x_s = j_s\}$, $J_1^s \in A^s$, is the initial probability distribution of the Markov chain of conditional order (1);

$$\nu^s_{l,y}(J_1^l) = \sum_{t=1}^{n-s} \mathrm{I}\{x_{t+s-l-y+1} = j_1, X_{t+s-l+2}^{t+s} = J_2^l\},\ l \geq 2,\ 0 \leq y \leq s-l+1,$$

is frequency of the state $J_1^l \in A^l$ with the time gap of length $y$ between the elements $j_1$ and $J_2^l$; $\nu_{s+1}(J_1^{s+1}) = \nu^s_{s+1,0}(J_1^{s+1})$ is frequency of $(s+1)$-tuple $J_1^{s+1}$.

At first, let us give ergodicity conditions for the Markov chain of conditional order.

**Theorem 1**. *The Markov chain of conditional order is ergodic if and only if there exists a number $m \in \mathbb{N}, s \leq m < \infty$, such that the following inequality holds:*

$$\min_{J_1^s, J_{1+m}^{s+m} \in A^s} \sum_{J_{s+1}^m \in A^{m-s}} \prod_{i=1}^{m} \sum_{k=0}^{K} \mathrm{I}\{< J_{i+s-L}^{i+s-1} >= k\} q^{(m_k)}_{j_{b_k+i-1}, j_{i+s}} > 0. \tag{3}$$

**Proof**. Consider the first-order vector-valued Markov chain

$$\{X_t = (x_t, x_{t+1}, \ldots, x_{t+s-1}) \in A^s : t \in \mathbb{N}\}$$

with the extended state space like in Doob (1953) which is equivalent to the $s$-order Markov chain $\{x_t \in A : t \in \mathbb{N}\}$. The transition matrix for $X_t$ has the following form:

$$\bar{P} = (\bar{p}_{J_1^{2s}}),\ J_1^{2s} \in A^{2s},\ \bar{p}_{J_1^{2s}} = \mathrm{I}\{J_2^s = J_{s+1}^{2s-1}\} p_{J_1^s j_{2s}}. \tag{4}$$

According to Kemeny and Snell (1963) the Markov chain $X_t$ is ergodic if and only if there exists a number $m \in \mathbb{N}$, such that the following inequality holds:

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

where $\bar{p}^{(c)}_{J_1^s J_{1+c}^{s+c}}$ is the $c$-step transition probability from $J_1^s$ to $J_{1+c}^{s+c}$ for the Markov chain $X_t$. Using properties of probability and definition (1) we come to the criterion (3). Theorem is proved.

In the sequel we will consider ergodic Markov chains. It is known, that the probability distribution of an ergodic Markov chain tends to a stationary probability distribution. The next theorem determines conditions under which the stationary distribution is uniform.

**Theorem 2**. *If the Markov chain of conditional order is ergodic, then its stationary distribution is uniform if and only if the following equations hold ($k = 0, 1, \ldots, K$):*

$$\begin{cases} q_{ij}^{(m_k)} = 1/N, \forall i, j \in A, & \text{if } s_k \in \{L+1, \ldots, s-1\}, \\ \sum\limits_{i \in A} q_{ij}^{(m_k)} = 1, \forall j \in A \text{ ( that is } Q^{(m_k)} \text{ is a doubly stochastic matrix)}, & \text{if } s_k = s. \end{cases} \tag{5}$$

**Proof**. As in the proof of Theorem 1 consider the first-order vector Markov chain $X_t$. It is known from Borovkov (1998b), that the stationary distribution for $X_t$ is uniform if and only if $\bar{P}$ is a doubly stochastic matrix, that is

$$\sum_{J_1^s \in A^s} \bar{p}_{J_1^{2s}} = 1,\ \forall J_{s+1}^{2s} \in A^s. \tag{6}$$

Define $k =< J_{2s-L}^{2s-1} >$ and transform (6) using (4) and (1):

$$\sum_{J_1^s \in A^s} \bar{p}_{J_1^{2s}} = \sum_{J_1^s \in A^s} \mathbf{I}\{J_2^s = J_{s+1}^{2s-1}\} q_{j_{b_k}, j_{2s}}^{(m_k)} = \sum_{j_1 \in A} q_{j_{b_k}, j_{2s}}^{(m_k)} = 1. \tag{7}$$

If $s_k = s$, then $b_k = 1$ and $\sum\limits_{j_1 \in A} q_{j_1, j_{2s}}^{(m_k)} = 1$. Hence $Q^{(m_k)}$ is a doubly stochastic matrix, and we have the second row in (5). If $s_k < s$, then $b_k > 1$, $\sum\limits_{j_1 \in A} q_{j_{b_k}, j_{2s}}^{(m_k)} = N q_{j_{b_k}, j_{2s}}^{(m_k)} = 1$, and we have the first row in (5). Theorem is proved.

We will use the likelihood function to estimate transition probability matrices $\{Q^{(m_k)}\}$ and conditional orders $\{s_k\}$. In order to build it we have to find $n$-dimensional probability distribution for the observed time series $X_1^n$ generated by the model (1).

**Lemma 1**. *The $n$-dimensional probability distribution ($n > s$) for the Markov chain of conditional order (1) has the following form:*

$$\mathrm{P}\{x_1 = j_1, \ldots, x_n = j_n\} = \pi_{J_1^s}^0 \prod_{t=s}^{n-1} \sum_{k=0}^{K} \mathrm{I}\{< J_{t-L+1}^t >= k\} q_{t-s+b_k, j_{t+1}}^{(m_k)}, \; j_1, \ldots, j_n \in A. \tag{8}$$

**Proof**. Using theorem on compound probabilities and the Markov property we have:

$$\mathrm{P}\{x_1 = j_1, \ldots, x_n = j_n\} = \pi^0(J_1^s) \prod_{t=s}^{n-1} p_{J_{t-s+1}^{t+1}}.$$

Hence, taking into account definition (1), we come to (8). Lemma is proved.

**Corollary 1**. *The loglikelihood function for the Markov chain of conditional order (1) has the following form:*

$$l_n(X_1^n, \{Q^{(i)}\}, L, \{s_k\}, \{m_k\}) = \ln \pi_{X_1^s}^0 + \\ + \sum_{J_0^{L+1} \in A^{L+2}} \sum_{k=0}^{K} \mathrm{I}\{< J_1^L >= k\} \nu_{L+2, s_k - L - 1}^s (J_0^{L+1}) \ln q_{j_0, j_{L+1}}^{(m_k)}.$$

Now we can construct maximum likelihood estimators (MLEs) for the transition probabilities $\{Q^{(m_k)} : k = 0, \ldots, K\}$ and the conditional orders $\{s_k : k = 0, \ldots, K\}$.

**Theorem 3**. *If the true values $s$, $L$, $\{s_k : k = 0, \ldots, K\}$ and $\{m_k : k = 0, \ldots, K\}$ are known, then the MLEs for the one-step transition probabilities $\{q_{j_0, j_{L+}}^{(m_k)}, j_0, j_{L+1} \in A : k = 0, \ldots, K\}$ are*

$$\hat{q}_{j_0, j_{L+1}}^{(m_k)} = \begin{cases} \dfrac{\sum\limits_{J_1^L \in M_{m_k}} \nu_{L+2, g(s_k, L)}^s (J_0^{L+1})}{\sum\limits_{J_1^L \in M_{m_k}} \nu_{L+1, g(s_k, L)}^s (J_0^L)}, \text{ if } \sum\limits_{J_1^L \in M_{m_k}} \nu_{L+1, g(s_k, L)}^s (J_0^L) > 0, \\ 1/N, \text{ if } \sum\limits_{J_1^L \in M_{m_k}} \nu_{L+1, g(s_k, L)}^s (J_0^L) = 0, \end{cases} \tag{9}$$

*where $M_i = \{J_1^L \in A^L : m_{<J_1^L>} = i\}$, $i = 1, \ldots, M$, $\bigcup\limits_{i=1}^{M} M_i = A^L$, $g(i, j) = i - j - 1$.*

**Proof**. In order to construct the MLEs we need to solve the following problem:

$$l_n(X_1^n, \{Q^{(i)}\}, L, \{s_k\}, \{m_k\}) \to \max_{\{Q^{(m_k)}\}_{1 \le m_k \le M}}, \\ \sum_{j_{L+1} \in A} q_{j_0, j_{L+1}}^{(m_k)} = 1, \; j_0 \in A, \; 1 \le m_k \le M.$$

This maximization problem splits into $N^{L+1}$ subproblems ($j_0 \in A, J_1^L \in A^L$):

$$\sum_{j_{L+1}\in A}\sum_{k=0}^{K} \mathrm{I}\{< J_1^L >= k\}\nu_{L+2,g(s_k,L)}(J_0^{L+1}) \ln q_{j_0,j_{L+1}}^{(m_k)} \to \max_{q_{j_0,j_{L+1}}^{(m_k)}},$$
$$\sum_{j_{L+1}\in A} q_{j_0,j_{L+1}}^{(m_k)} = 1.$$

Solve these subproblems with Lagrange multiplier method and come to the estimators (9). Theorem is proved.

In the rest of the paper we will assume that $M = K + 1$, i.e. $K + 1$ independent matrices correspond to $K + 1$ different values of BMF, and $m_k = k + 1, k = 0, 1, \ldots, K$. In this case estimators (9) have the following form:

$$\hat{q}_{j_0,j_{L+1}}^{(k+1)} = \begin{cases} \sum\limits_{J_1^L\in A^L} \mathrm{I}\{< J_1^L >= k\}\dfrac{\nu^s_{L+2,g(s_k,L)}(J_0^{L+1})}{\nu^s_{L+1,g(s_k,L)}(J_0^L)}, \text{ if } \nu^s_{L+1,g(s_k,L)}(J_0^L) > 0, \\ 1/N, \text{ if } \nu^s_{L+1,g(s_k,L)}(J_0^L) = 0. \end{cases} \tag{10}$$

We will also use the following notation for transition probabilities and their estimators:

$$q(J_0^{L+1}) = \sum_{k=0}^{K} \mathrm{I}\{< J_1^L >= k\} q_{j_0,j_{L+1}}^{(k+1)},\ \hat{q}(J_0^{L+1}) = \sum_{k=0}^{K} \mathrm{I}\{< J_1^L >= k\} \hat{q}_{j_0,j_{L+1}}^{(k+1)}.$$

According to Kharin and Maltsew (2011) we construct estimators for the conditional orders $\{s_k\}$.

**Theorem 4**. *If $s$ and $L$ are known, then the MLEs for conditional orders $\{s_k : k = 0, \ldots, K\}$ are*

$$\hat{s}_k = arg \max_{L+1\le y\le s} \sum_{J_1^L\in A^L} \mathrm{I}\{< J_1^L >= k\} \sum_{j_0,j_{L+1}\in A} \nu^s_{L+2,g(y,L)}(J_0^{L+1}) \ln(\hat{q}_{j_0,j_{L+1}}^{(k+1)}). \tag{11}$$

In order to estimate the order $s$ and the BMF length $L$ we use Bayesian information criterion (BIC) (see Csiszar and Shields 1999):

$$(\hat{s}, \hat{L}) = arg \min_{2\le s'\le S_+,\ 1\le L'\le L_+} BIC(s', L'), \tag{12}$$

$$BIC(s', L') = -2 \sum_{J_0^{L'+1}\in A^{L'+2}} \sum_{k=0}^{K} \mathrm{I}\{< J_1^{L'} >= k\} \nu^{s'}_{L'+2,\hat{g}(s_k,L')}(J_0^{L'+1}) \ln \hat{q}_{j_0,j_{L'+1}}^{(k+1)} +$$
$$+ d \ln(n - s'),$$

where $S_+ \ge 2$, $1 \le L_+ \le S_+ - 1$, are maximal admissible values of $s$ and $L$ respectively, $d$ is the number of independent parameters of the model (1) defined by formula (2).

## 4. Asymptotic properties of statistical estimators

Let us assume that the Markov chain (1) satisfies the stationarity condition. Define the probability distribution of the $l$-tuple $X_{t+l-1}^t \in A^l$, $l \in \mathbb{N}$:

$$\pi_l(J_1^l) = \mathrm{P}\{x_t = j_1, \ldots, x_{t+l-1} = j_l\},\ J_1^l \in A^l,\ t = 1, 2, \ldots.$$

At first, let us present results on consistency of the constructed statistical estimators from the previous section.

**Theorem 5**. *If Markov chain of conditional order (1) is stationary, then the statistical estimators (9) are consistent estimators as* $n \to \infty$*:*

$$\hat{q}_{ij}^{(k+1)} \xrightarrow{\mathrm{P}} q_{ij}^{(k+1)},\ i,j \in A,\ k = 0,\dots,K. \tag{13}$$

**Proof**. It is known from Basawa and Prakasa Rao (1980) that frequencies of the states for the first-order vector Markov chain $X_t$ (considered in the proof of Theorem 1) tend to the stationary probability distribution as $n \to \infty$:

$$\frac{1}{n-s}\sum_{t=1}^{n-s} \mathrm{I}\{X_t = J_1^s, X_{t+1} = J_2^{s+1}\} \xrightarrow{\mathrm{P}} \pi_{s+1}(J_1^{s+1}),\ J_1^{s+1} \in A^{s+1}.$$

Thus we can prove that $\hat{\pi}_{s+1}(J_1^{s+1}) = \nu_{s+1}(J_1^{s+1})/(n-s) \xrightarrow{\mathrm{P}} \pi_{s+1}(J_1^{s+1})$. Then we consider $\nu^s_{L+2,g(s_k,L)}(J_0^{L+1})$ and $\nu^s_{L+1,g(s_k,L)}(J_0^{L})$ as sums of the frequencies of $(s+1)$-tuples $\nu_{s+1}(J_1^{s+1})$:

$$\nu^s_{l+1,g(s_k,L)}(J_0^l) = \sum_{I_1^{s+1} \in A^{s+1}(g(s_k,L),J_0^l)} \nu_{s+1}(J_1^{s+1}),\ l \in \{L, L+1\},$$

where $A^{s+1}(y, J_0^l) = \{I_1^{s+1} \in A^{s+1} : i_1 = j_0, I_{y+2}^{y+l} = J_2^l\},\ y = 0,1,\dots$. So the following convergence holds:

$$\nu^s_{l+2,g(s_k,L)}(J_0^{l+1}) \xrightarrow{\mathrm{P}} \pi_{l+1,g(s_k,L)}(J_0^l) = \mathrm{P}\{x_t = j_0, X_{t+s_k-L}^{t+s_k-L+l-1} = J_1^l\}.$$

Note that $\pi_{L+2,g(s_k,L)}(J_0^{L+1}) = \sum\limits_{k=0}^{K} \mathrm{I}\{< J_1^L >= k\}\pi_{L+1,g(s_k,L)}(J_0^L)q_{j_0,j_{L+1}}^{(k+1)}$; using this equation and theorem on functional transformations of convergent random sequences from Borovkov (1998a), we come to (13). Theorem is proved.

**Theorem 6.** *Under conditions of Theorem 5 statistical estimators (11) are consistent as* $n \to \infty$*:*

$$\hat{s}_k \to s_k,\ k = 0,\dots,K+1. \tag{14}$$

**Proof**. Introduce the notation:

$$I_k(y) = \sum_{j_0,j_{L+1}\in A} \pi_{L+2,g(y,L)}(J_0^{L+1}) \ln \frac{\pi_{L+2,g(y,L)}(J_0^{L+1})}{\pi_{L+1,g(y,L)}(J_0^L)\pi_1(j_{L+1})}, y \in \{L+1,\dots,s\},$$

is the Shannon information on the random symbol $x_{L+1}$ contained in the random symbol $x_0$ under the fixed BMF $X_1^L = J_1^L$;

$$\bar{I}_k = \sum_{H_1^{s-L}\in A^{s-L}} \sum_{j_{L+1}\in A} \pi_{s+1}(H_1^{s-L}J_1^{L+1}) \ln \frac{\pi_{s+1}(H_1^{s-L}J_1^{L+1})}{\pi_s(H_1^{s-L}J_1^L)\pi_1(j_{L+1})}, y \in \{L+1,\dots,s\},$$

is the Shannon information on the random symbol $x_{s+1}$ contained in the $(s-L)$-tuple $X_1^{s-L}$ under the fixed BMF $X_{s-L+1}^s = J_1^L$;

$$\hat{I}_k(y) = \sum_{j_0,j_{L+1}\in A} \hat{\pi}_{L+2,g(y,L)}(J_0^{L+1}) \ln \frac{\hat{\pi}_{L+2,g(y,L)}(J_0^{L+1})}{\hat{\pi}_{L+1,g(y,L)}(J_0^L)\hat{\pi}_1(j_{L+1})}, y \in \{L+1,\dots,s\},$$

is the plug-in statistical estimator for $I_k(y)$. At first, note that

$$arg \max_{L+1\le y\le s} \sum_{j_0,j_{L+1}\in A} \nu^s_{L+2,g(y,L)}(J_0^{L+1}) \ln(\hat{q}_{j_0,j_{L+1}}^{(k+1)}) = arg \max_{L+1\le y\le s} \hat{I}_k(y), \tag{15}$$

where $< J_1^L >= k$. The second statement we need to prove the theorem, is the following:

$$I_k(s_k) = \bar{I}_k. \tag{16}$$

Using (16) and properties of Shannon information we can show that $I_k(s_k) \geq I_k(y)$, $\forall y \neq s_k$. Thus applying the first continuity theorem from Borovkov (1998a) and the equation (15) we come to (14). Theorem is proved.

**Theorem 7**. *Under conditions of Theorem 5 statistical estimators (12) are consistent as* $n \to \infty$*:*

$$(\hat{s}, \hat{L}) \xrightarrow{\mathrm{P}} (s, L).$$

**Proof**. Let $\pi_{l,y}(J_1^l) = \mathrm{P}\{x_t = j_1, X_{t+y+1}^{t+y+l-1} = J_2^l\}$, $l \geq 2$, $y \geq 0$. Then $q^{(k+1)}_{j_0,j_{L+1}} = \dfrac{\pi_{L+2,g(s_k,L)}(J_0^{L+1})}{\pi_{L+1,g(s_k,L)}(J_0^L)}$, where $< J_1^L >= k$. Note that if $X_1^{L'} = J_1^{L'}$ is fixed, then $-\sum\limits_{j_0,j_{L'}\in A} \pi_{L'+2,y}(J_0^{L'+1}) \ln \dfrac{\pi_{L'+2,y}(J_0^{L'+1})}{\pi_{L'+1,y}(J_0^{L'})}$ is a conditional entropy $H_{J_1^{L'},y}\{x_{L'+1}|x_0\}$ of $x_{L'+1}$ given $x_0$. Using asymptotic properties of the estimators (10) and (11) it is easy to show that for $n \to \infty$ the following asymptotics holds:

$$-\frac{1}{n} \sum_{J_0^{L'+1}\in A^{L'+2}} \sum_{k=0}^{K} \mathrm{I}\{< J_1^{L'} >= k\} \nu^{s'}_{L'+2,g(\hat{s}_k,L')}(J_0^{L'+1}) \ln \frac{\nu^{s'}_{L'+2,g(\hat{s}_k,L')}(J_0^{L'+1})}{\nu^{s'}_{L'+1,g(\hat{s}_k,L')}(J_0^{L'})} \xrightarrow{\mathrm{P}}$$

$$\xrightarrow{\mathrm{P}} \sum_{J_1^{L'}\in A^{L'}} \sum_{k=0}^{K} \mathrm{I}\{< J_1^{L'} >= k\} H_{J_1^{L'},g(y_k,L')}\{x_{L'+1}|x_0\},$$

where $L'+1 \leq y_k \leq s'$. Using properties of entropy and methods described in Csiszar and Shields (1999) we can prove that $\mathrm{P}\{(\hat{s}, \hat{L}) \in \{[2, S+] \times [1, L_+]\} \setminus \{(s, L)\}\} \xrightarrow{\mathrm{P}} 0$ at $n \to \infty$. Theorem is proved.

Now let us analyze the asymptotic normality property for estimators (10). Theorem 8 establishes asymptotic probability distribution of the normalized deviations of the statistical estimators for transition probabilities:

$$\bar{q}(J_0^{L+1}) = \sqrt{n-s}(\hat{q}(J_0^{L+1}) - q(J_0^{L+1}))\ , J_0^{L+1} \in A^{L+2}.$$

**Theorem 8**. *Under conditions of Theorem 5 as* $n \to \infty$ *the normalized deviations* $\{\bar{q}(J_0^{L+1}) : J_0^{L+1} \in A^{L+2}\}$ *have joint asymptotically normal probability distribution with zero mean and covariance matrix* $\Sigma_q = \Sigma_q(H_0^{L+1}, J_0^{L+1})$, $H_0^{L+1}$, $J_0^{L+1} \in A^{L+2}$:

$$\Sigma_q(H_0^{L+1}, J_0^{L+1}) = \mathrm{I}\{H_0^L = J_0^L\} q(H_0^{L+1}) \frac{\mathrm{I}\{h_{L+1} = j_{L+1}\} - q(H_0^L j_{L+1})}{\pi(H_0^L)}. \tag{17}$$

**Proof**. Let us give only a scheme of the proof. Complete proof can be found in Kharin and Maltsew (2012). The theorem is proved using asymptotic normality property for frequencies $\nu_{s+1}(J_1^{s+1})$ from Kharin and Petlitskii (2007). We represent the estimator $\bar{q}(J_0^{L+1})$ as a function of these frequencies. Therefore using the third continuity theorem from Borovkov (1998a) we can establish asymptotic normality property for estimators (10) and come to (17). Theorem is proved.

## 5. Statistical testing of hypotheses on the values of $\{Q^{(k)}\}$

Using the results of Section 4 let us construct a statistical test for two hypotheses:

$$H_0 = \{Q^{(1)} = Q_0^{(1)}, \ldots, Q^{(K+1)} = Q_0^{(K+1)}\}, H_1 = \bar{H}_0, \tag{18}$$

where $Q_0^{(1)}, \dots, Q_0^{(K+1)}$ are some fixed $K+1$ stochastic matrices of the order $N$.
For the decision making we will use the following statistic:

$$\rho = \rho(n) = \sum_{J_0^L \in A^{L+1}} \sum_{j_{L+1} \in Q(J_0^L)} \bar{q}_0^2(J_0^{L+1}) \pi_{L+1}(J_0^L)/q(J_0^{L+1}),$$

$$Q(J_0^L) = \{j_{L+1} \in A : q(J_0^{L+1}) > 0\},$$

where $\bar{q}_0^2(J_0^{L+1}) = \sqrt{n-s}(\hat{q}(J_0^{L+1}) - q_0(J_0^{L+1}))$.

**Theorem 9**. *Under conditions of Theorem 5 as $n \to \infty$ the probability distribution of the random variable $\rho(n)$ tends to the standard $\chi^2$-distribution with $u$ degrees of freedom,*

$$u = \sum_{J_0^L \in A^{L+1}} \left( |Q^(J_0^L)| - 1 \right).$$

**Proof**. Let us give only a scheme of the proof. Complete proof can be found in Kharin and Maltsew (2012). Since normalized deviations $\{\bar{q}(J_0^{L+1}) : J_0^{L+1} \in A^{L+1}\}$ have the joint asymptotically normal distribution according to Theorem 8, we can establish the probability distribution of $\rho(n)$ using the theorem on quadratic forms for multidimensional Gaussian vectors and the second continuity theorem from Borovkov (1998a). Theorem is proved.

Now we can construct the statistical test for the hypotheses (18) based on the statistic $\rho(n)$:

$$\text{accept the hypothesis} \begin{cases} H_0, \ if \ \rho(n) \le \Delta, \\ H_1, \ if \ \rho(n) > \Delta, \end{cases} \tag{19}$$

where $\Delta = G_u^{-1}(1-\alpha)$ is the $(1-\alpha)$-quantile of the standard $\chi^2$-distribution with $u$ degrees of freedom, $\alpha \in (0,1)$ is the given significance level.

**Corollary 2**. *Under conditions of Theorem 5 as $n \to \infty$ the asymptotic size of the test (19) is equal to the given significance level $\alpha \in (0,1)$:*

$$\alpha_n = \mathrm{P}\{\rho(n) > \Delta | H_0\} \xrightarrow[n\to\infty]{} \alpha.$$

Let us consider now the alternative hypothesis of the following special type:

$$H_{1n} = \{Q^{(1)} = Q_1^{(1)}, \dots, Q^{(K+1)} = Q_1^{(K+1)}\}, \tag{20}$$

$$Q_1^{(k)} = Q_0^{(k)} + \frac{1}{\sqrt{n-s}} \gamma^{(k)}, \ \gamma^{(k)} = (\gamma_{i,j}^{(k)}), i, j \in A, k = 1, \dots, K+1,$$

where $\{\gamma^{(k)}\}$ are some fixed square matrices of the order $N$, such that $\sum\limits_{j \in A} \gamma_{i,j}^{(k)} = 0$, $\sum\limits_{i,j \in A} (\gamma_{i,j}^{(k)})^2 > 0$. Formula (20) means that the alternative hypothesis $H_{1n}$ tends to the null hypothesis $H_0$ as $n \to \infty$; such a family of hypotheses $\{H_{1n} : n = 1, 2, \dots\}$ is called the family of contigual hypotheses (see Roussas 1972). For this case we can obtain the asymptotic power of the test (19). The next theorem is proved by analogy with Theorem 9. Complete proof is given in Kharin and Maltsew (2012).

**Theorem 10**. *If the Markov chain of conditional order (1) is stationary and the contigual family of alternatives (20) holds, then as $n \to \infty$ the probability distribution of the random variable $\rho(n)$ tends to the noncentral $\chi^2$-distribution with $u$ degrees of freedom and the noncentrality parameter $\lambda$:*

$$\lambda = \sum_{\substack{J_0^L \in A^{L+1}, \\ j_{L+1} \in Q(J_0^L)}} \frac{\pi_{L+1}(J_0^L)}{q(J_0^{L+1})} \gamma^2(J_0^{L+1}),$$

*where* $\gamma(J_0^{L+1}) = \sum_{k=1}^{K+1} \mathrm{I}\{< J_1^L >= k\}\gamma_{j_0,j_{L+1}}^{(k)}$.

**Corollary 3**. *Under conditions of Theorem 9 the power of the test (19) as* $n \to \infty$ *tends to the limit:*

$$w = 1 - G_{u,\lambda}(G_u^{-1}(1-\alpha)), \tag{21}$$

*where* $G_{u,\lambda}$ *is the distribution function of the noncentral* $\chi^2$*-distribution with* $u$ *degrees of freedom and the noncentrality parameter* $\lambda$ *and* $\alpha \in \{0,1\}$ *is the given significance level.*

Let us note that the power doesn't tend to 1 because the alternative hypothesis $H_{1n}$ tends to the null hypothesis as $n \to \infty$.

## 6. Computer experiments on hypothesis testing

**Simulated data**. At first, we evaluate the test (19) performance for contigual alternatives (20) in two series of computer experiments by the following scheme: $U = 1000$ realizations of the Markov chain of conditional order were simulated according to (1). Parameters of the model: $N = 2, A = \{0,1\}, s = 8, L = 2,\ M = 4, s_0 = 8, s_1 = 6, s_2 = 8, s_3 = 3$. The length of the time series $n \in \{1000, 1500, \ldots, 20000\}$. **In the first series** of experiments the transition probabilities were chosen randomly for the null hypothesis $H_0$. **In the second series** of experiments transition probabilities were chosen randomly to provide alternative hypothesis $H_1$. In both series the frequency of the decision "accept the hypothesis $H_1$" was calculated at the fixed value of $n$:

$$\nu_\rho = \frac{1}{U}\sum_{u=1}^{U} \mathrm{I}\{\rho_u(n) > \Delta\},$$

where $\rho_u(n)$ is the value of $\rho(n)$ calculated by the $u$-th realization. In the first series $\nu_\rho$ is the estimator of the error I probability, we will denote it $\hat{\alpha}$. In the second series $\nu_\rho$ is the estimator of the power, we will denote it $\hat{w}$. Results for the first series of experiments are presented in Figure 1; results for the second series of experiments are presented in Figure 2. On both figures horizontal axis corresponds to the time series length $n$, vertical axis corresponds to the value of $\nu_\rho$; in both cases $\alpha = 0.05$. Solid line in Figure 1 plots the significance level $\alpha$. Solid line in Figure 2 plots the theoretical power (21) of the test. As we can see, theoretical values of $\alpha$ and $w$ are close enough to their experimental values $\hat{\alpha}$, $\hat{w}$ respectively which are indicated by dark circles.

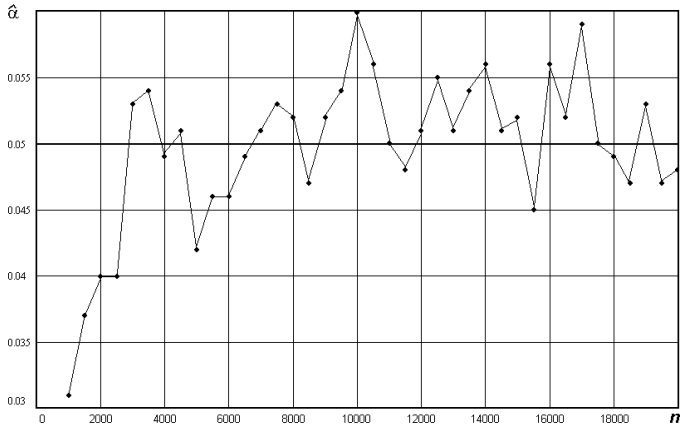


Figure 1: Dependence $\hat{\alpha}$ on $n$.

**Real data**. The real data we used is a genetic sequence from the human DNA. Splitting of genes into coding segments (exons) and noncoding segments (introns) is an important problem

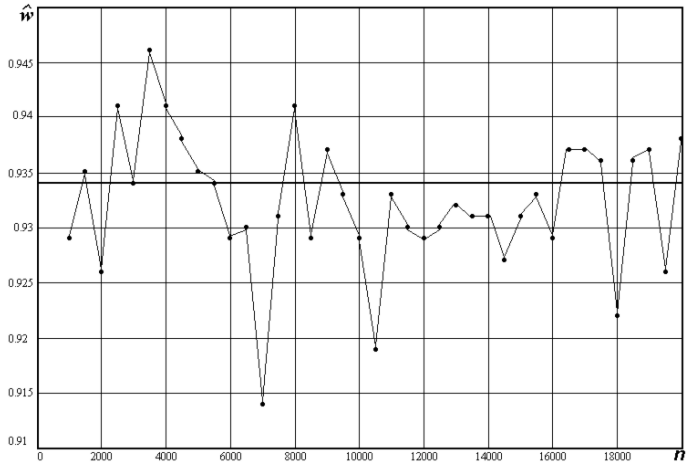


Figure 2: Dependence $\hat{w}$ on $n$.

in bioinformatics, and fitting a stochastic model for genetic sequence is a fruitful approach to this problem decsribed in Burge and Karlin (1997).

The sequence of introns from the human gene HSHMG17G taken from "Bioinformatics and genomics" (http://genome.crg.es/) was analyzed. The length of the sequence $n = 6922$, $S_+ \leq 6$, the size of the state space $A$ is 4 (0 corresponds to nucleotide A, 1 to C, 2 to G, 3 to T). We used in computer experiments the following three Markov chain models: fully-connected $s$-order Markov chain (MC($s$)), the Markov chain of order $s$ with $r$ partial connections (MC($s$, $r$)) and the Markov chain of conditional order with BMF length $L$ (MCCO($s$, $L$)). For each model the value of BIC was calculated. Results are presented in Table 1. Minimum value of BIC is marked by bold type.

Table 1: Values of BIC.

| model | BIC | model | BIC | model | BIC |
|---|---|---|---|---|---|
| MC(1) | 17792.7 | MC(4, 3) | 18162.9 | MCCO(3, 1) | 17557.5 |
| MC(2) | 17595.7 | MC(5, 1) | 18108.2 | MCCO(4, 1) | 17472.6 |
| MC(3) | 18293.1 | MC(5, 2) | 17553.8 | MCCO(4, 2) | 18205.2 |
| MC(4) | 22252.5 | MC(5, 3) | 18219.8 | MCCO(5, 1) | 17482.5 |
| MC(5) | 39894.1 | MC(5, 4) | 21896.6 | MCCO(5, 2) | 18170.6 |
| MC(6) | 116798.2 | MC(6, 1) | 18119.8 | MCCO(5, 3) | 22616.9 |
| MC(2, 1) | 18112.9 | MC(6, 2) | 17568.9 | **MCCO(6, 1)** | **17448.8** |
| MC(3, 1) | 18116.7 | MC(6, 3) | 18150.0 | MCCO(6, 2) | 18139.9 |
| MC(3, 2) | 17535.8 | MC(6, 4) | 21849.5 | MCCO(6, 3) | 22520.2 |
| MC(4, 1) | 18123.6 | MC(6, 5) | 26457.0 | MCCO(6, 4) | 41618.7 |
| MC(4, 2) | 17532.9 | | | | |

As we can see from Table 1, the most adequate model is the Markov chain of conditional order with parameters: $s = 6$, $L = 1$. Estimators for conditional orders are: $\hat{s}_0 = 4$, $\hat{s}_1 = 3$, $\hat{s}_2 = 3$, $\hat{s}_3 = 6$. Estimates for transition matrices for this MCCO(6, 1) model are:

$$\hat{Q}^{(1)} = \begin{pmatrix} 0.484 & 0.376 & 0.083 & 0.057 \\ 0.463 & 0.405 & 0.085 & 0.047 \\ 0.251 & 0.181 & 0.373 & 0.195 \\ 0.312 & 0.201 & 0.294 & 0.193 \end{pmatrix}, \quad \hat{Q}^{(2)} = \begin{pmatrix} 0.372 & 0.485 & 0.040 & 0.103 \\ 0.309 & 0.509 & 0.081 & 0.101 \\ 0.220 & 0.265 & 0.240 & 0.275 \\ 0.216 & 0.329 & 0.108 & 0.347 \end{pmatrix},$$

$$\hat{Q}^{(3)} = \begin{pmatrix} 0.254 & 0.210 & 0.270 & 0.266 \\ 0.170 & 0.370 & 0.285 & 0.175 \\ 0.205 & 0.320 & 0.320 & 0.155 \\ 0.196 & 0.253 & 0.306 & 0.245 \end{pmatrix}, \quad \hat{Q}^{(4)} = \begin{pmatrix} 0.201 & 0.181 & 0.331 & 0.287 \\ 0.099 & 0.326 & 0.276 & 0.299 \\ 0.125 & 0.230 & 0.342 & 0.303 \\ 0.125 & 0.230 & 0.342 & 0.303 \\ 0.193 & 0.206 & 0.215 & 0.386 \end{pmatrix}.$$

Let us note that the values of BIC close to the minimum are obtained for MCCO(4, 1) and MCCO(5, 1). These two models describe similar dependence to MCCO(6, 1), but they have shorter memory depth. Thus MCCO(6, 1) is chosen as the most adequate model, because the number of parameters for all three models is the same.

## 7. Conclusion

In this paper we consider a new parsimonious model for discrete-valued time series called Markov chain of conditional order. Probabilistic and statistical properties of the model are established. Ergodicity conditions and conditions under which the stationary probability distribution is uniform are found. Statistical estimators for parameters are constructedwhich and their consistency is proved. Asymptotic probability distribution of the estimators for the transition one-step probabilities is found. Statistical test for the values of transition matrices is constructed and its asymptotic power for contigual alternatives is evaluated. Computer experiments on simulated time series and on real DNA sequences are conducted.

## References

Basawa I, Prakasa Rao B (1980). *Statistical Inference for Stochastic Processes*. Academic Press, London.

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

Borovkov A (1998b). *Probability Theory*. Gordon and Breach, Amsterdam.

Buhlmann P, Wyner A (1999). "Variable Length Markov Chains." *The Annals of Statistics*, **27**(2), 480–513.

Burge C, Karlin S (1997). "Prediction of Complete Gene Structures in Human Genomic DNA." *J. Mol. Biol.*, **268**(1), 78–94.

Ching W (2004). "High-order Markov Chain Models for Categorical Data Sequences." *Naval Research Logistics*, **51**, 557–574.

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 (2005). "Robust Bayesian Prediction Under Distrtions of Prior and Conditional Distributions." *Journal of Mathematical Sciences*, **126**(1), 992–997.

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, Shlyk P (2009). "Robust Multivariate Bayesian Forecasting Under Functional Distortions in the Chi-square metric." *Journal of Statistical Planning and Inference*, **139**, 3842–3846.

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

Kharin Y, Maltsew M (2012). "Hypothesis Testing for Parameters of the Markov Chain of Conditional Order." *Proceedings of the National Academy of Sciences of Belarus. Series of physical-mathematical sciences*, **3**, 5–12 (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.

Li Y, Dong Y, Zhang H, Zhao H, Shi H, Zhao X (2010). "Spectrum Usage Prediction Based on High-order Markov Model for Cognitive Radio Networks." *10th IEEE International Conference on Computer and Information Technology*, pp. 2784–2788.

Raftery A (1985). "A Model for High-order Markov Chains." *J. Royal Statistical Society*, **B 47**, 528–539.

Roussas G (1972). *Contiguity of Probability Measures: Some Applications in Statistics*. University Press, Cambridge.

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

## Affiliation:

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

Mikhail Maltsau
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
published by the Austrian Society of Statistics
Volume 43/3-4
June 2014

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

*Submitted:* 2013-11-25
*Accepted:* 2014-03-10

Austrian Journal of Statistics
*June 2014, Volume 43/3-4, 217–228.*
*http://www.ajs.or.at/*

# On Drift Parameter Estimation in Models with Fractional Brownian Motion by Discrete Observations

**Yuliya Mishura**
Taras Shevchenko Nat. Univ. of Kyiv

**Kostiantyn Ralchenko**
Taras Shevchenko Nat. Univ. of Kyiv

**Abstract**

We study a problem of an unknown drift parameter estimation in a stochastic differential equation driven by fractional Brownian motion. We represent the likelihood ratio as a function of the observable process. The form of this representation is in general rather complicated. However, in the simplest case it can be simplified and we can discretize it to establish the a. s. convergence of the discretized version of maximum likelihood estimator to the true value of parameter. We also investigate a non-standard estimator of the drift parameter showing further its strong consistency.

*Keywords*: fractional Brownian motion, parameter estimation, stochastic differential equation, strong consistency, discretization.

## 1. Introduction

The models with long-range dependence are very popular now because they correspond to various processes in economy, finances and tele-traffic. From the mathematical point of view, long-range dependence can be modeled with the help of fractional Brownian motion with Hurst parameter $H \in \left(\frac{1}{2}, 1\right)$. More promising are so called mixed models involving both the standard Wiener process and the fractional Brownian motion. Similarly to the standard semimartingale models, the problem of parameter estimation arises immediately when we want to adapt the model with long-range dependence to the specific situation. In particular, the problem of the drift parameter estimation in the diffusion model with fractional Brownian motion is rather important. The standard maximum likelihood estimator was considered by many authors, see, e.g., Mishura (2008) and Prakasa Rao (2010). It is constructed by continuous observations on the whole interval. Asymptotic properties when the interval of observations increases to the whole half-axis, were established. However, in practical considerations the observations are never continuous. So, the problem of the discretization of the estimate appears. Some papers are devoted to the parameter estimation for the models with fBm and discrete observations, see, e.g., Hu and Nualart (2010), Xiao, Zhang, and Xu (2011a), Xiao, Zhang, and Zhang (2011b), Bishwal (2011), Tanaka (2013), Hu and Song (2013), Zhang, Xiao, Zhang, and Niu (2014) but only restricted classes of models, basically linear models were considered. The situation is such that in the general case the maximum-likelihood estimator has a very

complicated representation via the observed process and the dicretized version does not allow reasonable form for calculations. Therefore, we have to propose some non-standard approach to construct strongly consistent drift parameter estimators for the discrete observations of the models with long-range dependence. One of such approaches was demonstrated in Mishura, Ral'chenko, Seleznev, and Shevchenko (2014), where some specific discretized estimators were proposed. In the present paper we propose two approaches. One of them consists in direct discretization of maximum-likelihood parameter estimator, however, only for the case when drift and diffusion coefficients coincide. It is one of the cases when the discretization leads to the reasonable form of the estimator. Another approach is to discretize the non-standard drift parameter estimator that was introduced in Kozachenko, Melnikov, and Mishura (2013). This also leads to the consistent estimator. Strong consistency is established for both estimators and illustrated with some simulations.

## 2. Maximum-likelihood estimation

### 2.1. Model description

Let $B^H = \{B_t^H, t \geq 0\}$ be a fractional Brownian motion with Hurst index $H \in (1/2, 1)$, defined on the probability space $(\Omega, \mathcal{F}, \mathsf{P})$. Denote by $(\mathcal{F}_t)_{t\geq 0}$ the filtration generated by $B^H$. Consider the stochastic differential equation driven by fractional Brownian motion $B^H$:

$$\begin{aligned} dX_t &= \theta a(t, X_t)dt + b(t, X_t)dB_t^H, \quad 0 \leq t \leq T, \quad T > 0, \\ X\big|_{t=0} &= X_0 \in \mathbb{R}. \end{aligned} \tag{1}$$

Here $\theta \in \mathbb{R}$ is unknown parameter to be estimated.

Suppose that the following assumptions hold:

(I) there exist positive constants $C_1$, $C_2$ such that for all $t \in [0, T]$, $x, y \in \mathbb{R}$

$$\begin{gathered} |a(t,x) - a(t,y)| + |b(t,x) - b(t,y)| \leq C_1 |x - y|, \\ |a(t,x)| + |b(t,x)| \leq C_2(1 + |x|); \end{gathered}$$

(II) there exist constants $C_3 > 0$ and $\rho \in \left(\frac{1}{H} - 1, 1\right)$ such that for all $t \in [0, T]$, $x, y \in \mathbb{R}$

$$\left|b'_x(t,x) - b'_y(t,y)\right| \leq C_3 |x - y|^\rho;$$

(III) there exist constants $C_4 > 0$ and $\mu \in (1 - H, 1)$ such that for all $t, s \in [0, T]$, $x \in \mathbb{R}$

$$|b(t,x) - b(s,x)| + \left|b'_x(t,x) - b'_x(s,x)\right| \leq C_4 |t - s|^\mu.$$

According to (Nualart and Rascanu 2001, Theorem 2.1), under the conditions (I)–(III) there exists a unique solution $X$ of the stochastic equation (1).

In addition, suppose that the following conditions hold:

(IV) $b(t,x) \neq 0$;

(V) $a, b \in C([0,\infty) \times \mathbb{R})$.

Denote $\alpha = H - \frac{1}{2}$, $\widetilde{\alpha} = (1 - 2\alpha)^{-1}$, $C_H = \left(\frac{\Gamma(2-2\alpha)}{2H\Gamma(1-\alpha)^3\Gamma(\alpha+1)}\right)^{\frac{1}{2}}$, $l_H(t,s) = C_H s^{-\alpha}(t - s)^{-\alpha} I_{\{0<s<t\}}$, $\psi(t,x) = \frac{a(t,x)}{b(t,x)}$, $\varphi(t) = \psi(t, X_t)$, $I(t) = \int_0^t l_H(t,s)\varphi(s)ds$. Under the conditions (I), (III)–(V) $\varphi(t), t \in [0, T]$ is a continuous process with probability 1. Hence, it is Lebesgue integrable and for each $t \in [0, T]$ there exists an integral $\int_0^t l_H(t,s)\varphi(s)ds$.

Consider the new process $\widehat{B}_t^H := B_t^H + \theta \int_0^t \varphi(s)ds$. Suppose that the following assumptions hold.

(VI) There exists such function $\delta$ that belongs to $L_1[0,t]$ for all $t \in [0,T]$ a. s. and satisfies the equation

$$\theta \int_0^t l_H(t,s)\varphi(s)ds = (\widetilde{\alpha})^{-1/2} \int_0^t \delta_s ds;$$

(VII) $\mathsf{E}\int_0^t s^{2\alpha}\delta_s^2 ds < \infty$, $t \in [0,T]$;

(VIII) $\mathsf{E}\exp\left\{L_t - \frac{1}{2}\langle L\rangle_t\right\} = 1$, where $L_t = \int_0^t s^{\alpha}\delta_s d\widehat{B}_s$, and $\widehat{B}$ is Wiener process with respect to probability measure $\mathsf{P}_0(t)$ corresponding to the zero drift such that

$$\int_0^t l_H(t,s)\, d\widehat{B}_s^H = \widetilde{\alpha}^{-1/2} \int_0^t s^{-\alpha}\, d\widehat{B}_s.$$

(The existence of this Wiener process follows from the representation of fractional Brownian motion via Wiener process on a finite interval introduced in Norros, Valkeila, and Virtamo (1999).)

Then the likelihood ratio $\frac{d\mathsf{P}_\theta(t)}{d\mathsf{P}_0(t)}$ for the probability measure $\mathsf{P}_\theta(t)$ corresponding to our model and probability measure $\mathsf{P}_0(t)$ corresponding to the model with zero drift is equal to

$$\frac{d\mathsf{P}_\theta(t)}{d\mathsf{P}_0(t)} = \exp\left\{L_t - \frac{1}{2}\langle L\rangle_t\right\}.$$

Note that $L_t$ is a square-integrable martingale. Now we present likelihood ratio as a function of the observed process $X_t$.

### 2.2. The explicit form for the likelihood ratio and a discretized version of MLE

We can present likelihood ratio as a function of the observed process $X_t$.

$$L_t = \int_0^t s^{\alpha}\delta_s d\widehat{B}_s = \int_0^t s^{2\alpha}\delta_s dY_s, \tag{2}$$

where

$$Y_s = \int_0^s u^{-\alpha} d\widehat{B}_u = \widetilde{\alpha}^{1/2} \int_0^s l_H(s,u) b^{-1}(u, X_u) dX_u, \tag{3}$$

$$\delta_s = \theta\widetilde{\alpha}^{1/2}\left(\int_0^s l_H(s,u)\varphi(u)du\right)' = C_H\theta\widetilde{\alpha}^{1/2}\left(\int_0^s (s-u)^{-\alpha}u^{-\alpha}\varphi(u)du\right)' \tag{4}$$

or

$$\delta_s = C_H\theta\widetilde{\alpha}^{1/2}\left(\frac{\varphi(s)}{s^{2\alpha}} + \alpha\int_0^s \frac{s^{-\alpha}\varphi(s) - u^{-\alpha}\varphi(u)}{(s-u)^{\alpha+1}}du\right). \tag{5}$$

According to (Mishura 2008, formula (6.3.13)), the maximum-likelihood estimator has the form

$$\hat{\theta}_t^{(1)} = \frac{\widetilde{\alpha}^{-1/2}\int_0^t s^{\alpha} I'(s) d\widehat{B}_s}{\int_0^t s^{2\alpha}(I'(s))^2 ds}.$$

Since $I'(s) = \delta_s\theta^{-1}\widetilde{\alpha}^{-1/2}$, we obtain

$$\hat{\theta}_t^{(1)} = \frac{\theta L_t}{\int_0^t s^{2\alpha}\delta_s^2 ds}. \tag{6}$$

Using (2), (3), (5) and the definition of the kernel $l_H(t,s)$ we can write

$$\hat{\theta}_t^{(1)} = \frac{\theta \int_0^t s^{2\alpha}\delta_s dY_s}{\int_0^t s^{2\alpha}\delta_s^2 ds} = \frac{\int_0^t \left(\varphi(s) + \alpha s^{2\alpha}\int_0^s \frac{s^{-\alpha}\varphi(s) - u^{-\alpha}\varphi(u)}{(s-u)^{\alpha+1}}du\right) d\left(\int_0^s v^{-\alpha}(s-v)^{-\alpha}b^{-1}(v,X_v)dX_v\right)}{\int_0^t s^{2\alpha}\left(\frac{\varphi(s)}{s^{2\alpha}} + \alpha\int_0^s \frac{s^{-\alpha}\varphi(s) - u^{-\alpha}\varphi(u)}{(s-u)^{\alpha+1}}du\right)^2 ds}.$$

*Remark* 1. According to (Mishura 2008, Theorem 6.3.3), under assumptions (I)–(VIII) and

$$\int_0^\infty s^{2\alpha}(I'(s))^2 ds = \infty \quad \text{a. s.}$$

$\hat{\theta}_T^{(1)} \xrightarrow{\mathsf{P1}} \theta$, $T \to \infty$.

Let $t_k^n = \frac{k}{2^n}$, $k = 0, 1, 2, \ldots, 2^{2n}$. We can define a discretized version of the maximum-likelihood estimator

$$\hat{\theta}_n^{(2)} := \frac{\sum_{k=0}^{2^{2n}-1}\left(\varphi(t_k^n) + \alpha(t_k^n)^{2\alpha}\sum_{i=1}^{k-1}\frac{(t_k^n)^{-\alpha}\varphi(t_k^n) - (t_i^n)^{-\alpha}\varphi(t_i^n)}{(t_k^n - t_i^n)^{\alpha+1}}\frac{1}{2^n}\right)\left(\widetilde{Y}_{t_{k+1}^n} - \widetilde{Y}_{t_k^n}\right)}{\sum_{k=0}^{2^{2n}-1}(t_k^n)^{2\alpha}\left(\frac{\varphi(t_k^n)}{(t_k^n)^{2\alpha}} + \alpha\sum_{i=1}^{k-1}\frac{(t_k^n)^{-\alpha}\varphi(t_k^n) - t_i^{-\alpha}\varphi(t_i^n)}{(t_k^n - t_i^n)^{\alpha+1}}\frac{1}{2^n}\right)^2\frac{1}{2^n}} \tag{7}$$

where

$$\widetilde{Y}_{t_k} = \sum_{i=1}^{k-1}(t_i^n)^{-\alpha}(t_k^n - t_i^n)^{-\alpha}b^{-1}(t_i^n, X_{t_i^n})\left(X_{t_{i+1}^n} - X_{t_i^n}\right).$$

In the general case formula (7) is not suitable for applications because it involves a lot of weakly singular kernels and it is quite impossible to get its convergence to the true value of the parameter. But even if we get the convergence, the simulation error will be so great that annihilate our efforts in discretization. In order to avoid this technical difficulties, we start with the simplest case.

### 2.3. Estimation in the case $a = b$

Consider an equation

$$dX_t = \theta b(X_t)dt + b(X_t)dB_t^H. \tag{8}$$

In this case $\varphi \equiv 1$. So we get from (4) that

$$\delta_s = C_H\theta\widetilde{\alpha}^{1/2}\left(\mathrm{B}(1-\alpha, 1-\alpha)s^{1-2\alpha}\right)' = C_H\theta\widetilde{\alpha}^{-1/2}\mathrm{B}(1-\alpha,1-\alpha)s^{-2\alpha}.$$

Then (2) and (3) imply

$$L_t = C_H\theta\widetilde{\alpha}^{-1/2}\mathrm{B}(1-\alpha,1-\alpha)Y_t = C_H\theta\mathrm{B}(1-\alpha,1-\alpha)\int_0^t l_H(t,s)b^{-1}(X_s)dX_s.$$

Therefore the maximum-likelihood estimator (6) can be written as follows:

$$\hat{\theta}_t^{(1)} = \frac{\int_0^t l_H(t,s)b^{-1}(X_s)dX_s}{C_H\mathrm{B}(1-\alpha,1-\alpha)t^{1-2\alpha}}. \tag{9}$$

It follows from (8) that

$$\hat{\theta}_t^{(1)} = \frac{\theta\int_0^t l_H(t,s)ds + \int_0^t l_H(t,s)dB_s^H}{C_H\mathrm{B}(1-\alpha,1-\alpha)t^{1-2\alpha}} = \theta + \frac{\int_0^t l_H(t,s)dB_s^H}{C_H\mathrm{B}(1-\alpha,1-\alpha)t^{1-2\alpha}}.$$

Since $\int_0^t l_H(t,s)dB_s^H$ is a square integrable martingale with angle bracket $t^{1-2\alpha} \to \infty$ we see that $\frac{\int_0^t l_H(t,s)dB_s^H}{t^{1-2\alpha}} \xrightarrow[n\to\infty]{\mathsf{P1}} 0$. Hence the estimator $\hat{\theta}_t^{(1)}$ is strongly consistent.

Now we consider an estimator

$$\hat{\theta}_n^{(3)} = \frac{\sum_{k=1}^{2^{2n}-1}(t_k^n)^{-\alpha}(2^n - t_k^n)^{-\alpha} b^{-1}\left(X_{t_{k-1}^n}\right)\left(X_{t_k^n} - X_{t_{k-1}^n}\right)}{\mathrm{B}(1-\alpha, 1-\alpha)2^{n(1-2\alpha)}},$$

where $t_k^n = \frac{k}{2^n}$, $k = 0, 1, \ldots, 2^{2n}$. This estimator is a discretized version of the estimator (9).

**Theorem 1.** *Suppose that there exist positive constants $C_1$, $C_2$, $C_3$, $C_5$ and $\rho \in (1/H - 1, 1]$, such that*

*(a)* $|b(x) - b(y)| \le C_1 |x - y|$ *for all* $x, y \in \mathbb{R}$,

*(b)* $C_5 \le |b(x)| \le C_2(1 + |x|)$ *for all* $x \in \mathbb{R}$,

*(c)* $|b'(x) - b'(y)| \le C_3 |x - y|^\rho$ *for all* $x, y \in \mathbb{R}$

*Then* $\hat{\theta}_n^{(3)} \xrightarrow{\mathsf{P1}} \theta, n \to \infty$. *Moreover, for any* $\beta \in (1/2, H)$ *and* $\gamma > 1/2$ *there exists a random variable* $\eta = \eta_{\beta,\gamma}$ *with all finite moments such that* $\left|\hat{\theta}_n^{(3)} - \theta\right| \le \eta n^{\kappa+\gamma} 2^{-\tau n}$, *where* $\kappa = \gamma/\beta$, $\tau = (1 - H) \wedge (2\beta - 1)$.

*Proof.* It follows from (8) that

$$\begin{aligned} X_{t_k^n} - X_{t_{k-1}^n} &= \theta \int_{t_{k-1}^n}^{t_k^n} b(X_v)dv + \int_{t_{k-1}^n}^{t_k^n} b(X_v)dB_v^H \\ &= \theta \int_{t_{k-1}^n}^{t_k^n} b\left(X_{t_{k-1}^n}\right) dv + \theta \int_{t_{k-1}^n}^{t_k^n} \left(b(X_v) - b\left(X_{t_{k-1}^n}\right)\right) dv \\ &\quad + \int_{t_{k-1}^n}^{t_k^n} \left(b(X_v) - b\left(X_{t_{k-1}^n}\right)\right) dB_v^H + \int_{t_{k-1}^n}^{t_k^n} b\left(X_{t_{k-1}^n}\right) dB_v^H. \end{aligned} \tag{10}$$

Then

$$\hat{\theta}_n^{(3)} = \frac{\theta A_n + \theta B_n + D_n + E_n}{\mathrm{B}(1-\alpha, 1-\alpha)}, \tag{11}$$

where

$$\begin{aligned} A_n &= 2^{n(2\alpha-2)} \sum_{k=1}^{2^{2n}-1} (t_k^n)^{-\alpha}(2^n - t_k^n)^{-\alpha}, \\ B_n &= 2^{n(2\alpha-1)} \sum_{k=1}^{2^{2n}-1} (t_k^n)^{-\alpha}(2^n - t_k^n)^{-\alpha} b^{-1}\left(X_{t_{k-1}^n}\right) \int_{t_{k-1}^n}^{t_k^n} \left(b(X_v) - b\left(X_{t_{k-1}^n}\right)\right) dv, \\ D_n &= 2^{n(2\alpha-1)} \sum_{k=1}^{2^{2n}-1} (t_k^n)^{-\alpha}(2^n - t_k^n)^{-\alpha} b^{-1}\left(X_{t_{k-1}^n}\right) \int_{t_{k-1}^n}^{t_k^n} \left(b(X_v) - b\left(X_{t_{k-1}^n}\right)\right) dB_v^H, \\ E_n &= 2^{n(2\alpha-1)} \sum_{k=1}^{2^{2n}-1} (t_k^n)^{-\alpha}(2^n - t_k^n)^{-\alpha} \left(B_{t_k^n}^H - B_{t_{k-1}^n}^H\right). \end{aligned}$$

It is not hard to show that the sequence

$$A_n = \sum_{k=1}^{2^{2n}-1} \left(\frac{k}{2^{2n}}\right)^{-\alpha} \left(1 - \frac{k}{2^{2n}}\right)^{-\alpha} \frac{1}{2^{2n}}$$

converges to $\int_0^1 x^{-\alpha}(1-x)^{-\alpha}dx = \mathrm{B}(1-\alpha, 1-\alpha)$, moreover,

$$|A_n - \mathrm{B}(1-\alpha, 1-\alpha)| \le c_1 2^{-2n(1-\alpha)} \tag{12}$$

where $c_1$ is a constant. Indeed, $h(x) = x^{-\alpha}(1-x)^{-\alpha}$ is a decreasing function when $x \in \left(0, \frac{1}{2}\right]$, then

$$\int_0^{\frac{1}{2}} h(x)dx = \sum_{k=0}^{2^{2n-1}-1} \int_{\frac{k}{2^{2n}}}^{\frac{k+1}{2^{2n}}} h(x)dx < \int_0^{\frac{1}{2^{2n}}} h(x)dx + \sum_{k=1}^{2^{2n-1}} h\left(\frac{k}{2^{2n}}\right)\frac{1}{2^{2n}}.$$

On the other hand,

$$\int_0^{\frac{1}{2}} h(x)dx = \sum_{k=1}^{2^{2n-1}} \int_{\frac{k-1}{2^{2n}}}^{\frac{k}{2^{2n}}} h(x)dx > \sum_{k=1}^{2^{2n-1}} h\left(\frac{k}{2^{2n}}\right)\frac{1}{2^{2n}}.$$

So

$$0 < \int_0^{\frac{1}{2}} h(x)dx - \sum_{k=1}^{2^{2n-1}} h\left(\frac{k}{2^{2n}}\right)\frac{1}{2^{2n}} < \int_0^{\frac{1}{2^{2n}}} h(x)dx \le \left(1-\frac{1}{2^{2n}}\right)^{-\alpha}\frac{2^{-2n(1-\alpha)}}{1-\alpha}. \tag{13}$$

Similarly one can show that

$$0 < \int_{\frac{1}{2}}^1 h(x)dx - \sum_{k=2^{2n-1}+1}^{2^{2n}-1} h\left(\frac{k}{2^{2n}}\right)\frac{1}{2^{2n}} < \left(1-\frac{1}{2^{2n}}\right)^{-\alpha}\frac{2^{-2n(1-\alpha)}}{1-\alpha}. \tag{14}$$

Combining (13) and (14), we obtain (12).

By (Mishura *et al.* 2014, Lemma 2), there exist random variables $\xi_1$ and $\xi_2$ with all finite moments such that for all $n \ge 1$ and $k = 1, 2, \ldots, 2^{2n}$

$$\left|\int_{t_{k-1}^n}^{t_k^n} \left(b(X_u) - b(X_{t_{k-1}^n})\right) du\right| \le \xi_1 n^\kappa 2^{-n(\beta+1)}$$

and

$$\left|\int_{t_{k-1}^n}^{t_k^n} \left(b(X_u) - b(X_{t_{k-1}^n})\right) dB_u^H\right| \le \xi_2 n^{\gamma+\kappa} 2^{-2n\beta},$$

Then

$$|B_n| \le C_5^{-1}\xi_1 n^\kappa 2^{n(2\alpha-\beta-2)} \sum_{k=1}^{2^{2n}-1} (t_k^n)^{-\alpha}(2^n - t_k^n)^{-\alpha} = C_5^{-1}\xi_1 n^\kappa 2^{-n\beta} A_n \le c_2\xi_1 n^\kappa 2^{-n\beta}; \tag{15}$$

$$|D_n| \le C_5^{-1}\xi_2 n^{\gamma+\kappa} 2^{n(2\alpha-1-2\beta)} \sum_{k=1}^{2^{2n}-1} (t_k^n)^{-\alpha}(2^n - t_k^n)^{-\alpha} \le c_2\xi_2 n^{\gamma+\kappa} 2^{-n(2\beta-1)}. \tag{16}$$

Finally we estimate $E_n$. Start by writing

$$\mathsf{E}\left[E_n^2\right] = 2^{2n(2\alpha-1)}\mathsf{E}\left[\left(\sum_{k=1}^{2^{2n}-1}\int_{t_{k-1}^n}^{t_k^n} (t_k^n)^{-\alpha}\left(2^n - t_k^n\right)^{-\alpha} dB_s^H\right)^2\right].$$

According to (Mishura 2008, Corollary 1.9.4), for $f \in L_{1/H}[0,t]$ there exist a constant $C_H > 0$ such that

$$\mathsf{E}\left[\left(\int_0^t f(s)dB_s^H\right)^2\right] \le C_H\left(\int_0^t |f(s)|^{1/H} ds\right)^{2H}.$$

Hence,

$$\begin{aligned}\mathsf{E}\left[E_n^2\right] &\le c_3 2^{2n(2\alpha-1)}\left(\sum_{k=1}^{2^{2n}-1}\int_{t_{k-1}^n}^{t_k^n} (t_k^n)^{-\alpha/H}\left(2^n - t_k^n\right)^{-\alpha/H} ds\right)^{2H} \\ &= c_3 2^{2n(H-1)}\left(\sum_{k=1}^{2^{2n}-1}\left(\frac{k}{2^{2n}}\right)^{-\alpha/H}\left(1-\frac{k}{2^{2n}}\right)^{-\alpha/H}\frac{1}{2^{2n}}\right)^{2H}.\end{aligned}$$

As above,

$$\sum_{k=1}^{2^{2n}-1}\left(\frac{k}{2^{2n}}\right)^{-\alpha/H}\left(1-\frac{k}{2^{2n}}\right)^{-\alpha/H}\frac{1}{2^{2n}}\to \mathrm{B}\left(1-\alpha/H,1-\alpha/H\right),\ n\to\infty,$$

which implies that $\mathsf{E}\left[E_n^2\right]\le c_4 2^{2n(H-1)}$. Since $E_n$ is Gaussian, we have $\mathsf{E}\left[|E_n|^p\right]\le c_5(p)2^{pn(H-1)}$ for any $p\ge 1$. Therefore, for any $\nu>1$

$$\mathsf{E}\left[\sum_{n=1}^{\infty}\frac{|E_n|^p}{n^\nu 2^{pn(H-1)}}\right]=\sum_{n=1}^{\infty}\frac{\mathsf{E}\left[|E_n|^p\right]}{n^\nu 2^{pn(H-1)}}\le c_5(p)\sum_{n=1}^{\infty}n^{-\nu}<\infty.$$

Consequently,

$$\xi_3:=\sup_{n\ge 1}\frac{|E_n|}{n^{\nu/p}2^{n(H-1)}}<\infty$$

almost surely, moreover, by Fernique's theorem, all moments of $\xi_3$ are finite. Therefore,

$$|E_n|\le \xi_3 n^\delta 2^{-n(1-H)}, \tag{17}$$

where $\delta>0$ can be taken arbitrarily small.

Combining (11), (12) and (15)–(17) we obtain

$$\begin{aligned}\left|\hat\theta_n^{(3)}-\theta\right|&\le\frac{\theta\left|A_n-\mathrm{B}(1-\alpha,1-\alpha)\right|+\theta\left|B_n\right|+\left|D_n\right|+\left|E_n\right|}{\mathrm{B}(1-\alpha,1-\alpha)}\\&\le\frac{\theta c_1 2^{-2n(1-\alpha)}+\theta c_2\xi_1 n^\kappa 2^{-n\beta}+c_2\xi_2 n^{\gamma+\kappa}2^{-n(2\beta-1)}+\xi_3 n^\delta 2^{-n(1-H)}}{\mathrm{B}(1-\alpha,1-\alpha)}.\end{aligned}$$

Note that $2(1-\alpha)=3-2H>1-H\ge\tau$ and $\beta>1/2>1-H\ge\tau$. Then,

$$\left|\hat\theta_n^{(3)}-\theta\right|\le\eta n^{\kappa+\gamma}2^{-\tau n},$$

where $\eta\le c_6(\theta)(1+\xi_1+\xi_2+\xi_3)$. □

## 3. Non-standard estimators

In the paper Kozachenko *et al.* (2013) the following non-standard estimator for $\theta$ in the equation (1) was considered:

$$\hat\theta_t^{(4)}=\frac{\int_0^t a(s,X_s)b^{-2}(s,X_s)dX_s}{\int_0^t a^2(s,X_s)b^{-2}(s,X_s)ds}.$$

According to (Kozachenko *et al.* 2013, Theorem 4), if the assumptions (I)–(IV), (VI)–(VII) hold and there exist such $\beta>1-H$ and $p>1$ that

$$\frac{T^{H+\beta-1}(\log T)^p\int_0^T\left|(D_{0+}^\beta\varphi)(s)\right|ds}{\int_0^T\varphi_s^2 ds}\to 0\ \text{ a.s. as } T\to\infty,$$

then the estimator $\hat\theta_T^{(4)}$ is well-defined and strongly consistent as $T\to\infty$.

We define a discretized version of $\hat\theta_T^{(4)}$ for the equation

$$X_t=X_0+\theta\int_0^t a(X_s)ds+\int_0^t b(X_s)dB_s^H. \tag{18}$$

Put

$$\hat{\theta}_n^{(5)} := \frac{\sum_{k=1}^{2^{2n}} a\left(X_{t_{k-1}^n}\right) b^{-2}\left(X_{t_{k-1}^n}\right)\left(X_{t_k^n} - X_{t_{k-1}^n}\right)}{\sum_{k=1}^{2^{2n}} a^2\left(X_{t_{k-1}^n}\right) b^{-2}\left(X_{t_{k-1}^n}\right) \frac{1}{2^n}},$$

$$\widehat{\varphi}_n(t) := \sum_{k=0}^{2^{2n}-1} \varphi(t_k^n) I_{[t_k^n, t_{k+1}^n)}(t).$$

**Theorem 2.** *Suppose that there exist positive constants $C_1$, $C_3$, $C_5$, $C_6$ $\rho \in (1/H - 1, 1]$, $\beta > 1 - H$, $p > 1$ such that*

*(a)* $|a(x) - a(y)| + |b(x) - b(y)| \le C_1 |x - y|$ *for all* $x, y \in \mathbb{R}$,

*(b)* $C_5 \le |a(x)| \le C_6$, $C_5 \le |b(x)| \le C_6$ *for all* $x \in \mathbb{R}$,

*(c)* $|b'(x) - b'(y)| \le C_3 |x - y|^\rho$ *for all* $x, y \in \mathbb{R}$,

*(d)* $\frac{2^{n(H+\beta)} n^p \int_0^{2^n} \left|\left(D_{0+}^{\beta} \widehat{\varphi}_n\right)(s)\right| ds}{\sum_{k=1}^{2^{2n}} \varphi^2(t_{k-1}^n)} \to 0$ *a. s. as* $n \to \infty$.

*Then with probability one,* $\hat{\theta}_n^{(5)} \to \theta, n \to \infty$.

*Proof.* It follows from (18) that

$$\begin{aligned} X_{t_k^n} - X_{t_{k-1}^n} &= \theta \int_{t_{k-1}^n}^{t_k^n} a(X_v) dv + \int_{t_{k-1}^n}^{t_k^n} b(X_v) dB_v^H \\ &= \theta \int_{t_{k-1}^n}^{t_k^n} a\left(X_{t_{k-1}^n}\right) dv + \theta \int_{t_{k-1}^n}^{t_k^n} \left(a(X_v) - a\left(X_{t_{k-1}^n}\right)\right) dv \\ &\quad + \int_{t_{k-1}^n}^{t_k^n} \left(b(X_v) - b\left(X_{t_{k-1}^n}\right)\right) dB_v^H + \int_{t_{k-1}^n}^{t_k^n} b\left(X_{t_{k-1}^n}\right) dB_v^H. \end{aligned} \tag{19}$$

Then

$$\hat{\theta}_n^{(5)} = \theta + \frac{\theta B_n + E_n + D_n}{A_n},$$

where

$$\begin{aligned} A_n &= 2^{-n} \sum_{k=1}^{2^{2n}} \varphi^2(t_{k-1}^n), \\ B_n &= \sum_{k=1}^{2^{2n}} a\left(X_{t_{k-1}^n}\right) b^{-2}\left(X_{t_{k-1}^n}\right) \int_{t_{k-1}^n}^{t_k^n} \left(a(X_v) - a\left(X_{t_{k-1}^n}\right)\right) dv, \\ E_n &= \sum_{k=1}^{2^{2n}} a\left(X_{t_{k-1}^n}\right) b^{-2}\left(X_{t_{k-1}^n}\right) \int_{t_{k-1}^n}^{t_k^n} \left(b(X_v) - b\left(X_{t_{k-1}^n}\right)\right) dB_v^H, \\ D_n &= \sum_{k=1}^{2^{2n}} \varphi(t_{k-1}^n) \left(B_{t_k^n}^H - B_{t_{k-1}^n}^H\right). \end{aligned}$$

$D_n$ can be represented in the form

$$D_n = \int_0^{2^n} \widehat{\varphi}_n(s) dB_s^H.$$

Applying (Kozachenko *et al.* 2013, Theorem 3) we can estimate

$$\sup_{0 \le t \le 2^n} \left|\left(D_{2^n-}^{1-\beta} B_{2^n-}^H\right)(t)\right| \le \xi(p) 2^{n(H+\beta-1)} n^p (\log 2)^p.$$

Therefore

$$|D_n| \le \sup_{0\le t\le 2^n} \left|\left(D^{1-\beta}_{2^n-}B^H_{2^n-}\right)(s)\right| \cdot \int_0^{2^n} \left|\left(D^{\beta}_{0+}\widehat{\varphi}_n\right)(s)\right| ds$$
$$\le \xi(p)(\log 2)^p 2^{n(H+\beta-1)} n^p \int_0^{2^n} \left|\left(D^{\beta}_{0+}\widehat{\varphi}_n\right)(s)\right| ds.$$

Then

$$\left|\frac{D_n}{A_n}\right| \le \xi(p)(\log 2)^p \frac{2^{n(H+\beta)} n^p \int_0^{2^n} \left|\left(D^{\beta}_{0+}\widehat{\varphi}_n\right)(s)\right| ds}{\sum_{k=1}^{2^{2n}} \varphi^2(t^n_{k-1})} \to 0 \text{ a. s. as } n \to \infty.$$

Using the condition (b) we can write

$$\left|\frac{B_n}{A_n}\right| \le C_6^{-1} 2^{-n} \sum_{k=1}^{2^{2n}} \int_{t^n_{k-1}}^{t^n_k} \left|a(X_v) - a\left(X_{t^n_{k-1}}\right)\right| dv,$$
$$\left|\frac{E_n}{A_n}\right| \le C_6^{-1} 2^{-n} \sum_{k=1}^{2^{2n}} \int_{t^n_{k-1}}^{t^n_k} \left|b(X_v) - b\left(X_{t^n_{k-1}}\right)\right| dB^H_v.$$

It now follows from (Mishura *et al.* 2014, Lemma 2) that $\left|\frac{B_n}{A_n}\right| \to 0$, $\left|\frac{E_n}{A_n}\right| \to 0$ as $n \to \infty$. □

*Example* 1. Consider the model (8):

$$dX_t = \theta b(X_t)dt + b(X_t)dB^H_t.$$

Suppose that there exist positive constants $C_1$, $C_3$, $C_5$, $C_6$ $\rho \in (1/H - 1, 1]$, $\beta > 1 - H$, $p > 1$ such that

(a) $|b(x) - b(y)| \le C_1 |x - y|$ for all $x, y \in \mathbb{R}$,

(b) $C_5 \le |b(x)| \le C_6$ for all $x \in \mathbb{R}$,

(c) $|b'(x) - b'(y)| \le C_3 |x - y|^\rho$ for all $x, y \in \mathbb{R}$.

In this case the non-standard estimator $\hat{\theta}^{(5)}_n$ has the form

$$\hat{\theta}^{(6)}_n = 2^{-n} \sum_{k=1}^{2^{2n}} b^{-1}\left(X_{t^n_{k-1}}\right)\left(X_{t^n_k} - X_{t^n_{k-1}}\right), \tag{20}$$

$\widehat{\varphi}_n(t) = 1$. Then $\left(D^{\beta}_{0+}\widehat{\varphi}_n\right)(s) = \frac{1}{\Gamma(1-\beta)} \cdot s^{-\beta}$ and

$$\frac{2^{n(H+\beta)} n^p \int_0^{2^n} \left|\left(D^{\beta}_{0+}\widehat{\varphi}_n\right)(s)\right| ds}{\sum_{k=1}^{2^{2n}} \varphi^2(t^n_{k-1})} = \frac{n^p}{\Gamma(2-\beta) \cdot 2^{n(1-H)}} \to 0, \quad n \to \infty.$$

Consequently the conditions of Theorem 2 are satisfied and the estimator (20) is strongly consistent.

## 4. Simulations

In this section we illustrate quality of the estimators with the help of simulation experiments. We consider the equation (18) with $X_0 = 1$, $\theta = 1$. For each set of parameters, we simulate 100 trajectories of the solution. In the case $a = b$ we compute the average relative error

Table 1: $a(x) = b(x) = \sin x + 2$.

| $n$ | $H = 0.6$ | | $H = 0.7$ | | $H = 0.8$ | | $H = 0.9$ | |
|---|---|---|---|---|---|---|---|---|
| | $\delta_n^{(3)}$ | $\delta_n^{(5)}$ | $\delta_n^{(3)}$ | $\delta_n^{(5)}$ | $\delta_n^{(3)}$ | $\delta_n^{(5)}$ | $\delta_n^{(3)}$ | $\delta_n^{(5)}$ |
| 3 | 0.0929 | 0.0967 | 0.0935 | 0.1015 | 0.0926 | 0.0956 | 0.1144 | 0.0935 |
| 4 | 0.0512 | 0.0512 | 0.0510 | 0.0509 | 0.0497 | 0.0471 | 0.0522 | 0.0495 |
| 5 | 0.0262 | 0.0258 | 0.0264 | 0.0258 | 0.0251 | 0.0244 | 0.0287 | 0.0261 |
| 6 | 0.0122 | 0.0121 | 0.0120 | 0.0120 | 0.0123 | 0.0127 | 0.0125 | 0.0107 |

Table 2: $a(x) = b(x) = \cos x + 2$.

| $n$ | $H = 0.6$ | | $H = 0.7$ | | $H = 0.8$ | | $H = 0.9$ | |
|---|---|---|---|---|---|---|---|---|
| | $\delta_n^{(3)}$ | $\delta_n^{(5)}$ | $\delta_n^{(3)}$ | $\delta_n^{(5)}$ | $\delta_n^{(3)}$ | $\delta_n^{(5)}$ | $\delta_n^{(3)}$ | $\delta_n^{(5)}$ |
| 3 | 0.0898 | 0.0891 | 0.1006 | 0.0964 | 0.1021 | 0.0971 | 0.1088 | 0.0935 |
| 4 | 0.0567 | 0.0568 | 0.0501 | 0.0474 | 0.0544 | 0.0517 | 0.0611 | 0.0521 |
| 5 | 0.0228 | 0.0227 | 0.0276 | 0.0277 | 0.0255 | 0.0245 | 0.0280 | 0.0244 |
| 6 | 0.0139 | 0.0138 | 0.0122 | 0.0123 | 0.0130 | 0.0133 | 0.0140 | 0.0137 |

Table 3: $a(x) = b(x) = \frac{1}{1+x^2}$.

| $n$ | $H = 0.6$ | | $H = 0.7$ | | $H = 0.8$ | | $H = 0.9$ | |
|---|---|---|---|---|---|---|---|---|
| | $\delta_n^{(3)}$ | $\delta_n^{(5)}$ | $\delta_n^{(3)}$ | $\delta_n^{(5)}$ | $\delta_n^{(3)}$ | $\delta_n^{(5)}$ | $\delta_n^{(3)}$ | $\delta_n^{(5)}$ |
| 3 | 0.0969 | 0.0950 | 0.1027 | 0.1003 | 0.1079 | 0.0963 | 0.1198 | 0.0922 |
| 4 | 0.0467 | 0.0457 | 0.0473 | 0.0477 | 0.0485 | 0.0444 | 0.0489 | 0.0437 |
| 5 | 0.0257 | 0.0259 | 0.0289 | 0.0282 | 0.0235 | 0.0245 | 0.0307 | 0.0263 |
| 6 | 0.0123 | 0.0123 | 0.0129 | 0.0128 | 0.0120 | 0.0116 | 0.0130 | 0.0114 |

Table 4: $a(x) = b(x) = 1$.

| $n$ | $H = 0.6$ | | $H = 0.7$ | | $H = 0.8$ | | $H = 0.9$ | |
|---|---|---|---|---|---|---|---|---|
| | $\delta_n^{(3)}$ | $\delta_n^{(5)}$ | $\delta_n^{(3)}$ | $\delta_n^{(5)}$ | $\delta_n^{(3)}$ | $\delta_n^{(5)}$ | $\delta_n^{(3)}$ | $\delta_n^{(5)}$ |
| 3 | 0.0947 | 0.0925 | 0.1037 | 0.0998 | 0.1129 | 0.1106 | 0.1177 | 0.1052 |
| 4 | 0.0498 | 0.0513 | 0.0510 | 0.0504 | 0.0556 | 0.0522 | 0.0578 | 0.0504 |
| 5 | 0.0275 | 0.0271 | 0.0248 | 0.0255 | 0.0262 | 0.0260 | 0.0248 | 0.0236 |
| 6 | 0.0124 | 0.0125 | 0.0116 | 0.0118 | 0.0137 | 0.0138 | 0.0125 | 0.0120 |

$\delta_n^{(i)} = \left|\hat{\theta}_n^{(i)} - \theta\right| / \theta$ for each of estimators $\hat{\theta}_n^{(i)}$, $i = 3, 5$ (Tables 1–4). In the case $a \neq b$ we compute the average relative error only for the estimator $\hat{\theta}_n^{(5)}$ (Table 5).

In the case of equal coefficients we see that the estimators $\hat{\theta}_n^{(3)}$ and $\hat{\theta}_n^{(5)}$ have similar performance. The advantage of $\hat{\theta}_n^{(5)}$ is its independence of the parameter $H$ (which might be unknown). But in the case of known $H$ the estimator $\hat{\theta}_n^{(3)}$ is preferable because it is com-

Table 5: $a(x) = \sin x + 2$, $b(x) = \cos x + 2$.

| $n$ | $H = 0.6$ | $H = 0.7$ | $H = 0.8$ | $H = 0.9$ |
|---|---|---|---|---|
| | $\delta_n^{(5)}$ | $\delta_n^{(5)}$ | $\delta_n^{(5)}$ | $\delta_n^{(5)}$ |
| 3 | 0.0756 | 0.0792 | 0.0757 | 0.0751 |
| 4 | 0.0411 | 0.0361 | 0.0453 | 0.0459 |
| 5 | 0.0200 | 0.0199 | 0.0159 | 0.0200 |
| 6 | 0.0099 | 0.0113 | 0.0094 | 0.0100 |

putable faster.

Also the simulation results show that the rate on convergence probably does not depend on $H$. Moreover, it seems that it is around $2^{-n}$, so the bound in Theorem 1 is not optimal.

**Acknowledgement.** The work was carried out during a stay of the second author at the University of Bern in 2013-2014 and supported by Swiss Government Excellence Scholarship.

## References

Bishwal J (2011). "Minimum Contrast Estimation in Fractional Ornstein-Uhlenbeck Process: Continuous and Discrete Sampling." *Fractional Calculus and Applied Analysis*, **14**(3), 375–410. ISSN 1311-0454.

Hu Y, Nualart D (2010). "Parameter Estimation for Fractional Ornstein-Uhlenbeck Processes." *Statistics & Probability Letters*, **80**(11-12), 1030–1038.

Hu Y, Song J (2013). "Parameter Estimation for Fractional Ornstein-Uhlenbeck Processes with Discrete Observations." In F Viens, J Feng, Y Hu, E Nualart (eds.), *Malliavin Calculus and Stochastic Analysis*, volume 34 of *Springer Proceedings in Mathematics & Statistics*, pp. 427–442. Springer US. ISBN 978-1-4614-5905-7. doi:10.1007/978-1-4614-5906-4_19.

Kozachenko Y, Melnikov A, Mishura Y (2013). "On Drift Parameter Estimation in Models with Fractional Brownian Motion." Accepted by Statistics.

Mishura Y (2008). *Stochastic Calculus for Fractional Brownian Motion and Related Processes*, volume 1929 of *Lecture Notes in Mathematics*. Springer, Berlin.

Mishura Y, Ral'chenko K, Seleznev O, Shevchenko G (2014). "Asymptotic Properties of Drift Parameter Estimator Based on Discrete Observations of Stochastic Differential Equation Driven by Fractional Brownian Motion." Submitted to Modern Trends in Stochastics.

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.

Nualart D, Rascanu A (2001). "Differential Equations Driven by Fractional Brownian Motion." *Collectanea Mathematica*, **53**, 55–81.

Prakasa Rao B (2010). *Statistical Inference for Fractional Diffusion Processes*. Wiley Online Library: Books. Wiley. ISBN 9780470667132.

Tanaka K (2013). "Distributions of the Maximum Likelihood and Minimum Contrast Estimators Associated with the Fractional Ornstein-Uhlenbeck Process." *Statistical Inference for Stochastic Processes*, **16**(3), 173–192. ISSN 1387-0874.

Xiao W, Zhang W, Xu W (2011a). "Parameter Estimation for Fractional Ornstein-Uhlenbeck Processes at Discrete Observation." *Applied Mathematical Modelling*, **35**, 4196–4207.

Xiao W, Zhang W, Zhang XL (2011b). "Maximum-likelihood Estimators in the Mixed Fractional Brownian Motion." *Statistics*, **45**(1), 73–85.

Zhang P, Xiao W, Zhang XL, Niu P (2014). "Parameter Identification for Fractional Ornstein-Uhlenbeck Processes Based on Discrete Observation." *Economic Modelling*, **36**(C), 198–203.

**Affiliation:**

Yuliya Mishura and Kostiantyn Ralchenko
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

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 43/3-4 *Submitted:* 2013-11-25
June 2014 *Accepted:* 2014-03-10

Austrian Journal of Statistics
*June 2014, Volume 43/3-4, 229–236.*
*http://www.ajs.or.at/*

# Power-Law Random Graphs' Robustness: Link Saving and Forest Fire Model

**Marina Leri**
Russian Academy of Sciences

**Yury Pavlov**
Russian Academy of Sciences

**Abstract**

We consider random graphs with node degrees drawn independently from a power-law distribution. By computer simulation we study two aspects of graph robustness: preserving graph connectivity and node saving in the forest fire model, considering two types of graph destruction: the removal of nodes with the highest degrees and equiprobable node extraction.

*Keywords*: random graphs, power-law distribution, robustness, simulation modeling, forest fire model.

## 1. Introduction

The study of random graphs has been gaining interest in the past decades due to the wide use of these models for the description of massive data networks (see e.g. Aiello, Chung, and Lu 2000; Newman, Strogatz, and Watts 2001; Durrett 2007; Hofstad 2011). Such models can be used for representing transport, telephone and electricity networks, social relationships, telecommunications and, of course, the main global network – Internet. While considering these networks it has been noted that their topology could be described by random graphs, with the node degrees being independent and identically distributed (i.i.d.) random variables following the power-law distribution (Faloutsos, Faloutsos, and Faloutsos 1999; Reittu and Norros 2004, etc.).

The structure of present-day complex networks contains many elements, wherefore theoretical research in the field of power-law random graphs includes the study of the limit behaviour of different characteristics of such graphs' structure (Aiello *et al.* 2000; Pavlov 2007; Norros and Reittu 2008, , etc.). Furthermore, one of the important questions raised in the studies of these networks is how their structure and, therefore, functioning change if some of the nodes fall out. That is why one of the important trends in the random graph field has been the study of random graph robustness to different types of breakdowns (see e.g. Cohen, Erez, Ben-Avraham, and Havlin 2000; Bollobas and Riordan 2004; Durrett 2007; Norros and Reittu 2008).

Alongside with the theoretical approach simulation modeling has always been one of the tools for studying random graph objects (Reittu and Norros 2004; Leri 2009, , etc.). In our work

we consider two aspects of graph robustness: link saving or preserving graph connectivity, and node survival, which is closely connected to the study of forest fire models.

## 2. Power-law graph model

We consider power-law random graphs with the number of nodes that equals $N$ . Node degrees $\xi_1, \xi_2, \ldots, \xi_N$ are i.i.d. random variables drawn from the following distribution:

$$\mathbf{P}\{\xi \geq k\} = k^{-\tau}, \qquad k = 1, 2, \ldots, \quad \tau > 1, \tag{1}$$

For graph construction each node is given a certain degree in accordance with the degree distribution (1). Node degrees form stubs (or semiedges) that are numbered in an arbitrary order. The graph is constructed by joining all the stubs pairwise equiprobably to form links. If the sum of node degrees is odd one stub is added to a random vertex. Obviously these graphs have loops and multiple edges. Such construction gave these graphs one of their names – configuration graphs with i.i.d. degrees (Durrett 2007; Hofstad 2011).

## 3. Link saving

Research in the last decades (see Faloutsos *et al.* 1999; Reittu and Norros 2004) showed that configuration power-law random graphs with parameter $\tau$ of the node degree distribution (1) lying in the interval $(1, 2)$ are deemed to be a good implementation of Internet topology at both rooter and domain levels. That was one of the main reasons for us to study such an aspect of graph robustness as preserving graph connectivity or link saving. This issue is important because it is essential to know how the network structure will be influenced by the destruction of some nodes.

When $\tau \in (1, 2)$ the distribution (1) has finite expectation and infinite variance. Both theory (Reittu and Norros 2004; Durrett 2007; Pavlov 2007) and simulation (Reittu and Norros 2004; Leri 2009) agree on the fact that such graphs contain one so called giant component, which is a connected set of nodes the expectation of which is proportional to the number of graph nodes $N$.

For computer experiments we built a simulation model of power-law random graphs (Leri 2009) based on an algorithm introduced by Tangmunarunkit, Govindan, Jamin, Shenker, and Willinger (2002) using a pseudo random generator "Mersenne twister" (see Matsumoto and Nishimura 1998). Previously we showed (Leri 2009) that the structure of these graphs dramatically changes with the variations of the value of the node degree distribution parameter $\tau$ even within this small interval $(1, 2)$, but is much less dependent on the graph size $N$. With the value of parameter $\tau$ close to 1, the graph will be more connected and more than 95% of all graph nodes will form the giant component. On the other hand, the closer the value of parameter $\tau$ is to 2, the fewer nodes there will be in the giant component, i.e. only a half of all graph nodes. This does not however imply a significant growth of the other components. For example, the fraction of nodes in the second component will be rather small in comparison with the giant one. Even at its most, it will not exceed a little more than 1% of all graph nodes. In fact, other components will grow not in size, but in number. In particular, the structure of these power-law random graphs is one of the reasons why is it interesting to see how this structure changes when some graph nodes are removed.

In our work we consider two types of breakdowns: "random breakdown" when graph nodes are removed equiprobably, and "target attack", which means a removal of nodes with the highest degrees. For simulations we took graphs of ten sizes $N$ from 500 to 5000 and 9 values of parameter $\tau$ from the interval $(1, 2)$ with a step of 0.1 (for each pair $(N, \tau)$ 100 graphs were generated to form statistical data). The graph destruction process looks as follows. When a chosen node is destroyed, all the links going out of this node are also removed. And then all isolated nodes are taken away.

Let $\eta_1, \eta_2, \ldots, \eta_s$ be random variables that are equal to the sizes of graph components in decreasing order ($\eta_1$ – the size of the giant component, $\eta_2$ – the size of the second component, etc.), where $s$ is the total number of components. A graph is deemed destroyed when following event A occur: $\{\eta_1 \leq 2\eta_2\}$. Hence, when the size of the second biggest component becomes greater or equal to half the size of the giant component, the graph is considered destroyed.

Simulation results allowed us to derive regression dependencies of node percentages in the giant and the second biggest components ($\eta_1$ and $\eta_2$, respectively) and the total number of components $s$ on the graph size $N$, the parameter of the node degree distribution $\tau$ and the percentage of removed nodes $r$.

In the case of "random breakdown" the following relations were found:

$$\eta_1 = 129 - 36\ \tau - 1.1\ r,$$

$$\eta_2 = 2 - 0.25 \ln N + 0.42\ \tau - 0.017 \ln r,$$

$$\frac{s}{N} = -0.18 + 0.2\ \tau - 0.004\ r \ln \tau.$$

The determination coefficients ($R^2$) of these regression models are equal to 0.98, 0.7 and 0.98, respectively. The percentage of removed graph nodes has to be confined within the following bounds: $100/N \leq r \leq 117 - 32.7\tau$. Here in after the lower bound implies the removal of one node, and the upper bound means that the extraction of a higher percent of nodes will lead to complete graph breakdown. The results show that the percentage of nodes in the giant component does not depend on the graph size $N$ and the percent of nodes in the second component will not exceed 2% of the graph size.

The following regression dependencies were derived for "target attack" on the nodes with the highest degrees:

$$\eta_1 = 130 - 46\ \tau - 9\ r,$$

$$\eta_2 = 4.36 - 0.44 \ln N + \tau + 0.4 \ln r,$$

$$\ln s = -3.3 + \ln N + 2.3 \ln \tau + 0.1\ r,$$

with determination coefficients 0.95, 0.6 and 0.98, respectively, and the percentage of removed nodes confined within the following bounds: $100/N \leq r \leq 14 - 5.15\tau$. Here again, the percent of nodes in the giant component does not depend on the graph size $N$, and the percentage of nodes in the second component will not exceed 4% of the graph size.

Below are the results of the estimation of the regression dependence of the probability $\mathbf{P}\{\mathrm{A}\}$ of graph destruction on the percentage of removed nodes $r$ and parameter $\tau$ with $R^2 = 0.84$ and $R^2 = 0.76$, respectively.

For "random breakdown":

$$\mathbf{P}\{\mathrm{A}\} = \begin{cases} 0, & r < 37/\sqrt{\tau}, \\ -0.2 + 1.5 \cdot 10^{-4} \tau r^2, & 37/\sqrt{\tau} \leq r < 89/\sqrt{\tau}, \\ 1, & r \geq 89/\sqrt{\tau}, \end{cases}$$

For the "target attack":

$$\mathbf{P}\{\mathrm{A}\} = \begin{cases} 0, & \ln r < 1.85 - \tau, \\ -0.38 + 0.06 r e^{\tau}, & 1.85 - \tau \leq \ln r \leq 3.13 - \tau, \\ 1, & \ln r > 3.13 - \tau, \end{cases}$$

This means that in the "random breakdown" case (see Figure 1), for example, an estimated probability of graph destruction equals 0 when $\tau = 1.1$ for all $r < 35.3\%$, and when $\tau = 1.9$ for $r < 26.9\%$. And $\mathbf{P}\{\mathrm{A}\} = 1$ for $r > 84.8\%$ when $\tau = 1.1$, and when $\tau = 1.9$ for $r > 64.5\%$. On the other hand in the case of "target attack" (see Figure 2) $\mathbf{P}\{\mathrm{A}\} = 0$ when $\tau = 1.1$ for all $r < 2.12\%$, and when $\tau = 1.9$ for $r < 0.95\%$. And $\mathbf{P}\{\mathrm{A}\} = 1$ for $r > 7.6\%$ when $\tau = 1.1$, and when $\tau = 1.9$ for $r > 3.4\%$.

The results show that power-law random graphs with parameter $\tau \in (1, 2)$ are much more vulnerable to "target attacks" than to "random breakdowns". In order to destroy such a graph by deleting high-degree nodes it is enough to remove $3 - 7\%$ of them. If however graph nodes

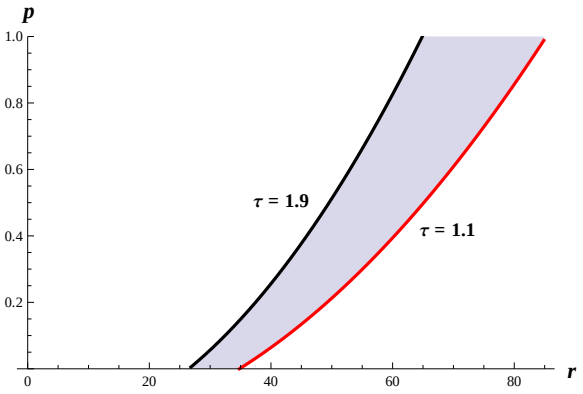


Figure 1: Probability of graph destruction for "random breakdown".

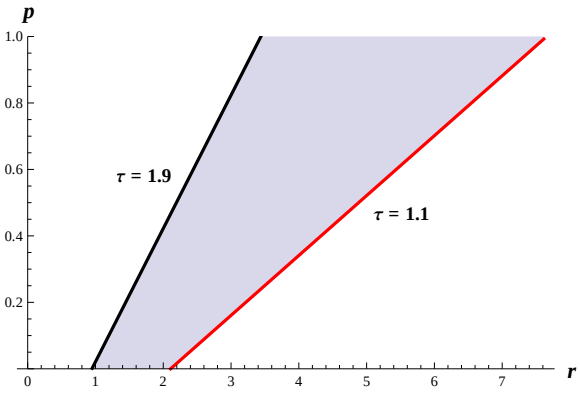


Figure 2: Probability of graph destruction for "target attack".

are broken randomly, it may be not ruined even if more than 60% of its vertices had been removed. Furthermore, robustness of these random graphs strongly depends on the value of parameter $\tau$. In both breakdown cases the graph proved to be more resistant if the value of $\tau$ was closer to 1 and more vulnerable as it moved closer to 2.

## 4. Node survival – forest fire model

The second aspect of power-law graphs' robustness we are considering here is node survival. This issue branched off the studies of forest fire models (see e.g. Drossel and Schwabl 1992; Bertoin 2012). Let us consider graph nodes as trees on a certain area of a real forest. Two nodes are connected if a fire can move on related trees from one tree to another (for proper implication it looks more like a crown fire). So, we pose the question of finding how many trees should initially be set on a certain area to ensure their maximum survival in case of a fire. This approach could be used not only for modeling forest fire dynamics. It also has other applications (Bertoin 2011), including modeling banking system defaults in order to minimize their negative effects (see e.g. Annakov 2008; Arinaminparty, Kapadia, and May 2012).

In this part of our work we consider the same configuration power-law random graphs with node degree distribution (1), but with parameter $\tau > 1$ with no the upper bound. Since we assume that the area of a forest is limited, we have to also restrain the number of trees growing there. So, to specify the graph topology, let graph vertices be placed in the nodes of a square lattice sized $100 \times 100$. Links connect nodes in the "closest neighbour" manner, so in a fully packed graph every inner tree (node) has 8 adjacent neighbours. This does not mean that in the following study we consider power-law random graphs with node degrees no higher than 8. The fact is that on the one hand high node degrees and, all the more, an average node degree have low probabilities, and on the other hand the graph may contain multiple edges that raise the probability of fire transfer from one neighbour to the other. That is why we introduce the lattice only to determine the relation between the initial number of nodes in the area and the parameter $\tau$ of the node degree distribution (1). If an average node degree $i$ is less than 8, some graph links are missing. Figure 3 shows a couple of examples of lattice-graph topology for two average inner node degrees.

Taking into consideration that graph node degrees are defined by distribution (1), and having determined the dependency between an average node degree $i$ and parameter $\tau$ on the interval $i \in (1, 8]$ as $i = \zeta(\tau)$ (where $\zeta$ is a Riemann zeta function) (see Table 1), we found that graph size $N \leq 10000$ is related to parameter $\tau$ by the following regression function (see Figure 4) with $R^2 = 0.97$:

$$N = 9256\ \tau^{-1.05}. \tag{2}$$

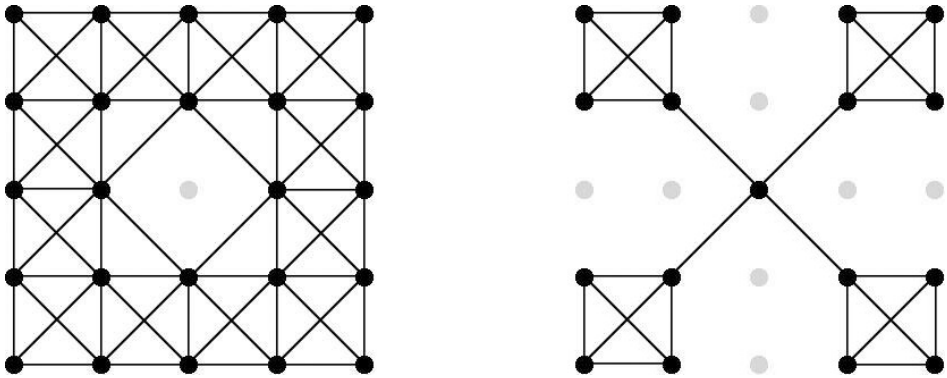

Figure 3: Lattice graph topology for $i = 7$ and $i = 4$, respectively.

Table 1: Calculated values of $\tau$ and $N$ for different $i$.

| $i$ | 1.01 | 1.21 | 1.33 | 1.42 | 1.5 | 1.6 | 2 | 2.66 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\tau$ | 6.75 | 2.96 | 2.53 | 2.32 | 2.19 | 2.05 | 1.73 | 1.49 | 1.42 | 1.29 | 1.23 | 1.18 | 1.16 | 1.13 |
| $N$ | 3350 | 3600 | 3750 | 3900 | 4000 | 4200 | 5000 | 4780 | 4489 | 5578 | 6700 | 8350 | 8911 | 10000 |

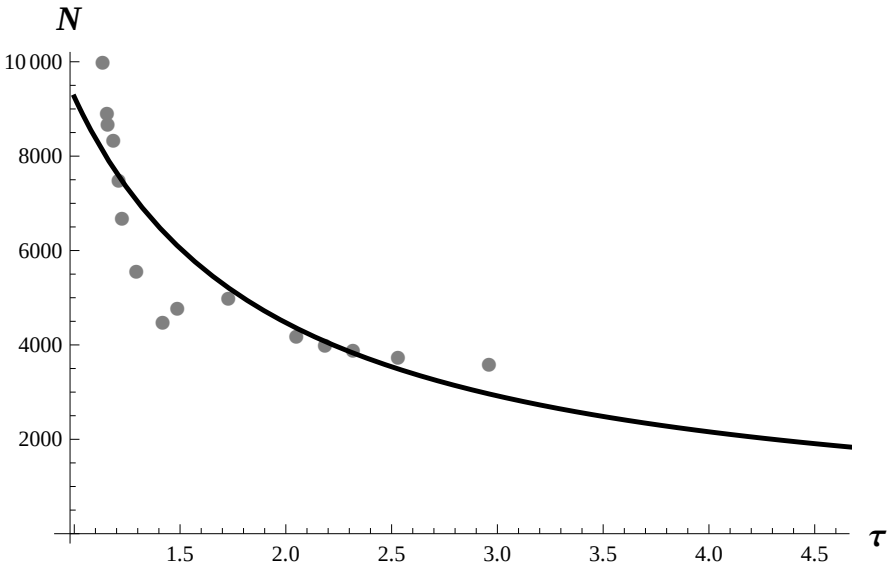


Figure 4: Regression relationship between $N$ and $\tau$.

For simulations we used a subset of configuration graphs the number of nodes in which is specified by relation (2). We assume that the graph destruction process (or fire) starts from some chosen node. As the first node is set on fire, it passes on along incident links to the connected nodes with a given probability $0 < p \leq 1$. Let's call it the probability of link destruction. This means that each link becomes inflammable with a probability $p$, and therefore a connected node is also set on fire. Otherwise a link becomes fire resistant and the node connected through this link remains intact. This does not mean however that the fire cannot reach this node via a parallel link (if any) with the same probability $p$.

The fire spreads over the graph until there appear inflammable links, and all burnt nodes and links are removed from the graph when it stops. The aim is to find the optimal values of parameter $\tau$ that secure maximum survival of nodes, and to find how they depend on the probability of link destruction.

We consider two cases of fire startup: "random fire start" when the first node to be removed is chosen equiprobably, and "targeted fire start" with fire starting from a node with the highest

degree. Let $g$ be the number of nodes remaining in a graph after a fire. Figure 5 and Figure 6 show the results obtained for $p = 1$ in both breakdown cases, respectively.

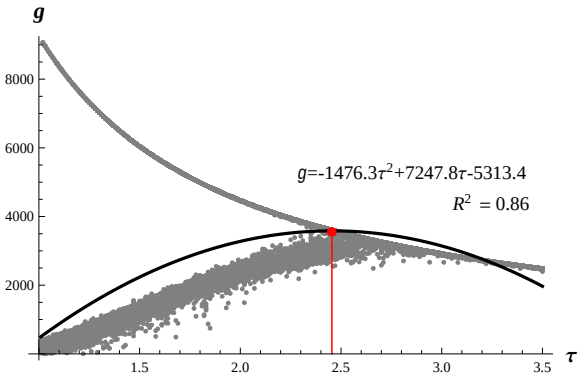


Figure 5: Relation between the number of remaining nodes $g$ and parameter $\tau$ ("random fire start", $p = 1$).

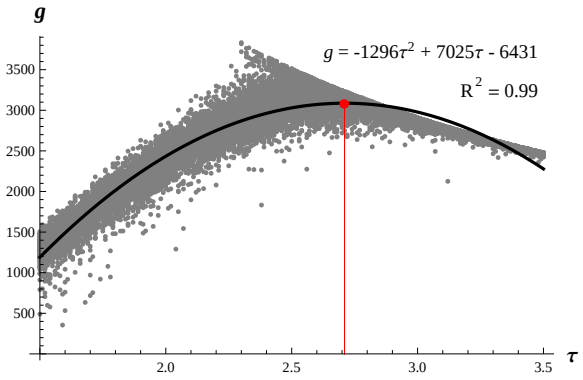


Figure 6: Relation between the number of remaining nodes $g$ and parameter $\tau$ ("targeted fire start", $p = 1$).

This means, for example, that in the "random" case the number of remaining nodes becomes maximal for the power-law graphs with parameter $\tau = 2.45$. The initial graph size $N$ then equals 3605 and an average number of nodes remaining after the fire is $g \approx 3580$.

For both cases of breakdown start were found regression dependencies of the number of nodes remaining in a graph $g$ on $\tau$ and the probability of fire spread $p$. Below we give these models for the cases of "random" start and "target" start, respectively:

$$g = 6008.8 - 1915.3\ p - 217.4\ \tau^2, \qquad (R^2 = 0.91);$$

$$g = 2938 - 894.2 \ln p - 74.5\ \tau^2, \qquad (R^2 = 0.95).$$

Obviously, the number of remaining nodes decreases as $p$ increases. Relations were found that describe the dependencies of $g$ on $\tau$ for different values of $p$ and dependencies of $g$ on $p$ for different $\tau$. This allowed to find the relation between $\tau_{max} = \tau_{max}(p)$ of parameter $\tau$ for which $g$ reaches its maximum $g_{max}$ on $p$ and, thus, find the values of $g_{max} = g_{max}(p)$. For example, for $p = 1$ and $p = 0.6$ under "random" start $\tau_{max}(1) = 2.46$, $\tau_{max}(0.6) = 1.1$, $g_{max}(1) = 3585$, $g_{max}(0.6) = 4468$, and under "target" start $\tau_{max}(1) = 2.74$, $\tau_{max}(0.6) = 2.23$, $g_{max}(1) = 3216$, $g_{max}(0.6) = 3995$.

Thus, in order to secure a maximum of unburnt trees in some specified territory in the case of a fire (either "random" or "targeted") the topology of their layout has to correspond to the topology of power-law random graph with parameter $\tau$ of node degree distribution (1) between values 2.4 and 2.7. Such graph will represent a multicomponent structure with no giant connected component with an average node degree 1.2 through 1.4. As for the difference between graphs robustness in the two breakdown cases, the graph will be more robust and more nodes will survive in a fire in the case of a "random" start than in the case of a "target" start.

## 5. Acknowledgments

The study was supported by the Russian Foundation for Basic Research, grant 13-01-00009 and by the Strategic Development Programme of the Petrozavodsk State University for years 2012–2016. Also we would like to thank professor A.M. Zubkov (Steklov Mathematical Institute of RAS) for a constructive discussion of the problem.

## References

Aiello W, Chung F, Lu L (2000). "A random graph model for massive graphs." *Proc. of the 32nd Annual ACM Symposium on Theory of Computing*, pp. 171–180.

Annakov B (2008). "Bank crisis and forest fire: What's in common?" *URL http://www.empatika.com/blog/agent_modeling_forest_fire*. In Russian.

Arinaminparty N, Kapadia S, May R (2012). "Size and complexity 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.

Bollobas B, Riordan O (2004). "Robustness and vulnerability of scale-free random graphs." *Internet Mathematics*, **1**(1), 1–35.

Cohen R, Erez K, Ben-Avraham D, Havlin S (2000). "Resilience of the Internet to Random Breakdowns." *Phys. Rev. Lett.*, **85**, 4626–4628.

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.

Hofstad R (2011). *Random Graphs and Complex Networks*. Eindhoven University of Technology.

Leri M (2009). "Modelling of random graphs of Internet-type." *Surveys in Applied and Industrial Mathematics*, **16**(5), 737–744. In Russian.

Matsumoto M, Nishimura T (1998). "Mersenne Twister: A 623-dimensionally equidistributed uniform pseudorandom number generator." *ACM Trans. on Modeling and Computer Simulation*, **8**(1), 3–30.

Newman M, Strogatz S, Watts D (2001). "Random graphs with arbitrary degree distribution and their applications." *Phys. Rev. E.*, **64**, 026118.

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 (2007). "The limit distribution of the size of a giant component in an Internet-type random graph." *Discrete Mathematics and Applications*, **17**(5), 425–437.

Reittu H, Norros I (2004). "On the power-law random graph model of massive data networks." *Performance Evaluation*, **55**, 3–23.

Tangmunarunkit H, Govindan R, Jamin S, Shenker S, Willinger W (2002). "Network topology generators: degree-based vs. structural." *Proceedings of the SIGCOMM'02*, pp. 147–159.

**Affiliation:**

Marina Leri and Yury Pavlov
Institute of Applied Mathematical
Research of the Karelian Research Centre
Russian Academy of Sciences
11, Pushkinskaya str.
Petrozavodsk Karelia
185910, Russia
E-mail: leri@krc.karelia.ru and pavlov@krc.karelia.ru

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 43/3-4
June 2014

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

*Submitted:* 2013-11-25
*Accepted:* 2014-03-10

Austrian Journal of Statistics
*June 2014, Volume 43/3-4, 237–245.*
*http://www.ajs.or.at/*

# A Few Remarks on Robust Estimation of Power Spectra

**Georgy Shevlyakov Nickolay Lyubomishchenko Pavel Smirnov**
St. Petersburg State Polytechnic University

**Abstract**

Various robust versions of the classical methods of power spectra estimation are considered. Their performance evaluation is studied in autoregressive models with contamination. It is found out that the best robust estimates of power spectra are based on robust highly efficient estimates of autocovariances. Several open problems for future research are formulated.

*Keywords*: power spectra, robustness, autoregressive models.

## A Few Remarks on Robust Estimation of Power Spectra

Georgy Shevlyakov, Nickolay Lyubomishchenko and Pavel Smirnov
St. Petersburg State Polytechnic University, Russia

**Abstract:** Various robust versions of the classical methods of power spectra estimation are considered. Their performance evaluation is studied in autoregressive models with contamination. It is found out that the best robust estimates of power spectra are based on robust highly efficient estimates of autocovariances. Several open problems for future research are formulated.

**Keywords:** power spectra, robustness, autoregressive models.

## 1. Introduction

Robust methods ensure high stability of statistical inference under uncontrolled deviations from the assumed distribution model. Much less attention is devoted in the literature to robust estimation of data spectra as compared to robust estimation of location, scale, regression and covariance (Huber, 1981; Hampel, Ronchetti, Rousseeuw, and Stahel, 1986; Maronna, Martin, and Yohai, 2006). However, it is necessary to study these problems due to their both theoretical and practical importance (estimation of time series power spectra in various applications, such as communication, geophysics, medicine, etc.), and also because of the instability of classical methods of power spectra estimation in the presence of outliers in the data (Kleiner, Martin, and Thomson, 1979).

There are several classical approaches to estimation of the power spectra of time series, e.g., via the nonparametric periodogram and the Blackman-Tukey formula methods, as well as via the parametric Yule-Walker and filter-based methods (Blackman and Tukey, 1958; Bloomfield, 1976; Brockwell and Davis, 1991). Thereafter, we may consider their various robust versions: to the best of our knowledge, a first systematic study of them is made in the dissertation of Bernhard Spangl (Spangl, 2008).

In what follows, we partially use the aforementioned study as a baseline, mostly follow the classification of robust methods of power spectra estimation given in (Spangl, 2008), specify them and propose some new approaches with their comparative performance evaluation. Basically, to obtain good robust estimates of power spectra, we use highly efficient robust estimates of scale and correlation (Shevlyakov and Smirnov, 2011).

Our main goals are both to outline the existing approaches to robust estimation of power spectra and to indicate open problems, so our paper is partially a review and partially a program for a future research.

The remainder of the paper is as follows. In Section 2, classical methods of power spectra estimation are briefly enlisted. In Section 3, robust modifications of classical approaches are formulated. In Section 4, a few preliminary results on the comparative study of the performance evaluation of various robust methods are represented. In Section 5, some conclusions and open problems for future research are drawn.

## 2. Classical estimation of power spectra

### 2.1. Nonparametric estimation of power spectra

The nonparametric approach to estimation of power spectra is based on smoothed periodograms (Blackman and Tukey, 1958; Bloomfield, 1976).

Let $x_t, t = 1, \dots, n$ be a second-order stationary time-series with zero mean. Assume that the time intervals between two consecutive observations are equally spaced with duration $\Delta t$. Then the periodogram is defined as follows:

$$\widehat{S}_P(f) = \Delta t/n \left| \sum_{t=1}^{n} x_t \exp\{-i2\pi f t \Delta t\} \right|^2 \tag{1}$$

over the interval $[-f_{(n)}, f_{(n)}]$, where $f_{(n)}$ is the Nyquist frequency: $f_{(n)} = 1/(2\Delta t)$.

The Blackman-Tukey formula gives the representation of formula (1) via the sample autocovariances $\hat{c}_{xx}$ of the time series $x_t$ (Blackman and Tukey, 1958):

$$\widehat{S}_P(f) = \widehat{S}_{BT}(f) = \Delta t \sum_{h=-(n-1)}^{n-1} \hat{c}_{xx}(h) \exp\{-i2\pi f h \Delta t\}\,. \tag{2}$$

It can be seen that the periodogram $\widehat{S}_P(f)$ (1) at the frequency $f = f_k = k/(n\Delta t)$, where $k$ is an integer such that $k \le \lfloor n/2 \rfloor$, is equal to the squared absolute value of the discrete Fourier transform $X(f_k)$ of the sequence $x_1, \dots, x_n$ given by the following formula

$$X(f_k) = \Delta t \sum_{t=1}^{n} x_t \exp\{-i2\pi f_k t \Delta t\}\,. \tag{3}$$

To reduce the bias and variance of the periodogram $\widehat{S}_P(f)$, the conventional techniques based on tapering and averaging of periodograms is used (Bloomfield, 1976).

### 2.2. Parametric estimation of power spectra

The widely used form of a parametric power spectra estimation procedure exploits an autoregressive model of order $p$ for the underlying power spectrum $S(f)$. A stationary $AR(p)$ process $x_t$ with zero mean is described by the following equation

$$x_t = \sum_{j=1}^{p} \phi_j x_{t-j} + \epsilon_t \,, \tag{4}$$

where $\epsilon_t$ are i.i.d. Gaussian white noises with zero mean and variance $\sigma_\epsilon^2$ . The power spectrum estimate $\widehat{S}_{AR}(f)$ has the form (Bloomfield, 1976)

$$\widehat{S}_{AR}(f) = \frac{\Delta t \hat{\sigma}_\epsilon^2}{\left|1 - \sum_{j=1}^{p} \hat{\phi}_j \exp\{-i2\pi f j \Delta t\}\right|^2} \,, \quad |f| \le f_{(n)}, \tag{5}$$

where $\hat{\phi}_1, \ldots, \hat{\phi}_p$ and $\hat{\sigma}_\epsilon^2$ are the maximum likelihood estimates of the model parameters.

## 3. Robust estimation of power spectra

### 3.1. Preliminaries

A natural way to provide robustness of the classical estimates of power spectra is based on using highly robust and efficient estimates of location, scale and correlation in the classical estimates. Here we enlist several highly robust and efficient estimates of scale and correlation.

*Robust Scale:* The median absolute deviation $MAD_n(x) = med\,|x - med\,x|$ is a highly robust estimate of scale with the maximal value of the breakdown point 0.5, but its efficiency is only 0.37 at the normal distribution (Hampel, Ronchetti, Rousseeuw, and Stahel, 1986). In (Rousseeuw and Croux, 1993), a highly efficient robust estimate of scale $Q_n$ has been proposed: it is close to the lower quartile of the absolute pairwise differences $|x_i - x_j|$, and it has the maximal breakdown point 0.5 as for $MAD_n$ but much higher efficiency 0.82. The drawback of this estimate is its low computation speed; the computation of $Q_n$ requires an order of greater time than of $MAD_n$.

In (Smirnov and Shevlyakov, 2010), an $M$-estimate of scale denoted by $FQ_n$ whose influence function is approximately equal to the influence function of the estimate $Q_n$ is proposed

$$FQ_n(x) = 1.483\, MAD_n(x)\, \left(1 - (Z_0 - n/\sqrt{2})/Z_2\right) \,, \tag{6}$$

$$Z_k = \sum_{i=1}^{n} u_i^k e^{-u_i^2/2} \,, \quad u_i = (x_i - med\,x)/(1.483\, MAD_n),\ k = 0, 2; \quad i = 1, \ldots, n.$$

The efficiency and breakdown point of $FQ_n$ are equal to 0.81 and to 0.5, respectively.

*Robust Correlation:* A remarkable robust minimax bias and variance $MAD$ correlation coefficient with the breakdown point 0.5 and efficiency 0.37 is given by

$$r_{MAD}(x, y) = (MAD^2(u) - MAD^2(v)) \Big/ (MAD^2(u) + MAD^2(v)), \tag{7}$$

where $u$ and $v$ are the robust principal variables (Shevlyakov and Smirnov, 2011)

$$u = \frac{x - med\,x}{\sqrt{2}\, MAD\, x} + \frac{y - med\,y}{\sqrt{2}\, MAD\, y}, \quad v = \frac{x - med\,x}{\sqrt{2}\, MAD\, x} - \frac{y - med\,y}{\sqrt{2}\, MAD\, y}.$$

Much higher efficiency 0.81 with the same breakdown point 0.5 can be provided by using the $FQ$ correlation coefficient (Shevlyakov and Smirnov, 2011)

$$r_{FQ}(x, y) = (FQ^2(u) - FQ^2(v)) \Big/ (FQ^2(u) + FQ^2(v)) \,. \tag{8}$$

### 3.2. Robust $L_p$-norm analogs of the discrete Fourier transform

Since computation of the discrete Fourier transform (DFT) (3) is the first step in periodogram estimation of power spectra, consider the following robust $L_p$-norm analogs of the DFT.

As the classical DFT (3) $X(f)$ can be obtained via the $L_2$-norm approximation to the data $y_t(f) = x_t \exp\{-i2\pi f\, t\Delta t\},\ t = 1, \ldots, n$:

$$X(f) \propto \arg\min_Z \sum_{t=1}^{n} \left| y_t(f) - Z \right|^2 ,$$

the $L_p$-norm analog of $X(f)$ (up to the scale factor) is defined as follows :

$$X_{L_p}(f) \propto \arg\min_Z \left\{ \sum_{t=1}^{n} \left| y_t(f) - Z \right|^p \right\}^{1/p}, \quad 1 \le p < \infty. \tag{9}$$

The case of $1 \le p < 2$, and especially the $L_1$-norm or the median Fourier transform, are of our particular interest (Pashkevich and Shevlyakov, 1995; Spangl and Dutter, 2005; Spangl, 2008):

$$X_{L_1}(f) \propto \arg\min_Z \left\{ \sum_{t=1}^{n} \left| y_t(f) - Z \right| \right\} . \tag{10}$$

The other possibilities such as the component-wise, spatial medians, and trimmed mean analogs of the DFT are also considered in (Pashkevich and Shevlyakov, 1995; Spangl, 2008).

### 3.3. Robust nonparametric estimation

Now we apply the aforementioned robust analogs of the DFT as well as highly robust and efficient estimates of scale and correlation to the classical nonparametric estimation of power spectra.

*Robust Nonparametric Estimation via Periodograms:* Here we apply the robust $L_p$-norm analogs of the DFT to the classical periodogram $\widehat{S}_P(f)$ (1):

$$\widehat{S}_{L_p}(f) \propto \left| X_{L_p}(f) \right|^2 . \tag{11}$$

In what follows, the $L_1$- or the median periodogram is of our particular interest.

*Robust Nonparametric Estimation via the Blackman-Tukey Formula:* In order to construct robust modifications of the Blackman-Tukey formula, we have to consider robust estimates of autocovariances $\widehat{c}_{xx}(h)$ instead of the conventional ones used in (2). These robust estimates are based on the highly robust $MAD$ and $FQ$ estimates of scale and correlation (6) - (8):

$$\widehat{c}_{MAD}(h) = r_{MAD}(x_t, x_{t-h}) MAD(x_t) MAD(x_{t-h}) = r_{MAD}(h) MAD^2(x) , \tag{12}$$

$$\widehat{c}_{FQ}(h) = r_{FQ}(x_t, x_{t-h}) FQ(x_t) FQ(x_{t-h}) = r_{FQ}(h) FQ^2(x) .$$

To provide the required Teplitz property (symmetry, semipositive definiteness, equal elements on sub-diagonals) of the autocovariance matrix $\widehat{C}_{xx}$ built of the element-wise robust autocovariances (12), a new effective transform is used (Letac, 2011). Thus, the Teplitz transformed estimates are substituted into formula (2), and the corresponding robust estimates of power spectra are denoted as $\widehat{S}_{MAD}(f)$ and $\widehat{S}_{FQ}(f)$, respectively.

### 3.4. Robust parametric estimation of power spectra via the Yule-Walker equations

A classical approach to estimation of autoregressive parameters $\phi_1, \ldots, \phi_p$ in (4) is based on

the solution of the linear system of the Yule-Walker equations (Bloomfield, 1976):

$$\begin{cases} \widehat{c}(1) = \widehat{c}(0)\widehat{\phi}_1 + \widehat{c}(1)\widehat{\phi}_2 + \cdots + \widehat{c}(p-1)\widehat{\phi}_p \\ \widehat{c}(2) = \widehat{c}(1)\widehat{\phi}_1 + \widehat{c}(2)\widehat{\phi}_2 + \cdots + \widehat{c}(p-2)\widehat{\phi}_p \\ \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \\ \widehat{c}(p) = \widehat{c}(p-1)\widehat{\phi}_1 + \widehat{c}(p-2)\widehat{\phi}_2 + \cdots + \widehat{c}(0)\widehat{\phi}_p \,. \end{cases} \tag{13}$$

The estimate of the innovation noise variance is defined by the following equation

$$\widehat{c}(0) = \widehat{c}(1)\widehat{\phi}_1 + \widehat{c}(2)\widehat{\phi}_2 + \cdots + \widehat{c}(p)\widehat{\phi}_p + \widehat{\sigma}_\epsilon^2 \,. \tag{14}$$

Substituting robust estimates of autocovariances (12) into (13) and (14), we get the robust analogs of the Yule-Walker equations. Solving these equations, we arrive at the robust estimate of power spectra in the form (5).

### 3.5. Robust parametric estimation via filtering

A wide collection of robust methods of power spectra estimation is given by various robust filters (Kalman, Masreliez, ACM-type, robust least squares, filter-cleaners, etc.) providing preliminary cleaning the data with the subsequent power spectra estimation. An extended comparative experimental study of robust filters is made in (Spangl and Dutter, 2005; Spangl, 2008); below we compare some of those results with ours.

## 4. Performance evaluation

### 4.1. Robustness of the median Fourier transform power spectra

The median Fourier transform power spectra estimate $\widehat{S}_{L_1}(f) \propto \left|X_{L_1}(f)\right|^2$ inherits the maximum value of the sample median breakdown point $\varepsilon^* = 1/2$.

**Theorem** *The breakdown point of $\widehat{S}_{L_1}(f)$ is equal to $1/2$. Here, the breakdown point $\varepsilon^*$ is understood as the maximal ratio of the number of unbounded observations in the data sample under which the estimate still remains bounded (Hampel, Ronchetti, Rousseeuw, and Stahel, 1986).*

Fig. 1 illustrates this phenomenon: the observed realisation is the mixture of $\sin(\pi t/4)$ and $\sin(\pi t/8)$ on the 40% and on the 60% of the interval of observation, respectively. In this case, the classical periodogram indicates the presence of both peaks whereas the median periodogram indicates only one spectrum peak, which corresponds to the dominating signal $\sin(\pi t/8)$.

### 4.2. Additive outlier contamination model

In Monte Carlo experiment, an autoregressive model is used because of, first, it is a direct stochastic counterpart of an ordinary differential equation, second, an autoregressive model is the maximum entropy parametric approximation to an arbitrary strictly stationary random process (Cover and Thomas, 1991).

In this paper, we use the autoregressive models AR(2): $x_t = x_{t-1} - 0.9\,x_{t-2} + \epsilon_t$ and AR(4): $x_t = x_{t-1} - 0.9\,x_{t-2} + 0.5\,x_{t-3} - 0.1\,x_{t-4} + \epsilon_t$ together with Gaussian additive outliers (AO) with pdf $N(x; 0, 10)$. The comparative study is performed on different sample sizes $n$ and numbers of trials $M$ (see, Figs. 2-4).

### 4.3. Disorder contamination model

In this paper, we propose a contamination model dubbed as a disorder contamination describing the violations of the thin structure of a random process, when an $AR$-process is shortly

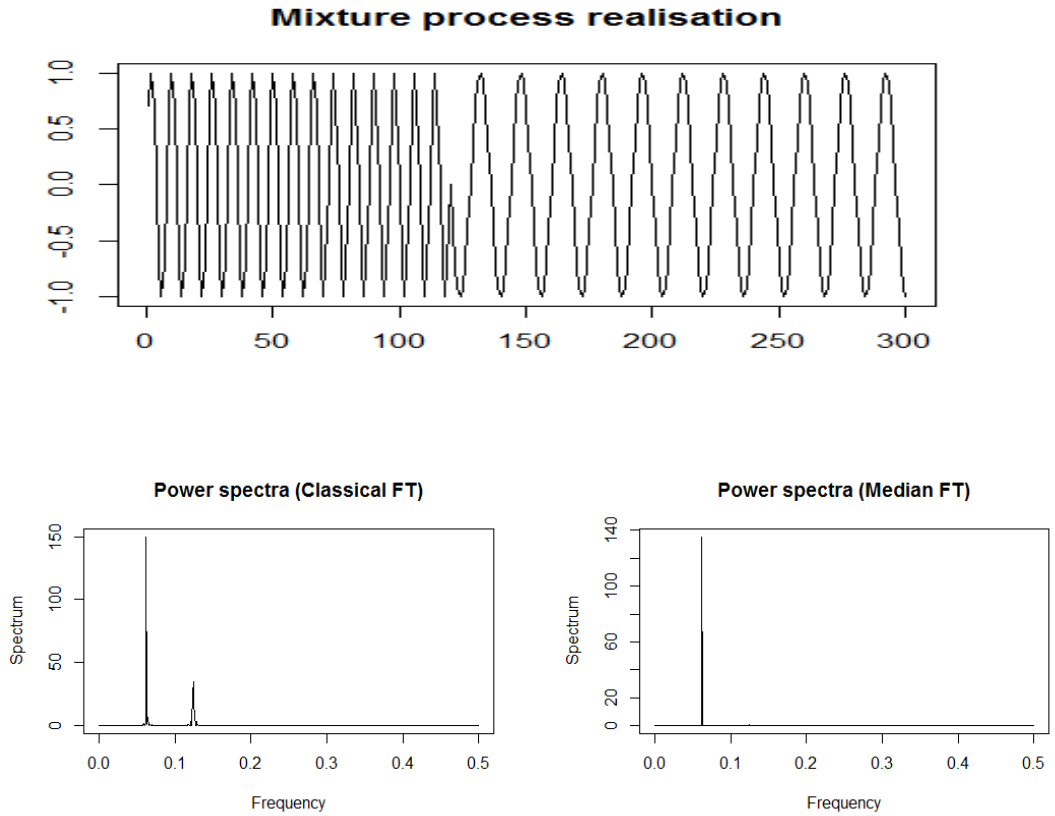


Figure 1: Median Fourier transform breakdown point $\varepsilon^* = 0.5$ property

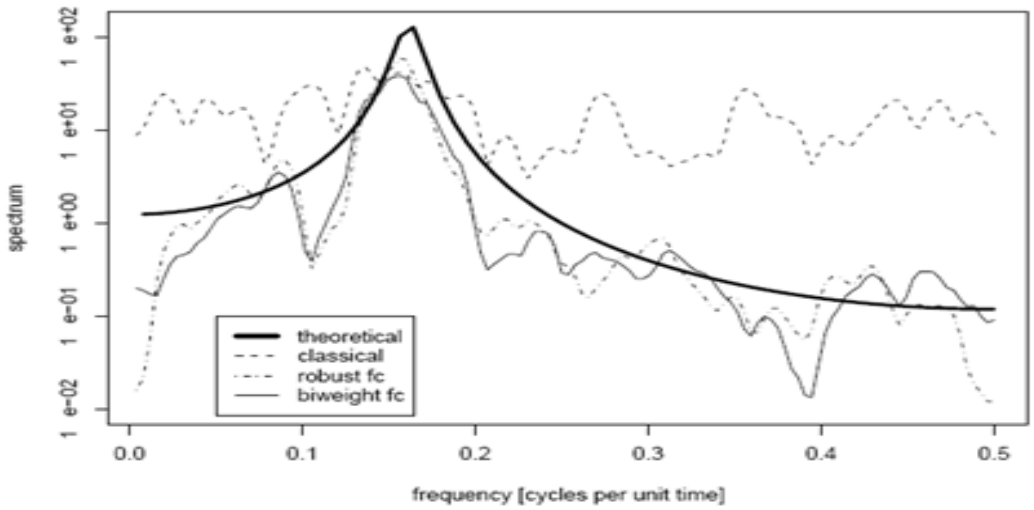


Figure 2: Power spectra estimation in $AR(2)$ model with 10% AO contamination by robust filter-cleaners: n=100, M=400

changed for another and then it returns to the previous state.

Below, the following disorder model is used: $x_t = -0.6\,x_{t-1} - 0.6\,x_{t-2} + \epsilon_t$ as the main process observed at $t = 0, 1, \ldots, 400$ and at $t = 512, \ldots, 1024$; $x_t = x_{t-1} - 0.9\,x_{t-2} + \epsilon_t$ as the disorder process at $t = 401, \ldots, 511$. The results of signal processing are exhibited in Figs. 5-6: the classical periodogram indicates two spectrum peaks of the main and contamination processes, whereas the median periodogram indicates only one peak of the main process.

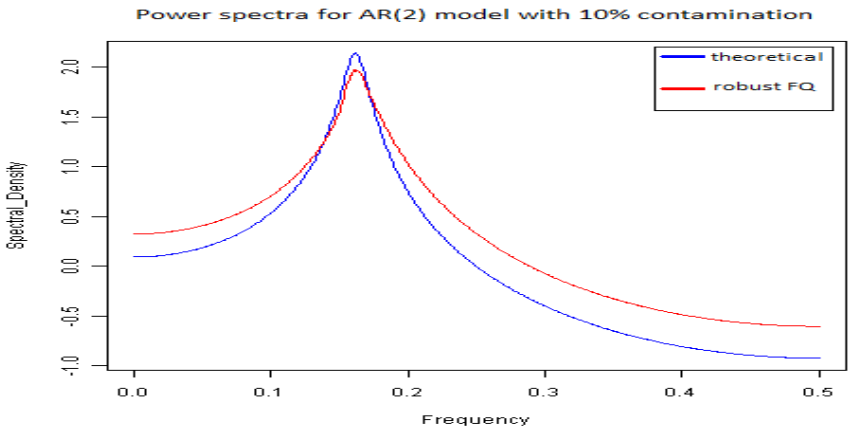


Figure 3: Robust Yule-Walker power spectra estimation in $AR(2)$ model with 10% AO contamination: n=128, M=2000

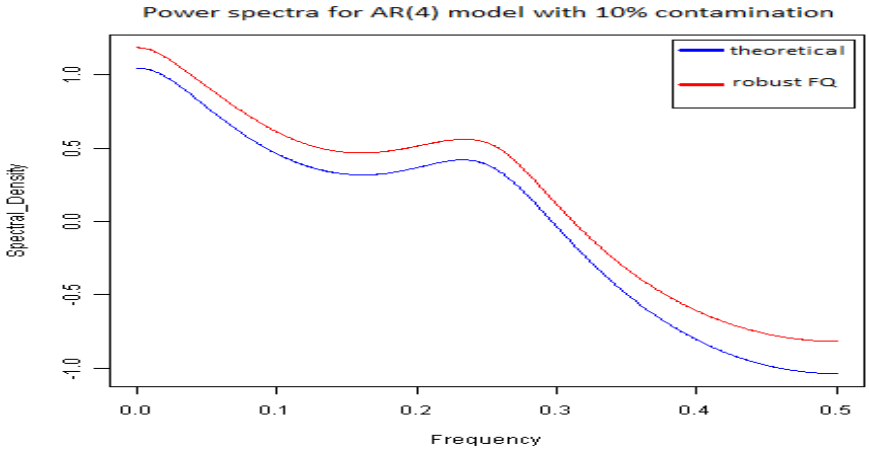


Figure 4: Robust Yule-Walker power spectra estimation in $AR(4)$ model with 10% AO contamination: n=128, M=2000

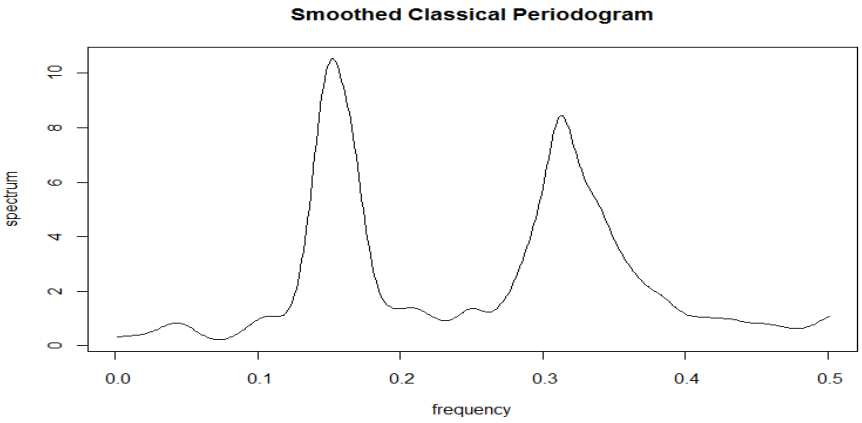


Figure 5: Smoothed classical power spectra estimation in disorder model with 10% contamination

## 5. Concluding remarks

1) From Figs. 2-3 it follows that the classical periodogram is catastrophically bad under contamination, and that the robust $FQ$ Yule-Walker estimate considerably outperforms robust

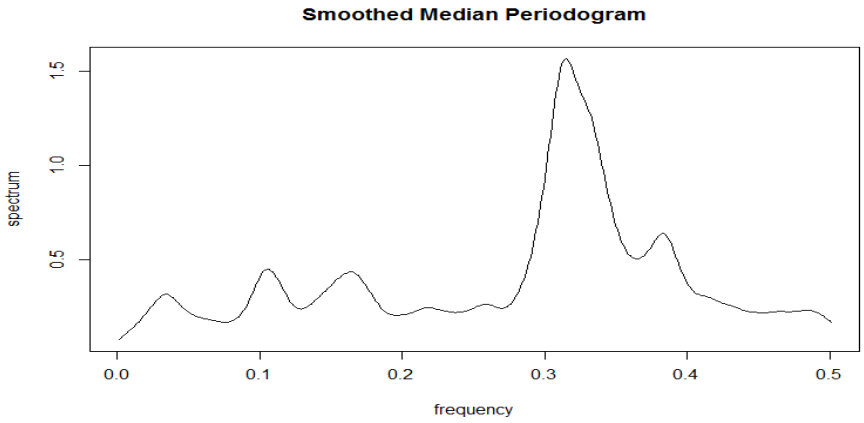


Figure 6: Smoothed robust power spectra estimation in disorder model with 10% contamination

filter methods.

2) From Fig. 4 it follows that the bias of estimation by the *FQ* Yule-Walker method increases with growing dimension and contamination. It can be also shown that under heavy contamination, the median periodogram and the robust Blackman-Tukey method outperform the *FQ* Yule-Walker method in estimating the peak location, although they have a considerable bias in amplitude.

3) The median periodogram exhibits high robustness both with respect to amplitude outliers and to disorder contamination.

4) The obtained results indicate many open problems: analysis of the asymptotic properties of the proposed estimates, reducing their bias and variance on finite samples, and study of the properties of the direct and inverse $L_p$-norm analogs of the Fourier transform.

## References

Blackman, R.B., and Tukey, J.W. (1958). *The Measurement of Power Spectra.* New York: Dover.

Bloofield, P. (1976). *Fourier Analysis of Time Series: An Introduction.* New York: Wiley.

Brockwell, P.J., and Davis, R.A. (1991). *Time Series: Theory and Methods.* New York: Springer.

Cover, T.M., and Thomas, J.A. (1991). *Elements of Information Theory.* New York: Wiley.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986). *Robust Statistics. The Approach Based on Influence Functions.* New York: Wiley.

Huber, P.J. (1981). *Robust Statistics.* New York: Wiley.

Kleiner, B., Martin, R.D., and Thomson, D.J. (1979). Robust Estimation of Power Spectra. *Journal of the Royal Statistical Society, B., 41*, 313-351.

Letac, G. (2011). Does there exist a copula in $n$ dimensions with a given correlation matrix? *International Conference on Analytical Methods in Statistics (AMISTAT 2011).* October 27-30, Prague, the Czech Republic.

Maronna, R., Martin, D., and Yohai, V. (2006). *Robust statistics. Theory and methods.* New York: Wiley.

Rousseeuw, P.J., and Croux, C. (1993). Alternatives to the Median Absolute Deviation. *Journal of American Statistical Association, 88*, 1273-1283.

Pashkevich, M.E., and Shevlyakov, G.L. (1995). The Median Analog of the Fourier Transform. In: *Proceedings of the CDAM 1995.* Minsk, Belarus. (in Russian)

Shevlyakov, G.L., Smirnov, P.O., Shin, V.I., and Kim, Kiseon. (2012). Asymptotically Minimax Bias Estimation of the Correlation Coefficient for Bivariate Independent Component Distributions. *Journal of the Multivariate Analysis, 111*, 59-65.

Shevlyakov, G.L., and Smirnov, P.O. (2011). Robust Estimation of the Correlation Coefficient: An Attempt of Survey. *Austrian Journal of Statistics, 40*, 147-156.

Smirnov, P.O., and Shevlyakov, G.L. (2010). On Approximation of the $Q_n$-Estimate of Scale by Fast $M$-Estimates. In: *Book of Astracts of the International Conference on Robust Statistics 2010 (ICORS 2010)*. Prague, the Czech Republic, pp. 94-95.

Spangl B., and Dutter R. (2005). On Robust Estimation of Power Spectra. *Austrian Journal of Statistics, 34*, 199-210.

Spangl B. (2008). *On Robust Spectral Density Estimation*. Dissertation. Technical University of Vienna.

**Affiliation:**

Georgy Shevlyakov, Nickolay Lyubomishchenko and Pavel Smirnov
St. Petersburg State Polytechnic University
St. Petersburg, Russia
E-mail: shev@gist.ac.kr

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 43/3-4 *Submitted:* 2013-11-25
June 2014 *Accepted:* 2014-03-10

Austrian Journal of Statistics
*June 2014, Volume 43/3-4, 247–254.*
*http://www.ajs.or.at/*

# Providing Data With High Utility And No Disclosure Risk For The Public and Researchers: An Evaluation By Advanced Statistical Disclosure Risk Methods

**Matthias Templ**
Vienna University of Technology

**Abstract**

The demand of data from surveys, registers or other data sets containing sensible information on people or enterprises have been increased significantly over the last years. However, before providing data to the public or to researchers, confidentiality has to be respected for any data set containing sensible individual information. Confidentiality can be achieved by applying statistical disclosure control (SDC) methods to the data. The research on SDC methods becomes more and more important in the last years because of an increase of the awareness on data privacy and because of the fact that more and more data are provided to the public or to researchers. However, for legal reasons this is only visible when the released data has (very) low disclosure risk.

In this contribution existing disclosure risk methods are review and summarized. These methods are finally applied on a popular real-world data set - the *Structural Earnings Survey* (SES) of Austria. It is shown that the application of few selected anonymisation methods leads to well-protected anonymised data with high data utility and low information loss.

*Keywords*: statistical disclosure control, data utility, disclosure risk, R.

## 1. Introduction

A microdata file is defined as a data set on individual level. For each observation a set of variables is typically available. Concerning SDC, these variables can be split into three categories.

- **Direct Identifiers:** Variables that definitely identify a statistical unit. For example, the social insurance number, name of companies or people or addresses are considered as direct identifiers.
- **Key variables:** A set of variables that - when considered together - may be used to identify an individual unit. For example with the combination of gender, age, region and occupation some individuals may be identified. Other examples for (confidential)

key variables could be income, health information, nationality or political preferences. For the description of the methods, it is advantageous to distinguish between categorical and continuous scaled key variables.

- **Non-confidential variables:** All variables that are not classified in any of the former two groups.

The goal of anonymizing a microdata set is to prevent that confidential information can be linked to a specific respondent. The ultimative aim is to release a safe microdata set that has both, low risk of linking confidential information to individual respondents and high data utility.

## 2. Measuring disclosure risk

Measuring risk in an microdata set is of course of great concern when having to decide on whether a microdata set is safe to be released. To be able to assess the disclosure risk it is required to make realistic assumptions on the information data users might have at hand to match against the microdata set. These assumptions are called 'disclosure risk scenarios'. Based on a specific disclosure risk scenario one must define a set of identifying variables (key variables) that can be used as input for the risk evaluation procedure.

Typically risk evaluation is based on the concept of "rareness/uniquenessss" in the sample and/or in the population. The interest is on units/individuals/observations that possess rare combinations of key variables. Those can be assumed to be identified easier and thus have higher risk. It is possible to cross tabulate all identifying variables and have a look at its cast. Patterns[1] with only very few individuals are in this sense considered risky if they have also low sampling weights, i.e. if the expected individuals with the same pattern is expected to be low in the population.

### 2.1. Frequencies counts

Consider a random sample of size $n$ drawn from a finite population of size $N$. Let $\pi_j, j = 1, \ldots, N$ be the (first order) inclusion probabilities, i.e. the probability that the element $u_j$ of a population of the size $N$ is chosen in a sample of the size $n$.

All possible combinations of categories in the key variables $X_1, \ldots, X_m$ can be calculated by cross tabulation of these categorical variables. Let $f_i, i = 1, \ldots, n$ be the frequency counts obtained by cross tabulation and let $F_i$ be the frequency counts of the population which belong to the same category. If $f_i = 1$ applies the corresponding observation is unique in the sample. If $F_i = 1$ applies then the observation is unique in the population. Note that $F_i$ is usually unknown since usually information on samples is collected and only few information about the population is known from registers and/or external sources.

### 2.2. The $k$-anonymity concept

Based on a set of key variables a desired characteristic of a protected microdata set might be to achieve $k$-anonymity (Samarati and Sweeney 1998; Sweeney 2002). This means that each possible combination of the values of the key variables features at least $k$ units in the microdata, meaning that all $f_i \geq 3$ , $i = 1, ..., n$ . A typical value is $k = 3$.

$k$-anonymity is typically provided by recoding categorical key variables and by additionally suppressing specific values in the key variables of individual units.

[1] a pattern is defined as a specific combination of values of all key variables

An extension of $k$-anonymity is $l$-diversity (Machanavajjhala, Kifer, Gehrke, and Venkitasubramaniam 2007). Consider for one group of observations with the same pattern in the key variables and let the group fulfill $k$-anonymity. A possible data intruder can therefore not identify an individual in this group. However, if all observations have the same entries in a sensitive variable (such as *cancer* in the variable *medical diagnosis*) then the attack is successful anyway.

### 2.3. Considering sample frequencies on subsets: SUDA2

SUDA (**S**pecial **U**niques **D**etection **A**lgorithm) estimates a disclosure risk for each individual. SUDA2 (see, e.g., Manning, Haglin, and Keane 2008) is a recursive algorithm for finding Minimal Sample Uniques. The algorithm generates all possible variable subsets of defined categorical key variables and scans them for unique patterns in the subsets of variables. The risk of an observation is then dependend on two aspects.

(a) The lower the amount of variables needed to receive uniquenesss, the higher the risk (and the higher the *suda score*) of the corresponding observation.

(b) The larger the number of minimal sample uniqueness contained within an observation, the higher the risk of the observation.

(a) is calculated for each observation $i$ by $l_i = \prod_{k=MSUmin_i}^{m-1}(m-k) \quad , i = 1, ..., n$, for $m$ the *depth* (the maximum size of variable subsets of the key variables), $MSUmin_i$ the number of minimal uniques of observation $i$ and $n$ the number of observations of the data set. Since each observation is treated independently, the $l_i$ that belongs to one pattern are summed up to result in a common suda score for each of the observation belonging to this pattern (this summation is the contribution of (b)).

To result in the final SUDA score, the suda score are normalized due division by $p!$, with $p$ being the number of key variables. The so called DIS suda score is then calculated from the suda and the so called DIS scores (we refer to Elliot 2000, for details). SUDA2 does not consider sampling weights and biased estimates may therefore result.

### 2.4. Considering population frequencies - the individual risk

To define if an individual unit is at risk, typically a threshold approach is used. If the individual risk of re-identification for an individual is above a certain threshold value, the unit is said to be at risk. To calculate the individual risks it is necessary to estimate the frequency of a given key in the population. In the previous section, Section 2.1, the population frequencies have already been estimated. However, one can show that these estimates almost always overestimate small population frequency counts (details can be found in Templ and Meindl 2010) and should not be used to estimate the disclosure risk.

A better approach is to use so-called super-population models in which population frequency counts are modeled given a certain distribution. The whole estimation procedure of sample counts given the population counts can be modeled, for example, by using a Negative Binomial distribution (see, e.g., Rinott and Shlomo 2006). It is out of scope of the paper to explain the final measurement of individual risk in this contribution but it can be found in Franconi and Polettini (2004) and Templ and Meindl (2010).

### 2.5. Measuring the global risk

Although the individual risk have to be respected since a data intruder should not be able to identify individuals, often also a measure of the global risk is estimated to express the risk of the whole data set with one number.

*Measuring the global risk based on the individual risks*

The first approach is to determine a threshold for the individual risk and to calculate the percentage of individuals that have larger individual risk than this threshold.

*Measuring the risk using log-linear models*

The sample frequencies, considered for each of $M$ patterns $m$, $f_m$ $, m = 1, ..., M$ can be modeled by a Poisson distribution, and the global risk may be defined as (see Skinner and Holmes 1998)

$$\tau_1 = \sum_{m=1}^{M} \exp\left(-\frac{\mu_m(1-\pi_m)}{\pi_m}\right) \quad , \text{ with } \mu_m = \pi_m \lambda_m \quad . \tag{1}$$

For simplicity, the inclusion probabilities are assumed to be equal, $\pi_m = \pi$ $, m = 1, ..., M$. $\tau_1$ can be estimated by log-linear models including the main effects and possible interactions. The model is

$$\log(\pi_m \lambda_m) = \log(\mu_m) = \mathbf{x}_m \beta \quad .$$

To estimate the $\mu_m$'s, the regression coefficients $\beta$ have to be estimated, for example, by using the iterative proportional fitting.Global risk measure 1 is then given by $\hat{r}_1 = \sum_{i=1}^{n} \mathbf{I}(f_i = 1) e^{-(1-\pi)\hat{\mu}}$ (corresponding to the risk $P(F_i = 1 | f_i = 1)$) and the second one by $\hat{r}_2 = \sum_{i=1}^{n} \mathbf{I}(f_i = 1) e^{1-(1-\pi)\hat{\mu}} / ((1-\pi)\hat{\mu})$ (corresponding to the risk $E(1/F_i | f_i = 1)$.

### 2.6. Measuring risk for continuous key variables

Applying the concept of uniquenesss and $k$-anonymity on quantitative variables results that every observation in the data set is unique. Hence, this approach will fail for continuous key variables.

If detailed information about a value of a continuous scaled variable is available, one may be able to identify (by linking information) and eventually gain further information about an individual. For continuous key variables it is assumed that an intruder has information about a statistical unit

*Distance-based record linkage*

By using distance based record linkage methods the aim is to find the nearest neighbors between observations from two data sets. Domingo-Ferrer and Torra (2001) has shown that these methods outperform probabilistic methods. Generally, it is evaluated if the original value falls within an interval centered on the masked value. Such an interval might be based on the standard deviation of the variable (see also Mateo-Sanz, Sebe, and Domingo-Ferrer 2004).

Almost all data sets from Official Statistics consists of statistical units whose values in at least one variable are quite different from the main part of the observations. This leads to the fact that these variables are very asymmetric distributed. Such outliers might be enterprises with a very large value for turnover, for example, or persons with extremely high income or even multivariate outliers. Other disclosure risk methods that are not used in this contribution take the "outlyingness" of an observation into account (for details, see, Templ and Meindl 2008).

## 3. Application to the statistics on earnings survey

The Structural Earnings Survey (SES) is conducted in almost all European countries and it includes variables on earnings of employees and other variables on employees and employment level (e.g. region, size of the enterprise, economic activities of the enterprise, gender and age of the employees, ...).

Generally such linked employer-employee data are used to identify the determinants/differentials of earnings but also some indicators are directly derived from the hourly earnings like the gender pay gap or the Gini coefficient (Gini 1912). The most classical example is the income inequality between genders as discussed in Groshen (1991), for example.

A correct identification of factors influencing the earnings could lead to relevant evidence-based policy decisions. The research studies are usually focused on examining the determinants of disparities in earnings.

The Austrian SES 2006 survey data consists of 199.909 observations obtained from a two-stage design - in the first stage of the design, the enterprises are chosen with certain inclusion probabilities depending on the enterprise size and location, in the second stage employee's in the selected enterprises are chosen with different inclusion probabilities (for more information have a look at Geissberger 2009).

### 3.1. Disclosure risk and information loss for SES

The following variables are chosen as key variables:

**Categorical key variables:** *size of enterprise* (5 ordered categories), *age* (66 ordered categories), *location* (3 categories), *economic activity* (53 categories)

**Continuous key variables:** *hourly earnings*, *earnings*

Table 1 shows the resulting disclosure risk and information loss of the SES data.

| risk, IL | orig | +rec1 | +rec2 | +rec3 | +supp | mdav | add | corr | sh |
|---|---|---|---|---|---|---|---|---|---|
| R:2-a | 2.49 | 0.47 | 0.24 | 0 | 0 | 0 | 0 | 0 | 0 |
| R:3-a | 5.65 | 1.12 | 0.56 | 0.01 | 0 | 0 | 0 | 0 | 0 |
| R:ind | 2.48 | 0.67 | 0.52 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 |
| R:suda | 0.87 | 0.15 | 0.1 | 0 | 0 | 0 | 0 | 0 | 0 |
| R:glob | 0.83 | 0.14 | 0.08 | 0 | 0 | 0 | 0 | 0 | 0 |
| R:glob | 1.35 | 0.23 | 0.13 | 0 | 0 | 0 | 0 | 0 | 0 |
| R:num | 100 | 100 | 100 | 100 | 100 | 99.73 | 7.86 | 61.86 | 12.26 |
| IL1 | - | - | - | - | - | 0 | 11.29 | 0.11 | 1.02 |
| IL:eig | - | - | - | - | - | 0 | 5.68 | 0.06 | 1.77 |
| IL:lm | 0 | 0.24 | 0.24 | 0.03 | 0.03 | 0.04 | 240.29 | 0.2 | 8.53 |

Table 1: Disclosure risk and information loss on SES

The the columns in Table 1 corresponds to the following data

**orig:** original data (key variables)

**rec1:** (recoding) the variable *economic activity* is recoded to 14 reasonable categories.

**rec2:** (recoding) additionally, the variable *size of employment* is recoded into three reasonable categories (10-49, 50-249, 250-...).

**rec3:** (recoding) additionally, age is discretised into six reasonable categories.

**supp:** (suppression) additionally, local suppression is applied so that no observation violates 3-anonymity.

**mdav:** microaggregation (method mdav, see e.g. Domingo-Ferrer and Mateo-Sanz (2002)) with aggregation level 3.

**add:** additive noise (noise parameter equals 10, see Templ, Kowarik, and Meindl (2013))

**corr:** correlated noise (defaults of Templ *et al.* (2013))

**sh:** shuffling (Muralidhar and Sarathy 2006)

The rows of Table 1 corresponds to disclosure risk and information loss measures - **R:2-a** (**R:3-a**): percent of observations violating 2(3)-anonymity, **R:ind**: percent of observations with individual risk below 0.01, **R:suda**: percent of observations having suda dis score lower than 0.1, **R:glob1**, **R:glob2**: global risks from log-linear models, **R:num**: distance-based disclosure risk, **IL1**: information loss IL1, **IL:eig**: information loss based on differences in the eigenvalues and **IL:lm**: model-based estimation information loss. The mentioned measures of information loss are briefly explained in the following.

**IL1:** $IL1 = \frac{1}{p}\sum_{i=1}^{p}\frac{|x_{ij}-x'_{ij}|}{\sqrt{2}S_j}$ , scaled distances between original and perturbed values for all $p$ continuous key variables.

**IL:eig:** The relative absolute differences between the eigenvalues of the covariance standardized continuous key variables of the original and the perturbed variables.

**IL:lm:** $|(\bar{\hat{y}}_w^o - \bar{\hat{y}}_w^p)/\bar{\hat{y}}_w^o|$, with $\bar{\hat{y}}_w$ the (Horwitz-Thompson) weighted mean of exponentials of the fitted values from the model $log(earningsHour) \sim age + Location + Sex + education + Occupation + economicActivity + Length + Size)$ (using weighted least squares estimation considering the sampling weights) obtained from the original (index $o$) and the perturbed data (index $p$).

Table 1 let us to the following interpretation. The original unmodified SES data contains about 5.35 % of observations that violate 3-anonymity and about 2.48 % of risky observations (using the individual risk approach). For the original data, the global model-based risk is 0.83 (and 1.35) which is quite similar to the percentage of observations having high dis suda score (0.87). Of course, the risk on continuous key variables is 100 % and the information loss on that variables is zero. When recoding *economic activity* into less categories, the risk reduces by almost the factor of 5. When additionally recoding the variable *age* the risk reduces dramatically. After applying local suppression additionally, the risk for all risk methods zero, expect the individual risk.

The risk on continuous variables is evaluated for any method independently. It is very low for adding additive noise to the data but in the same time the information loss is inacceptable large. The information loss is very small for adding correlated noise, but the risk is still high. For microaggregation, the information loss is (almost) zero, but the risk is high. However, always three observations are aggregated and therefore anonymisation might be fine but the disclosure risk method is not suitable for microaggregation. The performance of shuffling is good, but the model based estimates differ more than 8 % after shuffling the data.

Probably the most interesting information loss measure - the measure which accounts for fitting a linear model on the data (IL:lm) reports that the information loss very low expect for the adding additive noise method and shuffling.

## 4. Conclusion

In this contribution, popular disclosure risk methods have been summarized. We stressed to measure the disclosure risk after the application of any SDC method to the data. Because of the limit of pages we only briefly focused on measuring the data utility and information loss, but it should be clear that the aim is both, to provide a data set with low disclosure risk and high data utility.

In the practical example, a very popular data set was used and the disclosure risk and data utility/information loss is evaluated. Hereby, the whole range of disclosure risk methods has

been applied to the data, which is done the first time to our knowledge. The results show that by application of few selected anonymisation methods, the disclosure risk dramatically decreases and in the same time, the information loss is considerable small.

All estimations/calculations have been made with the R-package **sdcMicro** Templ *et al.* (2013). The SES data were provided by Statistics Austria.

## References

Domingo-Ferrer J, Mateo-Sanz J (2002). "Practical data-oriented microaggregation for statistical disclosure control." *IEEE Trans. on Knowledge and Data Engineering*, **14**(1), 189–201.

Domingo-Ferrer J, Torra V (2001). "A quantitative comparison of disclosure control methods for microdata." In *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, pp. 111–134.

Elliot M (2000). "DIS: A New Approach to the Measurement of Statistical Disclosure Risk." *Risk Management*, **2**(4), 39–48.

Franconi L, Polettini S (2004). "Individual risk estimation in $\mu$-Argus: a review." In J In: Domingo-Ferrer (ed.), *Privacy in Statistical Databases, Lecture Notes in Computer Science*, pp. 262–272. Springer.

Geissberger T (2009). *Verdienststrukturerhebung 2006, Struktur und Verteilung der Verdienste in Österreich*. Statistik Austria. ISBN 978-3-902587-97-8.

Gini C (1912). "Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche." *Studi Economico-Giuridici della R. Università di Cagliari*, **3**, 3–159.

Groshen E (1991). "The structure of the female/male wage differential." *Journal of Human Resources*, **26**, 455–472.

Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M (2007). "L-diversity: Privacy beyond k-anonymity." *ACM Trans. Knowl. Discov. Data*, **1**(1). ISSN 1556-4681. doi:10.1145/1217299.1217302. URL http://doi.acm.org/10.1145/1217299.1217302.

Manning A, Haglin D, Keane J (2008). "A recursive search algorithm for statistical disclosure assessment." *Data Mining and Knowledge Discovery*, **16**(2), 165–196. ISSN 1384-5810. doi:10.1007/s10618-007-0078-6. URL http://dx.doi.org/10.1007/s10618-007-0078-6.

Mateo-Sanz J, Sebe F, Domingo-Ferrer J (2004). "Outlier protection in continuous microdata masking." *Lecture Notes in Computer Science, Vol. Privacy in Statistical Databases, Springer Verlag*, **3050**, 201–215.

Muralidhar K, Sarathy R (2006). "Data Shuffling- A New Masking Approach for Numerical Data." *Management Science*, **52**(2), 658–670.

Rinott Y, Shlomo N (2006). "A generalized Negative Binomial smoothing model for sample disclosure risk estimation." In *Privacy in Statistical Databases. Lecture Notes in Computer Science. Springer*, pp. 82–93.

Samarati P, Sweeney L (1998). "Protecting privacy when disclosing information: $k$-anonymity and its enforcement through generalization and suppression." *Technical Report SRI-CSL-98-04*, SRI International.

Skinner C, Holmes D (1998). "Estimating the re-identification risk per record in microdata." *Journal of Official Statistics*, **14**, 361–372.

Sweeney L (2002). "$k$-anonymity: a model for protecting privacy." *Int J Uncertain Fuzziness Knowl Syst*, **10**(5), 557–570.

Templ M, Kowarik A, Meindl B (2013). *sdcMicro: Statistical Disclosure Control methods for the generation of public- and scientific-use files. Manual and Package.* R package version 4.1.0, URL http://CRAN.R-project.org/package=sdcMicro.

Templ M, Meindl B (2008). "Robust Statistics Meets SDC: New Disclosure Risk Measures for Continuous Microdata Masking." *Privacy in Statistical Databases. Lecture Notes in Computer Science. Springer*, **5262**, 113–126. ISBN 978-3-540-87470-6, DOI 10.1007/978-3-540-87471-3_10.

Templ M, Meindl B (2010). "Practical Applications in Statistical Disclosure Control Using R." In J Nin, J Herranz (eds.), *Privacy and Anonymity in Information Management Systems*, Advanced Information and Knowledge Processing, pp. 31–62. Springer London. ISBN 978-1-84996-238-4. 10.1007/978-1-84996-238-4_3, URL http://dx.doi.org/10.1007/978-1-84996-238-4_3.

**Affiliation:**

Matthias Templ
Department of Statistics and Probability Theory
Vienna University of Technology
Wiedner Hauptstr. 8–10
A-1040 Vienna, Austria &

Methods Unit
Statistics Austria
Guglgasse 13,

A-1110 Vienna, Austria
E-mail: matthias.templ@gmail.com
URL: http://www.statistik.tuwien.ac.at/public/templ

Austrian Journal of Statistics
*June 2014, Volume 43/3-4, 255–266.*
*http://www.ajs.or.at/*

# Software Tools for Robust Analysis of High-Dimensional Data

**Valentin Todorov**
UNIDO

**Peter Filzmoser**
Vienna University of Technology

**Abstract**

The present work discusses robust multivariate methods specifically designed for high dimensions. Their implementation in R is presented and their application is illustrated on examples. The first group are algorithms for outlier detection, already introduced elsewhere and implemented in other packages. The value added of the new package is that all methods follow the same design pattern and thus can use the same graphical and diagnostic tools. The next topic covered is sparse principal components including an object oriented interface to the standard method proposed by Zou, Hastie, and Tibshirani (2006) and the robust one proposed by Croux, Filzmoser, and Fritz (2013). Robust partial least squares (see Hubert and Vanden Branden 2003) as well as partial least squares for discriminant analysis conclude the scope of the new package.

*Keywords*: high dimensions, robustness, classification, PLS, PCA, outliers.

## 1. Introduction

High-dimensional data are typical in many contemporary applications in scientific areas like genetics, spectral analysis, data mining, image processing, etc. and introduce new challenges to the traditional analytical methods. First of all, the computational effort for the anyway computationally intensive robust algorithms increases with increasing number of observations $n$ and number of variables $p$ towards the limits of feasibility. Some of the robust multivariate methods available in R (see Todorov and Filzmoser 2009) are known to deteriorate rapidly when the dimensionality of data increases and others are not applicable at all when $p$ is larger than $n$.

The present work discusses robust multivariate methods specifically designed for high dimensions. Their implementation in R is presented and their application is illustrated on examples. A key feature of this extension of the framework is the object model which follows the one already introduced by **rrcov** and based on statistical design patterns. The first group of classes are algorithms for outlier detection, already introduced elsewhere and implemented in other packages. The value added of the new package is that all methods follow the same pattern and thus can use the same graphical and diagnostic tools. The next topic covered is sparse principal component analysis including an object oriented interface to the standard method proposed by Zou *et al.* (2006) and the robust one proposed by Croux *et al.* (2013). These

are presented and illustrated in Section 2. Robust partial least squares (Hubert and Vanden Branden 2003; Sernels, Croux, Filzmoser, and van Espen 2005) as well as partial least squares for discriminant analysis are presented in Section 3 and Section 4. Section 5 concludes.

## 2. Robust sparse principal component analysis

Principal component analysis (PCA) is a prominent technique for dimension reduction, and the principle is to find a smaller number $q$ of linear combinations of the originally observed $p$ variables while retaining most of the variability of the data. Dimension reduction by PCA is mainly used for: (i) visualization of multivariate data by scatter plots (in a lower dimensional space); (ii) transformation of highly correlated variables into a smaller set of uncorrelated variables which can be used by other methods (e.g. multiple or multivariate regression, linear or quadratic discriminant analysis); (iii) combination of several variables characterizing a given process into a single or a few *characteristic* variables or *indicators*. In some cases–in particular if the original variables have physical meaning–it is important to be able to interpret these new variables. The interpretation of the principal components needs to be based on the *loadings matrix*, which links the original variables with the principal components.

The standard approach to PCA identifies new directions which are linear combinations of the original variables in such a way, that the data projected on these directions have maximal variance. The different directions need to be orthogonal to each other, and the variance measure used for classical PCA is the empirical sample variance. Practically, the PCA directions can be found by computing the eigenvectors of the sample covariance or correlation matrix. The disadvantage of this approach is that outlying observations may even artificially increase the variance measure, thus leading to essentially uninformative directions. In other words, outliers may attract PCA directions, and the pattern of the data majority will not be covered by the few extracted classical components.

In contrast, the goal of robust PCA is to retain as much of the information of the data majority (and not of single outliers) as possible with fewer directions–the robust PCs. Different approaches to robust PCA are discussed in many review papers, see for example Todorov and Filzmoser (2009) and Filzmoser and Todorov (2013), and examples are given how these robust analyses can be carried out in R. Details about the methods and algorithms can be found in the corresponding references. However, PCA usually tends to provide PCs which are linear combinations of *all* the original variables, even if some of the loadings are small in absolute size. This is a disadvantage for high-dimensional data analysis, since PC directions will then in general be affected by all the variables, even if they are noise variables. It would be more useful to have a method which completely suppresses the influence of potential noise variables by assigning loadings of exactly zero to them. This is the goal of sparse PCA, and there are several proposals available nowadays. A straightforward informal method is to set to zeros those PC loadings which have absolute values below a given threshold (*simple thresholding*). Jolliffe, Trendafilov, and Uddin (2003) proposed SCoTLASS which applies a *lasso* penalty on the loadings in a PCA optimization problem, and recently Zou *et al.* (2006) reformulated PCA as a regression problem and used the *elastic net* to obtain a sparse version - SPCA.

The above mentioned proposals for sparse PCA are not robust with respect to outlying observations. They suffer from the same problem as classical (non-sparse) PCA, namely that the new directions will be attracted by outliers. To cope with the possible presence of outliers in the data, recently Croux *et al.* (2013) proposed a method which is sparse and robust at the same time. It utilizes the *projection pursuit* approach where the PCs are extracted from the data by searching the directions that maximize a robust measure of variance of data projected on it. An efficient computational algorithm was proposed by Croux, Filzmoser, and Oliveira (2007).

**Example** Sparse classical and robust PCA is illustrated here by the (low-dimensional) `cars` data set (Consumer Reports 1990, pp. 235–288); (Chambers and Hastie 1992, pp. 46–47), which is available in the package **rrcovHD** as the data frame `cars`. For $n = 111$ cars, $p = 11$ characteristics were measured, including the length, width, and height of the car. After looking at pairwise scatterplots of the variables, and computing pairwise Spearman rank correlations $\rho(X_i, X_j)$ we see that there are high correlations among the variables, for example, $\rho(X_1, X_2) = .83$ and $\rho(X_3, X_9) = .87$. Thus, PCA will be useful for reducing the dimensionality of the data set (see also Hubert, Rousseeuw, and Vanden Branden (2005)). The first four classical PCs explain more than 96% of the total variance and the first four robust PCs explain more than 95%, therefore we decide to retain four components in both cases. Next we need to choose the degree of sparseness which is controlled by a regularization parameter ($\lambda$). With sparse PCA we take a trade-off between sparseness of the loadings matrix and maximization of the explained variability. The appropriate tuning parameter can be chosen by computing sparse PCA for many different values of $\lambda$ and plotting the percentage of explained variance against $\lambda$. We choose $\lambda = 0.78$ for classical PCA and $\lambda = 2.27$ for robust PCA, thus attaining 83 and 82 percent of explained variance, respectively, which is only an acceptable reduction compared to the non-sparse PCA. Retaining $k = 4$ principal components as above and using the selected parameters $\lambda$, we can construct the so called *diagnostic plots* which are especially useful for identifying outlying observations. The *diagnostic plot* shows the orthogonal distances versus the score distance, and indicates with a horizontal and vertical line the cut-off values that allow to distinguish regular observations (those with small score and small orthogonal distance) from the different types of outliers: *bad leverage points* with large score and large orthogonal distance, *good leverage points* with large score and small orthogonal distance and *orthogonal outliers* with small score and large orthogonal distance (for detailed description see Hubert *et al.* (2005)). In Figure 1 the classical and robust diagnostic plot as well as their sparse counterparts are presented. The diagnostic plot for classical PCA reveals only several orthogonal outliers and identifies two observations as bad leverage points. Three more observations are identified as bad leverage points by sparse classical PCA which is already an improvement, but only the robust methods identify a large cluster of outliers. These outliers are masked by the non-robust score and orthogonal distances and cannot be identified by the classical methods. It is important to note that the sparsity feature added to the robust PCA did not influence its ability to detect properly the outliers.

## 3. Robust linear regression in high dimensions

The toolbox of linear regression methods and their robust counterparts becomes limited when the number of explanatory variables $p$ exceeds the number of observations $n$. The matrix of explanatory variables $\boldsymbol{X}$ is then said to be "flat". In that case, partial least squares (PLS) regression is known to work very well, in particular if the explanatory variables are highly correlated. In this section we will focus on PLS regression and robust versions thereof, since these are widely used tools in various areas.

PLS regression Wold (1975) can be used for the case of a univariate response (PLS1) as well as for a matrix of response variables (PLS2), here denoted by the $n \times q$ matrix $\boldsymbol{Y}$. In the latter case, the regression problem is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}, \tag{1}$$

with the regression coefficient matrix $\boldsymbol{B}$ and the errors $\boldsymbol{E}$. The basic idea is to decompose $\boldsymbol{X}$ and $\boldsymbol{Y}$ as follows,

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^\top + \boldsymbol{E}_{\boldsymbol{X}} \tag{2}$$

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{Q}^\top + \boldsymbol{E}_{\boldsymbol{Y}}, \tag{3}$$

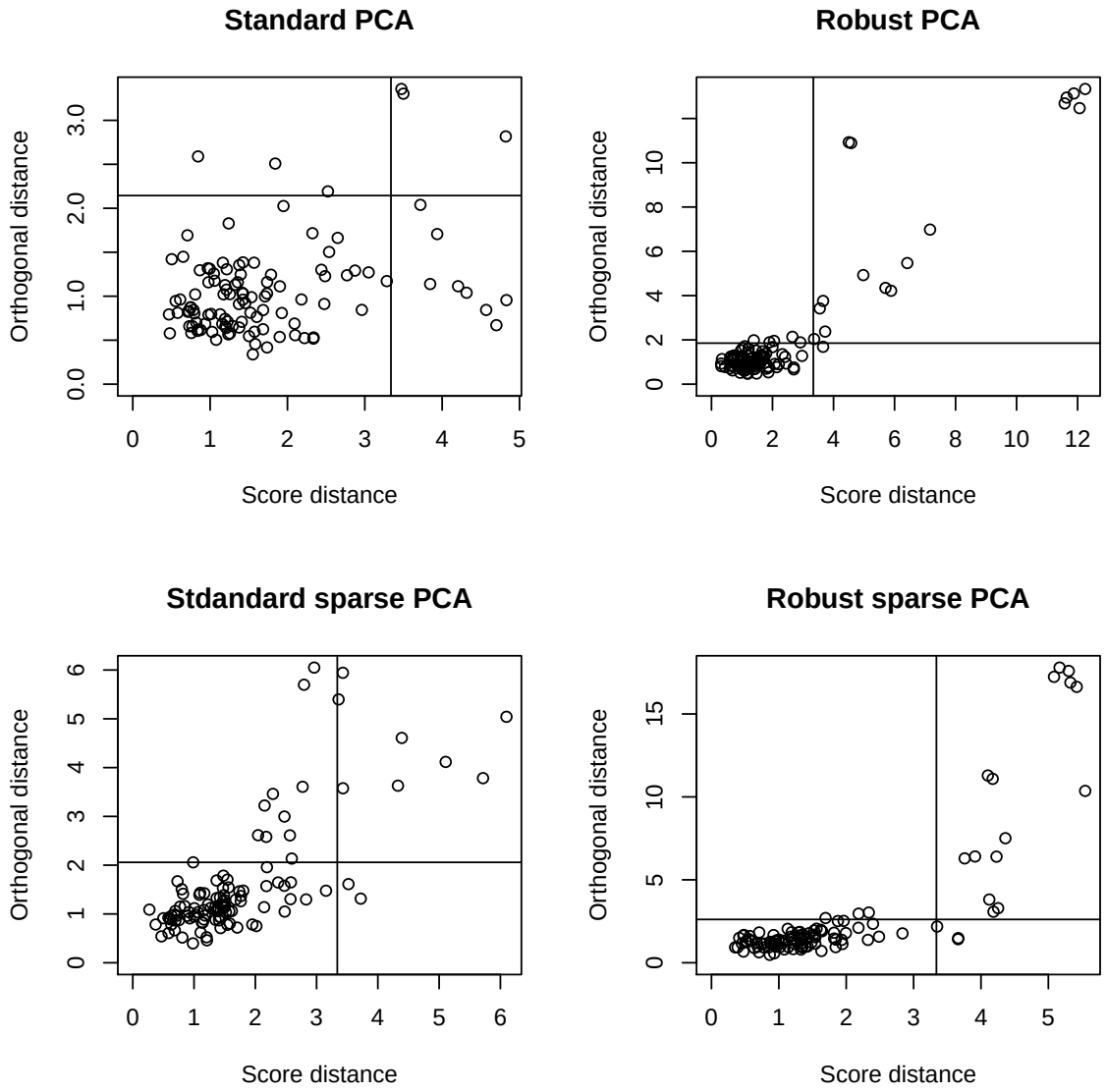


Figure 1: Distance-distance plots for standard and sparse PCA and their robust versions for the `cars` data.

with the scores matrices $\boldsymbol{T}$ and $\boldsymbol{U}$, and the loadings matrices $\boldsymbol{P}$ and $\boldsymbol{Q}$, each having $K$ columns, and the error matrices $\boldsymbol{E_X}$ and $\boldsymbol{E_Y}$. The number of components $K$ for the factorization is limited with $K \leq \min\{n, p, q\}$. The *inner relationship* connecting the scores is given by

$$\boldsymbol{U} = \boldsymbol{T}\boldsymbol{D} + \boldsymbol{H}, \tag{4}$$

with the diagonal matrix $\boldsymbol{D}$ and a residual matrix $\boldsymbol{H}$.

The key idea in PLS regression is to find a direction $\boldsymbol{w}$ in the $x$-space and a direction $\boldsymbol{c}$ in the $y$-space such that

$$\operatorname{cov}(\boldsymbol{X}\boldsymbol{w}, \boldsymbol{Y}\boldsymbol{c}) \longrightarrow \max \quad \text{with} \quad \|\boldsymbol{t}\| = \|\boldsymbol{X}\boldsymbol{w}\| = 1 \quad \text{and} \quad \|\boldsymbol{u}\| = \|\boldsymbol{Y}\boldsymbol{c}\| = 1, \tag{5}$$

where "cov" is an estimator for the covariance. The resulting $\boldsymbol{t}$ and $\boldsymbol{u}$ then form a column in the matrix $\boldsymbol{T}$ and $\boldsymbol{U}$, respectively.

The above procedure is carried out in a sequential manner. This means that the score vectors are computed one after the other, until $K$ vectors are extracted, hereby imposing appropriate constraints (e.g. uncorrelatedness). There are different proposals to solve problem (5), like

the NIPALS algorithm, the kernel algorithm, or the SIMPLS algorithm. For details we refer to Varmuza and Filzmoser (2009).

Hubert and Vanden Branden (2003) suggested a robust version of the SIMPLS algorithm. Since this algorithm is based on estimates of the covariance matrix of the $x$-variables, and of the joint covariance matrix between the $x$- and the $y$-variables, a first step is to robustify these estimates by employing robust PCA. In a second step, a multivariate robust regression method is used.

In case of PLS1, Sernels *et al.* (2005) proposed a robust version that is called partial robust M (PRM) regression. The main idea is to perform robust regression using an M-estimator of the response $\boldsymbol{y}$ on latent variables which are summarizing the explanatory variables. These latent variables, representing only partial information of the $x$-variables, are found in the same spirit as shown in criterion (5),

$$\text{cov}(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{a}) \longrightarrow \max, \tag{6}$$

with appropriate constraints on the loadings vector $\boldsymbol{a}$, and a robust estimator for "cov" using a certain weighting scheme for outlying observations. The loadings and scores are extracted sequentially, again with appropriate side constraints (see also Filzmoser and Todorov 2011).

**Example** To illustrate robust PLS regression we use a real data example, known from other studies on robust methods. The data set originates from 180 glass vessels Janssens, Deraedt, Freddy, and Veeckman (1998) and was analyzed also in Sernels *et al.* (2005); Hubert, Rousseeuw, and van Aelst (2008); Filzmoser, Maronna, and Werner (2008). In total, 1920 characteristics are available for each vessel, coming from an analysis by an electron-probe X-ray micro-analysis. The data set includes four different materials comprising the vessels, and we focus on the material forming the larger group of 145 observations. It is known from other studies on this data set that these 145 observations should form two groups, because during the measurement process the detector efficiency has been changed. In the original analysis, univariate PLS calibration was performed for all of the main constituents of the glass but here we will consider only the prediction of the sodium oxide concentration and will carry out classical (SIMPLS) and robust (RSIMPLS) PLS with $K = 8$ components. Since the response variable is univariate, regression diagnostic plots for both classical and robust PLS can be created, as shown in Figure 2. The vertical axis represents the standardized residuals $r_i/\hat{\sigma}$

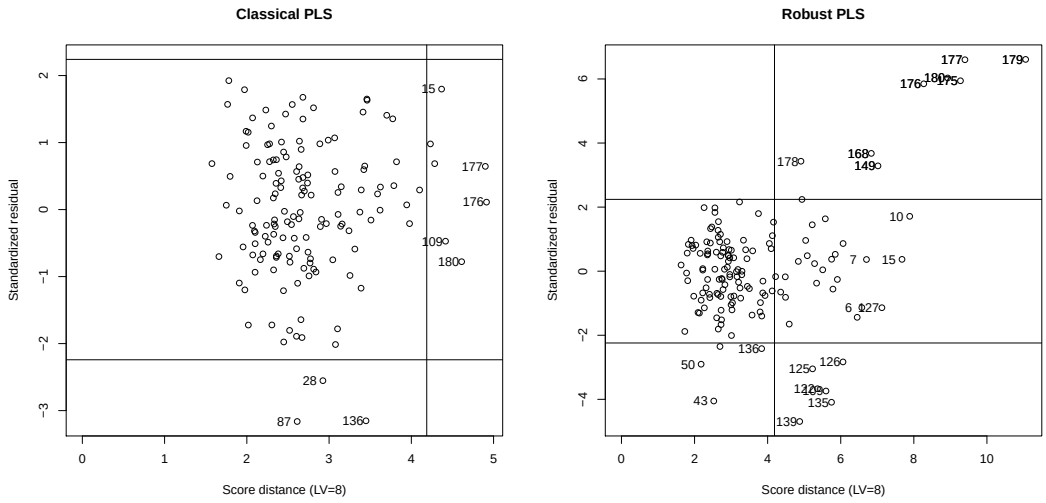


Figure 2: Regression diagnostic plots for the glass data set with (left) SIMPLS and (right) RSIMPLS.

with $r_i = y_i - \hat{\boldsymbol{\beta}}\boldsymbol{x}_i$ while on the horizontal axis the Mahalanobis distances of the data points

in the score space (therefore called *score distances*) are displayed. Outliers in the $t$-space are identified as data points with score distances exceeding the cutoff value of $\sqrt{\chi^2_{K,0.975}}$. Data points which have an absolute standardized residual exceeding $\sqrt{\chi^2_{1,0.975}}$ are flagged as regression outliers. The SIMPLS regression diagnostic plot identifies only three observations as regression outliers and several more as outlying according to the score distances while the robust plot identifies most of the outliers know from other studies. A detailed definition of

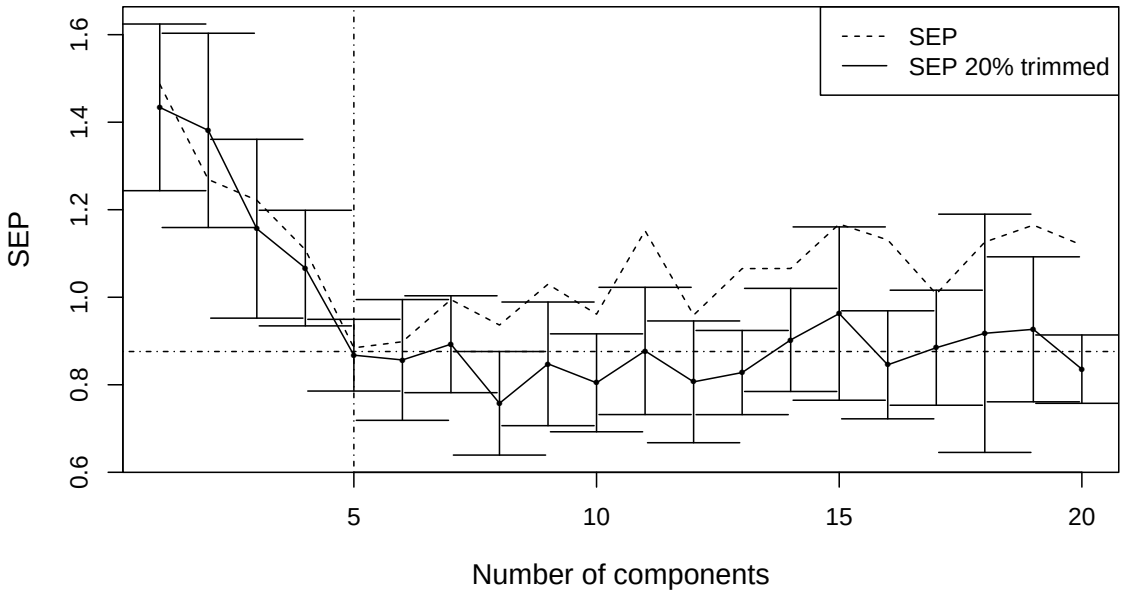


Figure 3: Results of 10-fold cross-validation for robust PLS for the glass data set. A model with 5 components is optimal.

this plot as well as its version for multivariate response variable, can be found in Hubert and Vanden Branden (2003).

For choosing an optimal number of PLS components, 10-fold cross-validation (CV) is used for a maximum of e.g. 20 components and the result is presented graphically in Figure 3. As a performance measure the *standard error of prediction* (SEP) value is used, and its 20% trimmed version.

$$\text{SEP} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\sum_{j=1}^{K}(y_{ij} - \hat{y}_{ij} - \text{bias})^2} \quad \text{with} \quad \text{bias} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{q}(y_{ij} - \hat{y}_{ij}). \tag{7}$$

Here, $\{\hat{y}_{ij}\} = \hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{B}}$ are the predicted values of the response variable, using the estimated regression parameters $\hat{\boldsymbol{B}}$ (see Varmuza and Filzmoser 2009). Note that the performance measure in (7) is not robust against outliers, because each observation gets the same contribution in the formulas. The influence of outliers to the performance measure can be reduced by trimming for example the 20% of the largest contributions. The dashed line presents the mean of SEP values from CV and the solid part presents the mean and standard deviation of 20% trimmed SEP values from CV. The vertical and horizontal lines correspond to the optimal number of components (after standard-error-rule) and the corresponding 20% trimmed SEP mean, respectively. The optimal number of components is selected as the lowest number whose prediction error mean is below the minimal prediction error mean plus one standard

error, see Varmuza and Filzmoser (2009). Here, 5 components are selected, leading to a prediction error of 0.95.

A more detailed model selection can be done with repeated double cross-validation (rdCV) (see Filzmoser, Liebmann, and Varmuza (2009); Liebmann, Filzmoser, and Varmuza (2010) for details). However, the procedure is rather time consuming. Within an "inner loop", $k$-fold CV is used to determine an optimal number of components, which then is applied to a "test set" resulting from an "outer loop". The procedure is repeated a number of times. The frequencies of the optimal numbers of components are shown in Figure 4. There is a clear peak at 5 components, meaning that a model with 5 components has been optimal in most of the experiments within rdCV. Note that here we obtain the same result as for single CV. In

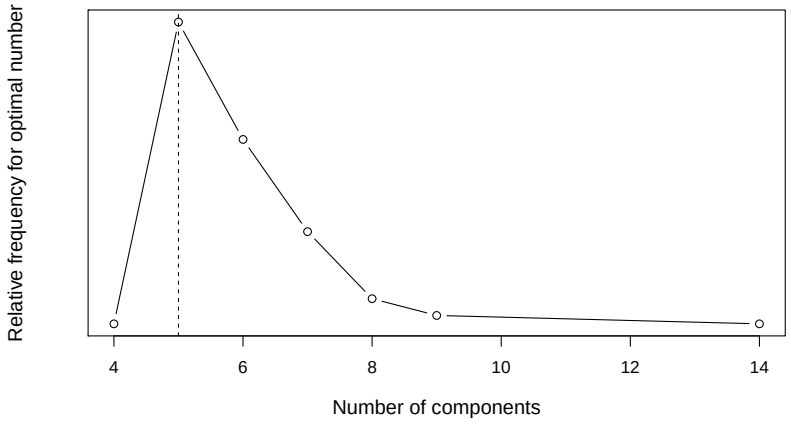


Figure 4: Results of rdCV of RSIMPLS. The optimal number of components is indicated by the vertical dashed line.

a next plot, Figure 5, the prediction performance measure, the 20% trimmed SEP, is shown. The gray lines correspond to the results of the 20 repetitions of the double CV scheme, while the black line represents the single CV result. Obviously, single CV is much more optimistic than rdCV. The estimated prediction error for 5 components is 0.85. Using the optimal number of 5 components, predictions and residuals can be computed. However, for rdCV there are predictions and residuals available for each replication (we used 20 replications). The diagnostic plot shown in 6 presents the predicted versus measured response values. The left panel is the prediction from a single CV, while in the right panel the resulting predictions from rdCV are shown. The latter plot gives a clearer picture of the prediction uncertainty. A similar plot can be generated for predicted values versus residuals (not shown here).

## 4. Robust classification in high dimensions

The prediction of group membership and/or describing group separation on the basis of a data set with known group labels (training data set) is a common task in many applications and linear discriminant analysis (LDA) has often been shown to perform best in such classification problems. However, very often the data are characterized by far more variables than objects and/or the variables are highly correlated which renders LDA (and the other similar standard methods) unusable due to their numerical limitations. Let us assume that $\boldsymbol{Y}$ is univariate and categorical, i.e. $\forall i, 1 \leq i \leq n : y_i \in \{1, \ldots, G\}$ where $G$ is the number of groups. For high dimensional data sets, classical linear discriminant analysis cannot be performed due to the singularity of the estimated covariance matrix $\hat{\boldsymbol{\Sigma}}$, as it requires the inverse of $\hat{\boldsymbol{\Sigma}}$. To overcome the high dimensionality problem in classification context one can reduce the dimensionality by either selecting a subset of "interesting" variables (variable selection)

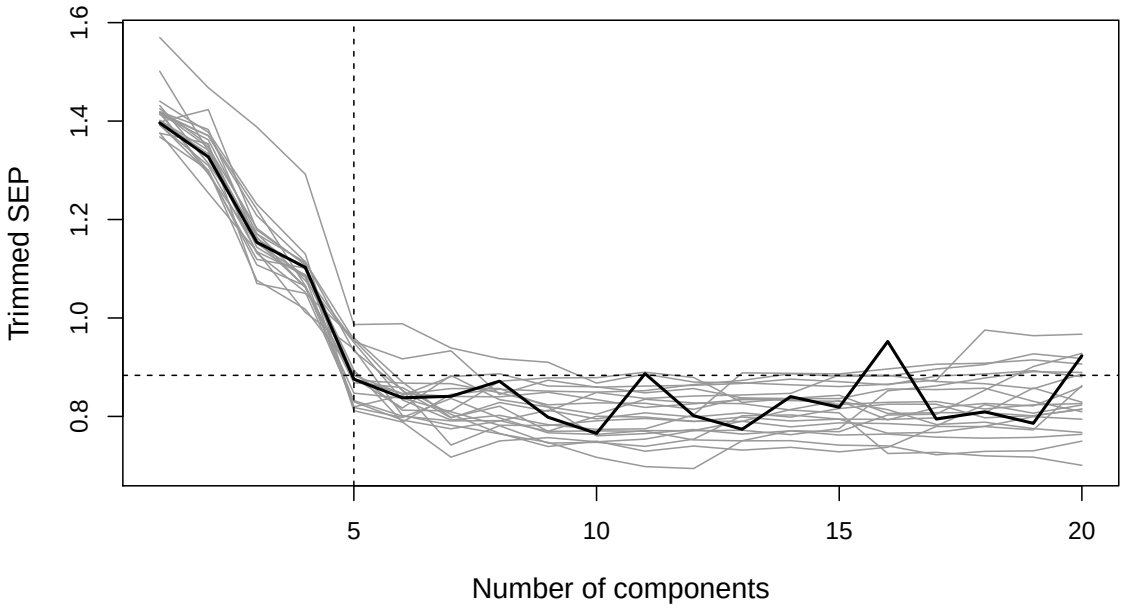


Figure 5: Results of rdCV for RSIMPLS. The gray lines result from repeated double CV, the black line from single CV.

or construct $K$ new components, $K \ll p$ which represent the original data with minimal loss of information (feature extraction, dimension reduction). Many methods for dimension reduction were considered in the literature but the two most popular are principal component analysis (PCA) and partial least squares (PLS). It is intuitively clear that a supervised method (which uses the group information while constructing the new components) like PLS should be preferred to unsupervised methods like PCA.

**SIMCA:** Instead of applying the dimension reduction method (e.g. PCA) to the full set of observations, one could fit a model to each of the groups (possibly with different number of components) and use these models to classify new observations. This method, called *Soft Independent Modeling of Class Analogies* (SIMCA), was introduced by Wold (1976) and nowadays is widely used as a discriminant technique in chemometrics, where typically $p$ is large relative to $n$. Since in SIMCA PCA is performed on each group, it provides additional information on the different groups, including the relevance of the different variables for groups separation, i.e. their discrimination power. In the original SIMCA method new observations are classified based on their deviation from the different PCA models. These deviations are the Euclidean distances of the observations to the PCA subspace, and thus they are called orthogonal distances. Vanden Branden and Hubert (2005) propose a slightly modified classification rule which better exploits the benefit of applying PCA to each group. This rule includes additionally the score distances, i.e. the Mahalanobis distances measured in the PCA (score) subspace. Furthermore, as a guard against outliers in the data, they propose to replace the classical PCA by a robust alternative. Both the classical and the robust version of the SIMCA method are available in the R package **rrcovHD**.

**Robust PLS-DA:** PLS was not originally designed to be used in the context of statistical discrimination but nevertheless was routinely applied with evident success by practitioners for this purpose. Taking into account the grouping variable(s) when decomposing the data one

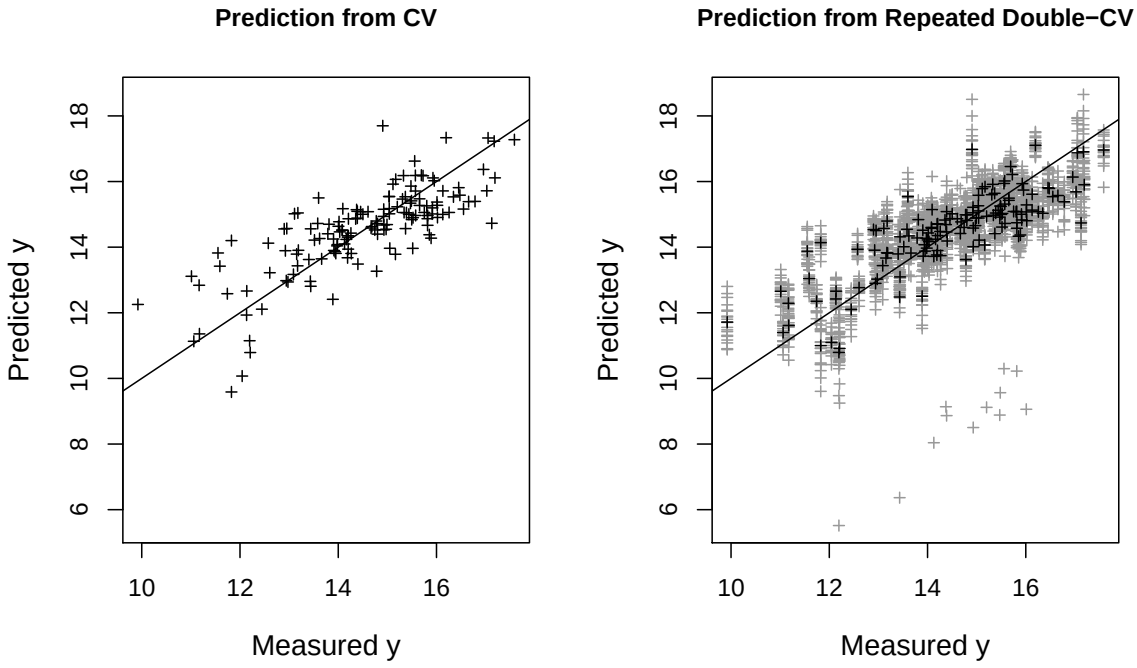


Figure 6: Predicted versus measured response values for RSIMPLS. The left panel shows the results from single CV, the right panel visualizes the results from repeated double CV.

would intuitively expect an improved performance for group separation. Since the response variable in case of a classification problem is a categorical variable, none of the robust PLS methods proposed above can be used. Therefore, in order to obtain a robust PLS-DA we proposed to apply any of the outlier detection methods described in Filzmoser and Todorov (2013), which are implemented in package **rrcovHD**, and then use classical PLS on the already cleaned data set. Hubert and Van Driessen (2004) used a data set containing the spectra of three different cultivars of the same fruit. The three cultivars (groups) are named D, M and HA, and their sample sizes are 490, 106 and 500 observations, respectively. The spectra are measured at 256 wavelengths. The fruit data is thus a high-dimensional data set which was used to illustrate a new approach for robust linear discriminant analysis, and it was studied again by Vanden Branden and Hubert (2005). From these studies it is known that the first two cultivars D and M are relatively homogenous and do not contain atypical observations, but the third group HA contains a subgroup of 180 observations which were obtained with a different illumination system. In Figure 7 are shown the prediction histograms for class D for the `fruit` data using classical and robust PLS-DA.

## 5. Summary and conclusions

An object oriented framework for robust multivariate analysis developed in the S4 class system of the programming environment R already exists implemented in the package **rrcov** and is described in Todorov and Filzmoser (2009). The main goal of this framework is to support the usage, experimentation, development and testing of robust multivariate methods as well as simplifying comparisons with related methods. In this article we investigated several robust multivariate methods specifically designed for high dimensions. The focus was on PCA and its sparse version, PLS, PLS for discrimination, and SIMCA. All considered methods and data sets are available in the R package **rrcovHD**. A key feature of this extension of the framework is that the object model follows the one already introduced by **rrcov** which is

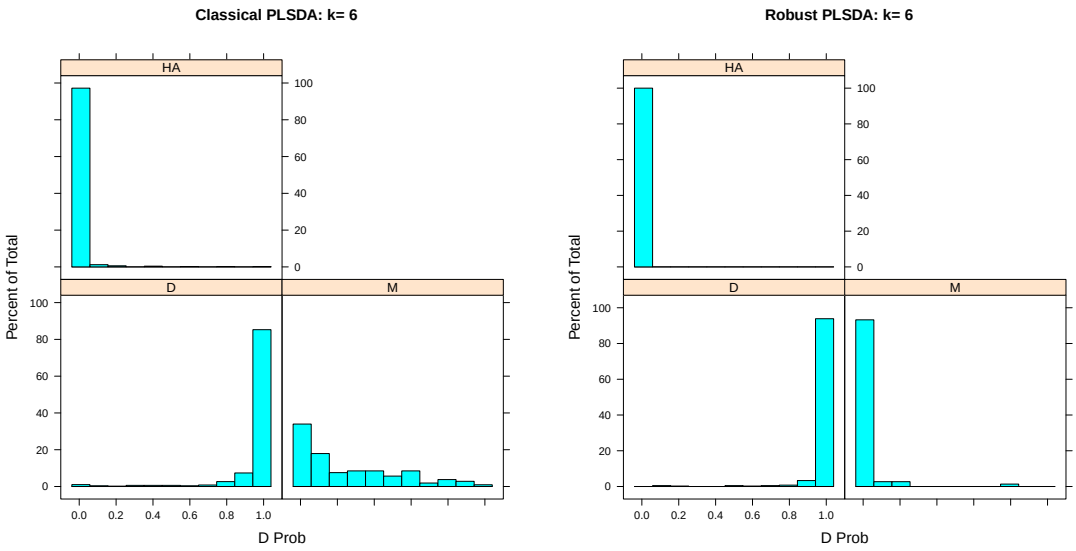


Figure 7: Prediction histograms for class D for the `fruit` data using classical and robust PLS-DA.

based on statistical design patterns. This makes it easy for the user to apply the methods, since they are following the same structure. A further advantage is that summaries, results, as well as diagnostic plots follow the same structure.

Finally, the strict design pattern used in the package **rrcovHD** is an advantage for extending the package with other methods developed for high-dimensional data–and for sure their robust versions will follow.

## Acknowledgements

The views expressed herein are those of the authors and do not necessarily reflect the views of the United Nations Industrial Development Organization.

## References

Chambers JM, Hastie TJ (1992). *Statistical Models in S*. Wadsworth and Brooks, Cole, Pacific Grove, CA.

Consumer Reports (1990). "Annual Auto Issue: The 1990 cars." http://backissues.com/issue/Consumer-Reports-April-1990. April.

Croux C, Filzmoser P, Fritz H (2013). "Robust Sparse Principal Component Analysis." *Technometrics*, **55**(2), 202–214.

Croux C, Filzmoser P, Oliveira M (2007). "Algorithms for Projection-pursuit Robust Principal Component Analysis." *Chemometrics and Intelligent Laboratory Systems*, **87**(218), 218–225.

Filzmoser P, Liebmann B, Varmuza K (2009). "Repeated Double Cross Validation." *Journal of Chemometrics*, **23**(4), 160–171.

Filzmoser P, Maronna R, Werner M (2008). "Outlier Identification in High Dimensions." *Computational Statistics and Data Analysis*, pp. 1694–1711.

Filzmoser P, Todorov V (2011). "Review of Robust Multivariate Statistical Methods in High Dimension." *Analytica Chimica Acta*, **705**, 2–14.

Filzmoser P, Todorov V (2013). "Robust Tools for the Imperfect World." *Information Sciences*, **245**, 4–20.

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

Hubert M, Rousseeuw PJ, van Aelst S (2008). "High-Breakdown Robust Multivariate Methods." *Statistical Science*, **23**, 92–119.

Hubert M, Van Driessen K (2004). "Fast and Robust Discriminant Analysis." *Computational Statistics & Data Analysis*, **45**, 301–320.

Hubert M, Vanden Branden K (2003). "Robust Methods for Partial Least Squares Regression." *Journal of Chemometrics*, **17**(10), 537–549.

Janssens K, Deraedt I, Freddy A, Veeckman J (1998). "Composition of 15-17$^{th}$ Century Archæological Glass Vessels Excavated in Antwerp, Belgium." *Mikrochimica Acta*, **15 (Suppl.)**, 253–267.

Jolliffe IT, Trendafilov NT, Uddin M (2003). "A Modified Principal Component Technique based on the LASSO." *J. Comput. Graph. Statist.*, **12**(3), 531–547.

Liebmann B, Filzmoser P, Varmuza K (2010). "Robust and Classical PLS Regression Compared." *Journal of Chemometrics*, **24**, 111–120.

Sernels S, Croux C, Filzmoser P, van Espen P (2005). "Partial Robust M-reression." *Chemometrics and Intellegent Laboratory Systems*, **79**, 55–64.

Todorov V, Filzmoser P (2009). "An Object Oriented Framework for Robust Multivariate Analysis." *Journal of Statistical Software*, **32**(3), 1–47. URL http://www.jstatsoft.org/v32/i03/.

Vanden Branden K, Hubert M (2005). "Robust Classification in High Dimensions Based on the SIMCA Method." *Chemometrics and Intellegent Laboratory Systems*, **79**, 10–21.

Varmuza K, Filzmoser P (2009). *Introduction to Multivariate Statistical Analysis in Chemometrics*. Taylor and Francis - CRC Press, Boca Raton, FL.

Wold H (1975). "Soft Modeling by Latent Variables: the Non-Linear Iterative Partial Least Squares Approach." In J Giani (ed.), *Perspectives in probability and statistics, papers in honor of M.S. Bartlett*, pp. 117–142. Academic Press, London.

Wold S (1976). "Pattern Recognition by means of Disjoint Principal Component Models." *Pattern Recognition*, **8**, 127–139.

Zou H, Hastie T, Tibshirani R (2006). "Sparse Principal Component Analysis." *Journal of Computational and Graphical Statistics*, **15**(2), 265–286.

## Affiliation:

Valentin Todorov
United Nations Industrial Organization (UNIDO), Austria
E-mail: valentin.todorov@unido.org
URL: http://www.unido.org

Peter Filzmoser
Department of Statistics and Probability Theory
Vienna University of Technology, Austria
E-mail: p.filzmoser@tuwien.ac.at
URL: http://www.statistik.tuwien.ac.at/public/filz

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 43/3-4
June 2014

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

*Submitted:* 2013-11-25
*Accepted:* 2014-03-10

Austrian Journal of Statistics
*June 2014, Volume 43/3-4, 267–277.*
*http://www.ajs.or.at/*

# Estimation of Time Series Models via Robust Wavelet Variance

**Stéphane Guerrier**
University of California, Santa Barbara

**Roberto Molinari**
Université de Genève

**Maria-Pia Victoria-Feser**
Université de Genève

**Abstract**

A robust approach to the estimation of time series models is proposed. Taking from a new estimation method called the Generalized Method of Wavelet Moments (GMWM) which is an indirect method based on the Wavelet Variance (WV), we replace the classical estimator of the WV with a recently proposed robust M-estimator to obtain a robust version of the GMWM. The simulation results show that the proposed approach can be considered as a valid robust approach to the estimation of time series and state-space models.

*Keywords*: maximum overlap discrete transform, M-estimator, generalized method of wavelet moments, composite stochastic processes, autoregressive processes.

## 1. Introduction

The robust estimation of time series parameters is still a widely open topic in statistics for various reasons. First of all, the robustness theory for dependent data is still not fully developed given that the classical robustness measures are not directly applicable in the time series context. In fact, for example, there is no unique definition of an influence function (Hampel 1974) for time series since there is no unique definition of outliers or, more specifically, there are different types of outliers which require to adapt such a measure (see Maronna, Martin, and Yohai 2006, for a detailed overview). Secondly, many of the existing methods for robust estimation of time series' parameters are limited in terms of the range of models that can be estimated and, above all, in terms of computation complexity as the models get larger or more complicated. Moreover, robust estimation of latent time series models (models made of the sum of several unobserved processes) has been largely ignored.

For robust estimation and inference for time series, a detailed list of references can be found in Maronna *et al.* (2006), Chapter 8. Most of the literature in this domain has dealt with standard time series models such as autoregressive and/or moving average processes, starting with the seminal work of Masreliez and Martin (1977), Denby and Martin (1979), Bustos and Yohai (1986) and Künsch (1984). Estimating robustly the parameters of latent models has not gone beyond the AR(1) plus white noise (Masreliez and Martin 1977), probably because of the difficulty in implementation of the different algorithms.

This paper intends to explore the possibilities opened up by combining two recently proposed approaches: the first concerning the robust estimation of the Wavelet Variance (WV) proposed by Mondal and Percival (2012) and the second proposed by Guerrier, Stebler, Skaloud, and Victoria-Feser (2013b) presenting a new method for the estimation of complex time series parameters based on the WV, called the generalized method of wavelet moments (GMWM). Since GMWM estimators are based on the matching between empirical and model based WV estimators, the use of a robust estimator for the WV will in principle ensure robustness of the model's parameters estimator, as done, for example, with the robust generalized method of moments (see Hansen 1982; Ronchetti and Trojani 2001).

The paper is organized as follows. In Section 2 we present a robust WV estimator proposed by Mondal and Percival (2012) that we modify to improve its robustness properties. In Section 3 we briefly present the GMWM and propose a robust version of this method, and in Section 4 we present a simulation study that involves several models, including latent time series models.

## 2. Robust Estimation of the Wavelet Variance

The WV is a quantity which is widely used throughout many scientific and engineering disciplines as a means to decompose, describe and summarize time series. For example, it has been used for over 30 years as a standard routine measure of frequency stability in lasers (see Fukuda, Tachikawa, and Kinoshita (2003)) or atomic clocks (see Allan (1987)). More recently, the WV has also been used with optical sensors (see Kebabian, Herndon, and Freedman (2005)), various types of gas monitoring spectrometers (see Bowling, Sargent, Tanner, and Ehleringer (2003); Werle, Mücke, and Slemr (1993)), sonic anemometer-thermometers (see Loescher, Ocheltree, Tanner, Swiatek, Dano, Wong, Zimmerman, Campbell, Stock, Jacobsen *et al.* (2005)), inertial sensors (see Guerrier (2009); El-Sheimy, Hou, and Niu (2008)), radio-astronomical instrumentation (see Schieder and Kramer (2001)). The WV was also used for example in Percival and Guttorp (1994) to analyse geophysics time series. This approach was also used for physiological signal analysis for example in Fadel, Orer, Barman, Vongpatanasin, Victor, and Gebber (2004) or in Gebber, Orer, and Barman (2006). In Whitcher (2004), discrete wavelet packet transforms are used to estimate one of the parameters of a seasonal long memory process for the analysis of atmospheric and economic time series.

The WV can be interpreted as the variance of a process after it has been subject to an approximate bandpass filter (Percival and Guttorp 1994). Let $\{X_t\}, t \in \mathbb{Z}$, be a stationary process, or a non-stationary process with stationary backward differences of order $d$. By applying a specific wavelet filter $\{\tilde{h}_{j,l}\}, j = 1, \ldots, J$ to this process we obtain the Maximum Overlap Discrete Wavelet Transform (MODWT) coefficients $\{W_{j,t}\}$ (see e.g. Percival and Walden 2000) as follows

$$W_{j,t} = \sum_{l=0}^{L_j - 1} \tilde{h}_{j,l} X_{t-l}, \; t \in \mathbb{Z} \tag{1}$$

where $j$ is the scale at which the filter is applied and $L_j = (2^j - 1)(L_1 - 1) + 1$ is the length of that filter with $L_1$ being the length of $\{\tilde{h}_{1,l}\}$. Given the wavelet coefficients, the WV at scale $j$ is defined as the variance of the wavelet coefficients at this scale

$$\nu_j = \text{var}\,(W_{j,t}) \tag{2}$$

Under the stationarity conditions defined above, it can be observed that the WV $\nu_j$ is not a function of $t$ (i.e. is time-invariant). This entails a series of properties, among which the following

$$\sum_{j=1}^{\infty} \nu_j = \sigma_X^2 \tag{3}$$

where $\sigma_X^2$ is the variance of $\{X_t\}$. Hence, the WV is a decomposition of the process variance and, as highlighted earlier, is consequently useful under many aspects if one is concerned by how the variance of a process is distributed across the different scales.

The MODWT estimator of the WV was defined in Percival (1995) and is given by

$$\hat{\nu}_j = \frac{1}{M(T_j)} \sum_{t \in T_j} W_{j,t}^2 \tag{4}$$

with $T_j$ being the set of time indices for which the wavelet coefficients are free of end effects, and $M(T_j) = T - L_j + 1$ being their number. This estimator of the WV is the most efficient asymptotically and it's properties were studied and proved in Serroukh, Walden, and Percival (2000).

An alternative estimator for the WV is based on the Discrete Wavelet Transform (DWT) coefficients (see Greenhall 1991; Percival and Guttorp 1994), for which the wavelet filter is applied to the process in (1) in a different manner. More specifically, the DWT filters a sequence $\{X_t\}$ on non-overlapping windows yielding the DWT wavelet coefficients

$$\overline{W}_{j,t} = 2^{-j/2} \sum_{l=0}^{L_j - 1} h_{j,l} X_{t-l} \tag{5}$$

where $t$ is taken at intervals of lag $L_j$.

However, in a recent article Mondal and Percival (2012) underline how even "a moderate amount of contamination often has a very adverse effect on conventional estimates of the wavelet variance". For this purpose they propose an M-estimator for the WV based on the transformation of the WV (a scale parameter) to a location parameter as follows

$$Q_{j,t} = \log\left(W_{j,t}^2\right) \tag{6}$$

They then propose to use the following M-estimator

$$\hat{\mu}_j = \underset{\mu_j \in \mathbb{R}}{\operatorname{argmin}} \left\{ \left| \sum_{t \in T_j} \psi(Q_{j,t} - \mu_j) \right| \right\} \tag{7}$$

which is then inversely transformed and corrected for bias in order to obtain a consistent estimator for $\nu_j$. Here $\psi(\cdot)$ is a function of bounded variation which guarantees the robustness of the estimator. Mondal and Percival (2012) suggest four types of $\psi$-functions and make use of the median-type function for their simulations, that is to say $\psi(z) = \text{sign}(z)$. This $\psi$-function is therefore the one which will be used in the Monte Carlo simulations presented further on in this paper.

Moreover, preliminary simulations have shown that in many cases the WV based on the DWT coefficients appear to be more appropriate for robustness purposes than the MODWT coefficients. Hence the Monte Carlo study will be done using the WV based on the DWT coefficients $\overline{W}_{j,t}$ by using the relationship between these and the MODWT coefficients as underlined in Percival (1995).

## 3. Robust Generalized Method of Wavelet Moments

Guerrier *et al.* (2013b) propose a method for the estimation of complex time series models, namely the GMWM. The method extends from the GMM setting and uses the implicit link which exists between the WV and the underlying assumed model $P_\theta$. The link is the following

$$\nu_j = \int_{-1/2}^{1/2} S_{W_j}(f) df = \int_{-1/2}^{1/2} |\overline{H}_j(f)|^2 S_{P_{\boldsymbol{\theta}}}(f) df \tag{8}$$

where $S_{W_j}(f)$ is the Power Spectral Density (PSD) function for the wavelet coefficients $W_j$ or $\overline{W}_{j,t}$, $\overline{H}_j(f) = \sum_{l=0}^{L_1-1} \tilde{h}_{j,l} e^{-i2\pi fl}$ denotes the transfer function of the wavelet filters $\tilde{h}_{j,l}$ (or $h_{j,l}$), with $|\cdot|$ being the modulus, and $S_{P_{\boldsymbol{\theta}}}$ is the PSD implied by $P_{\boldsymbol{\theta}}$. Hence there is a link between the WV and $P_{\boldsymbol{\theta}}$.

Let us define $\boldsymbol{\nu} = [\nu_1, \ldots, \nu_J]$ as the vector of WV and $\boldsymbol{\nu}(\boldsymbol{\theta})$ as the WV vector implied by the process $P_{\boldsymbol{\theta}}$. Taking advantage of the above link, Guerrier *et al.* (2013b) propose the following estimator

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmin}} \; (\hat{\boldsymbol{\nu}} - \boldsymbol{\nu}(\boldsymbol{\theta}))^T \, \boldsymbol{\Omega} \, (\hat{\boldsymbol{\nu}} - \boldsymbol{\nu}(\boldsymbol{\theta})) \tag{9}$$

where $\boldsymbol{\Omega}$ is an appropriate positive definite weighting matrix. The authors provide the proofs of consistency of the estimator for a number of time series models as well as of its asymptotic normality.

The idea behind this paper is to combine the estimation method presented in Section 2 with the GMWM. Hence, instead of using the classical estimator of the WV defined in (4), we propose to use the transformed and corrected version of the estimator in (7) using the DWT. We then use this estimator for $\hat{\boldsymbol{\nu}}$ in (9) to obtain a robust estimation method.

This proposed approach has its theoretical bases in the papers by Ronchetti and Trojani (2001) and Genton and Ronchetti (2003). Using a robust estimator of $\boldsymbol{\nu}$ implies a robust estimator for $\boldsymbol{\theta}$ with a bounded influence function since a bounded estimator for $\boldsymbol{\nu}$ bounds the function $(\hat{\boldsymbol{\nu}} - \boldsymbol{\nu}(\boldsymbol{\theta}))$.

The next section presents a Monte Carlo study to investigate the performance of this new approach on different stochastic processes.

## 4. Monte Carlo Study

In this section we present a Monte Carlo study of the estimator proposed in Section 3. We will investigate the performance of the estimator on three processes, namely a white noise process (WN), a first-order autoregressive process (AR1) and a composite stochastic process like the simulation presented in Guerrier *et al.* (2013b).

In addition to the wavelet moments used in the latter article, Guerrier, Stebler, Skaloud, and Victoria-Feser (2013a) suggest using additional moments of the processes to improve the performance of the GMWM estimator. Hence, the simulations will use the second moment in the case of the WN and AR1 processes and the first and second moments of the first-order difference of the composite process since the latter is stationary and has a non-zero expectation.

We will compare three estimators: the Maximum Likelihood Estimator (MLE), the classical GMWM estimator (GMWM) and the robust estimator proposed in the present paper (RGMWM). For the classic GMWM, the first and second moments will be estimated respectively via the classical estimators of mean and variance whereas the third estimator will use respectively the median and the M-estimator proposed in (7). For the classical and robust WV estimator, the DWT wavelet transform is used. In all studies, processes of length $L = 1000$ were simulated and the contaminations were additive (i.e. Gaussian noise with a specific variance $\sigma_\epsilon^2$ was added to an $\epsilon$-percentage of the observations of the underlying process).

### 4.1. White Noise

A white noise process can be written as $X_t \overset{iid}{\sim} N(0, \sigma^2)$.

Hence, the only parameter to be estimated is $\sigma^2$. The performance of the proposed estimators when there is no contamination is illustrated in Figure 1 where all estimators appear to be unbiased. The RGMWM however has a larger variance which is to be expected since efficiency is the price to pay for robustness.

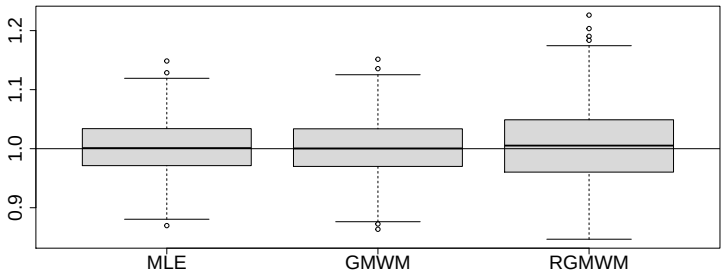


Figure 1: Finite sample performance of the MLE, GMWM and RGMWM estimators on an uncontaminated white noise process of length $L = 1,000$, with $\sigma = 1$. MLE represents the maximum likelihood estimator, GMWM represents the classic GMWM estimator with additional second moment of the process, RGMWM represents the robust GMWM based on the M-estimator proposed by Mondal and Percival (2012) with DWT wavelet transforms.

Table 1: Finite sample bias, variance and MSE of the MLE, GMWM and RGMWM estimators on an uncontaminated white noise process of length $L = 1,000$, with $\sigma = 1$. MLE represents the maximum likelihood estimator, GMWM represents the classic GMWM estimator with additional second moment of the process, RGMWM represents the robust GMWM based on the M-estimator proposed by Mondal and Percival (2012) with DWT wavelet transforms.

| | MLE | GMWM | RGMWM |
|---|---|---|---|
| Bias | $2.033 \cdot 10^{-3}$ | $1.768 \cdot 10^{-3}$ | $6.990 \cdot 10^{-3}$ |
| Variance | $2.021 \cdot 10^{-3}$ | $2.141 \cdot 10^{-3}$ | $4.520 \cdot 10^{-3}$ |
| MSE | $2.025 \cdot 10^{-3}$ | $2.144 \cdot 10^{-3}$ | $4.568 \cdot 10^{-3}$ |

Table 1 confirms these interpretations showing that in an uncontaminated setting, the best choice would be the MLE. However, by contaminating 5% of the sample with additive noise with $\sigma_\epsilon^2 = 100$ we can see how the MLE and the classical GMWM become highly biased and variable. Looking at Figure 2 and at the Mean Squared Errors (MSE) in Table 2, the advantage of using the RGMWM is evident.

### 4.2. First-Order Autoregressive

A first-order autoregressive process can be represented as follows

$$X_t = \phi X_{t-1} + \epsilon_t$$

where $\phi$ is the autoregressive parameter and $\epsilon_t \overset{iid}{\sim} N(0, \sigma^2)$.

Figure 3 shows how the proposed RGMWM estimator appears to confirm its robustness properties under a 1%-contaminated process with additive noise with $\sigma_\epsilon^2 = 9$. Its improved performance compared to the classical estimators is highlighted by the results in Table 3. The latter table appears to indicate that this robust approach is particularly convenient for estimating the $\sigma^2$ of the innovation process compared to the autoregressive parameter $\phi$.

### 4.3. Latent Time Series Model

The GMWM methodology was mainly developed to estimate models made up by latent processes. An example of such a process was given in Guerrier *et al.* (2013b) as a sum

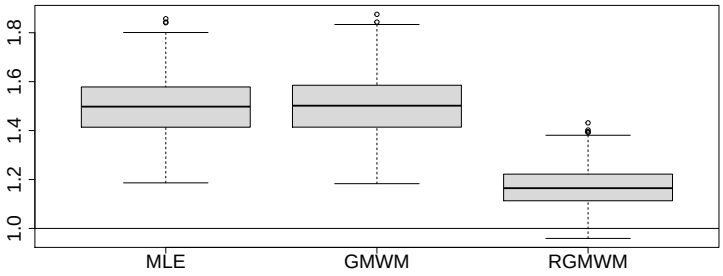


Figure 2: Finite sample performance of the MLE, GMWM and RGMWM estimators on a 5%-contaminated white noise process of length $L = 1,000$, with $\sigma = 1$ and contamination generated by adding Gaussian noise with $\sigma_\epsilon^2 = 100$. MLE represents the maximum likelihood estimator, GMWM represents the classic GMWM estimator with additional second moment of the process, RGMWM represents the robust GMWM based on the M-estimator proposed by Mondal and Percival (2012) with DWT wavelet transforms.

Table 2: Finite sample bias, variance and MSE of the MLE, GMWM and RGMWM estimators on a 5%-contaminated white noise process of length $L = 1,000$, with $\sigma = 1$ and contamination generated by adding Gaussian noise with $\sigma_\epsilon^2 = 100$. MLE represents the maximum likelihood estimator, GMWM represents the classic GMWM estimator with additional second moment of the process, RGMWM represents the robust GMWM based on the M-estimator proposed by Mondal and Percival (2012) with DWT wavelet transforms.

| | MLE | GMWM | RGMWM |
|---|---|---|---|
| Bias | $4.988 \cdot 10^{-1}$ | $5.018 \cdot 10^{-1}$ | $1.690 \cdot 10^{-1}$ |
| Variance | $1.450 \cdot 10^{-2}$ | $1.478 \cdot 10^{-2}$ | $6.448 \cdot 10^{-3}$ |
| MSE | $2.633 \cdot 10^{-1}$ | $2.666 \cdot 10^{-1}$ | $3.500 \cdot 10^{-2}$ |

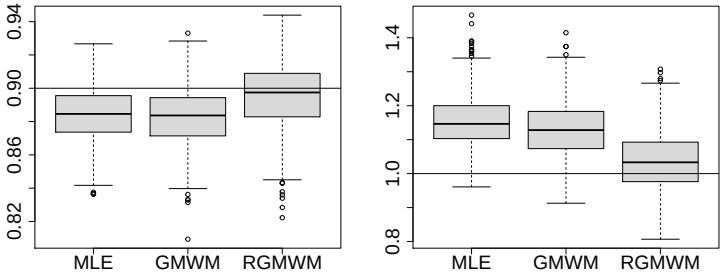


Figure 3: Finite sample performance of the MLE, GMWM and RGMWM estimators on a 1%-contaminated first-order autoregressive process of length $L = 1,000$ with $\sigma = 1$, $\phi = 0.9$ and contamination generated by adding Gaussian noise with $\sigma_\epsilon^2 = 9$. MLE represents the maximum likelihood estimator, GMWM represents the classic GMWM estimator with additional second moment of the process, RGMWM represents the robust GMWM based on the M-estimator proposed by Mondal and Percival (2012) with DWT wavelet transforms.

Table 3: Finite sample bias, variance and MSE of the MLE, GMWM and RGMWM estimators on a 1%-contaminated first-order autoregressive process of length $L = 1,000$ with $\sigma = 1$, $\phi = 0.9$ and contamination generated by adding Gaussian noise with $\sigma_\epsilon^2 = 9$. MLE represents the maximum likelihood estimator, GMWM represents the classic GMWM estimator with additional second moment of the process, RGMWM represents the robust GMWM based on the M-estimator proposed by Mondal and Percival (2012) with DWT wavelet transforms.

| | | MLE | GMWM | RGMWM |
|---|---|---|---|---|
| $\phi$ | Bias | $-1.629 \cdot 10^{-2}$ | $-1.759 \cdot 10^{-2}$ | $-4.946 \cdot 10^{-3}$ |
| | Variance | $2.764 \cdot 10^{-4}$ | $3.067 \cdot 10^{-4}$ | $3.667 \cdot 10^{-4}$ |
| | MSE | $5.419 \cdot 10^{-4}$ | $6.163 \cdot 10^{-4}$ | $3.912 \cdot 10^{-4}$ |
| $\sigma^2$ | Bias | $1.563 \cdot 10^{-1}$ | $1.313 \cdot 10^{-1}$ | $3.846 \cdot 10^{-2}$ |
| | Variance | $6.661 \cdot 10^{-3}$ | $6.904 \cdot 10^{-3}$ | $8.082 \cdot 10^{-3}$ |
| | MSE | $3.108 \cdot 10^{-2}$ | $2.414 \cdot 10^{-2}$ | $9.561 \cdot 10^{-3}$ |

of an autoregressive process, a drift process $\{\omega\}$ and a white noise process as follows

$$Y_t = \phi Y_{t-1} + \omega + u_t,\ u_t \overset{iid}{\sim} N(0, \sigma_{AR}^2)$$
$$X_t = Y_t + \epsilon_t,\ \epsilon_t \overset{iid}{\sim} N(0, \sigma_{WN}^2)$$

For these kind of processes, the GMWM (along with the robust version presented in this paper) presents important advantages over alternative approaches (see Guerrier *et al.* (2013b)). When contaminating this process (with $\phi = 0.95$, $\omega = 0.04$, $\sigma_{AR}^2 = 16$, $\sigma_{WN}^2 = 4$) with 5%-additive outliers with $\sigma_\epsilon^2 = 9$, the results seem to indicate that the classic GMWM does not appear to be greatly affected by this contamination for the first three parameters. However, it shows all the impact of the outliers for the estimation of the white noise innovation parameter $\sigma_{WN}^2$ where the RGMWM shows only a very slight bias.

The results presented in Table 4 show how the GMWM seems to perform better than the proposed RGMWM except for the parameter $\sigma_{WN}^2$ where the GMWM is completely biased. Therefore, although slightly biased for most of the parameters, the RGMWM limits this bias for all parameters whereas the classical GMWM loses this property for one parameter. The improved performance of the RGMWM on the innovation parameters could be explained by the fact that the latter process is especially identifiable at the first scales of the WV which are also the most informative (i.e. they have a larger number of wavelet coefficients).

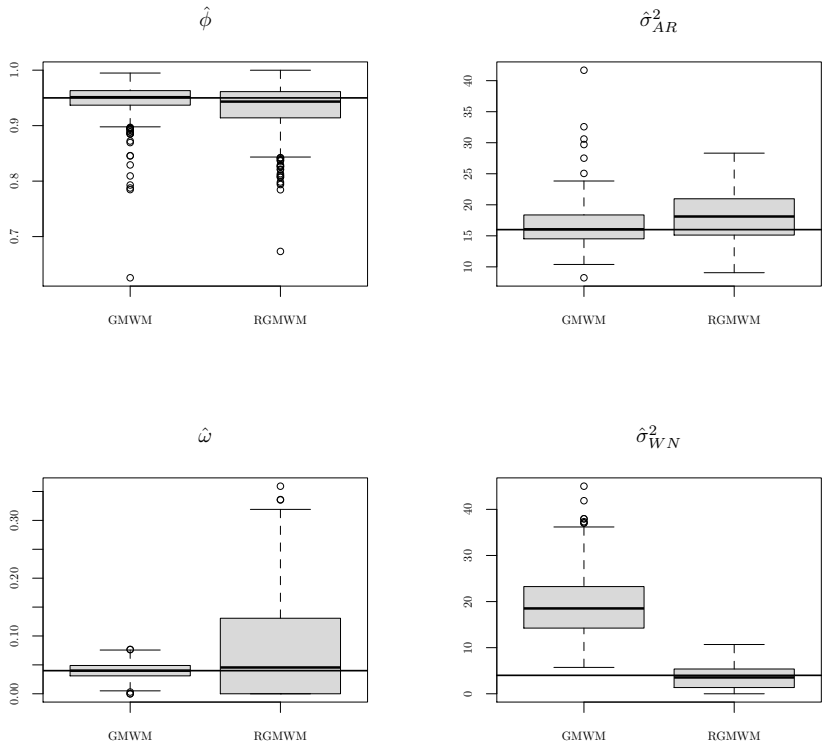


Figure 4: Finite sample performance of the MLE, GMWM and RGMWM estimators on a 5%-contaminated composite process (10) of length $L = 1,000$, with $\phi = 0.95$, $\omega = 0.04$, $\sigma^2_{AR} = 16$, $\sigma^2_{WN} = 4$ and contamination generated by adding Gaussian noise with $\sigma^2 = 9$. GMWM represents the classic GMWM estimator with additional first and second moment of the first-differenced process, RGMWM represents the robust GMWM based on the M-estimator proposed by Mondal and Percival (2012) with DWT wavelet transforms.

## 5. Conclusions and Perspectives

Given the theoretical bases and the results of the Monte Carlo studies, the proposed estimator appears to be an extremely valid candidate for the robust estimation of time series models. Knowing the theoretical WV $\boldsymbol{\nu}(\boldsymbol{\theta})$ of a process, it is possible to estimate the parameters $\boldsymbol{\theta}$ of this process in a robust manner.

The theoretical WV of many processes can be derived from the results in Zhang (2008) or, as an alternative, Guerrier *et al.* (2013b) suggest to use indirect inference to overcome the complexity of calculations for certain models. Hence, the proposed estimator is easily implemented and computationally inexpensive while at the same time providing a robust estimation method for many processes for whom robust estimation methods are scarce.

There are many possible developments for this method, including the study of its asymptotic properties. Given the variety of wavelet decompositions, different wavelets and filtering methods could be explored to understand if some of them could contribute more effectively to the robust estimation approach presented in this paper. Moreover, as highlighted earlier, Guerrier *et al.* (2013a) suggested additional adjustments to the GMWM methodology to improve its performance and its robust equivalents could be considered to improve the performance also of the approach proposed in this paper.

Table 4: Finite sample bias, variance and MSE of the GMWM and RGMWM estimators on a 5%-contaminated composite process (10) of length $L = 1,000$, with $\phi = 0.95$, $\omega = 0.04$, $\sigma^2_{AR} = 16$, $\sigma^2_{WN} = 4$ and contamination generated by adding Gaussian noise with $\sigma^2 = 9$. GMWM represents the classic GMWM estimator with additional first and second moment of the first-order difference of the process, RGMWM represents the robust GMWM based on the M-estimator proposed by Mondal and Percival (2012) with DWT wavelet transforms.

| | | GMWM | RGMWM |
|---|---|---|---|
| $\phi$ | Bias | $-3.709 \cdot 10^{-3}$ | $-1.670 \cdot 10^{-2}$ |
| | Variance | $9.269 \cdot 10^{-4}$ | $1.744 \cdot 10^{-3}$ |
| | MSE | $9.407 \cdot 10^{-4}$ | $2.023 \cdot 10^{-3}$ |
| $\sigma^2_{AR}$ | Bias | $5.035 \cdot 10^{-1}$ | 2.088 |
| | Variance | $1.030 \cdot 10^{1}$ | $1.436 \cdot 10^{1}$ |
| | MSE | $1.055 \cdot 10^{1}$ | $1.872 \cdot 10^{1}$ |
| $\omega$ | Bias | $4.896 \cdot 10^{-4}$ | $3.344 \cdot 10^{-2}$ |
| | Variance | $1.801 \cdot 10^{-4}$ | $6.897 \cdot 10^{-3}$ |
| | MSE | $1.803 \cdot 10^{-4}$ | $8.015 \cdot 10^{-3}$ |
| $\sigma^2_{WN}$ | Bias | $1.535 \cdot 10^{1}$ | $-4.975 \cdot 10^{-1}$ |
| | Variance | $4.332 \cdot 10^{1}$ | 6.333 |
| | MSE | $2.788 \cdot 10^{2}$ | 6.580 |

## References


Allan D (1987). "Time and Frequency (Time-Domain) Characterization, Estimation, and Prediction of Precision Clocks and Oscillators." *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, **UFFC-34**(6), 647–654.

Bowling D, Sargent S, Tanner B, Ehleringer J (2003). "Tunable Diode Laser Absorption Spectroscopy for Stable Isotope Studies of Ecosystem-Atmosphere CO2 Exchange." *Agricultural and Forest Meteorology*, **118**(1-2), 1–19. ISSN 0168-1923.

Bustos OH, Yohai VJ (1986). "Robust estimates for ARMA models." *Journal of the American Statistical Association*, **81**, 155–168.

Denby L, Martin RD (1979). "Robust estimation of the first-order autoregressive parameter." *Journal of the American Statistical Association*, **74**(365), 140–146.

El-Sheimy N, Hou H, Niu X (2008). "Analysis and Modeling of Inertial Sensors using Allan Variance." *IEEE Transactions on Instrumentation and Measurement*, **57**(1), 140–149. ISSN 0018-9456.

Fadel P, Orer H, Barman S, Vongpatanasin W, Victor R, Gebber G (2004). "Fractal Properties of Human Muscle Sympathetic Nerve Activity." *American Journal of Physiology- Heart and Circulatory Physiology*, **286**(3), H1076.

Fukuda K, Tachikawa M, Kinoshita M (2003). "Allan Variance Measurements of Diode Laser Frequency-Stabilized with a Thin Vapor Cell." *Applied Physics B : Lasers and Optics*, **77**, 823–827.

Gebber G, Orer H, Barman S (2006). "Fractal Noises and Motions in Time Series of Presympathetic and Sympathetic Neural Activities." *Journal of neurophysiology*, **95**(2), 1176–1184. ISSN 0022-3077.

Genton M, Ronchetti E (2003). "Robust Indirect Inference." *Journal of the American Statistical Association*, **98**(461), 67–76. ISSN 0162-1459.

Greenhall C (1991). "Recipes for Degrees of Freedom of Frequency Stability Estimators." In *IEEE Transactions on Instrumentation and Measurements*, volume 40, pp. 994–999. IEEE. ISSN 0018-9456.

Guerrier S (2009). "Improving Accuracy with Multiple Sensors: Study of Redundant MEMS-IMU/GPS Configurations." In *Proceedings of the 22nd International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS 2009)*, pp. 3114–3121. Savannnah, GA, USA.

Guerrier S, Stebler Y, Skaloud J, Victoria-Feser MP (2013a). "Limits of the Allan Variance and Optimal Tuning of Wavelet Variance based Estimators." *Submitted working paper*. URL http://www.hec.unige.ch/guerrier/allan.

Guerrier S, Stebler Y, Skaloud J, Victoria-Feser MP (2013b). "Wavelet variance based estimation for composite stochastic processes." *Journal of the American Statistical Association*. To appear.

Hampel FR (1974). "The Influence Curve and its Role in Robust Estimation." *Journal of the American Statistical Association*, **69**, 383–393.

Hansen L (1982). "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica: Journal of the Econometric Society*, **50**(4), 1029–1054. ISSN 0012-9682.

Kebabian P, Herndon S, Freedman A (2005). "Detection of Nitrogen Dioxide by Cavity Attenuated Phase Shift Spectroscopy." *Analytical chemistry*, **77**(2), 724–728. ISSN 0003-2700.

Künsch H (1984). "Infinitesimal Robustness for Autoregressive Processes." *The Annals of Statistics*, **12**, 843–863.

Loescher H, Ocheltree T, Tanner B, Swiatek E, Dano B, Wong J, Zimmerman G, Campbell J, Stock C, Jacobsen L, *et al.* (2005). "Comparison of Temperature and Wind Statistics in Contrasting Environments Among Different Sonic Anemometer-Thermometers." *Agricultural and forest meteorology*, **133**(1-4), 119–139. ISSN 0168-1923.

Maronna RA, Martin RD, Yohai VJ (2006). *Robust Statistics: Theory and Methods*. Wiley, Chichester, West Sussex, UK.

Masreliez C, Martin R (1977). "Robust Bayesian estimation for the linear model and robustifying the Kalman filter." *IEEE Transactions on Automatic Control*, **22**, 361–371.

Mondal D, Percival D (2012). "M-estimation of wavelet variance." *Annals of The Institute of Statistical Mathematics*, **64**(1), 27–53.

Percival D (1995). "On Estimation of the Wavelet Variance." *Biometrika*, **82**, 619–631.

Percival D, Guttorp P (1994). "Long-Memory Processes, the Allan Variance and Wavelets." *Wavelets in Geophysics*, **4**, 325–344.

Percival D, Walden A (2000). *Wavelet Methods for Time Series Analysis*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 32 Avenue of the Americas, New York, NY 10013-2473, USA.

Ronchetti E, Trojani F (2001). "Robust inference with GMM estimators." *Journal of Econometrics*, **101**, 37–69.

Schieder R, Kramer C (2001). "Optimization of Heterodyne Observations using Allan Variance Measurements." *Astronomy and Astrophysics*, **373**(2), 746–756. doi:10.1051/0004-6361:20010611.

Serroukh A, Walden AT, Percival D (2000). "Statistical Properties and Uses of the Wavelet Variance Estimator for the Scale Analysis of Time Series." *Journal of the American Statistical Association*, **95**, 184–196.

Werle P, Mücke R, Slemr F (1993). "The limits of signal averaging in atmospheric trace-gas monitoring by tunable diode-laser absorption spectroscopy (TDLAS)." *Applied Physics B: Lasers and Optics*, **57**(2), 131–139. ISSN 0946-2171.

Whitcher B (2004). "Wavelet-based Estimation for Seasonal Long-memory Processes." *Technometrics*, **46**(2), 225–238. ISSN 0040-1706. doi:10.1198/004017004000000275.

Zhang NF (2008). "Allan variance of time series models for measurement data." *Metrologia*, **45**, 549–561.

## Affiliation:

Stéphane Guerrier
Department of Statistics & Applied Probability
University of California, Santa Barbara, USA
E-mail: guerrier@pstat.ucsb.edu

Roberto Molinari and Maria-Pia Victoria-Feser
Geneva School of Economics and Management
Université de Genève, Switzerland
E-mail: Roberto.Molinari@unige.ch, Maria-Pia.VictoriaFeser@unige.ch

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 43/3-4
June 2014

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

*Submitted:* 2013-11-25
*Accepted:* 2014-03-10

Austrian Journal of Statistics
*June 2014, Volume 43/3-4, 279–292.*
*http://www.ajs.or.at/*

# On the Exact Two-Sided Tolerance Intervals for Univariate Normal Distribution and Linear Regression

**Viktor Witkovský**
Slovak Academy of Sciences

**Abstract**

Statistical tolerance intervals are another tool for making statistical inference on an unknown population. The tolerance interval is an interval estimator based on the results of a calibration experiment, which can be asserted with stated confidence level $1 - \alpha$, for example 0.95, to contain at least a specified proportion $1 - \gamma$, for example 0.99, of the items in the population under consideration. Typically, the limits of the tolerance intervals functionally depend on the tolerance factors. In contrast to other statistical intervals commonly used for statistical inference, the tolerance intervals are used relatively rarely. One reason is that the theoretical concept and computational complexity of the tolerance intervals is significantly more difficult than that of the standard confidence and prediction intervals.

In this paper we present a brief overview of the theoretical background and approaches for computing the tolerance factors based on samples from one or several univariate normal (Gaussian) populations, as well as the tolerance factors for the non-simultaneous and simultaneous two-sided tolerance intervals for univariate linear regression. Such tolerance intervals are well motivated by their applicability in the multiple-use calibration problem and in construction of the calibration confidence intervals. For illustration, we present examples of computing selected tolerance factors by the implemented algorithm in MATLAB.

*Keywords*: normal population, linear regression, tolerance factor, simultaneous tolerance intervals, multiple-use calibration, MATLAB algorithm.

## 1. Introduction

Statistical tolerance intervals are interval estimators used for making statistical inference on population(s), which can be fully described by a probability distribution from a given family of distributions (as e.g., the family of normal distributions). For more details on different types of statistical intervals consult, e.g., the following books: Hahn and Meeker (1991), Krishnamoorthy and Mathew (2009), and Liu (2011).

Although the concept of statistical tolerance intervals has been well recognized for a long time, surprisingly, it seems that their applications remain still limited. The reliable algorithms for

computing the exact tolerance factors are missing in the commonly used statistical packages (even for inferences on normal populations), however, more or less accurate approximations are available. Implementations of such algorithms (mainly based on approximate and/or Monte Carlo methods) are currently under fast development, as, e.g., in the package `tolerance` for `R`, see Young (2010).

Thus, possible applications should rely either on implemented approximate methods or on published collections of tables for tolerance factors (if available), see e.g. the book Odeh and Owen (1980), which gives many of the most required factors in the context of the normal distribution, however, with limited precision. Due to the recognized importance of statistical tolerance intervals in technical applications, ISO (the International Organization for Standardization) has currently prepared a revised version of the ISO standard 16269-6 (Statistical interpretation of data — Part 6: Determination of statistical tolerance intervals), which also provides detailed tables of tolerance factors for selected tolerance intervals.

The theory of statistical tolerance intervals, as well as the computational methods and algorithms, have been developed significantly during the last three decades. This, together with the fast growing computational power of the personal computers, allows development of fast, efficient and reliable implementations of the algorithms for highly precise computing of the exact tolerance factors and limits required for the statistical tolerance intervals. For a comprehensive overview of the recent advances and developments in this area see Krishnamoorthy and Mathew (2009).

In this paper we shall briefly overview the theoretical background and describe some computational approaches for computing the exact tolerance factors for two-sided statistical tolerance intervals based on sample(s) from normal (Gaussian) population(s). Moreover, we shall also present a method for computing the exact simultaneous two-sided tolerance intervals for normal linear regression by using the method for computing the simultaneous tolerance factors for several independent univariate normal populations.

Based on that, we have developed a MATLAB algorithm for efficient and highly precise computation of the exact tolerance factors for the non-simultaneous as well as simultaneous two-sided tolerance intervals for several independent univariate normal populations. This can be used also for computing the exact tolerance factors for the non-simultaneous two-sided tolerance intervals, and also (in combination with other optimization procedures, based on Monte Carlo simulations) for computing the exact simultaneous two-sided tolerance intervals for univariate normal linear regression.

The methods and algorithms can be further used in the multiple-use calibration problem for constructing the appropriate simultaneous interval estimators (calibration confidence intervals) for values of the variable of primary interest, say $x$, based on possibly unlimited sequence of future observations of the response variable, say $y$, and on the results of the given calibration experiment, which was modeled/fitted by a linear regression model. Such calibration intervals can be obtained by inverting the simultaneous tolerance intervals constructed for the regression (calibration) function. For more details see, e.g., Scheffé (1973), Mee, Eberhardt, and Reeve (1991), Mee and Eberhardt (1996), Mathew and Zha (1997), and Chvosteková (2013b).

## 2. Two-sided tolerance intervals for univariate normal distribution

First, let us consider a simple calibration experiment, say $\mathcal{E}$, which is represented by a random sample of size $n$ from a population whose distribution is characterized by a univariate normal distribution $N(\mu, \sigma^2)$, i.e. $Y_1, \ldots, Y_n$, where $Y_i$ are independent random variables normally distributed, $Y_i \sim N(\mu, \sigma^2)$, where $\mu$ and $\sigma^2$ are unknown parameters (mean and variance) of the population distribution.

Notice that the available information on distribution of the unknown population, based on the result of experiment $\mathcal{E}$, is fully characterized by the random sample, or equivalently by

the sufficient statistics: the sample mean, $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, and the sample variance, $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$. Under given assumptions, it is well known that the sufficient statistics are independent random variables and their distribution is given by $\bar{Y} \sim N(\mu, \delta^2\sigma^2)$, where $\delta^2 = \frac{1}{n}$, and $S^2 \sim \sigma^2 \frac{1}{\nu}\chi^2_\nu$, where $\nu = n - 1$ denotes the degrees of freedom (DFs) and $\chi^2_\nu$ represents a chi-square distribution with $\nu$ DFs.

## 2.1. Two-sided tolerance intervals for one univariate normal distribution

Given the result of the calibration experiment $\mathcal{E}$, we wish to construct a two-sided tolerance interval (i.e., a random interval $(L_\mathcal{E}, U_\mathcal{E})$, with its limits depending on the result of the experiment $\mathcal{E}$), which can be asserted with confidence level $1 - \alpha$ (for example 0.95) to contain at least a specified proportion $1 - \gamma$ (for example 0.99) of the items in the population under consideration.

That is, we wish to construct the two-sided $(1-\gamma, 1-\alpha)$-tolerance interval which will cover a pre-specified proportion of possibly infinite sequence of independent future realizations of the response variable $Y = \mu + \sigma\epsilon$ (with $\epsilon \sim N(0,1)$ assumed to be independent of the calibration experiment $\mathcal{E}$) such that

$$\mathsf{P}_{\{\mathcal{E}\}}\Big(\mathsf{P}_{\{Y\}}\Big(L_\mathcal{E} \le Y \le U_\mathcal{E} \,|\, \mathcal{E}\Big) \ge 1 - \gamma\Big) = 1 - \alpha. \tag{1}$$

Notice that the confidence level $1 - \alpha$ is related to the random nature of the outcome (result) of the calibration experiment $\mathcal{E}$. That is, the required two-sided tolerance interval will cover more than $(1 - \gamma) \times 100\%$ proportion of the items of the unknown (normal) population, and this will be true in $(1 - \alpha) \times 100\%$ cases of the hypothetical calibration experiments.

In general, there are potentially many possible approaches to finding a solution to the problem as specified by (1). There is no unique solution until the form of the tolerance limits of the two-sided tolerance interval $(L_\mathcal{E}, U_\mathcal{E})$ is reasonably restricted. Commonly, the tolerance limits are considered in the form

$$L_\mathcal{E} = \bar{Y} - \kappa\sqrt{S^2}, \qquad U_\mathcal{E} = \bar{Y} + \kappa\sqrt{S^2}, \tag{2}$$

where $\kappa$ denotes the tolerance factor (a subject of the required solution) which depend on the stated coverage and confidence probabilities ($1 - \gamma$ and $1 - \alpha$, respectively), and further on the parameters characterizing the design of the experiment, $\delta^2$ and $\nu$. So, if necessary, we can emphasize the dependence of the tolerance factor $\kappa$ on other parameters by writing either $\kappa(1 - \gamma, 1 - \alpha, \delta^2, \nu)$, or $\kappa(\delta^2, \nu)$, etc.

Consequently, the following conditional probability statement (conditional for given result of $\mathcal{E}$) should be fulfilled for $(1 - \alpha) \times 100\%$ of the possible results of the calibration experiment (i.e., $\bar{Y}$ and $S^2$)

$$\begin{aligned}
1 - \gamma &\le \mathsf{P}_{\{Y\}}\Big(L_\mathcal{E} \le Y \le U_\mathcal{E} \,|\, \mathcal{E}\Big) \\
&= \mathsf{P}_{\{\epsilon\}}\Big(\bar{Y} - \kappa\sqrt{S^2} \le \mu + \sigma\varepsilon \le \bar{Y} + \kappa\sqrt{S^2} \,|\, \bar{Y}, S^2\Big) \\
&= \mathsf{P}_{\{\epsilon\}}\Big((\bar{Y} - \mu)/\sigma - \kappa\sqrt{S^2/\sigma^2} \le \epsilon \le (\bar{Y} - \mu)/\sigma + \kappa\sqrt{S^2/\sigma^2} \,|\, \bar{Y}, S^2\Big) \\
&= \Phi\left(\delta Z + \kappa\sqrt{Q/\nu}\right) - \Phi\left(\delta Z - \kappa\sqrt{Q/\nu}\right) \\
&= \Phi\left(\delta|Z| + \kappa\sqrt{Q/\nu}\right) - \Phi\left(\delta|Z| - \kappa\sqrt{Q/\nu}\right) \equiv C(\kappa \,|\, Z, Q),
\end{aligned} \tag{3}$$

where $Z = (\bar{Y} - \mu)/\delta\sigma$, $Q = \nu S^2/\sigma^2$, and $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution. So, $C(\kappa \,|\, Z, Q)$ represents the proportion of the population covered by the tolerance interval for the given tolerance factor $\kappa$ and for the given result of the calibration experiment $\mathcal{E}$. $C(\kappa \,|\, Z, Q)$ is commonly known as a content function.

The content function $C(\kappa \,|\, Z, Q)$ cannot be evaluated directly for given $\kappa$ and the observed result of the experiment $\mathcal{E}$ ($\bar{Y}$, and $S^2$), as it depends on the unknown parameters $\mu$ and $\sigma^2$. However, if we are interested in the stochastic properties of the tolerance intervals based on a large number of results of the hypothetical calibration experiments (i.e., the variability of the results of independent calibration experiments is to be considered), then $Z$ and $Q$ are independent pivotal random variables with known probability distributions independent of the unknown parameters $\mu$ and $\sigma^2$, i.e. $Z \sim N(0,1)$ and $Q \sim \chi^2_\nu$.

So, the content function $C(\kappa; Z, Q)$, now considered as a random variable (a function of random variables $Z$ and $Q$), can be used directly for checking the stochastic properties (the true confidence level) of the tolerance intervals, for any candidate value of the tolerance factor $\kappa$.

In particular, the tolerance factor $\kappa$ is exact for the $(1-\gamma, 1-\alpha)$-tolerance interval $(L_\mathcal{E}, U_\mathcal{E})$, defined by (2), if

$$E_{\{Z,Q\}}\Big(\mathbb{I}\left(C(\kappa; Z, Q) \geq 1-\gamma\right)\Big) = 1-\alpha, \tag{4}$$

where $\mathbb{I}(\cdot)$ is an indicator function, with $\mathbb{I}(\mathsf{true}) = 1$ and $\mathbb{I}(\mathsf{false}) = 0$, and $E_{\{Z,Q\}}(\cdot)$ denotes the expectation operator with respect to the distribution of the random variables $Z$ and $Q$.

Consequently, by applying a suitable iterative optimization procedure, $C(\kappa; Z, Q)$ can be used for computing the exact value of the tolerance factor $\kappa$, such that it fulfills the required property given by (1), or (4), respectively. This may be realized either by using (repeated) Monte Carlo simulations, or two-dimensional numerical integrations.

The below presented formula for computing the exact tolerance factor $\kappa$ of the two-sided $(1-\gamma, 1-\alpha)$-tolerance intervals for a univariate normal distribution requires (repeated) evaluation of one-dimensional integral, only. As we shall discuss in more details in the next Sections, the approach can be generalized also for computing the tolerance factor for other models based on normal distribution (as, e.g., the non-simultaneous, point-wise tolerance intervals, as well as the simultaneous tolerance intervals for normal linear regression models), however, with possibly needed evaluation of multivariate integrals.

Derivation is based on the results presented in Krishnamoorthy and Mathew (2009) (for more details see the equations (1.2.3), (1.2.4), also (2.5.7) and (2.5.8)).

Notice that for a fixed $\delta$ and $Z$ the function $\Phi(\delta|Z|+r) - \Phi(\delta|Z|-r)$ is an increasing function of $r$. Let us denote by $r_{1-\gamma}$ the solution to the equation

$$\Phi(\delta|Z| + r_{1-\gamma}) - \Phi(\delta|Z| - r_{1-\gamma}) = 1-\gamma. \tag{5}$$

Then, $C(\kappa \,|\, Z, Q) \geq 1-\gamma$ if and only if $\kappa\sqrt{Q/\nu} > r_{1-\gamma}$ (or equivalently $Q \geq \frac{\nu r^2_{1-\gamma}}{\kappa^2}$).

Based on (5), the problem can be rewritten equivalently as

$$\mathsf{P}_{\{\epsilon\}}\Big((\epsilon - \delta|Z|)^2 \leq r^2_{1-\gamma} \,|\, |Z|\Big) = 1-\gamma \tag{6}$$

where $\epsilon \sim N(0,1)$. For fixed $Z$, the random variable $(\epsilon - \delta|Z|)^2 \sim \chi^2_1(\delta^2 Z^2)$, i.e. it has a non-central chi-square distribution with one degree of freedom and the noncentrality parameter $\delta^2 Z^2$. Consequently, $r_{1-\gamma} = \sqrt{\chi^2_{1;1-\gamma}\left(\delta^2 Z^2\right)}$, where $\chi^2_{1;1-\gamma}\left(\delta^2 Z^2\right)$ denotes the $(1-\gamma)$-quantile of the distribution $\chi^2_1(\delta^2 Z^2)$.

Thus, the tolerance factor $\kappa$ defined by (1) and (2) is given implicitly as a solution to the equation

$$\begin{aligned}
1-\alpha &= E_{\{|Z|\}}\left(\mathsf{P}_{\{Q\}}\left(Q \geq \frac{\nu}{\kappa^2}\chi^2_{1;1-\gamma}\left(\delta^2 Z^2\right)\right)\right) \\
&= E_{\{|Z|\}}\left(1 - F_{\chi^2_\nu}\left(\frac{\nu}{\kappa^2}\chi^2_{1;1-\gamma}\left(\delta^2 Z^2\right)\right)\right) \\
&= 2\int_0^\infty \left(1 - F_{\chi^2_\nu}\left(\frac{\nu}{\kappa^2}\chi^2_{1;1-\gamma}\left(\delta^2 z^2\right)\right)\right)\phi(z)\,dz
\end{aligned}$$

$$= \quad 2\int_0^\infty \Gamma\left(\frac{\nu}{2}, \frac{\nu}{2\kappa^2}\chi^2_{1;1-\gamma}\left(\delta^2 z^2\right)\right)\phi(z)\,dz, \tag{7}$$

where $E_{\{|Z|\}}(f(|Z|))$ denotes the expectation of the function $f(|Z|)$, with respect to the distribution of $|Z|$, where $Z \sim N(0,1)$, $F_{\chi^2_\nu}(\cdot)$ denotes the CDF of a chi-square distribution with $\nu$ degrees of freedom, $\Gamma(\cdot,\cdot)$ is the incomplete regularized upper gamma function, and $\phi(z)$ denotes the PDF (probability density function) of a standard normal distribution. From computational point of view, the value $\chi^2_{1;1-\gamma}(\delta^2 z^2) = r^2_{1-\gamma}$ can be computed more efficiently by directly solving the equation (5), i.e. $\Phi(\delta z + r_{1-\gamma}) - \Phi(\delta z - r_{1-\gamma}) = 1-\gamma$, than by using a dedicated algorithm for computing quantiles of the non-central chi-square distribution.

## 2.2. Two-sided tolerance intervals for several independent univariate normal distributions with common variance

Here we consider a calibration experiment $\mathcal{E} = \{\mathcal{E}_1, \ldots, \mathcal{E}_m\}$ which is based on $m+1$ sufficient statistics, $\bar{Y}_1, \ldots, \bar{Y}_m$ and $S^2$, where $\bar{Y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$ (the sample means), $S^2 = \frac{1}{\nu}\sum_{i=1}^m (n_i - 1)S_i^2$ (the pooled sample variance) with $S_i^2 = \frac{1}{n_i-1}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2$, and $\nu = \sum_{i=1}^m (n_i - 1)$, where $n_i$ is the sample size of the $i$th population.

We wish to construct a set of simultaneous two-sided tolerance intervals $(L_{\mathcal{E},i}, U_{\mathcal{E},i})$, with limits $L_{\mathcal{E},i} = \bar{Y}_i - \kappa_i\sqrt{S^2}$ and $U_{\mathcal{E},i} = \bar{Y}_i + \kappa_i\sqrt{S^2}$, such that

$$\mathsf{P}_{\{\mathcal{E}\}}\Big(\bigcap_{i=1}^m \Big\{\mathsf{P}_{\{Y_i\}}\Big(L_{\mathcal{E},i} \le Y_i \le U_{\mathcal{E},i} \,|\, \mathcal{E}\Big) \ge 1-\gamma\Big\}\Big) = 1-\alpha, \tag{8}$$

where $Y_i \sim N(\mu_i, \sigma^2)$ are mutually independent random variables, independent from the calibration experiment $\mathcal{E} = \{\mathcal{E}_1, \ldots, \mathcal{E}_m\}$.

For a given (candidate) set of the tolerance factors, say $\kappa_1, \ldots, \kappa_m$, the content function for the simultaneous tolerance intervals $(L_{\mathcal{E},i}, U_{\mathcal{E},i})$ is given by

$$\begin{aligned} &C(\kappa_1, \ldots, \kappa_m \,|\, Z_1, \ldots, Z_m, Q) = \\ &\quad \min_i \left(\Phi\left(\delta_i|Z_i| + \kappa_i\sqrt{Q/\nu}\right) - \Phi\left(\delta_i|Z_i| - \kappa_i\sqrt{Q/\nu}\right)\right), \end{aligned} \tag{9}$$

where $\delta_i^2 = \frac{1}{n_i}$, $Z_i = (\bar{Y}_i - \mu_i)/\delta_i\sigma \sim N(0,1)$ and $Q = \nu S^2/\sigma^2 \sim \chi^2_\nu$ are mutually independent pivot random variables.

The set of tolerance factors $\kappa_1, \ldots, \kappa_m$ is exact for the simultaneous $(1-\gamma, 1-\alpha)$-tolerance intervals $(L_{\mathcal{E},i}, U_{\mathcal{E},i})$ if

$$E_{\{Z_1,\ldots,Z_m,Q\}}\Big(\mathbb{I}\left(C(\kappa_1, \ldots, \kappa_m; Z_1, \ldots, Z_m, Q) \ge 1-\gamma\right)\Big) = 1-\alpha. \tag{10}$$

This may be checked either by a Monte Carlo simulation, or by $(m+1)$-dimensional numerical integration.

Notice, however, that the solution to the equation (10) is not unique, until further restrictions are imposed on the set of possible tolerance factors $\kappa_1, \ldots, \kappa_m$. Frequently, it is required to have a common tolerance factor $\kappa$ for all simultaneous tolerance intervals, i.e. $\kappa_1 = \cdots = \kappa_m = \kappa$.

Under such restriction, the formula (7) can be generalized for computing the exact common tolerance factor $\kappa$ of the simultaneous tolerance intervals, with limits $L_{\mathcal{E},i} = \bar{Y}_i - \kappa\sqrt{S^2}$ and $U_{\mathcal{E},i} = \bar{Y}_i + \kappa\sqrt{S^2}$, such that (8) holds true. However, a relatively simple generalization is possible only under further restrictive assumption that the calibration experiment $\mathcal{E} = \{\mathcal{E}_1, \ldots, \mathcal{E}_m\}$ is based on $m$ independent samples with common sample size $n$ for each population $N(\mu_i, \sigma^2)$, i.e. with $\nu = m(n-1)$. In particular, under this restriction we get the content function

$$C(\kappa \,|\, Z_1, \ldots, Z_m, Q) = \Phi\left(\delta \max_i |Z_i| + \kappa\sqrt{Q/\nu}\right) - \Phi\left(\delta \max_i |Z_i| - \kappa\sqrt{Q/\nu}\right), \tag{11}$$

as $\Phi(a+r)-\Phi(a-r)$ is a decreasing function of $|a|$.

Then, using the analogy of (7), the generalized formula is derived by considering the distribution of the random variable $Z^m_{\max}=\max(|Z_1|,\ldots,|Z_m|)$ (where $Z_i\sim N(0,1)$, $i=1,\ldots,m$, are independent random variables) instead of $|Z|$ (where $Z\sim N(0,1)$). In summary, the exact (simultaneous) tolerance factor $\kappa$ can be computed as a solution to the equation

$$1-\alpha=2m\int_0^\infty \Gamma\left(\frac{\nu}{2},\frac{\nu}{2\kappa^2}\chi^2_{1;1-\gamma}\left(\delta^2z^2\right)\right)\left(2\Phi(z)-1\right)^{m-1}\phi(z)\,dz, \tag{12}$$

where $\delta^2=\frac{1}{n}$, $\nu=m(n-1)$, and $\chi^2_{1;1-\gamma}(\delta^2z^2)$ denotes the $(1-\gamma)$-quantile of the non-central chi-square distribution with 1 degree of freedom and the non-centrality parameter $\sqrt{\delta^2z^2}$.

For $m=1$, the tolerance factor given by the solution to the equation (12) is equivalent to the factor given by the solution to the equation (7) with $\nu=m(n-1)$. Application of such a tolerance factor leads to the non-simultaneous tolerance intervals with limits $L_{\mathcal{E},i}=\bar{Y}_i-\kappa\sqrt{S^2}$ and $U_{\mathcal{E},i}=\bar{Y}_i+\kappa\sqrt{S^2}$ for the considered $m$ populations, each fulfilling the property as defined by (1), but formally different from the individual tolerance intervals defined by (2), i.e. $L_{\mathcal{E}_i}=\bar{Y}_i-\kappa\sqrt{S_i^2}$, $U_{\mathcal{E}_i}=\bar{Y}_i+\kappa\sqrt{S_i^2}$.

### 2.3. One-sided tolerance intervals

For completeness (however, without more details), we note that the tolerance factor for the one-sided $(1-\gamma,1-\alpha)$-tolerance interval $(L_{\mathcal{E}},\infty)$ (resp. $(-\infty,U_{\mathcal{E}})$) can be computed as a quantile of the non-central $t$-distribution. In particular, the exact tolerance factor for the (non-simultaneous) upper tolerance limit $U_{\mathcal{E}}=\bar{Y}+\kappa\sqrt{S^2}$, based on a simple calibration experiment $\mathcal{E}$, is given by

$$\kappa=\delta t_{\nu,\Delta;1-\alpha}, \tag{13}$$

where $\nu=n-1$, $\Delta=\frac{z_{1-\gamma}}{\delta}$ with $\delta^2=\frac{1}{n}$ and $z_{1-\gamma}$ being the $(1-\gamma)$-quantile of the standard normal distribution, and $t_{\nu,\Delta;1-\alpha}$ denotes the $(1-\alpha)$-quantile of the non-central $t$-distribution with $\nu$ degrees of freedom and the noncentrality parameter $\Delta$. For more details see Krishnamoorthy and Mathew (2009), equations (1.2.2) and (2.2.3).

We notice an interesting (technical) relationship of the right hand side expression of the equation (7) to the CDF of the noncentral $t$-distribution with $\nu$ degrees of freedom and the noncentrality parameter $\Delta$, say $F_{t_{\nu,\Delta}}(\cdot)$. In particular,

$$F_{t_{\nu,\Delta}}(x)=\Phi(-\Delta)+\int_{-\Delta}^\infty \Gamma\left(\frac{\nu}{2},\frac{\nu}{2x^2}\left(z+\Delta\right)^2\right)\phi(z)\,dz. \tag{14}$$

This relationship allows to use similar computational strategies for computing the required tolerance factor $\kappa$, as for computing the CDF of the noncentral $t$-distribution. For more details see Witkovský (2013a).

Based on (14), and using the analogy of (12), the exact common tolerance factor $\kappa$ for simultaneous one-sided upper tolerance limits $U_{\mathcal{E}_i}=\bar{Y}_i+\kappa\sqrt{S^2}$ (for $m$ independent, equally sampled normal populations with possibly different means $\mu_i$, common variance $\sigma^2$, and with common sample size $n$) can be computed as a solution to the equation

$$1-\alpha=\Phi\left(-\frac{z_{1-\gamma}}{\delta}\right)^m+m\int_{-\frac{z_{1-\gamma}}{\delta}}^\infty \Gamma\left(\frac{\nu}{2},\frac{\nu\delta^2}{2\kappa^2}\left(z+\frac{z_{1-\gamma}}{\delta}\right)^2\right)\left(\Phi(z)\right)^{m-1}\phi(z)\,dz, \tag{15}$$

where $\nu=m(n-1)$, $\delta^2=\frac{1}{n}$, and $z_{1-\gamma}$ is the $(1-\gamma)$-quantile of the standard normal distribution. For more details and alternative derivation see Krishnamoorthy and Mathew (2009), equation (2.5.3).

## 3. Two-sided tolerance intervals for univ. normal linear regression

Here we shall assume that the calibration experiment $\mathcal{E}$ is modeled by the linear regression model $Y = X\beta + \varepsilon$, where $Y$ is an $n$-dimensional random vector of responses measured for $n$ values $x_i$, $i = 1, \ldots, n$, of the explanatory variable $x \in \mathcal{X} \subseteq \mathbb{R}^r$. However, here we shall assume that the explanatory variable is one-dimensional, i.e. that $x \in (x_{\min}, x_{\max}) \subseteq \mathbb{R}$, what is a typical situation for the frequently used $p$-order polynomial regression models.

The matrix $X$ represents the $(n \times q)$-dimensional calibration experiment design matrix with rows $f(x_i)'$, for $i = 1, \ldots, n$, i.e. $q$-dimensional functions of $r$-dimensional vectors $x_i$. For example, in simple $p$-order polynomial linear regression model we get $q = p + 1$ and $f(x_i) = (1, x_i, x_i^2, \ldots, x_i^p)'$ for $x_i \in \mathcal{X} = (x_{\min}, x_{\max})$. For simplicity, here we shall assume that $X$ is a full-ranked matrix.

Further, $\beta$ is the $q$-dimensional vector of regression coefficients and $\varepsilon$ is an $n$-dimensional vector of measurement errors with its assumed distribution $\varepsilon \sim N(0, \sigma^2 I_n)$. Based on the calibration experiment $\mathcal{E}$, we get the sufficient statistics

$$\hat{\beta} = (X'X)^{-1}X'Y, \quad S^2 = \frac{1}{\nu}(Y - X\hat{\beta})'(Y - X\hat{\beta}), \tag{16}$$

and mutually independent pivot variables

$$Z_X = \frac{\hat{\beta} - \beta}{\sigma^2} \sim N\left(0, (X'X)^{-1}\right), \quad Q = \frac{\nu S^2}{\sigma^2} \sim \chi^2_\nu, \tag{17}$$

where $\nu = n - q$.

### 3.1. Non-simultaneous tolerance intervals

The non-simultaneous tolerance intervals for the possible future realizations of the response variable $Y(x) = f(x)'\beta + \sigma\epsilon$ (where $f(x)$ is a known $q$-dimensional model function of $x \in \mathcal{X}$ and $\epsilon \sim N(0, 1)$ is independent of the calibration experiment $\mathcal{E}$), say $(L_{\mathcal{E},x}, U_{\mathcal{E},x})$, are such that

$$\mathsf{P}_{\{\mathcal{E}\}}\Big(\mathsf{P}_{\{Y(x)\}}\Big(L_{\mathcal{E},x} \le Y(x) \le U_{\mathcal{E},x} \,|\, \mathcal{E}\Big) \ge 1 - \gamma\Big) = 1 - \alpha. \tag{18}$$

Similarly as in the univariate distribution case, the limits of the two-sided tolerance intervals for linear regression, $(L_{\mathcal{E},x}, U_{\mathcal{E},x})$ for $x \in \mathcal{X}$, are typically restricted to the form

$$L_{\mathcal{E},x} = f(x)'\hat{\beta} - \kappa_x\sqrt{S^2}, \qquad U_{\mathcal{E},x} = f(x)'\hat{\beta} + \kappa_x\sqrt{S^2}, \tag{19}$$

where by $\kappa_x$ we denote the required tolerance factor at $x \in \mathcal{X}$.

Then, for the given candidate of the tolerance factor, say $\kappa_x$, the content function for the non-simultaneous tolerance interval $(L_{\mathcal{E},x}, U_{\mathcal{E},x})$ is

$$\begin{aligned} C(\kappa_x \,|\, Z_X, Q) &= \Phi\left(|f(x)'Z_X| + \kappa_x\sqrt{Q/\nu}\right) - \Phi\left(|f(x)'Z_X| - \kappa_x\sqrt{Q/\nu}\right) \\ &= \Phi\left(\delta_x|Z| + \kappa_x\sqrt{Q/\nu}\right) - \Phi\left(\delta_x|Z| - \kappa_x\sqrt{Q/\nu}\right) \\ &\equiv C(\kappa_x \,|\, Z, Q), \end{aligned} \tag{20}$$

where $Z = \frac{1}{\delta_x}f(x)'Z_X \sim N(0, 1)$, $Q \sim \chi^2_\nu$, and $\delta_x^2 = f(x)'(X'X)^{-1}f(x)$ denotes the variance of the estimator $f(x)'\hat{\beta}$ at $x \in \mathcal{X}$.

By comparing (3) and (20), it is clear that the exact tolerance factor $\kappa_x$ for the two-sided non-simultaneous tolerance interval $(L_{\mathcal{E},x}, U_{\mathcal{E},x})$, evaluated at $x \in \mathcal{X}$, can be computed by solving the equation (7), with $\nu = n - q$ and $\delta^2 = \delta_x^2 = f(x)'(X'X)^{-1}f(x)$.

Notice that the value of the exact tolerance factor $\kappa_x$ does not depend directly on the vector $x \in \mathcal{X}$ and the model design matrix $X$. In fact, it depends on the model design only through $\nu = n - q$ and $\delta_x^2$. That is, $\kappa_x$ is equal for all $x \in \mathcal{X}$, such that $\delta_x^2 = f(x)'(X'X)^{-1}f(x)$ is equal. This allows creation of tables and/or efficient interpolation-based approximations for

computing the exact non-simultaneous tolerance factors $\kappa_x$ for the univariate normal linear regression models.

### 3.2. Simultaneous tolerance intervals

The simultaneous two-sided tolerance intervals for a possibly infinite sequence of the future realizations of the response variable $Y(x) = f(x)'\beta + \sigma\epsilon$, say $(L_{\mathcal{E}}(x), U_{\mathcal{E}}(x))$ for any $x \in \mathcal{X}$, are such that

$$\begin{aligned} 1-\alpha &= \mathsf{P}_{\{\mathcal{E}\}}\Big(\mathsf{P}_{\{Y(x)\}}\Big(L_{\mathcal{E}}(x) \le Y(x) \le U_{\mathcal{E}}(x) \,|\, \mathcal{E}\Big) \ge 1-\gamma, \text{ for all } x \in \mathcal{X}\Big) \\ &= \mathsf{P}_{\{\mathcal{E}\}}\Big(\min_{x\in\mathcal{X}} \mathsf{P}_{\{Y(x)\}}\Big(L_{\mathcal{E}}(x) \le Y(x) \le U_{\mathcal{E}}(x) \,|\, \mathcal{E}\Big) \ge 1-\gamma\Big). \end{aligned} \tag{21}$$

Similarly as before, we consider the limits of the tolerance intervals to be restricted to the form

$$L_{\mathcal{E}}(x) = f(x)'\hat{\beta} - \kappa(x)\sqrt{S^2}, \qquad U_{\mathcal{E}}(x) = f(x)'\hat{\beta} + \kappa(x)\sqrt{S^2}, \tag{22}$$

where by $\kappa(x)$ we denote the tolerance factor function defined for all $x \in \mathcal{X}$.

Then, for a given candidate of the tolerance factor function, say $\kappa(x)$, the content function for the simultaneous tolerance intervals $(L_{\mathcal{E}}(x), U_{\mathcal{E}}(x))$ is given by

$$\begin{aligned} &C(\kappa(x) \,|\, Z_X, Q) = \\ &\quad \min_{x\in\mathcal{X}} \left(\Phi\left(|f(x)'Z_X| + \kappa(x)\sqrt{Q/\nu}\right) - \Phi\left(|f(x)'Z_X| - \kappa(x)\sqrt{Q/\nu}\right)\right), \end{aligned} \tag{23}$$

where $Z_X = \frac{\hat{\beta}-\beta}{\sigma^2} \sim N\left(0, (X'X)^{-1}\right)$ and $Q \sim \chi^2_\nu$ with $\nu = n - q$. Notice that the content function (23) depends on the design matrix $X$, in particular through the matrix $(X'X)^{-1}$.

The tolerance factor function $\kappa(x)$ is exact for the simultaneous $(1-\gamma, 1-\alpha)$-tolerance intervals $(L_{\mathcal{E}}(x), U_{\mathcal{E}}(x))$, for all $x \in \mathcal{X}$, if

$$E_{\{Z_X,Q\}}\Big(\mathbb{I}\left(C(\kappa(x); Z_X, Q) \ge 1-\gamma\right)\Big) = 1-\alpha. \tag{24}$$

This may be checked either by a Monte Carlo simulation, or by $(q+1)$-dimensional numerical integration. In general, evaluation of (23) and/or (24) is a computationally demanding task, as it requires minimum search over $x \in \mathcal{X}$ for each evaluation at $Z_X, Q$.

The solution to the equation (24) is not unique, until further restrictions are imposed on the form of the tolerance factor function $\kappa(x)$. In accordance with Witkovský (2013b), here we suggest considering the family of the candidate tolerance factor functions $\kappa(x)$, parametrized by the scalar parameter $\tilde{m} \ge 1$, of the form

$$\kappa(x) = \kappa\left(\delta^2(x), \nu, \tilde{m}\right), \tag{25}$$

where the function $\kappa\left(\delta^2(x), \nu, \tilde{m}\right)$ is given implicitly, for each $x \in \mathcal{X}$, as a solution to the equation (12), by setting $\delta^2 = \delta^2(x) = f(x)'(X'X)^{-1}f(x)$, $\nu = n - q$, and with $m = \tilde{m}$.

Here, the parameter $\tilde{m}$ (the *simultaneousity parameter* to be determined) represents the complexity of the regression function $f(x)'\beta$ over the considered range $x \in \mathcal{X}$. The optimum value of $\tilde{m}$ depends on the model and the design of the calibration experiment $\mathcal{E}$: the model function (e.g., the polynomial of order $p$), the considered set $\mathcal{X}$, the design matrix $X$, and the degrees of freedom $\nu$. For example, in simple linear regression (polynomial of the order $p = 1$) the value $\tilde{m} = 2$ is a good starting point for the numerical (iterative) search procedure (i.e., the complexity of the simple linear regression function for all $x \in \mathcal{X}$ is assumed to be similar to the complexity of two independent normal populations).

Another possibility, suggested in Mee *et al.* (1991), is to consider the family of functions $\kappa(x) = \kappa(\delta(x))$, linear functions of $\delta(x) = \sqrt{f(x)'(X'X)^{-1}f(x)}$, parametrized by the scalar parameter $\lambda > 0$ (a parameter to be determined). In particular,

$$\kappa(x) = \kappa(\delta(x)) = \kappa(\delta(x), q, \lambda) = \lambda\left(z_{1-\frac{\gamma}{2}} + \delta(x)\sqrt{q+2}\right), \tag{26}$$

where $z_{1-\frac{\gamma}{2}}$ is the $(1-\frac{\gamma}{2})$-quantile of the standard normal distribution. Based on that, Mee *et al.* (1991) derived their optimum tolerance function $\kappa(\delta(x))$ (however, not exact) as a solution to the equation

$$E_{\{W,Q\}}\Big(\mathbb{I}\left(\hat{C}(\kappa(\delta(x)); W, Q) \geq 1-\gamma\right)\Big) = 1-\alpha, \tag{27}$$

by using the approximate content function

$$\hat{C}(\kappa(\delta(x)) \,|\, W, Q) = \min_{\delta(x)} \left(\Phi\left(\delta(x)\sqrt{W} + \kappa(\delta(x))\sqrt{Q/\nu}\right) - \Phi\left(\delta(x)\sqrt{W} - \kappa(\delta(x))\sqrt{Q/\nu}\right)\right), \tag{28}$$

where the range of $\delta(x)$ is considered for $x \in \mathcal{X}$, and $W \sim \chi^2_q$ is independent of $Q \sim \chi^2_\nu$. Notice that the content function (28) does not depend directly on the design matrix $X$.

### 3.3. Multiple-use calibration problem

A motivation for computing tolerance intervals for the univariate normal linear regression is the multiple-use calibration problem and the associated problem of computing the calibration confidence intervals.

In many experimental sciences, acquisition of the measurement results frequently requires measurement procedures involving instrument calibration which can be modeled as a linear (polynomial) regression problem. Then, the required measurement result, say $x_*$, is obtained through measuring the observable response variable, say $Y_* = Y(x_*) = f(x_*)'\beta + \sigma\epsilon$, and by inverting the fitted regression (calibration) function. A problem of constructing and computing the appropriate confidence intervals for the unobservable values of the explanatory variable $x_*$, based on a given fitted calibration function (a result of the calibration experiment), for a possibly unlimited sequence of future observations of the response variable $Y_*$, is known as the multiple-use calibration problem.

As proposed in Scheffé (1973), such calibration intervals for $x_*$ values can be obtained from the simultaneous tolerance intervals for the considered linear regression (calibration function), with warranted minimum $(1-\gamma)$-coverage (for all such intervals simultaneously), and with confidence at least $(1-\alpha)$ (i.e. for $(1-\alpha)\times 100\%$ of possible calibration experiments). For an overview of the problem and the used methods see, e.g., Mee *et al.* (1991), Mee and Eberhardt (1996), Mathew and Zha (1997), Krishnamoorthy and Mathew (2009), Chvosteková (2013a), Chvosteková (2013b), and Witkovský (2013b).

In particular, for given observation $Y_* = Y(x_*)$, we shall construct the calibration confidence interval for the unobservable value of the explanatory variable, say $x^* \in \mathcal{X}$, by inverting the simultaneous tolerance intervals. So, the calibration confidence interval for $x_*$ is given by the random set

$$\mathcal{S}(Y_*; \mathcal{E}) = \{x \in \mathcal{X} : Y_* \in (L_{\mathcal{E}}(x), U_{\mathcal{E}}(x))\}. \tag{29}$$

The set (29) is not necessarily an interval. However, for most practical situations where the calibration function is (significantly) strictly monotonic for $x \in \mathcal{X}$, the confidence set (29) typically results in an interval. Based on (21) and (29), we can directly characterize the stochastic properties of the calibration confidence intervals:

$$\mathsf{P}_{\{\mathcal{E}\}}\Big(\mathsf{P}_{\{Y(x_*)\}}\left(x_* \in \mathcal{S}(Y(x_*) \,|\, \mathcal{E})\right) \geq 1-\gamma\Big) = 1-\alpha. \tag{30}$$

We notice, however, that from the practical point of view, such calibration confidence intervals are considered to be too conservative, and consequently, as suggested in Mee and Eberhardt (1996), usage of the non-simultaneous two-sided tolerance intervals $(L_{\mathcal{E},x}, U_{\mathcal{E},x})$ is recommended in (29), instead of using the exact simultaneous two-sided tolerance intervals $(L_{\mathcal{E}}(x), U_{\mathcal{E}}(x))$.

## 4. MATLAB algorithm

Based on (12), we have developed the MATLAB algorithm `ToleranceFactorGK`, that computes the tolerance factors $\kappa$ for the two-sided tolerance intervals by using an adaptive Gauss-Kronod quadrature. Usage of the complementary incomplete Gamma function (for computing the CDF of chi-square distribution) and the complementary error function (for computing the CDF of standard normal distribution) allows precise evaluation of the tolerance factors even for extremely small values of the probabilities $\gamma$ and/or $\alpha$ (i.e. for extremely high coverage and confidence). The complementary error function is also used to find the solution (root) $r$, of the equation $[1 - (\Phi(x + r) - \Phi(x - r))] - \gamma = 0$, by using the Halley's method (root-finding algorithm based on two function derivatives). The current version of the algorithm is available at the web page http://www.mathworks.com/matlabcentral/fileexchange/24135-tolerancefactor.

For illustration and possible comparisons with other algorithms, here we present several values of the tolerance factor $\kappa$ (presented with up to 15 decimal places) computed by the algorithm `ToleranceFactorGK` for the two-sided $(1 - \gamma, 1 - \alpha)$-tolerance interval for univariate normal population(s), based on a calibration experiment characterized by the parameters $n$, $\delta^2$, $\nu$, and $m$.

**Example 1.** Let us consider the following parameters: $\gamma = 0.01$, $\alpha = 0.05$, $n = 10$, $m = 1$, $\nu = n - 1$, and $\delta^2 = \frac{1}{n}$. The tolerance factor, defined as a solution to the equation (12), is calculated in MATLAB by using the algorithm `ToleranceFactorGK`:

```
gamma = 0.01; alpha = 0.05;
n = 10; m = 1; nu = n-1; delta2 = 1/n;
kappa = ToleranceFactorGK(n,1-gamma,1-alpha,m,nu,delta2)

kappa = 4.436908728948544
```

**Example 2.** As was explained in Section 2, by solving the equation (12), it is possible to compute the common tolerance factor also for the simultaneous tolerance intervals of $m$ populations, assuming that the common sample size for all $m$ populations is $n$. Let us consider the following parameters: $\gamma = 0.01$, $\alpha = 0.05$, $n = 10$, $m = 4$, $\nu = m(n-1)$, and $\delta^2 = \frac{1}{n}$. The common tolerance factor for the simultaneous two-sided tolerance intervals is calculated by

```
gamma = 0.01; alpha = 0.05;
n = 10; m = 4; nu = m*(n-1); delta2 = 1/n;
options.Simultaneous = true;
kappa = ToleranceFactorGK(n,1-gamma,1-alpha,m,nu,delta2,options)

kappa = 3.574857233534562
```

**Example 3.** The information from $m$ independent sources can be effectively used also if we are interested in calculating a non-simultaneous tolerance interval for one particular population. However, we wish to use the pooled sample variance estimator (i.e. with more degrees of freedom than could be achieved from one sample). So, let us consider the following parameters: $\gamma = 0.01$, $\alpha = 0.05$, $n = 10$, $m = 4$, $\nu = m(n-1) = 36$ $\delta^2 = \frac{1}{n}$. Now, the (non-simultaneous) tolerance factor is $\kappa = 3.385579684948129$. Notice that the tolerance factor can be calculated also if we directly set $m = 1$ and $\nu = 36$ (if $m = 1$ the calculated tolerance factor is non-simultaneous).

```
gamma = 0.01; alpha = 0.05;
n = 10; m = 4; nu = m*(n-1); delta2 = 1/n;
options.Simultaneous = false;
kappa = ToleranceFactorGK(n,1-gamma,1-alpha,m,nu,delta2,options)

kappa = 3.385579684948129
```

**Example 4.** In order to illustrate the ability to compute the tolerance factors even for extremely large values of the coverage and confidence probabilities, let us consider the following

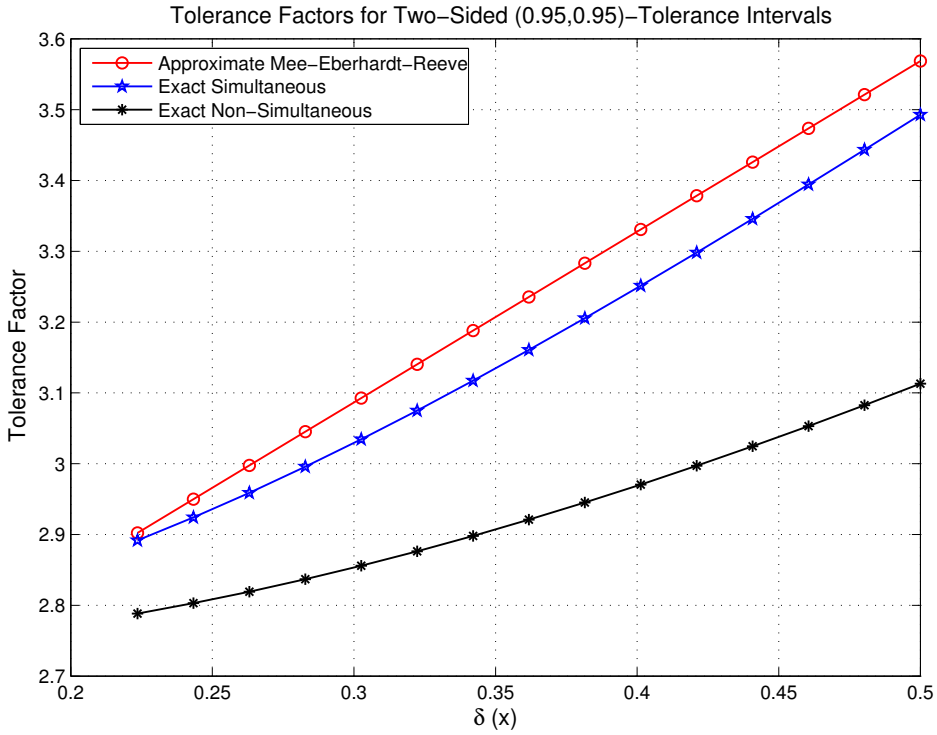


Figure 1: Tolerance factors for the two-sided $(0.95, 0.95)$-tolerance intervals evaluated at 15 equidistant points $\delta(x) \in (\delta_{\min}, \delta_{\max})$.

parameters: $\gamma = 10^{-5}$, $\alpha = 10^{-18}$, $n = 250$, $m = 1$, $\nu = n - 1$, $\delta^2 = \frac{1}{n}$. The calculated value of the tolerance factor is given by

```
gamma = 1e-5; alpha = 1e-18;
n = 250; m = 1; nu = n-1; delta2 = 1/n;
options.TailProbability = true;
kappa = ToleranceFactorGK(n,gamma,alpha,m,nu,delta2,options)

kappa = 6.967664575030617
```

**Example 5.** The algorithm can be used directly for computing the exact tolerance factors of the non-simultaneous two-sided tolerance intervals for normal linear regression models, and also, by using further optimization (used for finding the optimum value of $\tilde{m}$) for computing the exact tolerance factors of the simultaneous two-sided tolerance intervals.

For illustration, let us consider a calibration experiment for simple linear regression: $Y = X\beta + \varepsilon$, where $X$ is an $(n \times 2)$ design matrix with $n = 20$. The first column of $X$, representing the intercept, is a column of ones, the second column has two distinct elements: $-1$ for the first 10 rows and 1 for the last 10 rows. So, $(X'X)^{-1}$ is a diagonal matrix with both diagonal elements equal to $\frac{1}{n} = 0.05$, and consequently, $\delta(x) = \sqrt{(1, x)(X'X)^{-1}(1, x)'} = \sqrt{\frac{1}{n}(1 + x^2)}$.

We wish to compute the tolerance factors for the two-sided tolerance intervals with $x \in \mathcal{X} = (-2, 2)$, i.e. for $\delta(x) \in (\delta_{\min}, \delta_{\max}) = \left(\sqrt{\frac{1}{n}}, \sqrt{\frac{1}{n}(1 + 2^2)}\right) = (0.2236, 0.5)$.

Figure 1 plots the values of the exact non-simultaneous, the exact simultaneous and the approximate tolerance factors, calculated for 15 equidistant values of $\delta(x) \in (\delta_{\min}, \delta_{\max})$. The exact non-simultaneous tolerance factors were calculated by (12), with $n = 20$, $\nu = n - q = 18$, and $m = 1$. The exact simultaneous tolerance factors were calculated according to (24) and (25) with $\tilde{m} = 4.3$, found by Monte Carlo based optimization, and $\nu = n - q = 18$. The approximate tolerance factors were calculated by (26), with the optimum value of the parameter $\lambda = 1.2057$ taken from Table 2 in Mee *et al.* (1991), for $n = 20$ and $\tau = 2$. Here is the MATLAB code used for computing the tolerance factors:

```
%% Exact non-simultaneous tolerance factors:
gamma = 0.05; alpha = 0.05;
n = 20; q = 2; nu = (n-q); m = 1;
delta_min = sqrt(1/n); delta_max = sqrt((1+2^2)/n); N = 15;
delta = linspace(delta_min,delta_max,N)';
kappa_NonSim = zeros(N,1);
for i = 1:N
    kappa_NonSim(i) = ...
      ToleranceFactorGK(n,1-gamma,1-alpha,m,nu,delta(i)^2);
end

%% Exact simultaneous tolerance factors:
gamma = 0.05; alpha = 0.05;
n = 20; q = 2; nu = (n-q); m = 4.3;
options.Simultaneous = true;
kappa_Sim = zeros(N,1);
for i = 1:N
    kappa_Sim(i) = ...
      ToleranceFactorGK(n,1-gamma,1-alpha,m,nu,delta(i)^2,options);
end

%% Approximate Mee-Eberhardt-Reeve tolerance factors:
lambda_MER = 1.2334;
z_quantile = norminv(1-gamma/2);
kappa_MER = zeros(N,1);
for i = 1:N
    kappa_MER(i) = ...
      lambda_MER * (norminv(1-gamma/2) + sqrt(2+q)*delta(i));
end
```

## 5. Discussion

The motivation for computing the exact simultaneous tolerance intervals for univariate normal distributions and univariate normal linear regression models is rather strong. However, the required methods and algorithms for computing the tolerance factors are more complicated, than those for computing the non-simultaneous tolerance intervals. The efficient algorithms are still missing in the commonly used statistical packages.

The main goal of the paper was to advocate the usage of the exact and/or approximate tolerance intervals. We have presented a brief overview of the theoretical background and approaches for computing the tolerance factors based on samples from one or several univariate normal populations, and also presented the methods for computing the tolerance factors for the non-simultaneous and simultaneous two-sided tolerance intervals for univariate normal linear regression. For a more comprehensive overview of the models and methods for tolerance intervals and tolerance regions we suggest to consult the book Krishnamoorthy and Mathew (2009).

## Acknowledgements

This work is partly based on the invited lecture presented at the 10th International Conference on Computer Data Analysis & Modeling - CDAM 2013, Belarus State University, Minsk, September 10-14, 2013, see Witkovský (2013c). Financial support from the Slovak Research and Development Agency, projects APVV-0096-10, SK-AT-0025-12, and from the Scientific Grant Agency of the Ministry of Education of the Slovak Republic and the Slovak Academy of Sciences, projects VEGA 2/0038/12 and VEGA 2/0043/13, is greatly acknowledged.

## References

Chvosteková M (2013a). "Simultaneous Two-Sided Tolerance Intervals for a Univariate Linear Regression Model." *Communications in Statistics - Theory and Methods*, **42**, 1145–1152.

Chvosteková M (2013b). "Two-Sided Tolerance Intervals in a Simple Linear Regression." *Acta Universitatis Palackianae Olomucensis, Facultas Rerum Naturalium, Mathematica*, **52**(2), 31–41.

Hahn G, Meeker W (1991). *Statistical Intervals: A Guide for Practitioners*. Wiley Interscience, New York.

Krishnamoorthy K, Mathew T (2009). *Statistical Tolerance Regions: Theory, Applications, and Computation*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Liu W (2011). *Simultaneous Inference in Regression*. CRC Press, Taylor & Francis Group.

Mathew T, Zha W (1997). "Multiple Use Confidence Regions in Multivariate Calibration." *Journal of the American Statistical Association*, **92**, 1141–1150.

Mee R, Eberhardt K (1996). "A Comparison of Uncertainty Criteria for Calibration." *Technometrics*, **38**, 221–229.

Mee R, Eberhardt K, Reeve C (1991). "Calibration and Simultaneous Tolerance Intervals for Regression." *Technometrics*, **33**, 211–219.

Odeh R, Owen D (1980). *Tables for Normal Tolerance Limits, Sampling Plans, and Screening*. Marcel Dekker, Inc., New York.

Scheffé H (1973). "A Statistical Theory of Calibration." *Annals of Statistics*, **1**, 1–37.

Witkovský V (2013a). "A Note on Computing Extreme Tail Probabilities of the Noncentral $t$-Distribution with Large Noncentrality Parameter." *Acta Universitatis Palackianae Olomucensis, Facultas Rerum Naturalium, Mathematica*, **52**(2), 131–143.

Witkovský V (2013b). "On Exact Multiple-Use Linear Calibration Confidence Intervals." In *MEASUREMENT 2013, Proceedings of the 9th International Conference on Measurement*, pp. 35–38. Institute of Measurement Science, Slovak Academy of Sciences, Bratislava, Smolenice, Slovakia, May 27-30, 2013.

Witkovský V (2013c). "On the Exact Tolerance Intervals for Univariate Normal Distribution." In S Aivazian, P Filzmoser, Y Kharin (eds.), *CDAM 2013, Computer Data Analysis and Modeling: Theoretical and Applied Stochastics, Proceedings of the 10th International Conference*, volume 1, pp. 130–137. Belarusian State University, Minsk, Minsk, Belarus, September 10-14.

Young D (2010). "An R Package for Estimating Tolerance Intervals." *Journal of Statistical Software*, **36**, 1–39.

**Affiliation:**

Viktor Witkovský
Institute of Measurement Science
Slovak Academy of Sciences
Dúbravská cesta 9
SK-841 04 Bratislava, Slovakia
E-Mail: witkovsky@savba.sk

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 43/3-4
June 2014

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

*Submitted:* 2013-11-01
*Accepted:* 2014-03-10

# News and Announcements

---

Regular meetings of the Vienna R Meetup Group find place approx. every two months. The meetup group is supported by Revolutions Analytics and data-analysis OG. More information on past and future presentations at the meetup, the orgianisation of the meetup group, members and discussions can be found at http://www.meetup.com/ViennaR/ .

---

The annual useR!2014 conference takes place at UCLA campus in Los Angeles from June 30 – July 3, 2014. The registration fee is fair and tutorials are for free.

---

Funny stories and insights by Andreas Quatember to reflect on (wrongly presented) statistics in the (mostly Austrian) news can be found at http://www.jku.at/ifas/content/e101235. Worth reading.

---

Matthias Templ

ISSN 1026 - 597X

# Contents

**Page**