Springer Proceedings in Mathematics & Statistics

Matúš Maciak
Michal Pešta
Martin Schindler *Editors*

Analytical Methods in Statistics

AMISTAT, Liberec, Czech Republic, September 2019

Springer

Springer Proceedings in Mathematics & Statistics

Volume 329

Springer Proceedings in Mathematics & Statistics

This book series features volumes composed of selected contributions from workshops and conferences in all areas of current research in mathematics and statistics, including operation research and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

More information about this series at http://www.springer.com/series/10533

Matúš Maciak · Michal Pešta · Martin Schindler
Editors

Analytical Methods in Statistics

AMISTAT, Liberec, Czech Republic, September 2019

Springer

Editors
Matúš Maciak
Department of Probability
and Mathematical Statistics
Charles University
Prague, Czech Republic

Michal Pešta
Department of Probability
and Mathematical Statistics
Charles University
Prague, Czech Republic

Martin Schindler
Department of Applied Mathematics
Technical University of Liberec
Liberec, Czech Republic

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-030-48813-0 ISBN 978-3-030-48814-7 (eBook)
https://doi.org/10.1007/978-3-030-48814-7

Mathematics Subject Classification: 62B05, 62C12, 62F03, 62F12, 62F35, 62F40, 62G08, 62G20

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This proceeding volume follows the third workshop on Analytical Methods in Statistics (AMISTAT 2019) which took place in Liberec (Czech Republic) in September 16–19, 2019. The workshop was organized after two successful workshops AMISTAT 2015 and AMISTAT 2011 by the Department of Applied Mathematics at the Faculty of Science, Humanities and Education at the Technical University of Liberec and the Department of Probability and Mathematical Statistics at the Faculty of Mathematics and Physics at the Charles University.

The workshop brought together several scientists, researchers, and young scholars from around the World (Austria, Belgium, Canada, Finland, France, India, Russia, Serbia, Slovakia, Sweden, Turkey, UK, USA, and the Czech Republic) and it offered a great opportunity for a series of interesting and highly appreciated talks, many formal and informal discussions, and also some valuable time spent together among colleagues and friends. Many interesting topics and ideas arising from everyday problems were covered in various contributions focusing especially at the analytical methods in statistics, asymptotics, estimation and Fisher information, robustness, stochastic models and inequalities, and many others. A small part of the contributions, by those authors who considered their work being partially complete, is contained in the present book.

The joint motto of this proceeding, the workshop itself, and the talks is the "analytical statistics" with the main emphasizes on the fact that the statistics always provides mathematicians with new, challenging, and also exciting problems as they are usually all based in some real life situations. For such problems one can rarely determine any axioms or deterministic approaches, but rather analytical methods are needed to asses the problems properly. The statisticians, in general, utilize a knowledge from all areas of mathematics including abstract calculations, numerical and algorithmic computation, or formal interpretation of the results. Statisticians are, therefore, expected to find a solution to real problems and the answer that the solution does not exist is not plausible. At least a solution optimal under some acceptable constraints is used as a counterpart. The AMISTAT 2019 workshop was again full of such fresh ideas, proposals, and approaches. We are grateful for this opportunity and also for the contributions included in this proceeding book.

Last, but not least, we would like to express our great thanks to Dr. Veronika Rosteck, Springer Editor of Statistics, for her effort, encouragement, and help needed when preparing this book. We thank all the authors of the chapters and all referees for their work, consideration, and appreciation.

Prague, Czech Republic
April 2020

Matúš Maciak
Michal Pešta
Martin Schindler

Contents

Contributors

Rauf Ahmad Department of Statistics, Uppsala University, Uppsala, Sweden

Olcay Arslan Faculty of Science, Department of Statistics, Ankara University, Tandogan, Ankara, Turkey

Pei Geng Department of Mathematics, College of Arts and Sciences, Illinois State University, Normal, IL, USA

Yeşim Güney Faculty of Science, Department of Statistics, Ankara University, Tandogan, Ankara, Turkey

Jana Jurečková Faculty of Mathematics and Physics, Department of Statistics, Charles University in Prague, Prague, Czech Republic

Jan Kalina The Czech Academy of Sciences, Institute of Computer Science, Praha 8, Czech Republic

Hira L. Koul Department of Statistics and Probability, Michigan State University, East Lansing, MI, USA

Matúš Maciak Faculty of Mathematics and Physics, Department of Probability and Mathematical Statistics, Charles University, Prague, Czech Republic

Ivan Mizera University of Alberta, Mathematical and Statistical Sciences, Edmonton, Alberta, Canada

Klaus Nordhausen Institute of Statistics & Mathematical Methods in Economics, Vienna University of Technology, Vienna, Austria

Michal Pešta Faculty of Mathematics and Physics, Department of Probability and Mathematical Statistics, Charles University, Prague, Czech Republic

Una Radojičić Institute of Statistics & Mathematical Methods in Economics, Vienna University of Technology, Vienna, Austria

Petra Vidnerová The Czech Academy of Sciences, Institute of Computer Science, Praha 8, Czech Republic

Sebastiano Vitali Faculty of Mathematics and Physics, Department of Probability and Mathematical Statistics, Charles University, Prague, Czech Republic;
University of Bergamo, Department of Management, Economics and Quantitative Methods, Bergamo, Italy

Silvelyn Zwanzig Department of Mathematics, Uppsala University, Uppsala, Sweden

Averaged Autoregression Quantiles in Autoregressive Model

Yeşim Güney, Jana Jurečková, and Olcay Arslan

Abstract This paper considers the averaged autoregression quantile in autoregressive models. Our primary interest is its structure, qualities, and its applications. Moreover, under the local heteroscedasticity we investigate the properties of averaged autoregression quantile. For an illustration, a simulation study is provided.

Keywords Autoregressive model · Local heteroscedasticity · Quantile autoregression

1 Introduction

Quantile regression (QR) introduced by [1] has been increasingly used in many applied areas (for more details see [2–5]). The application of QR has subsequently moved into the areas of time-series as well as other subjects. In the time series context, there is also a rich literature including works [6–16] and the references therein. In addition to these studies, many authors also consider the time series that exhibit conditional heteroscedasticity. See [17] and [18] for related studies in linear autoregressive conditional heteroskedasticity (ARCH) models. Also, see [19] for quantile regression estimation of the linear generalized autoregressive conditional heteroskedasticity (GARCH) model and [20] for quantile regression for autoregressive moving average (ARMA) models with asymmetric GARCH (AGARCH) errors.

Y. Güney (✉) · O. Arslan
Faculty of Science, Department of Statistics, Ankara University, 06100 Tandogan,
Ankara, Turkey
e-mail: ydone@ankara.edu.tr

O. Arslan
e-mail: arslan@ankara.edu.tr

J. Jurečková
Faculty of Mathematics and Physics, Department of Statistics, Charles University in Prague,
Sokolovská 83, 186 75 Prague, Czech Republic
e-mail: jurecko@karlin.mff.cuni.cz

M. Maciak et al. (eds.), *Analytical Methods in Statistics*, Springer Proceedings
in Mathematics & Statistics 329, https://doi.org/10.1007/978-3-030-48814-7_1

Ignoring the heteroscedasticity in time series analysis may result in inefficient estimation of unconditional mean function and unreliable inferences. One approach to deal with the heteroscedasticity is to consider the conditional heteroscedasticity and to assume that the innovations follow ARCH or GARCH models as proposed by Engle [21] and Bollerslev [22]. Although the GARCH-type model is successful, it has a disadvantage of nonrobustness to the stationarity assumption [23]. An alternative way to handle the heteroscedasticity in time series models is to consider the unconditional heteroscedasticity. Comparing the conditional heteroscedastic models, the unconditional heteroscedasticity is easier to handle. Therefore, in this paper we consider the unconditional heteroscedastic autoregressive model.

Testing the heteroscedasticity in regression model has been considered in [24, 25]. It has been shown in [26] that the rank tests for regression are asymptotically insensitive to the local heteroscedasticity, provided the error distribution is symmetric. However, little attention has been paid to the issue of testing the heteroscedasticity in quantile autoregression models. Therefore, we consider testing the homoscedasticity against the local (Pitman) heteroscedasticity in the autoregressive model. Moreover, this study also focuses on the analogs of the averaged regression quantile in the autoregressive model.

The paper is organized as follows. Section 2 includes the definitions of α-autoregression quantile and autoregression rank scores for the autoregressive model. In Sect. 3, we introduce the averaged autoregression α-quantile (AAQ(α)) and give some of its properties. In Sect. 4, we consider the local heteroscedasticity in the autoregressive model. We give some numerical results to highlight the potential of the AAQ(α) in Sect. 5.

2 Autoregression Quantile and Autoregression Rank Scores for AR(p) Model

Consider the p-order autoregressive model (AR(p))

$$X_t = \phi_0 + \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \epsilon_t, \ t \in \mathbb{Z} \tag{1}$$

where $\boldsymbol{\phi} = \left(\phi_0, \phi_1, \ldots, \phi_p\right)$ is the unknown parameter vector. The order of the autoregressive model p is assumed to be finite and known. The following assumptions are imposed throughout this paper.

The innovations ϵ_t are assumed to be independently and identically distributed (iid) according to a continuous distribution function F with

(A1) $E\left(\epsilon_t\right) = 0, \ Var\left(\epsilon_t\right) = \sigma^2 < \infty.$

(A2) To gurantee the stationarity of the process, we assume that ϕ_j are such that all roots of the equation

$$z^p - \phi_1 z^{p-1} - \phi_2 z^{p-2} - \cdots - \phi_p = 0 \tag{2}$$

are inside the unit circle.

Under the assumptions given in (A1)–(A2), the process $\{X_t\}$ is casual and invertible [27]. The set of all parameter values satisfying (2) is denoted by $\mathbf{P}$.

The distribution function F of ϵ_t is unknown, but we assume that it is increasing on the set $\{\epsilon : 0 < F(\epsilon) < 1\}$. For any fixed $\alpha \in (0, 1)$, denote $\epsilon_{t\alpha} = \epsilon_t - F^{-1}(\alpha)$, $t = 1, \ldots, n$. Then $\epsilon_{1\alpha}, \ldots, \epsilon_{n\alpha}$ are iid with distribution function $F_\alpha(\epsilon) = F(\epsilon + F^{-1}(\alpha))$, $\epsilon \in \mathbb{R}$ and $F_\alpha^{-1}(\epsilon) = F^{-1}(\epsilon) - F^{-1}(\alpha)$, $0 < \epsilon < 1$ so that $F_\alpha^{-1}(\alpha) = 0$. The model given in Eq. (2) can be rewritten as

$$X_t = \phi_0(\alpha) + \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \epsilon_{t\alpha} \tag{3}$$

where $\phi_0(\alpha) = \phi_0 + F^{-1}(\alpha)$.

Before we proceed to the AAQ(α), we first illustrate the autoregression quantile estimation for the AR(p) model. The α-autoregression quantile was first considered by [9], and later studied by [28] and by [29]. For the sake of simplicity, we also use the notation $\mathbf{Y}^*_{t-1} = \left(X_{t-1}, \ldots, X_{t-p}\right)^T$ and $\mathbf{Y}_{t-1} = \left(1, X_{t-1}, \ldots, X_{t-p}\right)^T$. Define $\mathbf{Z}^*_n$ and $\mathbf{Z}_n$ matrices whose t-th rows are $\mathbf{Y}^{*T}_{t-1}$ and $\mathbf{Y}^T_{t-1}$ for $1 \le t \le n$, respectively.

Let $\left(X_{-p+1}, \ldots, X_0, X_1, \ldots, X_n\right)$ be the observed series from model (1). For the identifiability, assume that the first p observations $\left(X_{-p+1}, \ldots, X_0\right)$ are known. We work with the rest of the observations $(X_1, \ldots, X_n)$. The α-th autoregression quantile estimator $\widehat{\boldsymbol{\phi}}(\alpha) = \left(\widehat{\phi}_0(\alpha), \widehat{\boldsymbol{\phi}}^*(\alpha)^T\right)^T$ is defined by [9] as follows

$$\arg\min_{\mathbf{b} \in \mathbb{R}^{p+1}} \sum_{t=1}^{n} h_\alpha\left(X_t - \mathbf{Y}^T_{t-1}\mathbf{b}\right) \tag{4}$$

where $h_\alpha(u) = |u| \{\alpha I(u > 0) + (1 - \alpha) I(u < 0)\}$, $u \in \mathbb{R}$, $\alpha \in [0, 1]$ and $I(\cdot)$ is the indicator function. This minimization problem can also be written

$$\arg\min_{\mathbf{b} \in \mathbb{R}^{p+1}} \left\{ \sum_{t=1}^{n} \left(\alpha \left[X_t - \mathbf{Y}^T_{t-1}\mathbf{b}\right]^+ + (1 - \alpha)\left[X_t - \mathbf{Y}^T_{t-1}\mathbf{b}\right]^- \right) \right\}, \tag{5}$$

where $z^+ = \max\{z, 0\}$, and $z^- = \max\{-z, 0\}$ for $z \in \mathbb{R}$. Here $\widehat{\boldsymbol{\phi}}(\alpha) = \left(\widehat{\phi}_0(\alpha), \widehat{\boldsymbol{\phi}}^*(\alpha)^T\right)^T$ coincides with $\widehat{\mathbf{b}}$ part of the optimal solution $\left(\widehat{\mathbf{b}}, \widehat{\boldsymbol{\mu}}^+, \widehat{\boldsymbol{\mu}}^-\right)$ of the parametric programming problem

$$\begin{cases} \alpha \mathbf{1}^T_{n-p}\boldsymbol{\mu}^+ + (1 - \alpha)\,\mathbf{1}^T_n \boldsymbol{\mu}^- := \min \\ \mathbf{X}_n - \mathbf{Z}_n \mathbf{b} = \boldsymbol{\mu}^+ - \boldsymbol{\mu}^- \\ \mathbf{b} \in \mathbb{R}^{p+1}, \boldsymbol{\mu}^\pm \in \mathbb{R}^n_+, \ 0 < \alpha < 1 \end{cases} \tag{6}$$

where $\mathbf{1}_n = (1, 1, \ldots, 1)^T \in \mathbb{R}^n$ and $\mathbf{X}_n = \left(X_{p+1}, \ldots, X_n\right)^T$ ([9]).

The idea of [30] is adapted to the autoregressive context and the autoregression rank scores

$$\widehat{\mathbf{a}}(\alpha) = \left(\widehat{a}_{n;1}(\alpha), \ldots, \widehat{a}_{n;n}(\alpha)\right)^T, \ 0 \le \alpha \le 1 \tag{7}$$

are defined by [9] as the solution of the following dual program

$$\begin{cases} \mathbf{X}_n^T \mathbf{a} := \max \\ \mathbf{Z}_n^T \mathbf{a} = (1-\alpha)\, \mathbf{Z}_n^T \mathbf{1}_n \\ \mathbf{a} \in [0,1]^n\,, 0 < \alpha < 1. \end{cases} \tag{8}$$

Since $\widehat{\boldsymbol{\phi}}(\alpha)$ and $\widehat{\mathbf{a}}(\alpha)$ are dual to each other, the components of the dual solutions $\widehat{\mathbf{a}}(\alpha)$ can be expressed by

$$\widehat{a}_{n;t}(\alpha) = \begin{cases} 1, & X_t > \mathbf{Y}_{t-1}^T \widehat{\boldsymbol{\phi}}(\alpha)\,, \\ 0, & X_t < \mathbf{Y}_{t-1}^T \widehat{\boldsymbol{\phi}}(\alpha)\,,\ t = 1, \ldots, n \end{cases} \tag{9}$$

and if $X_t = \mathbf{Y}_{t-1}^T \widehat{\boldsymbol{\phi}}(\alpha)$ for some t then $0 < \widehat{a}_{n;t}(\alpha) < 1$. It is clear that $\widehat{\mathbf{a}}(\alpha)$ are continuous, piecewise linear, and such that $\widehat{a}_{n;t}(0) = 1$ and $\widehat{a}_{n;t}(1) = 0$.

Throughout the rest of the paper, we still impose the following conditions

(A3) Suppose that the distribution function F has a continuous density f that is positive on the support of F.

(A4) $0 < \int x^4 dF < \infty$

(A5) Define Σ_n and Σ_n^* matrices by $\Sigma_n = n^{-1}\mathbf{Z}_n^T \mathbf{Z}_n = n^{-1} \sum_{t=1}^{n} \mathbf{Y}_{t-1}\mathbf{Y}_{t-1}^T$ and $\Sigma_n^* = n^{-1}\mathbf{Z}_n^{*T} \mathbf{Z}_n^* = n^{-1} \sum_{t=1}^{n} \mathbf{Y}_{t-1}^* \mathbf{Y}_{t-1}^{*T}$. Assume that there exist positively definite matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^*$ such that, as $n \to \infty$,

$$\boldsymbol{\Sigma}_n \xrightarrow{P} \boldsymbol{\Sigma} \text{ and } \boldsymbol{\Sigma}_n^* \xrightarrow{P} \boldsymbol{\Sigma}^*.$$

Under the Assumptions (A1)–(A5), the following Bahadur type representation of the α-th autoregression quantile has been obtained in [9].

Theorem 1 *Under above assumptions, the α-autoregression quantile admits the following asymptotic representation:*

$$\begin{aligned} &\widehat{\boldsymbol{\phi}}(\alpha) - \left(\phi_0 + F^{-1}(\alpha)\,, \phi_1, \ldots, \phi_p\right)^T \\ &= n^{-1}\Sigma_n^{-1}\left(f\left(F^{-1}(\alpha)\right)\right)^{-1} \sum_{t=1}^{n} \mathbf{Y}_{t-1}\left(\alpha - I\left[\epsilon_t \le F^{-1}(\alpha)\right]\right) + O_p\left(n^{-1/2}\right) \end{aligned} \tag{10}$$

as $n \to \infty$, and the convergence is uniform over each subinterval $[\alpha^*, 1-\alpha^*] \subset (0,1)$.

3 Averaged Autoregression Quantile

The averaged regression quantile was first considered in [31] and defined in [32]. Its properties and the asymptotic equivalence to the α-quantile of the location model was considered in [32]. The averaged regression quantile was used in hypothesis testing and extreme events. For example, some tests based on weighted averaged regression quantile for testing of the Gumbel domain of attraction against Fréchet or Weibull domains are proposed in [33]. Jurečková [34] considered the averaged regression quantile in the context of extreme events and defined averaged extreme regression quantile.

An alternative to the averaged regression quantile is the averaged version of the two-step regression α-quantile, defined in [35]. There is also shown that the average regression quantile process is asymptotically equivalent to the location quantile process and that it converges to a Gaussian process in the Skorokhod topology. Further [36] investigated some properties of the averaged two-step regression quantile and considered the probabilistic risk assessment in the situation when the return depends on some exogenous variables.

Similar applies to the averaged autoregression quantile [the AAQ(α)], which is defined as follows.

$$\begin{aligned}\overline{B}_n(\alpha) &= \widehat{\phi}_0(\alpha) + \frac{1}{n}\sum_{t=1}^{n}\mathbf{Y}_{t-1}^{*T}\widehat{\boldsymbol{\phi}}^*(\alpha) \\ &= \overline{\mathbf{Z}}_n^T\widehat{\boldsymbol{\phi}}(\alpha). \qquad (11)\end{aligned}$$

The averaged AR quantile $\overline{B}_n(\alpha)$ has the following useful properties, analogous to those of the α-AR quantiles, proven in [12].

Lemma 1 *(i) If* $\alpha \in (0, 1)$ *is a continuity point of* $\overline{B}_n(\alpha)$*, then*

$$\overline{B}_n(\alpha) = -\frac{1}{n}\sum_{t=1}^{n} X_{t-1}\frac{d}{d\alpha}\widehat{a}_t(\alpha). \qquad (12)$$

(ii) $\overline{B}_n(\alpha)$ *is nondecreasing step-function of* $\alpha \in (0, 1)$.

Proof The duality between $\overline{B}_n(\alpha)$ and $\widehat{\mathbf{a}}(\alpha)$ given in (8) implies that

$$\sum_{t=1}^{n} h_\alpha\left(X_t - \mathbf{Y}_{t-1}^T\widehat{\boldsymbol{\phi}}(\alpha)\right) = \sum_{t=1}^{n} X_t\left(\widehat{a}_t(\alpha) - (1-\alpha)\right).$$

For $0 < \alpha_1 < \alpha_2 < 1$, we obtain

$$
\begin{aligned}
&\sum_{t=1}^{n} h_{\alpha_2}\left(X_t - \mathbf{Y}_{t-1}^T \widehat{\boldsymbol{\phi}}(\alpha_1)\right) - h_{\alpha_1}\left(X_t - \mathbf{Y}_{t-1}^T \widehat{\boldsymbol{\phi}}(\alpha_1)\right) \\
&= (\alpha_2 - \alpha_1) \sum_{t=1}^{n} \left(X_t - \mathbf{Y}_{t-1}^T \widehat{\boldsymbol{\phi}}(\alpha_1)\right) \\
&\geq \sum_{t=1}^{n} X_t \left(\widehat{a}_t(\alpha_2) - \widehat{a}_t(\alpha_1) + (\alpha_2 - \alpha_1)\right).
\end{aligned}
$$

Then we can write

$$
(\alpha_2 - \alpha_1) \sum_{t=1}^{n} \mathbf{Y}_{t-1}^T \widehat{\boldsymbol{\phi}}(\alpha_1) \leq -\sum_{t=1}^{n} \left(X_t \left(\widehat{a}_t(\alpha_2) - \widehat{a}_t(\alpha_1)\right)\right). \tag{13}
$$

Similarly, we get

$$
(\alpha_2 - \alpha_1) \sum_{t=1}^{n} \mathbf{Y}_{t-1}^T \widehat{\boldsymbol{\phi}}(\alpha_2) \geq -\sum_{t=1}^{n} \left(X_t \left(\widehat{a}_t(\alpha_2) - \widehat{a}_t(\alpha_1)\right)\right). \tag{14}
$$

Using (13) and (14), we obtain

$$
\overline{B}_n(\alpha_1) \leq -\frac{1}{n} \sum_{t=1}^{n} X_t \frac{\widehat{a}_t(\alpha_2) - \widehat{a}_t(\alpha_1)}{\alpha_2 - \alpha_1} \leq \overline{B}_n(\alpha_2) \tag{15}
$$

As $\alpha_2 \to \alpha_1$, we obtain the result. □

Next we will study the asymptotic property of $\overline{B}_n(\alpha)$.

Theorem 2 *Under the conditions of Theorem 1, for fixed* $\alpha \in (0, 1)$

$$
n^{1/2} \left(\overline{B}_n(\alpha) - \frac{1}{n} \sum_{t=1}^{n} \mathbf{Y}_{t-1}^T \, \boldsymbol{\phi} \boldsymbol{-} \, \epsilon_{n:[n\alpha]} \right) = O_p\left(n^{-1/4}\right) \tag{16}
$$

as n $\to \infty$, *where* $\epsilon_{n:1} \leq \epsilon_{n:2} \leq \cdots \leq \epsilon_{n:n}$ *are the order statistics corresponding to* $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$.

Proof Let $\widetilde{\boldsymbol{\phi}}(\alpha) = \left(\phi_0 + F^{-1}(\alpha), \phi_1, \ldots, \phi_p\right)^T$. Using Eq. (10), we can write

$$
\begin{aligned}
&n^{1/2} \overline{\mathbf{Z}}_n^T \left(\widehat{\boldsymbol{\phi}}_n(\alpha) - \widetilde{\boldsymbol{\phi}}(\alpha)\right) \qquad (17) \\
&= n^{-1/2} \left(f\left(F^{-1}(\alpha)\right)\right)^{-1} \mathbf{Z}_n^T \left(\mathbf{Z}_n^T \mathbf{Z}_n\right)^{-1} \sum_{t=1}^{n} \mathbf{Y}_{t-1} \left(\alpha - I\left[\epsilon_t \leq F^{-1}(\alpha)\right]\right) \\
&+ O_p\left(n^{-1/4}\right), \quad \text{and}
\end{aligned}
$$

$$
\begin{aligned}
& n^{1/2}\overline{\mathbf{Z}}_n^T\left(\widehat{\phi}_n(\alpha)-\widetilde{\phi}(\alpha)\right)\\
&=\frac{1}{n^{1/2}f\left(F^{-1}(\alpha)\right)}\sum_{t,k=1}^{n}\left[\mathbf{Y}_{t-1}^T\left(\overline{\mathbf{Z}}_n^T\overline{\mathbf{Z}}_n\right)^{-1}\mathbf{Y}_{k-1}\left(\alpha-I\left[\epsilon_t\le F^{-1}(\alpha)\right]\right)\right]\\
&+O_p\left(n^{-1/4}\right)\\
&=\frac{1}{n^{1/2}\left(f\left(F^{-1}(\alpha)\right)\right)}\mathbf{1}_n^T\widehat{\mathbf{H}}_n\mathbf{c}_n(\alpha)+O_p\left(n^{-1/4}\right)\\
&=\frac{1}{n^{1/2}\left(f\left(F^{-1}(\alpha)\right)\right)}\mathbf{1}_n^T\mathbf{c}_n(\alpha)+O_p\left(n^{-1/4}\right)\\
&=\frac{1}{n^{1/2}\left(f\left(F^{-1}(\alpha)\right)\right)}\sum_{t=1}^{n}\left(\alpha-I\left[\epsilon_t\le F^{-1}(\alpha)\right]\right)+O_p\left(n^{-1/4}\right)\\
&=\sqrt{n}\left(\epsilon_{n:[n\alpha]}-F^{-1}(\alpha)\right)+O_p\left(n^{-1/4}\right)
\end{aligned}
$$

where $\mathbf{c}_n(\alpha)=\left(\alpha-I\left[\epsilon_1\le F^{-1}(\alpha)\right],\ldots,\alpha-I\left[\epsilon_n\le F^{-1}(\alpha)\right]\right)^T$ and $\widehat{\mathbf{H}}_n$ is the projection matrix defined by $\widehat{\mathbf{H}}_n=\mathbf{Z}_n\left(\mathbf{Z}_n^T\mathbf{Z}_n\right)^{-1}\mathbf{Z}_n^T$. Consequently, we obtain

$$
n^{1/2}\left(\overline{B}_n(\alpha)-\frac{1}{n}\sum_{t=1}^{n}Y_{t-1}^T\,\boldsymbol{\phi}-\epsilon_{n:[n\alpha]}\right)=O_p\left(n^{-1/4}\right). \tag{18}
$$

□

4 Local Heteroscedasticity in Autoregressive Model

We explore the local heteroscedasticity defined as

$$
X_t=\phi_0+\phi_1X_{t-1}+\cdots+\phi_pX_{t-p}+\sigma_t\epsilon_t,\ t\in\mathbb{Z} \tag{19}
$$

where $\boldsymbol{\phi}=\left(\phi_0,\phi_1,\ldots,\phi_p\right)\in\mathbb{R}^{p+1}$ and $(\sigma_1,\sigma_2,\ldots,\sigma_n)\in\mathbb{R}^n$ are the unknown parameters and ϵ_t are iid with $Var(\epsilon_t)=1$ and unknown distribution function F. The σ_t are scaling constants which express the possible heteroscedasticity.

One usual way for modeling the time-varying volatility is to use a log linear form

$$
\sigma_t=\exp\left\{\mathbf{d}_t^T\gamma\right\}\ \text{or}\ \log\sigma_t=\mathbf{d}_t^T\gamma\ t=1,2,\ldots,n \tag{20}
$$

where $\mathbf{d}_t\in\mathbb{R}^q$, $t=1,2,\ldots,n$ is the covariate vector and $\boldsymbol{\gamma}\in\mathbb{R}^q$ is the unknown parameter vector. If $\boldsymbol{\gamma}=\mathbf{0}$, the model given in (19) will be homoscedastic. We assume that

$$\begin{cases} \sum\limits_{t=1}^{n} d_{tj},\ j=1,2,\ldots,q \\ \max_{1\le t\le n}\|d_t\| = o\left(n^{1/2}\right) \text{ as } n\to\infty \\ \lim_{n\to\infty}\mathbf{D}_n = \lim_{n\to\infty}\frac{1}{n}\sum\limits_{t=1}^{n}\mathbf{d}_t\mathbf{d}_t^T = \mathbf{D} \\ \max_{1\le t\le n}\left\{\mathbf{d}_t^T\left(\sum_{k=1}^{n}\mathbf{d}_k\mathbf{d}_k^T\right)^{-1}\mathbf{d}_t\right\} \to 0 \text{ as } n\to\infty. \end{cases} \tag{21}$$

We consider the following local heteroscedasticity

$$\boldsymbol{\gamma} = \boldsymbol{\gamma}_n = n^{-1/2}\boldsymbol{\delta},\ \ \boldsymbol{\delta} \in \mathbb{R}^q,\ \ \boldsymbol{\delta} \neq \mathbf{0},\ \ \|\boldsymbol{\delta}\| \le C < \infty. \tag{22}$$

In the following theorem, we show that under the local heteroscedasticity the averaged autoregression α-quantile is also asymptotically equivalent to the location α-quantile.

Theorem 3 *Consider the model given in equation (19) under the assumptions given in (20–22). Then the Eq. (16) is true for any fixed* $\alpha \in (0,1)$ *and*

$$\begin{aligned}&\sqrt{n}\overline{\mathbf{Y}}_{t-1}^T\left(\widehat{\boldsymbol{\phi}}_n(\alpha) - \boldsymbol{\phi} - \boldsymbol{e}_0F^{-1}(\alpha)\right) \\ &= \frac{1}{\sqrt{n}f\left(F^{-1}(\alpha)\right)}\sum_{t=1}^{n}\left(\alpha - I\left[\epsilon_t < F^{-1}(\alpha)\right]\right) + O_p\left(n^{-1/4}\right)\end{aligned} \tag{23}$$

with $\boldsymbol{e}_0 = (1,0,\ldots,0)^T \in \mathbb{R}^{p+1}$*; moreover*

$$\begin{aligned}&\sqrt{n}\left(\epsilon_{n:[n\alpha]} - F^{-1}(\alpha)\right) \\ &= \frac{1}{\sqrt{n}f\left(F^{-1}(\alpha)\right)}\sum_{t=1}^{n}\left(\alpha - I\left[\epsilon_t < F^{-1}(\alpha)\right]\right) + O_p\left(n^{-1/4}\right).\end{aligned} \tag{24}$$

In addition, under the local heteroscedasticity, both $\sqrt{n}\overline{\mathbf{Y}}_{t-1}^T\left(\widehat{\boldsymbol{\phi}}_n(\alpha) - \boldsymbol{\phi} - \mathbf{e}_0F^{-1}(\alpha)\right)$ *and* $\sqrt{n}\left(\epsilon_{n:[n\alpha]} - F^{-1}(\alpha)\right)$ *are asymptotically normally distributed* $\mathcal{N}\left(0, \frac{\alpha(1-\alpha)}{f^2\left(F^{-1}(\alpha)\right)}\right)$.

Proof Under the Assumptions (20)–(22) and assuming that the first p observations (initial values) $\left(x_1, x_2, \ldots, x_p\right)$ are completely known, the joint density of the data $X = (X_1, \ldots, X_n)^T$ is given by the following expression

$$q_{n\gamma} \propto \prod_{t=p+1}^{n}\exp\left\{\mathbf{d}_t^T\gamma\right\} f\left(x_t\exp\left\{\mathbf{d}_t^T\gamma\right\}\middle| x_{t-1},\ldots,x_{t-p},\boldsymbol{\phi},\sigma^2\right). \tag{25}$$

Under the local heteroscedasticity given in (22), the sequence of the densities $q_{n\gamma}$ is contiguous to q_{n0}. Thus the Eq. (16) is true. □

5 Simulation Study

In this section we provide a simulation study, to illustrate the performance of the AAQ(α). The simulation study is carried out using statistical software R. We considered the following scenarios.

Scenario 1: In this scenario, our autoregressive model is

$$X_t = \phi_0 + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \epsilon_t,\ t \in \mathbb{Z}, \tag{26}$$

with the true parameter values $(\phi_0, \phi_1, \phi_2)^T = (2, 0.8, -0.2)$. We generated the errors from the following distributions: (i) $\epsilon_t \sim N(0, 1)$, (ii) $\epsilon_t \sim Cauchy(0, 1)$, (iii) $\epsilon_t \sim t_3(0, 1)$, and (iv) $\epsilon_t \sim 0.15N(0, 1) + 0.85N(0, 9)$.

Scenario 2: In our second scenario, we generated the data from the following heteroscedastic autoregressive model with the same ϕ values as in Scenario 1

$$X_t = \phi_0 + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \sigma_t \epsilon_t,\ t \in \mathbb{Z} \tag{27}$$

where $\sigma_t = \exp\left\{\mathbf{d}_t^T \gamma\right\}, \gamma = n^{-1/2}\boldsymbol{\delta}$ and $\|\boldsymbol{\delta}\| \leq c < \infty$ and $\mathbf{d}_t$ is generated from standard normal distribution independently. Here $\boldsymbol{\delta}$ and c are taken as: (i) $\boldsymbol{\delta} = (0.5, 0, 0)^T$ and $c = 0.5$, (ii) $\boldsymbol{\delta} = (1, 1, 1)^T$ and $c = 2$ and (iii) $\boldsymbol{\delta} = (5, 5, 5)^T$ and $c = 10$. The errors were generated from the standard normal distribution.

We set the sample sizes as $n = 20$, 100 and 500. We repeat the simulation for 1000 times. The data from mixtures of the normal distribution is generated by using the R package "KScorrect" [37] and the autoregression α-quantiles are calculated by using the R package "quantreg" [38] (e.g., see [39]).

The AAQ(α) and the location α-quantile are computed. For the sake of comparison, the difference $\overline{B}_n\ (\alpha) - \frac{1}{n}\sum_{t=1}^{n} \mathbf{Y}_{t-1}^T\ \boldsymbol{\phi} - \epsilon_{n:[n\alpha]}$ are calculated and sorted for each replication. The empirical quantiles of these differences are plotted. Figures 1, 2, 3 and 4 show the median, 5%, 10%, 90% and 95%-quantiles in the sample differences with the error term has a normal and contaminated normal for $n = 20$, Cauchy and student-t distributions for $n = 500$, respectively.

Figures 1 and 2 are provided to compare the behavior of the difference between the AAQ(α) and location α-quantile for the normal and contaminated normal distributions in the small sample case. These figures display that this difference is not seriously affected by the contamination of data even for a small sample. Figures 3 and 4 show how much the difference between the AAQ(α) and location α-quantile is effected under the Student-t and Cauchy distribution assumptions in the large sample case. From these figures, it can be easily seen that this difference is not also severely affected by these distributional assumptions for large samples.

Mean, standard deviation and quantiles of difference $\overline{B}_n(\alpha) - \frac{1}{n}\sum_{t=1}^{n} \mathbf{Y}_{t-1}^T\ \boldsymbol{\phi} - \epsilon_{n:[n\alpha]}$ are calculated for Scenarios 1 and 2 for $\alpha = 0.05$ and 0.55. The simulation results for Scenarios 1 and 2 are given in Tables 1 and 2 and

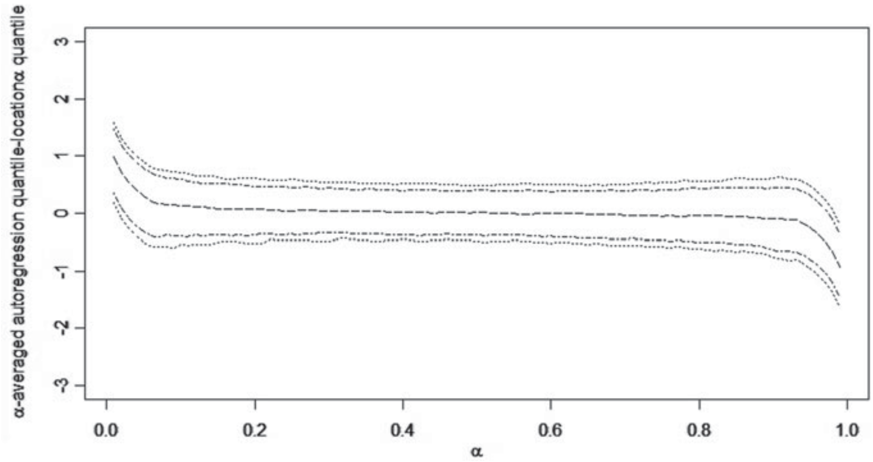


Fig. 1 5, 10, 50, 90 and 95%—quantiles in the sample of 1000 differences for normal distributed errors; $n = 20$

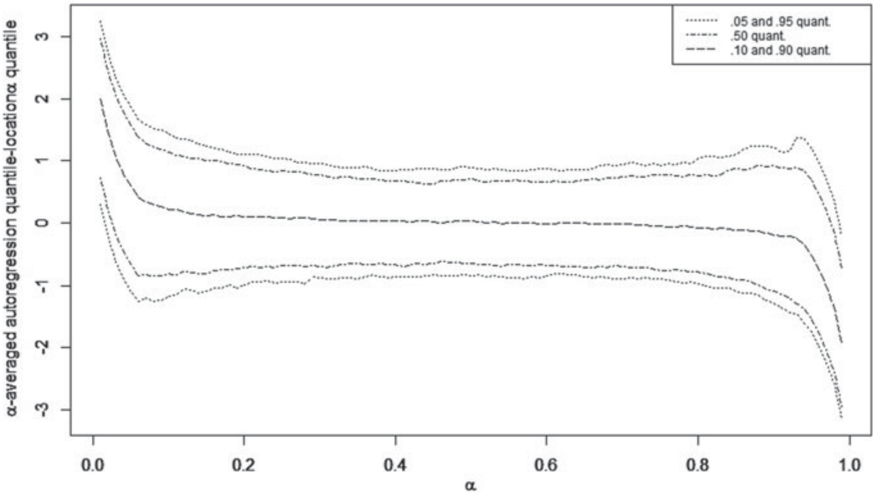


Fig. 2 5, 10, 50, 90 and 95%—quantiles in the sample of 1000 differences for contaminated normal distributed errors; $n = 20$

Tables 3 and 4, respectively. In Tables 1 and 2, N, C, t, and CN represent the normal, Cauchy, student-t and contaminated normal distributions.

The results given in Table 1 support the conclusion that, except for the Cauchy case, the AAQ(α) and location α-quantile are asymptotically equivalent. This can be easily seen from Table 1. Further, for the large case, similar results are obtained even for Cauchy distribution. Results obtained for $\alpha = 0.55$ are given in Table 2 which clearly confirm the asymptotic equivalency between the AAQ(α) and location

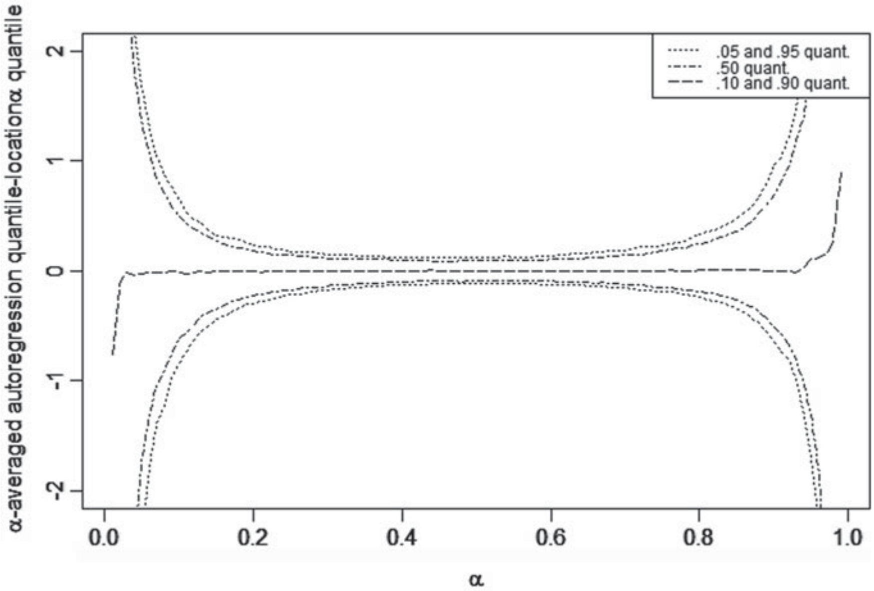


Fig. 3 5, 10, 50, 90 and 95%—quantiles in the sample of 1000 differences for Cauchy distributed errors; $n = 500$

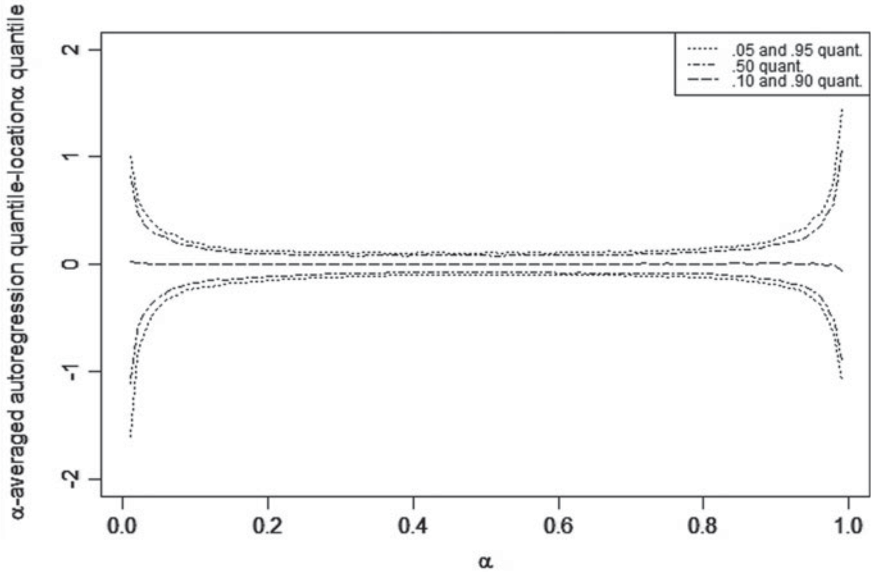


Fig. 4 5, 10, 50, 90 and 95%—quantiles in the sample of 1000 differences for student-t distributed errors; $n = 500$

Table 1 Mean, standard dev. and quantiles of differences for Scenario 1 and $\alpha = 0.05$

n	Dist.	Mean	Std. dev.	Quantiles				
				0	0.25	0.50	0.75	1
20	N	0.239	0.433	−1.449	−0.037	0.277	0.542	1.548
100	N	0.010	0.217	−0.741	−0.127	0.010	0.164	0.632
500	N	0.001	0.095	−0.309	−0.062	0.007	0.068	0.359
20	C	−16.629	89.434	−2193.43	−7.726	0.381	3.266	5.869
100	C	−1.145	4.296	−45.359	−2.265	−0.113	1.4733	4.547
500	C	−0.238	1.340	−6.461	−0.950	−0.068	0.700	2.724
20	t	−0.098	1.984	−27.365	−0.508	0.432	0.915	2.204
100	t	−0.023	0.527	−2.291	−0.325	0.058	0.341	1.284
500	t	−0.010	0.217	−0.847	−0.152	0.000	0.139	0.539
20	CN	0.512	0.891	−2.909	−0.023	0.591	1.133	3.114
100	CN	0.117	0.666	−1.881	−0.326	0.143	0.588	1.817
500	CN	0.010	0.293	−0.918	−0.174	0.020	0.219	0.832

Table 2 Mean, standard dev. and quantiles of differences for scenario 1 and $\alpha = 0.55$

n	Dist.	Mean	Std. dev.	Quantiles				
				0	0.25	0.50	0.75	1
20	N	0.001	0.294	−1.048	−0.195	−0.001	0.195	0.912
100	N	0.000	0.129	−0.412	−0.089	0.003	0.086	0.389
500	N	−0.000	0.055	−0.207	−0.038	−0.000	0.038	0.178
20	C	0.064	0.567	−1.678	−0.261	0.025	0.321	6.733
100	C	0.009	0.178	−0.941	−0.103	0.003	0.122	1.062
500	C	−0.003	0.073	−0.858	−0.047	−0.004	0.041	0.239
20	t	0.004	0.342	−0.983	−0.223	−0.001	0.221	1.079
100	t	0.006	0.134	−0.429	−0.082	0.011	0.098	0.535
500	t	0.000	0.060	−0.207	−0.039	0.002	0.042	0.190
20	CN	−0.002	0.522	−1.589	−0.348	−0.005	0.352	1.612
100	CN	0.014	0.310	−0.842	−0.200	0.005	0.213	1.120
500	CN	0.003	0.130	−0.460	−0.089	0.004	0.086	0.396

α-quantile for all the distributions and sample sizes that considered in our simulation study.

The results given in Tables 3 and 4 show that under the presence of low heteroscedasticity ($c = 0.5$), the AAQ(α) and the location α-quantile are very close to each other. Further, the results are the similar for the high heteroscedasticity case ($c = 2$ and 10).

Table 3 Mean, standard dev. and quantiles of differences for scenario 2 and $\alpha = 0.05$

c	n	Mean	Std. dev.	Quantiles				
				0	0.25	0.50	0.75	1
0.5	20	0.236	0.451	−1.863	−0.048	0.291	0.531	1.318
	100	0.020	0.208	−0.635	−0.107	0.023	0.160	0.676
	500	0.004	0.096	−0.335	−0.060	0.010	0.068	0.254
2	20	0.179	0.482	−1.717	−0.091	0.227	0.506	1.374
	100	−0.006	0.220	−0.763	−0.151	−0.008	0.142	0.588
	500	0.006	0.094	−0.306	−0.057	0.005	0.072	0.308
10	20	−0.416	1.090	−7.153	−0.907	−0.154	0.324	1.267
	100	−0.352	0.349	−2.071	−0.561	−0.320	−0.102	0.623
	500	−0.040	0.105	−0.406	−0.109	−0.036	0.032	0.281

Table 4 Mean, standard dev. and quantiles of differences for scenario 2 and $\alpha = 0.55$

c	n	Mean	Std. dev.	Quantiles				
				0	0.25	0.50	0.75	1
0.5	20	−0.021	0.283	−0.948	−0.207	−0.030	0.174	0.790
	100	0.001	0.128	−0.485	−0.081	−0.001	0.085	0.393
	500	−0.002	0.059	−0.181	−0.044	−0.001	0.039	0.151
2	20	0.001	0.289	−0.839	−0.194	−0.004	0.199	0.811
	100	0.001	0.120	−0.422	−0.080	−0.002	0.090	0.332
	500	0.001	0.057	−0.196	−0.039	−0.002	0.038	0.186
10	20	−0.040	0.208	−0.723	−0.176	−0.049	0.075	0.829
	100	−0.014	0.123	−0.407	−0.102	−0.017	0.076	0.411
	500	−0.006	0.055	−0.210	−0.044	−0.006	0.030	0.167

6 Conclusions

In this paper, we have proposed the scalar statistics AAQ(α) based on the autoregression quantiles. Some properties of the AAQ(α) have been discussed for the stationary autoregressive model and its asymptotic properties have been explored. All of these properties have shown that the AAQ(α) is closely related to the averaged regression quantile. Further, we have also considered the heteroscedastic autoregressive models and have investigated the behavior of the AAQ(α) for this model. We have provided a simulation study to illustrate the performance of the AAQ(α) and observed that the simulation study also confirms our findings.

Acknowledgements The first author thanks for the hospitality of Charles University in Prague during her six months visit as a Visiting Doctoral Scholar. She also thanks to Ankara University for allowing her to continue her studies at Charles University. The second author gratefully acknowledges the support of the Grant GAČR 18-01137S.

References

1. Koenker, R., Bassett, G.: Regression quantiles. Econometrica **46**, 33–49 (1978)
2. Koenker, R.: Galton, Edgeworth, Frisch and prospects for quantile regression in econometrics. J. Econ. **95**, 347–374 (2000)
3. Koenker, R., Hallock, K.: Quantile regression. J. Econ. Perspect. **15**, 143–156 (2001)
4. Koenker, R., Xiao, Z.: Inference on the quantile regression processes. Econometrica **70**, 1583–1612 (2002)
5. Koenker, R.: Quantile Regression, Econometric Society Monographs (No. 38). Cambridge University Press, New York (2005)
6. Knight, K.: Limit theory for autoregressive-parameter estimates in an infinite-variance random walk. Can. J. Stat. **17**, 261–278 (1989)
7. Weiss, A.: Estimating nonlinear dynamic models using least absolute error estimation. Econ. Theory **7**, 46–68 (1991)
8. Koul, H., Mukherjee, K.: Asymptotics of R-, MD- and LAD-estimators in linear regression models with long range dependent errors. Probab. Theory Rel. Fields **95**, 535–553 (1993)
9. Koul, H., Saleh, A.K.: Autoregression quantiles and related rank-scores processes. Ann. Stat. **23**(2), 670–689 (1995)
10. Knight, K.: Some limit theory for L1-estimators in autoregressive models under general conditions. In: Lecture Notes-Monograph Series. In: Dodge, Y. (ed.), L1-Statistical Procedures and Related Topics. California, vol. 31, pp. 315–328 (1997)
11. Knight, K.: Asymptotics for L1 regression estimates under general conditions. Ann. Stat. **26**, 755–770 (1998)
12. Jurečková, J., Hallin, M.: Optimal tests for autoregressive models based on autoregression rank scores. Ann. Stat. **27**, 1385–1414 (1999)
13. Koenker, R., Xiao, Z.: Unit root quantile regression inference. JASA **99**(467), 775–787 (2004)
14. Koenker, R., Xiao, Z.: Quantile autoregression. JASA **101**(475), 980–1006 (2006)
15. Knight, K.: Comment on: quantile autoregression. JASA **101**(475), 991–1001 (2006)
16. Li, G., Li, Y., Tsai, C.L.: Quantile correlations and quantile autoregressive modeling. JASA **110**(509), 246–261 (2015)
17. Granger, C.W.G., White, H., Kamstra, M.: Interval forecasting: an analysis based on ARCH-quantile estimators. J. Econ. **40**, 87–96 (1989)
18. Koenker, R., Zhao, Q.: Conditional quantile estimation and inference for ARCH models. Econ. Theory **12**, 793–813 (1996)
19. Xiao, Z., Koenker, R.: Conditional quantile estimation and inference for GARCH models. JASA **104**(488), 1696–1712 (2009)
20. Noh, J., Lee, S.: Quantile regression for location-scale time series models with conditional heteroscedasticity. Scandinavian J. Stat. **43**(3), 700–720 (2016)
21. Engle, R.: Autorregressive conditional heteroskedasticity with estimates of United Kingdom inflation. Econometrica **50**, 987–1008 (1982)
22. Bollerslev, T.: Generalized autorregressive conditional heteroskedasticity. J. Econ. **31**, 307–327 (1986)
23. Xu, K.L., Phillips, P.C.: Adaptive estimation of autoregressive models with time-varying variances. J. Econ. **142**(1), 265–280 (2008)
24. Gutenbrunner, C.: Tests for heteroscedasticity based on regression quantiles and regression rank scores. In: Mandl, P., Hušková, M. (eds.), Asymptotic Statistics: Proceedings of the 5th Prague Symposium. Physica-Verlag, Heidelberg (1994)
25. Gutenbrunner, C., Jurečková, J., Koenker, R.: Regression rank test for heteroscedasticity (1995) (unpublished)
26. Jurečková, J., Navratil, R.: Rank tests under uncertainty: regression and local heteroscedasticity. Syn. Soft Comput. Stat. Intel. Data Anal. Adv. Intel. Syst. Comput. **190**, 255–261 (2012)
27. Brockwell, P.J., Davis, R.A., Fienberg, S.E.: Time Series: Theory and Methods: Theory and Methods. Springer Science & Business Media (1991)

28. Hallin, M., Jurečková, J.: Optimal tests for autoregressive models based on autoregression rank scores. Annal. Stat. **27**(4), 1385–1414 (1999)
29. Hallin, M., Jurečková, J., Koul, H.L.: Serial Autoregression and regression rank scores statistics. In: Advances in Statistical Modeling and Inference: Essays in Honor of Kjell A. Doksum, pp. 335–362 (2007)
30. Gutenbrunner, C., Jurečková, J.: Regression rank scores and regression quantiles. Ann. Stat. **20**, 305–330 (1992)
31. Bassett Jr., G.W.: A property of the observations fit by the extreme regression quantiles. Comp. Stat. Data Anal. **6**, 353–359 (1988)
32. Jurečková, J., Picek, J.: Averaged regression quantiles. In: Lahiri, S.N., et al. (eds.), Contemporary Developments in Statistical Theory Springer Proceedings in Mathematics & Statistics, **68(12)**, 203–216 (2014)
33. Picek, J., Schindler, M.: The contribution of the averaged regression quantiles for testing max-domains of attractions. Business Administration, **91** (2015)
34. Jurečková, J.: Averaged extreme regression quantile. Extremes **19**(1), 41–49 (2016)
35. Jurečková, J.: Regression quantile and averaged regression quantile processes. In: Workshop on Analytical Methods in Statistics, pp. 53–62. Springer, Cham (2015)
36. Jurečková, J., Schindler, M., Picek, J.: Empirical regression quantile process with possible application to risk analysis (2017). arXiv:1710.06638
37. Novack-Gottshall, P., Wang, S.C., Novack-Gottshall, M.P.: R Package ‘KScorrect’ (2016)
38. Koenker, R., Portnoy, S., Ng, P.T., Zeileis, A., Grosjean, P., Ripley, B.D.: R Package ‘quantreg’ (2019)
39. Jurečková, J., Picek, J., Schindler, M.: Robust statistical methods with R. Chapman and Hall/CRC (2019)

Regression Neural Networks with a Highly Robust Loss Function

Jan Kalina and Petra Vidnerová

Abstract Artificial neural networks represent an important class of methods for fitting nonlinear regression to data with an unknown regression function. However, usual ways of training of the most common types of neural networks applied to nonlinear regression tasks suffer from the presence of outlying measurements (outliers) in the data. So far, only a few robust alternatives for training common forms of neural networks have been proposed. In this work, we robustify two common types of neural networks by considering robust versions of their loss functions, which have turned out to be successful in linear regression. Particularly, we extend the idea of using the loss of the least trimmed squares estimator to radial basis function networks. We also propose multilayer perceptrons and radial basis function networks based on the loss of the least weighted squares estimator. The performance of these novel methods is compared with that of standard neural networks on 4 datasets. The results bring arguments in favor of the novel robust approach based on the least weighted squares estimator with trimmed linear weights in terms of yielding the smallest robust prediction error in a variety of situations. Robust neural networks are even able to outperform the prediction ability of support vector regression.

Keywords Nonlinear regression · Neural networks · Robustness

1 Introduction

Nonlinear regression modeling, i.e. estimating (smoothing, fitting) a continuous response variable based on a set of regressors (features, independent variables) plays a crucial role in the analysis of real data in a tremendous variety of applications. An important task of regression modeling is also to predict a future development of the

J. Kalina (✉) · P. Vidnerová
The Czech Academy of Sciences, Institute of Computer Science,
Pod Vodárenskou věží 2, 182 07 Praha 8, Czech Republic
e-mail: kalina@cs.cas.cz

P. Vidnerová
e-mail: petra@cs.cas.cz

M. Maciak et al. (eds.), *Analytical Methods in Statistics*, Springer Proceedings
in Mathematics & Statistics 329, https://doi.org/10.1007/978-3-030-48814-7_2

response [6]. In practical applications, the nonlinear regression function is not known and is not assumed to be of any specific form. Recently, there is an increasing trend in applying machine learning methods to nonlinear regression modeling. In this paper, multilayer perceptrons (MLPs) and radial basis function (RBF) networks, i.e. two very important classes of feedforward artificial neural networks [10], are considered for the nonlinear regression task.

Real data across various disciplines, e.g. in numerous regression tasks of biomedicine, economics, engineering etc., are typically contaminated by the presence of outlying measurements (outliers). In some applications (e.g. in measurements of molecular genetic and metabolomic biomarkers [14]), outliers appear unavoidably, because severe measurement errors are immanent to the measurement technology. So far, most available applications of MLPs and RBF networks to regression tasks have not paid sufficient attention to the presence and influence of outliers; both these networks however implicitly assume the observed data not to be contaminated by outliers [2, 26]. Therefore, it is highly desirable to consider alternative robust approaches to training of MLPs and RBF networks. One direction of the robustification is based on an intrinsically performed detection of outliers [1]. Another direction for a possible robustification is inspired by the very rich experience of robust statistics with data contamination by outliers or anomalies (see [12]); this approach represents the interest of the current paper.

While there are some robust approaches to training neural networks available, they are mostly tailor-made the classification task; see ([17], p. 54) for discussion. Let us mention at least a few available robust approaches for the regression task. Compositions of sigmoidal activation functions were considered to robustify the performance for a rather specific task in [18] to estimate a response which is almost constant over relatively large intervals. If subtractive clustering (SC) is used for an automatic recommendation of the center vectors, a robustified loss function may be subsequently used [26]; still, the popular SC approach remains vulnerable to outliers and consecutive steps of the training cannot improve this. A recent approach to outlier detection for regression RBF networks was developed in [17], which is denoted as generalized edited nearest neighbor (ENN) algorithm; this was also combined with robust versions of the activation function. Robust loss functions based on least trimmed squares or least trimmed absolute values estimators were investigated in [24, 25], where they outperformed standard training approaches on contaminated data. We do not agree with the formulas for partial derivatives of the loss function published in [24], but this may not influence the results presented there, as practical computations typically exploit numerical approximations of derivatives (not relying on theoretical expressions). Nevertheless, even the extensive numerical computations in [25] do not compare robust neural networks with the (sophisticated and powerful) support vector regression.

The idea to apply a robust loss function in neural networks will be extended in the current paper by means of the least weighted squares estimator, which represents a natural generalization of the least trimmed squares and turns out to be a perspective and (possibly) highly robust tool for estimating parameters in linear regression. Section 2 recalls the least trimmed squares and least weighted squares estimators of parameters in linear regression and in the location model. Section 3 uses these estimators to propose novel robust versions of MLPs and RBF networks. Numerical examples presented in Sect. 4 illustrate the performance of the novel robust neural networks. Finally, Sect. 5 concludes the paper.

2 Highly Robust Estimation in Linear Models

This section recalls two (possibly highly) robust implicitly weighted estimators of parameters of the linear regression model (including the location model as a special case), namely the least trimmed squares and least weighted squares estimators. Highly robust estimators are defined as those, which attain a high value of the breakdown point; this measure of robustness of a statistical estimator of an unknown parameter represents a fundamental concept of robust statistics [12]. Formally, the finite-sample breakdown point evaluates the minimal fraction of data that can drive an estimator beyond all bounds when set to arbitrary values.

The standard linear regression model has the form

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + e_i, \quad i = 1, \ldots, n, \tag{1}$$

with a continuous response $Y_1, \ldots, Y_n$ explained by the total number of p regressors, and independent and identically distributed (not necessarily Gaussian) random errors $e_1, \ldots, e_n$.

The least trimmed squares (LTS) estimator [22, 23] of β represents a popular robust regression estimator with a high breakdown point. Consistency of the LTS and other properties were derived in [27]. The user must select the value of a trimming constant h $(n/2 \leq h < n)$. We will denote residuals corresponding to a particular $b = (b_0, \ldots, b_p)^T \in \mathbb{R}^{p+1}$ as

$$u_i(b) = Y_i - b_0 - b_1 X_{i1} - \cdots - b_p X_{ip} \tag{2}$$

and order statistics of their squares as

$$u^2_{(1)}(b) \leq \cdots \leq u^2_{(n)}(b). \tag{3}$$

The LTS estimator, formally obtained as

$$\operatorname*{arg\,min}_{b \in \mathbb{R}^{p+1}} \frac{1}{n} \sum_{i=1}^{h} u^2_{(i)}(b), \tag{4}$$

may attain a high robustness but cannot achieve a high efficiency. We may consider the LTS as an implicitly weighted estimator, namely as a special case of the least weighted squares with weights equal only to 0 or 1.

The least weighted squares (LWS) estimator (see e.g. [28]) for the model (1) represents a flexible natural extension of the LTS. The LWS estimator motivated by the idea to down-weight potential outliers based on ranks of residuals however remains much less known compared to the LTS. The LWS estimator may achieve a high breakdown point (with properly selected weights) and is robust to heteroscedasticity [28]. Its primary attention is focused on estimating β and not on outlier detection. The LWS estimator with given magnitudes of weights $w_1, \ldots, w_n$ is defined as

$$\mathbf{b}^{LWS} = (b_0^{LWS}, \ldots, b_p^{LWS})^T = \operatorname*{arg\,min}_{b \in \mathbb{R}^{p+1}} \sum_{k=1}^{n} w_k u_{(k)}^2(b). \tag{5}$$

The efficiency of the LWS is able to exceed the low efficiency of the LTS; if data-dependent adaptive weights of [4] are used, the estimator asymptotically attains the full efficiency of the least squares. The LWS estimator was successful in a variety of recent applications including denoising gene expression measurements acquired by the microarray technology [14] or image analysis based on landmarks measured within facial images [13]. There has been a good experience with implicit weighting also for multivariate robust estimation; the multivariate analogy of the LWS is the minimum weighted covariance determinant (MWCD) estimator proposed in [21].

The location model represent an important special case of (1) in the form

$$Y_i = \mu + e_i \quad \text{for} \quad i = 1, \ldots, n, \tag{6}$$

where $\mu \in \mathbb{R}$ represents a parameter of location (shift). In (6), the LWS estimator inherits the appealing properties of the LWS from (1). The performance of the LWS on real data in (6) was revealed as successful e.g. in the image analysis applications of [13], where the LWS estimator in (6) was also proven to correspond to the estimator with the smallest weighted variance. This allows a very efficient computation of the LWS in (6).

3 Robust Neural Networks with Implicitly Weighted Loss Functions

We consider the regression model

$$Y_i = f(X_i) + e_i, \quad i = 1, \ldots, n, \tag{7}$$

with an unknown nonlinear function f, where $Y_1, \ldots, Y_n$ are values of the response and $X_i \in \mathbb{R}^p$ (with $p \geq 1$) is a vector of regressors corresponding to the i-th observation. This is a nonlinear regression setup with a univariate continuous response $Y_1, \ldots, Y_n$, which is explained by means of p regressors. A novel robust tool for neural networks is proposed in this section, namely an MLP or an RBF network based on the loss function of the LWS estimator.

MLPs, which represent a very popular type of artificial neural networks, contain an input layer, one or more hidden layers with a fixed number of neurons, and an output layer. As we use the most standard form of multilayer perceptrons, we will not present their detailed model, as it can be found in numerous monographs (see e.g. [7, 9]). For a particular multilayer perceptron (with a selected architecture), let the fitted value of the response for the i-th measurement (i.e. estimate of Y_i) be denoted by $\hat{Y}_i$ for each $i = 1, \dots, n$.

Let us start by describing the training of a standard MLP in a symbolic (general but very simplified) way in Algorithm 1. There, we denote the whole (say m-dimensional) vector of all parameters of a given MLP with a specified architecture as $\theta \in \mathbb{R}^m$. Denoting the estimated version of f obtained by the MLP as $\hat{f}$, we may denote the vector of fitted values of Y by $\hat{Y} = \hat{f}(\hat{\theta})$ and the vector of residuals, which depend on $\hat{f}$, as $u = Y - \hat{Y}$. Concerning the stopping rule in Algorithm 1, our computations use a default version implemented in [3]. Algorithm 1 is formulated in such a way that it remains valid also for a robust version of an MLP, as it considers a general loss function.

Algorithm 1 MLP in the nonlinear regression model (1) with a selected (standard or robust) loss function ℓ

```
Input: $X_1, \dots, X_n$, where $X_i \in \mathbb{R}^p$ for each $i = 1, \dots, n$
Input: $Y_1, \dots, Y_n$, where $Y_i \in \mathbb{R}$ for each $i = 1, \dots, n$
Input: A chosen loss function $\ell$
Output: A fitted MLP based on minimizing a given loss $\ell$
  Choose $\hat{\theta}_0 \in \mathbb{R}^m$ as an initial estimate of $\theta$
  $i := 0$
  repeat
    $u^i = (u_1^i, \dots, u_n^i) := Y - f(\hat{\theta}_i)$
    $i := i + 1$
    $\hat{\theta}_i := \arg\min \ell(u_1^{i-1}, \dots, u_n^{i-1})$ (where the optimization over estimates of $\theta$ is solved by a
    stochastic gradient method)
  until a certain stopping rule is fulfilled
```

The most common way of training MLPs minimizes the sum of prediction errors in the form

$$\ell = \ell(u_1, \dots, u_n) := \min \sum_{i=1}^{n} u_i^2. \tag{8}$$

It corresponds to the least squares estimation in a location model. It is now natural to replace this quadratic loss function by one of available robust alternatives (again for the location model). We consider a method of [24] denoted here as LTS-MLP; for a fixed h, it is defined by replacing (8) in the form

$$\ell := \sum_{i=1}^{h} u_{(i)}^2. \tag{9}$$

We define a new version of MLP dentoted as LWS-MLP by choosing ℓ in the form

$$\ell := \sum_{i=1}^{n} w_i u_{(i)}^2 \tag{10}$$

for selected magnitudes of weights $w_1, \ldots, w_n$. We always consider the natural standardization to $\sum_{i=1}^{n} w_i = 1$. We consider three particular choices, namely the LWSa-MLP with linear weights

$$w_i = \frac{2(n+1-i)}{n(n+1)}, \quad i = 1, \ldots, n, \tag{11}$$

LWSb-MLP with trimmed linear weights

$$w_i = \frac{h-i+1}{h} \mathbb{1}[i \leq h], \quad i = 1, \ldots, n, \tag{12}$$

where we consider $h = \lfloor 3n/4 \rfloor$ and $\lceil x \rceil = \min\{n \in \mathbb{N};\ n \geq x\}$, and finally LWSc-MLP with weights generated by the (strictly decreasing) logistic function

$$w_i = \left(1 + \exp\left\{\frac{i-n-1}{n}\right\}\right)^{-1}, \quad i = 1, \ldots, n. \tag{13}$$

While LTS-MLP loss detects outliers and trims them away, LWS-MLP estimator does not do this but intrinsically arranges observations according to outlyingness.

Another alternative version denoted here as LTA-MLP was defined in [24], where a robust loss function corresponding to the least trimmed absolute value (LTA) estimator was used. LTA-MLP is defined for a fixed h ($n/2 \leq h < n$) by means of

$$\ell := \sum_{i=1}^{h} |u_{(i)}| \tag{14}$$

and according to [24] yields very similar results to those of LTS-MLP. We can say that the LTA estimator is practically unknown in the community of robust statistics; at the same time, it is not sufficiently discussed in the majority of monographs on robust estimation [12]. It is worth noting that, although we are not aware of systematic numerical comparison of the LTA estimator with other robust estimates in linear regression, it has been claimed that the performance of the LTA is very similar to that of the LTS in linear regression. Possible improvements of the LTA compared to the LTS are known not to be more than only marginal (see p. 429 of [29]). Still, the

LWS estimator seems to be much more promising in terms of both robustness and efficiency, as repeatedly discussed [5, 28].

Radial basis function (RBF) networks represent another important class of neural networks. They contain an input layer with p inputs, a single hidden layer with N RBF units (neurons), and a linear output layer. The user chooses N together with a radially symmetric function denoted here as ρ. The RBF network is based also on minimizing (8); using the Gaussian density as ρ, the residuals can be expressed as

$$u_i = Y_i - \sum_{j=1}^{N} a_j \rho(||X_i - c_j||), \quad i = 1, \ldots, n, \tag{15}$$

with parameters $c_1, \ldots, c_N \in \mathbb{R}^p$ and $a_1, \ldots, a_N \in \mathbb{R}$, and possibly with other parameters corresponding to ρ. We refer to [10, 16] for a detailed description of RBF networks. RBF networks can be expressed in an analogous way as MLPs in Algorithm 1 by means of minimizing the sum of squared residuals.

Robust versions of RBF networks, which will be denoted here as LTS-RBF, LWSa-RBF, LWSb-RBF, or LWSc-RBF networks, will be defined by means of the loss functions above. In other words, the are obtained by replacing the quadratic loss in (15) by the loss functions of the LTS or LWS estimators.

We implemented all the robust neural networks in Keras [3]. The implementation exploits a back-propagation algorithm, namely a stochastic gradient descent method, i.e. the same approach as in [24, 25], for optimization of all parameters for both standard and robust MLPs as well as RBF networks. As our experiments have demonstrated, also the loss function of LWS-MLP and LWS-RBF networks is in practice smooth enough for our gradient-based approach.

4 Numerical Experiments

The aim of the computations over 1 simulated and 3 real datasets is to illustrate the performance of the novel robust neural networks and compare it with other nonlinear regression tools.

4.1 Data Description

(A) The so-called Eckerle4 dataset publicly available in the package NISTnls of R software [20] has $p = 1$ regressor and $n = 35$ observations, including one apparent outlier. In Fig. 1, this real dataset is presented together with fitted trend, estimated by a standard MLP as well as LTS-MLP.

(B) A simulated dataset obtained by means of a sine function with a (rather artificial) contamination by a linearly decreasing trend with $p = 1$ and $n = 101$.

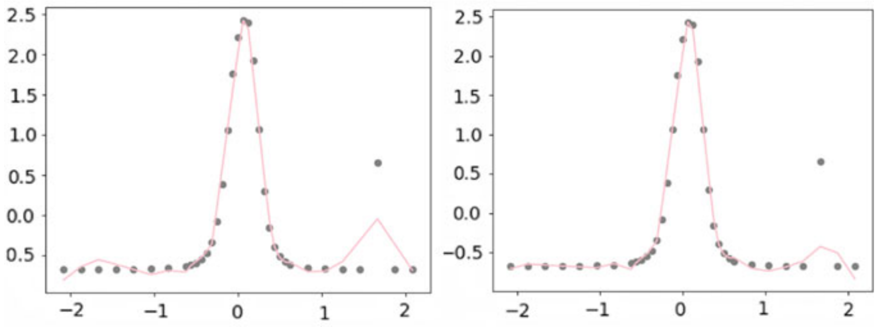


Fig. 1 Dataset Eckerle4. Horizontal axis: the regressor. Vertical axis: the response. The curve corresponds to the standard MLP (left) and LTS-MLP with $h = \lfloor 3n/4 \rfloor$ (right)

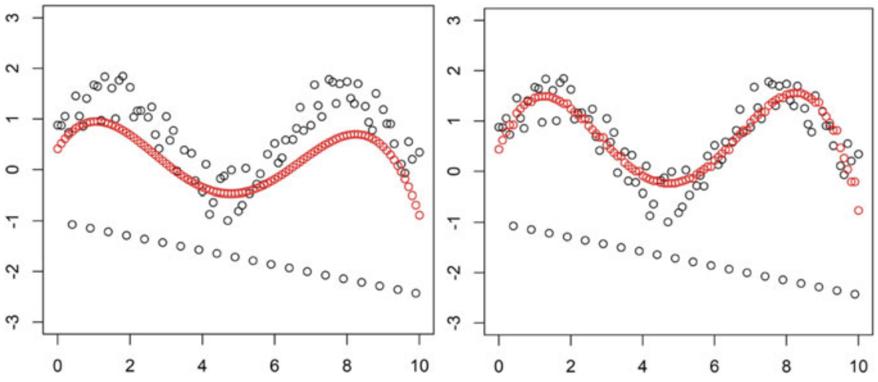


Fig. 2 The simulated dataset. Horizontal axis: the regressor. Vertical axis: the response. The curve corresponds to the standard RBF network (left) and LTS-RBF network with $h = \lfloor 3n/4 \rfloor$ (right)

The dataset is presented in Fig. 2, together with estimated trend, obtained by a standard RBF as well as LTS-RBF network.

(C) The Auto MPG dataset [8] with $p = 4$ continuous regressors and $n = 392$ observations after omitting all missing values (i.e. observations with index 33, 127, 331, 337, 355, and 375) from the original dataset. The consumption of each car in miles per gallon (MPG) is considered here as a response explained by engine displacement, horsepower, weight, and acceleration.

(D) The Boston Housing dataset [8] with $p = 11$ continuous regressors (omitting features 4, 7, and 9 from the original dataset) and $n = 506$ observations. The per capita crime rate by town (i.e. in each individual location) is considered as the response variable here.

4.2 Methods

The following methods will be used in the computations. For the description of standard machine learning methods, the reader may refer to monographs [9, 10].

- RBF network. The number N of RBF units used in particular examples is specified in Table 1.
- LTS-RBF network with the same architecture as the plain RBF network and $h = \lfloor 3n/4 \rfloor$.
- LWS-RBF (i.e. LWAa-RBF, LWSb-RBF, LWSc-RBF) networks with the same architecture as the plain RBF network.
- MLP with 1 or 2 hidden layers as specified in Table 1 for particular examples, together with the number of neurons in these layers. In every example, a sigmoid activation function is considered in every hidden layer. A linear output layer is always used.
- LTS-MLP with the same architecture as the plain MLP and $h = \lfloor 3n/4 \rfloor$.
- LWS-MLP with the same architecture as the plain ML.

Three different measures of prediction errors are evaluated for each situation within a ten-fold cross validation study, performed in a standard way. Because the standard MSE suffers from the presence of outliers in the data, we also consider the trimmed MSE (TMSE) and weighted MSE (WMSE) defined formally as

$$\mathsf{MSE} = \frac{1}{n}\sum_{i=1}^{n} r_i^2, \quad \mathsf{TMSE}(\alpha) = \frac{1}{h}\sum_{i=1}^{h} r_{(i)}^2, \quad \mathsf{WMSE} = \sum_{i=1}^{n} w_i r_{(i)}^2, \tag{16}$$

where $r_i = Y_i - \hat{Y}_i$ are prediction errors and $\hat{Y}_i$ denotes the fitted value of the i-th observation for $i = 1, \dots, n$. For TMSE, we choose h as the is integer part of $3n/4$, and squared prediction errors are arranged as $r_{(1)}^2 \leq \cdots \leq r_{(n)}^2$. WMSE requires to use some fixed non-increasing magnitudes of weights and we use here trimmed linear weights (12) with $\sum_{i=1}^{n} w_i = 1$.

4.3 Results

The results for standard as well as robust neural networks with the selected architectures and parameters are presented in Table 1. The number N of RBF units for all versions of RBF networks was selected as the most suitable one for plain RBF networks. The number of neurons in the hidden layers for all versions of MLPs was selected as the most suitable for plain MLPs.

The dataset Eckerle4, the simplest from the 4 datasets under considerations, is very simple with very much variability (except for an apparent outlier). Results over the two datasets with $p = 1$ are illustrated in Figs. 1 and 2. In these datasets with

Table 1 Results of numerical experiments. Three error measures (MSE, TMSE and WMSE) defined in (16) evaluated for various nonlinear regression methods for 4 datasets. The architectures (number of RBF units and neurons in hidden layers) are specified here for various versions of RBF networks and MLPs, respectively

Neural network	MSE	TMSE	WMSE	MSE	TMSE	WMSE
	Dataset Eckerle4			Simulated dataset (Fig. 2)		
	$p = 1, n = 35$			$p = 1, n = 101$		
	3 RBF units			10 RBF units		
RBF	0.03	<0.01	<0.01	0.18	0.055	0.050
LTS-RBF	0.04	<0.01	<0.01	0.29	0.035	0.033
LWSa-RBF	0.04	<0.01	<0.01	0.28	0.037	0.033
LWSb-RBF	0.04	<0.01	<0.01	0.31	0.032	0.029
LWSc-RBF	0.04	<0.01	<0.01	0.32	0.038	0.031
	1 hidden layer			1 hidden layer		
	with 4 neurons			with 8 neurons		
MLP	0.04	<0.01	<0.01	0.24	0.067	0.061
LTS-MLP	0.05	<0.01	<0.01	0.30	0.038	0.034
LWSa-MLP	0.05	<0.01	<0.01	0.32	0.038	0.033
LWSb-MLP	0.05	<0.01	<0.01	0.33	0.035	0.030
LWSc-MLP	0.05	<0.01	<0.01	0.35	0.037	0.032
	Auto MPG dataset			Boston housing dataset		
	$p = 4, n = 392$			$p = 11, n = 506$		
	40 RBF units			50 RBF units		
RBF	46.9	17.2	19.3	52.7	4.4	5.6
LTS-RBF	52.7	12.9	14.1	60.3	4.1	5.2
LWSa-RBF	54.1	14.4	13.8	62.1	4.1	5.0
LWSb-RBF	50.6	11.8	12.5	61.6	4.0	4.7
LWSc-RBF	53.7	14.0	13.4	62.4	4.3	4.9
	2 hidden layers			2 hidden layers		
	with 16 and 8 neurons			with 16 and 8 neurons		
MLP	60.8	28.9	31.0	57.9	5.3	6.3
LTS-MLP	69.4	14.3	17.6	67.2	4.3	5.9
LWSa-MLP	70.3	14.5	16.2	70.8	4.2	5.7
LWSb-MLP	71.6	13.9	15.8	68.8	4.1	5.5
LWSc-MLP	72.5	14.3	16.7	70.6	4.2	5.7

$p = 1$, TMSE is able to ignore the true outliers for robust but also for plain neural networks. This is because the regression task is not so difficult for these datasets and the outliers are exactly those points, which have large absolute values of the residuals. The situation becomes much more complex for the other datasets.

In all examples, robust versions of neural networks approaches are able to yield smaller values of robust prediction errors (TMSE and WMSE); this is true in spite of the fact that the architecture of the neural networks was optimized for the plain networks. On the other hand, standard versions of neural networks are superior in terms of conventional MSE. This does not mean that the robust methods are less suitable, because the MSE itself is vulnerable to the presence of outliers. Thus, only robust versions of MSE should be considered for data contaminated with outliers.

Comparing RBF networks with MLPs, RBF networks turn out to yield smaller values of the prediction errors for all 4 datasets. It is especially interesting for the two datasets with $p > 1$ from real applications that the superiority of robust neural networks compared to standard (non-robust) ones is revealed. Basically we can say that using (any) robust neural network brings benefits, while the results of LWSb-RBF networks are not overcome by any other method in the 4 datasets.

5 Conclusions

Robust alternatives to training neural networks are highly desirable because of the vulnerability of common types of neural networks to the presence of outliers in the data. We use highly robust estimators corresponding to the LTS and LWS estimators to formulate robust loss function of MLPs and RBF networks. Thus, we extend the idea of [24], who used the loss function of the LTS (only) within MLPs. To the best of our knowledge, our approach is the first application of the LWS estimator within neural networks. The novel methods assign implicit weights to individual observations and correspond to their outlyingness, which offers a possible interpretation of individual observations and their influence to the resulting estimated trend. Robust fitting of neural networks based on the loss function of the least weighted squares estimator is able to minimize robust measures of prediction error. The methods denoted as LWSb-MLPs and LWSb-RBF networks, i.e. those with trimmed linear weights, turn out to yield better results in terms of prediction accuracy compared to other choices of weights for the LWS loss.

The superior results of the neural networks based on the LWS estimator are in correspondence with recent findings of [15]. There, the LWS turned out to outperform other estimators in linear regression, including S-estimators and mainly MM-estimators, where the latter allow to tune paramters so that a high robustness and a high efficiency are reached simultaneously.

The robust neural networks considered in the paper appear suitable for all the 4 datasets considered in this paper and thus are recommendable for real datasets, where robustness to data contamination by outliers is desirable. All datasets analyzed here do contain outliers. If a new dataset should be analyzed, which does not seem to contain apparent outliers, the strategy common in linear regression may be adopted for neural networks as well; namely, the novel robust neural networks may serve as a diagnostic tool. In such a situation, the user may check if the results of a standard neural network are similar with results of robust ones. In case of remarkable discrepancies, the robust approach may be more suitable. As a limitation, however, it is necessary to state that the robust neural networks of this paper (just like any robust statistical method [12]) may not be suitable for certain datasets, e.g. when we are interested in every individual observation and ignoring specific observations (or their clusters) is not desirable.

Several possible directions recommendable for future research include adapting robust neural networks for heteroscedastic data, proposing an adaptive selection of h for the LTS-based loss function, considering robust and regularized neural networks, or proposing adaptive (data-dependent) selection of weights for the LWS-based loss. In addition, it would be desirable to perform a systematic comparison of robust approaches to training neural networks over a larger number of datasets, accompanied by a detailed statistical analysis of the data and by a thorough interpretation of the results on the level of individual observations.

Acknowledgements The work is supported by the projects GA19-05704S and GA18-23827S of the Czech Science Foundation. The authors are grateful to Jan Tichavský and Jiří Tumpach for technical help.

References

1. Alnafessah, A., Casale, G.: Artificial neural networks based techniques for anomaly detection in Apache Spark. Cluster Computing (2020) (online first)
2. Borş, A.G., Pitas, I.: Robust RBF networks. In: Howlett, R.J., Jain, L.C., Kacprzyk, J. (eds.), Radial basis function networks 1. Recent developments in theory and applications, pp. 123–133. Physica Verlag Rudolf Liebing KG, Vienna (2001)
3. Chollet, F.: Keras. Github repository (2015). https://github.com/fchollet/keras
4. Čížek, P.: Semiparametrically weighted robust estimation of regression models. Comput. Stat. Data Anal. **55**, 774–788 (2011)
5. Čížek, P.: Reweighted least trimmed squares: an alternative to one-step estimators. Test **22**, 514–533 (2013)
6. Davies, L.: Data analysis and approximate models. In: Nonparametric Regression and Image Analysis. CRC Press, Boca Raton, Model Choice, Location-scale, Analysis of Variance (2014)
7. Du, K.L., Swamy, M.N.S.: Neural Networks and Statistical Learning. Springer, London (2014)
8. Frank, A., Asuncion, A.: UCI Machine Learning Repository. University of California, Irvine (2010). http://archive.ics.uci.edu/ml
9. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning, 2nd edn. Springer, New York (2009)
10. Haykin, S.O.: Neural Networks and Learning Machines: A Comprehensive Foundation, 3rd edn. Prentice Hall, Upper Saddle River (2009)
11. Hubert, M., Rousseeuw, P.J., van Aelst, S.: High-breakdown robust multivariate methods. Stat. Sci. **23**, 92–119 (2008)
12. Jurečková, J., Picek, J., Schindler, M.: Robust Statistical Methods with R, 2nd edn. Chapman & Hall/CRC, Boca Raton (2019)

13. Kalina, J.: Implicitly weighted methods in robust image analysis. J. Math. Imag. Vis. **44**, 449–462 (2012)
14. Kalina, J.: A robust supervised variable selection for noisy high-dimensional data. BioMed. Res. Int. Article 320385 (2015)
15. Kalina, J., Tichavský, J.: On robust estimation of error variance in (highly) robust regression. Meas. Sci. Rev. **20**, 1–9 (2020)
16. Kalina, J., Vidnerová, P.: Robust training of radial basis function neural networks. In: Proceedings 18th International Conference ICAISC 2019, pp. 113–124 (2019)
17. Kordos, M., Rusiecki, A.: Reducing noise impact on MLP training—techniques and algorithms to provide noise-robustness in MLP network training. Soft Comput. **20**, 46–65 (2016)
18. Lee, C.C., Chung, P.C., Tsai, J.R., Chang, C.I.: Robust radial basis function neural networks. IEEE Trans. Syst. Man Cybern. B **29**, 674–685 (1999)
19. Mašíček, L.: Optimality of the least weighted squares estimator. Kybernetika **40**, 715–734 (2004)
20. R Core Team: R: A language and environment for statistical computing. In: R Foundation for Statistical Computing, Vienna (2019). https://www.R-project.org/
21. Roelant, E., van Aelst, S., Willems, G.: The minimum weighted covariance determinant estimator. Metrika **70**, 177–204 (2009)
22. Rousseeuw, P.J., Hubert, M.: Robust statistics for outlier detection. WIREs Data Mining Knowl. Discov. **1**, 73–79 (2011)
23. Rousseeuw, P.J., Van Driessen, K.: Computing LTS regression for large data sets. Data Mining Knowl. Discov. **12**, 29–45 (2006)
24. Rusiecki, A.: Robust learning algorithm based on LTA estimator. Neurocomputing **120**, 624–632 (2013)
25. Rusiecki, A., Kordos, M., Kamiński, T., Greń, K.: Training neural networks on noisy data. Lect. Notes Artif. Intel. **8467**, 131–142 (2014)
26. Su, M.J., Deng, W.: A fast robust learning algorithm for RBF network against outliers. Lect. Notes Comput. Sci. **4113**, 280–285 (2006)
27. Víšek, J.Á.: The least trimmed squares. Part I: consistency. Kybernetika **42**, 1–36 (2006)
28. Víšek, J.Á.: Consistency of the least weighted squares under heteroscedasticity. Kybernetika **47**, 179–206 (2011)
29. Wilcox, R.R.: Introduction to Robust Estimation and Hypothesis Testing, 2nd edn. Elsevier, Burlington (2005)

Weighted Empirical Minimum Distance Estimators in Berkson Measurement Error Regression Models

Hira L. Koul and Pei Geng

Abstract We develop analogs of the two classes of weighted empirical minimum distance (m.d.) estimators of the underlying parameters in linear and nonlinear regression models when covariates are observed with Berkson measurement error. One class is based on the integral of the square of symmetrized weighted empirical of residuals while the other is based on a similar integral involving a weighted empirical of residual ranks. The former class requires the regression and measurement errors to be symmetric around zero while the latter class does not need any assumption. The first class of estimators includes the analogs of the least absolute deviation and Hodges-Lehmann estimators while the second class includes an estimator that is asymptotically more efficient than these two estimators at some error distributions when there is no measurement error. In the case of linear model, no knowledge of the measurement error distribution is needed. We also develop these estimators for nonlinear models when the measurement error distribution is known and when it is unknown but validation data is available.

Keywords Analog of Hodges-Lehmann estimator · Validation data

1 Introduction

Statistical literature is replete with the various minimum distance estimation methods in the one and two sample location models. Beran [2, 3] and Donoho and Liu [7, 8] argue that the minimum distance estimators based on L_2 distances involving either density estimators or residual empirical distribution functions have some desirable

H. L. Koul (✉)
Department of Statistics and Probability, Michigan State University, East Lansing, MI, USA
e-mail: koul@msu.edu

P. Geng
Department of Mathematics, College of Arts and Sciences, Illinois State University, Normal, IL, USA
e-mail: pgeng@ilstu.edu

M. Maciak et al. (eds.), *Analytical Methods in Statistics*, Springer Proceedings in Mathematics & Statistics 329, https://doi.org/10.1007/978-3-030-48814-7_3

finite sample properties, tend to be robust against some contaminated models and are also asymptotically efficient at some error distributions.

In the classical regression models without measurement error in the covariates, classes of minimum distance estimators of the underlying parameters based on Cramér-von Mises type distances between certain weighted residual empirical processes were developed in Koul [12–15]. These classes include some estimators that are robust against outliers in the regression errors and asymptotically efficient at some error distributions.

In practice there are numerous situations when covariates are not observable. Instead one observes their surrogate with some error. The regression models with such covariates are known as the measurement error regression models. Fuller [9], Cheng and Van Ness [6], Carroll et al. [5] and Yi [19] discuss numerous examples of practical importance of these models.

Given the desirable properties of the above minimum distance (m.d.) estimators and the importance of the measurement error regression models, it is desirable to develop their analogs for these models. The next section describes the m.d. estimators of interest and their asymptotic distributions in the classical linear regression model. Their analogs for the linear regression Berkson measurement error (ME) model are developed in Sect. 3. The two classes of m.d. estimators are developed. One assumes the symmetry of the regression model error and ME error distributions and then basis the m.d. estimators on the symmetrized weighted empirical of the residuals. This class includes an analog of the Hodges-Lehmann estimator of the one sample location parameter, see Hodges and Lehmann (1963), and the least absolute deviation (LAD) estimator. The second class is based on a weighted empirical of residual ranks. This class of estimators does not need the symmetry of the errors distributions. This class includes an estimator that is asymptotically more efficient than the analog of Hodges-Lehmann and LAD estimators at some error distributions. Neither classes need the knowledge of the measurement error or regression error distributions.

Section 4 discusses analogs of these estimators in the Berkson measurement error nonlinear regression models, where the measurement error distribution is assumed to be known. Section 5 develops their analogs when the ME distribution is unknown but validation data is available. In this case the consistency rate of these estimators is $\min(n, N)^{1/2}$, where n and N are the primary data and validation data sample sizes, respectively. Section 6 provides an application of the proposed estimators to a real data example. Several proofs are deferred to the last section.

2 Linear Regression Model

In this section we recall the definition of the m.d. estimators of interest here in the no measurement error linear regression model and their known asymptotic normality results.

Accordingly, consider the linear regression model where for some $\theta \in \mathbb{R}^p$, the response variable Y and the p dimensional observable predicting covariate vector X obey the relation

$$Y = X'\theta + \varepsilon, \tag{1}$$

where ε is independent of X and symmetrically distributed around $E(\varepsilon) = 0$. For an $x \in \mathbb{R}$, x' and $\|x\|$ denote its transpose and Euclidean norm, respectively. Let (X_i, Y_i), $1 \le i \le n$ be a random sample from this model. The two classes of m.d. estimators of θ based on weighted empirical processes of the residuals and residual ranks were developed in Koul [12–15]. To describe these estimators, let G be a nondecreasing right continuous function from $\mathbb{R}$ to $\mathbb{R}$ having left limits and define

$$V(x, \vartheta) := n^{-1/2} \sum_{i=1}^{n} X_i \big\{ I(Y_i - X_i'\vartheta \le x) - I(-Y_i + X_i'\vartheta < x) \big\},$$

$$M(\vartheta) := \int \big\| V(x, \vartheta) \big\|^2 dG(x), \qquad \hat{\theta} := \text{argmin}_{\vartheta \in \mathbb{R}^p} M(\vartheta).$$

This class of estimators, one for each G, includes some well celebrated estimators. For example $\hat{\theta}$ corresponding to $G(x) \equiv x$ yields an analog of the one sample location parameter Hodges-Lehmann estimator in the linear regression model. Similarly, $G(x) \equiv \delta_0(x)$, the degenerate measure at zero, makes $\hat{\theta}$ equal to the least absolute deviation (LAD) estimator.

A class of m.d. estimators when the error distribution is not symmetric and unknown is obtained by using the weighted empirical of the residual ranks defined as follows. Write $X_i = (X_{i1}, X_{i2}, \ldots, X_{ip})'$, $i = 1, \ldots, n$. Let $\bar{X}_j := n^{-1} \sum_{i=1}^{n} X_{ij}$, $\bar{X} := (\bar{X}_1, \ldots, \bar{X}_p)'$ and $X_{ic} := X_i - \bar{X}$, $1 \le i \le n$. Let $R_{i\vartheta}$ denote the rank of the ith residual $Y_i - X_i'\vartheta$ among $Y_j - X_j'\vartheta$, $j = 1, \ldots, n$. Let Ψ be a distribution function on $[0, 1]$ and define

$$V(u, \vartheta) := n^{-1/2} \sum_{i=1}^{n} X_{ic} I(R_{i\vartheta} \le nu), \quad K(\vartheta) := \int_0^1 \|\mathcal{V}(u, \vartheta)\|^2 d\Psi(u),$$

$$\hat{\theta}_R := \text{argmin}_{\vartheta \in \mathbb{R}^p} K(\vartheta).$$

Yet another m.d. estimator, when error distribution is unknown and not symmetric, is

$$V_c(x, \vartheta) := n^{-1/2} \sum_{i=1}^{n} X_{ic} I(Y_i - X_i'\vartheta \le x),$$

$$M_c(\vartheta) := \int \big\| V_c(x, \vartheta) \big\|^2 dx, \qquad \hat{\theta}_c := \text{argmin}_{\vartheta \in \mathbb{R}^p} M_c(\vartheta).$$

If one reduces the model (1) to the two sample location model, then $\hat{\theta}_c$ is the median of pairwise differences, the so called Hodges-Lehmann estimator of the two sample location parameter. Thus in general $\hat{\theta}_c$ is an analog of this estimator in the linear regression model.

The following asymptotic normality results can be deduced from Koul [15] and [16, Sect. 5.4].

Lemma 1 *Suppose the model (1) holds and* $E\|X\|^2 < \infty$.
(a). *In addition, suppose* $\Sigma_X := E(XX')$ *is positive definite and the error d.f.* F *is symmetric around zero and has density* f. *Further, suppose the following hold.*

$$\begin{aligned} &G \textit{ is a nondecreasing right continuous function on } \mathbb{R} \textit{ to } \mathbb{R}, \\ &\textit{having left limits and } dG(x) = -dG(-x), \forall\, x \in \mathbb{R}. \end{aligned} \tag{2}$$

$$0 < \int f^j dG < \infty, \quad \lim_{z \to 0} \int \big[f(x+z) - f(x)\big]^j dG(x) = 0, \; j = 1, 2, \tag{3}$$

$$\int\limits_0^\infty (1 - F) dG < \infty.$$

Then

$$n^{1/2}(\hat{\theta} - \theta) \to_D N\big(0, \sigma_G^2 \Sigma_X^{-1}\big), \quad \sigma_G^2 := \frac{Var\Big(\int\limits_{-\infty}^{\varepsilon} f(x)dG(x)\Big)}{\big(\int f^2 dG\big)^2}.$$

(b). *In addition, suppose the error d.f.* F *has uniformly continuous bounded density* f, $\Omega := E\{(X - EX)(X - EX)'\}$ *is positive definite and* Ψ *is a d.f. on* $[0, 1]$ *such that* $\int\limits_0^1 f^2(F^{-1}(s))d\Psi(s) > 0$. *Then*

$$n^{1/2}\big(\hat{\theta}_R - \theta\big) \to_D N(0, \gamma_\Psi^2 \Omega^{-1}), \quad \gamma_\Psi^2 := \frac{Var\Big(\int\limits_0^{F(\varepsilon)} f(F^{-1}(s))d\Psi(s)\Big)}{\Big(\int\limits_0^1 f^2(F^{-1}(s))d\Psi(s)\Big)^2}.$$

(c). *In addition, suppose* Ω *is positive definite,* F *has square integrable density* f *and* $E|\varepsilon| < \infty$. *Then* $n^{1/2}(\hat{\theta}_c - \theta) \to_D N\big(0, \sigma_I^2 \Omega^{-1}\big)$, *where* $\sigma_I^2 := 1/12\big(\int f^2(x)dx\big)^2$.

Before proceeding further we now describe some comparison of the above asymptotic variances. Let $\sigma_{LAD}^2 := 1/(4f^2(0))$ and $\sigma_{LSE}^2 := \text{Var}(\varepsilon)$ denote the factors of the asymptotic covariance matrices of the LAD and the least squares estimators, respectively. Let γ_I^2 denote the γ_Ψ^2 when $\Psi(s) \equiv s$, i.e.,

$$\gamma_I^2 = \frac{\int\int \big[F(x \wedge y) - F(x)F(y)\big] f^2(x) f^2(y) dx dy}{\Big(\int\limits_0^1 f^3(x)dx\Big)^2}.$$

Table 1, obtained from Koul [16], gives the values of these factors for some distributions F. From this table one sees that the estimator $\hat{\theta}_R$ corresponding to $\Psi(s) \equiv s$

Table 1 A comparison of asymptotic variances

F	γ_I^2	σ_I^2	σ_{LAD}^2	σ_{LSE}^2
Double Exp.	1.2	1.333	1	2
Logistic	3.0357	3	4	3.2899
Normal	1.0946	1.0472	1.5708	1
Cauchy	2.5739	3.2899	2.46	∞

is asymptotically more efficient than the LAD at logistic error distribution while it is asymptotically more efficient than the Hodges-Lehmann type estimator at the double exponential and Cauchy error distributions. For these reasons it is desirable to develop analogs of $\hat{\theta}_R$ also for the ME models.

As argued in Koul (Chap. 5, [16]), the estimators $\{\hat{\theta}_G,\ G$ a d.f.$\}$ are robust against heavy tails in the error distribution in the general linear regression model. The estimator $\hat{\theta}_I$, where $G(x) \equiv x$, not a d.f., is robust against heavy tails and also asymptotically efficient at the logistic errors.

3 Berkson ME Linear Regression Model

In this section we shall develop analogs of the above estimators in the Berkson ME linear regression model, where the response variable Y obeys the relation (1) and where, instead of observing X, one observes a surrogate Z obeying the relation

$$X = Z + \eta. \tag{4}$$

In (4), Z, η, ε are assumed to be mutually independent and $E(\eta) = 0$. Note that η is $p \times 1$ vector of errors and its distribution need not be known.

Analog of $\hat{\theta}$. We shall first develop and derive the asymptotic distribution of the analogs of the estimators $\hat{\theta}$ in the Berkson ME linear regression model (1) and (4). Rewrite the model as

$$Y = Z'\theta + \xi, \qquad \xi := \eta'\theta + \varepsilon, \quad E(\xi) = 0, \quad \exists\, \theta \in \mathbb{R}. \tag{5}$$

Because Z, η, ε are mutually independent, ξ is independent of Z in (5).

Let H denote the distribution functions (d.f.) of η. Assume that the d.f. F of ε is continuous and symmetric around zero and that H is also symmetric around zero, i.e., $-dH(v) = dH(-v)$, for all $v \in \mathbb{R}^p$. Then the d.f. of ξ

$$L(x) := P(\xi \le x) = P(\eta'\theta + \varepsilon \le x) = \int F(x - v'\theta)dH(v)$$

is also continuous and symmetric around zero. This symmetry in turn motivates the following definition of the class of m.d. estimators of θ in the model (5), which mimics the definition of $\hat{\theta}$ by simply replacing X_i by Z_i. Define

$$\widetilde{V}(x,t) := n^{-1/2}\sum_{i=1}^{n} Z_i\big\{I(Y_i - Z_i't \le x) - I(-Y_i + Z_i't < x)\big\},$$

$$\widetilde{M}(t) := \int \big\|\widetilde{V}(x,t)\big\|^2 dG(x), \qquad \widetilde{\theta} := \operatorname{argmin}_{t\in\mathbb{R}^p}\widetilde{M}(t).$$

Because L is continuous and symmetric around zero and ξ is independent of Z, $E\widetilde{V}(x,\theta) \equiv 0$.

The following assumptions are needed for the asymptotic normality of $\widetilde{\theta}$.

$$E\|Z\|^2 < \infty \text{ and } \Gamma := EZZ' \text{ is positive definite.} \tag{6}$$

$$H \text{ satisfies } dH(v) = -dH(-v),\ \forall\, v \in \mathbb{R}^p. \tag{7}$$

$$F \text{ has Lebesgue density } f, \text{ symmetric around zero, and} \tag{8}$$

$$\text{such that } \ell(x) = \int f(x - v'\theta)dH(v) \text{ of } L \text{ satisfies the following:}$$

$$0 < \int \ell^j dG < \infty, \qquad \lim_{z\to 0}\int \big[\ell(y+z) - \ell(y)\big]^j dG(y) = 0, \quad j = 1, 2.$$

$$A := \int_0^\infty (1 - L)dG < \infty. \tag{9}$$

Under (6), $n^{-1}\sum_{i=1}^n Z_i Z_i' \to_p \Gamma$ and $n^{-1/2}\max_{1\le i\le n}\|Z_i\| \to_p 0$. Use these facts and argue as in Koul [15] to deduce that (2) and (6)–(9) imply

$$n^{1/2}\big(\widetilde{\theta} - \theta\big) \to_D \mathcal{N}(0, \tau_G^2\Gamma^{-1}), \qquad \tau_G^2 := \frac{\operatorname{Var}\big(\int_{-\infty}^{\xi}\ell dG\big)}{\big(\int \ell^2 dG\big)^2}. \tag{10}$$

Remark 1 We shall discuss some examples and some sufficient conditions for the above assumptions. The conditions (8) and (9) are satisfied by a large class of densities f, ME distributions H and integrating measure G. If G is a d.f., then f being uniformly continuous and bounded implies these conditions. In this case ℓ is also uniformly continuous, $\sup_x \ell(x) \le \sup_x f(x) < \infty$ so that $\int \ell^j dG \le \sup_x f^j(x) < \infty$ and $\int \big[\ell(y+z) - \ell(y)\big]^j dG(y) \le \sup_{|x-y|\le z}|\ell(y) - \ell(x)|^j \to 0$, as $z \to 0$. Moreover, here $A \le 1$. Thus these two assumptions reduce to assuming $\int \ell^j dG > 0$, $j = 1, 2$.

Given the importance of the two estimators corresponding to $G(x) \equiv x,\ G(x) \equiv \delta_0(x)$, it is of interest to provide some easy to verify sufficient conditions that imply conditions (8) and (9) for these two estimators.

Consider the case $G(x) \equiv x$. Assume f to be continuous and $\int f^2(x)dx < \infty$. Then because H is a d.f., ℓ is also continuous and symmetric around zero and $\int \ell(x+z)dx = \int \ell(x)dx = 1$. Moreover, by the Cauchy-Schwartz (C-S) inequality and Fubini's Theorem,

$$\begin{aligned} 0 < \int \ell^2(y)dy &= \int \Big(\int f(y - v'\theta)dH(v)\Big)^2 dy \\ &\leq \int\int f^2(y - v'\theta)dydH(v) = \int f^2(x)dx < \infty. \end{aligned}$$

Finally, because $\ell \in L_2$, by Theorem 9.5 in Rudin [18], it is shift continuous in L_2, i.e., (8) holds. Hence all conditions of (8) are satisfied.

Next, consider (9). The assumptions $E(\varepsilon) = 0$ and $E(\eta) = 0$ imply that $\int |x| f(x) dx < \infty$, $\int \|v\| dH(v) < \infty$ and hence

$$\int |y| dL(y) = \int |y| \int f(y - v'\theta)dH(v)dy = \int\int |x + v'\theta| f(x)dxdH(v) < \infty.$$

This in turn implies (9) in the case $G(x) \equiv x$.

To summarize, (6), (7), and F having continuous symmetric square integrable density f implies all of the above conditions needed for the asymptotic normality of the above analog of the Hodges-Lehmann estimator in the Berkson ME linear regression model. This fact is similar to the observation made in Berkson (1950) that the naive least square estimator, where one replace X_i's by Z_i's, continues to be consistent and asymptotically normal under the same conditions as when there is no ME. But, unlike in the no ME case, here the asymptotic variance

$$\tau_I^2 := \frac{\mathrm{Var}\big(L(\xi)\big)}{\big(\int \ell^2(y)dy\big)^2} = \frac{1}{12\big(\int \big(\int f(y - v'\theta)dH(v)\big)^2 dy\big)^2}$$

depends on θ. If H is degenerate at zero, i.e., if there is no ME, then $\tau_I^2 = \sigma_I^2$, the factor that appears in the asymptotic covariance matrix of the Hodges-Lehmann estimator in this case.

Next, consider the case $G(x) \equiv \delta_0(x)$—degenerate measure at 0. Assume f to be continuous and bounded from the above and

$$\ell(0) := \int f(v'\theta)dH(v) > 0. \tag{11}$$

Then the continuity and symmetry of f implies that as $z \to 0$,

$$\int \ell(y+z)dG(y) = \ell(z) = \int f(z - v'\theta)dH(v) \to \int f(-v'\theta)dH(v) = \ell(0),$$

$$\int \big[\ell(y+z)-\ell(y)\big]^2 dG(y) = \Big[\int \big\{f(z-v'\theta)-f(-v'\theta)\big\}dH(v)\Big]^2$$
$$\le \int \big\{f(z-v'\theta)-f(-v'\theta)\big\}^2 dH(v) \to 0.$$

Moreover, here $\int_0^\infty (1-L)dG = 1 - L(0) = 1/2$ so that (9) is also satisfied.

To summarize, (6), (7), (11) and f being continuous, symmetric around zero and bounded from the above imply all the needed conditions for the asymptotic normality of the above analog of the LAD estimator in the Berkson ME linear regression model. Moreover, here

$$\int_{-\infty}^{\xi} \ell(x)dG(x) = \ell(0)I(\xi \ge 0), \quad \int \ell^2(x)dG(x) = \ell^2(0),$$
$$\mathrm{Var}\Big(\int_{-\infty}^{\xi} \ell(x)dG(x)\Big) = \ell^2(0)/4.$$

Consequently, here the asymptotic covariance matrix also depends on θ via

$$\tau_0^2 = 1\big/4\ell^2(0) = 1\Big/4\Big(\int f(v'\theta)dH(v)\Big)^2.$$

In the case of no ME, $\Gamma^{-1}\tau_0^2$ equals the asymptotic covariance matrix of the LAD estimator. Unlike in the case of the previous estimator, here the conditions needed for f are a bit more stringent than those required for the asymptotic normality of the LAD estimator when there is no ME.

Analog of $\hat\theta_R$. Here we shall describe the analogs of the class of estimators $\hat\theta_R$ based on the residual ranks obtained from the model (5). These estimators do not need the errors ξ_i's to be symmetrically distributed. Let $\widetilde{R}_{i\vartheta}$ denote the rank of $Y_i - Z_i'\vartheta$ among $Y_j - Z_j'\vartheta,\ j = 1,\ldots,n$, $\bar Z := n^{-1}\sum_{i=1}^n Z_i$, $Z_{ic} := Z_i - \bar Z$, $1 \le i \le n$ and define

$$\widetilde{\mathcal{V}}(u,\vartheta) := n^{-1/2}\sum_{i=1}^n Z_{ic} I(\widetilde{R}_{i\vartheta} \le nu), \quad \widetilde{K}(\vartheta) := \int_0^1 \|\widetilde{\mathcal{V}}(u,\vartheta)\|^2 d\Psi(u),$$
$$\widetilde{\theta}_R := \mathrm{argmin}_{\vartheta\in\mathbb{R}^p}\widetilde{K}(\vartheta).$$

Use the facts $\sum_{i=1}^n Z_{ic} = 0$, $\Psi(\max(a,b)) = \max\{\Psi(a),\Psi(b)\}$ and $\max(a,b) = 2^{-1}[a+b+|a-b|]$, for any $a, b \in \mathbb{R}$, to obtain the computational formula

$$\widetilde{K}(t) = -\frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n Z_{ic}'Z_{jc}\Big|\Psi\Big(\frac{R_{it}}{n}-\Big)-\Psi\Big(\frac{R_{jt}}{n}-\Big)\Big|.$$

The following result can be deduced from Koul [15]. Suppose $E\|Z\|^2 < \infty$, $\widetilde{\Gamma} := E(Z - EZ)(Z - EZ)'$ is positive definite, density ℓ of the r.v. ξ is uniformly continuous and bounded and $\int\limits_0^1 \ell^2(L^{-1}(s))d\Psi(s) > 0$. Then $n^{-1/2}\max_{1\le i\le n}\|Z_i\| \to_p 0$, $n^{-1}\sum_{i=1}^n (Z_i - \bar{Z})(Z_i - \bar{Z})' \to_p \widetilde{\Gamma}$ and

$$n^{1/2}\big(\widetilde{\theta}_R - \theta\big) \to_D N\big(0, \widetilde{\tau}_\Psi^2 \widetilde{\Gamma}^{-1}\big), \quad \widetilde{\tau}_\Psi^2 := \frac{\mathrm{Var}\big(\int_0^{L(\xi)} \ell(L^{-1}(s))d\Psi(s)\big)}{\big(\int_0^1 \ell^2(L^{-1}(s))d\Psi(s)\big)^2}.$$

Density f of F being uniformly continuous and bounded implies the same for $\ell(x) = \int f(x - v'\theta)dH(v)$. It is also worth pointing out the assumptions on F, H and L needed here are relatively less stringent than those needed for the asymptotic normality of $\widetilde{\theta}$.

Of special interest is the case $\Psi(s) \equiv s$. Let $\widetilde{\tau}_I^2$ denote the corresponding $\widetilde{\tau}_\Psi^2$. Then by the change of variable formula,

$$\begin{aligned}\widetilde{\tau}_I^2 &= \frac{\mathrm{Var}\big(\int_0^{L(\xi)} \ell(L^{-1}(s))ds\big)}{\int_0^1 \ell^2(L^{-1}(s))ds} = \frac{\mathrm{Var}\big(\int_0^{\xi} \ell^2(x)dx\big)}{\big(\int_0^1 \ell^3(x)dx)^2} \\ &= \frac{\int\int \big[L(x \wedge y) - L(x)L(y)\big]\ell^2(x)\ell^2(y)dxdy}{\big(\int_0^1 \ell^3(x)dx\big)^2}.\end{aligned}$$

An analog of $\hat{\theta}_c$ here is $\widetilde{\theta}_c := \mathrm{argmin}_{\vartheta\in\mathbb{R}^p} \widetilde{M}_c(\vartheta)$, where

$$\widetilde{V}_c(x, \vartheta) := n^{-1/2}\sum_{i=1}^n Z_{ic} I(Y_i - Z_i'\vartheta \le x), \qquad \widetilde{M}_c(\vartheta) := \int \big\|\widetilde{V}_c(x, \vartheta)\big\|^2 dx.$$

Arguing as above one obtains that $n^{1/2}\big(\widetilde{\theta}_c - \theta\big) \to_D N\big(0, \tau_I^2 \widetilde{\Gamma}^{-1}\big)$.

4 Nonlinear Regression with Berkson ME

In this section we shall investigate the analogs of the above m.d. estimators in nonlinear regression models with Berkson ME.

Let $q \ge 1$, $p \ge 1$ be known positive integers, $\Theta \subseteq \mathbb{R}^q$ be a subset of the q-dimensional Euclidean space $\mathbb{R}^q$ and consider the model where the unobservable p-dimensional covariate X, its observable surrogate Z and the response variable Y obey the relations

$$Y = m_\theta(X) + \varepsilon, \qquad X = Z + \eta, \tag{12}$$

for some $\theta \in \Theta$. Here $m_\vartheta(x)$ is a known parametric function, nonlinear in x, from $\Theta \times \mathbb{R}^p$ to $\mathbb{R}$ with $E|m_\vartheta(X)| < \infty$, for all $\vartheta \in \Theta$. The r.v.'s ε, Z, η are assumed to be mutually independent, $E\varepsilon = 0$ and $E\eta = 0$. Unlike in the linear case, here we need to assume that the d.f. H of η is known. See Sect. 5 for the unknown H case.

Fix a θ for which (12) holds. Let $\nu_\vartheta(z) := E(m_\vartheta(X)|Z = z)$, $\vartheta \in \mathbb{R}^q$, $z \in \mathbb{R}^p$. Under (12), $E(Y|Z = z) \equiv \nu_\theta(z)$. Moreover, because H is known,

$$\nu_\vartheta(z) = \int m_\vartheta(z + s)dH(s)$$

is a known parametric regression function. Thus, under (12), we have the regression model

$$Y = \nu_\theta(Z) + \zeta, \qquad E(\zeta|Z = z) = 0, \qquad z \in \mathbb{R}^p.$$

Unlike in the linear case, the error ζ is no longer independent of Z in general.

To proceed further we assume there is a vector of p functions $\dot{m}_\vartheta(x)$ such that, with $\dot{\nu}_\vartheta(z) := \int \dot{m}_\vartheta(z + s)dH(s)$, for every $0 < b < \infty$,

$$\max_{1 \le i \le n, n^{1/2}\|\vartheta - \theta\| \le b} n^{1/2}\big|\nu_\vartheta(Z_i) - \nu_\theta(Z_i) - (\vartheta - \theta)'\dot{\nu}_\theta(Z_i)\big| = o_p(1), \tag{13}$$

$$E\|\dot{\nu}_\theta(Z)\|^2 < \infty. \tag{14}$$

Let

$$L_z(x) := P(\zeta \le x|Z = z), \qquad x \in \mathbb{R},\ z \in \mathbb{R}^p.$$

Assume the following. For every $z \in \mathbb{R}^p$,

$$L_z(\cdot) \text{ is continuous and } L_z(x) = 1 - L_z(-x),\ \forall x \in \mathbb{R}^p. \tag{15}$$

Let G be as before and define

$$U(x, \vartheta) := n^{-1/2} \sum_{i=1}^{n} \dot{\nu}_\vartheta(Z_i)\big\{I(Y_i - \nu_\vartheta(Z_i) \le x) - I(-Y_i + \nu_\vartheta(Z_i) < x)\big\}$$

$$D(\vartheta) := \int \big\|U(x, \vartheta)\big\|^2 dG(x), \qquad \widehat{\theta} := \text{argmin}_\vartheta D(\vartheta).$$

In the case $q = p$ and $m_\theta(x) = x'\theta$, $\widehat{\theta}$ agrees with $\widetilde{\theta}$. Thus the class of estimators $\widehat{\theta}$, one for each G, is an extension of the class of estimators $\widetilde{\theta}$ from the linear case to the above nonlinear case.

Next, consider the extension of $\hat{\theta}_R$ to the above nonlinear model (12). Let $S_{i\vartheta}$ denote the rank of $Y_i - \nu_\vartheta(Z_i)$ among $Y_j - \nu_\vartheta(Z_j)$, $j = 1, \ldots, n$ and define

$$\mathcal{U}_n(u,\vartheta) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\nu}_\vartheta(Z_i)\big\{I(S_{i\vartheta} \le nu) - u\big\},$$

$$\mathcal{K}(\vartheta) := \int \|\mathcal{U}_n(u,\vartheta)\|^2 d\Psi(u), \quad \widehat{\theta}_R := \text{argmin}_\vartheta \mathcal{K}(\vartheta).$$

The estimator $\widehat{\theta}_R$ gives an analog of $\hat{\theta}_R$ in the present set up.

Our goal here is to prove the asymptotic normality of $\widehat{\theta}$, $\widehat{\theta}_R$. This will be done by following the general method of Sect. 5.4 of Koul [16]. This method requires the two steps. In the first step we need to show that the defining dispersions $D(\vartheta)$ and $\mathcal{K}(\vartheta)$ are AULQ (asymptotically uniformly locally quadratic) in $\vartheta - \theta$ for $\vartheta \in \mathcal{N}_n(b) := \{\vartheta \in \Theta, n^{1/2}\|\vartheta - \theta\| \le b\}$, for every $0 < b < \infty$. The second step requires to show that $n^{1/2}\|\widehat{\theta} - \theta\| = O_p(1) = n^{1/2}\|\widehat{\theta}_R - \theta\|$.

4.1 Asymptotic Distribution of $\widehat{\theta}$

In this subsection we shall derive the asymptotic normality of $\widehat{\theta}$. To state the needed assumptions for achieving this goal we need some more notation. Let $\nu_{nt}(z) := \nu_{\theta+n^{-1/2}t}(z)$, $\xi_{it} := \nu_{nt}(Z_i) - \nu_\theta(Z_i)$, $1 \le i \le n$, $\dot{\nu}_{nt}(z) := \dot{\nu}_{\theta+n^{-1/2}t}(z)$, and $\dot{\nu}_{ntj}(z)$ denote the jth coordinate of $\dot{\nu}_{nt}(z)$, $1 \le j \le q$, $t \in \mathbb{R}^q$. For any real number a, let $a^\pm = \max(0, \pm a)$ so that $a = a^+ - a^-$. Also, let $\beta_i(x) := I(\zeta_i \le x) - L_{Z_i}(x)$ and $\alpha_i(x,t) := I(\zeta_i \le x + \xi_{it}) - I(\zeta_i \le x) - L_{Z_i}(x + \xi_{it}) + L_{Z_i}(x)$.

Because $dG(x) \equiv -dG(-x)$ and $U(x,\vartheta) \equiv U(-x,\vartheta)$, we have

$$D(\vartheta) \equiv 2\int_0^\infty \big\|U(x,\vartheta)\big\|^2 dG(x) \equiv 2\widetilde{D}(\vartheta), \qquad \text{say.} \tag{16}$$

We are now ready to state our assumptions.

$$\int_0^\infty E\Big(\|\dot{\nu}_\theta(Z)\|^2\big(1 - L_Z(x)\big)\Big) dG(x) < \infty. \tag{17}$$

$$\int_0^\infty E\Big(\|\dot{\nu}_{nt}(Z) - \dot{\nu}_\theta(Z)\|^2 L_Z(x)(1 - L_Z(x)\Big) dG(x) \to 0, \ \forall t \in \mathbb{R}^q.$$

$$\sup_{\|t\| \le b, 1 \le i \le n} \big\|\dot{\nu}_{nt}(Z_i) - \dot{\nu}_\theta(Z_i)\big\| \to_p 0. \tag{18}$$

Density ℓ_z of L_z exists for all $z \in \mathbb{R}^p$ such that (19)

$$0 < \int \ell_z(x) dG(x) < \infty, \ \forall z \in \mathbb{R}^p, \ \ 0 < \int E(\ell_Z^2(x)) dG(x) < \infty,$$

$$\int E\big(\|\dot{\nu}_\theta(Z)\|^2 \ell_Z^j(x)\big)dG(x) < \infty, \;\; j = 1, 2.$$

$$\lim_{u\to 0} \int_{-\infty}^{\infty} \big(\ell_z(x+u) - \ell_z(x)\big)^j dG(x) = 0, \;\; j = 1, 2, \forall\, z \in \mathbb{R}^p. \tag{20}$$

$$E\Big(\int_{-|\xi_t(Z)|}^{|\xi_t(Z)|} \|\dot{\nu}_{nt}(Z)\|^2 \int_{-\infty}^{\infty} \ell_Z(x+u)dG(x)du\Big) \to 0, \; \forall\, t \in \mathbb{R}^q, \tag{21}$$

where $\xi_t(z) := \nu_{nt}(z) - \nu_\theta(z)$.

With $\varGamma_\theta(x) := E\big(\dot{\nu}_\theta(Z)\dot{\nu}_\theta(Z)'\ell_Z(x)\big)$, the matrix (22)

$$\varOmega_\theta := \int_{-\infty}^{\infty} \varGamma_\theta(x)\varGamma_\theta(x)'dG(x) \text{ is positive definite.}$$

For every $\epsilon > 0$ there is a $\delta > 0$ and $N_\epsilon < \infty$ such that $\forall\, \|s\| \le b, n > N_\epsilon$,

$$P\Big(\sup_{\|t-s\|<\delta} \Big(n^{-1/2} \int \sum_{i=1}^{n} \big[\dot{\nu}_{ntj}^{\pm}(Z_i) - \dot{\nu}_{nsj}^{\pm}(Z_i)\big]\alpha_i(x,t)dG(x)\Big)^2 > \epsilon\Big) < \epsilon, \tag{23}$$

$$P\Big(\sup_{\|t-s\|<\delta} n^{-1} \int_0^{\infty} \Big\| \sum_{i=1}^{n} \{\dot{\nu}_{nt}(Z_i) - \dot{\nu}_{ns}(Z_i)\}\beta_i(x)\Big\|^2 dG(x) > \epsilon\Big) < \epsilon. \tag{24}$$

For every $\epsilon > 0$, $\alpha > 0$ there exists $N \equiv N_{\alpha,\varepsilon}$ and $b \equiv b_{\alpha,\epsilon}$ such that

$$P\Big(\inf_{\|t\|>b} D(\theta + n^{-1/2}t) \ge \alpha\Big) \ge 1 - \epsilon, \quad \forall n > N. \tag{25}$$

From now onwards we shall write ν and $\dot{\nu}$ for ν_θ and $\dot{\nu}_\theta$, respectively.

Remark 2 We shall now discuss the above assumptions when $m_\vartheta(x) = \vartheta' h(x)$, where $h = (h_1, \ldots, h_q)'$ is a vector of q function on $\mathbb{R}^p$ with $E\|h(X)\|^2 < \infty$, first for general G and then for some special cases of G. An example of this is the polynomial regression model with Berkson ME, where $p = 1, h_j(x) = x^j$, $j = 1, \ldots, q$. Let $\beta(z) := E(h(X)|Z = z)$. Then $\nu_\vartheta(z) = \vartheta'\beta(z)$ and $\dot{\nu}_\vartheta(z) \equiv \beta(z)$, a constant in ϑ. Therefore (13), (14), (18), (23) and (24) are all vacuously satisfied. The condition (25) also holds here, in a similar way as in the linear regression model, cf., Koul [16, Proof of Lemma 5.5.4, pp. 183–185]. Direct calculations show that (26)–(29) below imply the remaining assumptions (17), (19), (21) and (22), respectively.

$$\int_{0}^{\infty} E\Big(\|\beta(Z)\|^2\big(1 - L_Z(x)\big)\Big) dG(x) < \infty. \tag{26}$$

$\forall\, z \in \mathbb{R}^p$, density ℓ_z of L_z exists and satisfies (27)

$$0 < \int \ell_z^j(x)dG(x) < \infty, \;\; j = 1, 2, \quad 0 < \int E(\ell_Z^2(x))dG(x) < \infty,$$

$$\int E\big(\|\beta(Z)\|^2 \ell_Z^j(x)\big) dG(x) < \infty, \;\; j = 1, 2, \quad \text{and (20) holds.}$$

$$E\Big(\int_{n^{-1/2}b\|\beta(Z)\|}^{|n^{-1/2}b\|\beta(Z)\|} \|\beta(Z)\|^2 \int_{-\infty}^{\infty} \ell_Z(x+u)dG(x)du\Big) \to 0, \tag{28}$$

for every $0 < b < \infty$.

With $\mathcal{B}(x) := E\big(\beta(Z)\beta(Z)'\ell_Z(x)\big)$, the matrix (29)

$$\int_{-\infty}^{\infty} \mathcal{B}(x)\mathcal{B}(x)' dG(x) \text{ is positive definite.}$$

Consider further the case $G(x) \equiv x$. Let $\sigma := (E\varepsilon^2)^{1/2}$. Assume

$$\text{(a)}\;\; E\|h(X)\|^3 < \infty, \quad E\zeta^2 < \infty. \quad \text{(b)}\;\; C := \sup_{x\in\mathbb{R}, z\in\mathbb{R}^p} \ell_z(x) < \infty. \tag{30}$$

Then $E\|\beta(Z)\|^j \le E\|h(X)\|^j < \infty, \; j = 1, 2, 3.$ Let $\gamma(z) := 2\|\theta\|\,\|\beta(z)\| + \sigma$. Then

$$\begin{aligned} E\big(|\zeta|\big|Z = z\big) &= E\big(|Y - \theta'\beta(Z)|\big|Z = z\big) \\ &= E\big(|\theta' h(X) + \varepsilon - \theta'\beta(Z)|\big|Z = z\big) \le \gamma(z), \quad \forall\, z \in \mathbb{R}^p. \end{aligned} \tag{31}$$

Hence

$$\begin{aligned} &\int_{0}^{\infty} E\Big(\|\beta(Z)\|^2\big(1 - L_Z(x)\big)\Big)dx \\ &\le E\Big(\|\beta(Z)\|^2 E\big(|\zeta|\big|Z\big)\Big) \le E\Big(\|\beta(Z)\|^2 \gamma(Z)\Big) \\ &\le 2\|\theta\| E\big(\|\beta(Z)\|^3\big) + \sigma E\big(\|\beta(Z)\|^2\big) < \infty, \end{aligned}$$

thereby showing that (26) is satisfied. The assumption (30)(b) and $\ell_z(x)$ being a density in x for each z and Theorem 9.5 of Rudin [18] readily imply (27) here. The left hand side of (28) equals $2n^{-1/2}bE\big(\|\beta(Z)\|^3\big) \to 0$, by (30)(a).

Next, consider the case $G(x) = \delta_0(x)$- measure degenerate at zero. Assume

$$\lim_{u\to 0} \ell_z(u) = \ell_z(0) > 0, \ \forall z \in \mathbb{R}^p, \quad 0 < E\ell_Z^2(0) < \infty, \tag{32}$$
$$E\big(\|\beta(Z)\|^2 \ell_Z^j(0)\big) < \infty, \quad j = 1, 2.$$

Then the left hand side of (26) equals $(1/2)E\|\beta(Z)\|^2 < E\|h(X)\|^2 < \infty$. Condition (27) is trivially satisfied and the left hand side of (28) equals

$$E\Big(\|\beta(Z)\|^2\big[L_Z(n^{-1/2}b\|\beta(Z)\|) - L_Z(-n^{-1/2}b\|\beta(Z)\|)\big]\Big) \to 0,$$

by the DCT and the continuity of $L_z(\cdot)$, for each z.

To summarize, in the case $m_\vartheta(x) = \vartheta' h(X)$ and $G(x) \equiv x$, assumptions (30)(a), (b) and $\int \mathcal{B}(x)\mathcal{B}(x)'dx$ being positive definite imply all of the above assumptions (13), (14) and (17)–(25). Similarly, in the case $m_\vartheta(x) = \vartheta' h(X)$ and $G(x) \equiv \delta_0(x)$, $E\|h(X)\|^2 < \infty$, (32) and $\mathcal{B}(0)\mathcal{B}(0)'$ being positive definite imply all these conditions.

Remark 3 Because of the importance of the estimators $\widehat{\theta}$ when $G(x) = x$, and $G(x) = \delta_0(x)$, it is of interest to give some simple sufficient conditions for a general m_ϑ that imply the given assumptions for these two estimators.

Suppose G satisfies $dG(x) \equiv g(x)dx$, where $g_\infty := \sup_{x\in\mathbb{R}} g(x) < \infty$. Note that $G(x) \equiv x$ corresponds to the case $g(x) \equiv 1$. Consider the following assumptions.

$$\text{(a)} \ \ E\big\|\dot{m}_\theta(X)\big\|^4 < \infty, \tag{33}$$
$$\text{(b)} \ \ E\big\|\dot{m}_{\theta+n^{-1/2}t}(X) - \dot{m}_\theta(X)\big\|^2 \to 0, \ \forall t \in \mathbb{R}^q.$$

Density ℓ_z of L_z exists for all $z \in \mathbb{R}^p$ and satisfies (34)

$$0 < \int \ell_z^2(x)dx < \infty, \forall z \in \mathbb{R}^p, \quad 0 < \int E(\ell_Z^2(x))dx < \infty,$$

$$0 < \int E\big(\|\dot{\nu}(Z)\|^2 \ell_Z^2(x)\big)dx < \infty.$$
$$E\big(\|\dot{\nu}_{nt}(Z)\|^2 |\nu_{nt}(Z) - \nu(Z)|\big) \to 0, \quad \forall t \in \mathbb{R}^q. \tag{35}$$

Because $\big\|\dot{\nu}(Z)\big\|^j \le E\big(\|\dot{m}_\theta(X)\big\|^j\big|Z\big)$, $E\big\|\dot{\nu}(Z)\big\|^j \le E\big\|\dot{m}_\theta(X)\big\|^j < \infty, \ j = 1$, 2, 3, 4, by (33)(a). Similarly, for every $t \in \mathbb{R}^q$,

$$E\big\|\dot{\nu}_{nt}(Z) - \dot{\nu}(Z)\big\|^2 \le E\big\|\dot{m}_{nt}(X) - \dot{m}_\theta(X)\big\|^2 \to 0, \qquad \text{by (34)(b)}. \tag{36}$$

Next, similar to (31), $E\big(|\zeta|\big|Z\big) = E\big(|Y - \nu(Z)|\big|Z\big) \le 2|\nu(Z)| + \sigma$ implies that the left hand side of (17) is bounded from the above by

$$
\begin{aligned}
g_\infty E\Big(\|\dot\nu(Z)\|^2 E\big(|\zeta|\big|Z\big)\Big) &\le g_\infty E\Big(\|\dot\nu(Z)\|^2\big(2|\nu(Z)|+\sigma\big)\Big) \\
&\le g_\infty\big[2E^{1/2}(\|\dot\nu(Z)\|^4)E^{1/2}(m_\theta^2(X)) + \sigma E\|\dot\nu(Z)\|^2\big] \\
&< \infty,
\end{aligned}
$$

by (33)(a), thereby verifying (17) here. Similarly, with C denoting the above upper bound, for every $t \in \mathbb{R}^q$, the left hand side of (18) is bounded from the above $CE\Big(\|\dot\nu_{nt}(Z) - \dot\nu_\theta(Z)\|^2\Big) \to 0$, by (36). The left hand side of (21) is bounded from the above by $2g_\infty E\big(\|\dot\nu_{nt}(Z)\|^2|\nu_{nt}(Z) - \nu(Z)|\big) \to 0$, by (35).

In other words, in the case G has bounded Lebesgue density, conditions (33)–(35) imply assumptions (14), (17), (18), (19), (20), and (21). Not much simplification occurs in the remaining assumptions (18) and (22)–(25). See Remark 2 for some special cases.

Next consider the case when $G(x) = \delta_0(x)$ and the following assumptions.

$$
\sup_{x\in\mathbb{R}} \ell_z(x) < \infty, \;\; 0 < \lim_{u\to 0} \ell_z(u) = \ell_z(0) < \infty, \;\; \forall\, z \in \mathbb{R}^p. \tag{37}
$$

$$
\varGamma_\theta(0) \text{ is positive definite.} \tag{38}
$$

In this case (33), (35), (37) and (38) together imply the assumptions (14), (17)–(22). Not much simplification occurs in the remaining three assumptions (23)–(25), except in some special cases as in Remark 2.

We now resume the discussion about the asymptotic normality of $\widehat\theta$. First, we show that $E(D(\theta)) < \infty$, so that by the Markov inequality, $D(\theta)$ is bounded in probability. To see this, by (15), $EU(x,\theta) \equiv 0$ and, for $x \ge 0$,

$$
\begin{aligned}
E\|U(x,\theta)\|^2 &= E\Big(\|\dot\nu(Z)\|\big\{I(\zeta \le x) - I(\zeta > -x)\big\}\Big)^2 \\
&= 2E\Big(\|\dot\nu(Z)\|^2\big(1 - L_Z(x)\big)\Big).
\end{aligned}
$$

By the Fubini Theorem, (16) and (17),

$$
E(D(\theta)) = 2E(\widetilde D(\theta)) = 4\int_0^\infty E\Big(\|\dot\nu(Z)\|^2\big(1 - L_Z(x)\big)dG(x) < \infty. \tag{39}
$$

To state the AULQ result for D, we need some more notation. Let

$$
W(x,0) := n^{-1/2}\sum_{i=1}^n \dot\nu(Z_i)\big\{I\big(\zeta_i \le x\big) - L_{Z_i}(x)\big\}, \tag{40}
$$

$$
T_n := \int_{-\infty}^{\infty} \varGamma_\theta(x)\big\{W(x,0) + W(-x,0)\big\}dG(x), \quad \widehat t = -\varOmega_\theta^{-1}T_n/2,
$$

where $\varGamma_\theta(x)$ and $\varOmega_\theta$ are as in (22). We are ready to state the following lemma.

Lemma 2 *Suppose the above set up and assumptions (17)–(24) hold. Then for every $b < \infty$,*

$$\sup_{\|t\|\le b} \left| D(\theta + n^{-1/2}t) - D(\theta) - 4T_n' t - 4t'\varOmega_\theta t \right| \to_p 0. \tag{41}$$

If in addition (25) holds, then, with $\varSigma_\theta$ given at (45) below,

$$\begin{aligned} &\text{(a)} \quad \|n^{1/2}\big(\widehat{\theta} - \theta) - \widehat{t}\,\| \to_p 0. \\ &\text{(b)} \quad n^{1/2}\big(\widehat{\theta} - \theta) \to_D N\big(0, 4^{-1}\varOmega_\theta^{-1}\varSigma_\theta\varOmega_\theta^{-1}\big). \end{aligned} \tag{42}$$

Proof The proof of (41) appears in Sect. 6. The proof of the claim (42)(a), which uses (25), (39) and (41), is similar to that of Theorem 5.4.1 of Koul [16], where (25) and (39) are used to show that $n^{1/2}\|\widehat{\theta} - \theta\| = O_p(1)$.

Define, for $y \in \mathbb{R}$, $u \in \mathbb{R}^p$,

$$\psi_u(y) := \int_{-\infty}^{y} \ell_u(x)dG(x), \quad \varphi_u(y) := \psi_u(-y) - \psi_u(y). \tag{43}$$

By (19), $0 < \psi_u(y) \le \psi_u(\infty) = \int_{-\infty}^{\infty} \ell_u(x)dG(x) < \infty$, for all $u \in \mathbb{R}^p$. Thus for each u, $\psi_u(y)$ is an increasing continuous bounded function of y and $\psi_u(-y) \equiv \psi_u(\infty) - \psi_u(y)$, and $\varphi_u(y) = \psi_u(\infty) - 2\psi_u(y)$, for all $y \in \mathbb{R}$.

By (15), $E(\varphi_u(\zeta)|Z = z) = 0$, for all $u, z \in \mathbb{R}^p$. Let

$$\begin{aligned} C_z(u, v) &:= \mathrm{Cov}\big[\big(\varphi_u(\zeta), \varphi_v(\zeta)\big)\big|Z = z\big] = 4\mathrm{Cov}\big[\big(\psi_u(\zeta), \psi_v(\zeta)\big)\big|Z = z\big], \\ \mathcal{K}(u, v) &:= E\big(\dot{\nu}(Z)\dot{\nu}(Z)'C_Z(u, v)\big), \qquad u, v \in \mathbb{R}^p. \end{aligned}$$

Next let $\mu(z) := \dot{\nu}(z)\dot{\nu}(z)'$, Q denote the d.f. of Z and rewrite $\varGamma_\theta(x) = E\big(\dot{\nu}_\theta(Z)\dot{\nu}_\theta(Z)'\ell_Z(x)\big) = \int \mu(z)\ell_z(x)dQ(z)$. By the Fubini Theorem,

$$\begin{aligned} T_n &:= \int_{-\infty}^{\infty} \varGamma_\theta(x)\big\{W(x, 0) + W(-x, 0)\big\}dG(x) \\ &= \int \int_{-\infty}^{\infty} \mu(z)\big\{W(x, 0) + W(-x, 0)\big\}\ell_z(x)dG(x)dQ(z) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int \mu(z)\dot{\nu}(Z_i)\varphi_z(\zeta_i)dQ(z). \end{aligned} \tag{44}$$

Clearly, $ET_n = 0$ and by the Fubini Theorem, the covariance matrix of T_n is

$$\begin{aligned} \Sigma_\theta &:= ET_nT_n' \qquad (45)\\ &= E\Big\{\Big(\int \mu(z)\dot{\nu}(Z)\varphi_z(\zeta)dQ(z)\Big)\Big(\int \mu(v)\dot{\nu}(Z)\varphi_v(\zeta)dQ(v)\Big)'\Big\}\\ &= \int\int \mu(z)\mathcal{K}(z,v)\mu(v)'dQ(z)dQ(v). \end{aligned}$$

Thus T_n is a $p \times 1$ vector of independent centered finite variance r.v.'s. By the classical CLT, $T_n \to_D N(0, \Sigma_\theta)$. Hence, the minimizer $\widetilde{t}$ of the approximating quadratic form $D(\theta) + 4T_nt + 4t'\Omega_\theta t$ with respect to t satisfies $\tilde{t} = -\Omega_\theta^{-1}T_n/2 \to_D N\big(0, 4^{-1}\Omega_\theta^{-1}\Sigma_\theta\Omega_\theta^{-1}\big)$. The claim (42)(b) now follows from this result and (42)(a). □

4.2 Asymptotic Distribution of $\widehat{\theta}_R$

In this subsection we shall establish the asymptotic normality of $\widehat{\theta}_R$. For this we need the following assumptions, where $\mathcal{U}(b) := \{t \in \mathbb{R}^q;\ \|t\| \le b\}$, and $0 < b < \infty$.

$$\ell_z \text{ is uniformly continuous and bounded for every } z \in \mathbb{R}^p. \qquad (46)$$

$$n^{-1}\sum_{i=1}^n E\|\dot{\nu}_{nt}(Z_i) - \dot{\nu}(Z_i)\|^2 \to 0, \quad \forall\, t \in \mathcal{U}(b). \qquad (47)$$

$$n^{-1/2}\sum_{i=1}^n \|\dot{\nu}_{nt}(Z_i) - \dot{\nu}(Z_i)\| = O_p(1), \quad \forall\, t \in \mathcal{U}(b). \qquad (48)$$

$\forall\, \epsilon > 0,\ \exists\, \delta > 0$ and $n_\epsilon < \infty$ such that for each $s \in \mathcal{U}(b)$, $\forall\, n > n_\epsilon$,

$$P\Big(\sup_{t\in\mathcal{U}(b);\,\|t-s\|\le\delta} n^{-1/2}\sum_{i=1}^n \|\dot{\nu}_{nt}(Z_i) - \dot{\nu}_{ns}(Z_i)\| \le \epsilon\Big) > 1 - \epsilon. \qquad (49)$$

$\forall\, \epsilon > 0,\ 0 < \alpha < \infty,\ \exists\, N \equiv N_{\alpha,\epsilon}$ and $b \equiv b_{\epsilon,\alpha}$ such that

$$P\Big(\inf_{\|t\|>b} \mathcal{K}(\theta + n^{-1/2}t) \ge \alpha\Big) \ge 1 - \epsilon, \quad \forall\, n > N. \qquad (50)$$

Let

$$\bar{\dot{\nu}} := n^{-1}\sum_{i=1}^n \dot{\nu}(Z_i), \quad \dot{\nu}^c(Z_i) := \dot{\nu}(Z_i) - \bar{\dot{\nu}},$$

$$\widehat{\Gamma}_\theta(u) := E\Big(\dot{\nu}^c(Z)\dot{\nu}^c(Z)'\ell_Z(L_Z^{-1}(u))\Big), \quad \widehat{\Omega}_\theta := \int_0^1 \widehat{\Gamma}_\theta(u)\widehat{\Gamma}_\theta(u)'d\Psi(u),$$

$$\widehat{\mathcal{U}}(u) := n^{-1/2}\sum_{i=1}^n \dot{\nu}^c(Z_i)\big\{I(L_{Z_i}(\zeta_i) \le u) - u\big\}, \quad 0 \le u \le 1,$$

$$\widehat{T}_n := \int_0^1 \widehat{\Gamma}_\theta(u)\widehat{\mathcal{U}}(u)d\Psi(u), \quad \widehat{\mathcal{K}}(t) := \int_0^1 \big\|\widehat{\mathcal{U}}(u)\big\|^2 d\Psi(u) + 2\widetilde{T}_n't + t'\widetilde{\Omega}_\theta t.$$

We need to have an alternate representation of the covariance matrix of $\widehat{T}_n$. Let, for $z \in \mathbb{R}^p, \quad 0 \le v \le 1$,

$$\kappa_z(v) := \int_0^v \ell_z(L_z^{-1}(u))d\Psi(u), \quad k_z^c(v) = k_z(v) - \int_0^1 k_z(u)du.$$

By (46), κ_z is a uniformly continuous increasing and bounded function on [0, 1], for all $z \in \mathbb{R}^p$. Let U denote a uniform [0, 1] r.v. Conditionally, given Z, $L_Z(\zeta) \sim_D U$. Hence, $E\big(k_z\big(L_Z(\zeta)\big)\big|Z\big) = Ek_z(U)$ so that $E\big(k_z^c(L_Z(\zeta))\big|Z\big) = Ek_z^c(U) = 0$, a.s. Let $\mu^c(z) := \dot{\nu}^c(z)\dot{\nu}^c(z)'$. Argue as for (44) and use the facts that $\sum_{i=1}^n \dot{\nu}_c(Z_i) \equiv 0$ and $\int_0^1 udk_z(u) = k_z(1) - \int_0^1 k_z(u)du$ to obtain that

$$\widehat{T}_n = -n^{-1/2}\sum_{i=1}^n \int \mu^c(z)\dot{\nu}^c(Z_i)\kappa_z^c\big(L_{Z_i}(\zeta_i)\big)dQ(z).$$

Define

$$\widehat{C}_z(s,t) := E\big[\kappa_s^c(L_Z(\zeta))\kappa_t^c(L_Z(\zeta))\big|Z = z\big] = E\big[\kappa_s^c(U)\kappa_t^c(U)\big],$$
$$\widehat{K}(s,t) := E\big(\dot{\nu}^c(Z)\dot{\nu}^c(Z)'\widehat{C}_Z(s,t)\big).$$

Then argue as in (45) to obtain

$$\widehat{\Sigma}_\theta := E\widehat{T}_n\widehat{T}_n' = \int\int \mu^c(z)\widehat{K}(z,v)\mu^c(v)'dQ(z)dQ(v).$$

We are now ready to state the following asymptotic normality result for $\widehat{\theta}_R$.

Lemma 3 *Suppose the nonlinear Berkson measurement error model (12) and the assumptions (13), (14), (46)–(49) hold. Then the following holds.*

$$\sup_{\|t\|\le b}\left|\mathcal{K}(\theta+n^{-1/2}t)-\widehat{K}(t)\right|=o_p(1). \tag{51}$$

In addition, if (50) *holds and* $\widehat{\Omega}_\theta$ *is positive definite then* $n^{1/2}(\widehat{\theta}_R-\theta)\to_d N\left(0,\widehat{\Omega}_\theta^{-1}\widehat{\Sigma}_\theta\widehat{\Omega}_\theta^{-1}\right)$.

The proof of this lemma is similar to that of Theorem 1.2 of Koul [15], hence no details are given here. Assumption (50) is used to show that $n^{1/2}\|\widehat{\theta}_R-\theta\|=O_p(1)$.

Remark 4 As in Remark 1, let $m_\theta(x)=\theta' h(x)$. Then $\nu_\vartheta(z)=\vartheta'\beta(z)$, where $\beta(z):=E\big(h(X)|Z=z\big)$. Thus $\dot{\nu}_\vartheta(z)\equiv\beta(z)$ and the assumptions (47)–(49) are vacuously satisfied. The assumption (50) is shown to be satisfied by an argument similar to the one used in the proof of Lemma 5.4.4 of Koul [16, pp. 183–185]. This proof uses the monotonicity in t for every unit vector $e\in\mathbb{R}^p$ of simple linear rank statistics based on the ranks of $Y_i-te'h(X_i)$, $1\le i\le n$, see Hájek [10, Theorem II.7E].

For the asymptotic normality of $\widehat{\theta}_R$ here, one only needs (46) and Ψ to be a d.f. such that $\widehat{\Omega}$ is positive definite. Note that here $\mu^c(z)=\beta^c(z):=\beta(z)-\bar{\beta}$, $\bar{\beta}:=n^{-1}\sum_{i=1}^n\beta(Z_i)$ and

$$\widehat{K}(s,t):=E\big(\beta^c(Z)\beta^c(Z)'\widehat{C}_Z(s,t)\big),$$

$$\widehat{\Sigma}=\int\int\beta^c(z)\beta^c(z)'\widehat{K}(z,v)\beta^c(v)\beta^c(v)'dQ(z)dQ(v),$$

$$\widehat{\Gamma}(u):=E\Big(\beta^c(Z)\beta^c(Z)'\ell_Z(L_Z^{-1}(u))\Big),\quad \widehat{\Omega}=\int_0^1\widehat{\Gamma}(u)\widehat{\Gamma}(u)'d\Psi(u),$$

do not depend on θ. Clearly, these assumptions are far less stringent than those needed for the asymptotic normality of $\widehat{\theta}$ corresponding to $G(x)\equiv x$.

5 M.D. Estimators with Validation Data

In this section we develop the m.d. estimators of Sect. 4 when the d.f. H of the Berkson ME η is unknown but a validation data set is available. Not knowing H renders ν_θ to be an unknown function. Validation data is used to estimate this function, which in turn is used to define m.d. estimators.

Let N be a known positive integer. A set of r.v.'s $\{(\tilde{X}_k,\tilde{Z}_k),k=1,...,N\}$ is said to be validation data if these r.v.'s are independent of the original sample and both $\tilde{Z}_k$ and $\tilde{X}_k$ are observable and obey the model (12). Besides having the primary data set $\{(Y_i,Z_i),1\le i\le n\}$, we assume that a validation data set of the covariate $\{(\tilde{X}_k,\tilde{Z}_k),1\le k\le N\}$ is available. Then $\tilde{\eta}_k:=\tilde{X}_k-\tilde{Z}_k$, $1\le k\le N$ are observable and their empirical d.f. $H_N(s):=N^{-1}\sum_{k=1}^N I(\tilde{\eta}_k\le s)$, $s\in\mathbb{R}$, provides an estimate of H.

Under (13)–(15), we have the following estimates of ν_θ and $\dot{\nu}_\theta$.

$$\hat{\nu}_\vartheta(z) := N^{-1}\sum_{k=1}^{N} m_\vartheta(z+\tilde{\eta}_k), \qquad \hat{\dot{\nu}}_\vartheta(z) := N^{-1}\sum_{k=1}^{N} \dot{m}_\vartheta(z+\tilde{\eta}_k).$$

An analog of $\widehat{\theta}$ in the current set up is defined as follows. Let

$$\widehat{U}(x,\vartheta) := n^{-1/2}\sum_{i=1}^{n} \hat{\dot{\nu}}_\vartheta(Z_i)\{I(Y_i - \hat{\nu}_\vartheta(Z_i) \le x) - I(-Y_i + \hat{\nu}_\vartheta(Z_i) < x)\},$$
$$D_1(\vartheta) := \int \|\widehat{U}(x,\vartheta)\|^2 dG(x), \qquad \widehat{\theta_1} = \operatorname{argmin}_\vartheta D_1(\vartheta).$$

To define the analog of $\hat{\theta}_R$ here, let $\tilde{S}_{i\vartheta}$ be the rank of $Y_i - \hat{\nu}_\vartheta(Z_i)$ among $Y_j - \hat{\nu}_\vartheta(Z_j)$, $1 \le j \le n$ and define

$$\tilde{\mathcal{U}}_n(u,\vartheta) := \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \hat{\dot{\nu}}_\vartheta(Z_i)\{I(\tilde{S}_{i\vartheta} \le nu) - u\}, \quad 0 \le u \le 1,$$
$$\tilde{\mathcal{K}}(\vartheta) := \int_0^1 \|\tilde{\mathcal{U}}_n(u,\vartheta)\|^2 d\Psi(u), \quad \tilde{\theta}_R := \operatorname{argmin}_\vartheta \tilde{\mathcal{K}}(\vartheta).$$

The asymptotic distributions of $\widehat{\theta_1}$ and $\tilde{\theta}_R$ as $n \wedge N \to \infty$ are described in the next two subsections. In their derivations, the $\lim(n/N)$ of the ratio n/N plays an important role. Some of the proofs are similar to those of $\widehat{\theta}$ and $\widehat{\theta}_R$. Some key steps of the proof can be found in the Appendix.

5.1 Asymptotic Distribution of $\widehat{\theta}_1$

In this subsection we derive the asymptotic distribution of $\widehat{\theta_1}$. In addition to (13)–(15) and (17)–(25), the following assumptions are needed, where $\Delta_\vartheta(z) := \hat{\nu}_\vartheta(z) - \nu_\vartheta(z)$ and θ is as in (12).

$$E\Big\| E\big\{\dot{m}_\theta(X)[\nu_\theta(Z) - m_\theta(X)]|Z\big\}\Big\|^2 < \infty, \tag{52}$$
$$E\Big\| E\big\{\dot{m}_\theta(X)[\nu_\theta(Z) - m_\theta(X)]|\eta\big\}\Big\|^2 < \infty.$$

The matrix (53)

$$\Sigma_1 := \operatorname{Cov}\Big(E\big[\int\int \mu(z)\dot{\nu}_\theta(Z)\ell_z(x)\ell_Z(x)[m_\theta(Z+\eta) - \nu_\theta(Z)]dx dQ(z)\big|\eta\big]\Big)$$

is positive definite.

$$\lambda := \lim(n/N) \ge 0. \tag{54}$$

$$\max_{1\le i\le n}\Big|N^{-1}\sum_{k=1}^{N} m_\theta(Z_i+\tilde\eta_k)-\nu_\theta(Z_i)\Big| = o_p(1). \tag{55}$$

$$E\Big\{\|\dot\nu_\theta(Z)\|^2\big(m_\theta(X)-\nu_\theta(Z)\big)^2\Big\} < \infty. \tag{56}$$

$$\int_0^\infty E\Big(\|\hat{\dot\nu}_\theta(Z)\|^2\big(1-L_Z(x\pm\varDelta_\theta(Z))\big)\Big)dG(x) < \infty. \tag{57}$$

$$\begin{aligned}\int_0^\infty E\Big(\|\hat{\dot\nu}_{nt}(Z)-\hat{\dot\nu}_\theta(Z)\|^2 L_Z(x\pm\varDelta_\theta(Z)) \qquad\qquad & \qquad (58)\\ \times(1-L_Z(x\pm\varDelta_\theta(Z))\Big)dG(x)\to 0,\ \ \forall\, t\in\mathbb{R}^q.&\end{aligned}$$

We also assume that (18)–(24) and (25) hold with $\dot\nu_{nt}$, $\dot\nu_\theta$ and D replaced by $\hat{\dot\nu}_{nt}$, $\hat{\dot\nu}$ and D_1, respectively. We denote these assumptions as (18)*–(25)*.

Here we discuss some sufficient conditions for the above assumptions. By the C-S inequality, both the expressions of (52) are bounded from the above by $2E\big\|\dot m_\theta(X)\big\|^2E\big|m_\theta(X)\big|^2$. Thus (52) is implied by (33)(a) and having $E\big|m_\theta(X)\big|^2 < \infty$.

Next, under (33)(a), (57) is trivially satisfied when $G(x)\equiv\delta_0(x)$. In the case $dG(x)=g(x)dx$ with $g_\infty := \sup_{y\in\mathbb{R}} g(y) < \infty$, (57) is implied by (33)(a) and the following conditions.

$$E\big(\big\|\dot m_\theta(X)\|^2|m_\theta(X)|\big) < \infty. \tag{59}$$

To see this, note that $E\big(|\varDelta_\theta(Z)|\big|Z\big)\le 2E\big(|m_\theta(X)|\big|Z\big)$, and $\|\hat{\dot\nu}_\theta(Z)\|^2\le N^{-1}\sum_{k=1}^{N}\|\dot m_\theta(Z+\eta_k)\|^2$ so that $E\|\hat{\dot\nu}_\theta(Z)\|^2\le E\|\dot m_\theta(X)\|^2$. Now argue as in Remark 4.2 and use these facts to obtain that the left hand side of (57) is bounded from the above by

$$\begin{aligned}&g_\infty E\Big(\|\hat{\dot\nu}_\theta(Z)\|^2E\Big(|\zeta|+|\varDelta_\theta(Z)|\big|Z\Big)\Big)\\ &\le g_\infty\Big[CE\big\|\dot m_\theta(X)\|^2+2E\big(\big\|\dot m_\theta(X)\|^2|m_\theta(X)|\big)\Big] < \infty,\end{aligned}$$

by (33)(a) and (59). Similarly, the left hand side of (58) is bounded from the above by a constant multiple of

$$E\big\|\hat{\dot\nu}_{nt}(Z)-\hat{\dot\nu}_\theta(Z)\big\|^2\le E\big\|\dot m_{\theta+n^{-1/2}t}(X)-\dot m_\theta(X)\big\|^2\to 0,\ \ \text{by (34)(b).}$$

We now turn to proving the asymptotic normality of $\widehat{\theta}_1$. Similar to Sect. 4.1, we first prove that $E(D_1(\theta)) < \infty$. Recall $\Delta_\vartheta(z) := \hat{\nu}_\vartheta(z) - \nu_\vartheta(z)$ and rewrite

$$\widehat{U}(x,\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\dot{\nu}}_\theta(Z_i) \Big\{ I(\zeta_i \le x + \Delta_\theta(Z_i)) - I(-\zeta_i < x - \Delta_\theta(Z_i)) \Big\}.$$

By the independence of the primary and validation data and a conditioning argument, for every $x > 0$,

$$\begin{aligned} E\|\widehat{U}(x,\theta)\|^2 &= E\Big(\|\hat{\dot{\nu}}_\theta(Z)\| \{ I(\zeta \le x + \Delta_\theta(Z)) - I(-\zeta < x - \Delta_\theta(Z)) \} \Big)^2 \\ &= E\Big(\|\hat{\dot{\nu}}_\theta(Z)\|^2 \{ 1 - L_Z(x + \Delta_\theta(Z)) + 1 - L_Z(x - \Delta_\theta(Z)) \} \Big). \end{aligned}$$

Hence by (56), $ED_1(\theta) < \infty$.

Next we sketch the proof of the AULQ property of $D_1(\vartheta)$. Define

$$\begin{aligned} \tilde{W}(x,0) &:= n^{-1/2} \sum_{i=1}^{n} \hat{\dot{\nu}}(Z_i) \{ I(\zeta_i \le x + \Delta_\theta(Z_i)) - L_{Z_i}(x) \}, \\ \tilde{T}_n &:= \int \Gamma_\theta(x) \{ \tilde{W}(x,0) + \tilde{W}(-x,0) \} dG(x). \end{aligned}$$

In the Appendix, we show that $\tilde{T}_n$ is approximated by a U-statistic based on the two independent samples. Theorem 6.1.4 in Lehmann [17] yields

$$\begin{aligned} &(a) \quad \tilde{T}_n \to N(0, \Sigma_\theta + 4\lambda\Sigma_1), \quad \lambda < \infty, \\ &(b) \quad \sqrt{N/n}\,\tilde{T}_n \to N(0, 4\Sigma_1), \quad \lambda = \infty. \end{aligned} \tag{60}$$

Next, the assumptions (54)–(58) and (18)*–(24)* ensure that the analog of Lemma 5 holds here also. Hence (41) with T_n and $D(\vartheta)$ replaced by $\widetilde{T}_n$ and $D_1(\vartheta)$, respectively, holds. Moreover, analog of (42) can be shown to hold in a similar manner as in Sect. 4 under (25)*. Consequently, the asymptotic distribution of $\widehat{\theta}_1$ based on data sets $\{(Y_i, Z_i), 1 \le i \le n\}$ and $\{(\tilde{X}_k, \tilde{Z}_k), 1 \le k \le N\}$ described in the following lemma.

Lemma 4 *Suppose model* (12) *with H unknown holds and an independent validation data* $\{(\tilde{X}_k, \tilde{Z}_k), 1 \le k \le N\}$ *obeying* (12) *is available. In addition assume that* (17)–(25) *and* (52)–(57) *hold. Then*

$$\sqrt{n}(\widehat{\theta}_1 - \theta) \to N(0, 4^{-1}\Omega_\theta^{-1}(\Sigma_\theta + 4\lambda\Sigma_1)\Omega_\theta^{-1}), \text{ for } 0 \le \lambda < \infty;$$

$$\sqrt{N}(\widehat{\theta}_1 - \theta) \to N(0, 16^{-1}\Omega_\theta^{-1}\Sigma_1\Omega_\theta^{-1}), \text{ for } \lambda = \infty.$$

The above result shows that the estimation step of regression function $\nu_\theta(z)$ due to the unknown distribution H introduces more variation in the asymptotic distribution of the m.d. estimators. Moreover, the limiting ratio λ of the sample sizes plays a role in the additional variation. When $\lambda = \lim n/N = 0$, the additional covariance term vanishes, therefore it reduces to the case when the ME distribution is known. In other words, when the validation sample size N is sufficiently large, compared to the primary sample size n, both $\hat{\theta}$ and $\widehat{\theta_1}$ achieve the same asymptotic efficiency. On the other hand, when $\lambda = \infty$, i.e., when the validation data size is very limited compared to the primary data size, the estimation consistency rate is restricted to $\sqrt{N}$ instead of $\sqrt{n}$.

5.2 Asymptotic Distribution of $\tilde{\theta}_R$

In this subsection we present the asymptotic distribution of the class of estimators $\tilde{\theta}_R$. First, we provide the additional assumptions. Let $\hat{\dot{\nu}}_{nt}(z) = N^{-1}\sum_{k=1}^{N} \dot{\nu}_{\theta+n^{-1/2}t}(z + \tilde{\eta}_k)$. Consider the following assumptions.

$$n^{-1}\sum_{i=1}^{n} E\|\hat{\dot{\nu}}_{nt}(Z_i) - \hat{\dot{\nu}}(Z_i)\|^2 \to 0, \quad \forall\, t \in \mathcal{U}(b). \tag{61}$$

$$n^{-1/2}\sum_{i=1}^{n} \|\hat{\dot{\nu}}_{nt}(Z_i) - \hat{\dot{\nu}}(Z_i)\| = O_p(1), \quad \forall\, t \in \mathcal{U}(b). \tag{62}$$

$\forall\, \epsilon > 0,\ \exists\ \delta > 0$ and $n_\epsilon < \infty$ such that for each $s \in \mathcal{U}(b)$, $\forall\, n > n_\epsilon$,

$$P\Big(\sup_{t\in\mathcal{U}(b);\,\|t-s\|\le\delta} n^{-1/2}\sum_{i=1}^{n} \|\hat{\dot{\nu}}_{nt}(Z_i) - \hat{\dot{\nu}}_{ns}(Z_i)\| \le \epsilon\Big) > 1 - \epsilon. \tag{63}$$

For every $\epsilon > 0,\ 0 < \alpha < \infty$, there exist an N_ϵ and $b \equiv b_{\epsilon,\alpha}$ such that

$$P\Big(\inf_{\|t\|>b} \tilde{\mathcal{K}}(\theta + n^{-1/2}t) \ge \alpha\Big) \ge 1 - \epsilon, \quad \forall\, n > N_\epsilon. \tag{64}$$

The matrix (65)

$$\Sigma_2 := \mathrm{Cov}\Big(E\big[\int\int \mu^c(z)\{\dot{\nu}_\theta(Z) - E(\dot{\nu}_\theta(Z))\}\ell_z(x)\ell_Z(x) \times\{m_\theta(Z+\eta) - \nu_\theta(Z)\}dx\,dQ(z)\big|\eta\big]\Big)$$

is positive definite.

Next, define $\tilde{\dot{\nu}} := n^{-1}\sum_{i=1}^{n} \hat{\dot{\nu}}(Z_i), \qquad \tilde{\dot{\nu}}^c(Z_i) := \hat{\dot{\nu}}(Z_i) - \tilde{\dot{\nu}},$

$$\tilde{T}_{n,R} := \int_0^1 \widehat{\Gamma}_\theta(u)\tilde{\mathcal{U}}_R(u)d\Psi(u),$$

$$\tilde{\mathcal{K}}_R(t) := \int_0^1 \|\tilde{\mathcal{U}}_R(u)\|^2 d\Psi(u) + 2\tilde{T}'_{n,R}t + t'\widehat{\Omega}_\theta t,$$

where $\widehat{\Gamma}_\theta$ and $\widehat{\Omega}_\theta$ are defined in Sect. 4.2. Similar to Lemma 4. we have the following lemma.

Lemma 5 *Suppose model* (12) *with H unknown holds and an independent validation data $\{(\tilde{X}_k, \tilde{Z}_k), 1 \le k \le N\}$ obeying* (12) *is available. In addition assume that* (54)*,* (55)*,* (61)*–*(65) *hold. Then, for* $0 \le \lambda < \infty$*,*

$$\begin{aligned} &\text{(a)} \quad \tilde{T}_{n,R} \to N(0, \widehat{\Sigma}_\theta + \lambda\Sigma_2), \\ &\text{(b)} \quad n^{1/2}(\tilde{\theta}_R - \theta) \to N(0, \widehat{\Omega}_\theta^{-1}(\widehat{\Sigma}_\theta + \lambda\Sigma_2)\widehat{\Omega}_\theta^{-1}). \end{aligned} \tag{66}$$

Moreover, $N^{1/2}(\tilde{\theta}_R - \theta) \to N(0, \widehat{\Omega}_\theta^{-1}\Sigma_2\widehat{\Omega}_\theta^{-1})$*, for* $\lambda = \infty$*.*

See Appendix for some details of the proof.

6 Data Analysis

Example. We shall now compute the above estimators based on some real data. The data pertains to the study of the relationship between the enzyme reaction speed (Y) and the basal density (X) of the UDP-galactose, see Bates and Watts [1], p. 70. A suitable model commonly used to analyze this data is the Michaelis-Menten model

$$m_\theta(x) = \frac{\alpha x}{\beta + x}, \quad \theta := (\alpha, \beta)', \; \alpha > 0, \; \beta > 0, \; x > 0.$$

In the primary data, consisting of $n = 30$ observations, the basal density variable was measured using a simple chemical method. It was believed that this method caused measurement error in the observation. Hence, in the validation data, consisting of $N = 10$ observations, an expensive procedure with a precision machine tool was used to produce precise observations of the basal density. Let Z denote the basal-density obtained by the chemical method parts per millions (ppm), $\widetilde{Z}$ denote the basal-density obtained by the exact measure (ppm) and Y, the reaction speed (counts/min^2). The primary and validation data are as follows. Table 2 gives the m.d. estimators $\hat{\theta}_1$ with $G(x) = x$ and $\tilde{\theta}_R$ with $\Psi(u) = u$, based on the above primary and validation data, and the naive least squares estimators $\hat{\theta}_{\text{nLS}}$ obtained by ignoring measurement errors. The MSEs are calculated by using the following formulas, where $\widetilde{\eta}_k = \widetilde{X}_k - \widetilde{Z}_k$, $MSE(\hat{\theta}_1) = \frac{1}{n}\sum_{i=1}^n \Big[Y_i -$

Table 2 M.D. and naive estimators and their MSE

Estimators	$\hat{\theta}_1$	$\tilde{\theta}_R$	$\hat{\theta}_{\text{nLS}}$
α	217.30	217.53	212.7
γ	0.069	0.063	0.064
MSE	48.96	57.85	49.87

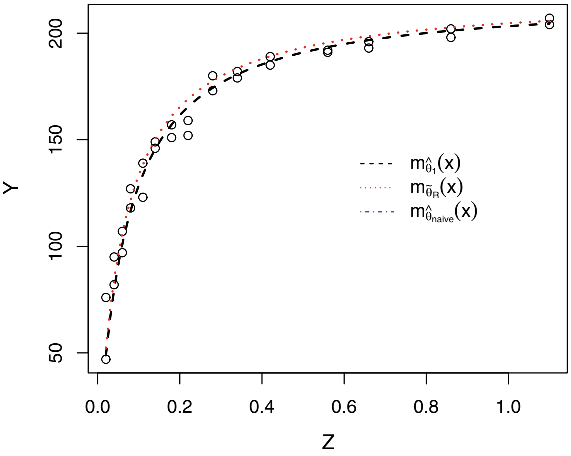


Fig. 1 Fitted regressions based on the three estimators

$\frac{1}{N}\sum_{k=1}^{N} m_{\hat{\theta}_1}(Z_i+\tilde{\eta}_k)\Big]^2$, $MSE(\tilde{\theta}_R)=\frac{1}{n}\sum_{i=1}^{n}\Big[Y_i-\frac{1}{N}\sum_{k=1}^{N} m_{\tilde{\theta}_R}(Z_i+\tilde{\eta}_k)\Big]^2$ and $MSE(\hat{\theta}_{\text{nLS}})=\frac{1}{n}\sum_{i=1}^{n}\Big[Y_i-m_{\hat{\theta}_{\text{nLS}}}(Z_i)\Big]^2$. Figure 1 presents the fitted regression curves using the three estimators.

Z	0.02	0.02	0.04	0.04	0.06	0.06	0.08	0.08	0.11	0.11	0.14	0.14	0.18	0.18	0.22
Y	76	47	82	95	97	107	118	127	123	139	146	149	157	151	159
Z	0.42	0.42	0.56	0.56	0.66	0.66	0.86	0.86	1.10	1.10	0.22	0.28	0.28	0.34	0.34
Y	185	189	191	192	193	196	198	202	207	2.04	152	173	180	179	182

$\widetilde{Z}$	0.04	0.07	0.20	0.30	0.38	0.48	0.60	0.76	0.95	1.110
$\widetilde{X}$	0.035	0.076	0.207	0.295	0.388	0.486	0.601	0.754	0.952	1.112

Appendix

This section contains some details of the proofs of the various results.

Proof of (41). Let $\widetilde{M}(t)=\widetilde{D}(\theta+n^{-1/2}t)$, where $\widetilde{D}(\vartheta)$ is as in (16). Define

$$
\begin{aligned}
\nu_{nt}(z) &:= \nu_{\theta+n^{-1/2}t}(z), \qquad \xi_{it} := \nu_{nt}(Z_i) - \nu_\theta(Z_i),\\
V_s(x,t) &:= \frac{1}{\sqrt{n}}\sum_{i=1}^n \dot{\nu}_{ns}(Z_i) I(Y_i - \nu_{nt}(Z_i) \le x)\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n \dot{\nu}_{ns}(Z_i) I\big(\zeta_i \le x + \xi_{it}\big),\\
V(x,t) &:= \frac{1}{\sqrt{n}}\sum_{i=1}^n \dot{\nu}(Z_i) I\big(\zeta_i \le x + \xi_{it}\big),\\
J(x,t) &:= \frac{1}{\sqrt{n}}\sum_{i=1}^n \dot{\nu}(Z_i) L_{Z_i}(x + \xi_{it}),\\
J_s(x,t) &:= \frac{1}{\sqrt{n}}\sum_{i=1}^n \dot{\nu}_{ns}(Z_i) L_{Z_i}(x + \xi_{it}), \quad W_s(x,t) := V_s(x,t) - J_s(x,t),\\
W(x,t) &:= V(x,t) - J(x,t), \quad s, t \in \mathbb{R}^q, x \in \mathbb{R}.
\end{aligned}
$$

Note that $EV_s(x,t) \equiv EJ_s(x,t),\ EW_s(x,t) \equiv 0$. By (15), $\forall s \in \mathbb{R}^q, x \in \mathbb{R}$,

$$
n^{-1/2}\sum_{i=1}^n \dot{\nu}_{ns}(Z_i)\big\{L_{Z_i}(x) + L_{Z_i}(-x)\big\} = n^{-1/2}\sum_{i=1}^n \dot{\nu}_{ns}(Z_i).
$$

Define

$$
\gamma_{nt}(x) := n^{-1/2}\sum_{i=1}^n \dot{\nu}_{nt}(Z_i)\xi_{it}\ell_{Z_i}(x), \quad g_n(x) := n^{-1}\sum_{i=1}^n \dot{\nu}(Z_i)\dot{\nu}(Z_i)'\ell_{Z_i}(x).
$$

Because of (15), $\gamma_{nt}(x) \equiv \gamma_{nt}(-x),\ g_n(x) \equiv g_n(-x)$ and we rewrite

$$
\begin{aligned}
\widetilde{M}(t) &= \int_0^\infty \Big\| V_t(x,t) + V_t(-x,t) - n^{-1/2}\sum_{i=1}^n \dot{\nu}_{nt}(Z_i)\Big\|^2 dG(x)\\
&= \int_0^\infty \Big\| \big\{W_t(x,t) - W_t(x,0)\big\} + \big\{W_t(x,0) - W(x,0)\big\}\\
&\qquad + \big\{W_t(-x,t) - W_t(-x,0)\big\} + \big\{W_t(-x,0) - W(-x,0)\big\}\\
&\qquad + \big\{J_t(x,t) - J_t(x,0) - \gamma_{nt}(x)\big\}\\
&\qquad + \big\{J_t(-x,t) - J_t(-x,0) - \gamma_{nt}(-x)\big\} + 2\big\{\gamma_{nt}(x) - g_n(x)t\big\}\\
&\qquad + \big\{W(x,0) + W(-x,0) + 2g_n(x)t\big\}\Big\|^2 dG(x).
\end{aligned}
$$

Expand the quadratic of the six summands in the integrand to obtain

$$\widetilde{M}(t) = M_1(t) + M_2(t) + \cdots + M_8(t) + 28 \text{ cross product terms},$$

where

$$M_1(t) := \int_0^\infty \left\| W_t(x, t) - W_t(x, 0) \right\|^2 dG(x),$$

$$M_2(t) := \int_0^\infty \left\| W_t(x, 0) - W(x, 0) \right\|^2 dG(x),$$

$$M_3(t) := \int_0^\infty \left\| W_t(-x, t) - W_t(-x, 0) \right\|^2 dG(x),$$

$$M_4(t) := \int_0^\infty \left\| W_t(-x, 0) - W(-x, 0) \right\|^2 dG(x),$$

$$M_5(t) := \int_0^\infty \left\| J_t(x, t) - J_t(x, 0) - \gamma_{nt}(x) \right\|^2 dG(x),$$

$$M_6(t) := \int_0^\infty \left\| J_t(-x, t) - J_t(-x, 0) - \gamma_{nt}(-x) \right\|^2 dG(x),$$

$$M_7(t) := 4 \int_0^\infty \left\| \gamma_{nt}(x) - g_n(x)t \right\|^2 dG(x),$$

$$M_8(t) := \int_0^\infty \left\| W(x, 0) + W(-x, 0) + 2g_n(x)t \right\|^2 dG(x).$$

Recall $\mathcal{U}(b) := \{t \in \mathbb{R}^q; \|t\| \le b\}$, $b > 0$. We shall prove the following lemma shortly.

Lemma 6 *Under the assumptions* (13) *to (18),* $\forall\, 0 < b < \infty$,

$$\sup_{t \in \mathcal{U}(b)} M_j(t) \to_p 0, \quad j = 1, 2, \ldots, 7, \tag{67}$$

$$\sup_{t \in \mathcal{U}(b)} M_8(t) \;=\; O_p(1). \tag{68}$$

Unless mentioned otherwise, all the supremum below are taken over $t \in \mathcal{U}(b)$. Lemma 6 together with the C-S inequality implies that the supremum over t of all the cross product terms tends to zero, in probability. For example, by the C-S inequality,

$$\sup_t \Big| \int_0^\infty \big\{W_t(x,t) - W_t(x,0)\big\}\big\{J_t(x,t) - J_t(x,0) - \gamma_{nt}(x)\big\} dG(x)\Big|^2 \\ \le \sup_t M_1(t) \sup_t M_5(t) = o_p(1),$$

by (67) used with $j = 1, 5$. Similarly, by (67) with $j = 1$ and (68),

$$\sup_t \Big| \int_0^\infty \big\{W_t(x,t) - W_t(x,0)\big\}\big\{W(x,0) + W(-x,0) + 2g_n(x)t\big\} dG(x)\Big|^2 \\ \le \sup_t M_1(t) \sup_t M_8(t) = o_p(1) \times O_p(1) = o_p(1).$$

Consequently, we obtain

$$\sup_t \big|\widetilde{M}(t) - M_8(t)\big| = o_p(1). \tag{69}$$

Expand the quadratic in M_8 to write

$$\begin{aligned} M_8(t) := &\int_0^\infty \big\|W(x,0) + W(-x,0)\big\|^2 dG(x) + \\ &4t' \int_0^\infty g_n(x)\big\{W(x,0) + W(-x,0)\big\} dG(x) + 4\int_0^\infty \big(t' g_n(x)\big)^2 dG(x) \\ = &\ \widetilde{M}(0) + 4t'\widetilde{T}_n + 4\int_0^\infty \big(t' g_n(x)\big)^2 dG(x), \end{aligned} \tag{70}$$

where $\widetilde{T}_n := \int_0^\infty g_n(x)\big\{W(x,0) + W(-x,0)\big\} dG(x)$. Let

$$T_n^* := \int_0^\infty \Gamma_\theta(x)\big\{W(x,0) + W(-x,0)\big\} dG(x).$$

By the LLNs and an Extended Dominated Convergence Theorem

$$\sup_t \left\| t'(g_n(x) - \Gamma_\theta(x)) \right\| \to_p 0, \quad \forall\, x \in \mathbb{R};$$
$$\sup_t \int_0^\infty \left\| t'(g_n(x) - \Gamma_\theta(x)) \right\|^2 dG(x) \to_p 0.$$

Moreover, recall $\widetilde{M}(0) = \widetilde{D}(\theta)$, so that by (39), $\widetilde{M}(0) = O_p(1)$. These facts together with the C-S inequality imply that

$$\begin{aligned}
\left\| \widetilde{T}_n - T_n^* \right\|^2 &= \left\| \int_0^\infty \left\{ g_n(x) - \Gamma_\theta(x) \right\} \left\{ W(x,0) + W(-x,0) \right\} dG(x) \right\|^2 \\
&\le \widetilde{M}(0) \int_0^\infty \left\| g_n(x) - \Gamma_\theta(x) \right\|^2 dG(x) \to_p 0.
\end{aligned}$$

These facts combined with (22), (69), (70) yield that

$$\sup_t \left| \widetilde{M}(t) - \widetilde{M}(0) - 4T_n^* t - 4t' \int_0^\infty \Gamma_\theta(x) \Gamma_\theta(x) dG(x)\, t \right| = o_p(1).$$

Now recall that $D(\vartheta) = 2\widetilde{D}(\vartheta)$, $\widetilde{M}(t) = \widetilde{D}(\theta + n^{-1/2}t)$, $\Omega_\theta = 2\int_0^\infty \Gamma_\theta \Gamma_\theta dG$ and $T_n = 2T_n^*$, see (40). Hence the above expansion is equivalent to

$$\begin{aligned}
&\sup_t \left| \widetilde{D}(\theta + n^{-1/2}t) - \widetilde{D}(\theta) - 2T_n t - 2t'\Omega_\theta t \right| = o_p(1), \\
&\sup_t \left| D(\theta + n^{-1/2}t) - D(\theta) - 4T_n t - 4t'\Omega_\theta t \right| = o_p(1),
\end{aligned}$$

which is precisely the claim (41). □

Proof of Lemma 6. Let $\delta_{it} := \xi_{it} - n^{-1/2} t' \dot{\nu}(Z_i)$. By (13) and (14),

$$\max_{1 \le i \le n,\, t} n^{1/2} \left| \delta_{it} \right| = o_p(1), \qquad \max_{1 \le i \le n} n^{-1/2} \|\dot{\nu}(Z_i)\| = o_p(1). \tag{71}$$

Hence,

$$\begin{aligned}
\max_{1 \le i \le n,\, t} \left| \xi_{it} \right| &\le \max_{1 \le i \le n,\, \|t\| \le b} \left| \delta_{it} \right| + \max_{1 \le i \le n,\, t} n^{-1/2} \left\| t \right\| \|\dot{\nu}(Z_i)\| \qquad (72) \\
&\le o_p(n^{-1/2}) + b \max_{1 \le i \le n} n^{-1/2} \|\dot{\nu}(Z_i)\| = o_p(1), \\
\sum_{i=1}^n \xi_{it}^2 &= \sum_{i=1}^n (\nu_{nt}(Z_i) - \nu(Z_i))^2 = \sum_{i=1}^n \delta_{it}^2 + n^{-1} \sum_{i=1}^n (t'\dot{\nu}(Z_i))^2,
\end{aligned}$$

$$\sup_t \sum_{i=1}^n \xi_{it}^2 \le n \max_{1\le i\le n,\, \|t\|\le b} \left|\delta_{it}\right|^2 + b^2 n^{-1} \sum_{i=1}^n \|\dot\nu(Z_i)\|^2 = O_p(1), \tag{73}$$

by (14). Moreover, by (14) and the Law of Large Numbers,

$$\begin{aligned}
&\sup_t \left\| n^{-1/2} \sum_{i=1}^n \dot\nu_\theta(Z_i)\xi_{it} \right\| \qquad (74)\\
&\le \max_{1\le i\le n,\, \|t\|\le b} n^{1/2}|\delta_{it}| n^{-1} \sum_{i=1}^n \|\dot\nu_\theta(Z_i)\| + bn^{-1} \left\| \sum_{i=1}^n \dot\nu_\theta(Z_i)\dot\nu_\theta(Z_i)' \right\| \\
&= o_p(1) + O_p(1) = O_p(1).
\end{aligned}$$

These facts will be use in the sequel.

Consider the term M_7. Write

$$\begin{aligned}
&\gamma_{nt}(x) - g_n(x)t \\
&= n^{-1/2} \sum_{i=1}^n \dot\nu_{nt}(Z_i)\xi_{it}\ell_{Z_i}(x) - n^{-1} \sum_{i=1}^n \dot\nu(Z_i)\dot\nu(Z_i)'\ell_{Z_i}(x)t \\
&= n^{-1/2} \sum_{i=1}^n \left[\dot\nu_{nt}(Z_i) - \dot\nu(Z_i)\right]\xi_{it}\ell_{Z_i}(x) + n^{-1/2} \sum_{i=1}^n \dot\nu(Z_i)\delta_{it}\ell_{Z_i}(x) \\
&= n^{-1/2} \sum_{i=1}^n \left[\dot\nu_{nt}(Z_i) - \dot\nu(Z_i)\right]\delta_{it}\ell_{Z_i}(x) \\
&\quad + n^{-1} \sum_{i=1}^n \left[\dot\nu_{nt}(Z_i) - \dot\nu(Z_i)\right]\dot\nu(Z_i)'\ell_{Z_i}(x)t + n^{-1/2} \sum_{i=1}^n \dot\nu(Z_i)\delta_{it}\ell_{Z_i}(x).
\end{aligned}$$

Hence

$$M_7 = \int_0^\infty \left\|\gamma_{nt}(x) - g_n(x)t\right\|^2 dG(x) \le 4\{M_{71}(t) + M_{72}(t) + M_{73}(t)\},$$

where

$$\begin{aligned}
M_{71}(t) &= n^{-1} \int_0^\infty \left\| \sum_{i=1}^n \left[\dot\nu_{nt}(Z_i) - \dot\nu(Z_i)\right]\delta_{it}\ell_{Z_i}(x) \right\|^2 dG(x), \\
M_{72}(t) &= n^{-2} \int_0^\infty \left\| \sum_{i=1}^n \left[\dot\nu_{nt}(Z_i) - \dot\nu(Z_i)\right]\dot\nu(Z_i)'\ell_{Z_i}(x)t \right\|^2 dG(x),
\end{aligned}$$

$$M_{73}(t) = n^{-1} \int_0^\infty \Big\| \sum_{i=1}^n \dot{\nu}(Z_i)\delta_{it}\ell_{Z_i}(x) \Big\|^2 dG(x).$$

But, by (18) and (71),

$$\sup_t M_{71}(t)$$
$$\leq n \sup_{t,1\leq i\leq n} \delta_{it}^2 \sup_{t,1\leq i\leq n} \big\|\dot{\nu}_{nt}(Z_i) - \dot{\nu}(Z_i)\big\|^2 \int_0^\infty n^{-1} \sum_{i=1}^n \ell_{Z_i}^2(x)dG(x) = o_p(1).$$

Similarly, by the C-S inequality,

$$\sup_t M_{72}(t) \leq b^2 \sup_{t,1\leq i\leq n} \big\|\dot{\nu}_{nt}(Z_i) - \dot{\nu}(Z_i)\big\|^2 n^{-1} \int_0^\infty \sum_{i=1}^n \|\dot{\nu}(Z_i)\|^2 \ell_{Z_i}^2(x)dG(x)$$
$$= o_p(1)O_p(1) = o_p(1),$$

by (18) and (19). Again, by (19) and (71),

$$\sup_t M_{73}(t) \leq \sup_{t,1\leq i\leq n} n|\delta_{it}|^2 n^{-1} \int_0^\infty \sum_{i=1}^n \|\dot{\nu}(Z_i)\|^2 \ell_{Z_i}^2(x)dG(x) = o_p(1).$$

These facts prove (67) for $j = 7$.

Next consider M_5. Let $D_{it}(x) := L_{Z_i}(x + \xi_{it}) - L_{Z_i}(x) - \xi_{it}\ell_{Z_i}(x)$. Then

$$M_5(t) := \frac{1}{n} \int_0^\infty \Big\| \sum_{i=1}^n \dot{\nu}_{nt}(Z_i) D_{it}(x) \Big\|^2 dG(x) \tag{75}$$
$$\leq \frac{1}{n} \sum_{i=1}^n \Big\|\dot{\nu}_{nt}(Z_i)\Big\|^2 \int_0^\infty \sum_{i=1}^n D_{it}^2(x)dG(x).$$

By (14) and (18),

$$\sup_t n^{-1} \sum_{i=1}^n \Big\|\dot{\nu}_{nt}(Z_i)\Big\|^2 \leq \sup_t \; n^{-1} \sum_{i=1}^n \Big\|\dot{\nu}_{nt}(Z_i) - \dot{\nu}(Z_i)\Big\|^2$$
$$+ \sup_t n^{-1} \sum_{i=1}^n \Big\|\dot{\nu}(Z_i)\Big\|^2 = o_p(1).$$

By the C-S inequality, Fubini Theorem, (20) and (73),

$$\begin{aligned}
&\int_0^\infty \sum_{i=1}^n D_{it}^2(x)dG(x) \\
&\leq \int_0^\infty \sum_{i=1}^n \Big(\int_{-|\xi_{it}|}^{|\xi_{it}|} \big(\ell_{Z_i}(x+u) - \ell_{Z_i}(x)\big)du\Big)^2 dG(x) \\
&\leq \int_0^\infty \sum_{i=1}^n |\xi_{it}| \int_{-|\xi_{it}|}^{|\xi_{it}|} \big(\ell_{Z_i}(x+u) - \ell_{Z_i}(x)\big)^2 dudG(x) \\
&\leq \max_{1\leq i\leq n,t} |\xi_{it}|^{-1} \int_{-|\xi_{it}|}^{|\xi_{it}|} \int_0^\infty \big(\ell_{Z_i}(x+u) - \ell_{Z_i}(x)\big)^2 dG(x)du \sum_{i=1}^n |\xi_{it}|^2 \\
&= o_p(1).
\end{aligned}$$

Upon combining these facts with (75) we obtain $\sup_t M_5(t) = o_p(1)$, thereby proving (67) for $j = 5$. The proof for $j = 6$ is exactly similar.

Now consider M_1. Let $\xi_t(Z) := \nu_{nt}(Z) - \nu(Z)$. Then

$$\begin{aligned}
EM_1(t) :\leq &\int_{-\infty}^\infty E\big\| W_t(x,t) - W_t(x,0)\big\|^2 dG(x) \\
\leq\ & n^{-1}\sum_{i=1}^n E\Big(\|\dot{\nu}_{nt}(Z_i)\|^2 \int_{-\infty}^\infty \Big|L_{Z_i}(x+\xi_{it}) - L_{Z_i}(x)\Big| dG(x)\Big) \\
\leq\ & n^{-1}\sum_{i=1}^n E\Big(\|\dot{\nu}_{nt}(Z_i)\|^2 \int_{-\infty}^\infty \int_{-|\xi_{it}|}^{|\xi_{it}|} \ell_{Z_i}(x+u)dudG(x) \\
=\ & E\Big(\int_{-|\xi_t(Z)|}^{|\xi_t(Z)|} \|\dot{\nu}_{nt}(Z)\|^2 \int_{-\infty}^\infty \ell_Z(x+u)dG(x)\,du\Big) \to 0,
\end{aligned}$$

by (21). Thus

$$M_1(t) = o_p(1), \qquad \forall\, t \in \mathcal{U}(b). \tag{76}$$

To prove that this holds uniformly in $t \in \mathcal{U}(b)$, because of the compactness of the ball $\mathcal{U}(b)$, it suffices to show that for every $\epsilon > 0$ there is a $\delta > 0$ and an N_ϵ such that for every $s \in \mathcal{U}(b)$,

$$P\Big(\sup_{\|t-s\|<\delta} \|M_1(t) - M_1(s)| \geq \epsilon\Big) \leq \epsilon, \quad \forall n > N_\epsilon. \tag{77}$$

Let $\dot{\nu}_{ntj}(z)$ denote the jth coordinate of $\dot{\nu}_{nt}(z)$, $j = 1, \dots, q$ and let

$$\alpha_i(x,t) := I(\zeta_i \leq x + \xi_{it}) - I(\zeta_i \leq x) - L_{Z_i}(x + \xi_{it}) + L_{Z_i}(x).$$

Then

$$\begin{aligned} M_1(t) &= \int_0^\infty \big\|W_t(x,t) - W_t(x,0)\big\|^2 dG(x) \\ &= \sum_{j=1}^q \int_0^\infty \Big(n^{-1/2} \sum_{i=1}^n \dot{\nu}_{ntj}(Z_i)\alpha_i(x,t)\Big)^2 dG(x) = \sum_{j=1}^q M_{1j}(t), \quad \text{say.} \end{aligned}$$

Thus it suffices to prove (77) with M_1 replaced by M_{1j} for each $j = 1, \dots, q$.

Any real number a can be written as $a = a^+ - a^-$, where $a^+ = \max(0,a)$, $a^- = \max(0,-a)$. Note that $a^\pm \geq 0$. Fix a $j = 1, \dots, q$, write $\dot{\nu}_{ntj}(Z_i) = \dot{\nu}^+_{ntj}(Z_i) - \dot{\nu}^-_{ntj}(Z_i)$ and define

$$\begin{aligned} &W^\pm_j(x,t) := n^{-1/2} \sum_{i=1}^n \dot{\nu}^\pm_{ntj}(Z_i)\alpha_i(x,t), \\ &D^\pm_j(x,s,t) := W^\pm_j(x,t) - W^\pm_j(x,s), \quad R^\pm_j(s,t) := \int_0^\infty \big(D^\pm_j(x,s,t)\big)^2 dG(x). \end{aligned}$$

Then

$$\begin{aligned} &\big|M_{1j}(t) - M_{1j}(s)\big| \qquad (78) \\ &= \Big| \int_0^\infty \big(W^+_j(x,t) - W^-_j(x,t)\big)^2 dG(x) \\ &\qquad - \int_0^\infty \big(W^+_j(x,s) - W^-_j(x,s)\big)^2 dG(x) \Big| \\ &\leq \int_0^\infty \big(D^+_j(x,s,t)\big)^2 dG(x) + \int_0^\infty \big(D^-_j(x,s,t)\big)^2 dG(x) \\ &\quad + 2\Big\{ \int_0^\infty \big(D^+_j(x,s,t)\big)^2 dG(x) \int_0^\infty \big(D^-_j(x,s,t)\big)^2 dG(x) \Big\}^{1/2} \end{aligned}$$

$$
\begin{aligned}
&+2\Big[\Big\{\int_0^\infty \big(D_j^+(x,s,t)\big)^2 dG(x)\Big\}^{1/2} \\
&\qquad\qquad +\Big\{\int_0^\infty \big(D_j^-(x,s,t)\big)^2 dG(x)\Big\}^{1/2}\Big] M_{1j}^{1/2}(s) \\
&= R_j^+(s,t) + R_j^-(s,t) + 2\big(R_j^+(s,t)R_j^-(s,t)\big)^{1/2} \\
&\qquad\qquad +\big\{(R_j^+(s,t))^{1/2} + (R_j^-(s,t)\big)^{1/2}\big\} M_{1j}^{1/2}(s).
\end{aligned}
$$

Write

$$
\begin{aligned}
D_j^+(x,s,t) &= n^{-1/2}\sum_{i=1}^n \dot{\nu}_{ntj}^+(Z_i)\alpha_i(x,t) - n^{-1/2}\sum_{i=1}^n \dot{\nu}_{nsj}^+(Z_i)\alpha_i(x,s) \\
&= n^{-1/2}\sum_{i=1}^n \big[\dot{\nu}_{ntj}^+(Z_i) - \dot{\nu}_{nsj}^+(Z_i)\big]\alpha_i(x,t) \\
&\qquad\qquad + n^{-1/2}\sum_{i=1}^n \dot{\nu}_{nsj}^+(Z_i)\big[\alpha_i(x,t) - \alpha_i(x,s)\big] \\
&= D_{j1}^+(x,s,t) + D_{j2}^+(x,s,t), \quad \text{say.}
\end{aligned}
$$

Hence

$$
R_j^+(s,t) \le 2\int_0^\infty \big(D_{j1}^+(x,s,t)\big)^2 dG(x) + 2\int_0^\infty \big(D_{j2}^+(x,s,t)\big)^2 dG(x). \tag{79}
$$

By (23), the first term here satisfies (77). We proceed to verify it for the second term. Fix an $s \in \mathcal{U}_b$, $\epsilon > 0$ and $\delta > 0$. Let

$$
\Delta_{ni} := n^{-1/2}\big(\delta\|\dot{\nu}(Z_i)\| + 2\epsilon\big), \quad B_n := \Big\{\sup_{t\in\mathcal{N}_b,\|t-s\|\le\delta} \big|\xi_{it} - \xi_{is}\big| \le \Delta_{ni}\Big\}.
$$

By (18), there exists an N_ϵ such that $P(B_n) > 1-\epsilon$, for all $n > N_\epsilon$. On B_n, $\xi_{is} - \Delta_{ni} \le \xi_{it} \le \xi_{is} + \Delta_{ni}$ and, by the nondecreasing property of the indicator function and d.f., we obtain

$$
\begin{aligned}
&I(\zeta_i \le x + \xi_{is} - \Delta_{ni}) - I(\zeta_i \le x) - L_{Z_i}(x - \xi_{is} + \Delta_{ni}) + L_{Z_i}(x) \\
&\qquad\qquad - L_{Z_i}(x + \xi_{is} + \Delta_{ni}) + L_{Z_i}(x + \xi_{is} - \Delta_{ni}) \\
&\le \alpha_i(x,t) = I(\zeta_i \le x + \xi_{it}) - I(\zeta_i \le x) - L_{Z_i}(x + \xi_{it}) + L_{Z_i}(x) \\
&\le I(\zeta_i \le x + \xi_{is} + \Delta_{ni}) - I(\zeta_i \le x) - L_{Z_i}(x + \xi_{is} + \Delta_{ni}) + L_{Z_i}(x) \\
&\qquad\qquad + L_{Z_i}(x + \xi_{is} + \Delta_{ni}) - L_{Z_i}(x + \xi_{is} - \Delta_{ni}).
\end{aligned}
$$

Let

$$\mathcal{D}_{j2}^{\pm}(x,s,a) := n^{-1/2}\sum_{i=1}^{n}\dot{\nu}_{njs}^{\pm}\Big\{I(\zeta_i \le x+\xi_{is}+a\varDelta_{ni}) - I(\zeta_i \le x) \\ -L_{Z_i}(x+\xi_{is}+a\varDelta_{ni}) + L_{Z_i}(x)\Big\}.$$

The above inequalities and $\dot{\nu}_{njs}^{+}(Z_i)$ being nonnegative yield that on B_n,

$$\begin{aligned}&\int_0^\infty \big(D_{j2}^{+}(x,s,t)\big)^2 dG(x)\\ &\le \int_0^\infty \big(\mathcal{D}_{j2}^{+}(x,s,1)-\mathcal{D}_{j2}^{+}(x,s,0)\big)^2 dG(x)\\ &\quad+\int_0^\infty \big(\mathcal{D}_{j2}^{+}(x,s,-1)-\mathcal{D}_{j2}^{+}(x,s,0)\big)^2 dG(x)\\ &\quad+\int_0^\infty \Big(n^{-1/2}\sum_{i=1}^{n}\dot{\nu}_{njs}^{+}(Z_i)\big\{L_{Z_i}(x+\xi_{is}+\varDelta_{ni})\\ &\qquad\qquad -L_{Z_i}(x+\xi_{is}-\varDelta_{ni})\big\}dG(x)\Big)^2.\end{aligned}$$

Note that $\max_{1\le i\le n}(|\xi_{is}|+\varDelta_{ni}) = o_p(1)$. Argue as for (76) to see that the first two terms in the above bound are $o_p(1)$, while the last term is bounded from the above by

$$\begin{aligned}&\int_0^\infty \Big(n^{-1/2}\sum_{i=1}^{n}\dot{\nu}_{njs}^{+}(Z_i)\int_{\xi_{is}-\varDelta_{ni}}^{\xi_{is}+\varDelta_{ni}} \ell_{Z_i}(x+u)du\, dG(x)\Big)^2 \qquad (80)\\ &\le 2n^{-1}\sum_{i=1}^{n}(\dot{\nu}_{njs}^{+}(Z_i))^2\sum_{i=1}^{n}\varDelta_{ni}\int_{\xi_{is}-\varDelta_{ni}}^{\xi_{is}+\varDelta_{ni}}\int_0^\infty \big[\ell_{Z_i}^2(x+u)-\ell_{Z_i}^2(x)\big]\, dG(x)\, du\\ &\quad+4n^{-1}\sum_{i=1}^{n}(\dot{\nu}_{njs}^{+}(Z_i))^2\sum_{i=1}^{n}\varDelta_{ni}^2\int_0^\infty \ell_{Z_i}^2(x)dG(x).\end{aligned}$$

The first summand in the above bound is bounded above by

$$2\max_{1\le i\le n}(2\Delta_{ni})^{-1}\int_{\xi_{is}-\Delta_{ni}}^{\xi_{is}+\Delta_{ni}}\int_0^\infty\big[\ell^2_{Z_i}(x+u)-\ell^2_{Z_i}(x)\big]du\,dG(x)$$
$$\times n^{-1}\sum_{i=1}^n(\dot{\nu}^+_{njs}(Z_i))^2\sum_{i=1}^n\Delta^2_{ni}=o_p(1),$$

because the first factor tends to zero in probability by (20) and the second factor satisfies

$$n^{-1}\sum_{i=1}^n(\dot{\nu}^+_{njs}(Z_i))^2\sum_{i=1}^n\Delta^2_{ni}\le n^{-1}\sum_{i=1}^n\|\dot{\nu}_{ns}\|^2\big(2n^{-1}\delta^2\sum_{i=1}^n\|\dot{\nu}(Z_i)\|^2+4\epsilon^2\big).$$

The second term in the upper bound of (80) is bounded from the above by

$$4n^{-1}\sum_{i=1}^n\|\dot{\nu}_{ns}(Z_i)\|^2\,n^{-1}\sum_{i=1}^n\big(\delta^2\|\dot{\nu}(Z_i)\|^2+4\epsilon^2\big)\int_0^\infty\ell^2_{Z_i}(x)dG(x)$$
$$\to_p E\|\dot{\nu}(Z)\|^2\big[\delta^2\int_0^\infty E(\|\dot{\nu}(Z)\|^2\ell^2_Z(x))dG(x)+4\epsilon^2\int_0^\infty E(\ell^2_Z(x))dG(x)\big].$$

Since the factor multiplying δ^2 is positive, the above term can be made smaller than ϵ by the choice of δ. Hence (77) is satisfied by the second term in the upper bound of (79). This then completes the proof of R^+_j satisfying (77). The details of the proof for verifying (77) for R^-_j are exactly similar. These facts together with the upper bound of (78) show that (77) is satisfied by M_{1j} for each $j=1,\ldots,q$. This also completes the proof of $\sup_t M_1(t)=o_p(1)$, thereby proving (67) for $j=1$. The proof for $j=3$ is similar.

Next, consider M_2. Recall $\beta_i(x):=I(\zeta_i\le x)-L_{Z_i}(x)$. Then

$$M_2(t):=n^{-1}\int_0^\infty\big\|\sum_{i=1}^n\{\dot{\nu}_{nt}(Z_i)-\dot{\nu}(Z_i)\}\beta_i(x)\big\|^2dG(x).$$

Because $E(\beta_i(x)|Z_i)\equiv0$, a.s., we have

$$EM_2(t)=\int_0^\infty E\Big(\big\|\dot{\nu}_{nt}(Z)-\dot{\nu}(Z)\|^2L_Z(x)(1-L_Z(x)\Big)dG(x)\to0,$$

by (18). Thus

$$M_2(t) = o_p(1), \qquad \forall\, t \in \mathbb{R}^q. \tag{81}$$

To prove this holds uniformly in $t \in \mathcal{U}(b)$, we shall verify (77) for M_2. Accordingly, let $\delta > 0$, $s \in \mathcal{U}(b)$ be fixed. Then forall $t \in \mathcal{U}(b)$ such that $\|t - s\| < \delta$,

$$\begin{aligned}
&\big|M_2(t) - M_2(s)\big| \\
&\le n^{-1}\int_0^\infty \Big\| \sum_{i=1}^n \{\dot\nu_{nt}(Z_i) - \dot\nu_{ns}(Z_i)\}\beta_i(x)\Big\|^2 dG(x) \\
&\quad + 2\Big(n^{-1}\int_0^\infty \Big\| \sum_{i=1}^n \{\dot\nu_{nt}(Z_i) - \dot\nu_{ns}(Z_i)\}\beta_i(x)\Big\|^2 dG(x)\Big)^{1/2} M_2(s)^{1/2}
\end{aligned}$$

This bound, (24) and (81) now readily verifies (77) for M_2, which also completes the proof of (67) for $j = 2$. The proof of (67) for $j = 4$ is precisely similar. This in turn completes the proof of Lemma 6. □

Proof of (60). Recall (43). Let $D(z, x) := m_\theta(z + x) - \nu_\theta(z)$. Use the fact $\widehat{U}(x, \theta) = \tilde{W}(x, 0) + \tilde{W}(-x, 0)$, to rewrite

$$\begin{aligned}
\tilde{T}_n &= \frac{1}{N^2\sqrt{n}} \sum_{i=1}^n \sum_{j,k=1}^N \int \mu(z)\dot m_\theta(Z_i + \tilde\eta_j)\{\varphi_z(\zeta_i) - 2\ell_z(\zeta_i)D(Z_i, \tilde\eta_k)\}dQ(z) \\
&\quad + \text{ higher order terms},
\end{aligned}$$

where $\varphi_z(x)$ is defined as in Sect. 4.1.

For further analysis of $\tilde{T}_n$, with the two independent samples $\{(Z_i, \zeta_i), 1 \le i \le n\}$ and $\{\tilde\eta_k, 1 \le k \le N\}$, define the symmetric kernel function ϕ and its projections as follows.

$$\begin{aligned}
&\phi(Z_1, \zeta_1, \tilde\eta_1, \tilde\eta_2) \\
&\quad := \int \mu(z)\dot m_\theta(Z_1 + \tilde\eta_1)\{\varphi_z(\zeta_1) - 2\ell_z(\zeta_1)D(Z_1, \tilde\eta_2)\}dQ(z) \\
&\qquad + \int \mu(z)\dot m_\theta(Z_1 + \tilde\eta_2)\{\varphi_z(\zeta_1) - 2\ell_z(\zeta_1)D(Z_1, \tilde\eta_1)\}dQ(z) \\
&E(\phi|Z_1, \zeta_1) = 2\int \mu(z)\dot\nu_\theta(Z_1)\varphi_z(\zeta_1)dQ(z), \quad E\phi(Z_1, \zeta_1, \tilde\eta_1, \tilde\eta_2) = 0, \\
&E(\phi|\tilde\eta_1) = -2\int \mu(z)E\{\dot\nu(Z)\ell_z(\zeta)D(Z, \tilde\eta_1)|\tilde\eta_1\}dQ(z).
\end{aligned}$$

Let $\tilde{T}_{n1}$ denote the first term in the right hand side of $\tilde{T}_n$. Then

$$
\begin{aligned}
\tilde{T}_{n1} &= \frac{1}{N^2\sqrt{n}}\sum_{i=1}^{n}\sum_{1\le j<k\le N}\phi(Z_i,\zeta_i,\tilde{\eta}_j,\tilde{\eta}_k)\\
&\quad+\frac{1}{N^2\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1}^{N}\int\mu(z)\dot{m}_\theta(Z_i+\tilde{\eta}_j)\{\varphi_z(\zeta_i)-2\ell_z(\zeta_i)D(Z_i,\tilde{\eta}_j)\}dQ(z)\\
&=: \tilde{T}_{n11}+\tilde{T}_{n12}.
\end{aligned}
$$

Note that that $\tilde{T}_{n11}$ is a U-statistic with permutation degree 1 in the primary sample $\{(Z_i,\zeta_i), 1\le i\le n\}$ and permutation degree 2 in the validation sample $\{\tilde{\eta}_k, 1\le k\le N\}$. Theorem 6.1.4 in Lehmann [17] and (52) yield that, for $0\le\lambda<\infty$,

$$
\begin{aligned}
\tilde{T}_{n11} &= \sqrt{n}\times\frac{N(N-1)}{2N^2}\times\frac{1}{\binom{n}{1}\binom{N}{2}}\sum_{i=1}^{n}\sum_{1\le j<k\le N}\phi(Z_i,\zeta_i,\tilde{\eta}_j,\tilde{\eta}_k)\\
&\to_D \frac{1}{2}N\Big(0,\,\mathrm{Var}(E(\phi|Z_1,\zeta_1))+4\,\lambda\,\mathrm{Var}(E(\phi|\tilde{\eta}_1))\Big)=N(0,\,\Sigma_\theta+4\lambda\Sigma_1).
\end{aligned}
$$

Moreover, for $\lambda=\infty$, Theorem 6.1.4 in Lehmann [17] also yields that $\sqrt{N/n}\,\tilde{T}_{n11}\to_D N(0,4\Sigma_1)$.

Similarly, $\tilde{T}_{n12}$ is a U-statistic with permutation degree 1 for both samples. Since (52) implies that $E\{\|\dot{m}_\theta(X)[m_\theta(X)-\nu_\theta(Z)]\|\}<\infty$, therefore $E\tilde{T}_{n12}=O(n^{-1/2})$. Moreover, Theorem 6.1.3 of U-statistics in Lehmann [17] implies that $\mathrm{Var}(\tilde{T}_{n12})=O(n^{-1})$ and hence $\tilde{T}_{n12}=o_p(1)$. Hence the claim (60).

Proof of (66). Let $\dot{D}_{ijk}:=\dot{m}_\theta(Z_i+\tilde{\eta}_k)-\dot{m}_\theta(Z_j+\tilde{\eta}_k)$. Based on the definitions of $\widehat{\Gamma}_\theta(u)$ and $\kappa_z(v)$, $T_{n,R}$ can be rewritten as

$$
\begin{aligned}
\tilde{T}_{n,R} &= \int\int_0^1 \mu^c(z)\tilde{\mathcal{U}}_R(u)\ell_z(L_z^{-1}(u))d\Psi(u)dQ(z)\\
&= -n^{-1/2}\sum_{i=1}^{n}\int\mu^c(z)\hat{\dot{\nu}}^c(Z_i)\kappa_z(L_{Z_i}(\zeta_i-\Delta(Z_i)))dQ(z)\\
&= -\frac{n^{-1/2}}{nN}\sum_{i=1}^{n}\sum_{j=1,j\ne i}^{n}\sum_{k=1}^{N}\int\mu^c(z)\dot{D}_{ijk}\kappa_z(L_{Z_i}(\zeta_i))dQ(z)\\
&\quad-\frac{n^{-1/2}}{nN}\sum_{i=1}^{n}\sum_{j=1,j\ne i}^{n}\sum_{k=1}^{N}\int\mu^c(z)\dot{D}_{ijk}\ell_z(\zeta_i)\Delta(Z_i)dQ(z)\\
&\quad+\text{ higher order} := T_{n,R1}+T_{n,R2}+\text{ higher order.}
\end{aligned}
$$

First, we study the asymptotic distribution of $T_{n,R1}$. Define for $1\le i,j\le n, i\ne j$, and $1\le k\le N$,

$$\psi_1(Z_i, \zeta_i, Z_j, \zeta_j, \tilde{\eta}_k) := \int \mu^c(z)\dot{D}_{ijk}\Big\{\kappa_z(L_{Z_i}(\zeta_i)) + \kappa_z(L_{Z_j}(\zeta_j))\Big\}dQ(z).$$

Then $T_{n,R1}$ can be rewritten as

$$T_{n,R1} = -\frac{n(n-1)}{2n^2} \times \frac{\sqrt{n}}{\binom{n}{2}\binom{N}{1}} \sum_{1\le i<j\le n} \sum_{k=1}^{N} \psi_1(Z_i, \zeta_i, Z_j, \zeta_j, \tilde{\eta}_k).$$

By the definition of U-statistics in Lehmann [17], $T_{n,R1}$ is a two sample U-statistic based on function ψ_1 with permutation degree of 2 on the sample $\{(Z_i, \zeta_i), 1 \le i \le n\}$ and permutation degree of 1 on the sample $\{\tilde{\eta}_k, 1 \le k \le N\}$. Because conditionally, $L_{Z_i}(\zeta_i)$, given Z_i, is a uniformly distributed r.v., we have $E(L_{Z_i}(\zeta_i)|Z_i) = \int_0^1 \kappa_z(u)du := K(z)$. Then the conditional expectations of ψ can be calculated as follows.

$$\begin{aligned} E(\psi_1|Z_1, \zeta_1) &= \int \mu^c(z)[\dot{\nu}_\theta(Z_1) - E(\dot{\nu}_\theta(Z))]\kappa_z^c(L_{Z_1}(\zeta_1))dQ(z), \\ E(\psi_1|Z_1, Z_2, \tilde{\eta}_1) &= \int \mu^c(z)\big\{\dot{D}_{121} + \dot{D}_{211}\big\}K(z)dQ(z) = 0, \\ E(\psi_1|\tilde{\eta}_1) &= 0. \end{aligned}$$

It can be seen that $\mathrm{Cov}(E(\psi_1|Z_1, \zeta_1)) = \widehat{\Sigma}_\theta$ as defined in Sect. 4.2. Then Theorem 6.1.4 in Lehmann [17] yields that

$$T_{n,R1} \to_D \frac{1}{2}N(0, 4\mathrm{Cov}(E(\psi|Z_1, \zeta_1))) = N(0, \widehat{\Sigma}_\theta).$$

Next, in order to study $T_{n,R2}$, define

$$\begin{aligned} &\psi_2(Z_i, \zeta_i, Z_j, \zeta_j, \tilde{\eta}_k, \tilde{\eta}_l) \\ &= \int \mu^c(z)[\dot{m}_\theta(Z_i + \tilde{\eta}_k) - \dot{m}_\theta(Z_j + \eta_k)]\ell_z(\zeta_i)D(Z_i, \tilde{\eta}_l)dQ(z) \\ &\quad + \int \mu^c(z)[\dot{m}_\theta(Z_j + \tilde{\eta}_k) - \dot{m}_\theta(Z_i + \eta_k)]\ell_z(\zeta_j)D(Z_j, \tilde{\eta}_l)dQ(z) \\ &\quad + \int \mu^c(z)[\dot{m}_\theta(Z_i + \tilde{\eta}_l) - \dot{m}_\theta(Z_j + \eta_l)]\ell_z(\zeta_i)D(Z_i, \tilde{\eta}_k)dQ(z) \\ &\quad + \int \mu^c(z)[\dot{m}_\theta(Z_j + \tilde{\eta}_l) - \dot{m}_\theta(Z_i + \eta_l)]\ell_z(\zeta_j)D(Z_j, \tilde{\eta}_k)dQ(z). \end{aligned}$$

Then $T_{n,R2}$ can be rewritten as a two sample U-statistic with permutation degree 2 for both primary sample and validation sample.

$$T_{n,R2}$$
$$= -\frac{n(n-1)}{2n^2}\,\frac{N(N-1)}{2N^2}\,\frac{\sqrt{n}}{\binom{n}{2}\binom{N}{2}}\sum_{1\le i<j\le n}\;\sum_{1\le k<l\le N}\psi_2(Z_i,\zeta_i,Z_j,\zeta_j,\tilde{\eta}_k,\tilde{\eta}_l).$$

The conditional expectations of ψ_2 are calculated as

$$E(\psi_2|\tilde{\eta}_1) = 2\int \mu^c(z)E\big\{[\dot{\nu}_\theta(Z) - E(\dot{\nu}_\theta(Z))]\ell_z(\zeta)D(Z,\tilde{\eta}_1)dQ(z)$$
$$E(\psi_2|Z_1,\zeta_1) = 0.$$

Then Theorem 6.1.4 in Lehmann [17] shows that, for $0 \le \lambda < \infty$,

$$T_{n,R2} \to_D \frac{1}{4}N(0, 4\lambda\mathrm{Cov}(E(\psi_2|\tilde{\eta}_1))) = N(0, \lambda\Sigma_2).$$

The two terms $T_{n,R1}$ and $T_{n,R2}$ are asymptotically independent becuase of the independence between the primary sample and validation sample. In fact, $T_{n,R1}$ is based on $E(\psi_1|Z_1,\zeta_1)$ and $T_{n,R2}$ is based on $E(\psi_2|\tilde{\eta}_1)$. Therefore, (66)(a) holds. An argument similar to one used for (51) yields that $\sup_{\|t\|\le b}|\tilde{\mathcal{K}}(\theta + n^{-1/2}t) - \tilde{\mathcal{K}}_R(t)| = o_p(1)$, which in turn yields the claim (66)(b) about $\tilde{\theta}_R$.

When $\lambda = \infty$, by Theorem 6.1.4 in Lehmann [17], $\sqrt{N/n}\,T_{n,R2} \to_D N(0, \Sigma_2)$. Then $\sqrt{N/n}\,\tilde{T}_{n,R} = \sqrt{N/n}\,\tilde{T}_{n,R1} + \sqrt{N/n}\,\tilde{T}_{n,R2} \to_D N(0, \Sigma_2)$. Therefore, we obtain that $\sqrt{N}(\tilde{\theta}_R - \theta) \to_D N(0, \widehat{\Omega}_\theta^{-1}\Sigma_2\widehat{\Omega}_\theta^{-1})$ for $\lambda = \infty$. □

References

1. Bates, D.M., Watts, D.G.: Nonlinear Regression Analysis and Its Applications. Wiley, New York (1998)
2. Beran, R.J.: Minimum Helinger distance estimates for parametric models. Ann. Statist. **5**, 445–463 (1977)
3. Beran, R.J.: An efficient and robust adaptive estimator of location. Ann. Statist. **6**, 292–313 (1978)
4. Berkson, J.: Are these two regressions? J. Amer. Statist. Assoc. **5**, 164–180 (1950)
5. Carroll, R.J., Ruppert, D., Stefanski, L.A., Crainiceanu, C.P.: Measurement Error in Nonlinear Models: A Modern Perspective, 2nd edn. Chapman & Hall/CRC, Bota Raton, FL (2006)
6. Cheng, C.L., Van Ness, J.W.: Statistical Regression with Measurement Error. Wiley, New York (1999)
7. Donoho, D.L., Liu, R.C.: Pathologies of some minimum distance estimators. Ann. Statist. **16**, 587–608 (1988a)
8. Donoho, D.L., Liu, R.C.: The "automatic" robustness of minimum distance functionals. Ann. Statist. **16**, 552–586 (1988b)
9. Fuller, W.A.: Measurement Error Models. Wiley, New York (1987)
10. Hájek, J.: Nonparametric Statistics. Holden Day, San Francisco, USA (1969)
11. Hodges Jr., J.L., Lehmann, E.L.: Estimates of location based on rank tests. Ann. Math. Statist. **34**, 598–611 (1963)

12. Koul, Hira L.: Weighted empirical processes and the regression model. J. Indian Statist. Assoc. **17**, 83–91 (1979)
13. Koul, Hira L.: Minimum distance estimation in multiple linear regression. Sankhya Ser. A., 47. Part **1**, 57–74 (1985a)
14. Koul, Hira L.: Minimum distance estimation in linear regression with unknown errors. Statist. Prob. Lett. **3**, 1–8 (1985b)
15. Koul, Hira L.: Asymptotics of some estimators and sequential residual empiricals in non-linear time series. Ann. Statist. **24**, 380–404 (1996)
16. Koul, H.L.: Weighted Empirical Processes in Dynamic Nonlinear Models. Lecture Notes Series in Statistics, 2nd edn., vol. 166. Springer, New York (2002)
17. Lehmann, E.L.: Elements of Large-Sample Theory. Springer, New York, N.Y., USA (1999)
18. Rudin, W.: Real and Complex Analysis, 2nd edn. McGraw-Hill, New York (1974)
19. Yi, G.: Statistical analysis with measurement error or misclassification. Strategy, method and application. With a foreword by Raymond J. Carroll. In: Springer Series in Statistics. Springer, New York (2017)

Implied Volatility Surface Estimation via Quantile Regularization

Matúš Maciak, Michal Pešta, and Sebastiano Vitali

Abstract The implied volatility function and the implied volatility surface are both key tools for analyzing financial and derivative markets and various approaches were proposed to estimate theses quantities. On the other hand, theoretical, practical, and also computational pitfalls occur in most of them. An innovative estimation method based on an idea of a sparse estimation and an atomic pursuit approach is introduced to overcome some of these limits: the quantile LASSO estimation implies robustness with respect to common market anomalies; the panel data structure allows for a time dependent modeling; changepoints introduce some additional flexibility in order to capture some sudden changes in the market and linear constraints ensure the arbitrage-free validity; last but not least, the interpolated implied volatility concept overcomes the problem of consecutive maturities when observing the implied volatility over time. Some theoretical backgrounds for the quantile LASSO estimation method are presented, the idea of the interpolated volatilities is introduced, and the proposed estimation approach is applied to estimate the implied volatility of the Erste Group Bank AG call options quoted in EUREX Deutschland Market.

Keywords Quantile regression · LASSO regularization · Panel data
Implied volitility · Options

M. Maciak (✉) · M. Pešta · S. Vitali
Faculty of Mathematics and Physics, Department of Probability and Mathematical Statistics,
Charles University, Prague, Czech Republic
e-mail: Matus.Maciak@mff.cuni.cz

M. Pešta
e-mail: Michal.Pesta@mff.cuni.cz

S. Vitali
e-mail: sebastiano.vitali@unibg.it

S. Vitali
University of Bergamo, Department of Management, Economics and Quantitative Methods,
Bergamo, Italy

M. Maciak et al. (eds.), *Analytical Methods in Statistics*, Springer Proceedings
in Mathematics & Statistics 329, https://doi.org/10.1007/978-3-030-48814-7_4

1 Introduction

The empirical econometrics and financial experts rely on many different analytical tools. For instance, considering option tradings and various derivative markets, the most fundamental tools are the option pricing strategies and the implied volatility estimation (see [3]). The most common approaches are usually based on the well-known Black-Scholes model introduced in [2] even despite the fact that this model is considered by practitioners to be unrealistic from the theoretical point of view. Many alternative approaches were, therefore, proposed in order to overcome some obvious drawbacks of the Black-Scholes model.

Semiparametric or nonparametric option pricing approaches are commonly considered instead (see, for instance, [1, 6], or [10]) while the arbitrage-free market validity is guaranteed by using some additional pre-defined shape constraints. On the other hand, the corresponding implied volatility function (or the implied volatility surface respectively) is usually obtained either similarly, in terms of some constrained optimization problem (for instance, [11, 19]), or alternatively, it can be interpolated directly from the estimated option pricing model (see [6] or [9]).

In this paper we focus on the implied volatility surface estimation (the implied volatility function which evolves in time over some fixed observational period) and we advocate and combine various ideas to construct the overall model: sparse estimation with LASSO regularization and changepoints, quantile regression with panel data structure, or interpolated implied volatility values with a constant maturity. From the theoretical point of view, the presented method is motivated by the concept of a regularized changepoint detection proposed in [7] and further elaborated for the conditional quantile estimation in [4, 5]. A similar idea of the sparse estimation was also recently presented in [18] to estimate the option price function using a standard squared loss objective function while the conditional quantile estimation approach was proposed in [12]. The quantile estimation is, in general, considered to be robust and it also offers a more complex insight into the underlying data as it can estimate any arbitrary conditional quantile rather than just the conditional mean. The robustness property is also useful as the final model is not too sensitive with respect to various market anomalies (such as bid-ask spreads, discrete ticks in price, non-synchronous trading, etc.). The panel data structure allows for a time dependent modeling and changepoints introduce some additional flexibility which is convenient for reflecting some occasional sudden changes in the market (caused by various financial, economical, political, or natural causes). Finally, the natural evolution of the implied volatility over time, from the issuing date of the option until its maturity, shows some increase in convexity of the implied volatility smile. In order to avoid this issue and to focus on the changes that are due to some exogenous effects the implied volatility of an artificial option with a constant maturity of 30 days is introduced. Such implied volatility is computed by interpolating the implied volatility of options at consecutive maturities. The artificial options with the constant maturity are later used to estimated the corresponding implied volatility surface.

The rest of the paper is organized as follows: the quantile LASSO model is described in Sect. 2 and two important model modifications for estimating a single implied volatility function or the overall time-dependent two-dimension surface are presented in Sect. 3. Both situations are considered: a model without arbitrage-free restrictions and also a model which complies with the financial theory on the arbitrage-free market scenarios. Finally, in Sect. 4, the model is applied to estimate the time dependent profile of the implied volatility function for the Erste Group Bank AG call options quoted in the EUREX Deutschland market and some inference tools are used to decide whether there is some significant sudden change over the given profile or not.

2 Quantile LASSO Model for Implied Volatilities

Let us firstly briefly summarize the idea of using the quantile estimation and the LASSO type regularization for the regression estimation in general. A standard linear regression model where, in addition, the unknown vector parameter can change along the available observations $i \in \{1, \ldots, n\}$, which are somehow naturally ordered, can be expressed as

$$Y_i = \mathbf{x}_i^\top \boldsymbol{\beta}_i + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{\beta}_i \in \mathbb{R}^p$ is a p-dimensional parameter (the dimension does not depend on $n \in \mathbb{N}$) and $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^\top$ is the subject's specific vector of covariates. The random error terms $\{\varepsilon_i\}_{i=1}^n$ are usually independent, centered, and identically distributed with some unknown distribution function F. It is also assumed that there is some form of sparsity in the unknown parameters $\boldsymbol{\beta}_i$'s, such that $\boldsymbol{\beta}_i = \boldsymbol{\beta}_{i-1}$, for most of the indexes $i \in \{2, \ldots, n\}$, but some few exceptions—changepoints. The model in (1) can be seen as a straightforward generalization of a simpler piece-wise constant model from [7] or, from the econometrics perspective, a more common trend model in [13]. The same model as in (1), however, for the dependent time series data, is also considered in [17].

The model in (1) is assumed to have $K^* \in \mathbb{N}$ changepoints in total, located at some unknown indexes $t_1^* < \cdots < t_{K^*}^* \in \{1, \ldots, n\}$, such that

$$\boldsymbol{\beta}_i = \boldsymbol{\beta}_{t_k^*}, \qquad \forall i = t_k^*, t_k^* + 1, \ldots, t_{k+1}^* - 1, \qquad k = 0, 1, \ldots, K^*, \tag{2}$$

with $t_0^* = 1, t_{K^*+1}^* = n$, and $\boldsymbol{\beta}_n = \boldsymbol{\beta}_{t_{K^*+1}^*}$. In general, the number of true changepoints $K^* \in \mathbb{N}$ and their locations $t_1^*, \ldots, t_{K^*}^*$ are all unknown. The true values of $\boldsymbol{\beta}_i$ are denoted by $\boldsymbol{\beta}_i^*$ and $K^* \equiv Card\{i \in \{2, \ldots, n\};\ \boldsymbol{\beta}_i^* \neq \boldsymbol{\beta}_{i-1}^*\}$. The idea of the estimation method is to recover the unknown changepoint locations and to estimate the underlying model phases—the vector parameters which are associated with the conditional quantiles of interest. For this purpose, the following optimization problem is formulated

$$\widehat{\boldsymbol{\beta}^n} = \underset{\substack{\boldsymbol{\beta}_i \in \mathbb{R}^p \\ \text{i = 1, …, n}}}{\text{Argmin}} \quad \sum_{i=1}^{n} \rho_\tau(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_i) + n\lambda_n \sum_{i=2}^{n} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_{i-1}\|_2, \tag{3}$$

where, for simplicity, $\widehat{\boldsymbol{\beta}^n} = (\widehat{\boldsymbol{\beta}}_1^\top, \ldots, \widehat{\boldsymbol{\beta}}_n^\top)^\top \in \mathbb{R}^{np}$, $\rho_\tau(u) = u(\tau - \mathbb{I}_{\{u<0\}})$, for $\tau \in (0, 1)$, is the standard check function used for the quantile regression, $\|\cdot\|_2$ stands for the classical L_2 norm, and $\lambda_n > 0$ is the tuning parameter which controls for the overall number of changepoints (the sparsity level) occurring in the final model: for $\lambda_n \to 0$ there will be $\widehat{\boldsymbol{\beta}}_i \neq \widehat{\boldsymbol{\beta}}_{i-1}$ for each $i \in \{2, \ldots, n\}$, while for $\lambda_n \to \infty$ no changepoints are expected to occur in the final model and, thus, $\widehat{\boldsymbol{\beta}}_i = \widehat{\boldsymbol{\beta}}_{i-1}$ for all $i \in \{2, \ldots, n\}$. The corresponding estimators for the changepoint locations are the observations $i \in \{2, \ldots, n\}$, where $\widehat{\boldsymbol{\beta}}_i \neq \widehat{\boldsymbol{\beta}}_{i-1}$. Let us, therefore, define the set

$$\widehat{\mathcal{A}}_n \equiv \{i \in \{2, \ldots, n\};\ \widehat{\boldsymbol{\beta}}_i \neq \widehat{\boldsymbol{\beta}}_{i-1}\} = \{\hat{t}_1 < \cdots < \hat{t}_{|\widehat{\mathcal{A}}_n|}\}, \tag{4}$$

and let $|\widehat{\mathcal{A}}_n|$ be the cardinality of $\widehat{\mathcal{A}}_n$. For each $k = 0, \ldots, |\widehat{\mathcal{A}}_n|$ we can also define the $(k+1)$-st model phase (observations indexed by the set $\{\hat{t}_k, \ldots, \hat{t}_{k+1} - 1\}$, where $\hat{t}_0 = 1$ and $\hat{t}_{\widehat{\mathcal{A}}_n+1} = n$), with the corresponding vector of estimated parameters $\widehat{\boldsymbol{\beta}}_{\hat{t}_k}$. The minimization problem formulated in (1) is convex and it can be effectively solved by using some standard optimization toolboxes (see, for instance, [8]). The theoretical properties are studied in detail in [5]. Under some reasonable assumptions, the method is consistent in terms of the changepoint detection and, also, in terms of the parameter estimation. Nevertheless, the regularization parameter in the LASSO problems should be chosen, in general, differently when aiming at the changepoint recovery or the underlying model estimation: for the former one, larger values are preferred to avoid the overestimation issue and false changepoint detection while for the estimation purposes, slightly smaller values of $\lambda_n > 0$ are needed in order to limit the shrinkage effect and to improve the estimation bias performance. The value of $\lambda_n > 0$ which satisfies the set of assumptions used in [5] is, for instance, $\lambda_n = (1/n) \cdot (\log n)^{5/2}$.

The role of the regularization parameter is crucial and various approaches can be used to determine a proper value for a given data. However, its importance can be suppressed by using some alternative regularization source. This is, for instance, also the case for the option pricing problem where the final model must satisfy some pre-defined shape restrictions in order to comply with the financial theory on the arbitrage-free market scenarios. In the next sections we present two important modifications of the quantile LASSO model which can be directly used to estimate the implied volatility function and the implied volatility surface respectively. Both these quantities serve as key tools for analyzing financial markets and derivative tradings in general.

3 Quantile LASSO and Implied Volatility Estimation

Firstly, we consider a situation where the implied volatility values are only observed for some specific day from the observational period. The data can be represented as a sample $\{(Y_i, x_i);\ i = 1, \ldots, n\}$, where Y_i stands for the observed implied volatility at the given strike x_i. Thus, there are $n \in \mathbb{N}$ observations in total for n unique strikes. The aim is to use the observed volatility values and to estimate the implied volatility function. The quantile fused LASSO presented in Sect. 2 is used, however, some modifications are needed in order to obtain the model which complies with the arbitrage free market conditions. Theses conditions imply that the estimated implied volatility function must be convex with respect to the strikes.

The quantile LASSO method provides a robust estimate which is not sensitive to various derivative market anomalies (such as bid-ask spreads, discrete ticks in price, or non-synchronous trading, or heavy tailed error distributions). The value of $\tau = 0.5$ is used to construct the conditional median in (3), which is, from the theoretical point of view, for a symmetric density of the error terms, the same quantity as the conditional mean. Nevertheless, the convexity of the final estimate is not automatically guaranteed in (3) and some additional linear constraints can be used to enforce the volatility smile in the final model.

3.1 Arbitrage-Free Market Restrictions

Let the available strikes $\{x_i\}_{i=1}^n$ be all from some compact domain $\mathcal{D}$ with some functional basis $\{\varphi_j(x);\ j = 1, \ldots, p\}$ defined on $\mathcal{D}$. Let $\boldsymbol{x}_i = (\varphi_1(x_i), \ldots, \varphi_p(x_i))^\top$ (for instance, let $\boldsymbol{x}_i = (1, x_i, x_i^2)^\top$, for $i = 1, \ldots, n$, where $p = 3$, which gives a standard quadratic fit). For each strike the quantile fused LASSO in (3) assumes the corresponding vector of unknown parameters $\boldsymbol{\beta}_i = (\beta_{i1}, \beta_{i2}, \beta_{i3})^\top \in \mathbb{R}^3$. This brings a huge amount of flexibility and the final model would be too much haphazard if no additional restrictions on the parameter vectors were imposed. Therefore, the regularization penalty in (3) is adopted. Another form of regularization can be applied if, for instance, some specific properties for the final fit are assumed (e.g., the convexity or the volatility smile respectively).

For $\mathbf{0} \in \mathbb{R}^p$ being a zero vector of the length $p \in \mathbb{N}$, we can easily define the model matrix

$$\mathbb{X} = \begin{bmatrix} \mathbf{x}_1^\top & \mathbf{0}^\top & \cdots & \mathbf{0}^\top \\ \mathbf{0}^\top & \mathbf{x}_2^\top & \cdots & \mathbf{0}^\top \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}^\top & \mathbf{0}^\top & \ldots & \mathbf{x}_n^\top \end{bmatrix}$$

and the model from (1) can be equivalently expressed as

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta}^n + \boldsymbol{\varepsilon}, \tag{5}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$, and $\boldsymbol{\beta}^n = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_n^\top)^\top \in \mathbb{R}^{np}$.

For the model in (5) we can directly use the minimization formulation in (3) but the solution is, in general, not smooth and the volatility smile required for the arbitrage-free market scenario is also not automatically guaranteed. The implied volatility function is assumed to be smooth and convex which can be both enforced by minimizing (3) with respect to some specific linear constraints defined for the functional basis $\{\varphi_j(x);\;\; j = 1, \ldots, p\}$. The overall smoothness property can be achieved by the right choice of the functional basis (e.g., polynomials, or splines of some specific degree, which is large enough). Moreover, it is easy to see, that the solution in (3) will be also convex if, in addition, the estimated vector of parameters $\widehat{\boldsymbol{\beta}^n} \in \mathbb{R}^{np}$ obeys

$$\boldsymbol{C}\widehat{\boldsymbol{\beta}^n} \geq \boldsymbol{0}, \tag{6}$$

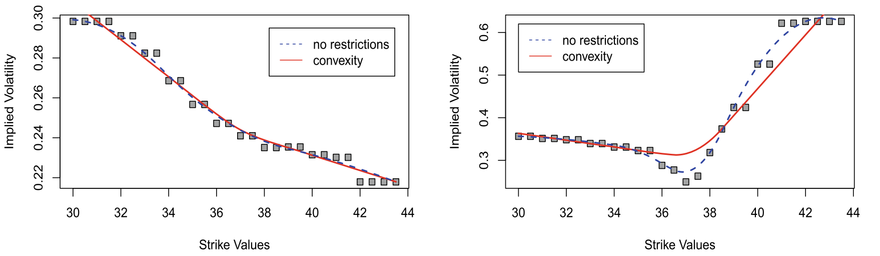
which holds element-wise for

$$\boldsymbol{C} = \begin{bmatrix} \widetilde{\mathbf{x}}_1^\top & \mathbf{0}^\top & \cdots & \mathbf{0}^\top \\ \mathbf{0}^\top & \widetilde{\mathbf{x}}_2^\top & \cdots & \mathbf{0}^\top \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}^\top & \mathbf{0}^\top & \ldots & \widetilde{\mathbf{x}}_n^\top \end{bmatrix},$$

where $\widetilde{\mathbf{x}}_i = (\varphi_1''(x_i), \ldots, \varphi_p{''}(x_i))^\top$ denotes the vector of the second derivatives of the functional basis functions $\varphi_j(x)$ for $j = 1, \ldots, p$ which are evaluated again at the given strike $x_i \in \mathcal{D}$, for $i = 1, \ldots, n$. The minimization (3) together with the linear constraints given in (6) is again convex and an effective solution can be obtained by adopting some standard optimization toolboxes.

For illustration, the quantile fused LASSO is applied for the Erste Group Bank AG call options quoted in EUREX Deutschland Market and the implied volatility function is estimated for two specific trading days—September 21st, 2018 and October 18th, 2018 (see Fig. 1 for illustration). It is clear from Fig. 1 that the arbitrage-free conditions are not automatically guaranteed by the data themselves and, indeed, the convexity property (volatility smile) must be enforced by the linear constraints in (6).

In practical applications, the estimated implied volatility function can change over time reflecting various trends or anomalies on the derivative market. Therefore, in the next section, we introduce another modification of the quantile LASSO model described above in order to estimate the implied volatility function for a set of consecutive days from some fixed period. We assume $n \in \mathbb{N}$ independent panels (one panel for each strike) and the strike specific implied volatilities are observed over some trading interval $[0, T]$, for some fixed $T > 0$.



(a) Implied volatility for $t = 1$ (September 21st, 2018) (b) Implied volatility for $t = 20$ (October 18th, 2018)

Fig. 1 The illustration of the Quantile LASSO performance when applied for the estimation of the implied volatility function. Two situations are considered: the first day of the observational period on the left panel, and $t = 20$ (volatility peak) on the right panel. The estimation is considered without any linear restrictions (dashed blue lines) and with the linear constraints—which enforce the volatility smile (solid red lines) and, thus, arbitrage-free validity

3.2 Time Dependent Implied Volatility Surface

The implied volatility values are now represented as a sample $\{(Y_{ti}, x_{it});\ t = 1, \dots, T;\ i = 1, \dots, n\}$, where Y_{it} stands for the implied volatility at some specific time $t \in \{1, \dots, T\}$ and the given strike $x_{it} \in \mathcal{D}$, for $i \in \{1, \dots, n\}$. For simplicity, the quoted strikes are common over time, therefore, we have that $x_{it} \equiv x_i$, for $i = 1, \dots, n$. For each quoted strike $x_i \in \mathcal{D}$ there is a strike specific panel of the implied volatilities observed over time $t \in \{1, \dots, T\}$. The value of $T \in \mathbb{N}$ represents, for instance, the number of trading days available in the data. The underlying panel data model takes the form

$$Y_{ti} = \boldsymbol{x}_i^\top \boldsymbol{\beta}_t + \varepsilon_{ti}, \quad \text{for } t = 1, \dots, T \text{ and } i = 1, \dots, n, \tag{7}$$

where again $\boldsymbol{x}_i = (\varphi_1(x_i), \dots, \varphi_p(x_i))^\top$ is the given functional basis on $\mathcal{D}$, and $\boldsymbol{\beta}_t = (\beta_{t1}, \dots, \beta_{tp})^\top \in \mathbb{R}^p$ is the vector of unknown parameters which can now also change over time $t \in \{1, \dots, T\}$. The error vectors $\boldsymbol{\varepsilon}_i = [\varepsilon_{1i}, \dots, \varepsilon_{Ti}]$ are assumed to be independently distributed across panels $i \in \{1, \dots, n\}$.

The time dependent implied volatility surface can be now estimated simultaneously, such that the final model will obey the shape restrictions required for the arbitrage-free market. The corresponding minimization problem takes the form

$$\underset{\substack{\boldsymbol{\beta}_t \in \mathbb{R}^p \\ \text{t = 1, ..., T}}}{\text{Minimize}} \quad \sum_{t=1}^{T} \sum_{i=1}^{n} \rho_\tau \Big(Y_{ti} - \boldsymbol{x}_i^\top \boldsymbol{\beta}_t \Big) + n\lambda_n \sum_{t=2}^{T} \|\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1}\|_2 \tag{8}$$

with respect to

$$\boldsymbol{C}\boldsymbol{\beta}_t \geq \boldsymbol{0}, \; t = 1, \ldots, T; \qquad \textit{(convexity in the strike over time)} \qquad (9)$$

where $\boldsymbol{C}$ is defined analogously as in (6). The overall vector of the estimated parameters $\widehat{\boldsymbol{\beta}^n} = (\widehat{\boldsymbol{\beta}}_1^\top, \ldots, \widehat{\boldsymbol{\beta}}_T^\top)^\top \in \mathbb{R}^{T \times p}$ represents the set of all panels while $\widehat{\boldsymbol{\beta}}_t \in \mathbb{R}^p$ is only associated with the estimated volatility function for some specific time $t \in \{1, \ldots, T\}$. The implied volatility function is obviously allowed to evolve over time to reflect possible changes on the market with no specific restrictions what so ever. Moreover, there is again a specific sparsity structure assumed: for situations where $\widehat{\boldsymbol{\beta}}_t \neq \widehat{\boldsymbol{\beta}}_{t-1}$ the estimated implied volatility function changes from time $(t-1)$ to time t to adapt for the situation at the market and, otherwise, the estimated implied volatility remains the same. The regularization parameter in (8) controls the amount of such changes and the shape constraints in (9) are responsible for the additional source of regularization by enforcing convexity of the estimated volatility function for each time point $t \in \{1, \ldots, T\}$. The minimization problem in (8) together with the linear constraints in (9) is again convex and the optimal solution can be obtained by the standard optimization software. The Karush-Kuhn-Tucker (KKT) optimality conditions can be easily derived and they are formulated by the following lemma.

Lemma 1 (a) *For any $l \in \{1, \ldots, |\widehat{\mathcal{A}}_n|\}$, $n \in \mathbb{N}$, and $\lambda_n > 0$ the following holds with probability one:*

$$\tau(T - \hat{t}_l + 1) \sum_{i=1}^{n} \mathbf{x}_i - \sum_{i=1}^{n} \sum_{k=\hat{t}_l}^{T} \mathbf{x}_i \mathbb{1}_{\{Y_{ik} \leq \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_k\}} = n\lambda_n \frac{\widehat{\boldsymbol{\theta}}_{\hat{t}_l}}{\|\widehat{\boldsymbol{\theta}}_{\hat{t}_l}\|_2},$$

for a reparametrization $\widehat{\theta_t} = \sum_{\iota=1}^{t} \widehat{\beta_\iota}$ for any $t \in \{1, \ldots, T\}$;

(b) *For any $t = \{1, \ldots T\}$, $n \in \mathbb{N}$, and $\lambda_n > 0$, the following holds with probability one:*

$$\left\| \tau(T - t + 1) \sum_{i=1}^{n} \mathbf{x}_i - \sum_{i=1}^{n} \sum_{k=t}^{T} \mathbf{x}_i \mathbb{1}_{\{Y_{ik} \leq \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_k\}} \right\|_2 \leq n\lambda_n.$$

The proof of Lemma 1 is straightforward and it is omitted. More details can be found in [12]. Let us, however, briefly state some technical assumptions which are needed to prove the the estimation consistency of the proposed method.

Assumptions:

(A1) The errors $\boldsymbol{\varepsilon}_i = [\varepsilon_{1i}, \ldots, \varepsilon_{Ti}]$ are independent copies of some strictly stationary sequence $\boldsymbol{\varepsilon} = [\varepsilon_1, \ldots, \varepsilon_T]$ with the continuous marginal distribution functions $F_{\varepsilon_t}(x)$ and $F_{(\varepsilon_t, \varepsilon_{t+k})}(x, y)$, for $x, y \in \mathbb{R}$, $t, \in \{1, \ldots, T\}$, and $k \geq 1$. Moreover, $F_{\varepsilon_t}(0) = \mathbb{P}[\varepsilon_t < 0] = \tau$, for $\tau \in (0, 1)$. The corresponding density functions $f(\cdot)$ and $f(\cdot, \cdot)$ are bounded and strictly positive in the neighborhood of zero;

(A2) There exist two constants $c, C \in \mathbb{R}$ such that

$$0 < c \leq \mu_{min}(\mathbb{E}[\mathbb{X}_n]) \leq \mu_{max}(\mathbb{E}[\mathbb{X}_n]) \leq C < \infty,$$

where μ_{min} and μ_{max} stand for the minimum and maximum eigenvalue of the matrix in the argument and $\mathbb{X}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^{\top}$. Moreover, $\max_{1 \le i \le n} \|\boldsymbol{x}_i\|_{\infty} < \infty$.

(A3) There are two deterministic positive sequences (λ_n) and (δ_n), such that $\lambda_n \to 0$, $\delta_n \to 0$, $n^{1/2}\delta_n \to \infty$, and $\lambda_n/\delta_n \to 0$ as $n \to \infty$.

Let us recall that in similar models (see, for instance, [5], [4], or [18]) there is an additional assumption which requires that the span between two consecutive changepoints increases. Analogously, the overall number of changes in the model is usually considered to be fixed. However, as far as $T \in \mathbb{N}$ is assumed to be fixed these two assumptions are irrelevant for our specific situation. Given the assumptions above, the consistency can be formulated by the next theorem.

Theorem 1 *Let the assumptions in (A1)–(A3) be all satisfied. Then, for any $t = 1, \ldots, T$, it holds that*

$$\|\widehat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t^*\|_1 = O_P\left(\sqrt{\frac{\log n}{n}}\right),$$

where $\widehat{\boldsymbol{\beta}}_t \in \mathbb{R}^p$ denotes the vector of the estimated parameters obtained by minimizing (8) and $\boldsymbol{\beta}_t^*$ is the corresponding vector of the true values.

The theorem above specifies a proper converge rate for the estimates obtained by minimizing (8). An example of sequences $\{\lambda_n\}$ and $\{\delta_n\}$, which satisfy Assumption (A3) are $\lambda_n = n^{-1} \cdot (\log n)^{1/2}$ and $\delta_n = (n^{-1} \log n)^{1/2}$. For the proof of the theorem we only refer to [12].

In the next section we discuss an application of the proposed modified quantile fused LASSO method to simultaneously estimate a set of implied volatility panels where each panel represents implied volatilities observed over time for a given quoted strike. The observed implied volatilities are, however, firstly interpolated over consecutive maturities in order to obtain artificial call options with a fixed expiry date (30 days). Such smoothing suppresses the natural dynamics of the market (such as increasing convexity of the volatility smile when progressing towards expiry dates) and it gives an opportunity to focus on exogenous effects (changepoints) only.

4 Application: Implied Volatility with Constant Maturity

The proposed quantile fused LASSO approach is applied to estimate the implied volatility surface and to detect possible changes over time for the call options written on Erste Group Bank AG and quoted in the EUREX Deutschland market. The implied volatilities $z_{i,t,k}$, where i represents the strike of the option, k its maturity, and t is the observing day, are downloaded from Thomson Reuters Datastream. The available call option strikes range from 30 Euro to 43.50 Euro with an equidistant step of 0.50 Euro,which gives 28 strikes all together ($n = 28$). There are three considered maturities, for October 19th, 2018, November 16th, 2018, and December 21st, 2018. There are 37 trading days within the analyzed period from September 21st, 2018 till November 12th, 2018, thus $T = 37$. Such period is long enough to capture the dynamics of the volatility smile and to investigate possible changes in its shape. However, similar analysis could be also conducted on some other time periods using the corresponding data.

The first aim is to construct panels that report, for each strike $i \in \{1, \dots, n\}$ and each observing day $t \in \{1, \dots, T\}$, the implied volatility $Y_{i,t}$ of an artificial option having always a constant maturity of K days. Therefore, for each day from the considered period, the observed implied volatilities $z_{i,t,k}$ of the two options that have their maturities immediately before and immediately after are interpolated using the linear combination defined as

$$Y_{it} = \frac{\dfrac{1}{(t+K)-k_b} \cdot z_{i,t,k_b} + \dfrac{1}{k_a-(t+K)} \cdot z_{i,t,k_a}}{\dfrac{1}{(t+K)-k_b} + \dfrac{1}{k_a-(t+K)}}, \tag{10}$$

where k_b is the maturity of the first option expiring before the time $t + K$, respectively, k_a is the maturity of the first option expiring after the time point $t + K$. For the Erste Group Bank AG call options quoted in EUREX Deutschland the fixed maturity of $K = 30$ days is considered.

For example, for the first observing day ($t = 1$), which is September 21st, 2018, the artificial option expires in $t + 30$ days, i.e. October 21st, 2018. The two options used for the artificial volatility interpolation are those with the expiry dates October 19th, 2018 (which is denoted as k_b) and November 16th, 2018 (denoted as k_a). In this case, the distance between the artificial maturity (October 21st, 2018) and the maturity of the first option is $(t + 30) - k_b = 2$ trading days and the distance between the artificial maturity and the maturity of the second option is $k_a - (t + 30) = 19$ trading days.

Therefore, the equation from (10) takes the form

$$Y_{it} = \frac{\dfrac{1}{2} \cdot x_{i,t,k_b} + \dfrac{1}{19} \cdot x_{i,t,k_a}}{\dfrac{1}{2} + \dfrac{1}{19}}. \tag{11}$$

The whole procedure is repeated for all strike panels $i \in \{1, \dots, n\}$ and all trading days from the observational period $t \in \{1, \dots, T\}$. The resulting panels of the artificial implied volatilities $\{Y_{it}\}_{i,t=1}^{n,T}$ are presented in Fig. 2. All together, there are 28 strike panels which are observed for $T = 37$ consecutive trading days.

In the second step, the proposed quantile fused LASSO estimation approach is used to estimate the overall time dependent implied volatility surface while the linear constraints from (9) are again employed to obtain the arbitrage-free valid model at the end. The linear constraints enforce the convexity of the estimated implied

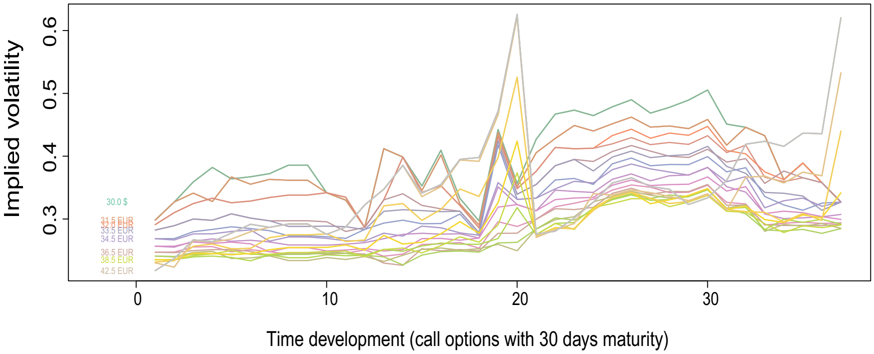


Fig. 2 The time development of the artificial implied volatility for the Erste Group Bank AG call options with a constant 30 days maturity. The analyzed period is from September 21st, 2018 till November 12th, 2018. All together, there are 28 strike panels ranging from 30.0 Euro up to 43.5 Euro observed for 37 consecutive trading days

volatility surface—the convexity (i.e., the volatility smile) with respect to the strikes simultaneously for every day from the analyzed period. The volatility smiles for consecutive days are assumed to be mostly the same with only a few exceptions where the implied volatility function from the time t changes at the time $t + 1$ to adapt for some existing underlying changes on the market. The estimated implied volatility panels are presented in Fig. 3. It is clear that the estimated implied volatility functions are, indeed, all convex for any time point $t \in \{1, \ldots, 37\}$. The overall surface is quite stable but there are also some obvious changes in the estimated volatility surface: some of them occur over time and others are present within the convexity of the volatility smile for some fixed trading days.

From the practical point of view, there are two different explanations for these changes: the changes occurring over time are most likely caused by some exogenous effects (such as the recent COVID-19 outbreaks or the President Trump tweets on additional 10 % tariff to be placed on Chinese imports) while the changes in the volatility smile (increasing convexity when approaching the expiry dates) are still due to some natural dynamics of the market (i.e., high spikes for high strikes in Fig. 3). From the theoretical point of view, these two cases can not be distinguished automatically therefore, we used the artificial options with the constant maturity of 30 days in order to suppress the changes caused by the natural dynamics of the market and, on the other hand, to highlight and detect the changes caused by the external causes. The natural market dynamics is still present in the estimated surface in term of a few high spikes, however, the rest of the surface can be effectively used to analyze the market with respect to external causes effecting the market.

Peripherally, one could be interested whether or not a change in the artificial implied volatilities occurred for some common trading day (cf. Fig. 3), assuming that the volatilities are approximately constant before and after the possible change for every strike. The ratio-type changepoint test statistics proposed in [14] as well as the bootstrap self-normalized changepoint test statistics form [15] both suggest

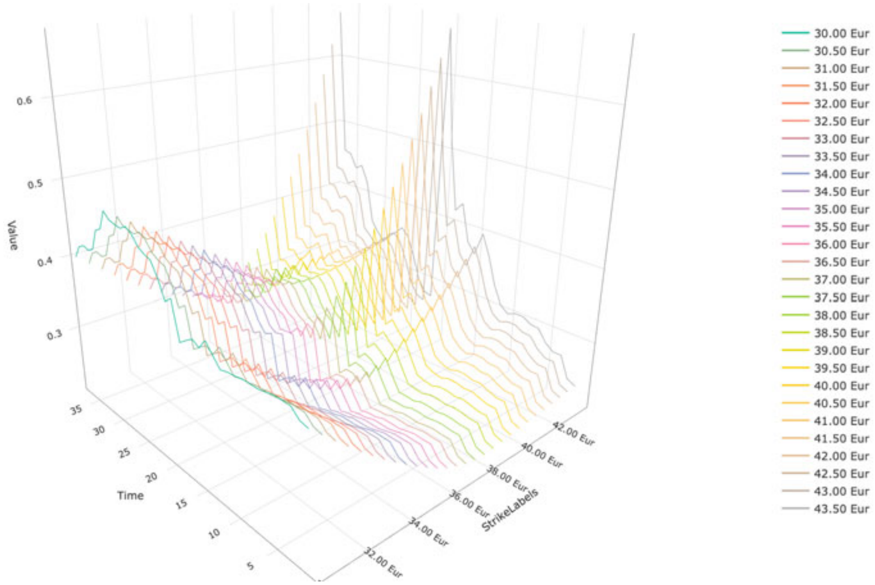
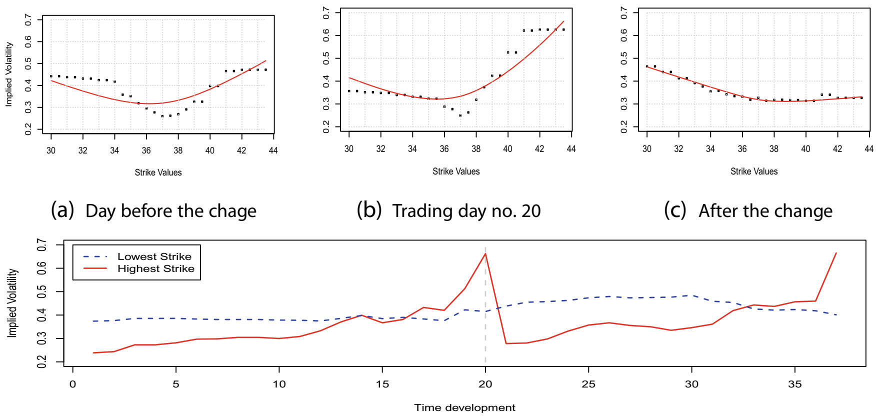


Fig. 3 The estimated panels of the implied volatilities: for each time point $t \in \{1, \ldots, 37\}$ the estimated implied volatility function is clearly convex in strikes and the overall surface is stable over time with just some few spikes for rather high strikes—detected changes in the volatility

to reject the null hypothesis of no change in the panel means. Furthermore, the changepoint estimator developed in [16] reveals a change on the 20th trading day (see Fig. 4).

Alternatively, one could be also interested in some individual tests whether there is a change in some specific strike panel when the panels are considered separately. From Fig. 3 it is obvious that a sudden change (a spike or a wave respectively) occurring on the 20th trading day is only observed for high strikes (roughly the strike values above 38 Euro) while no such behavior is observed for lower strikes. Such panel specific tests are, however, all significant as the overall variability of the observed implied volatility values is relatively high and, more importantly, the raw volatility values do not reflect the arbitrage free market scenario which is implicitly accounted for in our model. In addtion, the multiple testing problem should be considered taken care of properly. Therefore, more precise and more appropriate conclusions can be indeed drawn from the model presented in Fig. 3 rather than performing individual tests and considering individual strike panels separately.



(a) Day before the chage

(b) Trading day no. 20

(c) After the change

Fig. 4 Detection and estimation of the implied volatility change: In the top three panels, there is the implied volatility smile estimated for the day before the change (**a**), the day when the change is detected—the 20th trading day (**b**), and the day after the change (**c**). The left side of the smile seems to be stable while the convexity of the right part of the smile significantly increases for the 20$^{\mathrm{rm}}$ trading day. The overall time profiles for the lowest strike (in blue) and the highest strike (in red) are given in the lower panel. Indeed, the change in the implied volatility is due to the implied volatility recorded for rather higher strikes

5 Conclusion

The implied volatility function and the implied volatility surface are both fundamental tools for the empirical econometrics, the financial derivative markets in particular. A new method, based on the panel data structure, conditional quantile estimation, LASSO regularization, and artificial volatility interpolation is proposed to automatically estimate the time development of the implied volatility function over some specific (fixed) trading period.

This presented approach avoids some popular multistage techniques and nonparametric kernels which usually perform slowly. The sparsity principle and the LASSO fused-type penalty are used to firstly inflate the overall flexibility of the model but, later, the estimate is regularized in order to obtain the final model which fully complies with the financial theory developed for the arbitrage-free market scenarios. The model also implicitly incorporates a prior knowledge that the implied volatility function should not change too roughly and it should be, more or less, stable over time.

The main advantage of the proposed method is that it does not apriori assume the arbitrage-free input data. The estimated implied volatility function, which satisfies the arbitrage-free conditions (so called volatility smile) is obtained automatically in a straightforward and data-driven manner by minimizing the objective function together with some appropriate linear constraints which enforce the convex property. This is crucial for the implied volatility estimation because the volatilities violating the natural market conditions would have serious consequences.

The proposed quantile LASSO method for the panel data structures serves as an innovative and pioneering approach for the option pricing problem and the implied volatility estimation in particular. In addition, the interpolated implied volatilities with the fixed maturity over time offer a much more stable insight into the true market conditions. The proposed estimation approach can easily serve for both, the estimation under the arbitrage-free restrictions or the situation without such restrictions and the presented application shows a straightforward all-in-once implementation for real data cases.

Acknowledgements The authors would like to express thanks to the reviewers for their interesting comments and suggestions. The research of Matúš Maciak and Michal Pešta has been partially supported by the Czech Science Foundation project GAČR No. 18-00522Y. The work of Sebastiano Vitali was supported by MIUR-ex60% 2019 and 2020 sci. resp. Sebastiano Vitali.

References

1. Benko, M., Fengler, M., Härdle, W., Kopa, M.: On extracting information implied in options. Comput. Statistics **4**(22), 543–553 (2007)
2. Black, F., Scholes, M.: The pricing of options and corporate liabilities. J. Polit. Econ. **81**, 637–654 (1973)
3. Britten-Jones, M., Neuberger, A.: Option prices, implied price process and stochastic volatility. J. Finance **55**(2), 839–866 (2000)
4. Ciuperca, G., Maciak, M.: Change-point Detection by the Quantile LASSO Method. J. Statistical Theory Practice **14**(11) (2020). https://doi.org/10.1007/s42519-019-0078-z
5. Ciuperca, G., Maciak, M.: Change-point detection in a linear model by adaptive fused quantile method. Scand. J. Statistics (2019). https://doi.org/10.1111/sjos.12412
6. Fengler, M.R.: Semiparametric Modeling of Implied Volatility. Springer, Berlin, 1st edn. ISBN: 978-3-540-26234-3 (2005)
7. Harchaoui, Z., Lévy-Leduc, C.: Multiple change-point estimation with a total variation penalty. J. Am. Statistical Assoc. **105**(492), 1480–1493 (2010)
8. Huang, J., Ma, S., Xie, H., Zhang, C.: A group bridge approach for variable selection. Biometrika **96**, 339–355 (2009)
9. Hull, C.J., White, A.: The pricing of options on assets with stochastic volatilities. J. Finance **42**(1), 281–300 (1987)
10. Kahale, N.: An arbitrage-free interpolation of volatilities. Risk **5**(17), 102–106 (2004)
11. Kopa, M., Vitali, S., Tichý, T., Hendrych, R.: Implied volatility and state price density estimation: arbitrage analysis. Comput. Manage. Sci. **14**(4), 559–583 (2017)
12. Maciak, M.: Quantile LASSO with changepoints in panel data models applied to option pricing. Econometr. Statistics (2019). https://doi.org/10.1016/j.ecosta.2019.12.005
13. Maciak, M., Mizera, I.: Regularization techniques in joinpoint regression. Statistical Papers **57**(4), 939–955 (2016)
14. Maciak, M., Peštová, B., Pešta, M.: Structural breaks in dependent, heteroscedastic, and extremal panel data. Kybernetika **54**(6), 1106–1121 (2018)
15. Maciak, M., Pešta, M., Peštová, B.: Changepoint in dependent and non-stationary panels. Statistical Papers (2020) https://doi.org/10.1007/s00362-020-01180-6
16. Pešta, M., Peštová, B., Maciak, M.: Changepoint estimation for dependent and non-stationary panels. Appl. Math. (2020) https://doi.org/10.21136/AM.2020.0296-19
17. Qian, J., Su, L.: Structural change estimation in time series regression with endogenous variables. Econ. Lett. **125**, 415–421 (2014)

18. Qian, J., Su, L.: Shrinkage estimation of common breaks in panel data models via adaptive group fused Lasso. J. Econometr. **191**, 86–955 (2016)
19. Vitali, S., Kopa, M., Tichý, T.: State price density estimation for options with dividend yields. Central Europ. Rev. Econ. Issues **20**(3), 81–90 (2017)

A Remark on the Grenander Estimator

Ivan Mizera

Abstract We show rigorously that unlike in the case of s-concave densities, where different Rényi entropies yield different estimates, in the analogous estimation problem when the estimated density is assumed monotone, all Rényi entropies yield the same: the Grenander estimate, obtained as a special case, the maximum likelihood estimate of a monotone density.

Keywords Density estimation · Shape constraints · Convex optimization
Duality · Grenander estimator

1 Introduction

In this note, we study the behavior of a class of estimators of probability densities proposed by Koenker and Mizera [5–8], in a very specific situation, estimation of a monotone probability density.

1.1 *Probability Density Estimators via Rényi Entropies*

Probability density estimators considered in this note can be obtained as solutions of various optimization problems. Let $x_1, x_2, \ldots, x_n$ be the collection of datapoints that are believed to behave as outcomes of independent random variables, all with probability density h. We will be predominantly concerned with the case when $x_i \in \mathbb{R}$; nonetheless, we indulge in an introduction proceeding, even if a bit ambiguously, with $x_i \in \mathbb{R}^p$.

Research supported by the Natural Sciences and Engineering Research Council of Canada.

I. Mizera (✉)
University of Alberta, Mathematical and Statistical Sciences, Edmonton Alberta, Canada
e-mail: imizera@ualberta.ca

M. Maciak et al. (eds.), *Analytical Methods in Statistics*, Springer Proceedings
in Mathematics & Statistics 329, https://doi.org/10.1007/978-3-030-48814-7_5

Suppose that X is a convex set containg all x_i. One possible way to obtain the estimate of h is to consider the objective function

$$\Phi(g) = \frac{1}{n} \sum_{i=1}^{n} g(x_i) + \int_X \psi(g(x))\, dx, \tag{1}$$

and find $\hat{g}$ minimizing it—either under a constraint $J(g) \leq \Lambda$ for some suitable roughness/complexity penalty J, using a common equivalent formulation that adds the penalty term $\lambda J(g)$ to the objective Φ; or under a "hit-or-miss" penalty, a penalty attaining only two values, 0 and $+\infty$, and in this way expressing a *shape constraint*, a constraint on a qualitative form of the estimated density, typically translating into a constraint on g. As a rule, the minimizing $\hat{g}$ is not directly equal to the sought density estimate $\hat{h}$, but the latter can be obtained from $\hat{g}$ via certain transformation related to ψ. The form of this relationship, the possible repertory of the ψ's, as well as the domain of definition of and mathematical requirements on the putative g, and further potential details regarding the domain of integration X are to be discussed later.

Koenker and Mizera [5–7] call the just described way of obtaining the estimate of the density *primal*, to distinguish it from a different, *dual* way. While they originally pursued the penalized approach [5, 6], the possibility of shape-constrained alternatives was already mentioned in [6], and later fully developed, for the shape constraint expressed by the convexity of g, in [7]. This particular context turned out to be fortunate: its mathematical tractability enabled Koenker and Mizera [7] to obtain rigorous duality theory in the *functional setting*, and show that the repertory of functions ψ_α induced by the Rényi system of entropies nicely corresponds with the classes of so-called s-concave functions [4, 7].

The general duality theory for density estimation via Rényi entropies was for a broad class of density estimation problems, either penalized or under shape constraints, already developed by Koenker and Mizera [6]—but there only in the *discretized setting*. To understand the difference between the two, note first that as soon as ψ is convex, then Φ is a convex function; its minimization, either under a constraint on a convex penalty, or enforcing g to lie in a convex set is thus a convex optimization problem. A potent duality theory of such problems is very well developed for *finite-dimensional* problems, problems that operate over $\mathbb{R}^n$. As soon as g is defined over an infinite set—for instance, an interval (possibly bounded) in $\mathbb{R}$ or a convex set with nonempty interior in $\mathbb{R}^p$—then the ensuing convex optimization problem is infinite-dimensional. The duality theory for such problems is far from being that straightforward.

In the numerical context, however, g may need to be replaced by a function defined only on a finite set—the discrete set that "approximates" the original domain. For instance, if g is defined on an interval $[a, b]$, then such an approximation may define it instead merely on a *grid* $y_1 = a < y_2 < \cdots < y_{N-1} < y_N = b$. Intuitively, the approximation will be satisfactory if the distances between adjacent y_i and y_{i+1} are small—which entails that N is large; but modern optimization software and hardware is frequently able to handle such a situation.

Technically, the original datapoints x_i are either included in the grid, or $g(x_i)$ in (1) are replaced by interpolations from adjacent y_i; the integral term in (1) is represented by a suitable sum. The discretized problem typically remains convex; not only can it be then solved by the available software, but being finite-dimensional, its duality theory is typically much easier to develop.

The duality theory elucidated in this *discretized setting* can be interpreted in the functional setting—but this step is only heuristic: sums are replaced by integrals, differences by derivatives... The rigorous justification for the functional setting has yet to be given. For the penalized instances considered in [6] and revived in [8], such a justification is still an open problem. Although the result of such an exercise is entirely predictable, the functional forms of primals and duals are obvious from the discretized setting, the rigorous justification of strong duality—the fact strongly supported by numeric evidence—is still missing. (It is a question whether such rigorous justifications are necessary at all: that may be a viewpoint of authors coming from a computer science background, who typically do not bother, and in exclusively discretized setting concentrate on the more substantial, and also rewarding aspects, rather than arcane mathematical details. Such a viewpoint, however, may not be shared by the readers with more classical background in mathematical statistics, and possibly neither by the present author.)

1.2 The Repertory of ψ

The first instance of ψ in Φ, $\psi_1(x) = \mathrm{e}^{-x}$, came from the maximum likelihood formulation. Maximizing the likelihood of the estimated density h is equivalent to minimizing

$$\Phi(h) = \frac{1}{n}\sum_{i=1}^{n} - \log h(x_i) + \int_X h(x)\,dx. \tag{2}$$

The summation term is a standard negative log-likelihood for independent observations, the integral term replaces the constraint that h should integrate to 1; minimizing the objective with integral term automatically ensures that. Putting $g = -\log h$ yields ψ_1 for ψ in (1), as well as the relationship $h = \mathrm{e}^{-g}$ to the sought density estimate (which also ensures the nonnegativity of h).

The other instances of ψ were inspired by the duality theory. It was observed that the objective functions of the dual formulations of primal optimization problems involving (2) featured in one or another way Shannon entropy of the estimated density h, typically maximized under certain constraints. Replacing the Shannon entropy by the general repertory of Rényi entropies (which contain Shannon as a special case) broadens the repertory of estimation methods; after reverting back to the original primal formulation, we end up with a one-parameter family, ψ_α, of putative ψ. For the detailed functional forms of those, see [6–8]; here we only remark that they are all based on power functions, like $\psi_{1/2}(x) = 1/x$, with the only exception being $\psi_1(x) = \mathrm{e}^{-x}$ and $\psi_0(x) = -\log x$. In this note, we work only with certain mathematical properties elucidated in [7] and satisfied by all of relevant ψ_α.

For the formalism of convex analysis, see [9] or [2]. We allow infinities in the definitions of convex functions; all our convex functions are *proper*, never resulting in $-\infty$; hence their *effective domains* are the sets where their values are finite. From now on, every ψ satisfies

(A) ψ is nonincreasing and proper convex, with the open effective domain containing $(0, +\infty)$; on this domain, ψ is differentiable.

For every such ψ, we will have either

(A0) ψ satisfies all of (A), and $\psi(x_0) = 0$ for some x_0

or

(A1) ψ satisfies all of (A), $\psi(x) \geq 0$ for all x, and $\psi(x) \to 0$ for $x \to +\infty$.

The very last property is interpreted as $\psi(+\infty) = 0$. The only function in the Rényi system that has property (A0) and not (A1) is $\psi_0(x) = -\log x$; every other ψ_α satisfies (A1).

2 Estimation Under Monotonicity

In the implementation of the primal prescription minimizing (1) under the constraint that g is convex, we enforced convexity by requiring that

$$Dg(x) \geq 0 \quad \text{for all } x, \tag{3}$$

with D standing for the operator of the second derivative/difference (and $\geq$ in the multi-dimensional case with $p > 1$ understood in the sense of nonnegative definitess). In the functional setting, it is more prudent to work directly with the convexity of g (invoking D as the operator of second derivative may raise possible concerns about the differentiability of pertinent g, and indeed these would then need to be mitigated by employing derivatives in the generalized sense of distributions). In the discretized setting, however, (3) works well, and even lends the resulting algorithm certain flexibility: a possibility arises to replace D by some other difference operator. In particular, in the one-dimensional case (when $p = 1$ and $\geq$ is understood in the most straightforward way), replacing D by the operator of *first* difference enforces g via (3) to be nondecreasing. As we will see later, this eventually results in the *nonincreasing* estimate of the density h.

Estimation of a monotone density is a classical topic: the celebrated Grenander [3] estimator, maximum likelihood estimator under the assumption of noincreasing density, is nothing but our estimator under ψ equal to ψ_1, with nondecreasing g. To observe that, just note that if g is nondecreasing, then $h = \mathrm{e}^{-g}$ is noincreasing, and (2) is indeed the objective function arising from maximizing the likelihood.

As reiterated by [6–8], enforcing the convexity assumption on g yields for each ψ_α a different estimator, in a different class of probability densities. An intriguing question arose: in the monotonicity contex, do different ψ_α also lead to different estimators? The answer to this question obtained by numerical experimentation with various α seemed to be negative: all resulting estimates looked the same. Given, however, the limits of numerical experimentation, which inevitably depends on the particular data input and other factors—for instance, the coincidence may be observed just for those α for which the numerical algorithm is reasonably stable, and for some others the coincidence of the estimates may not appear that convincing (which was indeed the case, probably due to numerical woes)—it was of some interest whether the observed coincidence is a general rule, valid for all data inputs and all ϕ, a rule that can be established theoretically. This note gives an affirmative answer to this question.

2.1 Relevant Estimators and Duality

Thus, our datapoints $x_1, x_2, \dots, x_n$ emerge now in a domain that is not whole $\mathbb{R} = (-\infty, \infty)$, but it is bounded from below—which can be without loss of generality viewed as an interval with its left endpoint at 0. In view of the fact that one-dimensional density estimation admits a reduction to order statistics via sufficiency, we consider the x_i's already ordered: $x_1 \le x_2 \le \cdots \le x_n$. The objective Φ is still defined by (1); the integration domain X has left endpoint zero, and has to contain all x_i, but otherwise its exact specification for any ψ satisfying (A2) is irrelevant, and X may well be set to be the whole $[0, +\infty)$ interval—*an exception being, however,* functions satisfying only *(A1), for which the integration domain has to be set to* $[0, x_n]$.

The details of this nature are somewhat relevant to the preliminary characterization of estimates—which is in turn necessary to facilitate the subsequent mathematical treatment: we would like to know that the original general formulation can be reduced to one acting on bounded domain and employing specific classess of functions. These deliberations are not that essential in practice: after all, we know that in numerical algorithms the integration domain has always to be bounded, and in discretized setting every function can be considered to be piecewise constant or piecewise linear. Nonetheless, in theoretical setting they exemplify the natural character of these reductions.

So, the objective Φ is now minimized under the constraint that g is *nondecreasing*. To observe that this yields *nonincreasing* density estimates h, we have to invoke the relationship between the solution $\hat{g}$ of the optimization problem, and the desired estimate $\hat{h}$, which is

$$\hat{h} = -\psi'(\hat{g}). \tag{4}$$

As all ψ satisfying (A) are convex, the derivative $(-\psi'(x))' = -\psi''(x)$ is nonpositive for all x; therefore $-\psi'$ is nonincreasing, and nondecreasing $\hat{g}$ yields nonincreasing $\hat{h}$.

The relationship (4) is the outcome of the duality theory established for the shape-constrained density estimation via minimization of (1). For g enforced convex, the strong duality and subsequently (4) was established by Theorem 3.1 of [7]. The closer examination of its proof reveals that a cone $\mathcal{K}(X)$ of *convex* functions on X can be replaced by any convex cone *containing constant functions*. For convenience, we rephrase the generalized formulation here.

We suppose that all functions g relevant for objective Φ from (1) belong to a topological vector space $\mathcal{G}$ such that its topological dual $\mathcal{G}^*$, the set of continuous linear functionals on $\mathcal{G}$, contains evaluation functionals δ_x for all $x \in X$, continuous linear functionals such that $\delta_x(g) = g(x)$. Given the dataset $x_1, x_2, \ldots, x_n$, let P_n be the element of $\mathcal{G}^*$ defined as

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$$

—we customarily call P_n the empirical probability supported by the x_i's. Let $\mathcal{K}$ be a convex cone of functions from $\mathcal{G}$, a convex subset $\mathcal{G}$ closed under multiplication by positive constants. Its *polar cone* $\mathcal{K}^-$ is then the set of all $G \in \mathcal{G}^*$, such that $G(g) \leq 0$ for all $g \in \mathcal{G}$.

Proposition 1 Suppose that ψ satisfies (A), and let $\mathcal{K}$ be a convex cone of functions defined on X containing all constant functions. The strong (Fenchel) dual of a primal formulation minimizing (1) under the constraint that $g \in \mathcal{K}$ is the problem maximizing

$$- \int_X \psi^*(-h(y))dy \qquad \text{subject to} \qquad h = \frac{d(P_n - G)}{dy} \tag{5}$$

in the sense that the value, $\Phi(g)$, of the primal objective for any $g \in \mathcal{K}$ dominates the value, for any f satisfying the constraints of (5), of the objective function in (5); the minimal value of Φ and maximal value of the objective in (5) coincide. Moreover, there exists $\hat{h}$ attaining the maximal value of the problem defined by (5). Any dual feasible function h , that is, any h satisfying the constraints of (5) and yielding finite objective function in (5) is a probability density with respect to the Lebesgue measure: $h \geq 0$ and $\int_X h\,dx = 1$. The dual and primal optimal solutions satisfy (4).

Proof The proof is completely analogous to that of Theorem 3.1 in [7] and is omitted. We only remark that constant functions are first used to establish the constraint qualification for the strong dual, and then also to establish that every feasible h is a probability density.

To use this proposition in our case, we have now to establish the preliminary characterization of the estimates.

2.2 Preliminary Characterization of the Estimates

The preliminary characterization of the estimates amounts to finding a suitable collection $\tilde{\mathcal{G}}$ of *admissible* g, characterized by the property that for every initially relevant g we can find some $\tilde{g} \in \tilde{\mathcal{G}}$ such that $\Phi(\tilde{g}) \leq \Phi(g)$. Out of several possibilities, we are interested in $\tilde{\mathcal{G}}$ exhibiting favorable properties that may facilitate the analysis of solutions.

In the situation when g is assumed convex, such an admissible $\mathcal{G}$ is given in Sect. 2.4 of [7], and consists of g that are $+\infty$ outside the convex hull of the x_i's, and are polyhedral (that is, piecewise linear) on this convex hull, with the subdomains of linearity spanned by the x_i's; in one-dimensional case it means that the breakpoints can occur only at the x_i's.

This is quite a step beyond likelihood considerations which only require something like well-defined functional values $g(x_i)$. This, in fact, follows immediately from the convexity constraint which entails continuity (and thus well-defined functional values) on the relative interior of the effective domain. For ψ satisfying (A0), there is x such that $\psi(x) = 0$; any g identically equal to x yields an example of convex g such that $\Phi(g) < +\infty$. For ψ satisfying (A1), such g is any that is constant on the convex hull of the x_i's and equal to $+\infty$ elsewhere.

The existence of g with $\Phi(g) < +\infty$ entails that any $g(x_i)$ has to be finite. Once all the values $g(x_i)$, and thus also the summation term in (1) are fixed, then the integral term is minimized by those g that are maximal in the gaps between the x_i's, but still remain convex: that makes admissible g piecewise linear. Property (A1) drives the admissible g outside the convex hull of the x_i's to $+\infty$ rendering the exact boundary integration domain irrelevant, as $\psi(+\infty) = 0$. However, for ψ satisfying only (A0), such a reduction of the domain does not come automatically and has to be explicitly postulated.

Let us return to monotone g now. Monotonicity entails the existence of one-sided limits; the latter coincide in the points of continuity, and make $g(x_i)$ defined unambigously at these points. In the points of discontinuity, the $g(x_i)$'s could be anything between the two one-sided limits; the objective function Φ, however, pushes them eventually equal to the smaller of the two. The analogous argument with constant g as in the convex case entails that all values $g(x_i)$ must be finite; monotonicity implies that in the gaps between x_i any admissible g must be constant. Therefore, all admissible g are thus left- or right-continuous, depending on the sense of monotonicity. For nondecreasing g, which are the ones we will consider here, it is left continuity. Again, for ψ satisfying (A1), the admissible g are equal to $+\infty$ for all $x > x_n$, with the analogous consequences for the domain of integration X. And again, for ψ satisfying merely (A0), X needs to be set to $[0, x_n]$. We summarize all this in the following

Proposition 2 Suppose that the domain of all g is $X = [0, x_n]$. For all ψ satisfying (A), the admissible nondecreasing g are left-continuous and piecewise constant, with nonzero jumps allowed only at the x_i's.

With the help of the proposition, we can overcome the only remaining mathematical obstacle: applying Proposition 1 to our situation. The difficulty is of functional-analytic nature. The topological vector space $\mathcal{G}$ in Theorem 3.1 of [7] was $C(X)$, the space of continuous functions on a compact X; its topological dual $\mathcal{G}^*$ is the well-understood set of Radon measures on X. In our situation we deal with monotone functions that may not be continuous, hence not belong to $C(X)$. We have to take $\mathcal{G}$ instead to be the collection of all *bounded* (in view of Proposition 2 there are at most n jumps), left-continuous, nondecreasing functions on a compact interval $X = [0, x_n]$. The topology may be that of Skorokhod convergence—see [1]—but the usual sup topology of $C(X)$ works also here, in the possibly discontinuous case.

There are continuous functions on X that are also monotone, hence the intersection of our $\mathcal{G}$ with $C(X)$ is nonempty. It contains nondecreasing continuous functions on X, which are dense in $\mathcal{G}$. They are obviously not dense in $C(X)$; but the collection of all finite linear combinations of them is. To see that, consider all piecewise linear continuous functions on X, with finitely many pieces of linearity: each of them is a difference of two functions continuous and nondecreasing on X, and they constitute a dense subset in $C(X)$—this can be seen either directly, using uniform continuity, or via the Stone-Weierstrass theorem.

All this means that a continuous linear functional on $C(X)$ defines a continuous linear functional on its subset, its intersection with $\mathcal{G}$. As this subset is dense in $\mathcal{G}$, the functional can be unambiguously extended to a continuous functional on $\mathcal{G}$. Conversely, a continuous linear functional on $\mathcal{G}$ defines a continuous linear functional on its intersection with $C(X)$; as the finite linear combinations of the elements of this intersection are dense in $C(X)$, this functional can be unabiguously extended to that on $C(X)$.

In other words, there is a one-one correspondence between continuous linear functionals on $\mathcal{G}$ and those on $C(X)$, which means that their duals coincide: the elements of $\mathcal{G}^*$ are Radon measures, and we can proceed with the whole analysis along the same lines as in [7], where Φ was minimized under the constraint that g is convex.

2.3 The Main Result

As was noted above, numerical experiments with estimators using various Rényi entropies suggested that they all yield the same result—which is then the Grenander estimator, as (2) is a special case of (1). This is indeed the case, and can be proved formally.

Theorem *All ϕ satisfying (A) yield the same minimizer of the objective function* (1) *subject to g nondecreasing—and thus the same noincreasing estimate of the probability density.*

Proof In view of the characterization given by Proposition 2, we consider the new objective function

$$\Psi(g_1, g_2, \dots, g_n) = \frac{1}{n}\sum_{i=1}^{n} g_i + \sum_{i=1}^{n} w_i \psi(g_i) \tag{6}$$

The variables g_i now represent $g(x_i)$; in view of the piecewise constancy we set $w_i = x_i - x_{i-1}$, with x_0 interpreted as $x_0 = 0$. We minimize (6) under the condition

$$g_1 \le g_2 \le \cdots \le g_n \tag{7}$$

In view of Proposition 2, the minimizer of this finite-dimensional problem characterizes the minimizer of the original problem, of minimizing (1) on $X = [0, x_n]$, subject to g nondecreasing: it is a left-continuous, piecewise constant function $\hat{g}$ with $\hat{g}(x_i) = g_i$. We will show that the minimizer of the finite-dimensional problem is the same for all ψ satisfying (A).

We introduce new variables $q_1, q_2, \dots, q_n, q_{n+1}$ such that

$$\begin{array}{ll}
g_n = q_{n+1} - q_n & q_{n+1} \ge 0, q_n \ge 0, \\
g_{n-1} = q_{n+1} - q_n - q_{n-1} & q_{n-1} \ge 0, \\
\quad\vdots & \quad\vdots \\
g_3 = q_{n+1} - q_n - q_{n-1} - \cdots - q_3 & q_3 \ge 0, \\
g_2 = q_{n+1} - q_n - q_{n-1} - \cdots - q_3 - q_2 & q_2 \ge 0, \\
g_1 = q_{n+1} - q_n - q_{n-1} - \cdots - q_3 - q_2 - q_1 & q_1 \ge 0.
\end{array}$$

Note that the nonnegativity of $q_1, q_2, \dots, q_{n-1}$ follows from (7); for g_n, which in principle may be negative, we have to introduce two new variables, so that all q_k are nonnegative and we can use the conditions from page 142 of [2] characterizing optimum of our finite-dimensional problem. These conditions say that the partial derivatives of Ψ with respect to all q_k have to be nonnegative; for $k = 1, 2, \dots, n$, this yields

$$0 \le \frac{\partial \Psi}{\partial q_k} = (-1)\left(\frac{k}{n} + \sum_{i=1}^{k} w_i \psi'(g_i)\right) = (-1)\left(\frac{k}{n} - \sum_{i=1}^{k} w_i h_i\right),$$

putting $h_i = -\psi'(g_i)$ in accord with (4), and subsequently

$$\frac{k}{n} \le \sum_{i=1}^{k} w_i h_i \qquad \text{for all} \quad k = 1, 2, \dots, n. \tag{8}$$

For $k = n + 1$, we have

$$0 \le \frac{\partial \Psi}{\partial q_{n+1}} = 1 + \sum_{i=1}^{n} w_i \psi'(g_i) = 1 - \sum_{i=1}^{n} w_i h_i$$

yielding

$$1 \leq \sum_{i=1}^{n} w_i h_i,$$

which together with (8) for $k = 1$ gives that

$$1 = \sum_{i=1}^{n} w_i h_i, \tag{9}$$

Indeed: as the partial sum $\sum_{i=1}^{k} w_i h_i$ is the cumulative distribution function of the putative density estimate characterized by h_i, the equation (9) asserts that eventually such a distribution function ought to be equal to 1. More importantly, inequality (8) says that this distribution functions majorizes the empirical distribution function, equal to k/n at each x_k.

In view of the piecewise constancy of the density estimate, its cumulative distribution function is piecewise linear; we would like to know where it changes its slope. This is given by the complementarity condition on page 142 of [2]: it says that whenever $q_k > 0$, that is, whenever the putative density estimate has a jump at x_k, the the partial derivative of Ψ has to be zero:

$$0 = \frac{\partial \Psi}{\partial q_k} = \frac{k}{n} - \sum_{i=1}^{k} w_i h_i.$$

This says that whenever the cumulative distribution function of a putative estimate changes its slope, it is equal to the empirical distribution function, given by k/n—which it majorizes, so it is its least concave majorant. But this is the well-known characterization of the cumulative distribution function of the Grenander [3] estimate—which comes from the minimization of (2) subject to g nondecreasing.

But, as we obtained this characterization for every ψ under consideration, then obviously the result of minimization of (6) under (7), and thus also that of (1) under g nondecreasing is the same for all ψ satisfying (A): we have proved the theorem.

3 Conclusion

As Czech urban folkore has it, techno party is over—time is to move on. The fact that alternative Rényi entropies do not yield anything different in the monotonicity context other than tried-and-true maximum likelihood may be reassuring for many—for them, the world is now in harmony again, and they may well opine that it would be even without this note.

Instead, however, we hope this fact will be at least of some interest for those few intrigued by the alternative estimation possibilities offered by the Rényi entropies, if only in certain other situations; and if those do not find it minimally somewhat surprising, we hope they will at least appreciate the mathematical technology used to rigorously confirm otherwise clumsy numerical impressions.

References

1. Billingsley, P.: Convergence of Probability Measures. Wiley, New York (1968)
2. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
3. Grenander, U.: On the theory of mortality measurement. Scand. Actuarial J. 70–153 (1956)
4. Han, Q., Wellner, J.A.: Approximation and estimation of s-concave densities via Rényi divergences. Ann. Statistics **44**(3), 1332–1359 (2016)
5. Koenker, R., Mizera, I.: The alter egos of the regularized maximum likelihood density estimators: deregularized maximum-entropy, Shannon, Renyi, Simpson, Gini, and stretched strings. In: Hušková, M., Janžura, M. (eds.) Prague Stochastics 2006, Proceedings of the joint session of 7th Prague Symposium on Asymptotic Statistics and 15th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes, held in Prague from August 21 to 25, 2006, Matfyzpress, Prague, pp. 145–157 (2006)
6. Koenker, R., Mizera, I.: Primal and dual formulations relevant for the numerical estimation of a probability density via regularization. In: Pázman, A., Volaufová, J., Witkovský, V. (eds.) Proceedings of the conference on Tatra Mountains Mathematical Publications (ProbaStat '06), Smolenice, Slovakia, vol 39, pp. 255–264, 5–9 June 2006. Slovak Academy of Sciences (2008)
7. Koenker, R., Mizera, I.: Quasi-concave density estimation. Ann. Statistics **38**(5), 2998–3027 (2010)
8. Koenker, R., Mizera, I.: Shape constrained density estimation via penalized Renyi divergence. Statistical Sci. **33**(4), 510–526 (2018)
9. Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton (1970)

Non-Gaussian Component Analysis: Testing the Dimension of the Signal Subspace

Una Radojičić and Klaus Nordhausen

Abstract Dimension reduction is a common strategy in multivariate data analysis which seeks a subspace which contains all interesting features needed for the subsequent analysis. Non-Gaussian component analysis attempts for this purpose to divide the data into a non-Gaussian part, the signal, and a Gaussian part, the noise. We will show that the simultaneous use of two scatter functionals can be used for this purpose and suggest a bootstrap test to test the dimension of the non-Gaussian subspace. Sequential application of the test can then for example be used to estimate the signal dimension.

Keywords Independent component analysis · Scatter matrix · Bootstrap · Order determination

1 Introduction

Modern data sets contain often many variables making visualization and many other tasks concerning the data set very difficult. Therefore, dimension reduction methods gain popularity as they try to find a subspace of the data which is smaller and contains all interesting features. Three main issues are then here, (i) how to define what makes the data interesting, (ii) how large is the interesting subspace and (iii) how to find the subspace?

There are meanwhile many suggestions about how to define what is interesting and maybe the most used method is principal component analysis (PCA) [13] which defines as interesting subspace the one which accounts for as much of the variability in the data as possible. Another well-established approach is projection pursuit (PP) [11, 14] where usually univariate projections of the data, which maximize some

U. Radojičić (✉) · K. Nordhausen
Institute of Statistics & Mathematical Methods in Economics, Vienna University of Technology, Wiedner Hauptstr. 7, 1040 Vienna, Austria
e-mail: una.radojicic@tuwien.ac.at

K. Nordhausen
e-mail: klaus.nordhausen@tuwien.ac.at

M. Maciak et al. (eds.), *Analytical Methods in Statistics*, Springer Proceedings in Mathematics & Statistics 329, https://doi.org/10.1007/978-3-030-48814-7_6

criterion of non-Gaussianity specified by an projection index, are considered interesting. PCA is probably so popular as it is quite easy to compute and has many different guidelines on how to choose the dimension of the subspace of interest. PP on the other hand is, depending on the projection index used, often computationally expensive. Moreover, guidelines about how to choose the dimension of the interesting subspace are sparse. However, PP has been proven useful as a preprocessing step for, for example, clustering or outlier detection [9]. In general, it seems that the non-Gaussian subspace of the data is nowadays considered the subspace of interest and [4] suggested a general framework for this, denoted by non-Gaussian component analysis (NGCA). It divides the data into a non-Gaussian subspace and into a Gaussian subspace. While there are meanwhile many suggestions, like in [2, 3, 15, 33, 37, 41] to name a few, on how to perform NGCA there is not much research yet on how to estimate the dimensions of the two subspaces.

In this paper we will introduce a bootstrap test to test the dimension of the non-Gaussian subspace using two scatter matrices. For this purpose we will in the following first introduce scatter matrices and some of their relevant properties. Then, in Sect. 3 we will introduce the independent component (IC) model which is closely related to the NGCA model, which we will also define then there in detail. The bootstrap test is then introduced in Sect. 4 and evaluated in a simulation study in Sect. 5. Natural estimates of the signal dimension are found by successive conduction of the bootstrap test and two estimation strategies are discussed and evaluated in Sect. 6. Proofs of selected results are given in the Appendix.

2 Scatter Functionals

Scatter functionals are the main tools in our method and defined as follows:

Definition 1 Let $\mathbf{x}$ be a p-variate random vector with distribution function $F_{\mathbf{x}}$. Then a $p \times p$ matrix-valued functional $\mathbf{S}(F_{\mathbf{x}}) = \mathbf{S}(\mathbf{x})$ is called a scatter functional if it is symmetric, positive semi-definite and affine equivariant in the sense that

$$\mathbf{S}(\mathbf{A}\mathbf{x} + \mathbf{b}) = \mathbf{A}\mathbf{S}(\mathbf{x})\mathbf{A}^\top,$$

for all full rank $p \times p$ matrices $\mathbf{A}$ and all p-variate vectors $\mathbf{b}$.

Scatter functionals often come along with a location functional which is defined as:

Definition 2 Let $\mathbf{x}$ be a p-variate random vector with distribution function $F_{\mathbf{x}}$. Then a p-vector-valued functional $\mathbf{T}(F_{\mathbf{x}}) = \mathbf{T}(\mathbf{x})$ is called a location functional if it is affine equivariant in the sense that

$$\mathbf{T}(\mathbf{A}\mathbf{x} + \mathbf{b}) = \mathbf{A}\mathbf{T}(\mathbf{x}) + \mathbf{b},$$

for all full rank $p \times p$ matrices $\mathbf{A}$ and all p-variate vectors $\mathbf{b}$.

Thus, location and scatter functionals are a way to describe centrality and spread of the data and are then estimated by replacing $F_{\mathbf{x}}$ with the empirical distribution. Probably the most widely used pair of location and scatter functionals are the expected value $\mathbf{E}(\mathbf{x})$ and the covariance matrix $\mathbf{COV}(\mathbf{x})$.

The literature is however full of many alternatives which have different desirable properties, like robustness or efficiency, at specific models. A large family of functionals which we will use in the following are the M-estimators of location and scatter and are for example reviewed in [8].

Definition 3 M-functionals of location and scatter are defined by the two following implicit equations:

$$\mathbf{T}(\mathbf{x}) = \mathbf{E}(w_1(r))^{-1}\mathbf{E}(w_1(r)\mathbf{x})$$

and

$$\mathbf{S}(\mathbf{x}) = \mathbf{E}\left(w_2(r)\,(\mathbf{x} - \mathbf{T}(\mathbf{x}))\,(\mathbf{x} - \mathbf{T}(\mathbf{x}))^\top\right),$$

where $w_1(r)$ and $w_2(r)$ are nonnegative continuous functions of the Mahalanobis distance $r = ||\mathbf{S}(\mathbf{x})^{-1/2}(\mathbf{x} - \mathbf{T}(\mathbf{x}))||$.

Thus, M-functionals of location and scatter are weighted variants of the mean and the covariance matrix yielding them as special cases when choosing $w_1(r) = w_2(r) = 1$. Usually the weight functions are chosen to be non-increasing to obtain estimators that may be robust. Some popular members of the family of M-estimators have the following weight functions

- Huber's M-estimators [10]

$$w_1(r) = \begin{cases} 1 & r \le c \\ c/r & r > c \end{cases} \quad \text{and} \quad w_2(r) = \begin{cases} 1/\sigma^2 & r \le c \\ c/(r^2\sigma^2) & r > c \end{cases},$$

  where σ^2 is a scaling factor chosen so that $\mathrm{E}(Qw_2(\sqrt{Q})) = p$ and c is a tuning constant chosen to satisfy $q = Pr(Q \le c^2)$, where $Q \sim \chi^2_p$.
- M-estimators based on the likelihood of a t-distribution having $\nu \ge 1$ degrees of freedom [16]

$$w_1(r) = w_2(r) = \frac{p + \nu}{r^2 + \nu}.$$

Traditionally, M-estimators of location and scatter are computed via fixed point algorithms which are iterated from an initial starting point until the difference in successive functional values is less than some predetermined threshold. Depending on the weight functions there are however also other algorithms available, see e.g. [7].

A compromise here in the iterative process are the so called one-step M-estimators of location and scatter which start with a pair of location and scatter functionals and then use just one updating step to obtain weighted new functionals. A scatter functional from this family which we will consider later is the scatter matrix of fourth moments which starts with the pair $(\mathbf{T}_1,\ \mathbf{S}_1)$=$(\mathbf{E},\ \mathbf{COV})$ and yields eventually

$$\mathbf{COV}_4(\mathbf{x}) = \frac{1}{p+2}\mathbf{E}\left(r^2\left(\mathbf{x} - \mathbf{T}_1(\mathbf{x})\right)\left(\mathbf{x} - \mathbf{T}_1(\mathbf{x})\right)^\top\right),$$

thus having the weight function $w_2(r) = r^2/(p+2)$, where $r = ||\mathbf{S}_1(\mathbf{x})^{-1/2}(\mathbf{x} - \mathbf{T}_1(\mathbf{x}))||$.

Scatter functionals are mainly investigated in the context of elliptical distributions where it is a well-known fact that they are all proportional to each other given they exist [28]. However, as the Gaussian distribution is the only elliptical distribution with independent components, other properties of scatter functionals are of interest in NGCA. For example the properties of full and block independence for scatter functionals are defined in [28].

Definition 4 A scatter functional $\mathbf{S}(\mathbf{x})$ is said to have the full independence property if

$$\mathbf{S}(\mathbf{x}) = \mathbf{D}(\mathbf{x})$$

for all $\mathbf{x}$ having independent components where $\mathbf{D}(\mathbf{x})$ denotes a diagonal matrix.

If the p-variate vector $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_k)^\top$ has k independent blocks with corresponding block dimensions $p_1, \dots, p_k$, then a scatter functional $\mathbf{S}(\mathbf{x})$ is said to have the block independence property if

$$\mathbf{S}(\mathbf{x}) = \mathbf{B}(\mathbf{x}),$$

where $\mathbf{B}(\mathbf{x})$ is the block diagonal matrix with block dimensions $p_1, \dots, p_k$.

Most scatter functionals do not posses the full or block independence property, however **COV** and $\mathbf{COV}_4$ do. All scatter functionals are however diagonal and block diagonal in case when all but one of the independent parts are symmetric [28]. Exploiting the concept of symmetry, symmetrized scatter functionals can be defined.

Definition 5 Let $\mathbf{S}$ denote any scatter functional, then its symmetrized version is defined as

$$\mathbf{S}_{sym}(\mathbf{x}) := \mathbf{S}(\mathbf{x}^1 - \mathbf{x}^2),$$

where $\mathbf{x}^1$ and $\mathbf{x}^2$ are independent copies of $\mathbf{x}$.

For example [28] show that every symmetrized scatter functional possess the full and block independence property. Note also that **COV** and $\mathbf{COV}_4$ can actually be expressed as functions of pairwise differences and that symmetrized scatter functionals do not require a location functional. Actually, they are usually computed using all pairwise differences and computing the original scatter with respect to the origin. Symmetrized M-estimators of scatter are investigated in [35], while the computational issues are especially discussed in [7, 18].

3 NGCA and ICA

The non-Gaussian component analysis (NGCA) model we will consider in the following is defined as follows.

Definition 6 A (centered) p-variate vector $\mathbf{x}$ follows the NGCA model if it can be decomposed as

$$\mathbf{x} = \mathbf{A}\mathbf{z} = \mathbf{A}_1\mathbf{s} + \mathbf{A}_2\mathbf{n},$$

where $\mathbf{z} = (\mathbf{s}^\top \mathbf{n}^\top)^\top$ is a latent p-variate vector consisting of the q-variate non-Gaussian signal vector $\mathbf{s}$ and the $(p-q)$-variate Gaussian noise vector $\mathbf{n}$. The signal and noise vectors are independent and locations and scales are fixed using a pair of location and scatter functionals as $\mathbf{T}(\mathbf{z}) = \mathbf{0}$ and $\mathbf{S}(\mathbf{z}) = \mathbf{I}_p$, where $\mathbf{S}$. The full-rank $p \times p$ matrix $\mathbf{A}$ is called the mixing matrix and $\mathbf{A}_1$ and $\mathbf{A}_2$ are $p \times q$ and $p \times (p-q)$ matrices with ranks q and $p-q$ respectively and specify the signal and noise parts of $\mathbf{x}$.

The signal dimension q is the largest value separating between the signal and noise values. That is, there exists no q-variate vector $\mathbf{a}$ such that $\mathbf{a}^\top\mathbf{s}$ has a normal distribution, and also, q is the largest such number ensuring that $\mathbf{n}$ is a Gaussian noise vector. Still, the two matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ are not identifiable as both can be post-multiplied by $q \times q$ and $(p-q) \times (p-q)$ dimensional orthogonal matrices respectively and consequentially $\mathbf{A}$ is not identifiable either.

The goal of non-Gaussian component analysis is thus to find a $p \times p$ full rank unmixing block matrix

$$\mathbf{W} = (\mathbf{W}_1^\top \mathbf{W}_2^\top)^\top = \begin{pmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \end{pmatrix},$$

with submatrices $\mathbf{W}_1$ and $\mathbf{W}_2$, such that $\mathbf{W}_1\mathbf{x}$ recovers the non-Gaussian signal subspace and $\mathbf{W}_2\mathbf{x}$ the Gaussian noise subspace.

There are also several closely related models which we would like to introduce.

The independent component analysis (ICA) model can be seen as an extreme case of the NGCA model where all components of $\mathbf{s}$ are independent and q is either $p-1$ or p. In that case $\mathbf{A}$ is identifiable up to the order and the signs of its rows, and therefore, in this case, one can think of $\mathbf{W}$ as its inverse, keeping in mind that it is well defined up to the order and the signs of its rows. ICA is for example widely used in the analysis of biomedical signals and has many other applications; for details see for example [6, 21].

A compromise between NGCA and ICA is the non-Gaussian independent component model (NGICA) which is an NGCA model where all components of $\mathbf{s}$ are independent and the ICA model is thus a special case. The NGICA model has the advantage over the general NGCA model that the signal components of $\mathbf{s}$ are identifiable up to their order and signs. NGICA was for example considered in [12, 25, 32].

NGCA on the other hand can be seen as a special case of independent subspace analysis (ISA), where it is assumed that the latent vector $\mathbf{z}$ consists of k independent blocks and these subspaces need to be identified. For details about ISA see for example [20, 36].

As mentioned above, there are many methods to estimate the unmixing matrix in NGCA where many of them are based on projection pursuit ideas. The approach of interest in this paper is based however on the simultaneous use of two scatter functionals $\mathbf{S}_1$ and $\mathbf{S}_2$.

In the beginning we choose $\mathbf{S}_1 = \mathbf{COV}$ and $\mathbf{S}_2 = \mathbf{COV}_4$ and define the fourth-order-blind-identification (FOBI) functional as:

Definition 7 Let $\mathbf{x}$ be a p-variate random vector with finite fourth moments and set $\mathbf{S}_1 = \mathbf{COV}$ and $\mathbf{S}_2 = \mathbf{COV}_4$. Then the FOBI functional is defined as the $p \times p$ matrix-valued functional $\mathbf{W}$ which simultaneously diagonalizes $\mathbf{S}_1$ and $\mathbf{S}_2$. That means

$$\mathbf{W}(\mathbf{x})\mathbf{S}_1(\mathbf{x})\mathbf{W}(\mathbf{x})^\top = \mathbf{I}_p \quad \text{and} \quad \mathbf{W}(\mathbf{x})\mathbf{S}_2(\mathbf{x})\mathbf{W}(\mathbf{x})^\top = \mathbf{D}(\mathbf{x}),$$

where $\mathbf{D}(\mathbf{x})$ is a diagonal matrix with decreasing diagonal elements.

For convenience and when the context is clear the dependence on $\mathbf{x}$ of $\mathbf{S}_1$, $\mathbf{S}_2$, $\mathbf{W}$ and $\mathbf{D}$ will be omitted. The FOBI functional $\mathbf{W}$ is usually obtained by first whitening $\mathbf{x} \mapsto \mathbf{x}^{st} = \mathbf{S}_1(\mathbf{x})^{-1/2}(\mathbf{x} - \mathbf{E}(\mathbf{x}))$ and then performing an eigenvalue-eigenvector decomposition of $\mathbf{S}_2(\mathbf{x}^{st}) = \mathbf{U}\mathbf{D}\mathbf{U}^\top$. It can then be shown that $\mathbf{W} = \mathbf{U}\mathbf{S}_1^{-1/2}$, and that $\mathbf{D}$ in the eigenvalue-eigenvector decomposition of $\mathbf{S}_2(\mathbf{x}^{st})$ is the same $\mathbf{D}$ from the Definition 7 of the FOBI functional. The latent components $z_1, \ldots, z_p$ are then obtained as $\mathbf{z} = \mathbf{W}\mathbf{x}$. The intuition behind $\mathbf{W} = \mathbf{U}\mathbf{S}_1^{-1/2}$ is that $\mathbf{W} = \mathbf{U}\mathbf{S}_1^{-1/2}$ gives latent components $\mathbf{z} = \mathbf{W}\mathbf{x}$ obtained by first whitening $\mathbf{x}$ with respect to $\mathbf{S}_1$ and then choosing $\mathbf{z}$ to be the principal components, with respect to $\mathbf{S}_2$ of the whitened $\mathbf{x}$.

In [19] it is shown that in the ICA model the diagonal elements of $\mathbf{D}$, $d_1, \ldots, d_p$ correspond to kurtosis measures of latent variables $\mathbf{z}$ yielding $d_i = 1$ if and only if $\mathbf{E}(z_i^4) = 3$. Thus, in ICA, the FOBI functional is well-defined (up to signs) if all independent components have distinct kurtoses and in that case $\mathbf{z}$ corresponds to the original independent components up to signs and order.

FOBI was originally suggested as an ICA method in [5] and considered in an exploratory data analysis context in [22], and for NGCA and NGICA for example in [25], while recently reviewed in [29].

Recently it was discovered that not only the combination $\mathbf{COV}$ and $\mathbf{COV}_4$ is useful but that in general

$$\mathbf{W}(\mathbf{x})\mathbf{S}_1(\mathbf{x})\mathbf{W}(\mathbf{x})^\top = \mathbf{I}_p \quad \text{and} \quad \mathbf{W}(\mathbf{x})\mathbf{S}_2(\mathbf{x})\mathbf{W}(\mathbf{x})^\top = \mathbf{D}(\mathbf{x}),$$

is of interest outside of an elliptical model where $\mathbf{S}_1$ and $\mathbf{S}_2$ can be arbitrary scatter functionals or are sometimes required to satisfy certain properties. The reason why the combination $\mathbf{S}_1 - \mathbf{S}_2$ is considered especially outside an elliptical model is that if $\mathbf{x}$ has an elliptical distribution all scatters calculated at $\mathbf{x}$, provided that they exist, are proportional to each other.

In [23, 30] it is shown that any two scatter functionals which have the full independence property can be used to as an ICA method. The approach as a general exploratory method was introduced as invariant coordinate selection (ICS) [38] and useful for example for finding groups or outliers and as a transformation-retransformation method in multivariate nonparametrics [1, 24, 38]. For the exploratory use, there are also some guidelines provided in [38] on how to choose the two scatters while arguing that there is no general best combination.

For two squared dispersion measures S_1 and S_2, one can define a generalized kurtosis measure with respect to S_1–S_2 as $\mathrm{Ku}(x) = S_2(x)/S_1(x)$. Furthermore, for scatter functional $\mathbf{S}$ and random vector $\mathbf{z} = (z_1, \ldots, z_p)$, $S(z_i) := \mathbf{e}_i^\top \mathbf{S}(\mathbf{z})\mathbf{e}_i = \mathbf{S}(\mathbf{z})_{ii}$ is a squared dispersion measure for every $i = 1, \ldots, p$, where $\mathbf{e}_i$ is the i-th vector of canonical bases of $\mathbb{R}^p$. In that manner, for two scatters $\mathbf{S}_1$ and $\mathbf{S}_2$, and a latent vector $\mathbf{z} = (z_1, \ldots, z_p)$, $\mathbf{S}_2(\mathbf{z})_{ii}/\mathbf{S}_1(\mathbf{z})_{ii}$ can be interpreted as generalized kurtosis measures for the corresponding latent component z_i, with respect to $\mathbf{S}_1$–$\mathbf{S}_2$, for every $i = 1, \ldots, p$. Relevant for our purpose is that for any combination, $\mathbf{S}_1$ and $\mathbf{S}_2$, of scatter functionals and for any vector $\mathbf{u} \in \mathbb{R}^p$, the diagonal elements $d_1, \ldots, d_p$ of $\mathbf{D}$ satisfy,

$$\frac{\mathbf{u}^\top \mathbf{S}_2(\mathbf{z})\mathbf{u}}{\mathbf{u}^\top \mathbf{S}_1(\mathbf{z})\mathbf{u}} = \sum_{i=1}^{p} u_i^2 d_i .$$

Therefore, for each i, $d_i = \mathbf{S}_2(\mathbf{z})_{ii}/\mathbf{S}_1(\mathbf{z})_{ii}$, gives the marginal kurtosis of z_i with respect to $\mathbf{S}_1$–$\mathbf{S}_2$. In that manner, standard kurtosis can be considered a kurtosis measure with respect to $\mathbf{COV}$–$\mathbf{COV}_4$.

In the following we will give results on how to use other scatter functionals besides the FOBI combination for NGCA and NGICA. Prior to stating any formal results we will introduce the following ordering. Let $(d_1, \ldots, d_p)$ be the vector in $\mathbb{R}^p$ such that $p - q$ of its components are all equal and the rest, q of them, mutually distinct and distinct from the $p - q$ equal ones. We say that it is ordered in *decreasing-to-equal* order if $d_1 > d_2 > \cdots > d_q$ and $d_{q+1} = \cdots = d_p$.

As the basic NGCA model has two independent blocks where at least the noise block is symmetric, basically any two scatter functionals can be used for this purpose.

Result 1 *Let* $\mathbf{x}$ *follow an NGCA model formulated using location functional* $\mathbf{T}$ *and scatter functional* $\mathbf{S}_1$ *and let* $\mathbf{S}_2$ *be a scatter functional different from* $\mathbf{S}_1$*. Write* $\mathbf{W} = \mathbf{U}^\top \mathbf{S}_1(\mathbf{x})^{-1/2}$*, where* $\mathbf{U}$ *is the matrix of unit eigenvectors of* $\mathbf{S}_2\left(\mathbf{S}_1^{-1/2}(\mathbf{x} - \mathbf{T}(\mathbf{x}))\right)$ *(with corresponding eigenvalues in decreasing-to-equal order). If there exists no such* q*-variate vector* $\mathbf{u}$ *with* $\mathbf{u}^\top\mathbf{u} = 1$ *such that* $\mathbf{u}^\top\mathbf{s}$ *has the same kurtosis in the* $\mathbf{S}_1$–$\mathbf{S}_2$ *sense as a Gaussian component and if all non-Gaussian components* $\mathbf{s}$ *have mutually distinct kurtoses in* $\mathbf{S}_1$–$\mathbf{S}_2$ *sense, then*

$$\mathbf{W}\mathbf{x} = ((\mathbf{O}_1\mathbf{s})^\top\ (\mathbf{O}_2\mathbf{n})^\top)^\top,$$

where $\mathbf{O}_1$, $\mathbf{O}_2$ *are orthogonal matrices.*

There should be $p-q$ equal elements in $\mathbf{D}$ which give the directions for the Gaussian subspace, however the specific value which corresponds to a Gaussian component might depend on $\mathbf{S}_1$, $\mathbf{S}_2$ and $\mathbf{s}$ and might therefore be difficult to identify in a finite data setting. Also as in general in NGCA, only the subspaces can be identified. Making the stronger assumption of an NGICA model helps in this case, but the chosen scatters are then required to have the block independence property.

Result 2 *Let* $\mathbf{x}$ *follow an NGICA model formulated using location functional* $\mathbf{T}$ *and scatter functional* $\mathbf{S}_1$ *and let* $\mathbf{S}_2$ *be a scatter functional different from* $\mathbf{S}_1$*, where* $\mathbf{S}_1$ *and* $\mathbf{S}_2$ *have the block-independence property. Write* $\mathbf{W} = \mathbf{U}^\top \mathbf{S}_1(\mathbf{x})^{-1/2}$*, where* $\mathbf{U}$ *is the matrix of unit eigenvectors of* $\mathbf{S}_2\left(\mathbf{S}_1^{-1/2}(\mathbf{x}-\mathbf{T}(\mathbf{x}))\right)$ *(with corresponding eigenvalues in decreasing-to-equal order). If there exists no such* q*-variate vector* $\mathbf{u}$ *with* $\mathbf{u}^\top\mathbf{u} = 1$ *such that* $\mathbf{u}^\top\mathbf{s}$ *has the same kurtosis in the* $\mathbf{S}_1$*–*$\mathbf{S}_2$ *sense as a Gaussian component and if all non-Gaussian components* $\mathbf{s}$ *have mutually distinct kurtoses in* $\mathbf{S}_1$*–*$\mathbf{S}_2$ *sense, then*

$$\mathbf{Wx} = ((\mathbf{Js})^\top\ (\mathbf{On})^\top)^\top,$$

where $\mathbf{J}$ *is a diagonal matrix with diagonal elements* $1,\ -1$ *and* $\mathbf{O}$ *is an orthogonal matrix.*

The requirement of block independence property can be relaxed under certain circumstances.

Result 3 *Let* $\mathbf{x}$ *follow an NGICA model formulated using location functional* $\mathbf{T}$ *and scatter functional* $\mathbf{S}_1$ *such that all but one component of* $\mathbf{s}$ *are symmetric and let* $\mathbf{S}_2$ *be a scatter functional different from* $\mathbf{S}_1$*. Write* $\mathbf{W} = \mathbf{U}^\top \mathbf{S}_1(\mathbf{x})^{-1/2}$*, where* $\mathbf{U}$ *is the matrix of unit eigenvectors of* $\mathbf{S}_2\ (\mathbf{S}_1(\mathbf{x}))$ *(with corresponding eigenvalues in decreasing-to-equal order). If there exists no such* q*-variate vector* $\mathbf{u}$ *with* $\mathbf{u}^\top\mathbf{u} = 1$ *such that* $\mathbf{u}^\top\mathbf{s}$ *has the same kurtosis in the* $\mathbf{S}_1$*–*$\mathbf{S}_2$ *sense as a Gaussian component and if all non-Gaussian components* $\mathbf{s}$ *have mutually distinct kurtoses in* $\mathbf{S}_1$*–*$\mathbf{S}_2$ *sense, then*

$$\mathbf{Wx} = ((\mathbf{Js})^\top\ (\mathbf{On})^\top)^\top,$$

where $\mathbf{J}$ *is a diagonal matrix with diagonal elements* $1,\ -1$ *and* $\mathbf{O}$ *is an orthogonal matrix.*

To conclude this section we would, however, like to point out that in NGCA and NGICA the Gaussian subspace can still be separated from the non-Gaussian subspace if the kurtoses in $\mathbf{S}_1$–$\mathbf{S}_2$ sense of the signals are not distinct as long as they differ from the corresponding Gaussian value.

4 Testing the Signal Dimension in NGCA and NGICA

FOBI is such a popular functional since it is solely moment based and therefore analytical considerations are fairly easy. However, it requires strong moment assumptions and suffers from a lack of robustness. In the NGCA and NGICA context the

FOBI functional has the advantage that the values in $\mathbf{D}$ of Gaussian components are known to be one. Therefore, in these models, in [25, 27] is suggested the testing procedure to test the hypothesis

$$H_{0k} : \text{There are exactly } k \text{ non-Gaussian components}$$

by testing that there are $p - k$ eigenvalues in $\mathbf{D}$ equal to 1.

The criterion used in [25, 27], to identify the eigenvalues which are closest to 1, is $(d_i - 1)^2$, thus the variance of the $p - k$ elements of $\mathbf{D}$ closest to 1 is used as the test statistic. Denote $d_{(i)}$, $i = 1, \ldots, p$ the ascending ordered eigenvalues in the sense above, then the test statistic from [25, 27] for a sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is

$$T_k = n \sum_{i=1}^{p-k} \left(d_{(i)} - 1\right)^2 .$$

In [25, 27] it is then shown that assuming $\mathrm{E}(z_i^4)$ exist for $i = 1, \ldots, p$ and that there is no q-variate vector $\mathbf{u}$ with $\mathbf{u}^\top \mathbf{u} = 1$ such that $\mathrm{E}((\mathbf{u}^\top \mathbf{s})^4) = 3$, where $\mathbf{s}$ is the signal component, one can use FOBI for estimating the signal and noise subspaces in NGCA and NGICA models as well as making inference about their dimensions.

Before stating the result that gives the limiting distribution of the test statistic T_k and enables for testing of H_{0k}, $k \in \{1, \ldots, p\}$, we define $\mathbf{U}_k$ to be the $p \times k$ matrix of eigenvectors of $\mathbf{S}_2$ that correspond to the aforementioned $p - k$ eigenvalues in $\mathbf{D}$ that are closest to 1, and the statistic $T_k^* = n\,\mathrm{tr}(((\mathbf{0}, \mathbf{I}_{p-k})\mathbf{U}_k(\mathbf{S}_2 - \mathbf{I}_p)\mathbf{U}_k^\top(\mathbf{0}, \mathbf{I}_{p-k})^\top)^2)$. The statistic T_k^* then corresponds to the test statistic for testing H_{0k} in case where the noise part is known.

Result 4 *Under the previously stated assumptions and under* H_{0q}

1. *for* $k < q$, $(p + 2)^2 T_k \to_P c$ *for some* $c > 0$ *as* $n \to \infty$,
2. *for* $k = q$, $(p + 2)^2 T_k \to_d C_k$ *as* $n \to \infty$ *and*
3. *for* $k > q$, $(p + 2)^2 T_k \le (p + 2)^2 T_k^* \to_d C_k$, *as* $n \to \infty$,

where

$$C_k \sim 2\sigma_1 Q_1 + (2\sigma_1 + \sigma_2(p - k)) Q_2,$$

where Q_1, Q_2 *are independent, chi-squared distributed, random variables with* $(p - k - 1)(p - k + 2)/2$ *and* 1 *degrees of freedom respectively, and* $\sigma_1^2 = Var(||\mathbf{z}||^2) + 8$, $\sigma_2 = 4$.

The proof of the Result 4 can be found in [25]. In this setting, the null hypothesis is rejected if $T_k \ge c_{k,\alpha}$, where $c_{k,\alpha}$ is chosen so that $\mathrm{P}(C_k \ge c_{k,\alpha}) = \alpha$. Note that, in order to find $c_{k,\alpha}$ one must consistently estimate σ_1. If we write $\hat{\mathbf{z}}_i = \hat{\mathbf{W}}(\mathbf{x}_i - \bar{\mathbf{x}})$ $i = 1, \ldots, n$, then in the NGICA model we have $\sigma_1 = \sum_{k=1}^{p} \mathrm{E}(z_k^4) - p + 8$, with a consistent estimate

$$\hat{\sigma}_1 = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{p} (\hat{z}_i)_k^4 - p + 8.$$

In the wider NGCA model σ_1 can be consistently estimated by

$$\hat{\sigma}_1 = \tfrac{1}{n}\sum_{i=1}^{n} ||\hat{z}_i||^4 - p^2 + 8.$$

Besides T_k [25] proposes also alternative for this problem such as

$$\frac{(p+2)^2 T_{k,1}}{2\hat{\sigma}_1^2} \quad \text{and} \quad \frac{(p+2)^2 T_{k,2}}{2\hat{\sigma}_1^2 + 4(p-k)},$$

where $T_{k,1} = n\left(\sum_{i=1}^{p-k} d_{(i)}^2 - \left(\sum_{i=1}^{p-k} d_{(i)}\right)^2\right)$ and $T_{k,2} = n\left(\sum_{i=1}^{p-k}(d_{(i)}-1)\right)^2$. Under the true H_{0k}, proposed test statistics have chi-squared distributions with $(p-k-1)(p+2-k)/2$ and 1 degrees of freedom respectively. One can show that $T_{k,1} + T_{k,2} \sim \chi^2_{(p-k-1)(p+2-k)/2+1}$, and argue that $T_{k,1}$ provides a test statistic for testing the equality of $p-k$ eigenvalues closest to 1, while $T_{k,2}$ measures the deviation of the mean of those eigenvalues from the theoretical value of one. In [25] it is also argued that those two statistics use less information than T_k, and are therefore in most cases less powerful and that the limiting behaviour of their sum is quite similar to the one of T_k.

Result 4 gives the limiting distribution of T_k, and therefore when using it in practice, due to the involvement of higher order moments, one might need very large sample sizes for the result to hold. For the case of small sample sizes, in [25] is proposed to estimate the distribution of the test statistic under the null by bootstrapping samples from distribution for which the null hypothesis H_{0k} is true and which is as close as possible to the empirical distribution of observed sample.

Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ be a data sample, and let $\bar{\mathbf{x}}$ denote the sample mean vector. Further, let $\hat{\mathbf{W}} = (\hat{\mathbf{W}}_1^\top \, \hat{\mathbf{W}}_2^\top)^\top$ be the sample estimates of the FOBI unmixing matrices where the partition $(\hat{\mathbf{W}}_1^\top \, \hat{\mathbf{W}}_2^\top)^\top$ was done according to the descending order of the eigenvalues in $\hat{\mathbf{D}}$ in sense as described in Sect. 3. Furthermore, let $\hat{\mathbf{S}} = (\hat{\mathbf{s}}_1, \ldots, \hat{\mathbf{s}}_n) = \hat{\mathbf{W}}_1(\mathbf{X} - \bar{\mathbf{x}}\mathbf{1}_n^\top) \in \mathbb{R}^{k\times n}$ and $\hat{\mathbf{N}} = (\hat{\mathbf{n}}_1, \ldots, \hat{\mathbf{n}}_n) = \hat{\mathbf{W}}_2(\mathbf{X} - \bar{\mathbf{x}}\mathbf{1}_n^\top) \in \mathbb{R}^{(p-k)\times n}$ be the matrices of the estimated signal and noise vectors, $\hat{\mathbf{s}}_i$ and $\hat{\mathbf{n}}_i$, $i \in 1, \ldots, n$ respectively. $\mathbf{1}_n$ denotes here an n-vector full of ones. The proposed strategy in the NGICA model is using non-parametric bootstrap to create matrices $\mathbf{S}^* \in \mathbb{R}^{k\times n}$ by componentwise(row-wise)-independently sampling with replacement from $\hat{\mathbf{S}}$, and using parametric bootstrap to create $\mathbf{N}^* \in \mathbb{R}^{(p-k)\times n}$ as a random sample from $N(\mathbf{0}, \mathbf{I}_{p-k})$. Resulting bootstrap sample is then $\mathbf{X}^* = \hat{\mathbf{W}}^{-1}(\mathbf{S}^{*\top} \, \mathbf{N}^{*\top})^\top$.

A similar approach for NGCA model was introduced in [27]. The strategy is to initially sample with replacement an n-dimensional sample $\tilde{\mathbf{X}} \in \mathbb{R}^{p\times n}$ from $\mathbf{X}$ and then estimate its signal matrix $\mathbf{S}^* = \hat{\mathbf{W}}_1\tilde{\mathbf{X}}$. In order for the noise space to be Gaussian transform $\tilde{\mathbf{X}}$ into $\mathbf{X}^* = \hat{\mathbf{W}}^{-1}(\mathbf{S}^{*\top} \, \mathbf{N}^{*\top})^\top$, where $\mathbf{N}^* \in \mathbb{R}^{(p-k)\times n}$ is an n-dimensional random sample from $N(\mathbf{0}, \mathbf{I}_{p-k})$.

We showed earlier that using the general two scatter functionals approach is possible for NGCA and NGICA given the scatter functionals fulfill certain properties.

However it is already not in general possible to say which eigenvalues correspond to directions indicating the Gaussian subspace. Thus deriving a general asymptotic test for any scatter combination does not sound feasible. However the bootstrap strategy described above for FOBI can be adapted.

One of the alternative test statistics mentioned earlier which considers only the variance of the eigenvalues can be used here, when adding the additional assumption that the Gaussian subspace is larger than any set of the signal subspaces which would share the same eigenvalue, which is for example in NGICA anyway required.
where $d_{(1)}, \dots d_{(p-k)}$ are those $p-k$ eigenvalues of $\hat{\mathbf{S}}_2$ that have the smallest variance of all $p-k$ subsets of the set of eigenvalues of $\hat{\mathbf{S}}_2$. $\hat{t}_k$ is then the estimator of the variance of those $p-k$ eigenvalues of $\mathbf{S}_2$ that correspond to the Gaussian components.

Hence, for $k \in \{0, \dots, p-2\}$ one can test H_{0k} by examining the variance of the $p-k$ eigenvalues closest together in that sense. In that manner we propose a bootstrap procedure that uses two scatter matrices $\mathbf{S}_1$, $\mathbf{S}_2$ and a location functional $\mathbf{T}$ and starts with a sample $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n) \in \mathbb{R}^{p \times n}$. Using sample estimators $\hat{\mathbf{T}}$ and $\hat{\mathbf{S}}_1$ of $\mathbf{T}$ and $\mathbf{S}_1$ respectively, scatter estimator $\mathbf{S}_2$ as its sample estimate based on standardized sample $\hat{\mathbf{S}}_1^{-1/2}(\mathbf{X} - \hat{\mathbf{T}}\mathbf{1}_n^\top)$, and calculates corresponding unmixing matrix $\hat{\mathbf{W}}$ as discussed in Sect. 3. The test statistic used for testing H_{0k} is then

$$\hat{t}_k = n \sum_{i=1}^{p-k} \left(d_{(i)} - \frac{1}{p-k} \sum_{j=1}^{p-k} d_{(j)} \right)^2 ,$$

where $d_{(1)}, \dots, d_{(p-k)}$ are those $p-k$ eigenvalues of $\hat{\mathbf{S}}$ contained in $\hat{\mathbf{D}}$ that have the smallest variance of all $p-k$ - subsets of the set of eigenvalues in $\hat{\mathbf{D}}$. Thus, $\hat{t}_k$ is the estimator of the variance of those $p-k$ eigenvalues of $\hat{\mathbf{S}}_2$ that correspond to the Gaussian components.

Once the eigenvalues corresponding to the signal and noise space have been identified one can order the diagonal elements of $\hat{\mathbf{D}}$ in a way that the last $p-k$ eigenvalues form a $p-k$ - subset of set of all eigenvalues of $\hat{\mathbf{S}}_2$ with the minimal variance, and obtain the corresponding partitioning of $\hat{\mathbf{W}} = (\hat{\mathbf{W}}_1^\top \, \hat{\mathbf{W}}_2^\top)^\top$. Finally, the signal and the noise parts of the latent sample $\mathbf{Z}$ are estimated as $\hat{\mathbf{s}}_i = \hat{\mathbf{W}}_1(\mathbf{x}_i - \hat{\mathbf{T}})$ and $\hat{\mathbf{n}}_i = \hat{\mathbf{W}}_2(\mathbf{x}_i - \hat{\mathbf{T}})$ respectively yielding the matrices $\hat{\mathbf{S}} \in \mathbb{R}^{k \times n}$ and $\hat{\mathbf{N}} \in \mathbb{R}^{(p-k) \times n}$ which collect the estimated signal and noise vectors.

Since the bootstrapping strategy for the signal part is dependent on the model, in the NGCA model we use the non-parametric bootstrap to create the signal sample $\mathbf{S}^*$ by sampling with replacement from $\hat{\mathbf{S}}$.

In the NGICA model, where signal components are mutually independent, we use non-parametric bootstrap to create matrix $\mathbf{S}^* \in \mathbb{R}^{k \times n}$ by componentwise(row-wise)-independently sampling with replacement from $\hat{\mathbf{S}}$.

We also propose two strategies for sampling the noise component. Parametric bootstrap creates noise sample $\mathbf{N}^* \in \mathbb{R}^{(p-k)\times n}$ as a random sample from $N(\mathbf{0}, \mathbf{COV}(\mathbf{N}))$, while the nonparametric bootstrap creates noise sample $\mathbf{N}^* = (\mathbf{n}_1^*, \dots, \mathbf{n}_n^*) \in \mathbb{R}^{(p-k)\times n}$, such that $\mathbf{n}_i^* \leftarrow \mathbf{O}_i \hat{\mathbf{n}}_i$, $i = 1, \dots, n$, where $\mathbf{O}_i$ is a random orthogonal $p - k \times p - k$ matrix. The nonparametric strategy does not directly target a normal noise but assumes spherical noise as a proxy.

For the latent component sample $\mathbf{Z}^* = (\mathbf{S}^{*\top}\, \mathbf{N}^{*\top})^\top$ obtained by bootstrapping procedure explained above, set $\mathbf{X}^* = \mathbf{W}^{-1}\mathbf{Z}^*$. Finally, assuming that $\mathbf{X}_1^*, \dots, \mathbf{X}_M^*$ are M independent bootstrap samples obtained as described above and $\hat{t}_{i,k}^* = \hat{t}_k(\mathbf{X}_i^*)$ are the corresponding test statistics, the bootstrap p-value is given by

$$\hat{p} = \frac{\#(\hat{t}_{i,k}^* \geq \hat{t}_k) + 1}{M + 1}.$$

The bootstrapping procedure for the combination of any two scatters is given in a schematic view in Algorithm 1.

Algorithm 1 Algorithm for testing $H_{0k} : q = k$.

```
Set the proposed dimension $k$; Set the number of bootstrap samples $M$; Choose two scatter functionals $\mathbf{S}_1$ and $\mathbf{S}_2$ and location functional $\mathbf{T}$; Starting with the observed sample $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$, $\mathbf{x}_i \in \mathbb{R}^p$ estimate $\hat{\mathbf{T}} = \mathbf{T}(\mathbf{X})$, $\hat{\mathbf{S}}_1 = \mathbf{S}_1(\mathbf{X})$; Calculate centered and standardized sample $\mathbf{X}^c = (\mathbf{x}_1^c, \dots, \mathbf{x}_n^c)$ and $\mathbf{X}^{st} = (\mathbf{x}_1^{st}, \dots, \mathbf{x}_n^{st})$ respectively, where $\mathbf{x}_i^c = \mathbf{x}_i - \hat{\mathbf{T}}$, $\mathbf{x}_i^{st} = \hat{\mathbf{S}}_1^{-1/2}(\mathbf{x}_i - \hat{\mathbf{T}})$, $i = 1, \dots, n$; Estimate $\hat{\mathbf{S}}_2 = \mathbf{S}_2(\mathbf{X}^{st})$ and calculate its eigenvalue-eigenvector decomposition $\hat{\mathbf{S}}_2 = \hat{\mathbf{U}}\hat{\mathbf{D}}\hat{\mathbf{U}}^\top$; Calculate two-scatter functional $\hat{\mathbf{W}} = \hat{\mathbf{U}}\hat{\mathbf{S}}_1^{-1/2}$; Order eigenvalues in $\hat{\mathbf{D}}$ so that the variance of the last $p - k$ eigenvalues in $\hat{\mathbf{D}}$ is minimal and derive the corresponding partitioning of $\hat{\mathbf{W}} = (\hat{\mathbf{W}}_1^\top\, \hat{\mathbf{W}}_2^\top)^\top$; Compute the test statistic $\hat{t}_k = n \sum_{i=1}^{p-k} \left( d_i - \frac{1}{p-k} \sum_{j=1}^{p-k} d_j \right)^2$ as the estimate of the variance of the last $p - k$ eigenvalues in $\hat{\mathbf{D}}$; Calculate the signal estimate $\hat{\mathbf{S}} = (\hat{\mathbf{s}}_1, \dots, \hat{\mathbf{s}}_n) = \hat{\mathbf{W}}_1\mathbf{X}^c$ and the noise estimate $\hat{\mathbf{N}} = (\hat{\mathbf{n}}_1, \dots, \hat{\mathbf{n}}_n) = \hat{\mathbf{W}}_2\mathbf{X}^c$; Choose a bootstrapping strategy for the noise; Choose the model suitable for the signal; for $j \in \{1, \dots, M\}$ do
    if Strategy = parametric bootstrap then
        $\mathbf{n}_i^* \leftarrow N_{p-k}(\mathbf{0}, \mathbf{COV}(\hat{\mathbf{N}}))$, $i = 1, \dots, n$;
    if Strategy = nonparametric bootstrap then
        $\mathbf{n}_i^* \leftarrow \mathbf{O}_i\hat{\mathbf{n}}_i$, $i = 1, \dots, n$, where $\mathbf{O}_i$ is a random orthogonal $p - k \times p - k$ matrix;
    if Model = NGCA then
        Sample $\mathbf{S}^*$ with replacement from $\hat{\mathbf{S}}$;
    if Model = NGICA then
        For each $j = 1, \dots, k$ sample with replacement $j$-th signal component $(s_{j,1}^*, \dots, s_{j,n}^*) \leftarrow (\hat{s}_{j,1}, \dots, \hat{s}_{j,n})$, and set $\mathbf{S}^* = [s_{i,j}^*]$
    Compute $\mathbf{X}^* = \hat{\mathbf{W}}^{-1}(\mathbf{S}^{*\top}\, \mathbf{N}^{*\top})^\top$; Compute $\hat{t}_{j,k}^*$ based on $\mathbf{X}^*$;
Return bootstrap $p$-value: $\hat{p}_k = [\#(\hat{t}_{j,k}^* \geq \hat{t}_k) + 1]/(M + 1)$;
```

5 Performance Evaluation of the Test

The following simulation study is performed using R 3.6.1 [31] with the packages SpatialNP [34], ICtest [26], JADE [17], ICS [24], png [39], RcppRoll [40] and extraDistr [42], and it was conducted to compare the bootstrap FOBI test from [25] to four different testing procedures based on Algorithm 1 with the expectation as the location functional and the following pairs of scatter matrices:

1. *Cov-Cov4*: $\mathbf{S}_1 = \mathbf{COV}$, $\mathbf{S}_2 = \mathbf{COV}_4$.

Note that there is a difference between the "FOBI" and the *Cov-Cov4* testing procedures. In the "FOBI" denoted case the information that the noise eigenvalues should be one is used while in the *Cov-Cov4* denoted case Algorithm 1 is used ignoring this information.

2. *Cau-Hub*: $\mathbf{S}_1$ is M-estimator based on the likelihood of the t-distribution with one degree of freedom ($\nu = 1$), also known as the Cauchy distribution. $\mathbf{S}_2$ is an M-estimator based on Huber's weight function.
3. *sCau-sHub*: is the symmetrized version of the previous setting, thus a symmetrized M-scatter based on the Cauchy distribution and a symmetrized M-scatter based on Huber's weight function.

As estimation of both scatters in *sCau-sHub* is computationally very expensive and not feasible in the large data sets we follow a suggestion from [18] to base the symmetrized scatters not on all pairwise differences but only on an "incomplete" set which makes it much easier to compute. For details see [18].

4. *sCauI-sHubI*: is the incomplete combination of symmetrized scatters. We compute both scatters so that all observations are contained in 100 differences.

For more details on the computation of all the scatters see also the documention of the R-packages SpatialNP [34] and ICS [24].

Due to the computational costs and as it seems more natural to us, in all four settings always parametric bootstrap is used for the noise part.

To compare the bootstrap tests, we consider two different settings which both are 6-variate and have each 3 signal and 3 noise components. Model $M1$ follows an NGCA model and model $M2$ an NGICA model. In all cases the 6×6 matrix $\mathbf{A}$ was simulated in each iteration independently by filling it with random $N(0, 1)$ elements. The two models used are:

$M1$: An NGCA model with two non-Gaussian univariate components $\mathbf{s}_1$ and $\mathbf{s}_2$, representing x and y axis of the Greek letter Γ respectively, a non-Gaussian univariate component $\mathbf{s}_3$ with χ^2_1 distribution and three independent Gaussian components $N(0, 1)$. Hence, $p = 6$, $q = 3$. Figure 1 visualizes the three non-Gaussian components of this setting.

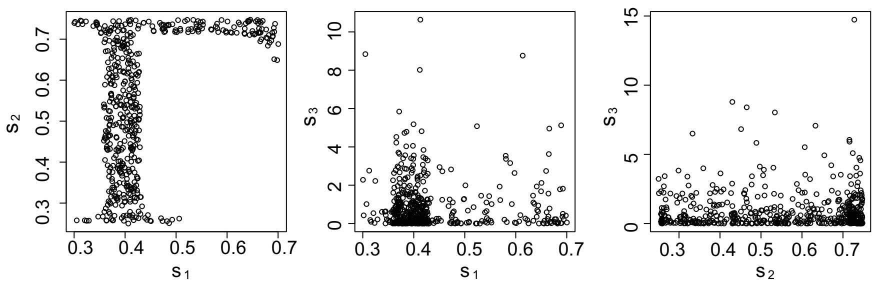


Fig. 1 Scatter plots of signal components in $M1$ based on a sample of 500

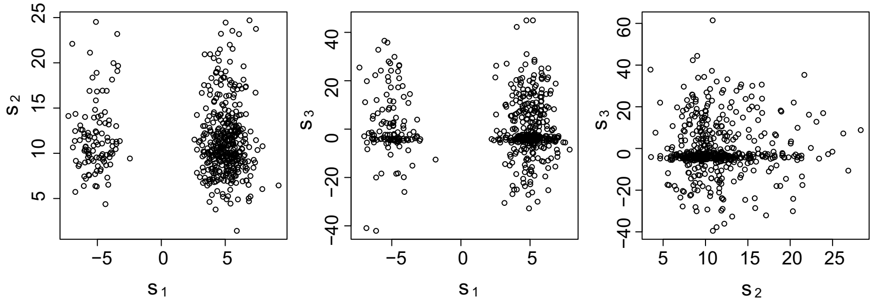


Fig. 2 Scatter plots of signal components in $M2$ based on a sample of 500

$M2$: An NGICA model with three independent components which all follow a Gaussian mixture model (GMM) with different parameter settings: $s_1 \sim (3+\sqrt{3})^{-1}\phi_{-5,1} + (1-(3+\sqrt{3})^{-1})\phi_{5,1}$, $s_2 \sim 0.7\phi_{10,2} + 0.3\phi_{15,5}$ and $s_3 \sim 0.4\phi_{-4,1} + 0.6\phi_{2,15}$, where $\phi_{\mu,\sigma}$ denotes the pdf of the normal distribution with mean μ and variance σ^2. The three noise components are independent $N(0,1)$. Therefore, $p = 6$, $q = 3$. For more insight into the shape of the non-Gaussian components see Fig. 2.

Note that if a random variable x comes from the two-component GMM, with equal variances for the components and the mixing probability is $(3+\sqrt{3})^{-1}$, then its kurtosis is equal to 3 for all choices of means of two components. Therefore, in the model M2, the kurtosis of the component s_1 equals 3. Hence, the requirements of Result 2 are violated when the scatter combination $\mathbf{S}_1 = \mathbf{COV}$ and $\mathbf{S}_2 = \mathbf{COV}_4$ is used. Thus it is to be expected that neither *Cov-Cov4* nor FOBI will be able to separate s_1 form the Gaussian components, which should result in very low rejection rates in testing for H_{02}.

In order to gain insight into the robustness of the proposed testing procedures we consider also the case when in the two settings small contamination is added. The perturbed models are denoted $M1_x$ and $M2_x$ respectively and are obtained by adding an additional perturbation (equal to $10\ \mathbf{1}_6$) to 0.5% of the mixed observations.

Table 1 $M1$ assuming NGCA model: rejection rates in 1000 repetitions for bootstrap tests of H_{02}, H_{03} (true) and H_{04}, with $\alpha = 0.05$

n	FOBI (boot)			*Cov-Cov4*			*Cau-Hub*			*sCau-sHub*			*sCauI-sHubI*		
	$k2$	$k3$	$k4$	$k2$	$k3$	$k4$	$k2$	$k3$	$k4$	$k2$	$k3$	$k4$	$k2$	$k3$	$k4$
500	0.363	0.035	0.016	0.360	0.088	0.031	0.954	0.061	0.017	0.788	0.055	0.027	0.233	0.091	0.083
1000	0.553	0.057	0.016	0.553	0.074	0.031	1.000	0.050	0.011	1.000	0.053	0.018	0.717	0.072	0.059
2000	0.839	0.049	0.015	0.801	0.065	0.024	1.000	0.051	0.012	1.000	0.044	0.016	0.994	0.055	0.045
4000	0.986	0.057	0.012	0.977	0.051	0.012	1.000	0.055	0.013					0.045	0.035

Table 2 $M1$ assuming NGICA model: rejection rates in 1000 repetitions for bootstrap tests of H_{02}, H_{03} (true) and H_{04}, with $\alpha = 0.05$

n	FOBI (boot)			*Cov-Cov4*			*Cau-Hub*			*sCau-sHub*			*sCauI-sHubI*		
	$k2$	$k3$	$k4$	$k2$	$k3$	$k4$	$k2$	$k3$	$k4$	$k2$	$k3$	$k4$	$k2$	$k3$	$k4$
500	0.384	0.047	0.033	0.402	0.128	0.061	0.967	0.190	0.119	0.813	0.148	0.078	0.302	0.168	0.149
1000	0.539	0.043	0.030	0.555	0.084	0.043	1.000	0.114	0.067	1.000	0.079	0.049	0.734	0.161	0.135
2000	0.831	0.035	0.029	0.804	0.054	0.040	1.000	0.077	0.057	1.000	0.046	0.035	0.996	0.096	0.072
4000	0.985	0.049	0.030	0.976	0.039	0.028	1.000	0.068	0.049					0.050	0.052

For all samples $\mathbf{X} \in \mathbb{R}^{n \times p}$ from models $M1$, $M2$, $M1_x$ and $M2_x$, with sample sizes $n = 500,\ 1000,\ 2000,\ 4000$, the bootstrap p-values based on $M = 200$ bootstrap samples were computed using the five tests described above where we use only parametric bootstrapping for the noise part. This is due to the computational complexity of the simulation and as it seems to be a more natural suggestion. We performed all the bootstrap tests once assuming an NGCA model and once assuming an NGICA model. 1000 repetitions where performed at the level $\alpha = 0.05$ and Tables 1, 2, 3, 4, 5, 6, 7 and 8 report the rejection rates for H_{02}, H_{03}(true) and H_{04} in all discussed settings. In the case $n = 4000$ also due to computational complexity the tests *sCau-sHub* have not been performed. In our settings the non-FOBI combinations should all be able to separate the signal and noise subspaces but only the symmetrised scatters would actually be able to recover the individual signal components in model $M2$.

Table 3 $M1_x$ assuming NGCA model: rejection rates in 1000 repetitions for bootstrap tests of H_{02}, H_{03} (true) and H_{04}, with $\alpha = 0.05$

n	FOBI (boot)			*Cov-Cov4*			*Cau-Hub*			*sCau-sHub*			*sCauI-sHubI*		
	$k2$	$k3$	$k4$	$k2$	$k3$	$k4$	$k2$	$k3$	$k4$	$k2$	$k3$	$k4$	$k2$	$k3$	$k4$
500	0.299	0.132	0.034	0.189	0.109	0.087	0.952	0.063	0.019	0.713	0.095	0.035	0.262	0.097	0.086
1000	0.183	0.092	0.006	0.149	0.103	0.072	1.000	0.054	0.012	0.963	0.188	0.045	0.674	0.129	0.096
2000	0.366	0.102	0.008	0.273	0.089	0.058	1.000	0.047	0.013	0.996	0.274	0.045	0.943	0.173	0.054
4000	0.705	0.159	0.019	0.568	0.147	0.049	1.000	0.062	0.021				0.996	0.306	0.074

Table 4 $M1_x$ assuming NGICA model: rejection rates in 1000 repetitions for bootstrap tests of H_{02}, H_{03} (true) and H_{04}, with $\alpha = 0.05$

n	FOBI (boot)			*Cov-Cov4*			*Cau-Hub*			*sCau-sHub*			*sCauI-sHubI*		
	$k2$	$k3$	$k4$	$k2$	$k3$	$k4$	$k2$	$k3$	$k4$	$k2$	$k3$	$k4$	$k2$	$k3$	$k4$
500	0.320	0.160	0.099	0.202	0.231	0.159	0.966	0.216	0.129	0.743	0.229	0.087	0.271	0.153	0.116
1000	0.191	0.131	0.042	0.145	0.206	0.150	1.000	0.133	0.061	0.961	0.343	0.110	0.677	0.244	0.116
2000	0.369	0.135	0.056	0.279	0.178	0.120	1.000	0.079	0.045	0.997	0.429	0.116	0.931	0.358	0.151
4000	0.708	0.194	0.049	0.581	0.213	0.120	1.000	0.083	0.039				0.998	0.472	0.115

Table 5 $M2$ assuming NGCA model: rejection rates in 1000 repetitions for bootstrap tests of H_{02}, H_{03} (true) and H_{04}, with $\alpha = 0.05$

n	FOBI (boot)			*Cov-Cov4*			*Cau-Hub*			*sCau-sHub*			*sCauI-sHubI*		
	$k2$	$k3$	$k4$	$k2$	$k3$	$k4$	$k2$	$k3$	$k4$	$k2$	$k3$	$k4$	$k2$	$k3$	$k4$
500	0.076	0.014	0.007	0.088	0.047	0.040	0.938	0.107	0.095	0.995	0.081	0.051	0.817	0.220	0.112
1000	0.076	0.025	0.009	0.075	0.021	0.015	0.999	0.067	0.055	1.000	0.076	0.064	0.993	0.113	0.091
2000	0.064	0.007	0.007	0.050	0.014	0.007	1.000	0.029	0.043	1.000	0.029	0.043	1.000	0.050	0.107
4000	0.064	0.016	0.003	0.052	0.017	0.010	1.000	0.052	0.021				1.000	0.048	0.055

Table 6 $M2$ assuming NGICA model: rejection rates in 1000 repetitions for bootstrap tests of H_{02}, H_{03} (true) and H_{04}, with $\alpha = 0.05$

n	FOBI (boot)			*Cov-Cov4*			*Cau-Hub*			*sCau-sHub*			*sCauI-sHubI*		
	$k2$	$k3$	$k4$	$k2$	$k3$	$k4$	$k2$	$k3$	$k4$	$k2$	$k3$	$k4$	$k2$	$k3$	$k4$
500	0.122	0.052	0.035	0.142	0.099	0.098	0.960	0.369	0.264	0.997	0.227	0.144	0.858	0.371	0.252
1000	0.112	0.051	0.028	0.095	0.064	0.039	0.999	0.247	0.195	1.000	0.115	0.100	0.995	0.196	0.159
2000	0.086	0.021	0.014	0.064	0.029	0.021	1.000	0.121	0.107	1.000	0.050	0.036	1.000	0.086	0.114
4000	0.072	0.036	0.028	0.052	0.031	0.029	1.000	0.117	0.095				1.000	0.060	0.058

Table 7 $M2_x$ assuming NGCA model: rejection rates in 1000 repetitions for bootstrap tests of H_{02}, H_{03} (true) and H_{04}, with $\alpha = 0.05$.

n	FOBI (boot)			*Cov-Cov4*			*Cau-Hub*			*sCau-sHub*			*sCauI-sHubI*		
	$k2$	$k3$	$k4$	$k2$	$k3$	$k4$	$k2$	$k3$	$k4$	$k2$	$k3$	$k4$	$k2$	$k3$	$k4$
500	0.264	0.069	0.033	0.403	0.200	0.095	0.923	0.110	0.083	0.993	0.263	0.105	0.763	0.237	0.130
1000	0.436	0.076	0.028	0.477	0.180	0.083	0.999	0.080	0.065	1.000	0.572	0.169	0.981	0.304	0.160
2000	0.700	0.107	0.036	0.686	0.221	0.093	1.000	0.064	0.050	1.000	0.900	0.143	1.000	0.579	0.171
4000	0.934	0.066	0.009	0.919	0.071	0.047	1.000	0.060	0.038				1.000	0.871	0.131

First we note in Tables 1-8 that the differences between FOBI and *Cov-Cov4* are rather small and probably mainly due to having different bootstrap samples. At least it is not obvious from these results that the knowledge of the value the eigenvalue of interest is of much relevance. It is however obvious that this combination of scatters does not work well in Model $M2$ as expected due to s_1.

Table 8 $M2_x$ assuming NGICA model: rejection rates in 1000 repetitions for bootstrap tests of H_{02}, H_{03} (true) and H_{04}, with $\alpha = 0.05$

n	FOBI (boot)			*Cov-Cov4*			*Cau-Hub*			*sCau-sHub*			*sCauI-sHubI*		
	$k2$	$k3$	$k4$	$k2$	$k3$	$k4$	$k2$	$k3$	$k4$	$k2$	$k3$	$k4$	$k2$	$k3$	$k4$
500	0.307	0.131	0.106	0.472	0.272	0.160	0.945	0.369	0.262	0.994	0.399	0.180	0.800	0.410	0.208
1000	0.489	0.164	0.116	0.555	0.271	0.135	0.999	0.244	0.193	1.000	0.653	0.205	0.992	0.436	0.208
2000	0.757	0.200	0.129	0.736	0.286	0.136	1.000	0.157	0.107	1.000	0.929	0.143	1.000	0.671	0.157
4000	0.940	0.083	0.060	0.917	0.128	0.112	1.000	0.183	0.119				1.000	0.876	0.157

Also from the robustness point of view the behaviour is as expected for this scatter combination and it performs poorly in the contaminated settings. In general, it seems that the combination *Cau-Hub* performs best. It works well in uncontaminated and contaminated cases while being more robust than the symmetrized counterparts. This is not a surprise as outliers have larger effects when symmetrizing and especially in the incomplete case. Comparing the symmetrized and incomplete symmetrized results it can be seen that the incomplete case starts to work in the uncontaminated settings when the sample sizes are sufficiently large, which is acceptable as it would anyway only be used when the usage of all pairwise differences would become too costly.

The knowledge whether the data actually follows an NGCA model or an NGICA model during bootstrap seems also only of minor relevance while the results in the NGICA model seem to be slightly worse than in the broader NGCA model, which can be simply due to difference in bootstrap samples. However, it is also possible that the difference in performance of bootstrap tests wrongly assuming NGICA and assuming NGCA would be larger in data sets where more dependence is introduced into signal components.

In Sect. 4 we suggested a strategy for testing the dimension of the signal space in NGCA and NGICA using any pair of scatter matrices. The simulation results show that under the null hypothesis of exactly $k = q$ non-Gaussian components, the alpha level is kept while the rejection frequencies are low if k is larger than q and high if k is smaller than q. This is in accordance with Result 4 which was derived however for FOBI only.

6 Estimation of the Signal Space Dimension

Usually the dimension q in NGCA or NGICA is unknown and therefore needs to be estimated from the data. The results from Sect. 5 encourage us to apply for this purpose the hypothesis tests successively. Different strategies for the successive testing are possible and while the test statistic is monotone in the dimension, its distribution is changing as can be seen from the FOBI results. Therefore different strategies might not yield the same dimension estimate.

In the following we will introduce two different strategies and compare them in a simulation study. The first strategy is denoted as the incremental strategy. This strategy assumes initially at least one Gaussian component and then tests successively, at level α, $H_{0k},\ k = p-2, \dots, 0$. The estimated $\hat{q}$ is the smallest k for which H_{0k} is not being rejected at level α, i.e.

$$\hat{q} = \min\{k \in \{0, \dots, p-2\} : H_{0k} \text{ is not being rejected}\}.$$

An algorithmic scheme is presented for this strategy in Algorithm 2.

Algorithm 2 Estimating dimension q of the signal subspace using an incremental approach

```
Set the proposed dimension $k = p - 2$; Set the significance level $\alpha$; Initiate the parameters of the Algorithm 1; repeat
    Test for $H_{0k}$ and compute bootstrap p-value $\hat{p}_k$ using Algorithm 1; if $\hat{p}_k > \alpha$ then
        $k = k - 1$
until $\hat{p}_k \le \alpha$ or $k = 0$;
Return the estimate $\hat{q} = k + 1$ of the signal dimension
```

For the incremental strategy the number of Gaussian components should be preferably small. If one suspects that this would not be the case, for example a divide and conquer strategy could be applied to find a point where acceptance switches to rejection at a specific level α. A possible variant for a divide and conquer strategy is presented in Algorithm 3.

Algorithm 3 Estimating dimension q of the signal subspace using divide and conquer strategy

```
Set the proposed dimension $k = \lceil \frac{p}{2} \rceil$; Set the significance level $\alpha$; Set $q^0_{min} = 1$ and $q^0_{max} = p - 1$;
Initiate the parameters of the Algorithm 1;
repeat
    Test $H_{0k}$ and $H_{0(k-1)}$ using Algorithm 1; if $H_{0k}$ is not rejected and $H_{0(k-1)}$ is rejected then
        Return $\hat{q} = k$;
    if $H_{0k}$ is not rejected and $H_{0(k-1)}$ is not rejected then
        $q^1_{min} \leftarrow q^0_{min}, q^1_{max} \leftarrow k - 1, k = \lceil \frac{q^1_{max} + q^1_{min}}{2} \rceil$
    if $H_{0k}$ is rejected then
        $q^1_{min} \leftarrow k + 1, q^1_{max} \leftarrow q^0_{max}, k = \lceil \frac{q^1_{max} + q^1_{min}}{2} \rceil$
    Update: $q^0_{min} \leftarrow q^1_{min}, q^0_{max} \leftarrow q^1_{max}$
until $q^0_{min} = q^0_{max}$;
Return the estimate $\hat{q} = k$ of the signal dimension
```

Naturally in both algorithms prior knowledge could be incorporated by adjusting the starting points of the procedures and also many other strategies are possible. As suggested in [27], a sequence of bootstrap test sizes α_k for testing H_{0k} can be determined so that the consistency of the procedure is preserved, but due to simplicity we will use fixed test sizes $\alpha_k = \alpha = 0.05,\ \forall k$.

We restrict ourselves to compare only these two strategies by adjusting models $M1$ and $M2$ slightly. In the adjusted models $M1^*$ and $M2^*$ the same signal components are used as in $M1$ and $M2$ respectively, but the number of Gaussian components is doubled to 6. As there was little difference in performance when bootstrapping an underlying NGCA or NGICA model, we restrict ourselves to assume an NGCA model. Moreover, encouraged by results presented in Tables 1-8 we compare only the scatter combinations *Cov-Cov4* and *Cau-Hub*, where all tests are executed at level $\alpha = 0.05$.

Based on 500 repetitions Fig. 3 shows the estimated signal dimensions.

Figure 3 shows that especially with increasing sample size correct dimensions are estimated in both models when using *Cau-Hub*, whereas as expected, *Cov-Cov4*

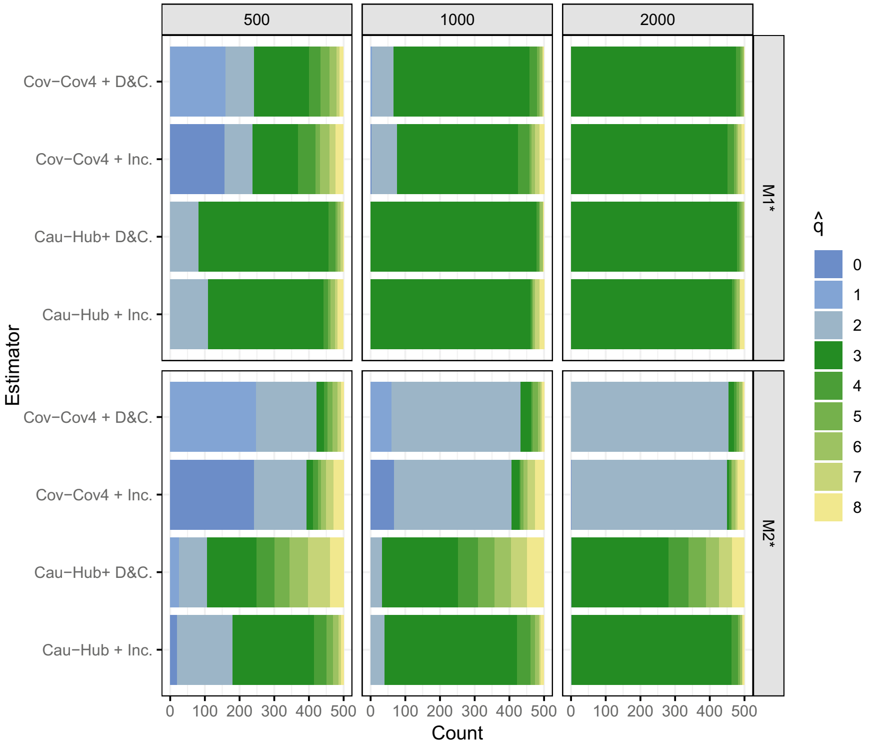


Fig. 3 Frequencies of estimated dimension of the non-Gaussian subspace for incremental strategy (Inc) and divide and conquer (D&C) strategy in models $M1^*$ and M2* based on 500 iterations when using different scatter combinations and different sample sizes

fails to recognize one signal in $M2^*$. It needs however also larger sample sizes compared to *Cau-Hub* in model $M1^*$. It also shows that there are differences between the strategies and at least here incremental strategy looks a bit better, which could possibly be justified by argumentation presented before in this section.

7 Conclusion

Dimension reduction is of increasing importance and quite often it is considered that the interesting subspace of the data is non-Gaussian. NGCA and NGICA are two dimension reduction approaches which follow these ideas and try to separate the Gaussian subspace from the non-Gaussian one. There are many methods suggested in the literature for NGCA and NGICA but usually they assume that the dimensions of the subspaces are known, which is rather unrealistic. In this paper we show under which conditions two different scatter matrices can be used to estimate the subspaces. Based on this approach we suggest also bootstrap tests to test for a specific subspace dimension and show how successive applications of the presented tests can be used to obtain an estimate of the dimensions of interest. A disadvantage of our suggestion is the computational complexity which also depends on the scatter matrices selected. Especially when using symmetrized scatters this becomes quite demanding, but if the sample sizes are large it seems that incomplete symmetrized scatters can be successfully used too. However, as we pointed out—usage of symmetrized scatters is actually not required if the goal is just to separate the two subspaces, since also non-symmetrized scatters can be rightfully used for the separation. It is just in the NGICA model that these combinations might not be able to recover the signals. Therefore, one strategy here could be to use computationally faster and often more robust regular scatter functionals in order to find the non-Gaussian subspace, and then to apply, on the estimated subspace, a regular ICA method, for example one based on two symmetrized scatter matrices, to estimate the independent components.

Acknowledgements The authors wish to express their gratitude to two anonymous referees, whose insightful comments greatly helped improving the quality of the manuscript. The work of KN was supported by the Austrian Science Fund (FWF) Grant number P31881-N32.

8 Appendix

Proof of the Result 1 Assume $\mathbf{x}$ follows an NGCA model formulated using location functional $\mathbf{T}$ and scatter functional $\mathbf{S}_1$, $\mathbf{x} = \mathbf{A}\mathbf{z} = \mathbf{A}_1\mathbf{s} + \mathbf{A}_2\mathbf{n}$, and let $\mathbf{S}_2$ be scatter functional different from $\mathbf{S}_1$.
Let $\mathbf{S}_2(\mathbf{x}^{st}) = \tilde{\mathbf{U}}\mathbf{D}\tilde{\mathbf{U}}^\top$ be eigen decomposition of $\mathbf{S}_2$, where $\mathbf{x}^{st} = \mathbf{S}_1(\mathbf{x})^{-1/2}\mathbf{x}$ and the eigenvalues in $\mathbf{D}$ are ordered so that $d_1 > \cdots > d_q$ and $d_{q+1} = \cdots = d_p$. Let $\mathbf{W} = \tilde{\mathbf{U}}^\top\mathbf{S}_1(\mathbf{x})^{-1/2}$ and $\mathbf{A} = \mathbf{U}\mathbf{L}\mathbf{V}$ be an SVD decomposition of mixing matrix $\mathbf{A}$. Since $\mathbf{x} = \mathbf{A}\mathbf{z}$,

$$\mathbf{S}_1(\mathbf{x})^{-1/2}\mathbf{x} = \mathbf{U}\mathbf{V}^\top\mathbf{z}, \quad \mathbf{S}_2(\mathbf{x}^{st}) = \mathbf{U}\mathbf{V}^\top\mathbf{S}_2(\mathbf{z})(\mathbf{U}\mathbf{V}^\top)^\top.$$

$\mathbf{S}_2(\mathbf{z})$ and $\mathbf{S}_2(\mathbf{x}^{st})$ are similar and thus have the same eigenvalues. Hence

$$\mathbf{S}_2(\mathbf{z}) = \mathbf{U}_B\mathbf{D}\mathbf{U}_B^\top \implies \mathbf{S}_2(\mathbf{x}^{st}) = \mathbf{U}\mathbf{V}^\top\mathbf{U}_B\mathbf{D}\mathbf{U}_B^\top(\mathbf{U}\mathbf{V}^\top)^\top,$$

where $\mathbf{U}_B$ is orthogonal matrix. Since $\mathbf{S}_2(\mathbf{x}^{st}) = \tilde{\mathbf{U}}\mathbf{D}\tilde{\mathbf{U}}^\top$ then $\tilde{\mathbf{U}} = \mathbf{U}\mathbf{V}^\top\mathbf{U}_B\mathbf{P}_B\mathbf{J}$, where $\mathbf{J}$ is a sign-changing matrix and $\mathbf{P}_B = \mathrm{diag}(\mathbf{I}_q, \mathbf{P}_{p-q})$ is block-diagonal matrix with the first block being an identity and the second block being a permutation matrix. Therefore

$$\mathbf{W} = (\mathbf{U}\mathbf{V}^\top\mathbf{U}_B\mathbf{P}_B\mathbf{J})^\top\mathbf{S}_1(\mathbf{x})^{-1/2}.$$

$\mathbf{S}_2(\mathbf{z})$ is a block-diagonal matrix implying that $\mathbf{U}_B$ is also block-diagonal, with orthogonal blocks $\mathbf{U}_{B1} \in \mathbb{R}^{q\times q}$ and $\mathbf{U}_{B2} \in \mathbb{R}^{(p-q)\times(p-q)}$. Hence,

$$\mathbf{W}\mathbf{x} = \mathbf{J}^\top((\mathbf{U}_{B1}^\top\mathbf{s})^\top\ (\mathbf{P}_{p-q}^\top\mathbf{U}_{B2}^\top\mathbf{n})^\top)^\top.$$

Proof of the Result 2 Assume $\mathbf{x}$ follows an NGICA model formulated using location functional $\mathbf{T}$ and scatter functional $\mathbf{S}_1$ with block-independence property, $\mathbf{x} = \mathbf{A}\mathbf{z} = \mathbf{A}_1\mathbf{s} + \mathbf{A}_2\mathbf{n}$, and let $\mathbf{S}_2$ be scatter functional different from $\mathbf{S}_1$ also having block-independence property.

Let $\mathbf{S}_2(\mathbf{x}^{st}) = \tilde{\mathbf{U}}\mathbf{D}\tilde{\mathbf{U}}^\top$ be eigen-decomposition of $\mathbf{S}_2(\mathbf{x}^{st})$, where $\mathbf{x}^{st} = \mathbf{S}_1(\mathbf{x})^{-1/2}\mathbf{x}$ and the eigenvalues in $\mathbf{D}$ are ordered so that $d_1 > \cdots > d_q$ and $d_{q+1} = \cdots = d_p$. Let $\mathbf{W} = \tilde{\mathbf{U}}^\top\mathbf{S}_1(\mathbf{x})^{-1/2}$ and $\mathbf{A} = \mathbf{U}\mathbf{L}\mathbf{V}$ be an SVD decomposition of mixing matrix $\mathbf{A}$. Since $\mathbf{x} = \mathbf{A}\mathbf{z}$,

$$\mathbf{S}_1(\mathbf{x})^{-1/2}\mathbf{x} = \mathbf{U}\mathbf{V}^\top\mathbf{z}, \quad \mathbf{S}_2(\mathbf{x}^{st}) = \mathbf{U}\mathbf{V}^\top\mathbf{S}_2(\mathbf{z})(\mathbf{U}\mathbf{V}^\top)^\top.$$

$\mathbf{S}_2(\mathbf{z})$ and $\mathbf{S}_2(\mathbf{x}^{st})$ are similar and thus have the same eigenvalues. Hence

$$\mathbf{S}_2(\mathbf{z}) = \mathbf{U}_B\mathbf{D}\mathbf{U}_B^\top \implies \mathbf{S}_2(\mathbf{x}^{st}) = \mathbf{U}\mathbf{V}^\top\mathbf{U}_B\mathbf{D}\mathbf{U}_B^\top(\mathbf{U}\mathbf{V}^\top)^\top,$$

where $\mathbf{U}_B$ is orthogonal matrix. Since $\mathbf{S}_2(x^{st}) = \tilde{\mathbf{U}}\mathbf{D}\tilde{\mathbf{U}}^\top$ then $\tilde{\mathbf{U}} = \mathbf{U}\mathbf{V}^\top\mathbf{U}_B\mathbf{P}_B\mathbf{J}$, where $\mathbf{J}$ is a sign-changing matrix and $\mathbf{P}_B = \mathrm{diag}(\mathbf{I}_q, \mathbf{P}_{p-q})$ is block-diagonal matrix with the first block being an identity and the second block being a permutation matrix. Therefore

$$\mathbf{W} = (\mathbf{U}\mathbf{V}^\top\mathbf{U}_B\mathbf{P}_B\mathbf{J})^\top\mathbf{S}_1(\mathbf{x})^{-1/2}.$$

$\mathbf{S}_2(\mathbf{z})$ is a block-diagonal matrix implying that $\mathbf{U}_B$ is also block-diagonal, with orthogonal blocks $\mathbf{I}_q \in \mathbb{R}^{q\times q}$ and $\mathbf{U}_{B2} \in \mathbb{R}^{(p-q)\times(p-q)}$. Hence,

$$\mathbf{W}\mathbf{x} = \mathbf{J}^\top(\mathbf{s}^\top\ (\mathbf{P}_{p-q}^\top\mathbf{U}_{B2}^\top\mathbf{n})^\top)^\top.$$

Proof of the Result 3 Assume $\mathbf{x}$ follows an NGICA model, $\mathbf{x} = \mathbf{A}\mathbf{z} = \mathbf{A}_1\mathbf{s} + \mathbf{A}_2\mathbf{n}$, and assume that all but one of one component of $\mathbf{s}$ are symmetric. Since $\mathbf{n}$ has Gaussian distribution, all but one of the independent blocks in $\mathbf{z}$ are symmetric implying that any scatter matrix $\mathbf{S}(\mathbf{z})$, provided that it exists at $\mathbf{z}$, has the block-independence property. Now, the Result 3 follows directly from Result 2.

References

1. Archimbaud, A., Nordhausen, K., Ruiz-Gazen, A.: ICS for multivariate outlier detection with application to quality control. Comput. Statistics Data Anal. **128**, 184–199 (2018)
2. Bean, D.M.: Non-Gaussian component analysis. Ph.D. thesis, University of California, Berkeley (2014)
3. Blanchard, G., Kawanabe, M., Sugiyama, M., Spokoiny, V., Müller, K.-R.: In search of non-Gaussian components of a high-dimensional distribution. J. Mach. Learn. Res. **7**, 247–282 (2006)
4. Blanchard, G., Sugiyama, M., Kawanabe, M., Spokoiny, V., Müller, K.-R.: Non-Gaussian component analysis: a semi-parametric framework for linear dimension reduction. In: Advances in Neural Information Processing Systems, pp. 131–138 (2005)
5. Cardoso, J.-F.: Source separation using higher order moments. In: International Conference on Acoustics, Speech, and Signal Processing, pp. 2109–2112 (1989)
6. Comon, P., Jutten, C.: Handbook of Blind Source Separation: Independent Component Analysis and Applications. Academic Press, Amsterdam (2010)
7. Dümbgen, L., Nordhausen, K., Schuhmacher, H.: New algorithms for M-estimation of multivariate scatter and location. J. Multivariate Anal. **144**, 200–217 (2016)
8. Dümbgen, L., Pauly, M., Schweizer, T.: M-functionals of multivariate scatter. Statistics Surv. **9**, 32–105 (2015)
9. Fischer, D., Berro, A., Nordhausen, K., Ruiz-Gazen, A.: REPPlab: an R package for detecting clusters and outliers using exploratory projection pursuit. In: Communications in Statistics—Simulation and Computation, pp. 1–23 (2019)
10. Huber, P.J.: Robust estimation of a location parameter. Ann Math Statistics 35:73–101 (1964)
11. Huber, P.J.: Projection pursuit. Ann. Statistics **13**, 435–475 (1985)
12. Jin, Z., Risk, B.B., Matteson, D.S.: Optimization and testing in linear non-Gaussian component analysis. Statistical Anal. Data Mining ASA Data Sci. J. **12**, 141–156 (2019)
13. Jolliffe, I.T.: Principal Component Analysis, 2nd edn. Springer, New York (2002)
14. Jones, M.C., Sibson, R.: What is projection pursuit? J. R. Statistical Soc. Ser. A **150**, 1–37 (1987)
15. Kawanabe, M., Sugiyama, M., Blanchard, G., Müller, K.-R.: A new algorithm of non-Gaussian component analysis with radial kernel functions. Ann. Inst. Statistical Math. **59**, 57–75 (2007)
16. Kent, J.T., Tyler, D.E.: Redescending M-estimates of multivariate location and scatter. Ann. Statistics **19**, 2102–2119 (1991)
17. Miettinen, J., Nordhausen, K., Taskinen, S.: Blind source separation based on joint diagonalization in R: the packages JADE and BSSasymp. J. Statistical Softw. **76**, 1–31 (2017)
18. Miettinen, J., Nordhausen, K., Taskinen, S., Tyler, D.E.: On the computation of symmetrized M-estimators of scatter. In: Agostinelli, C., Basu, A., Filzmoser, P., Mukherjee, D. (eds.) Recent Advances in Robust Statistics: Theory and Applications, pp. 151–167. Springer, New Delhi (2016)
19. Miettinen, J., Taskinen, S., Nordhausen, K., Oja, H.: Fourth moments and independent component analysis. Statistical Sci. **30**, 372–390 (2015)
20. Nordhausen, K., Oja, H.: Independent subspace analysis using three scatter matrices. Austr. J. Statistics **40**, 93–101 (2016)

21. Nordhausen, K., Oja, H.: Independent component analysis: a statistical perspective. Wiley Interdiscip. Rev. Comput. Statistics **10**, e1440 (2018)
22. Nordhausen, K., Oja, H., Ollila, E.: Multivariate models and the first four moments. In: Hunter, D.R., Richards, D.S.R., Rosenberger, J.L. (eds.) Nonparametric Statistics and Mixture Models: A Festschrift in Honour of Thomas P. Hettmansperger, pp. 267–287. World Scientific, Singapore (2011)
23. Nordhausen, K., Oja, H., Ollila, E.: Robust independent component analysis based on two scatter matrices. Austr. J. Statistics **37**, 91–100 (2016)
24. Nordhausen, K., Oja, H., Tyler, D.E.: Tools for exploring multivariate data: The package ICS. J. Statistical Softw. **28**, 1–31 (2008)
25. Nordhausen, K., Oja, H., Tyler, D.E., Virta, J.: Asymptotic and bootstrap tests for the dimension of the non-Gaussian subspace. IEEE Signal Process. Lett. **24**, 887–891 (2017)
26. Nordhausen, K., Oja, H., Tyler, D.E., Virta, J.: ICtest: estimating and testing the number of interesting components in linear dimension reduction. R package version 0.3-2 (2019)
27. Nordhausen, K., Oja, H., Tyler, D.E.: Asymptotic and bootstrap tests for subspace dimension. arXiv:1611.04908 (2017)
28. Nordhausen, K., Tyler, D.E.: A cautionary note on robust covariance plug-in methods. Biometrika **102**, 573–588 (2015)
29. Nordhausen, K., Virta, J.: An overview of properties and extensions of FOBI. Knowl.-Based Syst. **173**, 113–116 (2019)
30. Oja, H., Sirkiä, S., Eriksson, J.: Scatter matrices and independent component analysis. Austr. J. Statistics **35**, 175–189 (2006)
31. R Core Team.: R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2017)
32. Risk, Benjamin B., Matteson, David S., Ruppert, David: Linear non-Gaussian component analysis via maximum likelihood. J. Am. Statistical Assoc. **114**, 332–343 (2019)
33. Sasaki, H., Niu, G., Sugiyama, M.: Non-Gaussian component analysis with log-density gradient estimation. In: Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, pp. 1177–1185 (2016)
34. Sirkiä, S., Miettinen, J., Nordhausen, K., Oja, H., Taskinen, S.: SpatialNP: multivariate nonparametric methods based on spatial signs and ranks. R package version 1.1-4 (2019)
35. Sirkiä, S., Taskinen, S., Oja, H.: Symmetrised M-estimators of multivariate scatter. J. Multivariate Anal. **98**, 1611–1629 (2007)
36. Theis, F.J.: Towards a general independent subspace analysis. In: Schölkopf, B., Platt, J.C., Hoffman, T. (eds.) Advances in Neural Information Processing Systems 19, pp. 1361–1368. MIT Press, Cambridge, MA (2007)
37. Theis, F.J., Kawanabe, M., Müller, K.R.: Uniqueness of non-Gaussianity-based dimension reduction. IEEE Trans. Signal Process. **59**(9), 4478–4482 (2011)
38. Tyler, D., Critchley, F., Dümbgen, L., Oja, H.: Invariant coordinate selection. J. R. Statistical Soc. Ser. B **71**, 549–592 (2009)
39. Urbanek, S.: png: read and write PNG images. R package version 0.1-7 (2013)
40. Ushey, K.: RcppRoll: efficient rolling/windowed operations. R package version 0.3-0 (2018)
41. Virta, J., Nordhausen, K., Oja, H.: Projection pursuit for non-Gaussian independent components. arXiv preprint arXiv:1612.05445 (2016)
42. Wolodzko, T.: extraDistr: additional univariate and multivariate distributions. R package version 1.8-11 (2019)

A Comparison of Robust Model Choice Criteria Within a Metalearning Study

Petra Vidnerová, Jan Kalina, and Yeşim Güney

Abstract The methodology of automatic method selection (metalearning) allows to recommend the most suitable method (e.g. algorithm or statistical estimator) from several alternatives for a given dataset, based on information learned over a training database of datasets. Practitioners have become accustomed to using metalearning in the context of regression modeling, which is useful in a variety of applications in different fields. Still, none of previous metalearning studies on regression targeted at regression complexity issues and the majority of available metalearning studies for regression considered the standard mean square error as the prediction error measure. In this paper, a metalearning study focused on comparing different method selection criteria for the regression task is presented. A prediction rule, recommending the best regression estimator (possibly robust), is constructed over 31 training datasets. These are publicly available datasets, in which the linear model was carefully examined to be suitable. The results with the highest classification accuracy are obtained if the choice of the best estimator is based on robust versions of Akaike information criterion, particularly the version derived from MM-estimators. The work also advocates an implicitly weighted robust prediction mean square error.

Keywords Automatic method selection · Linear regression · Robust estimation
Robust Akaike criterion

P. Vidnerová · J. Kalina (✉)
The Czech Academy of Sciences, Institute of Computer Science,
Pod Vodárenskou věží 2, 182 07 Praha 8, Czech Republic
e-mail: kalina@cs.cas.cz

P. Vidnerová
e-mail: petra@cs.cas.cz

Y. Güney
Faculty of Science, Department of Statistics,
Ankara University, 06100 Tandogan, Ankara, Turkey
e-mail: ydone@ankara.edu.tr

M. Maciak et al. (eds.), *Analytical Methods in Statistics*, Springer Proceedings
in Mathematics & Statistics 329, https://doi.org/10.1007/978-3-030-48814-7_7

1 Introduction

In the standard linear regression model, the ordinary least squares estimator is notoriously known to be vulnerable to the presence of outliyng measurements (outliers) in the data. Because real data are typically (or perhaps always) contaminated by noise and/or outliers, a number of robust statistical estimators has been proposed and investigated as reliable resistant alternatives to the least squares [27]. Clearly, no robust method is able to be the most suitable for all possible datasets and the question of selecting the most suitable robust regression estimator for a particular datasets remains open. We believe that it is insolvable by means of analytical methods in statistics to acquire theoretical recommendations for choosing a particular robust estimator for a particular dataset. Still, MM-estimators are currently considered to be the most promising robust regression estimator, mainly because of their ability to combine high efficiency with high robustness [14]. The situation is however not unambiguous. Wilcox [43] namely admitted that no study of performance of MM-estimators under heteroscedasticity has been available. In addition, both efficiency and robustness of MM-estimators, which are theoretically obtained under stringent assumptions, become deteriorated even if the number of variables in the model is moderate (larger than small) [35].

Computer scientists invested a lot of effort into general approaches for developing tools for finding the best model (method) from a certain (given) class for a particular dataset. This so-called Algorithm Selection Problem (ASP) is one of crucial problems for solving intelligent systems for automatic data analysis in the rapidly developing field of Automatic Machine Learning (AutoML), completely removing the necessity to manually select a suitable method for a given dataset [9, 19]. Particular approaches for ASP have already been investigated together with their algorithms, properties, and effectiveness [19, 38].

Automatic method selection (also known as metalearning, automatic model choice, model selection, optimal algorithm selection, learning to learn etc.) represents one of the most important approaches to ASP. It can be described as a computational approach allowing to recommend the most suitable method or algorithm for a given dataset, based on information learned over a training database of real datasets. The information in the training database plays the role of prior knowledge, which can be exploited for a new independent dataset. Metalearning has become popular in recent works in classification, optimization, and also (but less frequently) regression tasks, including the analysis of big biomedical data [25]; in these tasks, metalearning contributes to making the analysis of real data more accessible to laymen without statistical expertise [25]. For these reasons, metalearning has established its position among practitioners, especially in specific situations when e.g. the user does not want to directly use methods too slow to be computed exactly [29].

Only within the last decade, metalearning was able to establish attractive applications [4]. The review [38] recalled 190 recent references on metalearning applications. Particularly, metalearning was successfully used to recommend the best optimization procedure [28] or the best variable selection method prior to a subsequent classification analysis [42]. Metalearning was used in [29] to answer the question whether optimization of parameters of a given classifier will improve its classification accuracy. Metalearning was also performed to compare various optimization tools (mainly hybrid algorithms) for business, economics and finance [39], or to recommend parameters for support vector regression based on regression complexity measures [24]. Metalearning has also been applied in the task of recommending the best robust estimator in the linear regression model in [18], where metalearning in its habitual form was revealed to be vulnerable to data contamination by noise or outlying values (outliers), and a robustified approach was proposed; still, the automatic character of metalearning does not allow a detailed (manual) analysis of outliers or other diagnostics.

Advantages of metalearning were summarized in the review [4]; still, we hold the opinion that a conscientiously critical evaluation of metalearning remains missing. In regression, standard mean square error was used in the seminal metalearning studies [6, 24], while alternative robust prediction errors (cf. [2]) have not been compared. Using the (non-robust) mean square error is however not suitable, if the data are contaminated by outliers [14]. Moreover, it was claimed in ([24], p. 238) that none of previous metalearning studies on regression explicitly addressed regression complexity, which remains (together with an application of robust information criteria [37]) a topic for future research. A mistake on p. 43 of [3], where a large number of features in metalearning is claimed to lead to overfitting, indicates that even experts on metalearning do not understand its statistical aspects.

Automatic method selection by means of metalearning is performed in this paper in the context of recommending a suitable robust regression estimator for a given dataset. We recall robust regression estimates and robust prediction errors or robust information criteria in Sect. 2 using these (standard or robust) criteria for the method selection task. The particular metalearning study is described in Sect. 3. Its results are presented in Sect. 4 and the discussion follows in Sect. 5.

2 Robust Regression

The metalearning study of this paper considers 31 datasets. For one (any) particular dataset with n observations, we have values of a continuous response denoted as $Y_1, \dots, Y_n$. We consider the standard linear model

$$Y_i = \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + e_i \quad i = 1, \dots, n. \tag{1}$$

with parameters $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T \in \mathbb{R}^p$, where we assume the vector of regressors (covariates, independent variables) for the i-th observation, denoted as $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})^T$, to always have $X_{i1} = 1$ for $i = 1, \dots, n$. Thus, the linear model always contains an intercept. We assume homoscedastic random errors $e_1, \dots, e_n$ with the common variance $\mathsf{var}\, e_i = \sigma^2$ for each $i = 1, \dots, n$, where $\sigma^2 > 0$ is an unknown nuisance parameter. We use the notation $\mathbf{X}$ for the matrix with elements

X_{ij}, where $i = 1, \dots, n$ and $j = 1, \dots, p$. Various available robust regression estimators of $\boldsymbol{\beta}$ will be described in Sect. 2.1. Evaluating the prediction error of a given estimator by means of robust Akaike information criteria is described in Sect. 2.2 and by means of robust mean square errors in Sect. 2.3.

2.1 Robust Estimation in the Linear Regression Model

We recall several important robust estimators of $\boldsymbol{\beta}$ in (1) in this section. While M-estimators and S-estimators are not used in the metalearning study, we will exploit robust information criteria (defined below in Sect. 2.2) based on their estimation principles. We will use the notation $u_1(\mathbf{b}), \dots, u_n(\mathbf{b})$ for residuals corresponding to a fixed vector $\mathbf{b} = (b_1, \dots, b_p)^T \in \mathbb{R}^p$. In other words,

$$u_i(\mathbf{b}) = Y_i - b_1 X_{i1} - \cdots - b_p X_{ip} = Y_i - \mathbf{X}_i^T \mathbf{b}, \quad i = 1, \dots, n, \tag{2}$$

M-estimators of $\boldsymbol{\beta}$ require to compute a scale statistic $\hat{\sigma}_M$, i.e. estimate of σ, which is regression invariant and scale equivariant, and to choose an absolutely continuous function $\rho_M : \mathbb{R} \to \mathbb{R}$, commonly assumed to be convex with a derivative $\psi_M(x) = d\rho(x)/dx$. Formally, M-estimators are defined by

$$\arg\min_{\mathbf{b} \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_M \left(\frac{u_i(\mathbf{b})}{\hat{\sigma}_M} \right), \tag{3}$$

where

$$\hat{\sigma}_M = C \cdot \mathrm{med}|u_i - \mathrm{med}(u_i)| \tag{4}$$

with med denoting the median. It is common to take $C = 1.4826$ under the assumption of normality (see p. 313 of [32]). In particular, when $\rho_M(t) = t^2/2$, the solution is equal to the least squares estimate.

It is not common to compute M-estimators by means of (3). If ψ_M is assumed to be continuous, M-estimators can be alternatively defined as one of solutions of

$$\arg\min_{\mathbf{b} \in \mathbb{R}^p} \sum_{i=1}^{n} \mathbf{X}_i \psi \left(\frac{u_i(\mathbf{b})}{S_n} \right) = \mathbf{0} \in \mathbb{R}^p, \tag{5}$$

i.e. of the set of p equations in the form

$$\arg\min_{\mathbf{b} \in \mathbb{R}^p} \sum_{i=1}^{n} X_{ij} \psi \left(\frac{u_i(\mathbf{b})}{S_n} \right) = 0, \quad j = 1, \dots, p. \tag{6}$$

At least one solution is a $\sqrt{n}$-consistent estimate of $\boldsymbol{\beta}$; see Sect. 5.5 of [15] for the discussion of asymptotic properties of M-estimators in (1). M-estimators have also been investigated under the assumption that ψ is a nondecreasing step function; see Sect. 4.3 of [14]. At any case, because M-estimators in (1) do not have a high breakdown point, i.e. are not highly resistant against severe outliers in the data, we do not use them in our metalearning study.

MM-estimators of $\boldsymbol{\beta}$ in (1) were proposed in [44] as tools allowing to tune robustness and efficiency simultaneously, which makes them so popular in current statistical applications [14, 27, 44]. The user must select another smooth function ρ_0 for the initial estimate and a smooth function ρ_{MM}; the function ψ_{MM} is defined as the derivative of ρ_{MM}, i.e. $\psi_{MM}(x) = d\rho_{MM}(x)/dx$. MM-estimators are computed within a three-stage procedure:

(a) Initial estimator $\hat{\mathbf{T}} = (\hat{T}_1, \dots, \hat{T}_n)^T \in \mathbb{R}^p$ with a high breakdown point is computed;

(b) Using residuals
$$u_i(\hat{\mathbf{T}}) = Y_i - \mathbf{X}_i^T \hat{\mathbf{T}}, \quad i = 1, \dots, n, \tag{7}$$
an M-scale (i.e. M-estimator of σ) $\hat{\sigma}_{MM} = \hat{\sigma}_{MM}(u_i(\hat{\mathbf{T}}))$ with a high breakdown point is computed;

(c) The MM-estimator is obtained as one of solutions of the set of equations (with variable $\mathbf{b} \in \mathbb{R}^p$)
$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \psi_{MM}\left(\frac{u_i(\mathbf{b})}{\hat{\sigma}_{MM}}\right) = \mathbf{0} \in \mathbb{R}^p. \tag{8}$$

We can see that the final expression (8) has the form of (5), but it is computed with a specific scale estimate $\hat{\sigma}_{MM}$ obtained in the first two stages. The consistency of MM-estimators under specific technical assumptions was derived in [44].

The LTS estimator [33] is defined as
$$\arg\min_{\mathbf{b}\in\mathbb{R}^p} \sum_{i=1}^{h} u_{(i)}^2(\mathbf{b}), \tag{9}$$
where
$$u_{(1)}^2(\mathbf{b}) \le u_{(2)}^2(\mathbf{b}) \le \cdots u_{(n)}^2(\mathbf{b}) \tag{10}$$
are values arranged in ascending order. We put h to be equal to $\lfloor 3n/4 \rfloor$, where $\lfloor x \rfloor$ denotes the integer part of $x \in \mathbb{R}$. This seems to be the most common choice in the literature (cf. [14]). An approximate algorithm for the LTS was proposed in [33].

The least weighted squares (LWS) estimator (see e.g. [41]) for the model (1) generalizes the LTS based on implicit weighting of individual measurements. It performs down-weighting of individual measurements through the idea to assign small (or zero) weights to potential outliers. The LWS estimator, which has acquired only much smaller attention compared to the popular LTS, may attain a high breakdown point, if a suitable weight function is chosen. The LWS estimator is at the same time robust to heteroscedasticity [41], but it its primary attention is focused on estimating $\boldsymbol{\beta}$ and not on outlier detection.

The definition of the LWS exploits the concept of weight function, which is defined as a function $\xi : [0, 1] \to [0, 1]$; it must be continuous on [0, 1] with $\xi(0) = 1$ and $\xi(1) = 0$. The weight function is assumed to have both one-sided derivatives existing in all points of (0, 1), where the one-sided derivatives are bounded by a common constant; also, the existence of a finite left derivative in 0 and finite right derivative in point 1 is assumed [40, 41].

The LWS estimator with a given ξ is defined as

$$b_{LWS} = \arg\min_{\mathbf{b}\in\mathbb{R}^p} \sum_{i=1}^{n} \xi\left(\frac{i-1/2}{n}\right) u_{(i)}^2(\mathbf{b}). \tag{11}$$

We may understand the quantities

$$w_i = \xi\left(\frac{i-1/2}{n}\right), \quad i = 1, \ldots, n, \tag{12}$$

as weights; we may express $w_1, \ldots, w_n$ as

$$\xi\left(\frac{1}{2n}\right), \ldots, \xi\left(\frac{2i-1}{2n}\right), \ldots, \xi\left(\frac{2n-1}{2n}\right). \tag{13}$$

Alternatively, we may start with choosing a fixed non-increasing sequence of non-negative weights $w_1, \ldots, w_n$ and formulate an equivalent definition of the LWS estimator of $\boldsymbol{\beta}$ in the form

$$\mathbf{b}_{LWS} = \arg\min_{\mathbf{b}\in\mathbb{R}^p} \sum_{i=1}^{n} w_i u_{(i)}^2(\mathbf{b}). \tag{14}$$

In this way, the observation with the smallest absolute residual obtains the largest weight w_1 etc. and the most outlying observation with the largest absolute residual obtains the smallest weight w_n. Note that we do not need to require the weights to be standardized to the condition $\sum_{i=1}^{n} w_i = 1$.

If we denote the ranks of (2) by $R_1(\mathbf{b}), \ldots, R_n(\mathbf{b})$, i.e. with $R_i(\mathbf{b})$ denoting the rank of $u_i^2(\mathbf{b})$ among $u_1^2(\mathbf{b}), \ldots, u_n^2(\mathbf{b})$, we may express the LWS estimator as

$$\mathbf{b}_{LWS} = \arg\min_{\mathbf{b}\in\mathbb{R}^p} \sum_{k=1}^{n} \xi\left(\frac{R_i(\mathbf{b})-1/2}{n}\right) u_{(i)}^2(\mathbf{b}). \tag{15}$$

It is meaningful to consider only weight functions which are non-increasing; only then, less reliable measurements obtain smaller weights. We use here the trimmed linear weights generated for a fixed $\tau \in [1/2, 1)$ by

$$\xi(t) = \left(1 - \frac{t}{\tau}\right) \cdot \mathbb{1}[t < \tau], \quad t \in [0, 1], \tag{16}$$

where $\mathbb{1}[.]$ denotes an indicator function. Here, τ is in relationship with the trimming, i.e. there are $\lfloor \tau n \rfloor$ measurements retained and the remaining measurements are ignored; this is analogous to α for the LTS. We use $\tau = 3/4$ in the computations.

For the computation of the LWS estimator, an analogy of the FAST-LTS algorithm of [33] exploiting the form (15), which is appealing from a computational perspective, can be formulated in a straightforward way. This approximate algorithm was characterized as reliable based on empirical evidence [18].

2.2 Robust Akaike Information Criterion

We recall Akaike information criterion (AIC) and its several robust versions here. AIC represents a general information–theoretical measure of quality of a regression fit, originally designed for model selection, which has also been recommended as a measure of prediction error, i.e. as a model (method) selection criterion. It was proposed in [1] as

$$\mathsf{AIC} = -2 \log L(y_i; \hat{\boldsymbol{\theta}}) + 2p \tag{17}$$

for a very general situation. Here, we consider the linear model (1). The value $\log L(y_i; \hat{\boldsymbol{\theta}})$ is the maximized log-likelihood function computed for the given model and $\hat{\boldsymbol{\theta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{\sigma})^T$ is the considered estimate of $\boldsymbol{\theta} = (\beta_1, \ldots, \beta_p, \sigma)^T$. In linear regression, the log-likelihood is typically (and also in our case) evaluated under the assumption of normally distributed errors. To select the most suitable model out of several possible choices, practitioners typically decide for the model with the smallest value of AIC.

Robust versions of AIC, denoted as M-AIC, S-AIC and MM-AIC, are based on M-estimators, S-estimators and MM-estimators, respectively. They were proposed in [37] and further investigated also in [13] or [14]. Robust versions of AIC, extending the pioneering ideas of [31], were successful also in models with autoregressive errors in the paper [11]. The definitions of the robust AIC criteria are formulated for a given model and given estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$. Residuals corresponding to the M-estimator $\boldsymbol{\beta}_M$, which has already been computed and is thus known at this moment, are denoted as $u^M = (u_1^M, \ldots, u_n^M)^T$. Residuals of the MM-estimator $\boldsymbol{\beta}_{MM}$ are denoted as $u^{MM} = (u_1^{MM}, \ldots, u_n^{MM})^T$. Scale estimates obtained by the M-estimation, S-estimation, and MM-estimation are denoted as $\hat{\sigma}_M$, $\hat{\sigma}_S$, and $\hat{\sigma}_{MM}$, respectively. The formal definitions proposed by [37] are

$$\text{M-AIC} = 2\sum_{i=1}^{n} \rho_M \left(\frac{u_i^M}{\hat{\sigma}_M}\right) + 2\text{tr}(J_n^{-1}K_n), \tag{18}$$

$$\text{S-AIC} = 2n\log(\hat{\sigma}_S) + 2\text{tr}(J_n^{-1}K_n), \tag{19}$$

and

$$\text{MM-AIC} = 2\sum_{i=1}^{n} \rho_{MM} \left(\frac{u_i^{MM}}{\hat{\sigma}_{MM}}\right) + 2\text{tr}(J_n^{-1}K_n); \tag{20}$$

here, tr denotes the matrix trace and the (rather technical) definitions of the empirical information matrices J_n and K_n can be found in [37]. It is worth noting that J_n and K_n are defined in the same manner for all robust versions of AIC. The loss functions ρ_M and ρ_{MM} corresponding to M- or MM-estimation were defined in Sect. 2.1.

In our computations, we use Huber's function for M-AIC and Tukey's biweight function for S-AIC. MM-AIC is used here with an initial S-estimator with Tukey's biweight function and with Huber's function playing the role of ρ_{MM}. Let us also recall that Huber's ρ function is defined as

$$\rho_H(t) = \begin{cases} \frac{1}{2}t^2, & \text{if } |t| \le c_1, \\ \\ c_1|t| - \frac{c_1^2}{2}, & \text{if } |t| > c_1, \end{cases} \tag{21}$$

where $c_1 > 0$ is a tuning constant; to obtain the 95 % asymptotic efficiency on the standard normal distribution one can take $c = 1.345$ [13]. Tukey's biweight function is defined as

$$\rho_T(t) = \begin{cases} \frac{t^2}{2} - \frac{t^4}{2c_2^2} + \frac{t^6}{6c_2^4}, & \text{if } |t| \le c_2, \\ \\ \frac{c_2^2}{6}, & \text{if } |t| > c_2, \end{cases} \tag{22}$$

where $c_2 > 0$ is a tuning constant which controls the breakdown point of the estimator. We use the usual choice $c_2 = 1.547$, which allows to reach the maximal breakdown point. The loss functions ρ_H and ρ_T are implemented in package robustbase of R software [22].

2.3 Robust Mean Square Error

The most standard choice of the prediction measure in (1) is the mean square (prediction) error (MSE). After computing a particular estimate for given data, let $r_1, \ldots, r_n$ be its prediction errors and let us consider them arranged (in squares) in ascending order as

$$r_{(1)}^2 \le \cdots \le r_{(n)}^2. \tag{23}$$

We also consider two robust versions of MSE.

The trimmed mean square error (TMSE) is defined for $\alpha \in [1/2, 1)$ as

$$\mathrm{TMSE}(\alpha) = \frac{1}{h} \sum_{i=1}^{h} r_{(i)}^2, \tag{24}$$

where $h = \lfloor \alpha n \rfloor$. We now define the weighted mean square error (WMSE) as

$$\mathrm{WMSE} = \sum_{i=1}^{n} \xi \left(\frac{R_i - 1/2}{n} \right) r_{(i)}^2, \tag{25}$$

where R_i denotes the rank of $r_{(i)}^2$ among $r_1^2, \ldots, r_n^2$. Equivalently, using the same idea as in the definition of the LWS estimator, we may perceive values (12) as magnitudes of (non-increasing) weights and denote WMSE as

$$\mathrm{WMSE} = \sum_{i=1}^{n} w_i r_{(i)}^2. \tag{26}$$

In the computations, we use the weight function ξ (16) corresponding to the trimmed linear weights. This allows to obtain a robust value of MSE trimming away about one quarter of the observations, which should hopefully be sufficient for real datasets to ignore true (but always unknown) outliers.

3 Description of the Metalearning Study

As characterized in Sect. 1, metalearning is a computational approach allowing to exploit information from previously observed datasets and to extend it to new datasets [3]. The aim of the metalearning study presented here is to recommend the most suitable regression estimator for a new dataset, exploiting the information from 31 datasets described in Sect. 3.1.

We will now describe the metalearning study with all its steps and chosen parameters. First, four linear regression estimators are fitted for each of the given datasets and the best estimator is found using various characteristics of Sects. 2.2 and 2.3. This constitutes the primary learning part of the study, which is described in Sect. 3.2. Then, a set of 9 selected features (Sect. 3.3) is retained for each of the individual datasets together with the result of the primary learning, which typically has the form of the index of the best method for each of the training datasets. The secondary learning part (Sect. 3.4) performed over these data (i.e. features and results of the primary learning) learns a classification rule by one of 7 different classifiers, allowing to predict the best regression method for a new dataset not present in the training database.

3.1 Data Acquisition and Pre-processing

We work with 31 publicly available datasets, presented in Table 1 together with some their basic characteristics. Only datasets with trustworthy documentation are considered, which are explicitly claimed to have been subjected to standard pre-processing (cleaning, suitable transforms of variables). Some of the datasets are well known as benchmarking datasets for the linear regression task. It is important that linear regression modeling has been presented in the literature for each of the 31 datasets (i.e. with the same response as is considered in our study) and the linear model has been found appropriate there. This study extends a preliminary study [18], which did not however consider any information criteria.

All observations (measurements) with any missing value are deleted here; this was however performed only for one dataset, while the others do not contain any missing values. Further, we performed an automatically detection of categorical variables. Categorical variables such with 3 or more categories were omitted for the purpose of computational complexity. Binary variables were replaced by a single dummy variables, interpreted as indicators of the first group. We use the intercept in (1) for each of the datasets. We do not perform any special treatment of multicollinearity, which does not seem to represent a major issue here due to small values of p. Actually, the computation of none of the estimators reported any warning, which would be however common in multicollinear data.

The response $Y_1, \dots, Y_n$ was transformed to contain values between 0 and 1 by the commonly used transform

$$Y_i \longmapsto \frac{Y_i - \min_i Y_i}{\max_i Y_i - \min_i Y_i}, \quad i = 1, \dots, n. \tag{27}$$

In the same way, all continuous regressors were transformed. Such transforms do not influence the prediction ability of the regression estimators, because all regression methods of this study are scale- and regression-equivariant, but do influence the features computed from each dataset. This is beneficial, as the original measurements differ greatly among datasets, as they also come from different fields and applications.

3.2 Primary Learning

In the first step of the metalearning, we consider the standard linear regression model (1) for each of the datasets. There are the following 4 regression estimators, described already in Sect. 2.1, fitted for each of the given datasets. Here we also specify the computational tools in R software used in our study.

- Least squares (LS), implemented in function lm.
- MM-estimator, implemented in function lmrob of the package robustbase. We use the default version, i.e. with breakdown point 0.5 and efficiency 0.95.

Table 1 The 31 datasets together with their basic characteristics. If missing values are present in a dataset, we omitted all observations for which the value of any of the variables is missing

Index	Dataset	Response variable	n	p	Source	Missing values
1	Aircraft	Cost	23	5	[26]	None
2	Ammonia	Unprocessed percentage	21	4	[36]	None
3	Auto MPG	Miles per gallon	392	5	[8]	Omitted
4	Boston housing	Crime rate	506	6	[8]	None
5	Building	Electricity consumption	4208	7	[34]	None
6	California housing	Median house price	20,640	9	[5]	None
7	Cirrhosis	Death rate	46	5	[36]	None
8	Coleman	Test score	20	6	[26]	None
9	Concrete compression strength	Concrete compression strength	1030	7	[8]	None
10	Delivery	Delivery time	25	3	[26]	None
11	Education	Education expenditures	50	4	[26]	None
12	Electricity	Output	16	4	[36]	None
13	Employment	# of employed people	16	7	[36]	None
14	Engel	Food expenditures	235	2	[21]	None
15	Furniture	Log relative wage	11	2	[20]	None
16	Houseprices	Selling price	28	6	[36]	None
17	Imports	Level of imports	18	4	[36]	None
18	Investment	Investment	22	2	[17]	None
19	Istanbul stock exchange	Istanbul index	536	8	[8]	None
20	Kootenay	Newgate	13	2	[26]	None
21	Livestock	Expenses	19	5	[36]	None
22	Machine	PRP	209	7	[8]	None
23	Murders	# of murders	20	4	[36]	None
24	NOx emissions	LNOx	8088	4	[26]	None
25	Octane	Octane rating	82	5	[36]	None
26	Pasture	Pasture rental price	67	4	[36]	None
27	Pension	Reserves	18	2	[26]	None
28	Petrol	Consumption	48	5	[36]	None
29	Stars CYG	Log temperature	47	2	[26]	None
30	Travel and tourism	TSI	141	13	[7]	None
31	Wood	Wood gravity	20	6	[26]	None

- Least trimmed squares (LTS), implemented in function ltsReg of the package robustbase, using $h = \lfloor 3n/4 \rfloor$.
- Least weighted squares (LWS) with trimmed linear weights (16), using our own implementation.

The best estimator is found using each of the information criteria of Sect. 2.2 and each of the prediction error measures of Sect. 2.3. The best estimators, i.e. the estimator (out of the 4 ones), which has the smallest value of the particular error measure or information criterion, is found always within a (standard) leave-one-out cross validation. The results are presented in Sect. 4.

3.3 Features

In the secondary learning described in Sect. 3.4 below, we compute the following set of 9 features for each dataset.

(A) The number of observations n;
(B) The value of p as denoted at the beginning of Sect. 2, i.e. the set of regressors contains a vector of ones and $p-1$ additional regressors.
(C) The ratio $n/(p-1)$;
(D) p-value of the (approximate) Shapiro-Wilk test for the least squares, evaluated by means of the function shapiro.test of R software. The test of Shapiro-Wilk is a well known test of normality of the errors $e_1, \dots, e_n$ in (1);
(E) Skewness of residuals of the least squares in (1);
(F) Kurtosis of residuals of the least squares in (1);
(G) Coefficient of determination R^2 for the least squares in (1), which is known as the most commonly used goodness-of-fit characteristic of the linear model;
(H) Estimated percentage of outliers in (1) evaluated by $\frac{1}{n}\sum_{i=1}^{n} I[u_i/\hat{\sigma} > 2.5]$, where $u_1, \dots, u_n$ are residuals obtained by the LTS with $h = 0.5$ and $\hat{\sigma}$ is the estimate of σ obtained by the LTS with $h = 0.5$. This rule suggested by [32] was repeatedly advocated in the literature on outlier detection (cf. [14]);
(I) p-value of the Breusch-Pagan test applied to the least squares residuals in (1), evaluated by means of the function bptest of the package lmtest of R software. This is a standard heteroscedasticity test of the random errors $e_1, \dots, e_n$. While the test is based on assessing variability in an auxiliary model, we use its default form, i.e. exactly all regressors from (1) appear in the auxiliary model in the form

$$\mathsf{var}\, e_i = \alpha_1 X_{i1} + \cdots + \alpha_p X_{ip} + \gamma_i, \quad i = 1, \dots, n, \tag{28}$$

with parameters $(\alpha_1, \dots, \alpha_p)^T$ and random errors $\gamma_1, \dots, \gamma_n$.

3.4 Secondary Learning: The Classification Task

In the primary learning, we learn which method is the best for each particular dataset. To be specific, the LWS estimator turns out to be best for the first dataset, the MM-estimator for the second dataset etc. Only this information from the primary learning, together with the features of Sect. 3.3, are used in the secondary learning. In other words, the full datasets are not used now any more. We can say that we consider a nominal categorical variable with 4 possible values; this has 31 values for the 31 datasets and the value for a given dataset states, which of the 4 estimators was the best for the given dataset in the primary learning (with a given method selection criterion). This variable plays now the role of a response in a classification task to 4 groups, where the features are the explanatory variables now. Such classification task does not suffer from multiple testing (i.e. no hypothesis testing is performed here in the task to recommend the best estimator for a particular dataset).

We use the following available classification methods for the metalearning task. The computation of most of them is performed exploiting available packages of R software, which are presented in Table 2.

- Support vector machine (SVM) classifier with Gaussian kernel,
- k-nearest neighbor classifier with $k = 5$,
- Multilayer perceptron (MLP) [12] with 2 hidden layers, which contain 6 and 3 neurons, with a sigmoid activation function in each hidden layer, and a linear output,
- Radial basis function (RBF) network [12, 23] with $N = 15$ RBF units,
- Linear discriminant analysis (LDA),

Table 2 Results of the secondary learning within the metalearning study performed over 31 datasets. The results are evaluated as classification accuracies in a leave-one-out cross validation

Model choice criterion				RBF network			MWCD–LDA
	SVM	k-NN	MLP		LDA	SCRDA	
R package	e1017	class	ANN2	RSNNS	MASS	rda	–
MSE	0.58	0.48	0.52	0.55	0.55	0.55	0.55
TMSE (0.75)	0.65	0.52	0.58	0.58	0.58	0.58	0.58
TMSE (0.85)	0.65	0.52	0.58	0.58	0.58	0.58	0.58
WMSE	0.65	0.48	0.58	0.58	0.58	0.58	0.61
AIC	0.58	0.45	0.55	0.55	0.55	0.55	0.55
M-AIC	0.58	0.48	0.58	0.58	0.58	0.58	0.58
S-AIC	0.58	0.48	0.48	0.52	0.55	0.58	0.58
MM-AIC	0.65	0.52	0.58	0.61	0.58	0.61	0.61

- Shrunken centroid regularized discriminant analysis (SCRDA) of [10],
- MWCD-LDA, which is obtained as a robust version of LDA based on the minimum weighted covariance determinant estimator (MWCD) of [30], with the weight function (16). This is the only classifier, for which we use our own implementation.

Hyperparameters for regularized classifiers (SVM, SCRDA, MWCD-LDA) were optimized in a standard 10-fold cross validation.

4 Results

We used the R software for the whole metalearning study. The results of the metalearning study, which solves a classification task to 4 groups, are presented in Table 2 evaluated in the form the classification accuracy, i.e. ratio of correctly classified cases (datasets). Among the 7 different classifiers, the SVM classifier turns out to yield the best performance. The best result is obtained for the approach exploiting MM-AIC, if the SVM classifier is used; in this situation, the best regression method is found correctly in 64 % of datasets.

We can see that approaches based on robust versions of MSE outperform those using the plain MSE; among different versions of robust MSE, WMSE seems superior to others. Robust versions of AIC yield superior results compared to those of plain AIC, while robust AIC based on MM-estimation turns out to be actually the very best tool in this study.

Comparing the relevance of individual features, those denoted as (D), (H) and (I) seem to be the most useful within the secondary learning. This (subjective) conclusion was made by comparing the performance of various classifiers over all possible subsets of features (of various sizes). The classification accuracies were, jointly for all 7 classifiers, the largest for situations when the features (D), (H) and (I) were all considered. This correspond to our intuition, because these features are connected to assumptions of the least squares (normality of errors, i.i.d. errors without contamination, and homoscedasticity) and their severe violation requires to use a (highly) robust approach instead.

5 Discussion

While various robust estimators are available for the linear regression model, there are no theoretical justifications for using a particular robust estimator for a given dataset. Actually, we believe that it remains impossible to analytically derive rules recommending to use a particular robust estimator for particular data. Therefore, this work is devoted to an automatic method selection study by means of the metalearning methodology, recommending the best (robust or non-robust) regression estimator for the linear regression model. The classification rule of the automatic method selection is learned over a training database of 31 carefully selected publicly available datasets. We can consider the performance of the constructed rule to be successful in recommending the most suitable nonlinear regression estimator, even in spite of the difficulty of the secondary learning into 4 groups.

While choosing a proper measure of prediction error in linear regression is known as a difficult task as such [2], the presented results bring novel arguments in favor of robust versions of AIC, especially MM-AIC. We can perceive AIC as well as its robust versions as conceptually simple, yet powerful tools reliable also under multicollinearity (in contrary to model choice based on hypothesis testing).

Reliable (resistant) metalearning should, based on our experience, desire to use the following.

- Homogeneous datasets.
- A small number of methods (estimators, algorithms). At the same time, these should be very distinct (i.e. well distinguishable, not just slightly modified version of the same approach).
- A suitable robust prediction error, particularly a robust version of AIC.
- A suitable classification method, preferably an SVM classifier with a Gaussian kernel.

As future research, we intend to work on the robustification and automation of the whole metalearning process, which would allow to perform metalearning in realistic scenarios over very large databases of datasets; this however requires to select the particular model (1) carefully for each dataset to avoid misspecification. Other open topics include the performance of robust AIC versions in nonlinear (robust) regression [16] or performance of robust estimators for high-dimensional data; especially MM-estimators are known to lose their robustness and efficiency for data with $n < p$ [35].

Acknowledgements The research was supported by the grants GA18-23827S (P. Vidnerová) and GA19-05704S (J. Kalina) of the Czech Science Foundation. The authors are grateful to Aleš Neoral for technical help and to the referee and the editor for their time and helpful suggestions.

References

1. Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: Petrov, B., Csaki, F. (eds.) Second International Symposium on Information Theory, pp. 267–281. Budapest, Academiai Kaido (1973)
2. Borra, S., Di Ciaccio, A.: Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods. Comput. Statist. Data Anal. **54**, 2976–2989 (2010)
3. Brazdil, P., Giraud-Carrier, C., Soares, C., Vilalta, E.: Metalearning: Applications to Data Mining. Springer, Berlin (2009)
4. Brazdil, P., Giraud-Carrier, C.: Metalearning and algorithm selection: progress, state of the art and introduction to the 2018 special issue. Mach. Learn. **107**, 1–14 (2018)
5. California housing dataset. https://github.com/ageron/handson-ml/tree/master/datasets/housing (2019)
6. Collins, A., Beel, J., Tkaczyk, D.: One-at-a-time: A meta-learning recommender-system for recommendation-algorithm selection on micro level. ArXiv:1805.12118 (2020)

7. Crotti, R., Misrahi, T.: The Travel & Tourism Competitiveness Report 2015. Growth Through Shocks. World Economic Forum, Geneva (2015)
8. Dua, D., Graff, C.: UCI Machine Learning Repository. University of California, Irvine. http://archive.ics.uci.edu/ml (2019)
9. Ewald, R.: Automatic Algorithm Selection for Complex Simulation Problems. Vieweg+Teubner Verlag, Wiesbaden (2012)
10. Guo, Y., Hastie, T., Tibshirani, R.: Regularized discriminant analysis and its application in microarrays. Biostatistics **8**, 86–100 (2007)
11. Güney, Y., Tuaç, Y., Özdemir, Ş., Arslan, O.: Conditional maximum Lq-likelihood estimation for regression model with autoregressive error terms. ArXiv:1804.07600 (2020)
12. Haykin, S.O.: Neural Networks and Learning Machines: A Comprehensive Foundation, 2nd edn. Prentice Hall, Upper Saddle River (2009)
13. Huber, P.J., Ronchetti, E.M.: Robust Statistics, 2nd edn. Wiley, New York (2009)
14. Jurečková, J., Picek, J., Schindler, M.: Robust Statistical Methods with R, 2nd edn. CRC Press, Boca Raton (2019)
15. Jurečková, J., Sen, P.K., Picek, J.: Methodology in Robust and Nonparametric Statistics. CRC Press, Boca Raton (2013)
16. Kalina, J.: On robust information extraction from high-dimensional data. Serb. J. Manage. **9**, 131–144 (2014)
17. Kalina, J.: Three contributions to robust regression diagnostics. J. Appl. Math. Stat. Inf. **11**(2), 69–78 (2015)
18. Kalina, J.: On Sensitivity of Metalearning: An Illustrative Study for Robust Regression. In: Proceedings ISNPS 2018. Accepted (in press) (2020)
19. Kersche, P., Hoos, H.H., Neumann, F., Trautmann, H.: Automated algorithm selection: survey and perspectives. Evol. Comput. **27**, 3–45 (2018)
20. Kmenta, J.: Elements of Econometrics. Macmillan, New York (1986)
21. Koenker, R.: Quantile Regression. Cambridge University Press, Cambridge (2005)
22. Koller, M., Mächler, M.: Defintions of ψ-functions available in Robustbase. https://cran.r-project.org/web/packages/robustbase/vignettes/ (2019)
23. Kudová, P.: Learning with Regularization Networks. Dissertation thesis. MFF UK, Prague (2006)
24. Lorena, A.C., Maciel, A.I., de Miranda, P.B.C., Costa, I.G., Prudêncio, R.B.C.: Data complexity meta-features for regression problems. Mach. Learn. **107**, 209–246 (2018)
25. Luo, G.: A review of automatic selection methods for machine learning algorithms and hyper-parameter values. Network Model. Anal. Health Inf. Bioinform. **5**, 5–18 (2016)
26. Maechler, M., Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibián-Barrera, M., Verbeke, T., Koller, M., Conceicao, E.L.T., di Palma, M.A.: Robustbase: Basic Robust Statistics R package version 0.92-7 (2016)
27. Maronna, R.A., Martin, R.D., Yohai, V.J., Salibián-Barrera, M.: Robust Statistics: Theory and Methods (with R), 2nd edn. Wiley, Oxford (2019)
28. Reif, M., Shafait, F., Dengel, A.: Meta-learning for evolutionary parameter optimization of classifiers. Mach. Learn. **87**, 357–380 (2012)
29. Ridd, P., Giraud-Carrier, C.: Using metalearning to predict when parameter optimization is likely to improve classification accuracy. In: Proceedings International Conference on Metalearning and Algorithm Selection MLAS'14, pp. 18–23 (2014)
30. Roelant, E., Van Aelst, S., Willems, G.: The minimum weighted covariance determinant estimator. Metrika **70**, 177–204 (2009)
31. Ronchetti, E.: Robust model selection in regression. Stat. Prob. Lett. **3**, 21–23 (1985)
32. Rousseeuw, P.J., Leroy, A.M.: Robust Regression and Outlier Detection. Wiley, New York (1987)
33. Rousseeuw, P.J., van Driessen, K.: Computing LTS regression for large datasets. Data Mining Knowl. Discovery **12**, 29–45 (2006)
34. Rusiecki, A., Kordos, M., Kamiński, T., Greń, K.: Training neural networks on noisy data. Lect. Notes Comput. Sci. **8467**, 131–142 (2014)

35. Smucler, E., Yohai, V.J.: Robust and sparse estimators for linear regression models. Comput. Stat. Data Anal. **111**, 116–130 (2017)
36. Spaeth, H.: Mathematical Algorithms for Linear Regression. Academic Press, Cambridge (1991)
37. Tharmaratnam, K., Claeskens, G.: A comparison of robust versions of the AIC based on M-S- and MM-estimators. Statistics **47**, 216–235 (2013)
38. Vanschoren, J.: Metalearning. In Hutter, F., Kotthoff, L., Vanschoren, J. (eds.): Automated Machine Learning. Methods, Systems, Challenges, Chap. 2, pp. 35–61. Springer, Cham (2019)
39. Vasant, P.M.: Meta-Heuristics Optimization Algorithms in Engineering, Business, Economics, and Finance. IGI Global, Hershey (2012)
40. Víšek, J.Á.: Robust error-term-scale estimate. IMS Collect. **7**, 254–267 (2010)
41. Víšek, J.Á.: Consistency of the least weighted squares under heteroscedasticity. Kybernetika **47**, 179–206 (2011)
42. Wang, G., Song, Q., Sun, H., Zhang, X., Xu, B., Zhou, Y.: A feature subset selection algorithm automatic recommendation method. J. Artif. Intell. Res. **47**, 1–34 (2013)
43. Wilcox, R.R.: Introduction to Robust Estimation and Hypothesis Testing, 3rd edn. Elsevier, Waltham (2012)
44. Yohai, V.J.: High breakdown-point and high efficiency robust estimates for regression. Ann. Stat. **15**, 642–656 (1987)

On Parameter Estimation for High Dimensional Errors-in-Variables Models

Silvelyn Zwanzig and Rauf Ahmad

Abstract Estimation of parameter vector for a linear model with errors-in-variables is considered when the number of regressors may exceed the sample size. As the classical approaches fail in this high-dimensional setting, new approaches are assessed. In particular, we address the problem from two perspectives. Assuming the usual functional model setting, the first solution concerns a generalization of the classical total least squares estimator. The second option assumes structural model and is based on estimating the unknown covariance matrix of large dimension. In both cases, only the exact solutions are considered so that no asymptotics are required. We assume normality, along with a few other mild assumptions, but do not assume any sparsity or related conditions.

Keywords Errors-in-variables models · Covariance estimation · Shrinkage

1 Introduction and Objectives

Consider the general linear model with error-in-variables (EIV). Let

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \ \text{ with } \ \mathbf{W} = \mathbf{X} + \boldsymbol{\Delta}, \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{n\times p}$ is the design matrix, $\mathbf{Y} \in \mathbb{R}^{n\times 1}$ is the response vector, $\boldsymbol{\beta} \in \mathbb{R}^{p\times 1}$ is the parameter vector, and $\boldsymbol{\varepsilon} \in \mathbb{R}^{n\times 1}$ is the error vector. The second part of Model (1) implies that $\mathbf{X}$ could not be observed for the model due to errors $\boldsymbol{\Delta} \in \mathbb{R}^{n\times p}$, so that $(\mathbf{W}, \mathbf{Y})$ are the actual observed data. For $i = 1, \ldots, n$, we can equivalently write

S. Zwanzig (✉)
Department of Mathematics, Uppsala University, Uppsala, Sweden
e-mail: zwanzig@math.uu.se

R. Ahmad
Department of Statistics, Uppsala University, Uppsala, Sweden
e-mail: rauf.ahmad@statistik.uu.se

M. Maciak et al. (eds.), *Analytical Methods in Statistics*, Springer Proceedings in Mathematics & Statistics 329, https://doi.org/10.1007/978-3-030-48814-7_8

$$Y_i = \mathbf{X}_i^T \boldsymbol{\beta} + \varepsilon_i,$$

as n independent observations of the model, where $\mathbf{W}_i = \mathbf{X}_i + \boldsymbol{\Delta}_i$. Here $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})'$ are the rows of $\mathbf{X}$. We allow $\mathbf{X}$ to be rank deficient, i.e.

$$r(\mathbf{X}) \leq r(\mathbf{W}) = n \ll p, \tag{2}$$

and set the following assumptions on Model (1) to proceed further. We will distinguish between a functional model where the design matrix $\mathbf{X}$ consists of fixed unknown nuisance parameters and the structural model where $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are unobserved (latent) iid random variables. The functional model is a more general and complicated model, so that estimators derived for the functional model also hold under structural model assumptions. For both cases we assume

(A1) $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}, \ \mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma_\varepsilon^2 \mathbf{I}$.
(A2) $\mathrm{E}(\boldsymbol{\Delta}) = \mathbf{O}, \ \mathrm{Cov}(\boldsymbol{\Delta}) = \sigma_w^2(\mathbf{I}_n \otimes \mathbf{I_p})$.
(A3) $\boldsymbol{\Delta} \perp\!\!\!\perp \boldsymbol{\varepsilon}$, where $\perp\!\!\!\perp$ denotes independence.

Note that, under A2, we are essentially assuming that the elements of $\boldsymbol{\Delta}$, denoted δ_{ij}, $i = 1, \ldots, n, \ j, \ldots, p$, are all iid with $\mathrm{Var}(\delta_{ij}) = \sigma_w^2$. A3 implies that δ_{ij} and ε_i are independent, and so are $\mathbf{W}_i$ and Y_i for fixed $\mathbf{X}$. Finally, A1-A3 suffice for most of our objectives. However, as we shall partly be dealing with likelihood estimation, a normality assumption need to be added for δ_{ij}, i.e. $\delta_{ij} \sim N(0, \sigma_w^2)$, so that A2 takes the form

$$\boldsymbol{\Delta} \sim N_{n,p}\left(\mathbf{O}, \sigma_w^2(\mathbf{I}_n \otimes \mathbf{I}_p)\right) \ \text{ and } \ \boldsymbol{\varepsilon} \sim N_p(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}).$$

Additionally, denoting $\mathbf{X} = (\mathbf{X}_1^T, \ldots, \mathbf{X}_n^T)$, with $\mathbf{X}_i, \ i = 1, \ldots, n$ iid vectors, we assume the following for the structural model.

(A2s) $\mathrm{E}(\mathbf{X}) = \mathbf{O}, \mathrm{Cov}(\mathbf{X}) = \mathbf{I}_n \otimes \boldsymbol{\Sigma}_X$.
(A3s) $\mathbf{X} \perp\!\!\!\perp \boldsymbol{\Delta}$, where $\perp\!\!\!\perp$ denotes independence.

Adding again the normality assumption for likelihood estimation, A3s extends to

$$\mathbf{X} \sim N_{n,p}\left(\mathbf{O}, (\mathbf{I}_n \otimes \boldsymbol{\Sigma}_X)\right).$$

Our main objective is to estimate $\boldsymbol{\beta}$ under the aforementioned set up, including particularly (2). As part of the theory rests on the modification of classical solutions to the EIV problem, we begin, in the next section, with a brief summary of these solutions. Two approaches, extending the theory to the high-dimensional case, are discussed in Sect. 3.

2 Classical Estimation Approaches

To motivate the classical approaches, we may begin with a naive approach, following the usual least-squares theory, so that

$$\widetilde{\boldsymbol{\beta}}_{\text{Naive}} \in \arg\min_{\beta \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{W}\boldsymbol{\beta}\|^2 . \tag{3}$$

Writing the complete set of naive solutions,

$$\left\{(\mathbf{W}^T\mathbf{W})^{-}\mathbf{W}^T\mathbf{Y} + \mathbf{Q}\mathbf{z} :\ \mathbf{z} \in \mathbb{R}^p\right\}$$

with $\mathbf{Q} = \mathbf{I}_p - (\mathbf{W}^T\mathbf{W})^{-}(\mathbf{W}^T\mathbf{W})$, the favorite estimate in this set is

$$\widehat{\boldsymbol{\beta}}_{\text{Naive}} = (\mathbf{W}^T\mathbf{W})^{-}\mathbf{W}^T\mathbf{Y}.$$

It can be shown that $\widehat{\beta}_{\text{Naive}}$ is inconsistent for p fixed and $n \to \infty$, but becomes consistent as $\sigma_w^2 \to 0$. It, however, follows that $\mathrm{E}(\widehat{\boldsymbol{\beta}}|\mathbf{X}, \boldsymbol{\Delta}) = (\mathbf{W}^T\mathbf{W})^{-}\mathbf{W}^T\mathbf{X}\boldsymbol{\beta} \neq \beta$. Assuming $\mathrm{Cov}(\boldsymbol{\Delta}_i)$ known, and in particular having a simple structure such as $\sigma_w^2\mathbf{I}$, a correction to $\widehat{\boldsymbol{\beta}}_{\text{Naive}}$ can be proposed such that the corrected estimator,

$$\widehat{\boldsymbol{\beta}}_{\text{Corr}} = \arg\min_{\beta \in \mathbb{R}^p} \left(\|\mathbf{Y} - \mathbf{W}\boldsymbol{\beta}\|^2 - \sigma_w^2\|\boldsymbol{\beta}\|^2\right)$$

is consistent for p fixed and $n \to \infty$ or for $\sigma_w^2 \to 0$. The impracticality of this approach stems, among other things, from the fact that the true error variance needs to be known. Adding other classical solutions, namely Ridge and total least squares (TLS) estimators, the family of estimators can be compactly represented as

$$\widehat{\boldsymbol{\beta}}_{\text{NAME}} = (\mathbf{W}'\mathbf{W} + \lambda\mathbf{I})^{-1}\mathbf{W}'\mathbf{Y} \quad \text{with} \quad \text{NAME} = \begin{cases} \text{Naive if } \lambda = 0 \\ \text{Corr} \quad \text{if } \lambda = -\sigma_w^2 \\ \text{Ridge if } \lambda = \lambda \\ \text{TLS} \quad \text{if } \lambda = -\lambda_{p+1}(\mathbf{M}) \end{cases} \tag{4}$$

where $\lambda_{p+1}(\mathbf{M}) = \lambda_{\min}$ is the minimum eigenvalue of the partitioned matrix

$$\mathbf{M} = \begin{pmatrix} \mathbf{W}'\mathbf{W} & \mathbf{W}'\mathbf{Y} \\ \mathbf{Y}'\mathbf{W} & \mathbf{Y}'\mathbf{Y} \end{pmatrix}. \tag{5}$$

The corresponding estimators for Ridge and TLS follow from

$$\widehat{\boldsymbol{\beta}}_{\text{Ridge}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left(\|\mathbf{Y} - \mathbf{W}\boldsymbol{\beta}\|^2 + \lambda\,\|\boldsymbol{\beta}\|^2\right)$$

and, given $\sigma_\varepsilon^2 = \sigma_w^2 = \sigma^2$,

$$\widehat{\boldsymbol{\beta}}_{\text{TLS}} \in \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p,\mathbf{X}} \left(\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \|\mathbf{X} - \mathbf{W}\|^2\right),$$

respectively. The TLS approach is by far the most commonly used in the literature. In fact, the TLS estimator is also MLE under normality assumption in the functional model. The minimum of TLS problem over $\mathbf{X}$ can be shown to be

$$\min_{\mathbf{X}} \left(\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \|\mathbf{X} - \mathbf{W}\|^2\right) = \frac{\|\mathbf{M}^{1/2}\boldsymbol{\eta}\|^2}{\|\boldsymbol{\eta}\|^2} \quad \text{where} \quad \boldsymbol{\eta} = \begin{pmatrix} \boldsymbol{\beta} \\ -1 \end{pmatrix} \in \mathbb{R}^{p+1} \qquad (6)$$

with $\mathbf{M}$ defined above. Now, for $n > p$, the minimum eigenvalue $\lambda_{\min} > 0$, a.s., under the assumptions, so that

$$\widehat{\boldsymbol{\beta}}_{\text{TLS}} = \left(\mathbf{W}^T\mathbf{W} - \lambda_{\min}\mathbf{I}_p\right)^{-1}\mathbf{W}^T\mathbf{Y}.$$

3 Proposed Estimation Approaches for $p \gg n$

There have been a few recent attempts in the literature to address the question of estimation of parameter vector for high-dimensional EIV models. Most of these approaches offer a parallel line of action to the theory for linear models without EIV problem, e.g. using regularization or shrinkage. Almost all these approaches address the problem from optimization perspective. For example, [2] consider a minimization problem using a robust *matrix uncertainty* estimator of the error component in the EIV model, where the idea was originally introduced in [10]. The solution is obtained under certain sparsity conditions.

Loh and Wainright [8] propose a regularized corrected LASSO estimator, using a non-convex objective function leading to a local minimum. A modification of [8], using convex objective function, is given in [4], which they name as convex conditional LASSO or CoCoLASSO; for details, see [4].

We note that, the aforementioned approaches impose a penalty term on the parameter vector to restrict the estimation space and then find a solution under certain strong constraints, usually in addition to the other conditions needed to deal with high-dimensional linear models. In fact, the treatment of EIV model in the literature has an additional issue, namely that the EIV model estimation is often put forth purely as an optimization problem, frequently non-convex, without earnestly taking care of the underlying statistical nature of the question. Consequently, the resulting estimators may usually have some large sample numerical properties, particularly consistency, but only at the cost of loosing the statistical spirit of the problem. This in turn results into its twin problem of non-interpretability of the obtained results.

We consider two alternative approaches to the EIV problem for high-dimensional case. The first of these attempts, assuming a functional model, generalizes the classical TLS solution introduced above. The second option assumes a structural model and is mainly based on a well-conditioned estimator of the true covariance matrix. We address these approaches in the next two subsections.

3.1 Generalized TLS Estimator for Functional Model

Recall the optimal TLS solution in the classical case discussed above, where $\widehat{\boldsymbol{\beta}}_{\text{TLS}}$ is a solution of the minimization problem

$$\min_{\mathbf{X}} \left(\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \|\mathbf{X} - \mathbf{W}\|^2\right) \;=\; \frac{1}{1 + \|\boldsymbol{\beta}\|^2}\|\mathbf{Y} - \mathbf{W}\boldsymbol{\beta}\|^2 \tag{7}$$

or, equivalently

$$\widehat{\boldsymbol{\beta}}_{\text{TLS}} \in \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p, \boldsymbol{\eta}^T=(\boldsymbol{\beta}^T,-1)} \frac{\|\mathbf{M}^{1/2}\boldsymbol{\eta}\|^2}{\|\boldsymbol{\eta}\|^2}, \tag{8}$$

with $\mathbf{M} \in \mathbb{R}^{(p+1)\times(p+1)}$ defined above. Whereas the unique solution in the $n > p$ case follows by using $\lambda_{\min}(\mathbf{M})$, the same does not hold for $p > n$ case whence $r(\mathbf{M}) = n$, a.s., so that $m = p + 1 - n$ of the eigenvalues are zero, i.e. $0, \ldots, 0 < \lambda_n \leq \lambda_{n-1} \leq \ldots \leq \lambda_1$, where $r(\cdot)$ denotes the rank of a matrix. Let $\mathbf{e}_j$ denote eigenvector corresponding to the eigenvalue λ_j, $j = 1, \ldots, p + 1$. If we let $\mathbf{e}_{n+1}, \ldots, \mathbf{e}_{p+1}$ to be $(p + 1)$-dimensional eigenvectors corresponding to the m zero eigenvalues, we have

$$\mathbb{R}^{p+1} = \mathscr{L}\{\mathbf{e}_1, \ldots, \mathbf{e}_n\} \oplus \mathscr{L}\{\mathbf{e}_{n+1}, \ldots, \mathbf{e}_{p+1}\},$$

where

$$\mathscr{L}_0 = \mathscr{L}\{\mathbf{e}_{n+1}, \ldots, \mathbf{e}_{p+1}\} = \mathscr{L}^{\perp}\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$$

is the space corresponding to the zero eigenvalues. For each $\mathbf{e} = (e_1, \ldots, e_{p+1})^T \in \mathscr{L}_0$ with $e_{p+1} \neq \mathbf{0}$, we obtain a TLS estimator by

$$\widehat{\beta}_{\text{TLS}} = -\left(\frac{e_1}{e_{p+1}}, \ldots, \frac{e_p}{e_{p+1}}\right)^T \in \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p, \boldsymbol{\eta}^T=(\boldsymbol{\beta}^T,-1)} \frac{\|\mathbf{M}^{1/2}\boldsymbol{\eta}\|^2}{\|\boldsymbol{\eta}\|^2}.$$

This implies that the minimum of the ratio in (8) is zero. Hence

$$\frac{1}{1 + \|\widehat{\boldsymbol{\beta}}_{\text{TLS}}\|^2}\left\|\mathbf{Y} - \mathbf{W}\widehat{\boldsymbol{\beta}}_{\text{TLS}}\right\|^2 = 0 \;\Leftrightarrow\; \left\|\mathbf{Y} - \mathbf{W}\widehat{\boldsymbol{\beta}}_{\text{TLS}}\right\|^2 = 0$$

so that

$$\{\widehat{\boldsymbol{\beta}}_{\text{TLS}}\} = \{\widetilde{\boldsymbol{\beta}}_{\text{Naive}}\}. \tag{9}$$

Now, using the bounds on the ratios of quadratic forms (see e.g. [5, Chap. 2]), we can write

$$\min_{\mathbf{e}\,\in\,\mathscr{L}_0^{\perp}} \frac{\|\mathbf{M}^{1/2}\mathbf{e}\|^2}{\|\mathbf{e}\|^2} \;=\; \lambda_n, \tag{10}$$

where λ_n is the minimum positive eigenvalue of $\mathbf{M}$. We therefore propose, as a possible solution, to use $-\lambda_n$ in (4), call it *generalized TLS* (GTLS), and assess its feasibility as an optimal solution. Hence, we define

$$\widehat{\boldsymbol{\beta}}_{\text{GTLS}} = \mathop{\arg\min}_{\boldsymbol{\eta}^T=(\boldsymbol{\beta}^T,-1),\boldsymbol{\eta}\in\mathscr{L}_0^{\perp}} \frac{\|\mathbf{M}^{1/2}\boldsymbol{\eta}\|^2}{\|\boldsymbol{\eta}\|^2}.$$

It then holds that

$$\widehat{\boldsymbol{\beta}}_{\text{GTLS}} = \left(\frac{e_{n,1}}{e_{n,p+1}},\ldots,\frac{e_{n,p}}{e_{n,p+1}}\right)^T,$$

where $\mathbf{e}_n$ is the eigenvector corresponding to λ_n. The GTLS solution now follows, using $\mathbf{M}$ in (5), from the eigenvalue equation

$$\mathbf{M}\begin{pmatrix}\widehat{\boldsymbol{\beta}}_{\text{GTLS}}\\-1\end{pmatrix} = \lambda_n\begin{pmatrix}\widehat{\boldsymbol{\beta}}_{\text{GTLS}}\\-1\end{pmatrix}$$

so that

$$\widehat{\boldsymbol{\beta}}_{\text{GTLS}} = \left(\mathbf{W}^T\mathbf{W} - \lambda_n\mathbf{I}_p\right)^{-1}\mathbf{W}^T\mathbf{Y}.$$

Note that, due to the use of the eigenvector $\mathbf{e}_n$ corresponding to λ_n, the GTLS solution above does not lead to any overfit, since

$$\min_{\mathbf{X}}\left(\|\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}}_{\text{GTLS}}\|^2 + \|\mathbf{X}-\mathbf{W}\|^2\right) \;=\; \lambda_n > 0.$$

3.2 Estimators for Structural Model

Recall the ith observation of the model

$$Y_i = \mathbf{X}_i^T\boldsymbol{\beta} + \varepsilon_i, \quad \mathbf{W}_i = \mathbf{X}_i + \boldsymbol{\Delta}_i,\; i = 1,\ldots,n.$$

Let $\mathbf{X}_i$ be iid with $\mathrm{E}(\mathbf{X}_i) = \mathbf{0}$, $\mathrm{Cov}(\mathbf{X}_i) = \boldsymbol{\Sigma}_X$. Then, for $\mathbf{Z}_i = (\mathbf{W}_i,\ Y_i)^T \in \mathbb{R}^{p+1}$, we have, under the assumptions A1, A2s and A3s, an iid sample of $\mathbf{Z}$ with the structured covariance matrix

$$\begin{aligned}\mathrm{Cov}(\mathbf{Z}) = \boldsymbol{\Sigma}(\boldsymbol{\beta}) &= \mathbf{L}_\beta \boldsymbol{\Sigma}_X \mathbf{L}_\beta^T + \sigma^2 \mathbf{I}_{p+1} \\ &= \begin{pmatrix} \boldsymbol{\Sigma}_X + \sigma^2 \mathbf{I} & \boldsymbol{\Sigma}_X \boldsymbol{\beta} \\ \boldsymbol{\beta}^T \boldsymbol{\Sigma}_X & \boldsymbol{\beta}^T \boldsymbol{\Sigma}_X \boldsymbol{\beta} + \sigma^2 \end{pmatrix},\end{aligned}$$

using $\sigma_\varepsilon^2 = \sigma_w^2 = \sigma^2$, where

$$\mathbf{L}_\beta = \left(\boldsymbol{\beta}^T \ \ \mathbf{I}_p\right)^T, \quad \mathbf{L}_\beta^\perp = \left(\boldsymbol{\beta}^T, \ -1\right)^T \quad \Rightarrow \quad \mathbf{L}_\beta \mathbf{L}_\beta^\perp = \mathbf{0}.$$

In particular, if we let $\boldsymbol{\Sigma}_X = \mathbf{I}$, then

$$\boldsymbol{\Sigma}(\boldsymbol{\beta}) = \mathbf{L}_\beta \mathbf{L}_\beta^T + \sigma^2 \mathbf{I}_{p+1} = \begin{pmatrix} (1+\sigma^2)\mathbf{I}_P & \boldsymbol{\beta} \\ \boldsymbol{\beta}^T & \|\boldsymbol{\beta}\|^2 + \sigma^2 \end{pmatrix} \tag{11}$$

with the eigenvalues $\sigma^2 + \lambda(\mathbf{L}_\beta \mathbf{L}_\beta^T)$, where $\lambda(\mathbf{L}_\beta \mathbf{L}_\beta^T)$ denotes the eigenvalues of $\mathbf{L}_\beta \mathbf{L}_\beta^T$, namely

$$0, 1, \ldots, 1, 1 + \|\boldsymbol{\beta}\|^2.$$

Note also that, $\mathbf{L}_\beta^\perp$ is the eigenvector corresponding to the smallest eigenvalue of $\boldsymbol{\Sigma}(\boldsymbol{\beta})$. It now follows that the solution under this alternative depends on an efficient, stable, estimator of $\boldsymbol{\Sigma}(\boldsymbol{\beta})$, say $\widehat{\boldsymbol{\Sigma}}$, which is validly applicable for $p \gg n$ case. To proceed further, we begin with the following definition of $\widehat{\boldsymbol{\beta}}$ that holds for such a well-defined estimator $\widehat{\boldsymbol{\Sigma}}$ of $\boldsymbol{\Sigma}(\boldsymbol{\beta})$.

Definition 1 Given an estimator $\widehat{\boldsymbol{\Sigma}}$, $\widehat{\boldsymbol{\beta}}$ is defined such that

$$\widehat{\boldsymbol{\Sigma}} \begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ -1 \end{pmatrix} = \widehat{\lambda}_{\min} \begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ -1 \end{pmatrix},$$

where $(\widehat{\boldsymbol{\beta}}, -1)^T$ is an eigenvector corresponding to the minimum eigenvalue of $\widehat{\boldsymbol{\Sigma}}$, i.e. $\widehat{\lambda}_{\min}$.

Obviously, Definition 1 leaves us a set of possible estimators corresponding to a variety of options of $\widehat{\boldsymbol{\Sigma}}$. For example, consider $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\mathbf{M}$ with $n > p$. Then $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{\mathrm{TLS}}$. In fact, analogous to the functional case, we can allow $\widehat{\boldsymbol{\Sigma}}$ to be partitioned, according to $\mathbf{Z}_i = (\mathbf{W}_i, \ Y_i)^T$, as

$$\widehat{\boldsymbol{\Sigma}} = \begin{pmatrix} \widehat{\boldsymbol{\Sigma}}_{WW} & \widehat{\boldsymbol{\Sigma}}_{WY} \\ \widehat{\boldsymbol{\Sigma}}_{YW} & \widehat{\boldsymbol{\Sigma}}_{YY} \end{pmatrix}, \tag{12}$$

so that a general solution can be stated as following.

(**A4**) Assume $\lambda_{\min}(\widehat{\boldsymbol{\Sigma}}) > 0$ and $\lambda_{\min}\{\widehat{\boldsymbol{\Sigma}}_{WW}\} > \lambda_{\min}(\widehat{\boldsymbol{\Sigma}})$.

Note that, the second part of Assumption A4 is the requirement of a sharp interlacing inequality (see e.g. [11]).

Theorem 1 *Given $\widehat{\boldsymbol{\Sigma}}$ partitioned as in (12), with $\lambda_{\min}(\widehat{\boldsymbol{\Sigma}}) = \widehat{\lambda}_{\min}$, let A4 hold. Then*

$$\widehat{\boldsymbol{\beta}} = \left(\widehat{\boldsymbol{\Sigma}}_{WW} - \widehat{\lambda}_{\min}\mathbf{I}\right)^{-1}\widehat{\boldsymbol{\Sigma}}_{WY}. \tag{13}$$

Proof The proof of Theorem 1 follows by considering the eigenvalue equation, $\widehat{\boldsymbol{\Sigma}}\boldsymbol{\eta} = \widehat{\lambda}_{\min}\widehat{\boldsymbol{\eta}}$, with $\widehat{\boldsymbol{\eta}} = \left(\widehat{\boldsymbol{\beta}} \ -1\right)^T$, so that

$$\widehat{\boldsymbol{\Sigma}}_{WW}\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\Sigma}}_{WY} = \widehat{\lambda}_{\min}\widehat{\boldsymbol{\beta}} \ \Rightarrow \ \left(\widehat{\boldsymbol{\Sigma}}_{WW} - \widehat{\lambda}_{\min}\mathbf{I}\right)\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\Sigma}}_{WY}.$$

Under Assumption A4, the matrix $\widehat{\boldsymbol{\Sigma}}_{WW} - \widehat{\lambda}_{\min}\mathbf{I}$ can be inverted which leads to $\widehat{\boldsymbol{\beta}} = \left(\widehat{\boldsymbol{\Sigma}}_{WW} - \widehat{\lambda}_{\min}\mathbf{I}\right)^{-1}\widehat{\boldsymbol{\Sigma}}_{WY}$.

3.3 Estimators of Covariance Matrix

In the sequel, we primarily focus on the problem of estimating $\boldsymbol{\Sigma}(\boldsymbol{\beta})$. Estimation of large covariance, and also precision, matrices have attracted huge attention of researchers, particularly due to application of such estimators in the multivariate theory, typically testing and classification. We specifically discuss a few approaches in the following which, for our perspective, are worth more attention than the others, and also because of their simplicity.

3.3.1 Shrinkage Covariance Estimators

Several attempts in the literature on the estimation of large covariance matrix use James-Stein shrinkage theory, so that for most of them the origin can be traced back to the idea Stein introduced in [13]. Further, in the context of shrinkage estimation, the idea most commonly followed is to define an objective function, either the likelihood function or as a linear, preferably convex, combination of the empirical covariance matrix and the target matrix the shrinkage is aimed at. Then the only difference in various proposals of different estimators mainly pertains to how the shrinkage intensity, in the form of constraint on the objective function, is set.

Stein's estimator

Stein [13] introduced his idea by re-writing the spectral decomposition of the empirical covariance estimator such that the eigenvalues are modified whereas the eigenvectors are left as usual, and then using a specially defined loss function. Precisely, for iid vectors $\mathbf{Z}_i \sim N_{p+1}(\mathbf{0}, \boldsymbol{\Sigma})$, $i = 1, \ldots, n$, it is well known that

$$\mathbf{M} = \sum_{i=1}^{n}\mathbf{Z}_i\mathbf{Z}_i^T \sim W_{p+1}(n, \boldsymbol{\Sigma}), \quad n \geq p+1.$$

It is further known that the joint distribution of the eigenvalues $\lambda_1 \geq \ldots \geq \lambda_{p+1}$ of $\mathbf{M}$ has a density proportional to (see e.g. [1])

$$\mathbf{\Pi}_{i<j}(\lambda_i - \lambda_j)\mathbf{\Pi}_{i=1}^{n} \exp\left(-\frac{\lambda^{2_i}}{4}\right) \text{ for } \lambda_1 \geq \ldots \geq \lambda_{p+1}.$$

The maximum likelihood estimator of $\boldsymbol{\Sigma}$ is $\mathbf{S} = \frac{1}{n}\mathbf{M}$ and the eigenvalues of $\boldsymbol{\Sigma}$, i.e. $\gamma_1 \geq \ldots \geq \gamma_{p+1}$ are estimated by $\widehat{\gamma}_j = \frac{1}{n}\lambda_j$. These estimated eigenvalues have the property that they underestimate the largest, γ_1, and overestimate the smallest, γ_{p+1}.

Write $\mathbf{M} = \mathbf{E}\mathbf{L}\mathbf{E}^T$, where $\mathbf{L} = \text{diag}(\lambda_1, \ldots, \lambda_{p+1})$, $\lambda_1 \geq \ldots \geq \lambda_{p+1}$, so that, the Stein estimator is

$$\widehat{\boldsymbol{\Sigma}}_{Stein} = \mathbf{E}\boldsymbol{\Phi}(\boldsymbol{\lambda})\mathbf{E}^T,$$

with $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{p+1})^T$, $\boldsymbol{\Phi}(\boldsymbol{\lambda}) = \text{diag}(\phi_1(\lambda), \ldots, \phi_{p+1}(\lambda))$, where $\phi_1(\lambda) \geq \ldots \geq \phi_{p+1}(\lambda)$ are defined as $\phi_j(\lambda) = \lambda_j/\alpha_j(\lambda)$, $j = 1, \ldots, p+1$, and

$$\alpha_j(\lambda) = n - p - 1 - 2j + 1 + 2\sum_{i>j} \frac{\lambda_i}{\lambda_j - \lambda_i} - 2\sum_{i<j} \frac{\lambda_j}{\lambda_i - \lambda_j}. \tag{14}$$

Stein used the aforementioned idea to provide estimators that outperform, for example the maximum likelihood estimator, subject to the following convex loss function which is invariant under a non-singular linear transformation.

$$L\left(\widehat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}\right) = \text{tr}\left(\boldsymbol{\Sigma}^{-1}\widehat{\boldsymbol{\Sigma}}\right) - \ln\left|\boldsymbol{\Sigma}^{-1}\widehat{\boldsymbol{\Sigma}}\right| - (p-1). \tag{15}$$

The key point here is to modify the eigenvalues but not the eigenvectors. Stein also provided an alternative version of $\phi_j(\lambda)$ which keeps the ordering of the eigenvalues; for details and proofs, see [13].

Condition-number regularized estimator

Won et al. [15] use the usual log-likelihood function

$$L(\boldsymbol{\Sigma}) \propto -\frac{1}{2}\text{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{M}) - \frac{n}{2}\ln|\boldsymbol{\Sigma}|$$

as an objective function subject to the constraint composed of condition number

$$\mathscr{M}_{\kappa_{max}} = \left\{\boldsymbol{\Sigma} : \frac{\gamma_{\max}}{\gamma_{\min}} \leq \kappa_{\max}\right\}$$

with $\gamma_{\max}$ and $\gamma_{\min}$ as the maximum and minimum eigenvalues of $\boldsymbol{\Sigma}$, respectively, and $\kappa_{\max} \geq 1$ is the tuning parameter (shrinkage intensity) such that $\kappa_{\max} = 1$ leads to the spherical matrix $\widehat{\sigma}\mathbf{I}$ as (unique) estimator of $\boldsymbol{\Sigma}$. In general, it leads to the estimator, named condition number regularized (CNR) estimator,

$$\widehat{\boldsymbol{\Sigma}}_{\text{CNR}} = \mathbf{E}\widetilde{\boldsymbol{\Lambda}}\mathbf{E}^T,$$

where $\widetilde{\boldsymbol{\Lambda}} = \text{diag}(\widetilde{\boldsymbol{\lambda}}_1, \ldots, \widetilde{\boldsymbol{\lambda}}_{p+1})$ with

$$\widetilde{\lambda}_i = \min\left\{\max\left(\tau^*, \frac{1}{n}\lambda_i\right), \kappa_{\max}\tau^*\right\} = \begin{cases} \tau^* & \text{if } \frac{1}{n}\lambda_i \leq \tau^* \\ \frac{1}{n}\lambda_i & \text{if } \tau^* < \frac{1}{n}\lambda_i < \kappa_{\max}\tau^* \\ \kappa_{\max}\tau^* & \text{if } \frac{1}{n}\lambda_i > \kappa_{\max}\tau^* \end{cases} \tag{16}$$

for some τ^* which is data-adaptively determined.

Ledoit-Wolf shrinkage estimator
Ledoit and Wolf [7] is a very recent proposal, whereas the authors have a history of related work on the estimation of high-dimensional covariance estimators, particularly their pioneering work in [6], where they introduced the objective function as a convex linear combination of the empirical estimator $\mathbf{S} = \frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i\mathbf{Z}_i^T$ and target estimator $\mathbf{T}$, i.e.

$$\widehat{\boldsymbol{\Sigma}}_{\text{LW}} = (1-\omega)\mathbf{S} + \omega\mathbf{T}.$$

Taking $\mathbf{T} = \widehat{\sigma}^2\mathbf{I}$, with $\widehat{\sigma}^2 = \text{tr}(\mathbf{S})/p$, the shrinkage intensity follows as

$$\omega = \frac{b^2}{d^2},$$

with

$$b^2 = \min(\overline{b}^2, d^2), \quad d^2 = \|\mathbf{S} - \widehat{\sigma}^2\mathbf{I}\|^2$$

where

$$\overline{b}^2 = \frac{1}{n^2}\sum_{i=1}^n \|\mathbf{Z}_i\mathbf{Z}_i^T - \mathbf{S}\|^2.$$

Bodnar et al. shrinkage estimator
Following [6], [3] use a similar approach to estimate covariance matrix which they call *generalized linear shrinkage estimator* (GLSE). They propose an oracle estimator depending on the true covariance matrix $\boldsymbol{\Sigma}$

$$\widehat{\boldsymbol{\Sigma}}_{\text{GLSE}} = \widehat{\alpha}\mathbf{S} + \widehat{\beta}\boldsymbol{\Sigma}_0$$

with $\boldsymbol{\Sigma}_0$ such that $\text{tr}(\boldsymbol{\Sigma}_0) \leq M$ and $\widehat{\alpha}$, $\widehat{\beta}$ measure the shrinkage intensity, computed as

$$\widehat{\alpha} = \frac{\text{tr}(\mathbf{S}\boldsymbol{\Sigma})\|\boldsymbol{\Sigma}_0\|^2 - \text{tr}(\mathbf{S}\boldsymbol{\Sigma}_0)\text{tr}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0)}{\|S\|^2\|\boldsymbol{\Sigma}_0\|^2 - \{\text{tr}(\mathbf{S}\boldsymbol{\Sigma}_0)\}^2}$$

$$\widehat{\beta} = \frac{\text{tr}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0)\|\mathbf{S}\|^2 - \text{tr}(\mathbf{S}\boldsymbol{\Sigma})\text{tr}(\mathbf{S}\boldsymbol{\Sigma}_0)}{\|S\|^2\|\boldsymbol{\Sigma}_0\|^2 - \{\text{tr}(\mathbf{S}\boldsymbol{\Sigma}_0)\}^2}.$$

Further they suppose a bona fide version of this estimator, where $\widehat{\alpha}, \widehat{\beta}$ are fitted by the data. Note that, depending on the structure of $\boldsymbol{\Sigma}_0$, the estimator may or may not be of Stein type.

3.3.2 MLE under Restrictions

Recall $\mathbf{Z}_i = (\mathbf{W}_i^T, Y_i) \in \mathbb{R}^{p+1}$ and define $\mathbf{M} = \sum_{i=1}^n \mathbf{Z}_i \mathbf{Z}_i^T \in \mathbb{R}^{(p+1)\times(p+1)}$. Using $\mathbf{Z}_i \sim N_{p+1}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\beta}))$, we consider the likelihood function

$$L(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-n/2} \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\mathbf{M}\boldsymbol{\Sigma}^{-1}\right)\right\}. \tag{17}$$

Further, denote the eigenvalues of $\boldsymbol{\Sigma}$ as $\gamma_1 \geq \gamma_2 \geq \ldots \geq \gamma_{p+1}$ and those of $\mathbf{M}$ as

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n > 0 \text{ and } \lambda_{n+1} = \ldots = \lambda_{p+1} = 0.$$

We prove the following lemma.

Lemma 1 *Given $L(\boldsymbol{\Sigma})$ in (17), we have*

$$L(\boldsymbol{\Sigma}) \leq \left(\prod_{i=n+1}^{p+1} \gamma_i\right)^{-\frac{n}{2}} \left(\prod_{i=1}^{n} \lambda_i\right)^{-\frac{n}{2}} n^{n^2} \exp\left(-\frac{n^2}{2}\right). \tag{18}$$

Proof First, it follows from von Neumann's inequality for eigenvalues [11, p. 120] that

$$\mathrm{tr}\left(\mathbf{M}\boldsymbol{\Sigma}^{-1}\right) \geq \sum_{i=1}^{p} \lambda_i \cdot \frac{1}{\gamma_i} = \sum_{i=1}^{n} \lambda_i \cdot \frac{1}{\gamma_i} = \sum_{i=1}^{n} \eta_i,$$

for the eigenvalues of $\mathbf{M}$ and $\boldsymbol{\Sigma}$ given above, where $\eta_i = \lambda_i/\gamma_i$. This implies that

$$\exp\left\{-\frac{1}{2}\mathrm{tr}\left(\mathbf{M}\boldsymbol{\Sigma}^{-1}\right)\right\} \leq \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\eta_i\right\} = \prod_{i=1}^{n} \exp\left\{-\frac{1}{2}\eta_i\right\}.$$

Further, using the properties of determinant, we can write

$$|\boldsymbol{\Sigma}|^{n/2} = \left(\prod_{i=1}^{p+1}\gamma_i\right)^{n/2} = \left(\prod_{i=1}^{n}\gamma_i\right)^{n/2} \left(\prod_{i=n+1}^{p+1}\gamma_i\right)^{n/2}.$$

Hence, from (17), we have

$$L(\boldsymbol{\Sigma}) \leq \left(\prod_{i=n+1}^{p+1} \gamma_i\right)^{-\frac{n}{2}} \left(\prod_{i=1}^{n} \lambda_i\right)^{-\frac{n}{2}} \left(\prod_{i=1}^{n} \eta_i\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} \eta_i\right\}$$
$$\leq \left(\prod_{i=n+1}^{p+1} \gamma_i\right)^{-\frac{n}{2}} \left(\prod_{i=1}^{n} \lambda_i\right)^{-\frac{n}{2}} \prod_{i=1}^{n}\left\{\eta_i^{n/2} \exp\left(-\frac{1}{2}\eta_i\right)\right\}. \tag{19}$$

Now, using the fact that the function

$$f(x) = x^{\frac{n}{2}} \exp\left(-\frac{1}{2}x\right)$$

attains its maximum at $x = n$, we obtain for the last part of (19), that $\eta_{\max} = n$ so that the bound also holds for every $\eta_i = n$, which in turn implies $\widehat{\gamma_i} = \lambda_i/n, i = 1, \ldots, n$. Finally, the last product in (19) can be bounded above at $n^{n/2}\exp(-n/2)$ such that

$$L(\boldsymbol{\Sigma}) \leq \left(\prod_{i=n+1}^{p+1} \gamma_i\right)^{-\frac{n}{2}} \left(\prod_{i=1}^{n} \lambda_i\right)^{-\frac{n}{2}} n^{\frac{n}{2}} \exp\left(-\frac{n}{2}\right). \tag{20}$$

Corollary 1 *Given $\boldsymbol{\Sigma}$ with $\gamma_{n+1} \leq \frac{1}{n}\lambda_n$ and $\gamma_j = \frac{1}{n}\lambda_j$ for $j = 1, \ldots, n$. Then $\boldsymbol{\Sigma}$ attains the bound in (18).*

We know that the true underlying covariance matrix is positive definite and has the eigenvalues $\gamma_1 \geq \gamma_2 \geq \ldots \geq \gamma_{p+1} > 0$. Nevertheless

$$\sup_{\boldsymbol{\Sigma},\gamma_{p+1}>0} L(\boldsymbol{\Sigma}) = \infty.$$

We define a constrained space, using some constants $c_{n+1} \geq \ldots \geq c_{p+1}$, by

$$\mathscr{M}_c = \{\boldsymbol{\Sigma} : \gamma_j \geq c_j, j = n+1, \ldots, p+1\}.$$

Then under $\gamma_{n+1} \leq \frac{1}{n}\lambda_n$ it holds that

$$\max_{\boldsymbol{\Sigma}\in\mathscr{M}_c} L(\boldsymbol{\Sigma}) = L(\widehat{\boldsymbol{\Sigma}}_c)$$

with

$$\widehat{\boldsymbol{\Sigma}}_c = \mathbf{S} + \mathbf{E}\boldsymbol{\Gamma}\mathbf{E}^T, \quad \boldsymbol{\Gamma} = \text{diag}(0, \ldots, 0, c_{n+1}, c_{n+2}, \ldots, c_{p+1}).$$

This is a Stein-type estimator because

$$\widehat{\boldsymbol{\Sigma}}_c = \mathbf{E}\boldsymbol{\Phi}_c\mathbf{E}^T$$

with $\boldsymbol{\Phi}_c = \mathrm{diag}(\frac{1}{n}\lambda_1, \ldots, \frac{1}{n}\lambda_n, c_{n+1}, c_{n+2}, \ldots, c_{p+1})$. Further, $\widehat{\boldsymbol{\Sigma}}_0 = \mathbf{E}\boldsymbol{\Gamma}\mathbf{E}^T$ is the orthogonal complement to $\mathbf{S}$ mainly depending on the constraints. It holds that $\widehat{\boldsymbol{\Sigma}}_0\mathbf{S} = \mathbf{0}$.

Theorem 2 *Given a Stein-type covariance estimator* $\widetilde{\boldsymbol{\Sigma}} = \mathbf{E}\widetilde{\boldsymbol{\Phi}}\mathbf{E}^T$, *where*

$$\widetilde{\boldsymbol{\Phi}} = diag(\phi_1, \ldots, \phi_{p+1})$$

with $\phi_1 \geq \phi_2, \ldots, \geq \phi_p \geq \phi_{p+1} > 0$. *Then the estimator* $\widehat{\beta}$ *given in Definition 1 coincides with the naive estimator* $\widehat{\beta}_{Naive}$.

Proof As the Stein estimator has the same eigenvectors as $\mathbf{S}$, therefore, $\widehat{\beta}$ is defined by $\mathbf{e}_n$, the eigenvector corresponding to the smallest eigenvalue. This means $\widehat{\beta} = \widehat{\beta}_{\mathrm{TLS}}$. The statement then follows from (3).

In fact, the result holds for all naive estimators since the order of the last m eigenvectors of $\mathbf{S}$ is arbitrary.

Note also that the estimator $\widehat{\boldsymbol{\Sigma}}_{\mathrm{GLSE}}$ is not of Stein-type, but the eigenvector corresponding to the minimal eigenvalue depends essentially on $\boldsymbol{\Sigma}_0$, which is chosen in advance.

One way out may be to take the set of covariance matrices

$$\mathscr{M}_{min} = \{\boldsymbol{\Sigma} : \gamma_j = \frac{1}{n}\lambda_n,\ j = n+1, \ldots, p+1\}.$$

Then the eigenvectors related to the smallest eigenvalue are $\mathbf{e}_n, \ldots, \mathbf{e}_{p+1}$. The only non-naive solution is then $\widehat{\beta}_{\mathrm{GTLS}}$.

References

1. Anderson, G.W., Guionnet, A., Zeitouni, O.: An Introduction to Random Matrices. Cambridge University Press, Cambridge (2016)
2. Belloni, A., Rosenbaum, M., Tsybakov, A.B.: An l_1, l_2, l_∞-regularization approach to high-dimensional errors-in-variables models. arXiv:1412.7216v1 (2016)
3. Bodnar, T., Gupta, A.K., Parolya, N.: On the strong convergence of the optimal linear shrinkage estimator for large dimensional covariance matrix. J. Multiv. Anal. **132**, 215–228 (2014)
4. Datta, A., Zou, H.: COCOLASSO for high-dimensional error-in-variables regression. arXiv:1510.071262v2 (Jan 2016)
5. Johnson, R.A., Wichern, D.W.: Applied Multivariate Data Analysis, 6th edn. Prentice Hall, NJ (2007)
6. Ledoit, O., Wolf, M.: A well-conditioned estimator for large dimensional covariance matrices. J. Multiv. Anal. **88**, 365–411 (2004)
7. Ledoit, O., Wolf, M.: Optimal estimation of a large-dimensional covariance matrix under Stein's loss. Bernoulli **24**(4B), 3791–3832 (2018)
8. Loh, P.-L., Wainright, M.J.: High-dimensional regression with noisy and missing data: provable guarantees with nonconvexity. Ann. Stat. **40**, 1637–1664 (2012)
9. Rajaratnam, B., Vincenzi, D.: A note on covariance estimation in the unbiased estimator of risk framework. J. Stat. Plann. Inf. **175**, 25–39 (2016)

10. Rosenbaum, M., Tsybakov, A.B.: Sparse recovery unde rmatrix uncertainty. Ann. Stat. **38**, 2620–2651 (2010)
11. Seber, G.A.F.: A Matrix Handbook for Statisticians. Wiley, New York, NY (2008)
12. Sørensen, Ø., Frigessi, A., Thoresen, M.: Measurement error in LASSO: impact and likelihood bias correction. Stat. Sin. **25**, 809–829 (2013)
13. Stein, C.: Lectures on the theory of estimation of many parameters. J. Math. Sci. **34**, 1373–1403 (1986)
14. Warton, D.I.: Penalized normal likelihood and ridge regularization of correlation and covariance matrices. J. Am. Stat. Assoc. **103**, 340–349 (2008)
15. Won, J.-H., Lim, J., Kim, S.-J., Rajaratnam, B.: Condition-number-regularized covariance estimation. JRSS B **75**, 427–450 (2013)
16. Zwanzig, S.: On a consistent rank estimate in a linear structural model. Tatra Mt. Math. Publ. **51**, 191–202 (2012)