ISSN 1645-6726

INSTITUTO NACIONAL DE ESTATÍSTICA
PORTUGAL

REVSTAT

Statistical Journal

Special Issue on Robust Statistics

Guest Editors:
Ana M. Pires
Manuela Souto de Miranda

Volume 5, No.1
March 2007

Catalogação Recomendada

REVSTAT. Lisboa, 2003-
Revstat : statistical journal / ed. Instituto Nacional de Estatística. - Vol. 1, 2003- . - Lisboa I.N.E., 2003- . - 30 cm
Semestral. - Continuação de : Revista de Estatística = ISSN 0873-4275. - edição exclusivamente em inglês
ISSN 1645-6726

CREDITS

- **EDITOR-IN-CHIEF**
  - *M. Ivette Gomes*
- **CO-EDITOR**
  - *M. Antónia Amaral Turkman*
- **ASSOCIATE EDITORS**
  - *António Pacheco*
  - *Barry Arnold*
  - *Dani Gamerman*
  - *David Cox*
  - *Dinis Pestana*
  - *Edwin Diday*
  - *Gilbert Saporta*
  - *Helena Bacelar Nicolau*
  - *Isaac Meilijson*
  - *Jef Teugels*
  - *João Branco*
  - *Ludger Rüschendorf*
  - *M. Lucília Carvalho*
  - *Marie Husková*
  - *Nazaré Mendes-Lopes*
  - *Radu Theodorescu*
  - *Susie Bayarri*
- **EXECUTIVE EDITOR**
  - *Maria José Carrilho*
- **SECRETARY**
  - *Liliana Martins*
- **PUBLISHER**
  - *Instituto Nacional de Estatística (INE)*
    Av. António José de Almeida, 2
    1000-043 LISBOA
    PORTUGAL
    Tel.: + 351 218 426 100
    Fax: + 351 218 426 364
    Web site: http://www.ine.pt
    Customer Support Service
    (National network): 808 201 808
    (Other networks): + 351 226 050 748
- **COVER DESIGN**
  - *Mário Bouçadas, designed on the stain glass window at INE by the painter Abel Manta*
- **LAYOUT AND GRAPHIC DESIGN**
  - *Carlos Perpétuo*
- **PRINTING**
  - *Instituto Nacional de Estatística*
- **EDITION**
  - *450 copies*
- **LEGAL DEPOSIT REGISTRATION**
  - *N.º 191915/03*

PRICE

[VAT 5% included]

- Single issue	€ 10
- Annual subscription (No. 1 Special Issue, No. 2 and No.3)...	€ 24
- Annual subscription (No. 2, No. 3)	€ 16

FOREWORD

The present issue of *REVSTAT – Statistical Review* contains a selection of papers, both invited and contributed, presented at the **International Conference on Robust Statistics – *ICORS2006*** that took place in Lisbon, Portugal, from 16 to 21 July, 2006.

The event was organized under the auspicious of the *Centre for Mathematics and its Applications – CEMAT*, an R&D from the Technical University of Lisbon, Portugal and the European Science Foundation (through the SACD Network).

The *ICORS* conference brought together leading researchers in Robust Statistics and described the state of the art in the area. The meeting also provided an important forum for discussion of future directions and applications of robust statistical methods. Thanks are due to our colleagues of the Scientific Committee (Ana Bianco from the University of Buenos Aires, Argentina, Graciela Boente from the University of Buenos Aires, Argentina, Frank Critchley from the Open University, UK, Christophe Croux from the Catholic University of Leuven, Belgium, Luisa Fernholz from the Temple University, USA, Peter Filzmoser from the Vienna University of Technology, Austria, Ursula Gather from the University of Dortmund, Germany, Ricardo Maronna from the National University of La Plata, Argentina, Hannu Oja from the University of Tampere, Finland, Elvezio Ronchetti from the University of Geneva, Switzerland, Peter Rousseeuw from the University of Antwerp, Belgium, David Tyler from the Rutgers University, USA, Roy Welsch from the Massachusetts Institute of Technology, USA and Victor Yohai from the University of Buenos Aires, Argentina) who organized the scientific program of the conference and to the referees of the submitted papers whose punctual contribution has been essential to the present edition. We also express our gratitude to all the speakers for the high scientific standards of the *ICORS2006*.

Unfortunately it has not been possible to condense all the contributions into a single special issue, due to lack of space. Only seven papers have been selected.

The papers in this volume give an overview of the use of robust procedures, taking into account theoretical approaches, computational analysis or challenging applied problems. Davies and Gather discuss the fundamental concept of breakdown point for equivariant functionals, while Hennig and Kutlukaya study the choice and the design of loss functions. Victoria-Feser or Spangl and Dutter deal with technical problems in the use of robust procedures for specific models – the former adapts indirect inference to a generalized linear latent variable model and the latter points out to robust spectral density estimation. Zioutas *et al.* look into robust regression. Robust discriminant analysis methods are compared in the study of Todorov and Pires. Welsch and Zhou discuss the use of robust methods in investment management.

Finally we would like to thank our colleagues of the Local Organizing Committee (Peter Rousseeuw from the University of Antwerp. Belgium. Conceição Amado. Rosário Oliveira and Isabel Rodrigues, all from the Technical University of Lisbon, Portugal and Carla Pereira from the University of Oporto, Portugal) for their diligent work on the organization of the conference.

Ana Pires
Manuela Souto de Miranda

INDEX

REVSTAT – Statistical Journal
Volume 5, Number 1, March 2007, 1–17

THE BREAKDOWN POINT — EXAMPLES AND COUNTEREXAMPLES

Authors: P.L. Davies
– University of Duisburg-Essen, Germany, and
Technical University Eindhoven, Netherlands
laurie.davies@uni-essen.de

U. Gather
– University of Dortmund, Germany
gather@statistik.uni-dortmund.de

Abstract:

- The breakdown point plays an important though at times controversial role in statistics. In situations in which it has proved most successful there is a group of transformations which act on the sample space and which give rise to an equivariance structure. For equivariant functionals, that is those functionals which respect the group structure, a non-trivial upper bound for the breakdown point was derived in Davies and Gather (2005). The present paper briefly repeats the main results of Davies and Gather (2005) but is mainly concerned with giving additional insight into the concept of breakdown point. In particular, we discuss the attainability of the bound and the dependence of the breakdown point on the sample or distribution and on the metrics used in its definition.

Key-Words:

- *equivariance; breakdown point; robust statistics.*

AMS Subject Classification:

- Primary: 62G07; Secondary: 65D10, 62G20.

1. INTRODUCTION

The breakdown point is one of the most popular measures of robustness of a statistical procedure. Originally introduced for location functionals (Hampel, 1968, 1971) the concept has been generalized to scale, regression and — with more or less success — to other situations.

In Huber's functional analytic approach to robustness breakdown is related to the boundedness of a functional and the breakdown point is defined in terms of the sizes of neighbourhoods on the space of distributions. A simple and intuitive definition of the breakdown point but one restricted to finite samples, the finite sample breakdown point, was introduced by Donoho (1982) and Donoho and Huber (1983). Successful applications of the concept of breakdown point have been to the location, scale and regression models in $\mathbb{R}^k$ and to models which are intimately related to these (see for example Ellis and Morgenthaler, 1992, Davies and Gather, 1993, Hubert, 1997, Terbeck and Davies, 1998, He and Fung, 2000, Müller and Uhlig, 2001). The reason for this is that such models have a rich equivariance structure deriving from the translation or affine group operating on $\mathbb{R}^k$. By restricting the class of statistical functionals to those with the appropriate equivariance structure one can prove the existence of non-trivial highest breakdown points (Davies and Gather, 2005), which in many cases can be achieved, at least locally (Huber, 1981, Davies, 1993).

It is the aim of this paper to provide some additional insight into the definition of the breakdown point, to point out the limits of the concept and to give some results on the attainment of the upper bound.

We proceed as follows: Chapter 2 summarizes the definitions and theorems of Davies and Gather (2005). Chapter 3 shows via examples that the breakdown point is a local concept. Chapter 4 is devoted to the attainability of the bound and Chapter 5 to the choice of metrics. Chapter 6 contains some concluding remarks.

2. DEFINITIONS AND BOUNDS FOR THE BREAKDOWN POINT

Let T be a functional defined on some subfamily $\mathcal{P}_T$ of the family $\mathcal{P}$ of all distributions on a sample space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ which takes its values in some metric space (Θ, D) with

$$\sup_{\theta_1,\theta_2\in\Theta} D(\theta_1,\theta_2) = \infty \ . \tag{2.1}$$

The finite sample breakdown point of T at a sample $\boldsymbol{x}_n = (x_1, ..., x_n)$, $x_i \in \mathcal{X}$, $i = 1, ..., n$, is defined as

$$\text{(2.2)} \qquad \operatorname{fsbp}(T, \boldsymbol{x}_n, D) \;=\; \frac{1}{n} \min\left\{k \in \{1, ..., n\} \colon \sup_{Q_{n,k}} D\bigl(T(P_n), T(Q_{n,k})\bigr) = \infty\right\}$$

where $P_n = \sum_{i=1}^n \delta_{x_i}/n$ and $Q_{n,k}$ is the empirical distribution of a replacement sample with at least $n-k$ points from the original sample $\boldsymbol{x}_n$.

Example 2.1. If T is the median functional T_{med} defined on $\mathcal{P}_T = \mathcal{P}$ with $\Theta = \mathbb{R}$, and $D(\theta_1, \theta_2) = |\theta_1 - \theta_2|$, then

$$\text{(2.3)} \qquad \operatorname{fsbp}(T_{\text{med}}, x, D) \;=\; \left\lfloor \frac{n+1}{2} \right\rfloor / \, n \; .$$

A distributional definition of the breakdown point requires a metric d on $\mathcal{P}$ with

$$\sup_{P,Q \in \mathcal{P}} d(P, Q) \;=\; 1 \; .$$

The breakdown point of a functional T at a distribution $P \in \mathcal{P}_T$ w.r.t. d and D is then defined by

$$\text{(2.4)} \qquad \epsilon^*(T, P, d, D) \;=\; \inf\left\{\epsilon > 0 \colon \sup_{d(P,Q) < \epsilon} D\bigl(T(P), T(Q)\bigr) = \infty\right\}$$

where $D\bigl(T(P), T(Q)\bigr) := \infty$ if $Q \notin \mathcal{P}_T$.

Example 2.2. Let $\mathcal{P}$ and D be as in Example 2.1 and d be the Kolmogorov-metric $d_k(P, Q) = \sup_x |F_P(x) - F_Q(x)|$. For the expectation functional T_E

$$T_E(P) \;=\; E(P) := \int x \, dP(x) \;, \qquad \mathcal{P}_T \;=\; \Bigl\{P \in \mathcal{P} \colon E(P) \text{ exists}\Bigr\}$$

we have $\epsilon^*(T_E, P, d, D) = 0$ for any $P \in \mathcal{P}_T$, in contrast to the median for which $\epsilon^*(T_{\text{med}}, P, d, D) = 1/2$.

As already pointed out in the introduction the derivation of a non-trivial upper bound for the breakdown point requires a group structure. Assume that G is a group of measurable transformations of the sample space $\mathcal{X}$ onto itself. Then G induces a group of transformations of $\mathcal{P}$ onto itself via $P^g(B) = P(g^{-1}(B))$ for all sets $B \in \mathcal{B}(\mathcal{X})$. Let $H_g = \{h_g \colon g \in G\}$ be the group of transformations $h_g \colon \Theta \to \Theta$ which describes the equivariance structure of the problem. A functional $T \colon \mathcal{P}_T \to \Theta$ is called equivariant with respect to G if and only if $\mathcal{P}_T$ is closed under G and

$$\text{(2.5)} \qquad T(P^g) = h_g\bigl(T(P)\bigr) \qquad \text{for all} \quad g \in G, \;\; P \in \mathcal{P}_T \; .$$

Let

$$G_1 := \Big\{ g \in G \colon \lim_{n\to\infty} \inf_{\theta\in\Theta} D\big(\theta, h_{g^n}(\theta)\big) = \infty \Big\} \tag{2.6}$$

and define

$$\Delta(Q) := \sup\Big\{ Q(B) \colon \ B \in \mathcal{B}(\mathcal{X}), \ g_{|B} = \iota_{|B} \ \text{ for some } g \in G_1 \Big\} \tag{2.7}$$

where ι is the unit element of G. We cite the main result from Davies and Gather (2005):

Theorem 2.1. *Suppose that the metrics d and D satisfy the properties given above and additionally*

$$d\Big(\alpha P + (1-\alpha) Q_1, \ \alpha P + (1-\alpha) Q_2\Big) \le 1-\alpha\,, \quad P, Q_1, Q_2 \in \mathcal{P}, \ 0 < \alpha < 1\,, \tag{2.8}$$

$$G_1 \neq \emptyset\ . \tag{2.9}$$

Then for all G-equivariant functionals $T\colon \mathcal{P}_T \to \Theta$, for all $P \in \mathcal{P}_T$ and for all $\boldsymbol{x}_n$ we have respectively

a) $\epsilon^*(T, P, d, D) \le \dfrac{\big(1-\Delta(P)\big)}{2}\,,$

b) $\operatorname{fsbp}(T, \boldsymbol{x}_n, D) \le \left\lfloor \dfrac{n - n\,\Delta(P_n) + 1}{2} \right\rfloor \Big/ n\,.$

Proof: **a**) cf. Davies and Gather (2005).

b) The proof is similar to a) but it is not given in Davies and Gather (2005). We present it here as it illustrates the simplicity of the idea of the finite sample breakdown point. The basic idea of all such proofs may be found in Huber (1981) although it was clearly known to Hampel (1975) who stated the breakdown point of what is now known as the LMS estimator (see Rousseeuw, 1984). Donoho and Huber (1983) give the first calculations for the finite sample breakdown point both for multivariate location and for a high breakdown linear regression estimator based on the multivariate location estimator of Donoho (1982). The corresponding calculations for the LMS estimator may be found in Rousseeuw (1984). Firstly we note that there are exactly $n\,\Delta(P_n)$ points in $\boldsymbol{x}_n$ for which $g(x_i) = x_i$ for some $g \in G_1$. We assume without loss of generality that these are the sample points $x_1, ..., x_{n\Delta(P_n)}$. If $\Delta(P_n) = 0$ there are no such points and some obvious alterations to the following proof are required. To ease the notation we write

$$l(n) = \left\lfloor \frac{n - n\,\Delta(P_n) + 1}{2} \right\rfloor .$$

We consider the sample $\boldsymbol{y}^*_{n,k}$ given by

$$\boldsymbol{y}^*_{n,k} = \Big(x_1, ..., x_{n\Delta(P_n)}, ..., x_{n-l(n)}, \ g^m(x_{n-l(n)+1}), ..., g^m(x_n) \Big)$$

for some $m \ge 1$ and some $g \in G_1$. We denote its empirical distribution by $Q^*_{n,k}$.

The sample $\boldsymbol{y}^*_{n,k}$ contains at least $n-l(n)$ points of the original sample $\boldsymbol{x}_n$. The transformed sample $g^{-m}(\boldsymbol{y}^*_{n,k})$ is equal to

$$\Big(x_1,...,x_{n\Delta(P_n)},\; g^{-m}(x_{n\Delta(P_n)+1}),...,g^{-m}(x_{n-l(n)}),\; x_{n-l(n)+1},...,x_n\Big)\;.$$

It contains at least $n\,\Delta(P_n)+l(n)$ points of the original sample $\boldsymbol{x}_n$ and as

$$n\,\Delta(P_n)+l(n)\;\geq\;n-l(n)$$

it contains at least $n-l(n)$ points of $\boldsymbol{x}_n$. By the equivariance of T we have

$$T(Q^{*g^{-m}}_{n,k})\;=\;h_{g^{-m}}\big(T(Q^*_{n,k})\big)$$

from which it follows

$$D\Big(h_{g^{-m}}\big(T(Q^*_{n,k})\big),T(Q^*_{n,k})\Big)\;\leq\;D\big(T(P_n),T(Q^*_{n,k})\big)+D\big(T(P_n),T(Q^{*g^{-m}}_{n,k})\big)\;.$$

From $\lim\limits_{n\to\infty}\inf\limits_{\theta} D(\theta,h_{g^n}(\theta))=\infty$ for all $g\in G_1$ we have

$$\lim_{m\to\infty} D\Big(h_{g^{-m}}\big(T(Q^*_{n,k})\big),T(Q^*_{n,k})\Big)\;=\;\infty$$

and hence $D\big(T(P_n),T(Q^*_{n,k})\big)$ and $D\big(T(P_n),T(Q^{*g^{-m}}_{n,k})\big)$ cannot both remain bounded. We conclude that for any $k\geq\big\lfloor\frac{n-n\Delta(P_n)+1}{2}\big\rfloor$

$$\sup_{Q_{n,k}} D\big(T(P_n),T(Q_{n,k})\big)\;=\;\infty$$

from which the claim of the theorem follows. □

For examples of Theorem 2.1 we refer to Davies and Gather (2005).

3. THE BREAKDOWN POINT IS A LOCAL CONCEPT

As seen above the median T_{med} has a finite sample breakdown point of $\lfloor(n+1)/2\rfloor$ at every real sample $\boldsymbol{x}_n$ and this is the highest possible value for translation equivariant location functionals. If we consider scale functionals then the situation is somewhat different. The statistical folklore is that the highest possible finite sample breakdown point for any affine equivariant scale functional is $\lfloor n/2\rfloor/n$ and that this is attained by the median absolute deviation functional T_{MAD}. Some authors (Croux and Rousseeuw, 1992, Davies, 1993) are aware that this is not correct as is shown by the following sample

$$\boldsymbol{x}_{11}\;=\;\big(1.0,\,1.8,\,1.3,\,1.3,\,1.9,\,1.1,\,1.3,\,1.6,\,1.7,\,1.3,\,1.3\big)\;. \tag{3.1}$$

The fsbp of T_{MAD} at this sample is 1/11. This can be seen by replacing the data point 1.0 by 1.3 so that for the altered data set $T_{\text{MAD}} = 0$ which is conventionally defined as breakdown. If a sample has no repeated observations then T_{MAD} has a finite sample breakdown point of $\lfloor n/2 \rfloor / n$ and this is indeed the highest possible finite sample breakdown point for a scale functional. The difference between the maximal finite sample breakdown points for location and scale functionals is explained by Theorem 2.1. For the sample (3.1) we have $\Delta(P_n) = 5/11$ and the theorem gives

$$\text{fsbp}\big(T_{\text{MAD}}, \boldsymbol{x}_{11}, D\big) \;\le\; 3/11\ .$$

For a sample $\tilde{\boldsymbol{x}}_{11}$ without ties we have $\Delta(P_n) = 1/n$ and the theorem yields

$$\text{fsbp}\big(T_{\text{MAD}}, \tilde{\boldsymbol{x}}_{11}, D\big) \;\le\; \left\lfloor \frac{n}{2} \right\rfloor / n \;=\; 5/11\ .$$

From the above it follows that T_{MAD} may or may not attain the upper bound. We study this in more detail in the next chapter.

4. ATTAINING THE BOUND

4.1. Location functionals

From Theorem 2.1 above it is clear that the maximum breakdown point for translation equivariant location functionals is 1/2. This bound is sharp as is shown by the location equivariant L_1-functional

$$T(P) \;=\; \text{argmin}_\mu \int \big(\|x-\mu\| - \|x\|\big)\, dP(x)\ . \tag{4.1}$$

In general the L_1-functional is not regarded as a satisfactory location functional as it is not affine equivariant in dimensions higher than one. For an affinely equivariant location functional the set G_1 of (2.6) is now the set of pure non-zero translations and it follows that $\Delta(P) = 0$ for any distribution P. Theorem 2.1 gives an upper bound of 1/2 which is clearly attainable in one dimension. It is not however clear whether this bound is attainable in higher dimensions. Work has been done in this direction but it is not conclusive (Rousseeuw and Leroy, 1987, Niinimaa, Oja and Tableman, 1990, Lopuhaä and Rousseeuw, 1991, Gordaliza, 1991, Lopuhaä, 1992, Donoho and Gasko, 1992, Davies and Gather, 2005, Chapter 5 and the Discussion of Rousseeuw in Davies and Gather, 2005).

We first point out that the bound 1/2 is not globally sharp. Take a discrete measure in $\mathbb{R}^2$ with point mass 1/3 on the points $x_1 = (0, 1)$, $x_2 = (0, -1)$, $x_3 = (\sqrt{3}, 0)$. The points form a regular simplex. For symmetry reasons every affinely equivariant location functional must yield the value $(1/\sqrt{3},0)$. On replacing $(\sqrt{3},0)$ by $(\eta\sqrt{3},0)$ it is clear that each affinely equivariant location functional must result in $(\eta/\sqrt{3},0)$. On letting $\eta \to \infty$ it follows that the breakdown point of every affinely equivariant location functional cannot exceed 1/3. In k dimensions one can prove in a similar manner that $1/(k+1)$ is the maximal breakdown point for points on a regular simplex with $k+1$ sides.

In spite of the above example we now show that there are probability distributions at which the finite sample replacement breakdown point is 1/2 even if this cannot be obtained globally. We consider a sample $\boldsymbol{x}_n = (x_1, ..., x_n)$ of size n in $\mathbb{R}^k$ and form the empirical measure P_n given by $P_n = 1/n \sum_{i=1}^n \delta_{x_i}$. To obtain our goal we define an appropriate affinely equivariant location functional T at P_n^A for all affine transformations A and also at all measures of the form P_n^{*A}. Here P_n^* is any empirical measure obtained from $\boldsymbol{x}_n$ by altering the values of at most $\lfloor (n-1)/2 \rfloor$ of the x_i. The new sample will be denoted by $\boldsymbol{x}_n^* = (x_1^*, ..., x_n^*)$. We have to show that the values of $T(P_n^{*A})$ can be defined in such a way that

$$T(P_n^A) = A\big(T(P_n)\big) \ , \tag{4.2}$$

$$T(P_n^{*A}) = A\big(T(P_n^*)\big) \tag{4.3}$$

and

$$\sup_{P_n^*} \big|T(P_n) - T(P_n^*)\big| < \infty \ . \tag{4.4}$$

This is done in Appendix A.

We note that the Sample conditions 1 and 2 in Appendix A are satisfied for an i.i.d. Gaussian sample of size n if n is sufficiently large. We indicate how this may be shown in Appendix B.

4.2. Scatter functionals

At the sample (3.1) above the median absolute deviation T_{MAD} has a finite sample breakdown point of 1/11 compared with the upper bound of 3/11 given by Theorem 2.1. We consider a modification of T_{MAD} as defined in Davies and Gather (2005) which attains the upper bound.

For a probability measure P the interval $I(P,\lambda)$ is defined by

$$I(P,\lambda) = \Big[\mathrm{med}(P) - \lambda,\ \mathrm{med}(P) + \lambda\Big] \ .$$

We write

$$\Delta(P,\lambda) = \max\Big\{P(\{x\})\colon\ x \in I(P,\lambda)\Big\} \ .$$

The new scale functional T^*_{MAD} is defined by

$$T^*_{\mathrm{MAD}}(P) = \min\Big\{\lambda\colon\ P\big(I(P,\lambda)\big) \ge \big(1 + \Delta(P,\lambda)\big)/2\Big\} \ .$$

We shall show

$$\text{fsbp}\big(T^*_{\text{MAD}}, \boldsymbol{x}_n, D\big) \;=\; \left\lfloor \frac{n - n\,\Delta(P_n) + 1}{2} \right\rfloor \big/ n\ . \tag{4.5}$$

We consider a replacement sample $\boldsymbol{x}'_n$ with

$$n_1 + n_2 \;=\; m \;<\; \Big\lfloor \big(n - n\,\Delta(P_n) + 1\big)/2 \Big\rfloor$$

points replaced and with empirical distribution P'_n. We show firstly that $T^*_{\text{MAD}}(P'_n)$ does not explode. Let λ' be such that the interval $\big[\text{med}(P'_n) - \lambda', \text{med}(P'_n) + \lambda'\big]$ contains the original sample $\boldsymbol{x}_n$. As the median does not explode we see that λ' remains bounded over all replacement samples. Clearly if $T^*_{\text{MAD}}(P'_n)$ is to explode $\boldsymbol{x}'_n$ must contain points outside of this interval. We denote the number of such points by n_1. We use n_2 points to increase the size of the largest atom of $\boldsymbol{x}'_n$ in the interval. This is clearly done by placing these points at the largest atom of $\boldsymbol{x}_n$. The size of the largest atom of $\boldsymbol{x}'_n$ in the interval is therefore at most $\Delta(P_n) + n_2/n$. It follows that $T^*_{\text{MAD}}(P'_n) \le \lambda'$ if the interval contains at least $\big(n + n\,\Delta(P_n) + n_2\big)/2$ observations. This will be the case if $n - n_1 \ge \big(n + n\,\Delta(P_n) + n_2\big)/2$ which reduces to $n_1 + n_2/2 \le n\big(1 - \Delta(P_n)\big)/2$ which holds as

$$n_1 + n_2/2 \;\le\; n_1 + n_2 \;<\; \Big\lfloor n\big(1 - \Delta(P_n) + 1\big) \Big\rfloor \big/ 2\ .$$

It remains to show that $T^*_{\text{MAD}}(P'_n)$ does not implode to zero. For this to happen we would have to be able to construct a replacement sample for which the interval $I(P', \lambda)$ is arbitrarily small but for which $P'\big(I(P', \lambda)\big) \ge \big(1 + \Delta(P', \lambda)\big)/2$. In order for the interval to be arbitrarily small it must contain either no points of the original sample $\boldsymbol{x}_n$ or just one atom. In the latter case we denote the size of the atom by $\Delta_1(P_n)$. Suppose we replace $n_1 + n_2$ points and that the n_2 points form the largest atom in the interval $I(P', \lambda)$. We see that if $n_2 \ge n\,\Delta_1(P_n)$ then

$$n_1 + n_2 + n\,\Delta_1(P_n) \;\ge\; (n + n_2)/2$$

which implies

$$2n_1 + 2n_2 \;\ge\; 2n_1 + n_2 + n\,\Delta_1(P_n) \;\ge\; n \;>\; n - n\,\Delta(P_n)$$

which contradicts $n_1 + n_2 < \big\lfloor n\big(1 - \Delta(P_n) + 1\big)\big\rfloor/2$. If the n_2 replacement points do not compose the largest atom then this must be of size at least $\Delta_1(P_n)$ which implies

$$n_1 + n_2 + n\,\Delta_1(P_n) \;\ge\; \big(n + n\,\Delta_1(P_n)\big)/2$$

and hence

$$2n_1 + 2n_2 \;\ge\; n - n\,\Delta_1(P_n) \;\ge\; n - n\,\Delta(P_n)$$

which again contradicts $n_1 + n_2 < \big\lfloor n\big(1 - \Delta(P_n) + 1\big)\big\rfloor/2$. We conclude that $T^*_{\text{MAD}}(P'_n)$ cannot implode, and thus (4.5) is shown.

5. THE CHOICE OF THE METRICS d AND D

5.1. The metric d

Considering the parts a) and b) of Theorem 2.1 we note that there is in fact a direct connection between the two results. We consider the total variation metric d_{tv} defined by

$$d_{tv}(P,Q) \;=\; \sup_{B\in\mathcal{B}(\mathcal{X})} \big|P(B)-Q(B)\big| \;.$$

If $\mathcal{B}(\mathcal{X})$ "shatters" every finite set of points in $\mathcal{X}$ then

$$d_{tv}(P_n, P_n^*) \;=\; k/n$$

where P_n denotes the empirical measure deriving from $(x_1,...,x_n)$ and P_n^* that deriving from $(x_1^*,...,x_n^*)$ with the two samples differing in exactly k points. Suppose now that $\epsilon^*(T,P_n,d_{tv},D) = \big(1-\Delta(P_n)\big)/2$. If $k < n\big(1-\Delta(P_n)\big)/2$ then breakdown in the sense of finite sample breakdown point cannot occur and we see that

$$\text{(5.1)} \qquad \operatorname{fsbp}\big(T,\boldsymbol{x}_n,D\big) \;\geq\; \left\lfloor \frac{n-n\,\Delta(P_n)}{2} \right\rfloor \Big/ n \;.$$

Unfortunately the inequality of Theorem 2.1 b) seems not to be provable in the same manner.

We point out that the breakdown point is not necessarily the same for all metrics d. A simple counterexample is provided by the scale problem in $\mathbb{R}$. If we use the Kolmogorov metric then the breakdown point of T_{MAD} at an atomless distribution is 1/4 (Huber, 1981, page 110). However if we use the Kuiper metric d_{ku}^1 defined in (5.3) below then the breakdown point is 1/2 in spite of the fact that both metrics satisfy the conditions of the theorem. More generally if d' and d'' are two metrics satisfying $\sup\limits_{P,Q\in\mathcal{P}} d(P,Q)=1$ and (2.8) and such that $d'\leq d''$ then

$$\text{(5.2)} \qquad \epsilon^*(T,P,d',D) \;\leq\; \epsilon^*(T,P,d'',D) \;\leq\; \big(1-\Delta(P)\big)/2 \;.$$

In particular if $\epsilon^*(T,P,d',D) = \big(1-\Delta(P)\big)/2$ then $\epsilon^*(T,P,d'',D) = \big(1-\Delta(P)\big)/2$. A class of ordered metrics is provided by the generalized Kuiper metrics d_{ku}^m defined by

$$\text{(5.3)} \qquad d_{ku}^m(P,Q) \;=\; \sup\left\{ \left| \sum_{k=1}^m \big(P(I_k)-Q(I_k)\big) \right| :\; I_1,...,I_m \text{ disjoint intervals} \right\} .$$

We have

$$\text{(5.4)} \qquad d_{ku}^1 \leq \,...\, \leq d_{ku}^m \;.$$

For further remarks on the choice of d we refer to Davies and Gather (2005), Rejoinder and for a related but different generalization of the Kuiper metric of use in the context of the modality of densities we refer to Davies and Kovac (2004).

5.2. The metric D

As we have seen in the case of d above there seems to be no canonical choice: different choices of d can lead to different breakdown points. A similar problem exists with respect to the metric D on Θ. In the discussion of Tyler in Davies and Gather (2005) it was also pointed out that it might be difficult to achieve (2.1) when Θ is a compact space. This problem is discussed in the Rejoinder of Davies and Gather (2005), Chapter 6, and solved in Davies and Gather (2006) with applications to directional data.

We now indicate a possibility of making D dependent on d. The idea is that two parameter values θ_1 and θ_2 are far apart with respect to D if and only if the corresponding distributions are far apart with respect to d. We illustrate the idea using the location problem in $\mathbb{R}$. Suppose we have data with empirical distribution P_n and two values of the location parameter θ_1 and θ_2. We transform the data using the translations θ_1 and θ_2 which gives rise to two further distributions $P_n(\cdot - \theta_1)$ and $P_n(\cdot - \theta_2)$. If these two distributions are clearly distinguishable then $d\big(P_n(\cdot - \theta_1), P_n(\cdot - \theta_2)\big)$ will be almost one. An opposed case is provided by an autoregressive process of order one. The parameter space is $\Theta = (-1, 1)$ and this may be metricized in such a manner that $D(\theta_1, \theta_2)$ tends to infinity for fixed θ_1 as θ_2 tends to the boundary. However values of θ close to, on, or even beyond the boundary, may not be empirically distinguishable from values of θ in the parameter space. A sample of size $n = 100$ generated with $\theta_1 = 0.95$ is not easily distinguishable from a series generated with $\theta_2 = 0.9999$ even though $D(\theta_1, \theta_2)$ is large.

We now give a choice of D in terms of d and such that (2.1) is satisfied. We set

$$G(\theta_1, \theta_2) = \Big\{ g \in G\colon\ h_g(\theta_1) = \theta_2 \Big\}$$

and then define D by

$$D(\theta_1, \theta_2) = D_P(\theta_1, \theta_2) = \inf_{g \in G(\theta_1, \theta_2)} \big| \log\big(1 - d(P^g, P)\big) \big| \,. \tag{5.5}$$

The interpretation is that we associate P with the parameter value θ_1 and P^g with the parameter value θ_2. The requirement (2.1) will only hold if $d(P^g, P)$ may be arbitrarily close to one so that the distributions associated with θ_1 and θ_2 are as far apart as possible. It is easily checked that D defines a pseudometric on Θ, which is sufficient for our purposes; namely $D_P \geq 0$, D_P is symmetric and satisfies the triangle inequality. In some situations it seems reasonable to require that d and D be invariant with respect to the groups G and H_G respectively. If d is G-invariant, i.e.

$$d(P,Q) = d(P^g, Q^g), \qquad \text{for all} \quad P, Q \in \mathcal{P}, \quad g \in G \;,$$

then D, defined by (5.5), inherits the invariance, i.e.

$$D(\theta_1, \theta_2) = D\big(h_g(\theta_1), h_g(\theta_2)\big), \qquad \text{for all} \quad \theta_1, \theta_2 \in \Theta, \quad g \in G \;.$$

The G-invariance of d can often be met.

6. FINAL REMARKS

We conclude with a small graphic showing the connections between all ingredients which are necessary for a meaningful breakdown point concept.

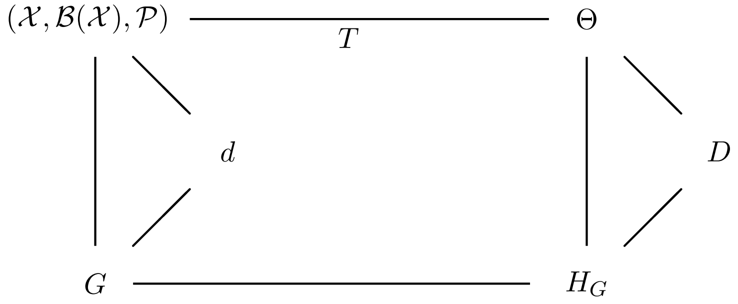

Figure 1: Connections.

We point out that each object in this graphic has an important influence on the breakdown point and its upper bound:

- $\epsilon^*(T, P, d, D)$ depends on P as shown in Chapter 3, and it depends on the metrics d and D as discussed in Chapter 5.
- It is the equivariance structure w.r.t. the group G which allows to prove an upper bound for $\epsilon^*(T, P, d, D)$ and it is the condition $G_1 \neq \emptyset$ which provides the main step in the proof. In particular, the choice of the group G determines $\Delta(P)$, thereby the upper bound, as well as its attainability. For many $\mathcal{P}$, T and G the attainability of the bound remains an open problem.

APPENDIX A

We consider the constraints imposed upon us when defining $T(P_n^*)$. We start with the internal constraints which apply to each P_n^* without reference to the other measures.

- **Case 1:** $P_n^{*A_1} \neq P_n^{*A_2}$ for any two different affine transformations A_1 and A_2. This is seen to reduce to $P_n^{*A} \neq P_n^*$ for any affine transformation A which is not the identity. If this is the case then there are no restrictions on the choice of $T(P_n^*)$. Having chosen it we extend the definition of T to all the measures P_n^{*A} by $T(P_n^{*A}) = A\big(T(P_n^*)\big)$.

- **Case 2:** $P_n^{*A} = P_n^*$ for some affine transformation A which is not the identity. If this is the case then A is unique and there exists a permutation π of $\{1, ..., n\}$ such that $A(x_i) = x_{\pi(i)}$. This implies that for each i we can form cycles
$$\Big(x_i, A(x_i), ..., A^{m_i-1}(x_i)\Big)$$
with $A^{m_i}(x_i) = x_i$. From this we see that for some sufficiently large m $A^m(x_i) = x_i$ for all i. On writing $A(x) = \alpha(x) + a$ we see that if the x_i, $i = 1, ..., n$, span $\mathbb{R}^k$ then $\alpha^m = I$ where I denotes the identity transformation on $\mathbb{R}^k$. This implies that α must be an orthogonal transformation and that
$$\sum_{j=0}^{m-1} \alpha^j(a) = 0 \ . \tag{A.1}$$
It follows that if we set $T(P_n^*) = \mu$, we must have $A(\mu) = \mu$ for any affine transformation for which $P_n^{*A} = P_n^*$. The choice of μ is arbitrary subject only to these constraints. Having chosen such a μ the values of $T(P_n^{*B})$ are defined to be $B(\mu)$ for all other affine transformations B.

The above argument shows the internal consistency relationships which must be placed on T so that $T(P_n^{*A}) = A\big(T(P_n)\big)$ for any affine transformation A. We now consider what one may call the external restrictions.

- **Case 3:** Suppose that P_n^* is such that there does not exist a $P_n'^*$ and an affine transformation A such that $P_n^{*A} = P_n'^*$. In this case the choice of $T(P_n^*)$ is only restricted by the considerations of Case 2 above if that case applies and otherwise not at all.

- **Case 4:** Suppose that P_n^* is such that there exists a $P_n'^*$ and an affine transformation A such that $P_n^* = P_n'^{*A}$. In this case we require $T(P_n^*) = A\big(T(P_n'^*)\big)$.

We now place the following conditions on the sample $\boldsymbol{x}_n$:

Sample condition 1: There do not exist two distinct subsets of $\boldsymbol{x}_n$ each of size at least $k+2$ and an affine transformation A which transforms one subset into the other.

Sample condition 2: If

$$\big|A(\boldsymbol{x}_n)\cap B(\boldsymbol{x}_n)\big| \;\geq\; \lfloor (n+1)/2\rfloor - 2k$$

for two affine transformations A and B then $A = B$.

Sample condition 3: $k < \lfloor (n-1)/2\rfloor$.

We now construct a functional T which satisfies (4.2), (4.3) and (4.4). If the sample conditions hold then for any affine transformation $A \neq I$ we have $P_n^A \neq P_n^*$ where P_n^* derives from a subset $\boldsymbol{x}_n^*$ which differs from $\boldsymbol{x}_n$ by at least one and at most $\lfloor (n-1)/2\rfloor$ points. This follows on noting that at most $k+1$ of the $A(x_i)$ belong to $\boldsymbol{x}_n$ by Sample condition 1. Because of this we can define $T(P_n)$ without reference to the values of $T(P_n^*)$. We set

$$T(P_n) \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i \;.$$

If P_n^* satisfies the conditions of Case 3 above we set

$$T(P_n^*) \;=\; \frac{1}{n^*}\sum_{i=1}^{n^*} x_{\pi(i)}$$

where the $x_{\pi(i)}$ are those $n^* \geq \lceil (n+1)/2\rceil$ points of the sample $\boldsymbol{x}_n$ which also belong to the sample $\boldsymbol{x}_n^*$. Finally we consider Case 4 above. We show that the sample assumptions and the condition $P_n^* = P_n^{\prime *A}$ uniquely determine the affine transformation A. To see this we suppose that there exists a second affine transformation B and a distribution $P_n^{\prime\prime *}$ such that $P_n^* = P_n^{\prime\prime *B}$. Let $x_{\pi(1)}^*, ..., x_{\pi(N')}^*$ denote those points of $\boldsymbol{x}_n^*$ not contained in the sample $\boldsymbol{x}_n$. Because of Sample condition 1 this set contains at least $\lceil (n+1)/2\rceil - k - 2$ points of the form $A(x_i)$. Similarly it also contains at least $\lceil (n+1)/2\rceil - k - 2$ points of the form $B(x_i)$. The intersection of these two sets is of size at least $\lfloor (n+1)/2\rfloor - 2k$ and we may conclude from Sample condition 2 that $A = B$. The representation is therefore unique. Let $x_{\pi(1)}, ..., x_{\pi(m)}$ be those points of $\boldsymbol{x}_n$ which belong to the sample $\boldsymbol{x}_n^{\prime *}$ and for which $A(x_{\pi(1)}), ..., A(x_{\pi(m)})$ belong to the sample $\boldsymbol{x}_n$. It is clear that $m \geq 1$. We define

$$T(P_n^{\prime *}) \;=\; \frac{1}{m}\sum_{i=1}^{m} x_{\pi(i)}$$

and by equivariance

$$T(P_n^*) \;=\; \frac{1}{m}\sum_{i=1}^{m} \mathcal{A}\big(x_{\pi(i)}\big) \;.$$

It follows that $T(P_n^*)$ is well defined and in both cases the sums involved come from the sample $\boldsymbol{x}_n$. The functional T is extended to all $P_n{*}^B$ and $P_n'{*}^B$ by affine equivariance. In all cases the definition of $T(P_n^*)$ is as the mean of a subset of $\boldsymbol{x}_n$. From this it is clear that (4.4) is satisfied.

APPENDIX B

We now show that Sample conditions 1 and 2 hold for independent random samples $X_1,...,X_n$ with probability one. Let $\mathcal{A}=A+a$ and $\mathcal{B}=B+b$ with A and B nonsingular matrices and a and b points in $\mathbb{R}^k$. We suppose that $A\neq B$. On taking differences we see that there exist variables $X_{i_1},...,X_{i_{k+1}}$ and $X_{j_1},...,X_{j_{k+1}}$ such that

$$A(X_{i_l}-X_{i_{k+1}})=B(X_{j_l}-X_{j_{k+1}})\ , \qquad j=1,...,k\ .$$

This implies that $B^{-1}A$ and $B^{-1}(b-a)$ are functions of the chosen sample points

$$\text{(B.1)}\qquad \begin{aligned} B^{-1}A \;&=\; C\bigl(X_{i_1},...,X_{i_{k+1}},\,X_{j_1},...,X_{j_{k+1}}\bigr)\ ,\\ B^{-1}(b-a) \;&=\; c\bigl(X_{i_1},...,X_{i_{k+1}},\,X_{j_1},...,X_{j_{k+1}}\bigr)\ . \end{aligned}$$

For n sufficiently large there exist four further sample points X_i, $i=1,...,4$ which are not contained in $\bigl\{X_{i_1},...,X_{i_{k+1}},\,X_{j_1},...,X_{j_{k+1}}\bigr\}$ and for which

$$A(X_1)+a \;=\; B(X_2)+b\ , \qquad A(X_3)+a \;=\; B(X_4)+b\ .$$

This implies

$$\text{(B.2)}\qquad B^{-1}A(X_3-X_1) \;=\; X_4-X_2\ .$$

However as the X_i, $i=1,...,4$, are independent of $X_{i_1},...,X_{i_{k+1}},\,X_{j_1},...,X_{j_{k+1}}$ it follows from (B.1) that (B.2) holds with probability zero. From this we conclude that $A=B$. Similarly we can show that $a=b$ and hence $\mathcal{A}=\mathcal{B}$.

ACKNOWLEDGMENTS

Research supported in part by Sonderforschungsbereich 475, University of Dortmund.

REFERENCES

[1] Croux, C. and Rousseeuw, P.J. (1992). A class of high-breakdown estimators based on subranges, *Commun. Statist. – Theory Meth.*, **21**, 1935–1951.

[2] Davies, P.L. (1993). Aspects of robust linear regression, *Ann. Statist.*, **21**, 1843–1899.

[3] Davies, P.L. and Gather, U. (1993). The identification of multiple outliers (with discussion and rejoinder), *J. Amer. Statist. Assoc.*, **88**, 782–801.

[4] Davies, P.L. and Gather, U. (2005). Breakdown and groups (with discussion and rejoinder), *Ann. Statist.*, **33**, 977–1035.

[5] Davies, P.L. and Gather, U. (2006). Addendum to the discussion of breakdown and groups, *Ann. Statist.*, **34**, 1577–1579.

[6] Davies, P.L. and Kovac, A. (2004). Densities, spectral densities and modality, *Ann. Statist.*, **32**, 1093–1136.

[7] Donoho, D.L. (1982). *Breakdown properties of multivariate location estimators*, Ph. D. qualifying paper, Dept. Statistics, Harvard University.

[8] Donoho, D.L. and Gasko, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness, *Ann. Statist.*, **20**, 1803–1827.

[9] Donoho, D.L. and Huber, P.J. (1983). *The notion of breakdown point.* In "A Festschrift for Erich Lehmann" (P.J. Bickel, K. Doksum and J.L. Hodges, Jr., Eds.), Wadsworth, Belmont, CA, 157–184.

[10] Ellis, S.P. and Morgenthaler, S. (1992). Leverage and breakdown in L_1 regression, *J. Amer. Statist. Assoc.*, **87**, 143–148.

[11] Gordaliza, A. (1991). On the breakdown point of multivariate location estimators based on trimming procedures, *Statist. Probab. Letters*, **11**, 387–394.

[12] Hampel, F.R. (1968). *Contributions to the theory of robust estimation*, Ph. D. thesis, Dept. Statistics, Univ. California, Berkeley.

[13] Hampel, F.R. (1971). A general qualitative definition of robustness, *Ann. Math. Statist.*, **42**, 1887–1896.

[14] Hampel, F.R. (1975). Beyond location parameters: robust concepts and methods (with discussion), *Proceedings of the 40th Session of the ISI*, Vol. 46, Book 1, 375–391.

[15] He, X. and Fung, W.K. (2000). High breakdown estimation for multiple populations with applications to discriminant analysis, *J. Multivariate Anal.*, **72**, 151–162.

[16] Huber, P.J. (1981). *Robust Statistics*, Wiley, New York.

[17] Hubert, M. (1997). The breakdown value of the L_1 estimator in contingency tables, *Statist. Probab. Letters*, **33**, 419–425.

[18] Lopuhaä, H.P. (1992). Highly efficient estimators of multivariate location with high breakdown point, *Ann. Statist.*, **20**, 398–413.

[19] Lopuhaä, H.P. and Rousseeuw, P.J. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices, *Ann. Statist.*, **19**, 229–248.

[20] Müller, C.H. and Uhlig, S. (2001). Estimation of variance components with high breakdown point and high efficiency, *Biometrika*, **88**, 353–366.

[21] Niinimaa, A.; Oja, H. and Tableman, M. (1990). The finite-sample breakdown point of the Oja bivariate median and of the corresponding half-samples version, *Statist. Probab. Letters*, **10**, 325–328.

[22] Rousseeuw, P.J. (1984). Least median of squares regression, *Journal of the American Statistical Association*, **79**, 871–880.

[23] Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection*, Wiley, New York.

[24] Terbeck, W. and Davies, P.L. (1998). Interactions and outliers in the two-way analysis of variance, *Ann. Statist.*, **26**, 1279–1305.

REVSTAT – Statistical Journal
Volume 5, Number 1, March 2007, 19–39

SOME THOUGHTS ABOUT THE DESIGN OF LOSS FUNCTIONS

Authors: Christian Hennig
– Department of Statistical Science,
University College London,
United Kingdom
chrish@stats.ucl.ac.uk

Mahmut Kutlukaya
– Strategy Development Department,
Banking Regulation and Supervision Agency,
Kavaklidere-Ankara, Turkey
mkutlukaya@bddk.org.tr

Abstract:

- The choice and design of loss functions is discussed. Particularly when computational methods like cross-validation are applied, there is no need to stick to "standard" loss functions such as the L_2-loss (squared loss). Our main message is that the choice of a loss function in a practical situation is the translation of an informal aim or interest that a researcher may have into the formal language of mathematics. The choice of a loss function cannot be formalized as a solution of a mathematical decision problem in itself. An illustrative case study about the location of branches of a chain of restaurants is given. Statistical aspects of loss functions are treated, such as the distinction between applications of loss functions to prediction and estimation problems and the direct definition of estimators to minimize loss functions. The impact of subjective decisions to the design of loss functions is also emphasized and discussed.

Key-Words:

- *prediction; estimation; decision theory; M-estimator; MM-estimator; linear regression.*

AMS Subject Classification:

- 62A01, 62C05, 62G09, 62J05, 62M20.

1. INTRODUCTION

Most statistical problems are defined in terms of loss functions in the sense that loss functions define what a "good" estimator or a "good" prediction is. This paper discusses some aspects of the choice of a loss function. The main message of the paper is that the task of choosing a loss function is about the translation of an informal aim or interest that a researcher may have in the given application into the formal language of mathematics. The choice of a loss function cannot be formalized as a solution of a mathematical decision problem in itself, because such a decision problem would require the specification of another loss function. Therefore, the choice of a loss function requires informal decisions, which necessarily have to be subjective, or at least contain subjective elements. This seems to be acknowledged somewhat implicitly in the decision theoretic literature, but we are not aware of any sources where this is discussed in detail.

Several different uses of loss functions can be distinguished.

(**a**) In *prediction problems*, a loss function depending on predicted and observed value defines the quality of a prediction.

(**b**) In *estimation problems*, a loss function depending on the true parameter and the estimated value defines the quality of estimation. As opposed to prediction problems, this assumes a statistical model to hold, which defines the parameter to be estimated. The true parameter value in an estimation problem is generally unobservable, while in a prediction problem the "truth" is observable in the future.

(**c**) *Definition of estimators*: many estimators (such as least squares or M-estimators) are defined as optimizers of certain loss functions which then depend on the data and the estimated value. Note that this is essentially different from (a) and (b) in the sense that the least squares estimator is not necessarily the estimator minimizing the mean squared estimation error or the squared prediction error.

There are several further uses of loss functions, which won't be treated in the present paper, for instance defining optimal testing procedures, Bayesian risk, etc.

While general loss functions have been treated in the literature[1], versions of the squared loss function are used in a vast majority of applications of prediction and estimation problems (note that UMVU estimation is a restricted optimization of a squared loss function). Main reasons for this seem to be the simplicity of the mathematics of squared loss and the self-confirming nature of the frequent

[1] See, for instance, Lehmann and Casella ([6]), who mainly use squared loss, but discuss alternatives in several chapters.

use of certain "standard" methods in science. However, if prediction methods are compared using nonparametric resampling techniques such as cross-validation and bootstrap, there is no computational reason to stick to the squared loss, and other loss functions can be used. Robustness aspects of loss functions have been discussed previously by Ronchetti, Field and Blanchard ([9]) and Leung ([7]).

In Section 2, the subject-matter dependent design of a loss function in a business application using robust regression is discussed to give an illustrating example of the "translation problem" mentioned above and to motivate some of the discussion in the following sections.

In Section 3, the implications of the different statistical uses of loss functions (a), (b) and (c) above are explored in more detail. The question whether the negative loglikelihood can be considered as the "true" objective loss function in estimation is discussed.

In Section 4, some philosophical aspects are treated. In particular, the concepts of subjectivity and objectivity, emphasizing the role of subjective decisions in the choice of loss functions, and the standardizing role of communication in the scientific community are discussed. Finally, a brief conclusion is given.

2. LOCATIONS OF RESTAURANTS: A CASE STUDY

The case study presented in this section is about a prediction problem in a business application. Because the original study is confidential, the story presented here is made up, and the original data are not shown. The values and rankings in Tables 1 and 2, however, are authentic (absolute and squared losses have been multiplied by a constant).

A restaurant chain wanted to predict the turnover for new branches, depending on the following six independent variables:

- number of people living in a (suitably defined) neighborhood,
- number of people working or shopping at daytime in the neighborhood,
- number of branches of competitors in the neighborhood,
- size of the branch,
- a wealth indicator of the neighborhood,
- distance to the next branch of the same chain.

The results are to be used to support decisions such as where to open new branches, and what amount of rents or building prices can be accepted for particular locations. Data from 154 already existing branches on all the variables were available. In our study we confined ourselves to finding a good linear regression type prediction rule, partly because the company wanted to have a simple formula, and partly because an alternative (regression trees) had already been explored in a former project.

The data are neither apparently nonlinear, nor heteroscedastic in any clear systematic way. However, there are obvious outliers. We decided to choose the best out of several more or less robust linear regression estimators using leave-one-out cross-validation (LOO-CV). In the real study, choice of transformations of variables and variable selection have also been considered, but this doesn't add to the discussion of interest here.

Note that LOO-CV processes all data points in the same manner, which means that all observations are treated as if they were a representative sample from the underlying population of interest. Particularly, outliers are treated in the same way as seemingly more typical data points (but may be weighted down implicitly, see below). This makes sense if there is no further subject matter information indicating that the outliers are erroneous or atypical in a way that we would not expect similar observations anymore in the future. In the given case study, outlying observations are not erroneous and stem from restaurants at some locations with special features. It may well be possible that further outliers occur in the future for similar reasons.

The estimators we took into account were

- the least squares (LS)-estimator,
- the least median of squares (LMS)-estimator as suggested by Rousseeuw ([10]),
- Huber's M-estimator for linear regression with tuning constant $k=1.345$ to produce 95% efficiency for normal samples, see Huber ([5]),
- an M-estimator for linear regression using the "bisquared" objective function with tuning constant $k = 4.685$ to produce 95% efficiency for normal samples, see Western ([13]),
- the MM-estimator suggested by Yohai ([14]) tuned to 95% efficiency for normal samples.

In principle, it is reasonable to include M-/MM-estimators tuned to smaller efficiency as well, which will then potentially downweight some further outliers. However, we compared several tunings of the MM-estimator in one particular situation, from which we concluded that not too much gain is to be expected from smaller tunings than 95% efficiency (larger efficiencies can be better, but our results on this are quite unstable).

All estimators were used as implemented in R (`www.R-project.org`), but the implementations we used for this project have been replaced by new ones in the meantime (in packages “MASS” and “robustbase”). The scale estimator used for the two M-estimators was a re-scaled median absolute deviance (MAD) based on the residuals (as implemented in the function `rlm`).

The estimators have been compared according to the estimated expected prediction error

$$\frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{y}_{-i}) \;, \tag{2.1}$$

where n is the number of observations, $y_1, ..., y_n$ are the observed turnovers, and $\hat{y}_{-i}$ is the predicted value of the turnover for y_i from applying the linear regression method to the data omitting the i^{th} observation. L is a loss function, of which the design will be discussed in the following.

Note that (2.1) already implies some decisions. Firstly, L is defined here to depend on y_i and y_{-i} only, but not directly on the values of the independent variables of the i^{th} observation. In general, this restriction is not required, but it is justified in the present setup by the fact that the company didn’t specify any particular dependence of their tolerance of prediction errors on the values of the independent variables, and there is no obvious subject-matter reason in the present study for such a dependence to be needed. This is a first illustration of our major principle to translate the informal interests and aims of those who use the results in the formal mathematical language.

Secondly, it is part of the design of the loss function not just to choose L, but also to decide about how the values of $L(y_i, \hat{y}_{-i})$ should be aggregated. Their mean is used in (2.1), but instead, their maximum, their median, another quantile or a trimmed mean could be chosen as well. Note that there is some interaction between the choice of L and the choice of how the values of L are to be aggregated. For example, under the assumption that we would like to do something robust against outliers, the choice of a bounded L-function bounds the influence of extreme prediction errors in itself and allows therefore the aggregation of the L-values in a less robust manner such as taking their mean. For the present study, we confine ourselves to the mean, of which the interpretation is that the prediction error of every single observation is judged as equally important to us, and we will deal with the influence of extreme observations via the choice of L.

As mentioned before, the “standard” loss function for this kind of problem is defined by $L_2(y, \hat{y}) = (y - \hat{y})^2$, but because we use LOO-CV, there is no mathematical reason to use $L = L_2$.

One of the decisions to make is whether L should be symmetric. This means that a negative prediction error is judged as causing the same loss as a positive error of the same absolute value. This is difficult to judge in the present situation.

It could be argued that it is not as bad for the company to underestimate the turnover at a particular location than to overestimate it, because the money spent by the company on a branch with overestimated turnover may be lost.

However, because the prediction should guide the decision whether a branch should be opened in the first place, how much rent should be paid and also how the branch will be initially equipped, underestimation of the turnover may have serious consequences as well, as offers for good locations may be turned down or under-equipped. Though the effects of over- and underestimation can be considered to be asymmetric in the present setup, we decided to stick to symmetric loss functions, meaning that the loss of paid money is treated as equally bad as the loss of money which is not earned because of a missed opportunity.

A main feature of L_2 is its convexity, which means that the differences between high prediction errors are assessed as more important than differences between small prediction errors. As an example, consider two prediction rules that only differ with respect to their cross-validated predictions of two data points, y_1 and y_2. Suppose that for rule 1, $y_1 - \hat{y}_{-1} = 10{,}000$, $y_2 - \hat{y}_{-2} = -10$, and for rule 2, $y_1 - \hat{y}_{-1} = 9{,}990$, $y_2 - \hat{y}_{-2} = -20$ (the units of y don't have a particular meaning here because we have to use artificial values anyway, but you may imagine them to mean £ 1,000 a year). L_2 favours rule 2 in this situation. But is this adequate?

Going back to the discussion above, if the values could be interpreted as earned (or lost) money, the L_1-loss ($L_1(y, \bar{y}) = |y - \bar{y}|$) seemed to be more adequate, because it assesses both rules as equally good, based on the fact that they both cause the same direct or indirect financial loss of 10,010 units. For the restaurant case, switching from L_2 to L_1-loss makes a big difference in terms of the quality ranking of the methods, as can be seen in Table 1.

Table 1: Ranking of regression methods and loss function values (multiplied by two different constants, for L_1 and L_2); the higher the rank, the better the result in terms of (2.1), using $L = L_2$ and $L = L_1$, evaluated on the restaurant data.

L_2		L_1	
1. M-Huber	3203	1. MM	2205
2. LS	3235	2. M-Huber	2219
3. MM	3524	3. M-Bisquare	2247
4. M-Bisquare	3651	4. LS	2274
5. LMS	3692	5. LMS	2278

However, the situation is more complex. Firstly, the data made available to us are about turnover and not about profit (a reason for this may be that for the accurate prediction of profits factors carry a higher weight that rather have to do with management decisions than with the location of the branch). Usually, profits are less sensitive against differences between two large values of turnovers than against the same absolute differences between two smaller values of turnovers. Therefore, more tolerance is allowed in the prediction of larger y_i-values.

Secondly, the data give turnovers over a long period (three years, say), and after a new branch has been opened, if it turns out after some months that the turnover has been hugely wrongly predicted, the management has several possibilities of reaction, ranging from hiring or firing staff over special offers and campaigns attracting more customers to closing the branch.

Therefore, if predictions are hugely wrong, it matters *that* they are hugely wrong, but it doesn't matter too much *how* wrong they exactly are. This means that, at least for large absolute errors, the loss function should be concave if not constant. Actually we chose a function which is constant for large absolute errors, because we could give the lowest absolute error above which the loss function is constant a simple interpretation: above this error value, predictions are treated as "essentially useless" and it doesn't matter how wrong they precisely are. This interpretation could be communicated to the company, and the company was then able to specify this limiting value. The design of a concave but strictly increasing function would have involved much more complicated communication.

The company initially specified the critical value for "usefulness" as 10% of the true turnover, i.e., they were concerned about relative rather than absolute error, which motivated the following loss function:

$$L_c(y,\hat{y}) = \begin{cases} \frac{(y-\hat{y})^2}{y^2} & : \quad \frac{(y-\hat{y})^2}{y^2} \le c^2 \\ c^2 & : \quad \frac{(y-\hat{y})^2}{y^2} > c^2 , \end{cases}$$

$c = 0.1$. Below the cutoff value c, we have used a squared function of the relative error. Two intuitive alternatives would be to choose the L_1-norm of the relative error below c or a concave function, possibly the square root, see Figure 1. Of course, an infinite number of other convex or concave functions could be chosen, but for pragmatic reasons it is necessary to discuss just a small number of possible choices, between which the differences can be given a clear interpretation.

The interpretation of L_1 here is again that all differences between relative errors are treated as equally important, be they between relatively large or relatively small errors. The concave function considers differences between small errors as more important. To optimize this function, it would be advantageous to predict some (maybe very few) observations very well, while the precise relative error values for all observations causing a bit larger prediction don't matter too much. Optimizing the convex square function, on the other hand, means to try as much as possible observations to achieve a relative prediction error below c, while differences between small errors don't have a large influence. Because the company is interested in useful information about many branches, rather than to predict few branches very precisely, we chose the squared function below c.

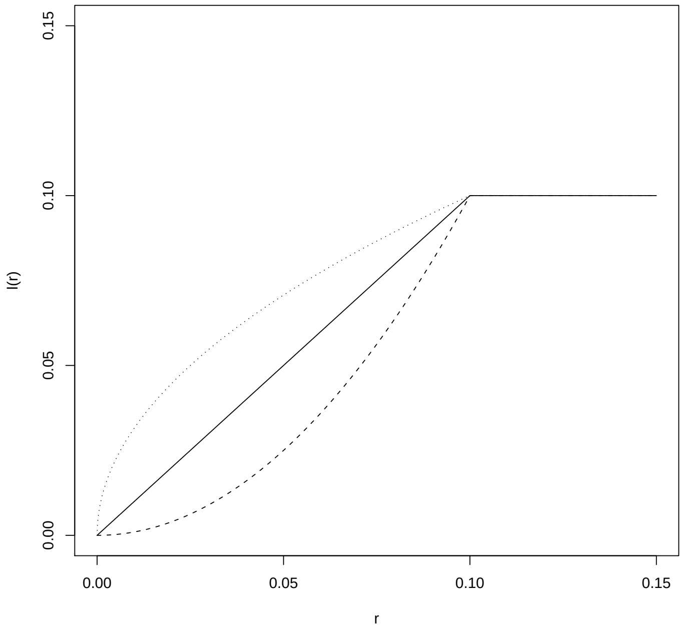

Figure 1: Bounded functions of the relative prediction error r, the lower part being squared, L_1 and square root.

Unfortunately, when we carried out the comparison, it turned out that the company had been quite optimistic about the possible quality of prediction. Table 2 (left side) shows the ranking of the estimators, but also the number of observations of which the relative prediction error has been smaller than c, i.e., for which the prediction has not been classified as "essentially useless".

Table 2: Ranking and loss function values of regression methods in terms of (2.1), using $L = L_c$ with $c = 0.1$ and $c = 0.2$. The number of observations of which the prediction has not been classified as "essentially useless" is also given.

Ranking $c = 0.1$	# obs. $\frac{(y-\hat{y})^2}{y^2} \le 0.1^2$	$10^7 *$ $L_{0.1}$	Ranking $c = 0.2$	# obs. $\frac{(y-\hat{y})^2}{y^2} \le 0.2^2$	$10^6 *$ $L_{0.2}$
1. M-Huber	42	8117	1. MM	85	2474
2. M-Bisquare	49	8184	2. M-Bisquare	86	2482
3. LS	38	8184	3. M-Huber	83	2494
4. MM	49	8195	4. LMS	75	2593
5. LMS	39	8373	5. LS	81	2602

With $n = 154$, this is less than a third of the observations for all methods. Confronted with this, the company decided to allow relative prediction errors up to 20% to be called "useful", which at least made it possible to obtain reasonable predictions for more than half of the observations. The company accepted this result (which can be seen on the right side of Table 2) though we believe that accepting even larger relative errors for more branches as "useful" would be reasonable, given the precision of the data at hand. One could also think about using a squared function of the relative error below $c = 0.2$, constant loss above $c = 0.4$ and something concave in between, which, however, would have been difficult to negotiate with the company. The question whether it would be advantageous to use an estimator that directly minimizes $\sum L(y, \hat{y})$, given a loss function L, instead of comparing other estimators in terms of L is treated in Section 3.1.

The considered loss functions lead to quite different rankings of methods. Figure 2 gives an illustration how the choice of the loss function affects the optimality of the estimator. It shows artificially generated heterogeneous data, coming from four different groups, all generated by normal errors along some regression line. The groups are indicated by four different symbols: circles (150 points), pluses (30 points), crosses (30 points) and triangles (3 points).

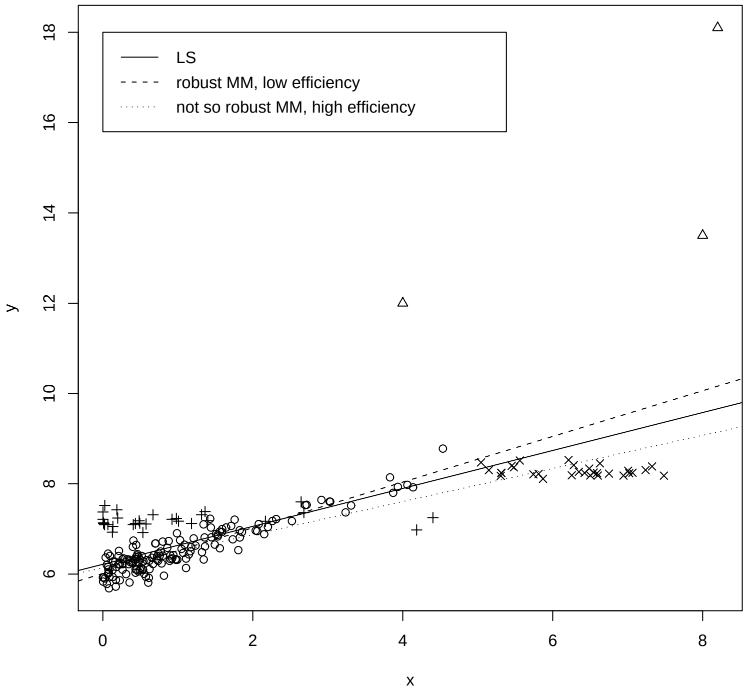

Figure 2: Artificial heterogeneous data with fits of three different regression estimators, giving full weight to all data (LS), only the majority group (circles; low efficiency MM) and about 80% of the data (high efficiency MM).

The plot has a rough similarity with some of the scatterplots from the original restaurants data. If the aim is to fit some points very well, and the loss function is chosen accordingly, the most robust "low efficiency MM-estimator" in Figure 2 is the method of choice, which does the best job for the majority of the data. A squared loss function would emphasize to make the prediction errors for the outlying points (triangles) as small as possible, which would presumably favour the LS-estimator here (this is not always the case, see Section 3). However, if the aim is to yield a good relative prediction error for more data than fitted well by the robust estimator, the less robust, but more efficient MM-estimator (or an estimator with breakdown point of, say, 75%) leads to a fit that does a reasonable job for circles, crosses, and some of the pluses. The decision about the best approach here is depending on the application. For instance, an insurance company may be interested particularly in large outliers and will choose a different loss function from a company which considers large prediction errors as "essentially useless". But even such a company may not be satisfied by getting only a tight majority of the points about right.

3. STATISTICAL ASPECTS

Though Section 2 was about prediction, methods have been compared that were originally introduced as parameter estimators for certain models, and that are defined via optimizing some objective (loss) functions. Therefore the applications (a), (b) and (c) of loss functions mentioned in the introduction were involved. Here are some remarks about differences and relations between these uses.

3.1. Prediction loss vs. objective functions defining estimators

First of all, the estimator defined by minimizing $\sum L(y, \hat{y})$ is not always the best predictor in terms of $\sum L(y, \hat{y}_{-i})$. Consider the situation in Figure 3, given that $L = L_2$, the squared loss function. Compare the LS-estimator with a robust estimator giving zero weight to the outlier at $(1.5, -2)$, the LMS-estimator, say, using LOO-CV. Whenever a non-outlier is deleted, the LMS-estimator computed from the remaining points will give an almost perfect fit, while the LS-estimator will be strongly influenced by the outlier. This means that the LMS estimator will be much better in terms of $L_2(y, \hat{y}_{-i})$. If the outlier is left out, LMS- and LS-estimator will get about the same line, which gives a bad prediction for the outlier. Adding the loss values up, the LMS-estimator gives a much smaller estimated L_2-prediction error. This is not mainly due to the use of LOO-CV, but will happen with any resampling scheme which is based on the prediction of a subsample of points by use of the remaining points. The situation changes (for LOO-CV) when further outliers are added at about $(-1.5, 2)$. In this case, the LS-estimator is better in terms of the estimated L_2-prediction error, because this is dominated by the outliers, and if one outlier is left out, the further outliers at about the same place enable LS to do a better job on these than the robust estimator. The situation is again different when outliers are added at other locations in a way that none of the outliers provides useful information to predict the others. In this situation, it depends strongly on where exactly the outliers are whether LOO-CV prefers LS or LMS. Here, the assessment of the prediction error itself is non-robust and quite sensitive to small changes in the data.

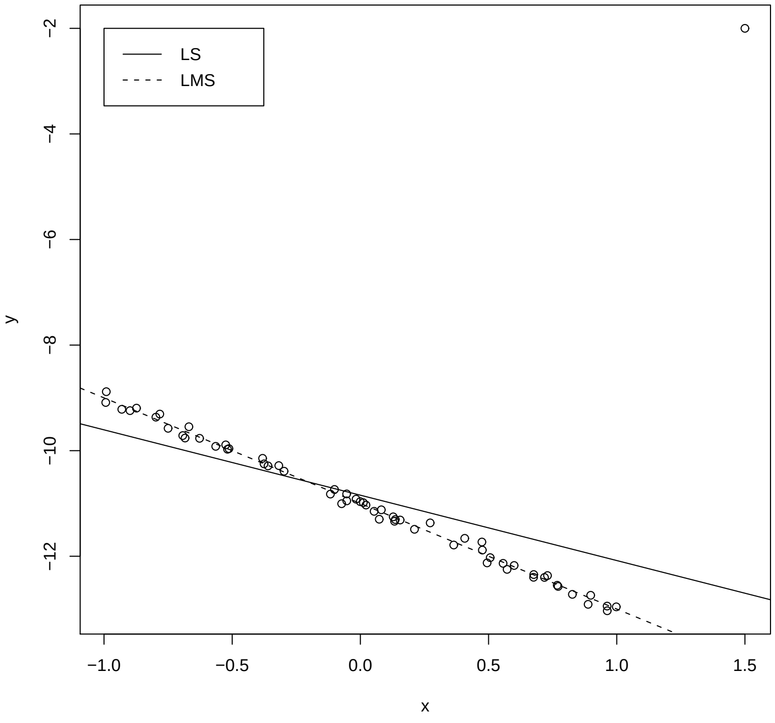

Figure 3: Artificial data with fits of LS and LMS estimator.

From a theoretical point of view, apart from the particular use of LOO-CV to estimate the prediction error, LS is clearly better than LMS in terms of L_2-prediction loss, in a "normal model plus outliers" situation, if the outliers make it possible to find a suitable compromise between fitting them and the majority, while it is bad for LS if the outliers are scattered all over the place and one outlier doesn't give useful information about the prediction of the others (as for example in a linear model with Cauchy random term). Whether the L_2-loss is reasonable or the LMS-fit should be preferred because it predicts the "good" majority of the data better even in cases where the outliers can be used to predict each other depends on subject-matter decisions.

Asymptotically, using empirical process theory, it is often possible to show that the estimator defined by minimizing $\sum L(y, \hat{y})$ is consistent for θ minimizing $EL(y, \theta)$ (in such situations, optimal prediction optimizing L and estimation of θ are equivalent). Therefore, for a given loss function, it makes at least some sense to use the estimator defined by the same objective function. However, this is often not optimal, not even asymptotically, as will be shown in the next section.

3.2. Prediction and maximum likelihood-estimation

Suppose that the data have been generated by some parametric model. Then there are two different approaches to prediction:

1. find a good prediction method directly, or

2. estimate the true model first, as well as possible, solve the prediction problem theoretically on the model and then plug in the estimated parameter into the theoretical prediction rule.

As an example, consider i.i.d. samples from an exponential(λ)-distribution, and consider prediction optimizing L_1-loss. The sample median suggests itself as a prediction rule, minimizing $\sum L_1(y - \hat{y})$. The theoretical median (and therefore the asymptotically optimal prediction rule) of the exponential(λ)-distribution is $\log 2/\lambda$, and this can be estimated by maximum likelihood as $\log 2/\bar{X}_n$, $\bar{X}_n$ being the arithmetic mean. We have simulated 10,000 samples with $n = 20$ observations from an exponential(1)-distribution. The MSE of the sample median has been 0.566 and the MSE of the ML-median has been 0.559. This doesn't seem to be a big difference, but using the paired Mann-Whitney test (not assuming a particular loss function), the advantage of the ML-median is highly significant with $p < 10^{-5}$, and the ML-median was better than the sample median in 6,098 out of 10,000 simulations.

Therefore, in this situation, it is advantageous to estimate the underlying model first, and to derive predictions from the estimator. There is an asymptotic justification for this, called the "convolution theorem" (see, e.g., Bickel et al, [1], p. 24). A corollary of it says that under several assumptions

$$\liminf_{n\to\infty} E_\theta L\Big(\sqrt{n}\big(T_n - q(\theta)\big)\Big) \geq E_\theta L\Big(M_n - q(\theta)\Big) , \tag{3.1}$$

where $q(\theta)$ is the parameter to be estimated (which determines the asymptotically optimal prediction rule), T_n is an estimator and M_n is the ML-estimator. This holds for every loss function L which is a function of the difference between estimated and true parameter satisfying

$$L(x) = L(-x) , \qquad \Big\{x\colon L(x) \leq c\Big\} \text{ convex } \forall\, c > 0 . \tag{3.2}$$

(3.2) is somewhat restrictive, but not strongly so. For example, it includes all loss functions discussed in Section 2 (applied to the estimation problem of the optimal prediction rule instead of direct prediction, however).

This fact may provoke three misinterpretations:

1. estimation is essentially equivalent to prediction (at least asymptotically — though the exponential example shows that the implications may already hold for small n),

2. the negative loglikelihood can be seen as the "true" loss function belonging to a particular model. In this sense the choice of the loss function would rather be guided by knowledge about the underlying truth than by subjective subject-matter dependent decisions as illustrated in Section 2,

3. all loss functions fulfilling (3.2) are asymptotically equivalent.

Our view is different.

1. The main assumption behind the convolution theorem is that we know the true parametric model, which is obviously not true in practice. While the ML-median performed better in our simulation, prediction by $\log 2/\bar{X}_n$ can be quite bad in terms of L_1-loss if the true distribution is not the exponential. The sample median can be expected to perform well over a wide range of distributions (which can be backed up by asymptotic theory, see above), and other prediction rules can turn out to be even better in some situations using LOO-CV and the like, for which we don't need any parametric assumption.

The basic difference between prediction and estimation is that the truth is observable in prediction problems, while it is not in estimation problems. In reality, it can not even be assumed that any probability model involving an i.i.d. component holds. In such a case, estimation problems are not well defined, while prediction problems are, and there are prediction methods that are not based on any such model. Such methods can be assessed by resampling methods as well (though LOO-CV admittedly makes the implicit assumption that the data are exchangeable).

Apart from this, there are parametric situations, in which the assumptions of the convolution theorem are not satisfied and optimal estimation and optimal prediction are even asymptotically different. For example, in many model selection problems, the BIC estimates the order of a model consistently, as opposed to the AIC (Nishii [8]). But often, the AIC can be proved to be asymptotically better for prediction, because for this task underestimation of the model order matters more than overestimation (Shibata [11], [12]).

2. The idea that the negative loglikelihood can be seen as the "true" loss function belonging to a particular model (with which we have been confronted in private communication) is a confusion of the different applications of loss functions. The negative loglikelihood *defines* the ML estimator, which is, according to the convolution theorem, asymptotically optimal with respect to several loss functions *specifying an estimation problem.* These loss functions are assumed to be symmetric. In some applications asymmetric loss functions may be justified, for which different estimators may be optimal (for example shrinked or inflated ML-estimators; this would be the case in Section 2 if the company had a rather conservative attitude, were less keen on risking money by opening new branches and would rather miss opportunities as long as they are not obviously excellent). This may particularly hold under asymmetric distributions, for which not even the negative loglikelihood itself is symmetric. (The idea of basing the loss function on the underlying distribution, however, could make some sense, see Section 3.4.)

In the above mentioned simulation with the exponential distribution, LOO-CV with the L_1-loss function decided in 6,617 out of 10,000 cases that the ML-median is a better predictor than the sample median. This shows that in a situation where the negative loglikelihood is a good loss function to *define* a predictor, LOO-CV based on the loss function in which we are really interested is able to tell us quite reliably that ML is better than the predictor based on direct optimization of this loss function (which is the sample median for L_1).

3. The idea that all loss functions are asymptotically equivalent again only applies to an estimation problem of a given parameter assuming that the model is known. The convolution theorem doesn't tell us in which parameter $q(\theta)$ in (3.1) we should be interested. The L_1-loss for the prediction problem determines that it is the median.

3.3. Various interpretations of loss functions

According to our main hypothesis, the choice of a loss function is a translation problem. An informal judgment of a situation has to be translated into a mathematical formula. To do this, it is essential to keep in mind how loss functions are to be interpreted. This depends essentially on the use of the loss function, referring to (a), (b) and (c) in the introduction.

(a) In prediction problems, the loss function is about how we measure the quality of a predicted value, having in mind that a true value exists and will be observable in the future. As can be seen from the restaurant example, this is not necessarily true, because if a prediction turns out to be very bad early, the company will react, which prevents the "true value" under the prediction model from being observed (it may further happen that the very fact that the company selects locations based on a new prediction method changes the underlying distribution). However, the idea of an observable true value to be predicted, enables a very direct interpretation of the loss function in terms of observable quantities.

(**b**) The situation is different in estimation problems, where the loss function is a function of an estimator and an underlying, essentially unobservable quantity. The quantification of loss is more abstract in such a situation. For example, the argument used in Section 2 to justify the boundedness of the loss function was that if the prediction is so wrong that it is essentially useless, it doesn't matter anymore how wrong it exactly is. Now imagine the estimation of a treatment effect in medicine. It may be that after some study to estimate the treatment effect, the treatment is applied regularly to patients with a particular disease. Even though, in terms of the prediction of the effect of the treatment on one particular patient, it may hold that it doesn't matter how wrong a grossly wrong prediction exactly is, the situation for the estimation of the overall effect may be much different. Under- or overestimation of the general treatment effect matters to quite a lot of patients, and it may be of vital importance to keep the estimation error as small as possible in case of a not very good estimation, while small estimation errors could easily be tolerated. In such a case, something like the L_2-loss could be adequate for estimation, while a concave loss is preferred for pointwise prediction. It could be argued that, at least in some situations, the estimation loss is nothing else than an accumulated prediction loss. This idea may justify the choice of the mean (which is sensitive to large values) to summarize more robust pointwise prediction losses, as in (2.1). Note that the convolution theorem compares *expected values* of losses, and the expectation as a functional is in itself connected to the L_2-loss. Of course, all of this depends strongly on the subject matter.

(**c**) There is also a direct interpretation that can be given to the use of loss functions to define methods. This is about measuring the quality of data summary by the method. For example, the L_2-loss function defining the least squares estimator defines how the locations of the already observed data points are summarized by the regression line. Because L_2 is convex, it is emphasized that points far away from a bulk of the data are fitted relatively well, to the price that most points are not fitted as precisely as would be possible. Again, a decision has to be made whether this is desired.

> As a practical example, consider a clustering problem where a company wants to assign k storerooms in order to deliver goods to n shops so that the total delivery distance is minimized. This is an L_1-optimization problem (leading to k-medoids) where neither prediction nor estimation are involved. Estimation, prediction and robustness theory could be derived for the resulting clustering method, but they are irrelevant for the problem at hand.

3.4. Data dependent choice of loss functions

In the restaurant example, the loss function has been adjusted because, having seen the results based on the initial specification of c, the company realized that a more "tolerant" specification would be more useful.

Other choices of the loss function dependent on the data or the underlying model (about which the strongest information usually comes from the data) are imaginable, e.g., asymmetric loss for skew distributions and weighting schemes depending on random variations where they are heteroscedastic.

In terms of statistical theory, the consequences of data dependent changes of loss functions can be expected to be at least as serious as data dependent choices of models and methods, which may lead to biased confidence intervals, incoherent Bayesian methodology and the like. Furthermore, the consequences of changing the loss function dependent on the data cannot be analyzed by the same methodology as the consequences of the data dependent choice of models, because the latter analysis always assumes a true model to hold, but there is no single true loss function. It may be argued, though, that the company representatives have a "true subjective" loss function in mind, which they failed to communicate initially.

However, as with all subjective decisions, we have to acknowledge that people change their point of view and their assessment of situations when new information comes in, and they do this often in ways which can't be formally predicted in the very beginning (unforeseen prior-data conflicts in Bayesian analysis are an analogous problem).

Here, we just emphasize that data dependent choice of the loss function may lead to some problems which are not fully understood at the moment. In situations such as the restaurant example, we are willing to accept these problems if the impression exists that the results from the initial choice of the loss function are clearly unsatisfactory, but loss functions should not be changed without urgency.

4. PHILOSOPHICAL ASPECTS

The term “subjective” has been used several times in the present paper. In science, there are usually some reservations against subjective decisions, because of the widespread view that objectivity is a main aim of science.

We use “subjectivity” here in a quite broad sense, meaning any kind of decision which can’t be made by the application of a formal rule of which the uniqueness can be justified by rational arguments. “Subjective decisions” in this sense should take into account subject-matter knowledge, and can be agreed upon by groups of experts after thorough discussion, so that they could be called “inter-subjective” in many situations and are certainly well-founded and not “arbitrary”. However, even in such situations different groups of experts may legitimately arrive at different decisions. This is similar to the impact of subjective decisions on the choice of subjective Bayesian prior probabilities.

For example, even if there are strong arguments in a particular situation that the loss function should be convex, it is almost always impossible to find decisive arguments why it should be exactly equal to L_2. In the restaurant example it could be argued that the loss function should be differentiable (because the sharp switch at c is quite artificial) or that it should not be exactly constant above c. But there isn’t any clear information suggesting how exactly it should behave around c.

Note that the dependent variable in the restaurant example is an amount of money, which, in principle, can be seen as a clear example of a high quality ratio scale measurement. But even this feature doesn’t make the measurement of loss in any way trivial or objective, as has been discussed in Section 2. The fact that it is a non-scientific business application does also not suffice as a reason for the impact of subjective decisions in this example. The argument not to take the absolute value as loss function was that in case of very wrong predictions it may turn out that the prediction is wrong early enough so that it is still possible to react in order to keep the effective loss as small as possible. But this may apply as well in several scientific setups, e.g., in medical, technical and ecological applications. In such a situation there is generally no way to predict exactly what the loss of grossly wrong prediction will be. If it is not possible to predict a given situation reliably, it is even less possible to predict accurately the outcome of possible reactions in case that the initial prediction turns out to be grossly wrong. Furthermore, there are generally no objective rules about how to balance underestimation and overestimation in situations which are not clearly symmetric. Therefore, the need for subjective decisions about the choice of loss functions is general and applies to “objective” science as well.

As emphasized before, a loss function cannot be found as a solution of a formal optimization problem, unless another loss function is invented to define this problem. There is no objectively best loss function, because the loss function defines what “good” means.

The quest for objectivity in science together with a certain misconception of it has some undesirable consequences. Experience shows that it is much easier to get scientific work published which makes use of standard measurements such as the L_2-loss, even in situations in which it is only very weakly (if at all) justified, than to come up with a rather idiosyncratic but sensible loss function involving obviously subjective decisions about functional shapes and tuning constants. It is almost certain that referees will ask for objective justifications or at least sensitivity analyses in the latter case. We are not generally against such sensitivity analyses, but if they are demanded in a situation where authors come up with an already well thought over choice of a loss function, it would be much more urgent to carry out such analyses if “standard” choices have been made without much reflection.

It seems that many scientists see “general agreement” as a main source of objectivity, and therefore they have no doubts about it in case that somebody does “what everybody else does” without justification, while obviously personal decisions, even if discussed properly, are taken as a reason for suspicion. This is clearly counterproductive.

It is important to acknowledge that there is some reason for this general attitude. By changing the loss function, it may actually be possible to arrive at very different results, including results previously desired by the researcher. This is made more difficult by insisting on the use of widespread standard measures that have proven useful under a range of different situations.

We see this as a legitimate, but in no way decisive argument. Science is essentially about reaching stable rational agreement. Certainly, agreement based on the unreflected choice of standard methods cannot be expected to be stable, and it may be controversial at best whether it can be seen as rational. On the other hand, more subjective decisions will not enable agreement as long as they are not backed up by clear comprehensible arguments. Therefore, such arguments have to be given. If for some decisions, there are no strong arguments, it makes sense to stick to standard choices. Therefore, if there are strong arguments that a loss function should be convex, but there is no further clear information how exactly it should look like, the standard choice L_2 should be chosen on grounds of general acceptance. But even if L_2 is chosen in such a situation, convexity should still be justified and it makes even sense to admit that, apart from convexity, L_2 has been chosen purely for the above reason. This is as well a subjective, but rational decision in the sense given in the beginning of this section.

A more sophisticated but often impractical approach would start from a list of characteristics (axioms) that the loss function in a particular application should fulfill, and then investigate the range of results obtained by the whole class of such loss functions.

The perhaps most important aspect of scientific agreement is the possibility to communicate in an unambiguous way, which is mainly ensured by mathematical formalism. Therefore, the subjective design of more or less idiosyncratic loss functions, including their detailed discussion, contributes to the clarity of the viewpoint of the researcher. Her subjective decisions become transparent and are accessible to rational discussion. Making the subjective impact clear in this way actually helps scientific discussion much more than to do what everybody else does without much discussion.

We don't know whether and to what extent our attitude to science is already present in the philosophical literature, but it seems to be quite close to what Ernest ([2]) wrote in his chapter about "the social construction of objective knowledge". Some more elaboration can be found in Hennig ([3]).

5. CONCLUSION

We hope that the present paper encourages researchers to choose or design loss functions which reflect closely their expert's view of the situation in which the loss function is needed. Instead of being "less objective", this would be rather quite helpful for scientific discussion.

Robustness is not treated as an aim in itself here, but rather as an implicit consequence of the decision of the researchers about the formalization of the prediction loss for atypical observations.

There are other problems in data analysis where similar principles can be applied. One example is the design of dissimilarity measures, see Hennig and Hausdorf ([4]). Combination of different loss criteria (such as efficiency and robustness in estimation) has not been treated in the present paper, but could be approached in a similar spirit.

ACKNOWLEDGMENTS

This paper is based on material presented by the first author at the International Conference on Robust Statistics, Lisbon 2006. We'd like to thank several participants of this conference for very valuable comments, which influenced the present paper quite a lot.

REFERENCES

[1] Bickel, P.J.; Klaassen, C.A.J.; Ritov, Y. and Wellner, J.A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*, Springer, New York.

[2] Ernest, P. (1998). *Social Constructivism as a Philosophy of Mathematics*, State University of New York Press.

[3] Hennig, C. (2003). *How wrong models become useful – and correct models become dangerous.* In "Between Data Science and Applied Data Analysis" (M. Schader, W. Gaul and M. Vichi, Eds.), Springer, Berlin, 235–243.

[4] Hennig, C. and Hausdorf, B. (2006). *Design of dissimilarity measures: a new dissimilarity measure between species distribution ranges.* In "Data Science and Classification" (V. Batagelj, H.-H. Bock, A. Ferligoj, A. Ziberna, Eds.), Springer, Berlin, 29–38.

[5] Huber, P.J. (1981). *Robust Statistics*, Wiley, New York.

[6] Lehmann, E.L. and Casella, G. (1998). *Theory of Point Estimation* (2^{nd} ed.), Springer, New York.

[7] Leung, D.H.-Y. (2005). Cross-validation in nonparametric regression with outliers, *Annals of Statistics*, **33**, 2291–2310.

[8] Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression, *Annals of Statistics*, **12**, 758–765.

[9] Ronchetti, E.; Field, C. and Blanchard, W. (1997). Robust linear model selection by cross-validation, *Journal of the American Statistical Association*, **92**, 1017–1023.

[10] Rousseeuw, P.J. (1984). Least median of squares regression, *Journal of the American Statistical Association*, **79**, 871–880.

[11] Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process, *Annals of Statistics*, **8**, 147–164.

[12] Shibata, R. (1981). An optimal selection of regression variables, *Biometrika*, **68**, 45–54.

[13] Western, B. (1995). Concepts and suggestions for robust regression analysis, *American Journal of Political Science*, **39**, 786–817.

[14] Yohai, V.J. (1987). High breakdown-point and high efficiency robust estimates for regression, *Annals of Statistics*, **15**, 642–656.

REVSTAT – Statistical Journal
Volume 5, Number 1, March 2007, 41–61

ESTIMATING SPECTRAL DENSITY FUNCTIONS ROBUSTLY

Authors: Bernhard Spangl
– Group of Applied Statistics and Computing,
University of Natural Resources and Applied Life Sciences,
Vienna, Austria
bernhard.spangl@boku.ac.at

Rudolf Dutter
– Department of Statistics and Probability Theory,
Vienna University of Technology,
Austria
R.Dutter@tuwien.ac.at

Abstract:

- We consider in the following the problem of robust spectral density estimation.

  Unfortunately, conventional spectral density estimators are not robust in the presence of additive outliers (cf. [18]). In order to get a robust estimate of the spectral density function, it turned out that cleaning the time series in a robust way first and calculating the spectral density function afterwards leads to encouraging results. To meet these needs of cleaning the data we use a robust version of the Kalman filter which was proposed by Ruckdeschel ([26]). Similar ideas were proposed by Martin and Thomson ([18]).

  Both methods were implemented in R (cf. [23]) and compared by extensive simulation experiments. The competitive method is also applied to real data. As a special practical application we focus on actual heart rate variability measurements of diabetes patients.

Key-Words:

- *robustness; spectral density function; AO-model.*

AMS Subject Classification:

- 62F35, 62M15, 60G35.

1. INTRODUCTION

Our research has been motivated by the frequency-domain analysis of short-term heart rate variability (HRV) measurements. This is a non-invasive method which has been increasingly used in medicine (cf. [8, 22]). To analyze biosignals or, generally speaking, time series, the spectral density function is commonly used in many application areas. Further areas of applications besides medicine are signal processing (cf. [31]) and geophysics (cf. [4, 9]).

The additive outlier model (AO model) which was introduced by Fox ([6]) is a commonly used model for outliers in time series. The AO model consists of a stationary core process, x_t, to which occasional outliers are added. The observed process $\{y_t,\ t=1,...,n\}$ is said to have additive outliers if it is defined by

$$y_t \;=\; x_t + v_t \tag{1.1}$$

where the contaminations v_t are independent and identically distributed. For the methods presented in this paper, it is convenient to model the univariate distribution of v_t by a contaminated normal distribution with degenerated central component, i.e.,

$$\mathcal{CN}(\gamma,0,\sigma^2) \;=\; (1-\gamma)\,\mathcal{N}(0,0) + \gamma\,\mathcal{N}(0,\sigma^2)\ . \tag{1.2}$$

Hence, the core process x_t is observed with probability $1-\gamma$ whereas the core process plus a disturbance v_t is observed with probability γ. We shall also assume that x_t and v_t are independent.

The AO model seems to be an appropriate model when analyzing heart rate variability data. To access the variability of heart rate in the frequency domain the spectral density function of the tachogram is estimated. The tachogram is the series of time intervals between consecutive heart beats, the so called R-R-intervals (e.g. Figure 1). The R-R-interval denotes the period between an R-peak and the next R-peak in an electrocardiogram.

Non-sinus ectopic beats and other artifacts can cause outlying observations in the tachogram. If, during the recording and sampling, an R-peak is missed in the electrocardiogram (ECG) this will result in a very large value in the tachogram. Or, if an ectopic beat occurs, i.e., if there is an extra heart beat between two regular beats, the amplitude in the ECG of the heart beat following the ectopic beat will be very low and therefore this beat will usually be missed. This results in a lower tachogram value followed by a higher one.

The aim of accessing the heart rate variability is accomplished by estimating the spectral density function of the tachogram robustly in order to be insensitive against outlying tachogram values caused by ectopic beats and other artifacts.

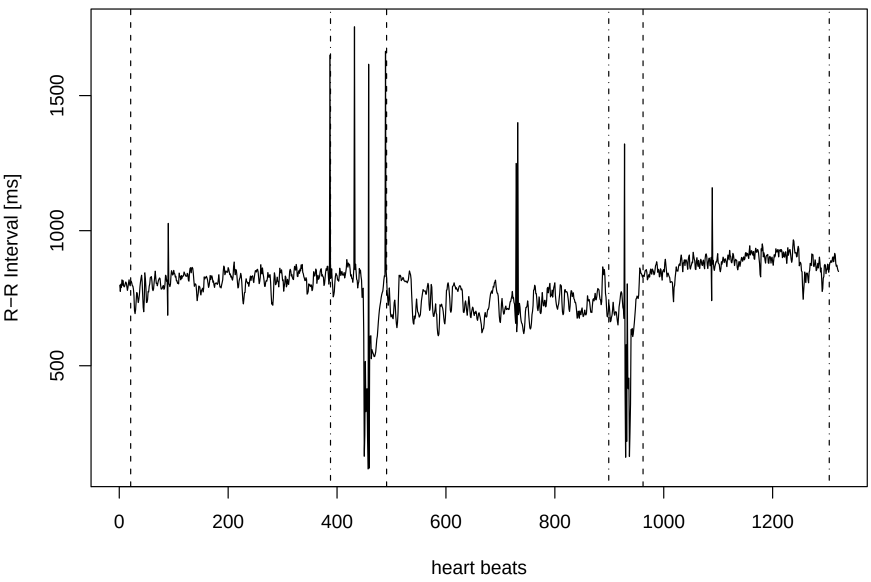

Figure 1: Tachogram of 1321 consecutive heart beats.

We do not compute the spectral density function of the entire tachogram series, but estimate several within overlapping windows to assure stationarity in each window (cf. also [28]). Each slice in Figure 2 represents the spectral density estimate of the corresponding time interval.

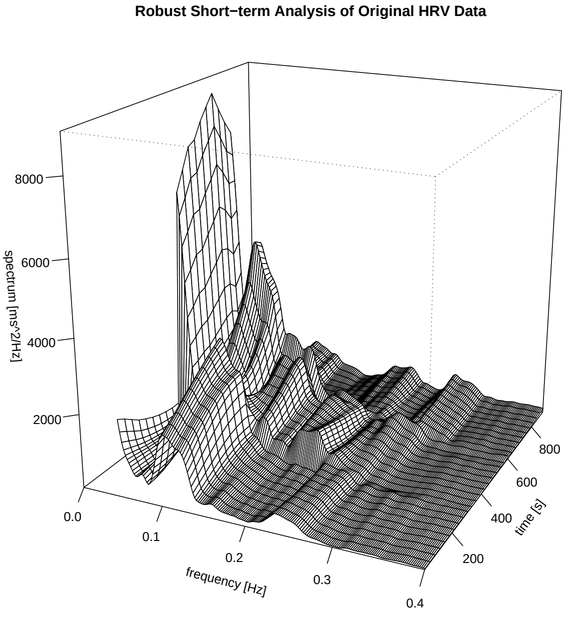

Figure 2: Robust dynamic Fourier analysis of the original short-term HRV data.

Although we do not use the entire tachogram series but several overlapping windows to access the heart rate variability we only focus on an analysis in the frequency domain. We are not interested in modeling the heart rate in the time domain nor in forecasting as this is often the aim in the context of online-monitoring.

In the present paper we consider the problem of estimating the spectral density function robustly. Unfortunately, conventional spectral density estimators are not robust in the presence of additive outliers. See [12] or [18] for details. To obtain a robust estimate of the spectral density function we present two different multi-step procedures. The first procedure was proposed by Martin and Thomson ([18]) and incorporates an important robust filtering operation which is accomplished by an approximate conditional-mean (ACM) type filter. For the second multi-step procedure we suggest to replace the ACM type filter and use the rLS filter proposed by Ruckdeschel ([26]) instead. Both filters are robustified versions of the Kalman filter. In order to compare both approaches we implement them in R.

In the next section we state the definitions of the state-space model and the classical Kalman filter which is the basis of the robustifying approaches proposed by Martin and Thomson ([18]) and Ruckdeschel ([26]). Both methods are described in Section 3. In Section 4 we give an outline of our simulation study and the results are presented in Section 5. Some remarks are given in Section 6.

2. PRELIMINARIES

2.1. State-space models

Let us assume we observe a q-dimensional, vector-valued process $\boldsymbol{y}_t$, $t=1,...,n$, which is only a linear transformation of an unobserved p-dimensional signal $\boldsymbol{x}_t$ with some noise added. Then the *state-space model* can be defined as follows:

$$\begin{aligned} \boldsymbol{x}_t &= \boldsymbol{\Phi}\,\boldsymbol{x}_{t-1} + \boldsymbol{\varepsilon}_t\ , \\ \boldsymbol{y}_t &= \boldsymbol{H}\boldsymbol{x}_t + \boldsymbol{v}_t\ , \end{aligned} \tag{2.1}$$

where $\boldsymbol{x}_t$ is the unobserved p-dimensional vector called the *state vector*. The first equation in (2.1) is called *state equation* and the second is called the *observation equation*. It is assumed that $\boldsymbol{\varepsilon}_t$ has dimension p, $\boldsymbol{\Phi}$ is a $p{\times}p$ matrix and $\boldsymbol{H}$ is a $q{\times}p$ matrix. We further assume that $\boldsymbol{x}_t$ is independent of future $\boldsymbol{\varepsilon}_t$, and that $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{v}_t$ are individually zero mean independent and identically distributed (iid) sequences which also are mutually independent but could be non-Gaussian.

A more general definition of state-space models considering correlated errors as well as more complex models including exogenous variables or selection matrices can be found in [27] and [5].

2.2. The classical Kalman filter

The primary aim of any analysis using state-space models as defined by (2.1) is to produce estimators of the underlying unobserved signal $\boldsymbol{x}_t$, given the data $\boldsymbol{Y}_s = \{\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_s\}$, up to time s. If $s < t$, $s = t$ or $s > t$, the problem is called *prediction*, *filtering* or *smoothing*, respectively.

In addition, we want to get estimators $T_t(\boldsymbol{Y}_s)$ of $\boldsymbol{x}_t$ which are best in the sense of the minimum mean-squared error, i.e.,

$$E\big(\|\boldsymbol{x}_t - T_t(\boldsymbol{Y}_s)\|^2\big) \;=\; \min_{T_t}! \; . \tag{2.2}$$

The solution is the conditional mean of $\boldsymbol{x}_t$ given $\boldsymbol{Y}_s$, i.e.,

$$T_t(\boldsymbol{Y}_s) \;=\; E(\boldsymbol{x}_t \,|\, \boldsymbol{Y}_s) \;, \tag{2.3}$$

and will further on be denoted by $\boldsymbol{x}_{t|s}$.

However, in general the conditional mean is hard to calculate and therefore we restrict ourselves to the class of linear estimators. Then the solution to these problems is accomplished via the *Kalman filter* and *smoother* (cf. [10, 11]). The estimators we obtain are the minimum mean-squared error estimates within the class of linear estimators.

In the following we will just focus on the Kalman filter. Its advantage is that it specifies how to update the filter values from $\boldsymbol{x}_{t-1|t-1}$ to $\boldsymbol{x}_{t|t}$ once a new observation $\boldsymbol{y}_t$ is obtained, without having to reprocess the entire data set $\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_t$. The Kalman filter recursions can be split into three steps:

(i) Initialization $(t=0)$:

$$\boldsymbol{x}_{0|0} = \boldsymbol{\mu}_0 \;, \qquad \boldsymbol{P}_0 = \boldsymbol{\Sigma}_0 \;, \tag{2.4}$$

where $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ are the unconditional mean and $p{\times}p$ covariance matrix of $\boldsymbol{x}_0$;

(ii) Prediction $(t \geq 1)$:

$$\begin{aligned} \boldsymbol{x}_{t|t-1} &= \boldsymbol{\Phi}\, \boldsymbol{x}_{t-1|t-1} \;, \\ \boldsymbol{M}_t &= \boldsymbol{\Phi}\, \boldsymbol{P}_{t-1} \boldsymbol{\Phi}^\top + \boldsymbol{Q} \;; \end{aligned} \tag{2.5}$$

(iii) Correction $(t \geq 1)$:

$$\begin{aligned} \boldsymbol{x}_{t|t} &= \boldsymbol{x}_{t|t-1} + \boldsymbol{K}_t(\boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{x}_{t|t-1}) \ , \\ \boldsymbol{P}_t &= \boldsymbol{M}_t - \boldsymbol{K}_t\boldsymbol{H}\boldsymbol{M}_t \ , \\ \text{with} \qquad \boldsymbol{K}_t &= \boldsymbol{M}_t\boldsymbol{H}^\top(\boldsymbol{H}\boldsymbol{M}_t\boldsymbol{H}^\top + \boldsymbol{R})^{-1} \ . \end{aligned} \tag{2.6}$$

The $p\times q$ matrix $\boldsymbol{K}_t$ is called the *Kalman gain*. The $p\times p$ matrix $\boldsymbol{M}_t$ is the conditional prediction error covariance matrix,

$$\boldsymbol{M}_t = E\Big((\boldsymbol{x}_t - \boldsymbol{x}_{t|t-1})\,(\boldsymbol{x}_t - \boldsymbol{x}_{t|t-1})^\top \mid \boldsymbol{Y}_{t-1}\Big) \ , \tag{2.7}$$

and the conditional filtering error covariance matrix $\boldsymbol{P}_t$ is given by

$$\boldsymbol{P}_t = E\Big((\boldsymbol{x}_t - \boldsymbol{x}_{t|t})\,(\boldsymbol{x}_t - \boldsymbol{x}_{t|t})^\top \mid \boldsymbol{Y}_t\Big) \ . \tag{2.8}$$

Moreover, the $p\times p$ matrix $\boldsymbol{Q}$ and the $q\times q$ matrix $\boldsymbol{R}$ denote the covariance matrices of $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{v}_t$, respectively.

3. ROBUST SPECTRAL DENSITY ESTIMATION

In order to obtain a robust estimate of the spectral density function, we clean the data in a robust way first and compute the spectral density function afterwards using a prewhitened spectral density estimate. This approach was proposed by Martin and Thomson ([18]) and leads to encouraging results. The data-cleaning operation wherein the robustness is introduced is accomplished by a robustified version of the Kalman filter.

Martin and Thomson ([18]), based on the work of Martin ([15]), propose to modify the calculation of the filter estimate as well as of the conditional filtering error covariance matrix in the correction step (2.6). In [15] Martin, motivated by Masreliez's result ([20]), only considers autoregressive models. This limitation to univariate signals and several approximations lead to a simplification of the correction step that enables a robust estimation of the filter estimate as well as of the conditional filtering error covariance matrix.

Another approach, proposed by Ruckdeschel ([26]), preserves the general concept of the Kalman filter, that allows for multivariate signals, and modifies only the updating of the filter estimate.

3.1. Robust prewhitening

Let $\{y_t,\, t=1,...,n\}$ again denote the observed process which is assumed to be second-order stationary and to have mean zero. The cleaning operator C maps the original data y_t into the cleaned data Cy_t. In the context of the AO model (1.1), we want the Cy_t to reconstruct the core process x_t, and so we will use the labeling $Cy_t = \widehat{x}_{t|t}$, where $\widehat{x}_{t|t}$ denotes an estimate of x_t at time t. The second index of $\widehat{x}_{t|t}$ should indicate that the kind of data cleaning procedure we have in mind here is a robust filtering procedure which uses the past and present data values $y_1,...,y_t$ to produce a cleaned filter estimate $\widehat{x}_{t|t}$ of x_t, $t=1,...,n$. For AO models with a fraction of contamination γ not too large, it turns out that the data cleaner has the property that $Cy_t = y_t$ most of the time, that is about $(1-\gamma)\times 100$ percent of the time.

The filter-cleaner procedure involves a robust estimation of an autoregressive approximation to the core process x_t of order p, with estimated coefficients $\widehat{\phi}_1,...,\widehat{\phi}_p$. Now, the residual process

$$r_t \,=\, Cy_t - \sum_{i=1}^{p} \widehat{\phi}_i\, Cy_{t-i}\;, \qquad t=p+1,...,n\;, \tag{3.1}$$

can easily be formed. Since cleaned data are used to obtain these residuals, and the $\widehat{\phi}_i$ are robust estimates, the transformation (3.1) is called a robust prewhitening operation. The benefit in the use of prewhitening in the context of spectral density estimation is to reduce the bias, i.e., the transfer of power from one frequency region of the spectral density function to another, known as leakage (cf. [3]).

The robust spectral density estimate is based on the above robust prewhitening as follows. Let

$$\widehat{H}_p(f) \,=\, 1 - \sum_{j=1}^{p} \widehat{\phi}_j\, e^{-i2\pi jf} \tag{3.2}$$

be the transfer function of the prewhitening operator (3.1) at frequency f, and let $\widehat{S}_r^{(lw)}(f)$ denote a lag window spectral estimate based on the residual process r_t. Then the spectral density estimate is

$$\widehat{S}(f) \,=\, \frac{\widehat{S}_r^{(lw)}(f)}{\left|\widehat{H}_p(f)\right|^2}\;, \tag{3.3}$$

where $\widehat{S}(f)$ is evaluated at the Fourier frequencies $f_k = k/n,\ k=0,...,[n/2]$.

3.2. The robust filter-cleaner algorithm

The robust filter-cleaner proposed by Martin and Thomson ([18]) is an approximate conditional-mean (ACM) type filter motivated by Masreliez's result ([20]).

3.2.1. The robust filter-cleaner

The filter-cleaner algorithm as presented in the paper of Martin and Thomson ([18]) relies on the p-th order autoregressive approximation of the underlying process x_t, which can be represented in state-space form (2.1) as follows. Assuming that x_t satisfies

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + \varepsilon_t$$

the state space model can be written as

$$\begin{aligned} \boldsymbol{x}_t &= \boldsymbol{\Phi}\, \boldsymbol{x}_{t-1} + \boldsymbol{\varepsilon}_t \ , \\ y_t &= x_t + v_t \ , \end{aligned} \tag{3.4}$$

with

$$\boldsymbol{x}_t = (x_t, x_{t-1}, ..., x_{t-p+1})^\top \ , \tag{3.5}$$

$$\boldsymbol{\varepsilon}_t = (\varepsilon_t, 0, ..., 0)^\top \tag{3.6}$$

and

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_1 & \cdots & \phi_{p-1} & \phi_p \\ 1 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 \end{pmatrix} . \tag{3.7}$$

Additionally, we set

$$\operatorname{cov}(\boldsymbol{\varepsilon}_t) = \boldsymbol{Q} = \begin{pmatrix} \sigma_\varepsilon^2 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} \tag{3.8}$$

and

$$\operatorname{var}(v_t) = \boldsymbol{R} = \sigma_0^2 \ . \tag{3.9}$$

The algorithm computes robust estimates $\widehat{\boldsymbol{x}}_{t|t}$ of the unobservable $\boldsymbol{x}_t$ according to the following recursion:

$$\widehat{\boldsymbol{x}}_{t|t} = \boldsymbol{\Phi}\, \widehat{\boldsymbol{x}}_{t-1|t-1} + \frac{\boldsymbol{m}_{.1,t}}{s_t^2}\, s_t\, \psi\left(\frac{y_t - \widehat{y}_{t|t-1}}{s_t}\right) \tag{3.10}$$

with $\boldsymbol{m}_{.1,t}$ being the first column of $\boldsymbol{M}_t$ which is computed recursively as

$$\boldsymbol{M}_{t+1} = \boldsymbol{\Phi}\,\boldsymbol{P}_t\,\boldsymbol{\Phi}^\top + \boldsymbol{Q}\ , \tag{3.11}$$

$$\boldsymbol{P}_t = \boldsymbol{M}_t - w\left(\frac{y_t - \widehat{y}_{t|t-1}}{s_t}\right)\frac{\boldsymbol{m}_{.1,t}\,\boldsymbol{m}_{.1,t}^\top}{s_t^2}\ . \tag{3.12}$$

The weight function w is defined by

$$w(r) = \frac{\psi(r)}{r}\ , \tag{3.13}$$

where ψ stands for some psi-function described below. The scale s_t is set to

$$s_t^2 = m_{11,t} \tag{3.14}$$

and $\widehat{y}_{t|t-1}$ denotes a robust one-step-ahead prediction of y_t based on $\boldsymbol{Y}_{t-1} = \{y_1,...,y_{t-1}\}$, and is given by

$$\widehat{y}_{t|t-1} = (\boldsymbol{\Phi}\,\widehat{\boldsymbol{x}}_{t-1|t-1})_1\ . \tag{3.15}$$

Finally, the cleaned process at time t results in

$$\widehat{x}_{t|t} = (\widehat{\boldsymbol{x}}_{t|t})_1\ . \tag{3.16}$$

It should be noted that if ψ is the identity function, which gives $w \equiv 1$, and (3.14) is replaced by $s_t^2 = m_{11,t} + \sigma_0^2$ with $\sigma_0^2 = \mathrm{var}(v_t)$ in the AO model, the above recursions are those of the Kalman filter. The use of $\sigma_0^2 = 0$ in (3.14) corresponds to the assumptions that $v_t = 0$ a large fraction of time and that a contaminated normal distribution with degenerated central component (1.2) provides a reasonable model. Correspondingly, $\boldsymbol{M}_t$ and $\boldsymbol{P}_t$ are the prediction and filtering error-covariance matrices as described in the previous section (Section 2). Again, in order to agree with the definition of the classical Kalman filter recursions, we specify the initial conditions for the above recursions by setting $\widehat{\boldsymbol{x}}_{0|0} = 0$ and $\boldsymbol{P}_0 = \widehat{\boldsymbol{C}}_{\boldsymbol{x}}$ where $\widehat{\boldsymbol{C}}_{\boldsymbol{x}}$ is an estimate of the $p{\times}p$ covariance matrix of the state process. We note that there also exists another way to specify those initial conditions (see [17]).

The psi-function ψ and the weight function w which are essential to obtain robustness should be bounded and continuous. Additionally, it is highly desirable that both have zero values outside a bounded, symmetric interval around the origin. Furthermore, $\psi(s)$ is odd and should look like the identity function for small values of s (see [15]). Boundedness assures that no single observation has an arbitrarily large effect on the filter-cleaner. Continuity assures that small variations, e.g., due to rounding, will not have a major effect. Compact support results in the following behavior which is desirable for a filter-cleaner: if an observation y_t deviates from its prediction $\widehat{y}_{t|t-1}$ by a sufficiently large amount, then $\widehat{\boldsymbol{x}}_{t|t}$ will be the pure prediction $\widehat{\boldsymbol{x}}_{t|t} = \boldsymbol{\Phi}\,\widehat{\boldsymbol{x}}_{t-1|t-1}$ and the filtering error covariance

$\boldsymbol{P}_t$ is set equal to the prediction error covariance $\boldsymbol{M}_t$. Martin and Thomson ([18]) proposed to use for ψ a special form of *Hampel's three-part redescending psi-function* ([7]),

$$\psi_{HA}(s) \;=\; \begin{cases} s & \text{if} \;\; |s| \le a\,, \\ a\,\mathrm{sgn}(s) & \text{if} \;\; a < |s| \le b\,, \\ \dfrac{a}{b-c}\big(s - c\,\mathrm{sgn}(s)\big) & \text{if} \;\; b < |s| \le c\,, \\ 0 & \text{if} \;\; c < |s|\;, \end{cases} \tag{3.17}$$

namely, Hampel's two-part redescending psi-function, with $b = a$, which has all the desirable properties.

3.2.2. An approximate optimality result

There is an approximate optimality result for the filter described above if we replace (3.14) by

$$s_t^2 \;=\; m_{11,t} + \sigma_0^2\;, \tag{3.18}$$

and ψ and w in (3.10) and (3.13), respectively, by

$$w(r) \;=\; \psi'(r) \;=\; \frac{\partial}{\partial r}\psi(r)\;. \tag{3.19}$$

Namely, under the assumption that the state prediction density $f_{\boldsymbol{x}_t}(\,.\,|\,\boldsymbol{Y}_{t-1})$ is Gaussian and that $\psi(r) = -(\partial/\partial r)\log g(r)$, where g is an approximation of the observation prediction density $f_{y_t}(\,.\,|\,\boldsymbol{Y}_{t-1})$, the filter is the conditional-mean filter proposed by Masreliez ([20]). The preceding assumption will never hold exactly under an AO model where v_t is non-Gaussian (see [15], Sec. 5). However, there is some evidence that $f_{\boldsymbol{x}_t}(\,.\,|\,\boldsymbol{Y}_{t-1})$ is nearly Gaussian and that the filter is a good approximation to the exact conditional-mean filter. Therefore the filter is referred to as an approximate conditional-mean (ACM) type filter. More details can be found in [15]. The results therein suggest that the use of Hampel's two-part redescending psi-function is reasonable when the observation noise v_t has a contaminated normal distribution. However, the weight function w given by (3.19) is discontinuous if using Hampel's two-part redescending psi-function, and therefore Martin and Thomson ([18]) prefer to specify w by (3.13).

3.2.3. Fixed-lag smoother-cleaners

As mentioned in [15], if one uses the last coordinate of the filter estimate $\widehat{\boldsymbol{x}}_{t|t}$ to produce cleaned data, then one has that $\widehat{x}_{t-p+1} = (\widehat{\boldsymbol{x}}_{t|t})_p$ is an estimate of x_{t-p+1} based on the observations $\boldsymbol{Y}_t$ up to time t. Such an estimate is usually called a fixed-lag smoother, with lag $p-1$ in this case.

3.2.4. Estimation of hyper parameters

To use the filter-cleaner algorithm we need robust estimates $\widehat{\boldsymbol{\phi}}$, $\hat{\sigma}_\varepsilon$ and $\widehat{\boldsymbol{C}}_{\boldsymbol{x}}$ of the AR(p) parameter vector $\boldsymbol{\phi} = (\phi_1, ..., \phi_p)^\top$, the innovations scale σ_ε and the $p{\times}p$ covariance matrix of the state process, respectively. Martin and Thomson ([18]) proposed to get initial estimates using bounded-influence autoregression (BIAR) via the *iteratively reweighted least squares* (IWLS) *algorithm*. Details about BIAR may be found in [19], [16] or [29].

3.2.5. Selection of order p

Martin and Thomson ([18]) propose the following procedure to select the order p of the autoregressive approximation. For increasing orders p BIAR estimates are computed and the estimated innovation scale estimates $\hat{\sigma}_\varepsilon(p)$ are examined for each order. The final order is selected as that value of p for which $\hat{\sigma}_\varepsilon(p+1)$ is not much smaller than $\hat{\sigma}_\varepsilon(p)$, e.g., less than a 10-percent decrement as suggested by Martin and Thomson ([18]).

Another robust order-selection rule based on BIAR estimates and motivated by Akaike's minimization criterion ([1]) was proposed by Martin ([16]).

3.3. The robust Least Squares (rLS) filter algorithm

In the following we describe a robustified version of the Kalman filter which was proposed by Ruckdeschel ([26]).

3.3.1. Robustified optimization problem

The idea is to reduce in the correction step (2.6) of the classical Kalman filter the influence of an observation $\boldsymbol{y}_t$ that is affected by an additive outlier. Instead of $\boldsymbol{K}_t \Delta \boldsymbol{y}_t$ with $\Delta \boldsymbol{y}_t = \boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{x}_{t|t-1}$ we use a huberized version of it, i.e.,

$$H_{b_t}(\boldsymbol{K}_t \, \Delta \boldsymbol{y}_t) \; = \; \boldsymbol{K}_t \, \Delta \boldsymbol{y}_t \; \min\left\{1, \; \frac{b_t}{\|\boldsymbol{K}_t \, \Delta \boldsymbol{y}_t\|}\right\} , \tag{3.20}$$

so that the obtained result will be equal to the one of the classical Kalman filter, if $\|\boldsymbol{K}_t \Delta \boldsymbol{y}_t\|$ is not too large, whereas if $\|\boldsymbol{K}_t \Delta \boldsymbol{y}_t\|$ is too large, the direction will remain unchanged and it will be projected on the q-dimensional ball with radius b_t.

This leads to a robustified optimization problem given by

$$E\big(\|\Delta \boldsymbol{x}_t - H_{b_t}(\boldsymbol{K}_t\,\Delta \boldsymbol{y}_t)\|^2\big) \;=\; \min_{\boldsymbol{K}_t}! \;, \tag{3.21}$$

where $\Delta \boldsymbol{x}_t = \boldsymbol{x}_t - \boldsymbol{x}_{t|t-1}$ denotes the prediction error. The above optimization problem is equivalent to the optimization problem (2.2) of the classical Kalman filter and its solution is named $\boldsymbol{K}_t^{\mathrm{rLS}}$.

3.3.2. The rLS filter

Hence, this gives us the following filter recursions:

(i) Initialization $(t=0)$:

$$\boldsymbol{x}_{0|0}^{\mathrm{rLS}} \;=\; \boldsymbol{\mu}_0 \;; \tag{3.22}$$

(ii) Prediction $(t\geq 1)$:

$$\boldsymbol{x}_{t|t-1}^{\mathrm{rLS}} \;=\; \boldsymbol{\Phi}\,\boldsymbol{x}_{t-1|t-1}^{\mathrm{rLS}} \;; \tag{3.23}$$

(iii) Correction $(t\geq 1)$:

$$\boldsymbol{x}_{t|t}^{\mathrm{rLS}} \;=\; \boldsymbol{x}_{t|t-1}^{\mathrm{rLS}} + H_{b_t}\big(\boldsymbol{K}_t^{\mathrm{rLS}}(\boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{x}_{t|t-1}^{\mathrm{rLS}})\big) \;. \tag{3.24}$$

The above filter recursions will be named *r̲obust l̲east s̲quares* (rLS) *filter*.

Because the calculation of $\boldsymbol{K}_t^{\mathrm{rLS}}$ is computationally extensive Ruckdeschel ([26]) proposes to use $\boldsymbol{K}_t^{\mathrm{KK}}$ instead where $\boldsymbol{K}_t^{\mathrm{KK}}$ denotes the Kalman gain obtained by the classical Kalman filter recursions. Simulation studies therein have shown that the worsening, in sense of a larger mean-squared error, is only small if using $\boldsymbol{K}_t^{\mathrm{KK}}$ instead of $\boldsymbol{K}_t^{\mathrm{rLS}}$. Hence, this simplifying modification almost yields the classical Kalman filter recursions with the only exception of replacing the first line of the correction step (2.6) by

$$\boldsymbol{x}_{t|t} \;=\; \boldsymbol{x}_{t|t-1} + H_{b_t}\big(\boldsymbol{K}_t^{\mathrm{KK}}(\boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{x}_{t|t-1})\big) \;. \tag{3.25}$$

From now on, if speaking of the rLS filter, we will only consider this modified version.

Moreover, Ruckdeschel ([26]) proved that the rLS filter is SO-optimal under certain side conditions. SO stands for s̲ubstitutive o̲utlier and means that, instead of disturbing v_t, contamination effects y_t directly, replacing it by an arbitrarily distributed variable y'_t with some low probability. For further details we refer the reader to [26].

Still, the open problem of fixing the clipping height b_t remains.

3.3.3. Fixing the clipping height b_t

In order to properly choose b_t Ruckdeschel ([26]) proposes an assurance criterion: How much efficiency in the ideal model relative to the optimal procedure, i.e., the Kalman filter, am I ready to pay in order to get robustness under deviations from the ideal model? This loss of efficiency, which we will obtain if we use a robust version instead of the classical Kalman filter, is quantified as the relative worsening of the mean-squared error in the ideal model. Hence, for a given relative worsening $\delta > 0$ we solve

$$E\Big(\big\|\Delta\boldsymbol{x}_t - H_{b_t}(\boldsymbol{K}_t^{\mathrm{rLS}}\Delta\boldsymbol{y}_t)\big\|^2\Big) \overset{!}{=} (1+\delta)\,E\Big(\big\|\Delta\boldsymbol{x}_t - \boldsymbol{K}_t^{\mathrm{KK}}\Delta\boldsymbol{y}_t\big\|^2\Big)\ . \tag{3.26}$$

The symbol $\overset{!}{=}$ means that b_t is chosen in a way to achieve equality.

Again, we use the simplifying modifications just mentioned and replace $\boldsymbol{K}_t^{\mathrm{rLS}}$ by $\boldsymbol{K}_t^{\mathrm{KK}}$. Moreover, in most time-invariant situations, the sequence of $\boldsymbol{M}_t$ (and hence also of $\boldsymbol{P}_t$ and $\boldsymbol{K}_t^{\mathrm{KK}}$) stabilizes due to asymptotic stationarity. Thus, once $\boldsymbol{M}_t$ does not change for more than a given tolerance level, we can stop calibration and use the last calculated b_t for all subsequent times s, $s > t$. The Kalman gain and filtering error covariance matrix used in this last calibration step will be denoted by $\boldsymbol{K}_\infty^{\mathrm{KK}}$ and $\boldsymbol{P}_\infty$, respectively. For details we refer to [2] and [21]. Further we make another simplifying modification and assume that for all t

$$\Delta\boldsymbol{x}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{M}_t) \qquad \text{and} \qquad \boldsymbol{v}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R})\ . \tag{3.27}$$

Thus, we may solve

$$\begin{aligned} E\Big(\big\|\Delta\boldsymbol{x} - H_b(\boldsymbol{K}_\infty^{\mathrm{KK}}\Delta\boldsymbol{y})\big\|^2\Big) &\overset{!}{=} (1+\delta)\,E\Big(\big\|\Delta\boldsymbol{x} - \boldsymbol{K}_\infty^{\mathrm{KK}}\Delta\boldsymbol{y}\big\|^2\Big) \\ &= (1+\delta)\ \mathrm{tr}\,\boldsymbol{P}_\infty\ , \end{aligned} \tag{3.28}$$

in b, uniquely for a given loss of efficiency δ, where $\mathrm{tr}\,\boldsymbol{P}_\infty$ denotes the trace of the conditional filtering error covariance matrix. We note that the relative time-expensive calibration, i.e., finding b to a given δ, can be done beforehand. Additional details may be found in [26] and [25].

4. SIMULATION STUDY

The outline of our simulation study is as follows: First we simulate a core process x_t of length $n=100$. x_t is chosen to be an autoregressive process of order 2 given by

$$x_t \;=\; x_{t-1} - 0.9\,x_{t-2} + \varepsilon_t \;, \tag{4.1}$$

with $\varepsilon_t \sim \mathcal{N}(0,1)$. The variance of the core process x_t, i.e., the value of the autocovariance function at lag zero can be calculated by numerical integration and is given approximately by $\mathrm{var}(x_t) \approx 7.27$. Additionally, the additive outliers are simulated from a contaminated normal distribution with degenerate central component (1.2) with $\sigma^2 = 10^2$. The contamination γ is varied from 0% to 20% by steps of 5%. That means that with probability γ, v_t is an additive outlier with $v_t \neq 0$. To obtain the contaminated process y_t, the v_t's are added to the core process x_t. For each level of contamination this was done 400 times.

For each of the contaminated series, estimates of the hyper parameter, i.e., the innovations scale $\widehat{\sigma}_\varepsilon$, the autoregressive parameters $\widehat{\phi}_1, ..., \widehat{\phi}_p$ and the $p\times p$ covariance matrix $\widehat{\boldsymbol{C}}_{\boldsymbol{x}}$ of the state process $\boldsymbol{x}_t$, are computed via bounded-influence autoregression. The order p of the autoregressive approximation is chosen according to the order-selection criterion proposed by Martin and Thomson ([18]), which yields values of p from 2 to 3 subject to the contamination level. In order to be able to compare the results we choose an equal order p for all levels of contamination and fix it equal to 3. Using an order $p=2$ in cases of lower contamination levels, where this is appropriate, we obtain almost perfect fits for both filtering algorithms. But, although the simulated core process is of order 2, the estimated BIAR parameters we obtain setting p equal to 3 are similar to the ones of the original core process, i.e., the first two AR parameters are close to the original ones and the third AR parameter is almost zero, as one would expect.

Then each process is cleaned using the ACM-type filter and the rLS filter proposed by Martin and Thomson ([18]) and Ruckdeschel ([26]), respectively. Afterwards, the hyper parameters of the filtered series are estimated again.

Those re-estimated hyper parameters are used to calculate a prewhitened spectral density estimate for each process. Last, the deviation of each estimated spectral density function from the true spectral density function is measured in the sense of the squared L_2-norm, i.e.,

$$err^2_{\widehat{S}(f)} := \big\|\widehat{S}(f) - S(f)\big\|^2 = \int \big(\widehat{S}(f) - S(f)\big)^2\, df \;, \tag{4.2}$$

where $\widehat{S}(f)$ and $S(f)$ denote the estimated and true spectral density functions.

5. RESULTS

Regarding the computation time the rLS filter performs better than the ACM-type filter as we expected. This is due to the fact that additional weights have to be computed within the correction step of the ACM-type filter.

Figure 3 tries to visualize the results of our simulation study. For both methods and contamination levels 0%, 10% and 20% seven curves are plotted on a logarithmic scale. The thick line represents the true spectral density function, whereas the thin line is the spectral density estimate of one realization out of 400. Moreover, we may calculate the minimum and maximum, at each frequency, the first and third quartile and median value of all spectral density estimates. Connecting all median values we obtain the grey line, to which we will refer hereafter as median spectral density function. In the same sense we refer to all minimum values as minimum spectral density function, and so on. Hence, the lower and upper dotted lines are the minimum and maximum spectral density functions, whereas the lower and upper dashed lines represent the first and third quartile spectral density functions. The results obtained by using the ACM-type filter are plotted in the left column, whereas the results of the rLS filter are displayed in the right column.

As expected, for both methods the dispersion of the spectral density estimates becomes greater the higher the contamination. However, this effect is more visible, especially at higher frequencies, when using the ACM-type filter.

Next, we try to visualize the squared errors of the estimated spectral density functions. First, the logarithm of the squared errors is taken. For both methods Figure 4 shows boxplots of the squared errors in eight equally-sized frequency bands as well as the total squared errors (bottom right) for all different levels of contamination. Again, the squared errors become greater the higher the contamination, especially at higher frequencies. And, this effect again is greater, when using the ACM-type filter. However, these errors are very small and, looking at the total squared errors for different contamination levels, we see that the ACM-type filter performs better than the rLS filter. The greatest contribution to the total squared error is the amount of the frequency band where the spectral density function has its peak. There the squared errors using the rLS filter are higher than the ones using the ACM-type filter. Moreover, we see that all squared errors are in the same range for all contamination levels.

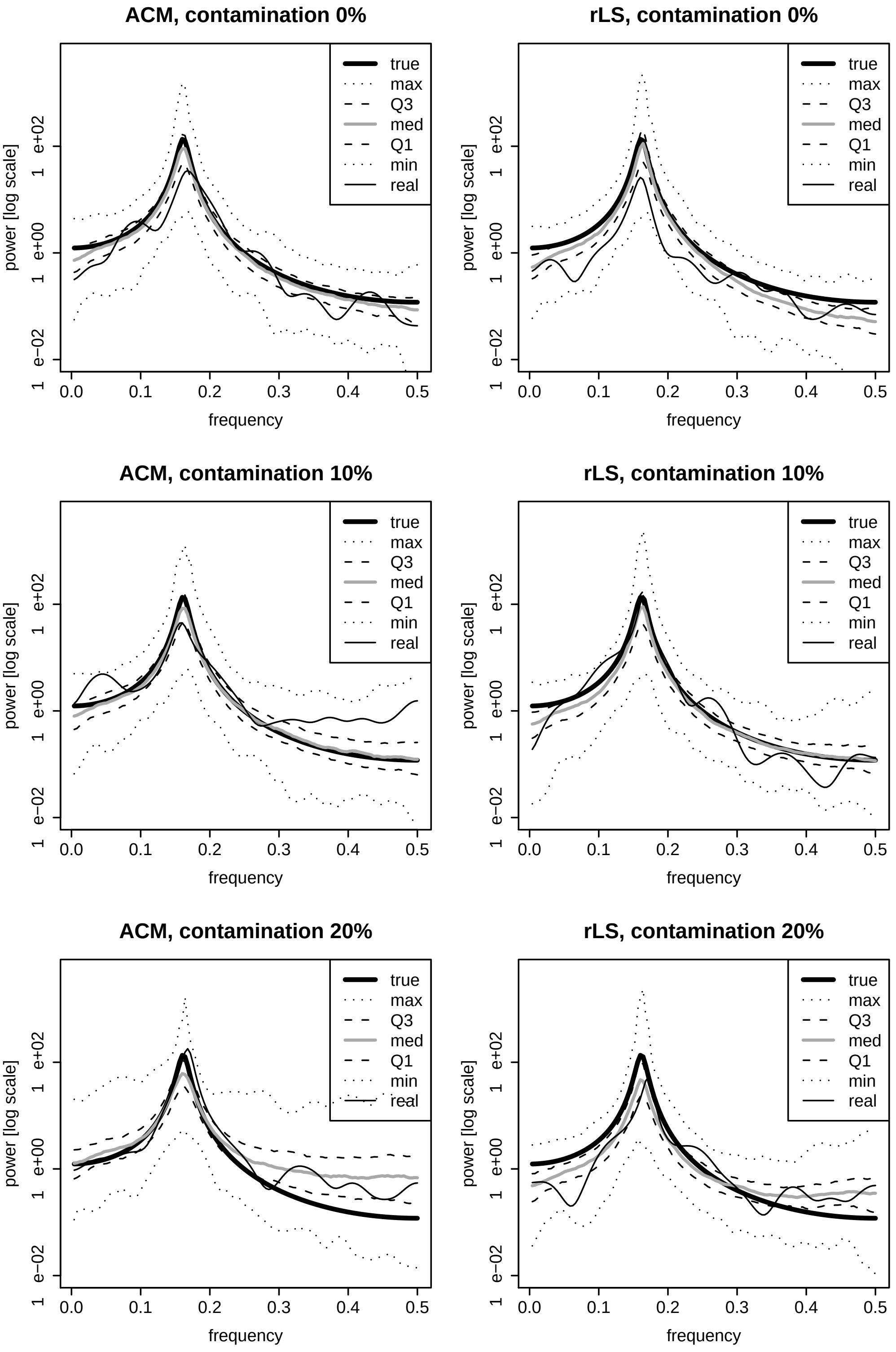

Figure 3: Robust spectral density estimates of the simulated data, left column 'ACM', right 'rLS'.

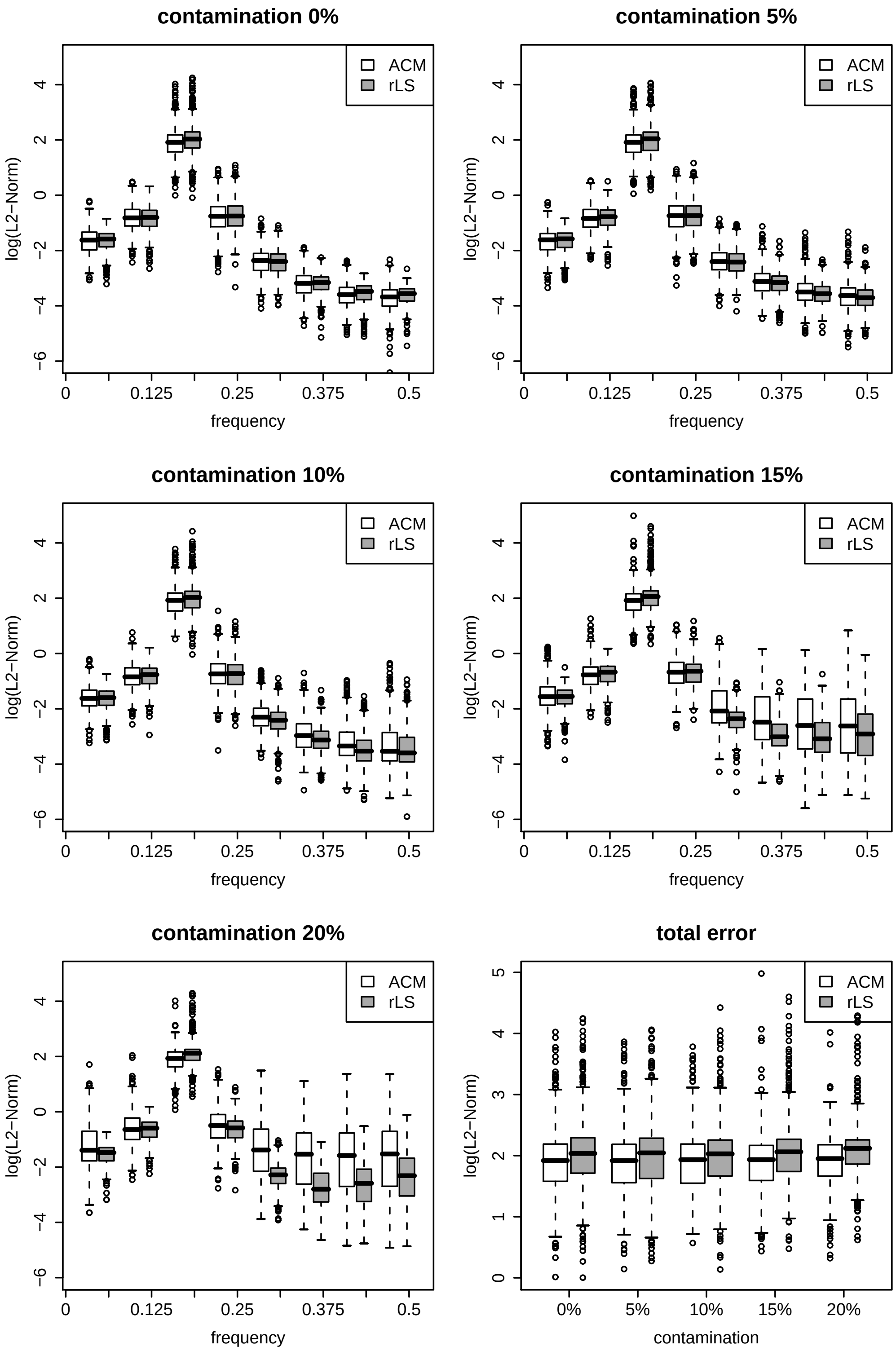

Figure 4: Boxplots of the errors.

6. DISCUSSION

In order to get a robust estimate of the spectral density function, it turns out that cleaning the series in a robust way first and calculating a prewhitened spectral density estimate afterwards leads to encouraging results. This data-cleaning operation wherein the robustness is introduced, is solved by two different robustified versions of the Kalman filter. Although, as far as we know, there exist no theoretical results on the statistical properties of both proposed multi-step procedures, the empirical results based on simulations and real data sets promise those procedures to be of high quality. The results of the simulation study suggest that the ACM-type filter algorithm performs slightly better than the rLS filter algorithm. Hence, the ACM-type filter algorithm was used to compute the robust spectral density estimates shown in Figure 2.

In [28] we compare the ACM-type filter approach with another approach proposed by Tatum and Hurvich ([30]). This procedure, called biweight filter-cleaner, also yields good results, but tends to underestimate the core process slightly. Moreover it is computational intensive.

The problem of estimating the hyper parameters was accomplished by bounded-influence autoregression. An alternative way would be to use a highly robust autocovariance function estimator (cf. [13]) and calculate estimates of the hyper parameters via the Yule-Walker equations. Hyper parameters may also be obtained by computing a robust covariance matrix via the MCD algorithm (cf. [24]) and estimate the parameters again using the Yule-Walker equations. Recently, Maronna et al. ([14]) propose to use τ-estimates. Our experience by now is that all these different approaches (except the last one, which we have not tried yet, although it seems worthwhile) leads to similar results.

The simulation study was only done for one specific autoregressive model of order 2. Other models seem worth trying. Further research and additional simulation studies have already been done, but, as well as the applications to the motivating real data, are not published here.

ACKNOWLEDGMENTS

The computation and all graphics were done with R (cf. [23]). The rLS filter algorithm was implemented by Peter Ruckdeschel. Moreover, an R-package, containing the BIAR algorithm, the ACM-type filter algorithm and the rLS filter algorithm, exists and is available at `www.uni-bayreuth.de/departments/math/org/mathe7/robKalman/`. We also like to thank two referees for their valuable suggestions.

REFERENCES

[1] Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, **AC-19**, 716–722.

[2] Anderson, B. and Moore, J. (1979). *Optimal Filtering*, Prentice Hall, Englewood Cliffs, NJ.

[3] Blackman, R. and Tukey, J. (1958). *The Measurement of Power Spectra*, Dover, New York.

[4] Chave, A.; Thomson, D. and Ander, M. (1987). On the robust estimation of power spectra, coherences, and transfer functions, *J. Geophys. Res.*, **92**(B1), 633–648.

[5] Durbin, J. and Koopman, S. (2001). *Time Series Analysis by State Space Methods*, Oxford University Press, New York.

[6] Fox, A. (1972). Outliers in time series, *J. Royal Statist. Soc.*, **34**(3), 350–363.

[7] Hampel, F. (1968). *Contributions to the Theory of Robust Estimation*, PhD thesis, University of California, Berkeley.

[8] Hartikainen, J.; Tahvanainen, K. and Kuusela, T. (1998). *Short-term measurement of heart rate variability.* In “Clinical Guide to Cardiac Autonomic Tests” (Malik, Ed.), Kluwer, Dordrecht, 149–176.

[9] Jones, A. and Hollinger, K. (1997). Spectral analysis of the KTB sonic and density logs using robust nonparametric methods, *J. Geophys. Res.*, **102**(B8), 18391–18403.

[10] Kalman, R. (1960). A new approach to linear filtering and prediction problems, *J. Basic Eng. – Trans. ASME*, **82**, 35–45.

[11] Kalman, R. and Bucy, R. (1961). New results in filtering and prediction theory, *J. Basic Eng. – Trans. ASME*, **83**, 95–108.

[12] Kleiner, R.; Martin, R. and Thomson, D. (1979). Robust estimation of power spectra, *J. Royal Statist. Soc. B*, **41**(3), 313–351.

[13] Ma, Y. and Genton, M. (2000). Highly robust estimation of the autocovariance function, *J. Time Series Analysis*, **21**(6), 663–684.

[14] Maronna, R.; Martin, R. and Yohai, V. (2006). *Robust Statistics: Theory and Methods*, John Wiley, New York.

[15] Martin, R. (1979). *Approximate conditional-mean type smoothers and interpolators.* In “Smoothing Techniques for Curve Estimation” (Gasser and Rosenblatt, Eds.), Springer, New York.

[16] Martin, R. (1980). *Robust estimation of autoregressive models.* In “Directions in Time Series” (D. Brillinger and G. Tiao, Eds.), Inst. Math. Statist. Publications, Haywood, CA, 228–254.

[17] Martin, R. (1981). *Robust methods for time series.* In “Applied Time Series II” (Findley, Ed.), Academic Press, New York.

[18] Martin, R. and Thomson, D. (1982). Robust-resistant spectrum estimation, *Proceedings of the IEEE*, **70**, 1097–1115.

[19] Martin, R. and Zeh, J. (1978). *Generalized M-estimates for autoregressions, including small-sample efficiency robustness*, Technical Report 214, Dept. of Electrical Engineering, Univ. Washington, Seattle.

[20] Masreliez, C. (1975). Approximate non-Gaussian filtering with linear state and observation relations, *IEEE Transactions on Automatic Control*, **AC-20**, 107–110.

[21] Moore, J. and Anderson, B. (1980). Coping with singular transition matrices in estimation and control stability theory, *Int. J. Control*, **31**, 571–586.

[22] Pumprla, J.; Howorka, K.; Groves, D.; Chester, M. and Nolan, J. (2002). Functional assessment of heart rate variability: physiological basis and practical applications, *Int. J. Cardiology*, **84**, 1–14.

[23] R Development Core Team (2005). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

[24] Rousseeuw, P. and Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator, *Technometrics*, **41**, 212–223.

[25] Ruckdeschel, P. (2000). *Robust kalman filtering.* In “XploRe. Application Guide” (Härdle, Hlávka and Klinke, Eds.), Springer, New York, Chapter 18, 483–516.

[26] Ruckdeschel, P. (2001). *Ansätze zur Robustifizierung des Kalman-Filters*, Volume 64 of *Bayreuther Mathematische Schriften*, Mathematisches Institut, Universität Bayreuth, Bayreuth, PhD thesis.

[27] Shumway, R. and Stoffer, D. (2000). *Time Series Analysis and Its Applications*, Springer, New York.

[28] Spangl, B. and Dutter, R. (2005). On robust estimation of power spectra, *Austrian Journal of Statistics*, **34**(2), 199–210.

[29] Stockinger, N. and Dutter, R. (1987). Robust time series analysis: A survey, *Kybernetika*, Supplement **23**, 1–90.

[30] Tatum, L. and Hurvich, C. (1993). *A frequency domain approach to robust time series analysis.* In “New Directions in Statistical Data Analysis and Robustness” (Morgenthaler, Ronchetti and Stahel, Eds.), Birkhäuser-Verlag, Basel.

[31] Thomson, D. (1994). An overview of multiple-window and quadratic-inverse spectrum estimation methods, *Proceedings of the IEEE ICASSP*, **6**, 185–194.

REVSTAT – Statistical Journal
Volume 5, Number 1, March 2007, 63–83

COMPARATIVE PERFORMANCE OF SEVERAL ROBUST LINEAR DISCRIMINANT ANALYSIS METHODS [*]

Authors: Valentin Todorov
– Austro Control GmbH,
Vienna, Austria
valentin.todorov@chello.at

Ana M. Pires
– Departamento de Matemática and CEMAT, Instituto Superior Técnico,
Technical University of Lisbon (TULisbon), Portugal
apires@math.ist.utl.pt

Abstract:

- The problem of the non-robustness of the classical estimates in the setting of the quadratic and linear discriminant analysis has been addressed by many authors: Todorov *et al.* [19, 20], Chork and Rousseeuw [1], Hawkins and McLachlan [4], He and Fung [5], Croux and Dehon [2], Hubert and Van Driessen [6]. To obtain high breakdown these methods are based on high breakdown point estimators of location and covariance matrix like MVE, MCD and S. Most of the authors use also one step re-weighting after the high breakdown point estimation in order to obtain increased efficiency. We propose to use M-iteration as described by Woodruff and Rocke [22] instead, since this is the preferred means of achieving efficiency with high breakdown. Further we experiment with the pairwise class of algorithms proposed by Maronna and Zamar [10] which were not used up to now in the context of discriminant analysis. The available methods for robust linear discriminant analysis are compared on two real data sets and on a large scale simulation study. These methods are implemented as R functions in the package for robust multivariate analysis *rrcov*.

Key-Words:

- *discriminant analysis; robustness; MCD; S-estimates; M-estimates; R.*

AMS Subject Classification:

- 62G35, 62H30.

[*] The presentation of material in this article does not imply the expression of any opinion whatsoever on the part of any organization and is the sole responsibility of the authors.

1. INTRODUCTION

The problem of discriminant analysis arises when one wants to assign an individual to one of g populations on the basis of a p-dimensional feature vector $\boldsymbol{x}$. Usually it is considered that the p-dimensional vectors $\boldsymbol{x}_{ik}$ come from multivariate normal populations π_k

$$(1.1)\qquad \boldsymbol{x}_{ik}\colon\ \pi_k \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \qquad (i=1,...,n_k;\ \ k=1,...,g)\ .$$

Here n_k is the size of the sample from population k for each of the g different groups. If it is further assumed that all covariance matrices are equal ($\boldsymbol{\Sigma}_1 = ... = \boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$), the overall probability of misclassification is minimized by assigning a new observation $\boldsymbol{x}$ to population π_k which maximizes

$$(1.2)\qquad d_k(\boldsymbol{x}) \;=\; \frac{1}{2}\,(\boldsymbol{x}-\boldsymbol{\mu}_k)^t\,\boldsymbol{\Sigma}^{-1}\,(\boldsymbol{x}-\boldsymbol{\mu}_k) + \log(\alpha_k) \qquad (k=1,...,g)\ ,$$

where α_k is the prior probability that an individual comes from population π_k. If the means $\boldsymbol{\mu}_k$, $k=1,...,g$, and the common covariance matrix $\boldsymbol{\Sigma}$ are unknown, which is usually the case, a training set consisting of samples drawn from each of the populations is required.

The problem of the non-robustness of the classical estimates in the setting of the quadratic and linear discriminant analysis has been addressed by many authors: Todorov *et al.* [19, 20], replaced the classical estimates by MCD estimates; Chork and Rousseeuw [1] used MVE instead; Hawkins and McLachlan [4] defined the Minimum Within Covariance Determinant estimator (MWCD) especially for the case of linear discriminant analysis; He and Fung [5] and Croux and Dehon [2] used S estimates; Hubert and Van Driessen [6] applied the MCD estimates computed by the FAST MCD algorithm.

Most of the authors use one step re-weighting after the high breakdown point estimation in order to obtain increased efficiency. We propose to use M-iteration as described by Woodruff and Rocke [22] instead, since this is the preferred means of achieving efficiency with high breakdown and the time necessary for the M-iteration is negligible when compared to the time necessary for the MCD estimation, even using the FAST-MCD algorithm. Further we want to experiment with the pairwise class of algorithms proposed by Maronna and Zamar [10] which have not been used up to now in the context of discriminant analysis.

In most of the cited papers, apart from the theoretical results, the proposed methods are illustrated on one or two data sets and only a limited simulation is performed, i.e. only a few contamination configurations are used and the new method is compared to one or two of the already known ones on the basis of these configurations. Todorov *et al.* [20] carried out a more extended simulation, using a general model and varying a number of parameters but this study was restricted only to scale contaminations of the training samples in case of two groups.

The purpose of this work is to review the recent results in robust linear discriminant analysis and to compare the available methods on a large scale simulation study. The discriminant analysis is considered in a prediction context and the performance of the discrimination rules is evaluated by misclassification probabilities obtained by simulation.

The paper is organized as follows. In the next section we describe the robust linear discriminant analysis methods used. In Section 3 we illustrate the application of these methods with two real data sets. In Section 4 we describe the simulation study and present the results. The paper ends with a brief summary and conclusions. The discussed methods for robust linear discriminant analysis are implemented as R functions in the package for robust multivariate analysis *rrcov*.

2. ROBUST ESTIMATORS FOR LINEAR DISCRIMINANT ANALYSIS

In order to obtain a robust procedure with high breakdown point for linear discriminant analysis the classical estimators are replaced by different robust estimators. To overcome the low efficiency of the most high breakdown point estimators, their reweighted version is used.

The Minimum Covariance Determinant (MCD) Estimator introduced by Rousseeuw [16] looks for a subset of h observations whose covariance matrix has the lowest determinant. The MCD location estimate $\boldsymbol{T}$ is defined as the mean of that subset and the MCD scatter estimate $\boldsymbol{C}$ is a multiple of its covariance matrix. The multiplication factor is selected so that $\boldsymbol{C}$ is consistent at the multivariate normal model and unbiased at small samples — see Pison and Willems [11]. This estimator is not very efficient at normal models, especially if h is selected so that maximal breakdown point is achieved, but in spite of its low efficiency it is the mostly used robust estimator in practice, mainly because of the existing efficient algorithm for computation as well as the readily available implementations in most of the well known statistical software packages like R, S-Plus, SAS and Matlab.

We start by finding initial estimates of the group means $\boldsymbol{m}_k^0$ and the common covariance matrix $\boldsymbol{C}_0$ based on the reweighted MCD estimates. There are several methods for estimating the common covariance matrix based on a high breakdown point estimator.

The easiest one is to obtain the estimates of the group means and group covariance matrices from the individual groups $(\boldsymbol{m}_k, \boldsymbol{C}_k)$, $k=1,...,g$, and then pool them to yield the common covariance matrix

$$\boldsymbol{C} = \frac{\sum_{k=1}^{g} n_k \boldsymbol{C}_k}{\sum_{k=1}^{g} n_k - g} \ . \tag{2.1}$$

This method, using MVE and MCD estimates, was proposed by Todorov *et al.* [19] and [20] and was also used, based on the MVE estimator by Chork and Rousseeuw [1]. Croux and Dehon [2] applied this procedure for robustifying linear discriminant analysis based on S estimates. A drawback of this method is that the same trimming proportions are applied to all groups which could lead to a loss of efficiency if some groups are outlier free. We will denote this method as A and the corresponding estimator as XXX-A. For example in the case of the MCD estimator this will be MCD-A.

Another method was proposed by He and Fung [5] for the S estimates and was later adapted by Hubert and Van Driessen [6] for the MCD estimates. Instead of pooling the group covariance matrices, the observations are centered and pooled to obtain a single sample for which the covariance matrix is estimated. It starts by obtaining the individual group location estimates $\boldsymbol{t}_k$, $k=1,...,g$, as the reweighted MCD location estimates of each group. These group means are swept from the original observations to obtain the centered observations

$$\boldsymbol{Z} = \{\boldsymbol{z}_{ik}\} \ , \qquad \boldsymbol{z}_{ik} = \boldsymbol{x}_{ik} - \boldsymbol{t}_k \ . \tag{2.2}$$

The common covariance matrix $\boldsymbol{C}$ is estimated as the reweighted MCD covariance matrix of the centered observations $\boldsymbol{Z}$. The location estimate $\boldsymbol{\delta}$ of $\boldsymbol{Z}$ is used to adjust the group means $\boldsymbol{m}_k$ and thus the final group means are

$$\boldsymbol{m}_k = \boldsymbol{t}_k + \boldsymbol{\delta} \ . \tag{2.3}$$

This process could be iterated until convergence, but since the improvements from such iterations are negligible (see [5], [6]) we are not going to use it. This method will be denoted by B and as already mentioned, the corresponding estimator as XXX-B, for example MCD-B.

The third approach is to modify the algorithm for high breakdown point estimation itself in order to accommodate the pooled sample. He and Fung [5] modified Ruperts's SURREAL algorithm for S estimation in case of two groups. Hawkins and McLachlan [4] defined the Minimum Within-group Covariance Determinant estimator (MWCD) which does not apply the same trimming proportion to each group but minimizes directly the determinant of the common within groups covariance matrix by pairwise swaps of observations. Unfortunately their estimator is based on the Feasible Solution Algorithm (see [4] and the references therein), which is extremely time consuming as compared to the FAST-MCD algorithm. Hubert and Van Driessen [6] proposed a modification of this algorithm taking advantage of the FAST-MCD, but it is still necessary to compute the MCD for each individual group. This method will be denoted by MCD-C.

Using the estimates $\boldsymbol{m}_k^0$ and $\boldsymbol{C}_0$ obtained by one of the methods, we can calculate the initial robust distances (Rousseeuw and van Zomeren [17])

$$RD_{ik}^0 = \sqrt{(\boldsymbol{x}_{ik} - \boldsymbol{m}_k^0)^t\, \boldsymbol{C}_0^{-1}\, (\boldsymbol{x}_{ik} - \boldsymbol{m}_k^0)}\ . \tag{2.4}$$

With these initial robust distances we can define a weight for each observation $\boldsymbol{x}_{ik}$, $i=1,...,n_k$ and $k=1,...,g$, by setting the weight to 1 if the corresponding robust distance is less or equal to a suitable cut-off, usually $\sqrt{\chi^2_{p,0.975}}$, and to 0 otherwise, i.e.

$$w_{ik} = \begin{cases} 1 & RD_{ik}^0 \le \sqrt{\chi^2_{p,0.975}} \\ 0 & \text{otherwise}\ . \end{cases} \tag{2.5}$$

With these weights we can calculate the final reweighted estimates of the group means, $\boldsymbol{m}_k$, and the common within-groups covariance matrix, $\boldsymbol{C}$, which are necessary for constructing the robust classification rules,

$$\begin{aligned} \boldsymbol{m}_k &= \left(\sum_{i=1}^{n_k} w_{ik}\, \boldsymbol{x}_{ik}\right) \Big/ \nu_k\ , \\ \boldsymbol{C} &= \frac{1}{\nu - g} \sum_{k=1}^{g} \sum_{i=1}^{n_k} w_{ik}\, (\boldsymbol{x}_{ik} - \boldsymbol{m}_k)\, (\boldsymbol{x}_{ik} - \boldsymbol{m}_k)^t\ , \end{aligned} \tag{2.6}$$

where ν_k are the sums of the weights within group k, for $k=1,...,g$, and ν is the total sum of weights,

$$\nu_k = \sum_{i=1}^{n_k} w_{ik}\ , \qquad \nu = \sum_{k=1}^{g} \nu_k\ .$$

Table 1 summarizes the methods to be considered in this study. It has already been shown by simulations that the reweighted versions of most of the estimators, at least in the case of one sample, are by far more efficient. This has also been shown for the common covariance matrix in the framework of linear discriminant analysis for the S estimates by He and Fung [5] and for the MCD estimates by Hubert and Van Driessen [6]. Therefore in the following sections we will prefer the reweighted estimates whenever possible without explicitly mentioning this.

Some of the methods are extremely slow which to some extent prevented us from performing the complete simulation on them. These are particularly the MWCD of Hawkins and McLachlan [4] and the S-estimates computed by

Ruppert's SURREAL algorithm. The FAST S algorithm, whose implementation is similar to the one proposed by Salibian-Barrera and Yohai [18] for the case of regression is promising, but since the available implementation is in pure R, it cannot compete with MCD (in FORTRAN) and OGK, for example. A C or FORTRAN implementation of this algorithm will allow its more frequent use. Note also that, because of the large amount of results, not all of them can be reported here.

Table 1: Estimators for the group means and the common covariance matrix which will be considered in this study.

Algorithm	Comment
FSA	Minimum Within-group Covariance Determinant estimator [4] computed by the FSA algorithm
MCD-A	method A MCD
MCD-B	method B MCD
MCD-C	method C MCD
M-tb	M estimator with translated biweight function [15]
M-bw	M estimator with biweight function [15]
OGK	Pairwise estimators — [10] (method B)
S	S estimates computed by Ruppert's SURREAL
Sfast	S estimates computed by the fast algorithm proposed for regression by [18] (method B)

3. EXAMPLES

3.1. The Fish catch data

As a first example for illustration of the robust approach to linear discriminant analysis we use a data set containing measurements on 159 fish caught in the lake Laengelmavesi, Finland. The data set is available from [12]. It is also included in the R package *rrcov* — see Todorov [21]. For the 159 fishes of 7 species the weight, length, height, and width were measured. Three different length measurements are recorded: from the nose of the fish to the beginning of its tail, from the nose to the notch of its tail and from the nose to the end of its tail. The height and width are calculated as percentages of the third length variable. This results in 6 observed variables, listed in Table 2. Observation 14 has a missing value in variable Weight, therefore this observation was excluded from the analysis. The 7 species are listed in Table 3. The last column of this table gives the number of observations in each class. In the six dimensional problem presented by this data set, classes 2 (with 6 observations) and 4 (with 11 observations) will cause a problem to the half-sample based robust methods. Therefore we will consider three cases: (i) all 7 classes, (ii) 6 classes, with class 2 removed and (iii) 5 classes, with classes 2 and 4 removed.

Table 2: Fish measurements data: Variables.

1	Weight	Weight of the fish (in grams)
2	Length1	Length from the nose to the beginning of the tail (in cm)
3	Length2	Length from the nose to the notch of the tail (in cm)
4	Length3	Length from the nose to the end of the tail (in cm)
5	Height%	Maximal height as % of Length3
6	Width%	Maximal width as % of Length3

Table 3: Fish measurements data: Names of the species in Finish and English. The last column shows the number of objects in each class.

	Finish	English	#
1	Lahna	Bream	34
2	Siika	Whitewish	6
3	Saerki	Roach	20
4	Parkki	Parkki	11
5	Norssi	Smelt	14
6	Hauki	Pike	17
7	Ahven	Perch	56

In order to evaluate and compare the considered linear discriminant rules we have to determine their performance in the classification of future observations, i.e. we need an estimate of the overall probability of misclassification. A number of methods to estimate this probability exist in the literature — see for example Lachenbruch [7]. The *apparent error rate* (known also as resubstitution error rate or reclassification error rate) is the most straightforward estimator of the actual (true) error rate in discriminant analysis and is calculated by applying the classification criterion to the same data set from which it was derived and then counting the number of misclassified observations. It is well known that this method is too optimistic (the true error is likely to be higher). If there are plenty of observations in each class the error rate can be estimated by splitting the data into training and validation sets. The first one is used to estimate the discriminant rules and the second to estimate the misclassification error.

This method is fast and easy to apply but it is wasteful of data which would be critical in our case. Another method is the *leaving-one-out* or the *cross-validation* method (Lachenbruch and Michey [8] which proceeds by removing one observation from the data set, estimating the discriminant rule using the remaining $n-1$ observations and than classifying this observation with the estimated discriminant rule. For the classical linear discriminant analysis there exist updating formulas which avoid the recomputation of the discriminant rule at each step but no such formulas are available for the robust methods. Thus the estimation of the error rate by this method can be very time consuming depending on the size of the data set. Nevertheless, for the sake of our example, we will afford the time and will use the leaving-one-out method to evaluate the considered discriminant rules. Table 4 shows the results. The apparent error rate is also computed and given for comparison.

Table 4: Fish measurements data: Apparent Error rate (APR) and Leaving-One-Out (CV) estimate of the error rate for the classical (MLE) and eight robust discriminant rules.

Method	**All Classes**		**Without 2**		**Without 2 and 4**	
	APR	CV	APR	CV	APR	CV
MLE	0.0127	0.0190	0.0132	0.0132	0.0142	0.0142
FSA	0.0949	0.1139	0.0197	0.0197	0.0142	0.0142
MCD-A	—	—	—	—	0.0851	0.0780
MCD-B	—	—	—	—	0.0638	0.0638
MCD-C	—	—	—	—	0.0496	0.0451
M-tb	—	—	—	—	0.0071	0.0142
M-bw	—	—	—	—	0.0142	0.0142
S	—	—	0.0132	0.0132	0.0142	0.0142
OGK	0.0126	0.0696	0.0066	0.0132	0.0142	0.0142

For the complete data set, apart from the MLE estimates, we could compute only the FSA and OGK which do not need a half-sample based estimates of each group. The estimated error rates (0.1139 and 0.0696 respectively) are higher than the error rate for MLE — 0.0190. If we remove class 2 which has only six observations, it is possible to compute also the S estimates. Now only FSA has slightly higher error rate, while the other rules (MLE, S and OGK) give the same (cross-validation) error rate of 0.0132. After removing also class 4 with only 11 observations all robust estimates are available. The MLE discriminant rule as well as most of the robust rules give the same error rate of 0.0142 and only the three versions based on FAST-MCD give somewhat higher values. As expected, in general the apparent error rate is lower than the leaving-one-out estimate.

Since there is no difference in the estimated error rates, it seems that both robust and non-robust methods perform equally well on this data set. As already noted by Hawkins and McLachlan [4] this does not mean that robust methods are not necessary, but on the contrary, this means that the robust methods, while providing safeguard against possible outliers in the data, do not perform worse when the data are outlier-free.

3.2. The Diabetes data

As a second example, we use the Diabetes Data, which was analyzed by [13] in an attempt to examine the relationship between chemical diabetes and overt diabetes in 145 nonobese adult subjects. The analysis was focused on three primary variables and the 145 individuals were classified initially on the basis of their plasma glucose levels into three groups: normal subjects, chemical diabetes and overt diabetes. This data set was also analyzed by [4] in the context of the robust linear discriminant analysis. The data set is available in several R packages: *diabetes* in package *mclust*, *chemdiab* in package *locfit* and *diabetes.dat* in *Rfwdmv*. We used the first one for which the value of the second variable, *insulin*, on the 104-th observation, is 45 while for the other data sets this value is 455 (note that 45 is more likely to be an outlier in this variable than 455). As in the first example, the discriminant rules based on MLE and the eight robust methods were applied. The corresponding apparent error rates and the leaving-one-out estimates of the error rate are shown in Table 5.

Table 5: Diabetes data: Apparent Error rate (APR) and Leaving-One-Out (CV) estimate of the error rate for the classical (MLE) and eight robust discriminant rules. The last two columns give the error rate estimates for the raw (not reweighted) methods.

Method	**Reweighted**		**Raw**	
	APR	CV	APR	CV
MLE	0.1310	0.1310	—	—
FSA	0.0483	0.0552	0.0621	0.0552
MCD-A	0.1241	0.1379	0.1379	0.1379
MCD-B	0.1034	0.1172	0.0966	0.1172
MCD-C	0.0699	0.0802	0.0965	0.0803
M-tb	0.0965	0.1103	—	—
M-bw	0.1034	0.1172	—	—
S	0.0965	0.1034	0.1034	0.1034
OGK	0.0689	0.1103	0.1034	0.1034

All the robust methods identify the outliers and show smaller error rates than the MLE discriminant rule. The FSA estimator performs best followed by MCD-C (i.e. the FAST-MCD analogue of MWCD as defined by Hubert and Van Driessen [6]). Table 5 also shows the results of the raw (not-reweighted) estimates but for this data set they differ only slightly from the reweighted ones.

4. SIMULATION

4.1. Distributions

The estimators considered will be evaluated on data sets generated from a variety of settings with different dimensions $p = 2, 6, 10$, different number of groups $g = 2, 3$ and different size of the training samples $n = \sum_{j=1}^{g} n_j$. In all cases the class distributions are normal, but the generated data sets differ in the shapes of the group populations and in the separation between the means of the groups. The various combinations of the parameters of these classification problems were to some extent motivated by the studies performed by Friedman [3] to test his regularized discriminant analysis method. These data structures are denoted by **Di** and are the following:

- **D1**. *Equal spherical covariance matrices.* In this situation all groups π_j, $j = 1, ..., g$, have the same spherical covariance matrix $\boldsymbol{I}_p$. The mean of the first group is the origin, the mean of the second group is at distance $d = 3.0$ and the mean of the third group is at the same distance $d = 3.0$, but in an orthogonal direction. More precisely, the data sets are generated from the following p-dimensional normal distributions, where each group π_j, $j = 1, ..., g$, has a separate mean $\boldsymbol{\mu}_j$ and all of them have the same covariance matrix $\boldsymbol{I}_p$,

$$\pi_j \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \ , \qquad j = 1, ..., 3 \ , \tag{4.1}$$

with

$$\begin{aligned} \boldsymbol{\mu}_1 &= (0, 0, ..., 0) \ , \\ \boldsymbol{\mu}_2 &= (3, 0, ..., 0) \ , \\ \boldsymbol{\mu}_3 &= (0, 3, ..., 0) \ , \\ \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3 &= \boldsymbol{I}_p \ . \end{aligned}$$

These distributions will be contaminated in the following two ways

- scale contamination:

$$\pi_j \sim (1-\varepsilon)\, N_p(\boldsymbol{\mu}_j, \boldsymbol{I}_p) + \varepsilon\, N_p(\boldsymbol{\mu}_j, \kappa \boldsymbol{I}_p)\ , \qquad j=1,...,g\ , \tag{4.2}$$

- location contamination

$$\pi_j \sim (1-\varepsilon)\, N_p(\boldsymbol{\mu}_j, \boldsymbol{I}_p) + \varepsilon\, N_p(\hat{\boldsymbol{\mu}}_j, 0.25^2 \boldsymbol{I}_p)\ , \qquad j=1,...,g\ , \tag{4.3}$$

$$\begin{aligned} \hat{\boldsymbol{\mu}}_j &= \boldsymbol{\mu}_j + (\nu Q_p, ..., \nu Q_p)\ , \\ Q_p &= \sqrt{\chi^2_{p;0.001}/p}\ , \end{aligned}$$

where $\varepsilon = \{0, 0.1, 0.25, 0.4\}$, $\kappa = \{9, 100\}$, and $\nu = 5, 10$ are parameters of the simulation. The shift of the location outliers is measured in terms of the unit measure $Q = \sqrt{\chi^2_{p;0.001}}$. The outliers are placed at distance νQ by adding νQ_p to each component of the location vector $\boldsymbol{\mu}$, where $Q_p = \sqrt{\chi^2_{p;0.001}/p}$ (see Rocke and Woodruff [15]).

The variation of the parameters g, p, n, ε, ν and κ results in 234 data distributions (18 uncontaminated, 108 location contaminated and 108 scale contaminated).

- **D2**. *Unequal spherical covariance matrices.* In this situation each group π_j, $j=1,...,g$, has a spherical covariance matrix $j\,\boldsymbol{I}_p$, i.e. the first group has as covariance matrix the identity matrix $\boldsymbol{I}_p$ and the covariance matrix of each other group is a multiple of the identity matrix $\boldsymbol{I}_p$ with inflation factor equal to the number of the group. The mean of the first group is the origin, the mean of the second is at distance $d = 3.0$ as in the situation **D1** and the mean of the third is at distance $d = 4.0$, but in an orthogonal direction. The data sets in this situation are generated from the distributions given in equation (4.1), where each group π_j, $j=1,...,g$, has a separate mean $\boldsymbol{\mu}_j$ and their covariance matrices $\boldsymbol{\Sigma}_j$ are spherical and proportional,

$$\begin{aligned} \boldsymbol{\mu}_1 &= (0,0,0,...,0)\ , \\ \boldsymbol{\mu}_2 &= (3,0,0,...,0)\ , \\ \boldsymbol{\mu}_3 &= (0,4,0,...,0)\ , \\ \boldsymbol{\Sigma}_1 &= \boldsymbol{I}_p\ , \\ \boldsymbol{\Sigma}_2 &= 2\,\boldsymbol{I}_p\ , \\ \boldsymbol{\Sigma}_3 &= 3\,\boldsymbol{I}_p\ . \end{aligned}$$

Only location contamination will be applied to these distributions, as described in equation (4.3). This results in altogether 136 data distributions (18 uncontaminated and 108 location contaminated).

- **D0**. In order to calibrate the simulation we will start with the same configurations as described by He and Fung [5] and then repeated by Croux and Dehon [2] and later by Hubert and Van Driessen [6]. In these classification problems data were generated from two groups $g=2$ with $p=3$ and different sizes of the training samples. In most of the cases the populations have the same spherical covariance matrices $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}_3$. The mean of the first group is the origin, $\boldsymbol{\mu}_1 = (0,0,0)$, the mean of the second group is $\boldsymbol{\mu}_2 = (1,1,1)$. Location and scale contaminations are applied as described in **D1**. More precisely, the following five data structures are used.

  - **A**: $n_1 = n_2 = 50$, $\varepsilon = 0$, no contamination.
  - **B**: $n_1 = n_2 = 50$, $\varepsilon = 0.2$, location contamination with $\hat{\boldsymbol{\mu}}_1 = (5,5,5)$ and $\hat{\boldsymbol{\mu}}_2 = (-4,-4,-4)$.
  - **C**: $n_1 = 100$, $n_2 = 10$, $\varepsilon = 0.2$, location contamination with $\hat{\boldsymbol{\mu}}_1 = (5,5,5)$ and $\hat{\boldsymbol{\mu}}_2 = (-4,-4,-4)$.
  - **D**: $n_1 = n_2 = 20$, $\varepsilon = 0.2$, scale contamination with $\kappa = 25$.
  - **E**: $n_1 = 70$, $n_2 = 30$, $\varepsilon = 0.2$, unequal covariance matrices $\boldsymbol{\Sigma}_1 = \boldsymbol{I}_3$ and $\boldsymbol{\Sigma}_2 = 4\,\boldsymbol{I}_3$, location contamination with $\hat{\boldsymbol{\mu}}_1 = (5,5,5)$ and $\hat{\boldsymbol{\mu}}_2 = (-10,-10,-10)$.

4.2. Criteria

The described linear discriminant analysis estimators can be evaluated with regard to the following two aspects of the discriminant analysis:

- the quality of the estimates of the group means and the common covariance matrix and thus the quality of the discriminant functions and scores and
- in a prediction context the performance of the discrimination rules evaluated by their misclassification probabilities obtained by simulation.

Although the quality of the estimates is important since it entirely determines the robustness of the discriminant rules towards outliers, in this study we will concentrate only on the second aspect, the evaluation of the classification performance of the rules, and leave the detailed study of the estimates for further work.

The discrimination performance of the estimated classification rules is evaluated by the Overall Probability of Misclassification (OPM) which can be estimated by simulation (similar as in He and Fung [5] and Hubert and Van Driessen [6]). For this purpose we generate a test sample consisting of 2000 observations from each group (with the known distribution), classify them using each of the estimated discrimination rules and obtain the corresponding proportions of the misclassified observations. This procedure is repeated $N = 100$ times and the mean and standard error of the probability of misclassification are calculated for each method. Whenever possible, the robust estimates are based on one-step reweighting.

4.3. Simulation results

In this section we present the results of the simulation study for the robust linear discriminant rules as well as the classical MLE one. Also the results of MLE applied to the clean data are shown and are denoted by MLE-C.

Table 6: Simulation results: Mean probability of misclassification for the classical and robust estimators under different cases of contaminated distributions, as described in He and Fung [5].

	Estimators								
	FSA	M-bw	MCD-A	MCD-B	MCD-C	OGK	S	MLE-C	MLE
A	0.202	0.202	0.203	0.203	0.208	0.202	0.203	0.202	0.202
B	0.207	0.207	0.209	0.203	0.208	0.211	0.202	0.202	0.661
C	0.210	0.263	0.217	0.222	0.218	0.215	0.213	0.211	0.617
D	0.240	0.221	0.226	0.223	0.225	0.227	0.219	0.215	0.260
E	0.462	0.283	0.285	0.283	0.291	0.289	0.279	0.277	0.558

First we consider the results of the simulation study following He and Fung [5]. Table 6 shows the estimated overall probability of misclassification for the discriminant rules in the five different distribution setups. Note that the S estimator, computed by the method B which we are using in this study is equivalent to the estimator denoted by S2A in [5]. In the case of clean data — setup **A** — all estimators perform almost equally well. In case **B** with 20% location contamination the MLE breaks down, but all the robust estimators perform equally well, following closely MLE-C. In the third case, with unequal sample sizes — setup **C** — the robust estimators are worse than MLE-C, although they do better than MLE. The best are FSA, OGK and S. In case **D**, with 20% scale contamination, S and M-bw perform best. In the last case, with unequal sample sizes per group, $n_1 = 70$ and $n_2 = 30$, and unequal covariance matrices most of the robust estimators also perform similar to MLE-C, only FSA breaks down. There is no estimator which performs best in all cases but S and M-bw are the ones that perform best in most of the cases being almost equally good (however, M-bw is much faster taking advantage of the existing fast algorithm for MCD). The next best estimator is OGK and it is even faster than MCD.

As far as the main simulation is concerned, let us start by investigating the case of clean data — i.e. $\varepsilon = 0$ — and consider the dependence of the estimators on the data dimension and the sample size. Figure 1 presents the mean overall probability of misclassification of the classical and robust discriminant rules when applied to clean data (the results for S and M-tb are not presented, since they are almost the same as those for M-bw). For $n_1 = n_2 = 50$ and $n_1 = n_2 = 100$ all robust estimators follow closely the MLE. For $n_1 = n_2 = 20$ only the smooth estimators and OGK are very near to the MLE. No substantial difference between two and three groups can be noted.

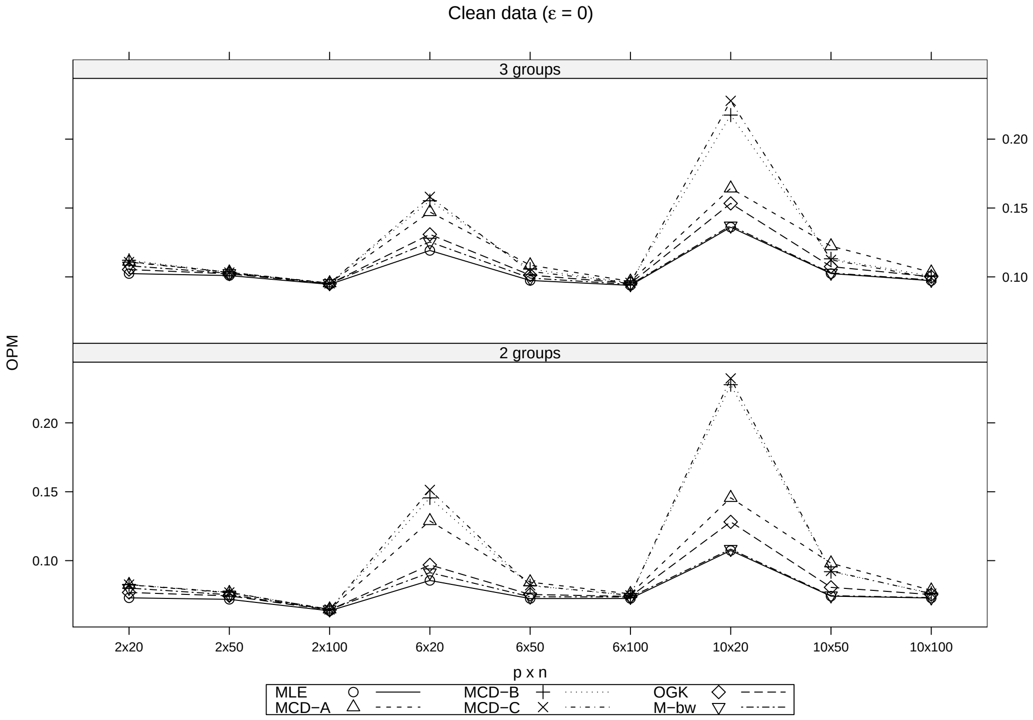

Figure 1: Mean Overall probability of misclassification for distribution setup $D1$ without contamination for different dimensions and sample sizes in case of 2 and 3 groups.

The next three tables display the estimated overall probability of misclassification (OPM) as a function of the contamination proportion ε in different simulation situations and different types of contamination. First we will consider the performance of the estimators in the case of scale contamination. Table 7 shows some of the results for two and three groups, where $\varepsilon = 0$, 10, 25 and 40% scale contamination is added to all groups with scale inflation factor $\kappa = 9$ and 100 respectively. Only the reweighted estimators are shown — the raw versions were slightly worse. Only the constrained M-estimates with Tukey's biweight function (M-bw) is shown, since it was slightly better than the M-estimates with the translated biweight function (Rocke [14]) in almost all of the cases. The results for the S estimates, as described by He and Fung [5] where these estimates are denoted S2A, were not computed for all cases because of their computational restrictions (the results actually computed are quite similar to those for M-bw). The S estimates computed by an algorithm similar to the one proposed by Salibian-Barrera and Yohai [18] for regression which they call *Fast S* are quite promising, but the available R implementation is rather slow (comparable with Rupperts SURREAL). A native implementation in C is under development. In the right hand part of the table, representing the results for 3 groups only MLE, M-bw and OGK are shown.

For $g = 2$, $p = 2$ and $n_1 = n_2 = 20$ in the case $\kappa = 9$ there is only slight gain in performance and only OGK and M-bw are better than the classical MLE estimates. The picture changes completely when $\kappa = 100$ where all robust methods show similar discrimination performance (closely following MLE-C). When the sample size is increased to $n_1 = n_2 = 50, 100$ the performance of MLE improves, but so does the performance of the robust estimators. When the number of variables p is increased ($p = 6, 10$), keeping the same sample size $n_1 = n_2 = 20$ only OGK remains usable, which is not surprising since all other robust estimators are based on a half sample. When the sample size is increased to $n_1 = n_2 = 50, 100$, all robust estimators perform very well again. In the case of three groups the results (shown in the right three columns of Table 7) are quite similar to the two-group situation, but the estimated overall probability of misclassification is slightly higher for all estimators, including MLE.

Next we will consider the performance of the estimators in case of location contamination. Table 8 shows some of the results for two and three groups, where $\varepsilon = 0$, 10, 25 and 40% location contamination is added to all groups with location shift parameter $\nu = 5$ and 10 respectively. The "uninteresting" cases where the dimension is high compared to the sample size, and to which we know that the robust estimators cannot be applied, are not shown. OGK preforms best in almost all cases except for 40% contamination with shift factor $\nu = 5$, where it always breaks down.

Table 9 shows some of the results for two and three groups for distribution setup **D2** (unequal spherical covariance matrices) when location contamination is added to the data. The situation is quite similar to the case of equal spherical covariance matrices, but the estimated probability of misclassification increases for all estimators including MLE-C.

Table 7: Mean probability of misclassification for Setup D1 with scale contamination in the case of two and three groups (for three groups not all of the estimators are shown).

ε	κ	MLE	M-bw	MCD-A	MCD-B	MCD-C	OGK	MLE	M-bw	OGK
		$p = 2, n_1 = n_2 = 20, g = 2$						$g = 3$		
0	—	0.073	0.084	0.082	0.083	0.083	0.077	0.102	0.113	0.105
0.10	9	0.079	0.083	0.083	0.084	0.084	0.074	0.109	0.112	0.106
0.25	9	0.087	0.089	0.091	0.088	0.089	0.083	0.109	0.103	0.103
0.40	9	0.094	0.095	0.091	0.092	0.090	0.088	0.119	0.119	0.114
0.10	100	0.124	0.084	0.082	0.083	0.083	0.079	0.166	0.114	0.109
0.25	100	0.197	0.084	0.083	0.083	0.082	0.080	0.252	0.110	0.107
0.40	100	0.218	0.082	0.083	0.082	0.082	0.080	0.324	0.103	0.102
		$p = 6, n_1 = n_2 = 20, g = 2$						$g = 3$		
0	—	0.086	0.160	0.129	0.146	0.151	0.097	0.119	0.146	0.131
0.10	9	0.108	0.158	0.126	0.134	0.140	0.106	0.138	0.144	0.129
0.25	9	0.115	0.146	0.126	0.131	0.127	0.102	0.151	0.149	0.136
0.40	9	0.123	0.131	0.131	0.126	0.124	0.114	0.159	0.146	0.138
0.10	100	0.186	0.165	0.129	0.145	0.149	0.107	0.230	0.138	0.125
0.25	100	0.225	0.130	0.112	0.118	0.123	0.106	0.322	0.137	0.128
0.40	100	0.291	0.123	0.136	0.123	0.108	0.105	0.366	0.155	0.129
		$p = 10, n_1 = n_2 = 20, g = 2$						$g = 3$		
0	—	0.107	0.245	0.146	0.228	0.232	0.128	0.136	0.245	0.153
0.10	9	0.123	0.212	0.144	0.196	0.204	0.131	0.159	0.224	0.151
0.25	9	0.153	0.186	0.149	0.179	0.182	0.142	0.173	0.207	0.158
0.40	9	0.158	0.166	0.173	0.165	0.164	0.150	0.196	0.196	0.175
0.10	100	0.153	0.210	0.143	0.198	0.203	0.135	0.236	0.224	0.158
0.25	100	0.237	0.168	0.123	0.163	0.170	0.134	0.325	0.213	0.166
0.40	100	0.283	0.167	0.203	0.167	0.162	0.156	0.419	0.202	0.171
		$p = 10, n_1 = n_2 = 50, g = 2$						$g = 3$		
0	—	0.074	0.089	0.098	0.092	0.092	0.081	0.102	0.112	0.107
0.10	9	0.093	0.094	0.101	0.096	0.095	0.087	0.124	0.118	0.114
0.25	9	0.100	0.097	0.099	0.097	0.098	0.091	0.128	0.116	0.113
0.40	9	0.102	0.098	0.096	0.098	0.099	0.093	0.133	0.122	0.120
0.10	100	0.173	0.094	0.102	0.098	0.098	0.089	0.201	0.117	0.114
0.25	100	0.179	0.096	0.100	0.096	0.096	0.092	0.242	0.116	0.113
0.40	100	0.251	0.097	0.095	0.096	0.097	0.095	0.315	0.120	0.120
		$p = 10, n_1 = n_2 = 100, g = 2$						$g = 3$		
0	—	0.073	0.075	0.078	0.076	0.076	0.075	0.097	0.100	0.100
0.10	9	0.086	0.081	0.083	0.082	0.082	0.081	0.110	0.103	0.103
0.25	9	0.085	0.079	0.081	0.080	0.079	0.079	0.110	0.103	0.103
0.40	9	0.085	0.079	0.079	0.079	0.079	0.078	0.115	0.107	0.107
0.10	100	0.125	0.079	0.081	0.079	0.079	0.079	0.154	0.102	0.102
0.25	100	0.134	0.077	0.078	0.077	0.077	0.077	0.177	0.099	0.100
0.40	100	0.157	0.078	0.079	0.078	0.078	0.078	0.212	0.107	0.107

Table 8: Mean probability of misclassification for Setup D1 with location contamination.

ε	κ	MLE	M-bw	MCD-A	MCD-B	MCD-C	OGK	MLE	M-bw	OGK
		$p = 2, n_1 = n_2 = 20, g = 2$						$g = 3$		
0	—	0.073	0.084	0.082	0.083	0.083	0.077	0.102	0.115	0.106
0.10	5	0.139	0.080	0.079	0.080	0.080	0.075	0.192	0.113	0.106
0.25	5	0.144	0.082	0.079	0.080	0.080	0.080	0.196	0.103	0.102
0.40	5	0.158	0.091	0.091	0.091	0.091	0.144	0.198	0.117	0.200
0.10	10	0.150	0.085	0.083	0.084	0.085	0.080	0.201	0.114	0.109
0.25	10	0.144	0.073	0.070	0.069	0.069	0.070	0.199	0.110	0.108
0.40	10	0.146	0.075	0.074	0.074	0.074	0.080	0.196	0.101	0.110
		$p = 2, n_1 = n_2 = 50, g = 2$						$g = 3$		
0	—	0.071	0.074	0.074	0.074	0.074	0.072	0.102	0.105	0.104
0.10	5	0.135	0.074	0.073	0.073	0.073	0.072	0.175	0.096	0.095
0.25	5	0.150	0.075	0.074	0.074	0.074	0.076	0.192	0.096	0.097
0.40	5	0.144	0.063	0.063	0.063	0.063	0.137	0.196	0.097	0.195
0.10	10	0.144	0.071	0.071	0.070	0.070	0.070	0.204	0.096	0.095
0.25	10	0.150	0.079	0.079	0.079	0.079	0.081	0.193	0.095	0.096
0.40	10	0.144	0.071	0.071	0.071	0.071	0.084	0.198	0.103	0.111
		$p = 2, n_1 = n_2 = 100, g = 2$						$g = 3$		
0	—	0.073	0.075	0.074	0.074	0.074	0.074	0.094	0.096	0.095
0.10	5	0.137	0.067	0.067	0.066	0.066	0.067	0.182	0.096	0.096
0.25	5	0.149	0.065	0.065	0.065	0.065	0.068	0.196	0.096	0.098
0.40	5	0.151	0.076	0.076	0.076	0.076	0.150	0.197	0.098	0.198
0.10	10	0.139	0.072	0.072	0.072	0.072	0.071	0.185	0.096	0.095
0.25	10	0.144	0.065	0.065	0.065	0.065	0.066	0.192	0.095	0.094
0.40	10	0.139	0.067	0.067	0.067	0.067	0.075	0.201	0.098	0.122
		$p = 6, n_1 = n_2 = 50, g = 2$						$g = 3$		
0	—	0.073	0.076	0.083	0.080	0.079	0.075	0.096	0.100	0.099
0.10	5	0.096	0.087	0.091	0.090	0.089	0.085	0.128	0.110	0.109
0.25	5	0.098	0.098	0.106	0.106	0.104	0.083	0.127	0.130	0.108
0.40	5	0.104	0.129	0.131	0.132	0.132	0.118	0.124	0.151	0.138
0.10	10	0.092	0.076	0.082	0.080	0.080	0.077	0.124	0.106	0.105
0.25	10	0.094	0.079	0.082	0.079	0.080	0.078	0.126	0.106	0.106
0.40	10	0.100	0.127	0.129	0.132	0.132	0.104	0.121	0.141	0.131
		$p = 6, n_1 = n_2 = 100, g = 2$						$g = 3$		
0	—	0.072	0.073	0.075	0.075	0.074	0.073	0.094	0.095	0.096
0.10	5	0.094	0.077	0.079	0.078	0.078	0.078	0.117	0.099	0.099
0.25	5	0.089	0.089	0.097	0.097	0.096	0.075	0.120	0.122	0.104
0.40	5	0.092	0.103	0.110	0.112	0.111	0.101	0.122	0.132	0.131
0.10	10	0.096	0.081	0.082	0.081	0.081	0.081	0.117	0.094	0.095
0.25	10	0.093	0.071	0.071	0.071	0.071	0.073	0.119	0.098	0.100
0.40	10	0.093	0.104	0.113	0.114	0.114	0.098	0.125	0.134	0.130
		$p = 10, n_1 = n_2 = 100, g = 2$						$g = 3$		
0	—	0.071	0.074	0.076	0.074	0.074	0.074	0.096	0.098	0.098
0.10	5	0.084	0.078	0.080	0.078	0.078	0.078	0.120	0.108	0.108
0.25	5	0.088	0.095	0.112	0.121	0.121	0.080	0.117	0.123	0.108
0.40	5	0.088	0.108	0.114	0.117	0.117	0.099	0.117	0.129	0.124
0.10	10	0.083	0.078	0.080	0.078	0.078	0.078	0.110	0.101	0.101
0.25	10	0.086	0.093	0.112	0.114	0.113	0.080	0.110	0.117	0.102
0.40	10	0.082	0.103	0.111	0.111	0.110	0.092	0.111	0.125	0.120

Table 9: Mean probability of misclassification for Setup D2 with location contamination.

ε	κ	MLE	M-bw	MCD-A	MCD-B	MCD-C	OGK	MLE	M-bw	OGK
		$p=2, n_1=n_2=20, g=2$						$g=3$		
0.00	—	0.112	0.128	0.125	0.127	0.126	0.116	0.138	0.155	0.146
0.10	5	0.189	0.131	0.127	0.126	0.126	0.123	0.202	0.144	0.137
0.25	5	0.193	0.127	0.124	0.125	0.126	0.125	0.221	0.151	0.145
0.40	5	0.190	0.132	0.131	0.131	0.131	0.183	0.224	0.197	0.226
0.10	10	0.190	0.126	0.123	0.122	0.124	0.117	0.227	0.155	0.149
0.25	10	0.196	0.122	0.121	0.120	0.121	0.122	0.229	0.152	0.149
0.40	10	0.184	0.118	0.115	0.116	0.116	0.137	0.225	0.149	0.187
		$p=2, n_1=n_2=50, g=2$						$g=3$		
0.00	—	0.109	0.113	0.114	0.113	0.113	0.111	0.137	0.138	0.137
0.10	5	0.175	0.109	0.109	0.108	0.108	0.108	0.197	0.132	0.129
0.25	5	0.187	0.108	0.107	0.107	0.107	0.110	0.218	0.137	0.137
0.40	5	0.184	0.115	0.115	0.114	0.115	0.185	0.221	0.154	0.222
0.10	10	0.188	0.113	0.112	0.112	0.113	0.112	0.214	0.133	0.130
0.25	10	0.191	0.120	0.120	0.120	0.120	0.121	0.216	0.134	0.133
0.40	10	0.183	0.112	0.112	0.112	0.112	0.140	0.224	0.142	0.201
		$p=2, n_1=n_2=100, g=2$						$g=3$		
0.00	—	0.100	0.102	0.102	0.102	0.102	0.102	0.134	0.136	0.135
0.10	5	0.171	0.108	0.107	0.107	0.107	0.108	0.201	0.130	0.129
0.25	5	0.184	0.105	0.104	0.104	0.104	0.109	0.220	0.136	0.136
0.40	5	0.169	0.100	0.100	0.100	0.100	0.169	0.227	0.143	0.227
0.10	10	0.179	0.110	0.109	0.109	0.109	0.109	0.214	0.140	0.139
0.25	10	0.182	0.106	0.105	0.105	0.105	0.106	0.216	0.132	0.130
0.40	10	0.178	0.103	0.102	0.102	0.102	0.133	0.213	0.135	0.201

5. SUMMARY AND CONCLUSIONS

In this paper we have reviewed the recent methods for robust LDA and have proposed several new ones — based on the Constrained M estimates as defined by Rocke [14] and on the pairwise estimator OGK of Maronna and Zamar [10]. It is shown with examples that the proposed robust LDA procedures behave very well on data sets with and without outlying observations. The simulation study of He and Fung [5] was repeated for all estimators and it showed S, M-bw and OGK as the best performers (the estimators are shown in increasing order of their speed). A large scale simulation study covering a variety of settings with different distributions and contaminations was performed, and showed that in most of the cases the robust LDA procedures behave similarly to the MLE procedure when applied on clean data — i.e. remain uninfluenced by the presence of outliers in the data unlike the classical rules.

Although the OGK estimator seems to be the best performer in terms of probability of misclassification as well as of speed, a more thorough study is necessary because of its non-affine equivariance. Also, the evaluation of the quality of the estimators of the group means and the common covariance matrix in the context of the linear discriminant analysis deserves further work.

All computations were performed by software developed in the statistical environment R, which is available in the package *rrcov* — Todorov [21].

ACKNOWLEDGMENTS

This research was partially supported by the Center for Mathematics and its Applications, Lisbon, Portugal, through *Programa Operacional "Ciência, Tecnologia, Inovação"* (POCTI) of the *Fundação para a Ciência e a Tecnologia* (FCT), cofinanced by the European Community fund FEDER.

We also acknowledge the valuable suggestions from the referees.

REFERENCES

[1] Chork, C. and Rousseeuw, P.J. (1992). Integrating a high breakdown option into discriminant analysis in exploration geochemistry, *Journal of Geochemical Exploration*, **43**, 191–203.

[2] Croux, C. and Dehon, C. (2001). Robust linear discriminant analysis using s-estimators, *The Canadian Journal of Statistics*, **29**, 473–492.

[3] Friedman, J.H. (1989). Regularized discriminant analysis, *Journal of the American Statistical Association*, **84**, 165–175.

[4] Hawkins, D.M. and McLachlan, G. (1997). High-breakdown linear discriminant analysis, *Journal of the American Statistical Association*, **92**, 136–143.

[5] He, X. and Fung, W. (2000). High breakdown estimation for multiple populations with applications to discriminant analysis, *Journal of Multivariate Analysis*, **72**, 151–162.

[6] Hubert, M. and Van Driessen, K. (2004). Fast and robust discriminant analysis, *Computational Statistics and Data Analysis*, **45**, 301–320.

[7] Lachenbruch, P.A. (1975). *Discriminant Analysis*, Hafner, New York.

[8] Lachenbruch, P.A. and Michey, M. (1968). Estimation of error rates in discriminant analysis, *Technometrics*, **10**, 1–11.

[9] Maronna, R. and Yohai, V. (1998). *Robust estimation of multivariate location and scatter.* In "Encyclopedia of Statistical Sciences", Updated Volume 2 (S.C.R. Kotz and D. Banks, Eds.), Wiley, New York, 589–596.

[10] Maronna, R. and Zamar, R. (2002). Robust estimation of location and dispersion for high-dimensional datasets, *Technometrics*, **44**, 307–317.

[11] Pison, G.; Van Aelst, S. and Willems, G. (2002). Small sample corrections for LTS and MCD, *Metrika*, **55**, 111–123.

[12] Puranen, J. (2006). Fish catch data set,
http://www.amstat.org/publications/jse/datasets/fishcatch.txt

[13] Reaven, G.M. and Miller, R.G. (1979). An attempt to define the nature of chemical diabetes using a multidimensional analysis, *Diabetologia*, **16**, 17–24.

[14] Rocke, D.M. (1996). Robustness properties of S-estimators of multivariate location and shape in high dimension, *Annals of Statistics*, **24**, 1327–1345.

[15] Rocke, D.M. and Woodruff, D.L. (1996). Identification of outliers in multivariate data, *Journal of the American Statistical Association*, **91**, 1047–1061.

[16] Rousseeuw, P. (1984). Least median of squares regression, *Journal of the American Statistical Association*, **79**, 851–857.

[17] Rousseeuw, P.J. and van Zomeren, B.C. (1991). *Robust distances: Simulation and cutoff values.* In: "Directions in Robust Statistics and Diagnostics", Part II (W. Stahel and S. Weisberg, Eds.), Springer Verlag, New York.

[18] Salibian-Barrera, M. and Yohai, V. (2005). A fast algorithm for S-regression estimates. To appear in the *Journal of Computational and Graphical Statistics.*

[19] Todorov, V.; Neykov, N. and Neytchev, P. (1990). *Robust selection of variables in the discriminant analysis based on mve and mcd estimators.* In: "Proceedings in Computational Statistics, COMPSTAT", Physica Verlag, Heidelberg.

[20] Todorov, V.; Neykov, N. and Neytchev, P. (1994). Robust two-group discrimination by bounded influence regression, *Computational Statistics and Data Analysis*, **17**, 289–302.

[21] Todorov, V.K. (2006). *rrcov: Scalable Robust Estimators with High Breakdown Point*, R package version 0.3-05.

[22] Woodruff, D.L. and Rocke, D.M. (1994). Computable robust estimation of multivariate location and shape in high dimension using compound estimators, *Journal of the American Statistical Association*, **89**, 888–896.

REVSTAT – Statistical Journal
Volume 5, Number 1, March 2007, 85–96

DE-BIASING WEIGHTED MLE VIA INDIRECT INFERENCE: THE CASE OF GENERALIZED LINEAR LATENT VARIABLE MODELS

Authors: Maria-Pia Victoria-Feser
– University of Geneva,
Switzerland
maria-pia.victoriafeser@hec.unige.ch

Abstract:

- In this paper we study bias-corrections to the weighted MLE (Dupuis and Morgenthaler, 2002), a robust estimator simply defined through a weighted score function. Indeed, although the WMLE is relatively simple to compute, for most models it is not consistent and hence not very helpful. For example, the model we consider in this paper is the generalized linear latent variable model (GLLVM) proposed in Moustaki and Knott (2000) (see also Moustaki, 1996, Sammel, Ryan, and Legler, 1997 and Bartholomew and Knott, 1999). The score functions of this model are very complicated. They contain integrals that need to be evaluated. Moreover, they are highly nonlinear in the parameters which makes the use of complicated robust estimator quite impossible in practice. Moustaki and Victoria-Feser (2006) propose to use a weighted MLE and develop indirect inference (Gouriéroux, Monfort, and Renault, 1993, Gallant and Tauchen, 1996 and also Genton and de Luna, 2000, Genton and Ronchetti, 2003) to remove the bias. It can be computed in a simple iterative fashion. In this paper, we actually focus on indirect inference for bias correction in general. We rely heavily on the findings of Moustaki and Victoria-Feser (2006).

Key-Words:

- *factor analysis; latent variables; M-estimators.*

AMS Subject Classification:

- 62G35.

1. INTRODUCTION

Consider a general class of weighted MLE (WMLE) proposed by Dupuis and Morgenthaler (2002) belonging to the class of M-estimators (Huber, 1981) defined as the solution in $\boldsymbol{\theta}$ of

$$\frac{1}{n}\sum_{i=1}^{n}\psi_c(\boldsymbol{x}_i;\boldsymbol{\theta}) \;=\; \frac{1}{n}\sum_{i=1}^{n} s(\boldsymbol{x}_i;\boldsymbol{\theta})\, w(\boldsymbol{x}_i,c) \;=\; \mathbf{0}\;,$$

with the underlying assumption that $\boldsymbol{x}_i \sim F_{\boldsymbol{\theta}}$ and the weights $w(\boldsymbol{x}_i,c)$ are such that smaller weights are given to observations with larger score function $s(\boldsymbol{x}_i;\boldsymbol{\theta}) = \partial/\partial\boldsymbol{\theta}\,\log\bigl(\partial/\partial\boldsymbol{x}\, F_{\boldsymbol{\theta}}(\boldsymbol{x})\bigr)$. A typical choice for the weights is Huber type weights, which for a given tuning parameter c are given by

$$w(\boldsymbol{x};c) \;=\; \min\left(1;\; \frac{c}{\|s(\boldsymbol{x};\boldsymbol{\theta})\|}\right), \tag{1.1}$$

where $\|...\|$ denotes the Euclidean norm. If $F_{\boldsymbol{\theta}}$ and/or the weight function are not symmetric, then the resulting estimator is not consistent. Based on a first order development of the bias, Dupuis and Morgenthaler (2002) propose a bias correction given by

$$B(\widehat{\boldsymbol{\theta}}) \;=\; -\left.\frac{\displaystyle\int s(\boldsymbol{x};\boldsymbol{\theta})\, w(\boldsymbol{x};\boldsymbol{\theta})\; dF_{\boldsymbol{\theta}}(\boldsymbol{x})}{\displaystyle\int\left(\frac{\partial}{\partial\boldsymbol{\theta}}\, s(\boldsymbol{x};\boldsymbol{\theta})\, w(\boldsymbol{x};\boldsymbol{\theta}) + s(\boldsymbol{x};\boldsymbol{\theta})\,\frac{\partial}{\partial\boldsymbol{\theta}}\, w(\boldsymbol{x};\boldsymbol{\theta})\right) dF_{\boldsymbol{\theta}}(\boldsymbol{x})}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \tag{1.2}$$

to be added to the inconsistent WMLE $\widehat{\boldsymbol{\theta}}$. The computation of two integrals is still needed (and can be done by means of simulations) as well as the derivative of the weight function. Alternatively, one can consider estimators of the type

$$\frac{1}{n}\sum_{i=1}^{n} s(\boldsymbol{x}_i;\boldsymbol{\theta})\, w(\boldsymbol{x}_i,c) - a(\boldsymbol{\theta}) \;=\; \mathbf{0}\;,$$

with

$$a(\boldsymbol{\theta}) \;=\! \int s(\boldsymbol{x};\boldsymbol{\theta})\, w(\boldsymbol{x},c)\; dF_{\boldsymbol{\theta}}(x)$$

and hence estimate simultaneously the bias correction with the estimator. This can become very complicated depending on the form of the score function. In the following section, a bias correction for a WMLE is presented, in the same spirit as (1.2) but based on the theory of indirect inference.

2. INDIRECT INFERENCE FOR BIAS REDUCTION

Indirect estimation (Gouriéroux, Monfort, and Renault, 1993, Gallant and Tauchen, 1996) has been proposed as an estimation procedure for a complex model $F_{\boldsymbol{\theta}}$ with intractable likelihood functions. It involves the computation of an estimator $\widehat{\boldsymbol{\pi}}$ for the parameters of an auxiliary model $F_{\boldsymbol{\pi}}$ that does not provide a consistent estimator of $\boldsymbol{\theta}$. In particular, let $\widehat{\boldsymbol{\pi}}$ be an M-estimator defined implicitly by

$$\int \psi(\boldsymbol{x}; \widehat{\boldsymbol{\pi}})\, dF_n(\boldsymbol{x}) \;=\; \mathbf{0} \;.$$

As the sample size tends to infinity, this auxiliary estimate defines a mapping from the parameter space of $\boldsymbol{\theta}$ to the parameter space of the auxiliary model, i.e. $\boldsymbol{\theta} \to \boldsymbol{\pi}(\boldsymbol{\theta})$, defined by

$$\int \psi(\boldsymbol{x}; \boldsymbol{\pi}(\boldsymbol{\theta}))\, dF_{\boldsymbol{\theta}}(\boldsymbol{x}) \;=\; \mathbf{0} \;. \tag{2.1}$$

With indirect inference one tries in some sense to invert this map, i.e. $\boldsymbol{\pi} \to \boldsymbol{\theta}(\boldsymbol{\pi})$, and use this inverse to calculate the estimator $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}(\widehat{\boldsymbol{\pi}})$. The latter can be obtained implicitly by the solution in $\boldsymbol{\theta}$ of

$$\int \psi(\boldsymbol{x}; \widehat{\boldsymbol{\pi}})\, dF_{\boldsymbol{\theta}}(\boldsymbol{x}) \;=\; \mathbf{0} \;. \tag{2.2}$$

This indirect estimator results as a particular case of the general minimization problem defining indirect estimators, i.e.

$$\widehat{\boldsymbol{\theta}} \;=\; \arg\min_{\boldsymbol{\theta}} \big(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}(\boldsymbol{\theta})\big)^T \,\Omega \big(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}(\boldsymbol{\theta})\big) \;,$$

with $\boldsymbol{\pi}(\boldsymbol{\theta})$ obtained as the solution of (2.1). The matrix Ω can be chosen on efficiency arguments but for simplicity, on can choose $\Omega = I$. The estimation of $\boldsymbol{\pi}(\boldsymbol{\theta})$ is the difficulty in applying the indirect method. If it is possible to create samples $\widetilde{\boldsymbol{x}}_i(\boldsymbol{\theta})$, $i = 1, ..., s \cdot n$, simulated (with fixed seed) from $F_{\boldsymbol{\theta}}$ for a given $\boldsymbol{\theta}$, then a Monte Carlo estimate of $\boldsymbol{\pi}(\boldsymbol{\theta})$ can be used. This estimate is defined as the solution in $\boldsymbol{\pi}(\boldsymbol{\theta})$ of

$$\frac{1}{sn} \sum_{i=1}^{sn} \psi\big(\widetilde{\boldsymbol{x}}_i(\boldsymbol{\theta}); \boldsymbol{\pi}(\boldsymbol{\theta})\big) \;=\; \mathbf{0} \;.$$

Gouriéroux, Monfort, and Renault (1993) show that this estimator is asymptotically equivalent to the one proposed by Gallant and Tauchen (1996) (available since 1992 as a working paper) defined by

$$\widehat{\boldsymbol{\theta}} \;=\; \arg\min_{\boldsymbol{\theta}} \left(\frac{1}{sn} \sum_{i=1}^{sn} \psi\big(\widetilde{\boldsymbol{x}}_i(\boldsymbol{\theta}); \widehat{\boldsymbol{\pi}}\big) \right)^T \Delta \left(\frac{1}{sn} \sum_{i=1}^{sn} \psi\big(\widetilde{\boldsymbol{x}}_i(\boldsymbol{\theta}); \widehat{\boldsymbol{\pi}}\big) \right) , \tag{2.3}$$

with again Δ chosen on efficiency arguments. When $\dim(\boldsymbol{\theta}) = \dim(\boldsymbol{\pi})$ and $\Delta = I$, the solution of (2.3) is given by the solution in $\boldsymbol{\theta}$ of (2.2) in which the integral is estimated by the mean over a simulated sample. We also note that when ψ is the score function, then $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\pi}}$.

Indirect inference has already been used with robust statistics: see Genton and de Luna (2000) and Genton and Ronchetti (2003). Similar ideas can be found in Cabrera and Fernholz (1999).

$\widehat{\boldsymbol{\theta}}$ can be found iteratively using a Newton step. For that we need

$$\frac{\partial}{\partial \boldsymbol{\theta}} \int \psi(\boldsymbol{x}; \widehat{\boldsymbol{\pi}})\, dF_{\boldsymbol{\theta}}(\boldsymbol{x}) \;=\; \int \psi(\boldsymbol{x}; \widehat{\boldsymbol{\pi}})\, s^T(\boldsymbol{x}; \boldsymbol{\theta})\, dF_{\boldsymbol{\theta}}(\boldsymbol{x}) \;.$$

Then the Newton step is given by

$$\widehat{\boldsymbol{\theta}}^{(k+1)} \;=\; \widehat{\boldsymbol{\theta}}^{(k)} - S^{-1}\big(\widehat{\boldsymbol{\pi}}, \widehat{\boldsymbol{\theta}}^{(k)}\big) \sum_{i=1}^{sn} \psi\big(\widetilde{\boldsymbol{x}}_i(\widehat{\boldsymbol{\theta}}^{(k)}); \widehat{\boldsymbol{\pi}}\big) \;, \tag{2.4}$$

where

$$S\big(\widehat{\boldsymbol{\pi}}, \widehat{\boldsymbol{\theta}}\big) \;=\; \sum_{i=1}^{sn} \psi\big(\boldsymbol{x}_i(\widehat{\boldsymbol{\theta}}); \widehat{\boldsymbol{\pi}}\big)\, s^T\big(\boldsymbol{x}_i(\widehat{\boldsymbol{\theta}}); \widehat{\boldsymbol{\theta}}\big)$$

and $\widehat{\boldsymbol{\pi}}$ is the (inconsistent) M-estimator. With this indirect estimator, there is hence no need for simultaneous estimation of bias (computation of $a(\boldsymbol{\theta})$). This estimator has been proposed by Moustaki and Victoria-Feser (2006) in the context of generalized linear latent variable models.

3. GENERALIZED LINEAR LATENT VARIABLE MODELS (GLLVM)

Latent variable models are widely used in social sciences for studying the interrelationships among observed variables. More specifically, latent variable models are used for reducing the dimensionality of multivariate data, for assigning scores to sample members on the latent dimensions identified by the model as well as for the construction of measurement scales (e.g. in educational testing and psychometrics). Moustaki and Knott (2000) proposed a generalized linear latent variable model (GLLVM) framework for any type of observed data (metric and categorical) in the exponential family. They extended the work of Moustaki (1996) and Sammel, Ryan, and Legler (1997) for mixed binary and metric variables (the latter with covariate effects as well) and Bartholomew and Knott (1999) for categorical variables. A similar framework is also discussed by Skrondal and Rabe-Hesketh (2004) that includes multilevel models (random-effects models) as a special case.

Formally, given a set of response variables $x_1, ..., x_p$, there exists a (smaller) set of latent variables or factors $z_1, ..., z_q$ that account for the dependencies among the response variables. In other words, given the latent variables, the manifest ones are conditionally independent. Factor analysis is the simplest case. In general we suppose that the conditional distribution of the manifest variables given the latent ones belongs to the exponential family, i.e.

$$g_m\big(x_m \,|\, \boldsymbol{z}, \boldsymbol{\theta}_m\big) \;=\; \exp\left\{ \frac{x_m\, \boldsymbol{\alpha}_m\, \boldsymbol{z}^*}{\phi_m} - \frac{b\big(\boldsymbol{\alpha}_m\, \boldsymbol{z}^*\big)}{\phi_m} + c(x_m, \phi_m) \right\}, \qquad m = 1, ..., p \;,$$

with $\boldsymbol{\alpha}_m = [\alpha_{m0}, ..., \alpha_{mq}]$, $m=1,...,p$, the so-called loadings, ϕ_m, $m=1,...,p$, the scale parameters (for example for normal manifest variables),

$$\boldsymbol{z}^* = [1, z_1, ..., z_q]^T = [1, \boldsymbol{z}^T]^T$$

and hence $\boldsymbol{\theta}_m = (\boldsymbol{\alpha}_m, \phi_m)^T$. The latent variables $\boldsymbol{z}$ are supposed standard multivariate normal with density $\varphi(\boldsymbol{z})$ (but the independence assumption can be relaxed), hence, the marginal distribution is

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \int \cdots \int \left[\prod_{m=1}^{p} g_m(x_m \,|\, \boldsymbol{z}, \boldsymbol{\theta}_m) \right] \varphi(\boldsymbol{z}) \; d\boldsymbol{z} \; .$$

The score functions become

$$\begin{aligned} \text{(3.1)} \qquad s_m^{(1)}(\boldsymbol{x};\boldsymbol{\theta}) &= \frac{\partial}{\partial \boldsymbol{\alpha}_m} \log\big(f(\boldsymbol{x};\boldsymbol{\theta})\big) \\ &= \frac{1}{f(\boldsymbol{x};\boldsymbol{\theta})} \int \cdots \int g(\boldsymbol{x}\,|\,\boldsymbol{z},\boldsymbol{\theta}) \left(\frac{x_m - b'(\boldsymbol{\alpha}_m \boldsymbol{z}^*)}{\phi_m} \right) \boldsymbol{z}^* \varphi(\boldsymbol{z}) \; d\boldsymbol{z} \; , \end{aligned}$$

$$\begin{aligned} \text{(3.2)} \qquad s_m^{(2)}(\boldsymbol{x};\boldsymbol{\theta}) &= \frac{\partial}{\partial \phi_m} \log\big(f(\boldsymbol{x};\boldsymbol{\theta})\big) \\ &= \frac{1}{f(\boldsymbol{x};\boldsymbol{\theta})} \int \cdots \int g(\boldsymbol{x}\,|\,\boldsymbol{z},\boldsymbol{\theta}) \\ &\quad \cdot \left(- \frac{x_m \boldsymbol{\alpha}_m \boldsymbol{z}^* - b(\boldsymbol{\alpha}_m \boldsymbol{z}^*)}{\phi_m^2} + \frac{\partial}{\partial \phi_m} c(\phi_m, x_m) \right) \varphi(\boldsymbol{z}) \; d\boldsymbol{z} \; , \end{aligned}$$

for $m=1,...,p$. The integrals in (3.1) and (3.2) can be approximated using fixed Gauss–Hermite quadrature (see e.g. Bock and Liberman, 1970), adaptive quadrature points (see e.g. Bock and Schilling, 1997, Schilling and Bock, 2005), Monte Carlo approximations (see e.g. Sammel, Ryan, and Legler, 1997) or Laplace approximation (see e.g. Huber, Ronchetti, and Victoria-Feser 2004). All these approximations lead to approximate ML estimators. The models we consider here are one factor models and although it is known that Gauss–Hermite rule can give biased estimators in some situations, we will nevertheless use it to compute the integrals.

Moustaki and Victoria-Feser (2006) study the robustness properties of the (approximated) MLE by means of the Influence Function (Hampel, 1968, 1974). Not surprisingly, even with binary data, the MLE can be biased by data contamination, which in this context appear as unexpected binary responses. Since the (approximate) MLE is already quite complicate computationally, Moustaki and Victoria-Feser (2006) propose to use a WMLE with consistency correction via indirect inference. The WMLE $\widehat{\boldsymbol{\pi}}$ is computed with Huber type weights (1.1). The consistent estimator $\widehat{\boldsymbol{\theta}}$ is obtained using indirect inference and called Indirect Globally Weighted Robust (IGWR) estimator. Its (approximate) asymptotic covariance is also given in Moustaki and Victoria-Feser (2006) which is used for inference and also for choosing the tuning constant c of the Huber weights on efficiency arguments.

4. SIMULATION STUDY

We report here the simulation study presented in Moustaki and Victoria-Feser (2006). The model we consider is the one-factor model ($q=1$) fitted to two binary ($m=1,2$) and three normal ($m=3,4,5$) manifest variables with parameter values

- $\boldsymbol{\alpha}_1 = [1.0, 0.7]\,,$
- $\boldsymbol{\alpha}_2 = [0.8, 1.0]\,,$
- $\boldsymbol{\alpha}_3 = [2.0, 0.6]$ and $\phi_3 = 1\,,$
- $\boldsymbol{\alpha}_4 = [2.5, 0.7]$ and $\phi_4 = 1\,,$
- $\boldsymbol{\alpha}_5 = [3.0, 0.8]$ and $\phi_5 = 1\,.$

150 samples of size 200 where generated and contaminated in three ways:

- 3% of the first normal variable (i.e. observations of x_3) set to an arbitrary value (20) (pointmass 1);
- 3% of all three normal variables set to an arbitrary value (20) (pointmass 3);
- 3% of the data from the mixed GLLVM with $\boldsymbol{\alpha}_5 = [3.0, 8]^T$ instead of $\boldsymbol{\alpha}_5 = [3.0, 0.8]^T$ (model deviation).

The MLE, IGWR and IGWR1 which is defined by the iterative procedure given in (2.4) with only one iteration, were computed. The tuning constant was set to $c=3.5$, which corresponds to an efficiency level of 95% with respect to the MLE.

Figure 1 presents the distributions of the different estimators for the loading of the first manifest variable (binary) α_{11} with all types of contamination (including no contamination). Even if the contamination occurs on the normal manifest variables, the MLE can be biased as can bee seen with the pointmass 3 contamination type. Figures 2 and 3 present the distribution of the different estimators for respectively the mean of the third manifest variable (normal) α_{30} and the loading of the fifth manifest variable (normal) α_{51} with all types of contamination. The bias on the MLE appears quite large, while both robust estimators remain very stable. Without contamination, there is no apparent difference in distribution between the MLE and the robust estimators. Figure 4 presents the same analysis but for the estimators of the scale parameter for the first normal variable ϕ_3. The MLE of the scale parameter seems to be affected only when the contamination occurs only on the corresponding manifest variable. Again, the behavior of the robust estimators show great stability.

It should be noted that Moustaki and Victoria-Feser (2006) conclude that although the IGWR1 seems to perform very well with the examples of this simulations study, its bias increases more rapidly that the one of the IGWR as the WMLE is more biased, i.e. as the tuning constant c decreases.

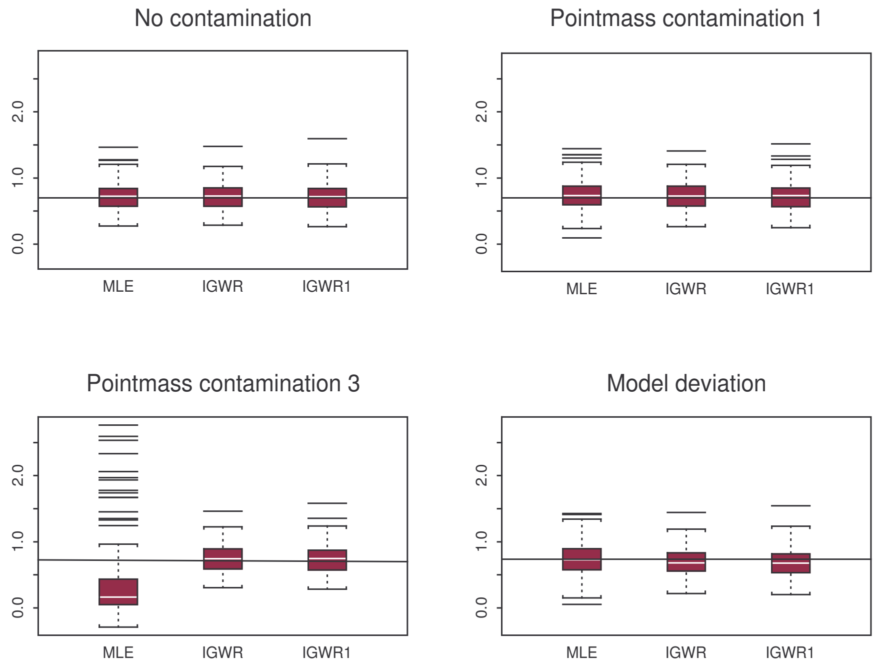

Figure 1: Distribution of the estimators for the loading on the first binary manifest variable. The horizontal line gives the true value.

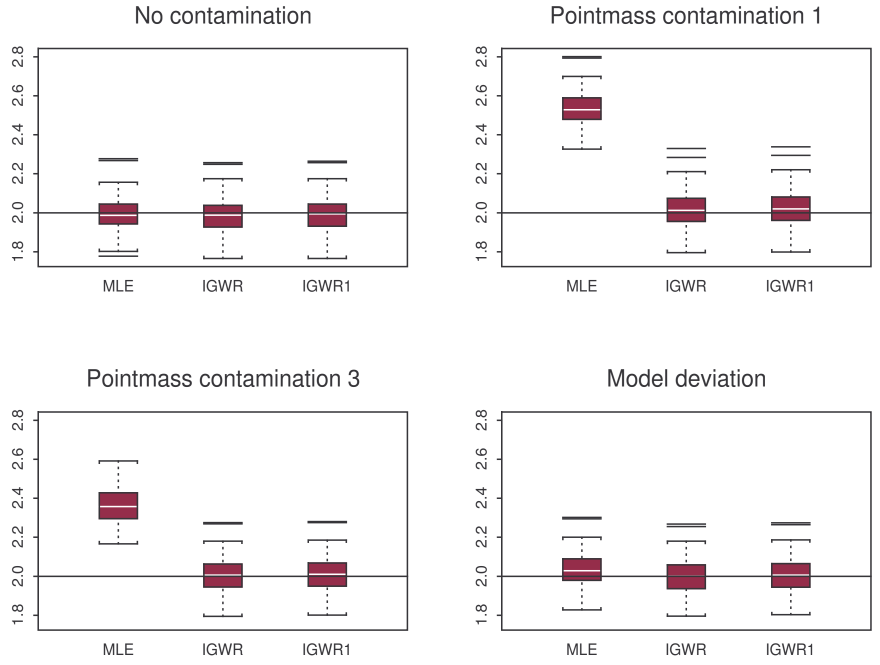

Figure 2: Distribution of the estimators for the mean on the first normal manifest variable. The horizontal line gives the true value.

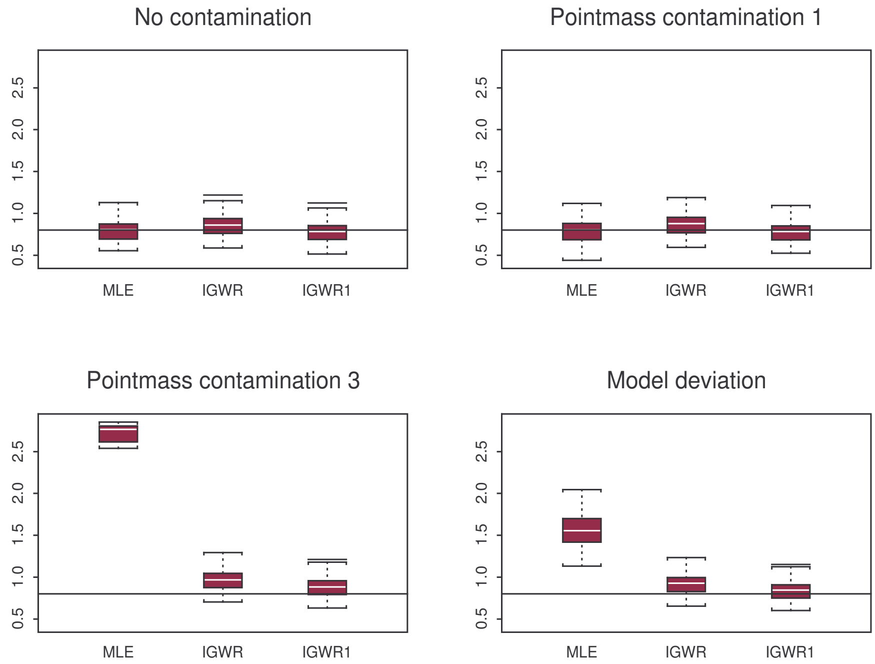

Figure 3: Distribution of the estimators for the loading on the third normal manifest variable. The horizontal line gives the true value.

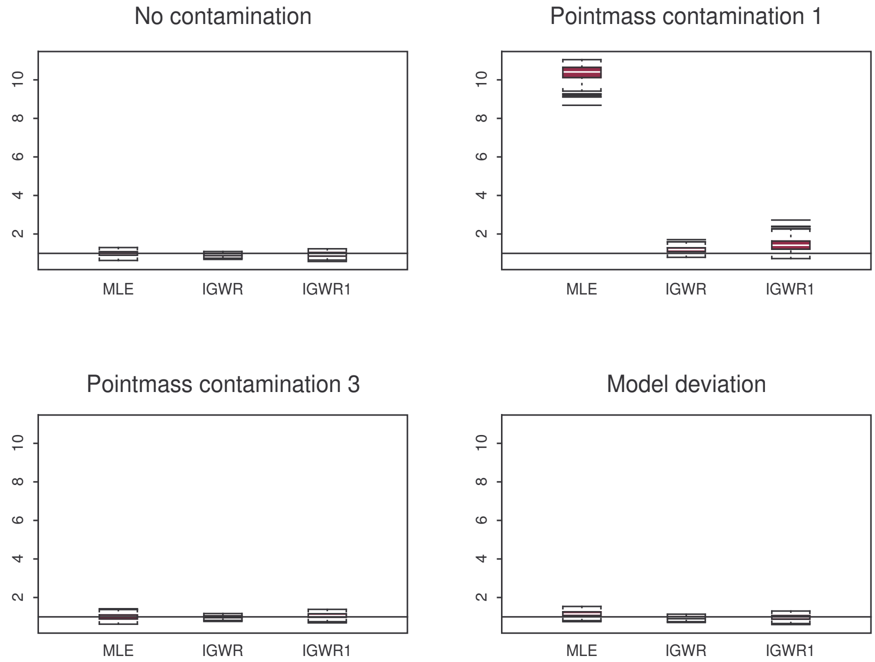

Figure 4: Distribution of the estimators for the scale on the first normal manifest variable. The horizontal line gives the true value.

5. ANALYSIS OF WEALTH DATA

Moustaki and Victoria-Feser (2006) also present an example based on a sub-sample of size 100 households of the 1990 consumption survey in Switzerland, provided by the Swiss Federal Statistical Office. The aim is to construct a measurement scale for the level of wealth, and for the purpose of this exercise, five variables have been selected. These are:

- purchase of a dishwasher (1/0) (Dishwasher)
- purchase of a car (1/0) (Car)
- equivalent food expenditure in logarithm (Food)
- equivalent expenditures for clothing in logarithm (Clothing)
- equivalent expenditures for housing in logarithm (Housing)

The continuous variables are treated as normal variables. Variables from the same survey have been analyzed before using the GLLVM by Moustaki and Knott (1997), Bartholomew and Knott (1999) and Huber, Ronchetti, and Victoria-Feser (2004). A one-factor model using both the ML and the IGWR estimators is fitted to the data. The bounding constant c has been set to 5 corresponding to an efficiency level of 94% (computed on the parameter values provided by the IGWR). The parameter values estimated by the ML and the IGWR estimators are presented in Table 1 together with their standard errors (the values in bold correspond to significant variables at the 5% level).

Table 1: Parameter's estimates and standard errors for the GLLVM on the wealth data.

Parameters		MLE		IGWR, $c = 5$	
		Estimate	Stand. Err.	Estimate	Stand. Err.
Constants	α_{10}	**−0.506**	0.23	**−0.589**	0.26
	α_{20}	**−0.623**	0.23	**−0.537**	0.23
	α_{30}	**6.922**	0.23	**6.887**	0.28
	α_{40}	**5.353**	0.32	**5.332**	0.32
	α_{50}	**7.087**	0.33	**7.140**	0.29
Loadings	α_{11}	0.466	0.26	**0.679**	0.28
	α_{21}	−0.167	0.24	0.216	0.25
	α_{31}	**1.021**	0.18	**1.098**	0.21
	α_{41}	**1.412**	0.31	**1.415**	0.28
	α_{51}	**1.044**	0.33	**1.064**	0.27
Variances	ϕ_3	0.289	0.16	**0.426**	0.17
	ϕ_4	**1.280**	0.27	**1.056**	0.20
	ϕ_5	**1.475**	0.22	**0.935**	0.14

The ML estimator shows that only the variables (Food, Clothing and Housing) are indicators of wealth, whereas the IGWR adds the variable Dishwasher. Both analyses exclude the variable Car. Variables Food and Housing are found with both methods to be indicators of the latent variable, whereas the association is stronger with the Clothing variable. For a diagnostics analysis, the weights given in (1.1) have been computed for each observation at the IGWR values and plotted in Figure 5. There are apparently (only) 5 outliers.

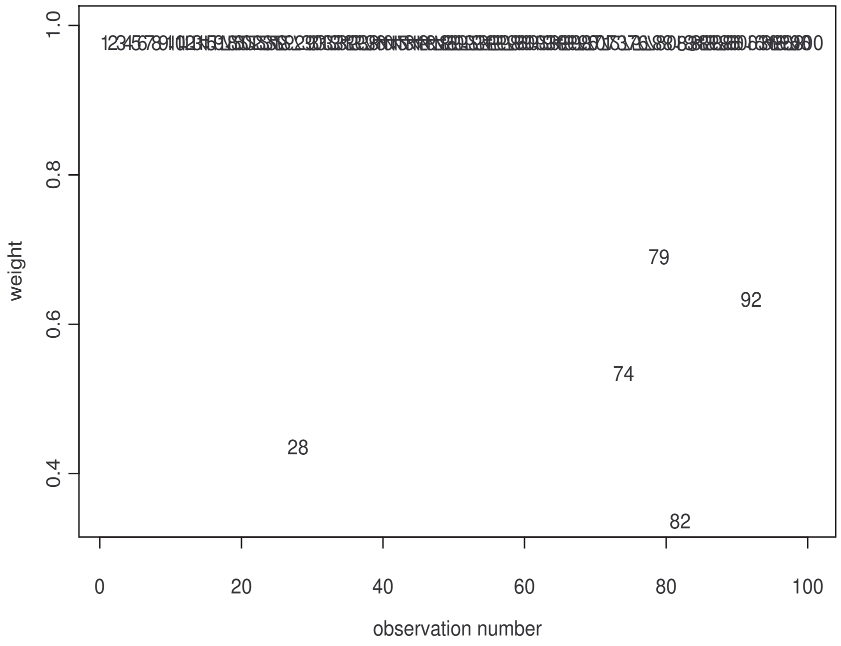

Figure 5: IGWR's weights against observation number for the wealth data.

ACKNOWLEDGMENTS

The author is partially supported by the Swiss National Science Foundation (grant #100012-109532).

REFERENCES

[1] Bartholomew, D.J. and Knott, M. (1999). *Latent Variable Models and Factor Analysis*, Arnold, London.

[2] Bock, R.D. and Liberman, M. (1970). Fitting a response model for n dichotomously scored items, *Psychometrika*, **35**, 179–197.

[3] Bock, R.D. and Schilling, S.G. (1997). *High-dimensional full-information item factor analysis.* In "Latent Variable Modelling and Applications to Causality" (M. Berkane, Ed.), Springer, New York, 164–176.

[4] Cabrera, J. and Fernholz, L.T. (1999). Target estimation for bias and mean square error reduction, *The Annals of Statistics*, **27**, 1080–1104.

[5] Dupuis, D.J. and Morgenthaler, S. (2002). Robust weighted likelihood estimators with an application to bivariate extremevalue problems, *Canadian Journal of Statistics*, **30**, 17–36.

[6] Gallant, A.R. and Tauchen, G. (1996). Which moments to match, *Econometric Theory*, **12**, 657–681.

[7] Genton, M.G. and de Luna, X. (2000). Robust simulation-based estimation, *Statistics and Probability Letters*, **48**, 253–259.

[8] Genton, M.G. and Ronchetti, E. (2003). Robust indirect inference, *Journal of the American Statistical Association*, **98**, 1–10.

[9] Gouriéroux, C.; Monfort, A. and Renault, A.E. (1993). Indirect inference, *Journal of Applied Econometrics*, **8** (supplement), S85–S118.

[10] Hampel, F.R. (1968). *Contribution to the Theory of Robust Estimation*, Ph.D. thesis, University of California, Berkeley.

[11] Hampel, F.R. (1974). The influence curve and its role in robust estimation, *Journal of the American Statistical Association*, **69**, 383–393.

[12] Huber, P.; Ronchetti, E. and Victoria-Feser, M.-P. (2004). Estimation of generalized latent trait models, *Journal of the Royal Statistical Society, Series B*, **66**, 893–908.

[13] Huber, P.J. (1981). *Robust Statistics*, John Wiley, New York.

[14] Moustaki, I. (1996). A latent trait and a latent class model for mixed observed variables, *British Journal of Mathematical and Statistical Psychology*, **49**, 313–334.

[15] Moustaki, I. and Knott, M. (1997). *Generalized latent trait models*, Statistics research report 36, London School of Economics.

[16] Moustaki, I. and Knott, M. (2000). Generalized latent trait models, *Psychometrika*, **65**, 391–411.

[17] Moustaki, I. and Victoria-Feser, M.-P. (2006). Bounded-bias robust inference for generalized linear latent variable models, *Journal of the American Statistical Association*, **101**, 644–653.

[18] Sammel, M.D.; Ryan, L.M. and Legler, J.M. (1997). Latent variable models for mixed discrete and continuous outcomes, *Journal of the Royal Statistical Society, Series B*, **59**, 667–678.

[19] Schilling, S. and Bock, D.R. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature, *Psychometrika*, **70**, 533–555.

[20] Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*, Chapman and Hall, London.

REVSTAT – Statistical Journal
Volume 5, Number 1, March 2007, 97–114

APPLICATION OF ROBUST STATISTICS TO ASSET ALLOCATION MODELS

Authors: Roy E. Welsch
– Sloan School of Management,
Massachusetts Institute of Technology, USA
rwelsch@mit.edu

Xinfeng Zhou
– Global Index and Markets Group,
Barclays Global Investors, USA
xinfeng.zhou@barclaysglobal.com

Abstract:

- Many strategies for asset allocation involve the computation of the expected value and the covariance matrix of the returns of financial instruments. How much of each instrument to own is determined by an attempt to minimize risk — the variance of linear combinations of investments in these financial assets — subject to various constraints such as a given level of return, concentration limits, etc. The covariance matrix contains many parameters to estimate and two main problems arise. First, the data will very likely have outliers that will seriously affect the covariance matrix. Second, with so many parameters to estimate, a large number of return observations are required and the nature of markets may change substantially over such a long period. In this paper we discuss using robust covariance procedures, FAST-MCD, Iterated Bivariate Winsorization and Fast 2-D Winsorization, to address the first problem and penalization methods for the second. When back-tested on market data, these methods are shown to be effective in improving portfolio performance. Robust asset allocation methods have great potential to improve risk-adjusted portfolio returns and therefore deserve further exploration in investment management research.

Key-Words:

- *robust statistics; asset allocation; FAST-MCD; bivariate Winsorization; penalization.*

AMS Subject Classification:

- 62G35, 62P20, 91B28.

1. INTRODUCTION

Asset allocation is the process that investors use to determine the asset classes in which to invest and the weight for each asset class. Past studies have shown that asset allocation explains 75 − 90% of the return variation and is the single most important factor determining the variability of portfolio performance. Among all the asset allocation models, Harry Markowitz's mean-variance portfolio theory is by far the most well-known and well-studied model for both academic researchers and practitioners alike [17, 18]. The crux of mean-variance portfolio theory assumes that investors prefer lower standard deviations/variances for a given level of expected return. Portfolios that provide the minimum standard deviation for a given expected return are termed efficient portfolios and those that do not are termed inefficient portfolios.

For a portfolio with N risky assets to invest in, the portfolio return is the weighted average return of each asset

$$r_p \equiv w_1 r_1 + w_2 r_2 + \cdots + w_N r_N = \boldsymbol{w}'\boldsymbol{r} \tag{1.1}$$

and the expected return and the variance of the portfolio can be expressed as

$$\begin{aligned} \mu_p &= w_1\mu_1 + w_2\mu_2 + \cdots + w_N\mu_N = \boldsymbol{w}'\boldsymbol{\mu} \; , \\ \mathrm{var}(r_p) &= \mathrm{var}\big(w_1 r_1 + w_2 r_2 + \cdots + w_N r_N\big) = \boldsymbol{w}'\boldsymbol{\Sigma}\,\boldsymbol{w} \; , \end{aligned} \tag{1.2}$$

where w_i, $\forall\, i=1,...,N$, is the weight of the i-th asset in the portfolio; r_i is the return of the i-th asset in the portfolio; μ_i is the expected return of the i-th asset in the portfolio; $\boldsymbol{w}$ is a $N{\times}1$ column vector of w_i's; $\boldsymbol{r}$ is a $N{\times}1$ column vector of r_i's; $\boldsymbol{\mu}$ is a $N{\times}1$ column vector of μ_i's; and $\boldsymbol{\Sigma}$ is the $N{\times}N$ covariance matrix of the returns of N assets.

We can formulate the following problem to assign optimal weight to each asset and identify the efficient portfolio:

$$\min_{\boldsymbol{w}} \boldsymbol{w}'\boldsymbol{\Sigma}\,\boldsymbol{w} \qquad \text{s.t.} \quad \boldsymbol{w}'\boldsymbol{\mu} = \mu_p, \quad \boldsymbol{w}'\boldsymbol{e} = 1 \; , \tag{1.3}$$

where μ_p is the expected portfolio return and $\boldsymbol{e}$ is $N{\times}1$ column vector with all elements 1. For each specified μ_p, the problem can be solved in closed form using the method of Lagrange [23]. The simple mean-variance optimization only requires two inputs–expected return vector and expected covariance matrix. The model is based on a formal quantitative objective that will always give the same solution with the same set of parameters. These all explain its popularity and its contribution to modern portfolio theory (MPT).

Nevertheless, the original form of mean-variance portfolio optimization has rarely been applied in practice because of several drawbacks. The method uses variance as the risk measure, which is often considered to be a simplistic measurement when the asset returns do not follow normal distributions. In reality, many of the financial assets' returns do have fat tails or are skewed. Besides, the one-period nature of static optimization also does not take dynamic factors into account, and some researchers argue for more complicated models based on stochastic processes and dynamic programming. However, the most serious problem of the mean-variance efficient frontier is probably the method's instability. The mean-variance frontier is very sensitive to the inputs, and these inputs are subject to random errors in the estimation of expected return and covariance. Small and statistically insignificant changes in these estimates can lead to a significant change in the composition of the efficient frontier. This may lead us to frequently and mistakenly rebalance our portfolio to stay on this elusive efficient frontier, incurring unnecessary transaction costs.

The Markowitz portfolio optimization estimates the expected return and the covariance matrix from historical return time series and treats them as true parameters for portfolio selection. The historical returns for N assets over T periods are denoted as $\boldsymbol{R}$, a $T \times N$ matrix where each column vector $\boldsymbol{r}_i$, $\forall\, i = 1, ..., N$, represents the returns of asset i over different periods and each row vector $\boldsymbol{R}_t$, $\forall\, t = 1, ..., T$, represents the returns of different assets at period t. The simple sample mean and covariance matrix are used as the parameters since they are the best unbiased estimators under the assumption of multivariate normality. Despite the simple computation involved, this approach has high complexity (large number of parameters). It suffers from the problem of high variance, which means the estimation errors can be significant and generate erroneous mean-variance efficient frontiers. This naïve "certainty equivalence" mean-variance approach often leads to extreme portfolio weights (instead of a diversified portfolio as the method anticipates) and dramatic swings in weights when there is a minor change to the expected returns or the covariance matrix [7, 10, 12]. The problem is further exacerbated if the number of observations is of the same order as the number of assets, which is often the case in financial applications to select industry sectors or individual securities.

A number of alternative models have been developed to improve parameter estimation. For example, factor-based models try to reduce the model complexity (number of parameters) by explaining asset return variances/covariances using a limited number of common factors. Multivariate GARCH models try to address fat tails and volatility clustering by incorporating the time dependence of returns in the covariance matrix. But neither approach effectively reduces or eliminates the influences of outliers in the data. A small percentage of outliers, in some cases even a single outlier, can distort the final estimated variance and covariance. Evidence has shown that the most extreme (large positive or negative) coefficients in the estimated covariance matrix often contain the largest error and as a result, mean-variance optimization based on such a matrix routinely gives the heaviest weights — either positive or negative — to those coefficients that are most unreliable. This "error-maximization" phenomenon [24] causes the mean-variance technique to behave very badly unless such errors are corrected.

In this study, we focus on investigating robust statistical approaches to reduce the influence of outliers, to increase the stability of the portfolio and to reduce asset turnover. The remainder of the paper is organized as follows. In Section 2, we investigate and extend some robust statistical methods such as FAST-MCD, Iterated Bivariate Winsorization, and Fast 2-D Winsorization to estimate the covariance matrix. We also explore penalization methods as a direct way to reduce asset turnovers. In Section 3, we apply these methods to construct US industrial selection portfolios and show that these robust methods dramatically improve risk-adjusted portfolio performance, especially when transaction costs are taken into consideration. In Section 4, we conclude this paper by summarizing our findings and offering possible directions for future research.

2. METHODS

During the past decade, statisticians have developed a variety of robust estimation methods to estimate both the mean and the covariance matrix [4, 8, 19, 20]. However, the use of robust estimators has received relatively little attention in the finance literature overall, and in the context of estimating the expected value and the covariance matrix of asset returns in particular [13, 22]. In this study, we take the initiative to investigate the value of some robust approaches to asset allocation problems.

2.1. FAST-MCD

The general principle of robust statistical estimation is to give full weights to observations assumed to come from the main body of the data, but to reduce or completely eliminate weights for the observations from tails of the contaminated data. The minimum covariance determinant (MCD) method [3], a robust estimator introduced by Rousseeuw in 1985, eliminates perceived outliers from the estimation of the mean and the covariance matrix. It uses the mean and the covariance matrix of h data points $(T/2 \leqslant h < T)$ with the smallest determinant to estimate the population mean and the covariance matrix. The method has a break-down value of $(T-h)/T$. If the data come from a multivariate normal distribution, the average of the optimal subset is an unbiased estimator of the population mean. The resulting covariance matrix is biased, but a finite sample correction factor $(c_{h,T} \geq 1)$ can be used to make the covariance matrix unbiased. The multiplication factor $c_{h,T}$ can be determined through Monte Carlo simulation. For our specific purpose, the bias by itself does not affect the asset allocation since all pairs of covariances are underestimated by the same factor.

MCD has rarely been applied to high-dimensional problems because it is extremely difficult to compute. MCD estimators are solutions to highly non-convex optimization problems that have exponential complexity of the order 2^N in terms of the dimension N of the data. Therefore, these original methods are not suitable for asset allocation problems when $N > 20$. Yet, in practice, asset allocation problems often include dozens of industrial classes or hundreds of individual securities, which makes the MCD method computationally infeasible. In order to cope with computational complexity problems, a heuristic FAST-MCD algorithm developed by Rousseeuw and Van Driessen [25], provides an efficient alternative. A naïve MCD approach would compute the MCD for up to $\binom{T}{h}$ subsets, while FAST-MCD uses sampling to reduce the computation and usually offers a satisfactory heuristic estimation. Other equivariant robust covariance methods are discussed in a recent book [20] and we are experimenting with the S-estimator they recommend, SR-05.

The key step of the FAST-MCD algorithm takes advantage of the fact that, starting from any approximation to the MCD, it is possible to compute another approximation with a determinant no higher than the current one. The method is based on the following theorem related to a concentration step (C-step):

Let $H_1 \subset \{1, ..., n\}$ be any h-subset of the original cross-sectional data, put $\hat{\mu}_1 = \frac{1}{h}\sum_{t\in H_1}\boldsymbol{R}_t$ and $\hat{\boldsymbol{\Sigma}}_1 = \frac{1}{h}\sum_{t\in H_1}(\boldsymbol{R}_t - \hat{\boldsymbol{\mu}}_1)\,(\boldsymbol{R}_t - \hat{\boldsymbol{\mu}}_1)'$. If $\det(\hat{\boldsymbol{\Sigma}}_1) \neq 0$, define the distance $d_1(t) = \sqrt{(\boldsymbol{R}_t - \hat{\boldsymbol{\mu}}_1)\,\hat{\boldsymbol{\Sigma}}_1^{-1}(\boldsymbol{R}_t - \hat{\boldsymbol{\mu}}_1)'}$, $t = 1, ..., T$. Now take H_2 such that $\{d_1(i);\ i \in H_2\} := \{(d_1)_{1:T}, ..., (d_1)_{h:T}\}$ where $(d_1)_{1:T} \leq (d_1)_{2:T} \leq \cdots \leq (d_1)_{T:T}$ are the ordered distances, and compute $\hat{\boldsymbol{\mu}}_2$ and $\hat{\boldsymbol{\Sigma}}_2$ based on H_2. Then $\det(\hat{\boldsymbol{\Sigma}}_2) \leq \det(\hat{\boldsymbol{\Sigma}}_1)$ with equality if and only if $\hat{\boldsymbol{\mu}}_2 = \hat{\boldsymbol{\mu}}_1$ and $\hat{\boldsymbol{\Sigma}}_2 = \hat{\boldsymbol{\Sigma}}_1$.

If $\det(\hat{\boldsymbol{\Sigma}}_1) > 0$, the C-step yields $\hat{\boldsymbol{\Sigma}}_2$ with $\det(\hat{\boldsymbol{\Sigma}}_2) \leq \det(\hat{\boldsymbol{\Sigma}}_1)$. Basically the theorem indicates the sequence of determinants obtained through C-steps converges in a finite number of steps from any original h-subset to a subset satisfying $\det(\hat{\boldsymbol{\Sigma}}_{m+1}) = \det(\hat{\boldsymbol{\Sigma}}_m)$. Afterward, running the C-step no longer reduces the determinant. However, this process only guarantees that the resulting $\det(\hat{\boldsymbol{\Sigma}})$ is a local minimum instead of the global one. To yield the h-subset with global minimum $\det(\hat{\boldsymbol{\Sigma}})$ or at least close to optimal, many initial choices (often > 500) of H_1 are taken and C-steps are applied to each.

Simulated and empirical results showed that FAST-MCD typically gives "good" results and is orders of magnitude faster than exact MCD methods. Yet, the FAST-MCD method still requires substantial running times for large N and T, and the probability of retaining outliers in the final h-subset increases when N becomes large. We use the FAST-MCD as an affine equivariant benchmark for faster non-equivariant methods. Other examples of its use are contained in [26, 30].

2.2. Iterated bivariate Winsorization (I2D-Winsor)

The FAST-MCD estimator for the covariance matrix is positive semidefinite and affine equivariant, which means the estimator behaves properly under affine transformations of the data. If the affine equivariance requirement is dropped, much faster estimators with high breakdown points can be computed. These methods are often based on pair-wise robust correlation or covariance estimates such as coordinate-wise outlier insensitive transformations (e.g. Huber-function transformation, quadrant correlation) and bivariate outlier resistant models. All these methods have quadratic complexity in the number of variables and linear complexity in the number of observations, so they reduce the computational complexity to $O(N^2 T)$.

Huber's function, defined as $H_c(x) = \min\{\max\{-c, x\}, c\}$, $c > 0$, has been widely used to shrink outliers towards the median by the transformation

$$\tilde{r}_{ti} = m_i + s_i \times H_c\big((r_{ti} - m_i)/s_i\big) \ , \tag{2.1}$$

where m_i and s_i are the median and the median absolute deviation from the median of return vector $\boldsymbol{r}_i$. Essentially Huber's function brings the outliers of each variable to the boundary $m_i \pm c \times s_i$ and, as a result, reduces the impact of outliers.

The one-dimensional Winsorization approach using the Huber function has been a popular method in finance because of its intuitive appeal and easy computation. Yet for covariance analysis, the method fails to take the orientation of the bivariate data into consideration. To address the problem, bivariate Winsorization methods have also been investigated. For each pair of variables, outliers are shrunken to the border of an ellipse which includes the majority of the data by using the bivariate transformation

$$\tilde{\boldsymbol{r}}_{t,i,j} = \boldsymbol{\mu}_0 + \min\Big(\sqrt{c/D(\boldsymbol{r}_{t,i,j})},\, 1\Big)\,(\boldsymbol{r}_{t,i,j} - \boldsymbol{\mu}_0) \ , \tag{2.2}$$

where, for each pair of $\boldsymbol{r}_i$ and $\boldsymbol{r}_j$, $\boldsymbol{r}_{t,i,j} = \begin{bmatrix} r_{ti} \\ r_{tj} \end{bmatrix}$; $\boldsymbol{\mu}_0 = \begin{bmatrix} m_i \\ m_j \end{bmatrix}$; $D(\boldsymbol{r}_{t,i,j})$ is the Mahalanobis distance based on an initial bivariate covariance matrix $\boldsymbol{\Sigma}_0$ and location $\boldsymbol{\mu}_0$: $(\boldsymbol{r}_{t,i,j} - \boldsymbol{\mu}_0)'\, \boldsymbol{\Sigma}_0^{-1}\, (\boldsymbol{r}_{t,i,j} - \boldsymbol{\mu}_0)$; c is a positive constant. The transformation shrinks the outlier towards $\boldsymbol{\mu}_0$ when $D(\boldsymbol{r}_{t,i,j}) > c$.

Based on the idea of shrinking data toward the border of a two-dimensional ellipse, Chilson et al. developed an iterated bivariate Winsorization (I2D-Winsor) method to estimate covariance and applied the method to cluster correlated genes [5]. The method includes the following three steps:

Step A. For each pair of variables $\boldsymbol{r}_i$ and $\boldsymbol{r}_j$, compute a simple robust mean and adjusted MAD for each column and construct the initial estimate of mean and covariance matrix as

$$\boldsymbol{\mu}_0 = \begin{bmatrix} m_i \\ m_j \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{\Sigma}_0 = \begin{bmatrix} \dfrac{s_i}{0.6745} & 0 \\ 0 & \dfrac{s_j}{0.6745} \end{bmatrix} . \tag{2.3}$$

Step B. For each $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, calculate the Mahalanobis distance for each return pair

$$D_{t,k} = (\boldsymbol{r}_{t,i,j} - \boldsymbol{\mu}_k)' \, \boldsymbol{\Sigma}_k^{-1} \, (\boldsymbol{r}_{t,i,j} - \boldsymbol{\mu}_k) \tag{2.4}$$

and then calculate the weight for each $\boldsymbol{r}_{t,i,j}$ as

$$z_t = \min\Big(\sqrt{c/D_{t,k}}\,,\, 1\Big) , \tag{2.5}$$

where the constant c is chosen as 5.99 (the 95% quantile of the χ_2^2 distribution).

Step C. Update $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ to $\boldsymbol{\mu}_{k+1}$ and $\boldsymbol{\Sigma}_{k+1}$ using equations

$$\begin{aligned} \boldsymbol{\mu}_{k+1} &= \sum_{i=1}^{T} z_t \, \boldsymbol{r}_{t,i,j} \bigg/ \sum_{i=1}^{T} z_t \, , \\ \boldsymbol{\Sigma}_{k+1} &= \sum_{i=1}^{T} z_t^2 \, (\boldsymbol{r}_{t,i,j} - \boldsymbol{\mu}_{k+1}) \, (\boldsymbol{r}_{t,i,j} - \boldsymbol{\mu}_{k+1})' \bigg/ \sum_{i=1}^{T} z_t^2 \, . \end{aligned} \tag{2.6}$$

This iteration is repeated until $\boldsymbol{\mu}_{k+1}, \boldsymbol{\Sigma}_{k+1}$ and $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ converge as determined by the sum of absolute differences between two consecutive $\boldsymbol{\Sigma}$ being less than a predefined error. The covariance matrix of variables $\boldsymbol{r}_i$ and $\boldsymbol{r}_j$ is then set to $\boldsymbol{\Sigma}_{k+1}$. Diagonal elements of the covariance matrix are obtained using bias adjusted median absolute deviations from the median.

The I2D-Winsor method allowed parallel computation of high dimensional correlation and covariance matrices for different gene expressions and obtained good performance in heterogeneous cluster studies. But the method suffers the drawback of failing to guarantee positive semidefiniteness of the covariance matrix — a crucial requirement for mean-variance portfolio optimization. Maronna et al. [21] proposed an adjustment method to obtain a positive semidefinite covariance matrix using a pair-wise robust covariance matrix. The method is based on the observation that any positive semidefinite covariance matrix $\boldsymbol{C}$ can be expressed as $\boldsymbol{C} = \sum \hat{\lambda}_i \, \hat{\boldsymbol{a}}_i \, \hat{\boldsymbol{a}}_i'$, where $0 \le \hat{\lambda}_1 \le \cdots \le \hat{\lambda}_N$ are the eigenvalues and $\hat{\boldsymbol{a}}_i$ $(i = 1, ..., N)$ are the corresponding eigenvectors. If $\boldsymbol{C}$ is not positive semidefinite, then one or more of the eigenvalues are negative. To convert such a matrix to a positive semidefinite one, a natural approach is to replace these negative eigenvalues with positive ones. When $\boldsymbol{C}$ is the sample correlation, $\hat{\lambda}_i$'s are the variances of the projected data on the direction of the corresponding eigenvectors.

This indicates that in order to get rid of possibly negative eigenvalues in the quadrant covariance matrix $\hat{\boldsymbol{C}}_0$, one can replace the $\hat{\lambda}_i$'s in $\boldsymbol{C}_0 = \sum \hat{\lambda}_i \, \hat{\boldsymbol{a}}_i \, \hat{\boldsymbol{a}}_i'$ by the square of robust standard deviation estimates for the projected data. We can compute the decomposition of $\hat{\boldsymbol{C}}_0$: $\hat{\boldsymbol{C}}_0 = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}'$, where $\boldsymbol{Q}$ is the orthogonal matrix of eigenvectors and $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues. Then we can transform $\boldsymbol{R}$ to $\tilde{\boldsymbol{R}}$ using the new basis $\boldsymbol{Q}$: $\tilde{\boldsymbol{R}} = \boldsymbol{R}\boldsymbol{Q}'$ and compute the robust standard deviation estimate $(\tilde{s}_j/0.6745)$ of the columns of $\tilde{\boldsymbol{R}}$. Let $\tilde{\boldsymbol{D}}$ be the diagonal matrix whose elements are $(\tilde{s}_j/0.6745)^2$ ordered from largest to smallest. The final positive definite robust covariance matrix is $\hat{\boldsymbol{\Sigma}} = \boldsymbol{Q}\tilde{\boldsymbol{D}}\boldsymbol{Q}'$.

By transforming the I2D-Winsor robust covariance matrix using Maronna's adjustment method, we guarantee the positive semidefiniteness of the final covariance matrix and make it directly applicable to asset allocation problems.

2.3. Fast 2-D Winsorization (F2D-Winsor)

Khan et al. [11] proposed a fast two-step, two-dimensional Winsorization method (F2D-Winsor) while investigating ways to make least-angle regression (LARS) robust. Instead of repeated iteration of step B in I2D-Winsor, which is computationally expensive, Khan's method only implements step B once. In order to achieve a similar level of robustness as I2D-Winsor, F2D-Winsor constructs an informative initial covariance matrix. We again combine F2D-Winsor ideas from Khan's paper and Maronna's method to guarantee the positive semidefiniteness of the covariance matrix and design the following F2D-Winsor method:

Step A. *Initial covariance estimate.* For each pair of variables $\boldsymbol{r}_i$ and $\boldsymbol{r}_j$, compute simple robust location (median) and scale (adjusted MAD) estimates for each variable. We then compute an initial covariance matrix using Khan's adjusted Winsorization method that is more resistant to bivariate outliers [11]. In the adjusted Winsorization method, two tuning parameters are used with c_1 for the two quadrants (separated by m_i and m_j) that contain the majority of the data and a smaller constant c_2 for the other two quadrants. For example, c_1 can be taken to be 1.96 ($\mu \pm 1.96\,\sigma$ includes 95% of the data from the normal distribution) and $c_2 = hc_1$ where $h = n_2/n_1$ with n_1 the number of observations in the major quadrants and $n_2 = T - n_1$, where T is the total number of observations. As shown in Figure 1, the data are now shrunk to the boundary of the four smaller rectangles instead of a large rectangle. As a result, the adjusted Winsorization method handles bivariate outliers better than the univariate Winsorization. However, it does raise a problem that the initial covariance matrix constructed from pairwise covariance may not be positive definite. To address the problem, Maronna's transformation is applied to convert the initial covariance matrix $\boldsymbol{\Sigma}_0$ to a positive definite one.

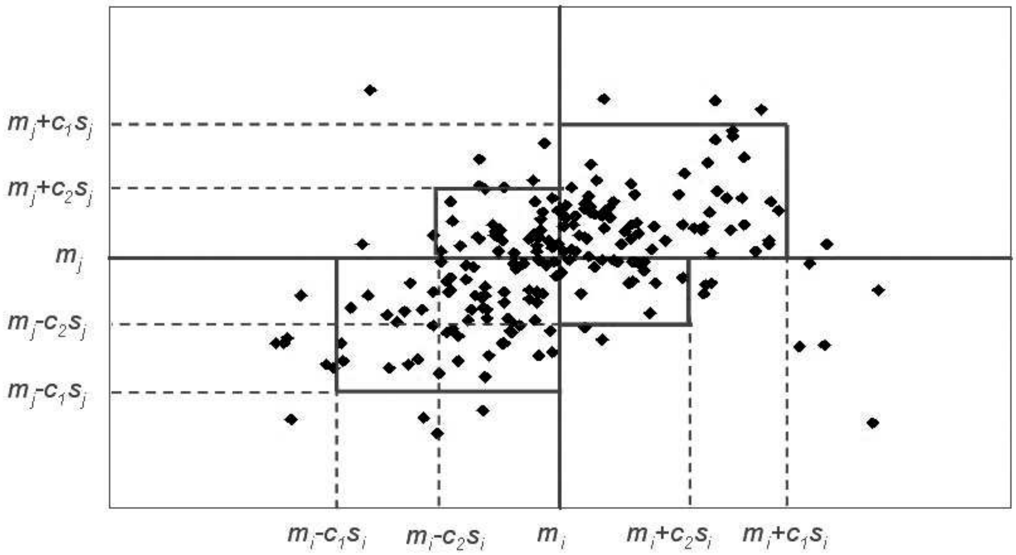

Figure 1: Adjusted Winsorization (for initial covariance) with $c_1 = 1.96$, where s_i and s_j are estimated from adjusted MAD.

Step B. *2D-Winsorization based covariance matrix.* For each pair of $(\boldsymbol{r}_i, \boldsymbol{r}_j)$, outliers are shrunk to the border of an ellipsoid by using the transformation $\tilde{\boldsymbol{r}}_{t,i,j} = \boldsymbol{\mu}_0 + \min\bigl(\sqrt{c/D_{t,0}}\,, 1\bigr)\,(\boldsymbol{r}_{t,i,j} - \boldsymbol{\mu}_0)$, with constant $c = 5.99$ (the 95% quantile of the χ^2_2 distribution). The covariance for each pair is calculated using this modified data. Maronna's transformation is again applied to guarantee the positive definiteness of the final covariance matrix.

2.4. L1-penalized mean-variance method (V1)

All these robust covariance matrix estimation methods try to increase the stability of the allocation model by increasing the stability of the mean and covariance matrix of returns over time. Since the influence of outliers is reduced, the updated return data tend to have less impact on the robust mean and covariance matrix, even if some of the new return vectors contain extreme values. In this sub-section, we also implement a different class of penalization-based robust estimators to directly increase model stability and reduce turnover.

If the expected return and covariance matrix are estimated from the historical sample $\boldsymbol{R}_1, ..., \boldsymbol{R}_T$, the original mean-variance portfolio optimization problem

$$\min_{\boldsymbol{w}} \boldsymbol{w}'\,\boldsymbol{\Sigma}\,\boldsymbol{w} \qquad \text{s.t.} \quad \boldsymbol{w}'\boldsymbol{\mu} = \mu_p, \quad \boldsymbol{w}'\boldsymbol{e} = 1 \ , \tag{2.7}$$

can be rewritten as

$$\min_{\boldsymbol{w},q} \frac{1}{T}\sum_{i=1}^{T}(\boldsymbol{w}'\boldsymbol{R}_t - q)^2 \qquad \text{s.t.} \quad \boldsymbol{w}'\boldsymbol{\mu} = \mu_p, \quad \boldsymbol{w}'\boldsymbol{e} = 1\ . \tag{2.8}$$

Lauprete [14] and Lauprete, et al. [15] proposed penalizing deviations from the market weights $(w_{m,i} = M_i/\sum_{j=1}^{N} M_j$ with M_j being the market value of asset $j)$ as a possible way to reduce the influences of outliers and to reduce turnover. These authors also considered using robust loss functions (M-estimators) in place of least-squares loss in (2.8). A recent paper by DeMiguel and Nogales [6] replaces M-estimators with S-estimators but omits any penalty term. If the market is efficient (or nearly efficient as many researchers believe), a penalty term serves as the prior in our optimization problem. We should penalize the final cost function if the proposed asset weights deviate from the prior. As a result, extreme deviations from the prior are unlikely. In this study, we focused on an L1 regularization method, which was the penalty function used in LASSO regression [27]. The regularized portfolio estimator can be expressed as [14]:

$$\begin{aligned}\big(\boldsymbol{w}(\lambda), q(\lambda)\big) \;&=\; \arg\min_{q\in R}\left(\frac{1}{T}\sum_{i=1}^{T}(\boldsymbol{w}'\boldsymbol{R}_t - q)^2 + \lambda|\boldsymbol{w} - \boldsymbol{w}_m|\right) \\ &\text{s.t.} \quad \boldsymbol{w}'\boldsymbol{\mu} = \mu_p, \quad \boldsymbol{w}'\boldsymbol{e} = 1\ ,\end{aligned} \tag{2.9}$$

where $\lambda > 0$ is the regularization parameter; $|\boldsymbol{w} - \boldsymbol{w}_m|$ is the L_1-norm of $\boldsymbol{w} - \boldsymbol{w}_m$: $\sum_{i=1}^{N}|w_i - w_{m,i}|$.

The term $\lambda|\boldsymbol{w} - \boldsymbol{w}_m|$ reflects the investor's a priori confidence in the market portfolio $\boldsymbol{w}_m$. A large λ means large penalty for any deviation and strong confidence in $\boldsymbol{w}_m$; a small λ reflects weak confidence in $\boldsymbol{w}_m$. We choose the parameter λ using 5-fold cross validation. For any given λ, we implement the following steps:

Step A. Divide the T observations randomly into 5 subsets of $T/5$ observations. Call these subsets $T(i)$ for $i = 1, ..., 5$. For every i, run the optimization to yield the optimal $\big(\hat{\boldsymbol{w}}(\lambda), \hat{q}(\lambda)\big)$ for the in-sample data:

$$\begin{aligned}\big(\hat{\boldsymbol{w}}(\lambda), \hat{q}(\lambda)\big) \;&=\; \arg\min_{q\in R}\left(\frac{1}{0.8\,T}\sum_{t\in T\setminus T(i)}(\boldsymbol{w}'\boldsymbol{R}_t - q)^2 + \lambda|\boldsymbol{w} - \boldsymbol{w}_m|\right) \\ &\text{s.t.} \quad \boldsymbol{w}'\boldsymbol{\mu} = \mu_p, \quad \boldsymbol{w}'\boldsymbol{e} = 1\ .\end{aligned} \tag{2.10}$$

Step B. For every $i = 1, ..., 5$ apply $\big(\hat{w}(\lambda), \hat{q}(\lambda)\big)$ to the out-of-sample data to calculate a sum of squared errors, $PE_\lambda(i) = \sum_{t\in T(i)}\big[\big(\hat{\boldsymbol{w}}_\lambda(i)'\boldsymbol{R}_t - \hat{q}_\lambda(i)\big)^2\big]$.

Step C. Calculate the total sum of squared errors $PE_\lambda = \sum_{i=1}^{5} PE_\lambda(i)$.

A series of candidate values of λ from 0.01 to 2 are tested to yield a value of λ with minimum total sum of squared errors PE_λ. Once λ is selected, $\boldsymbol{w}(\lambda)$ and $q(\lambda)$ can be solved as the "optimal" solution to the corresponding quadratic optimization problem. The lower bound of 0.01 was found by experimentation and may be different for other data sets.

3. APPLICATION RESULTS

In this section, we show a real asset allocation application using daily returns on 51 MSCI US industry sector indexes, from 01/03/1995 to 02/07/2005 (2600 trading days of data). Combining the stocks in these industry indexes ($\sim$700 stocks included) forms a general index for US equity markets broader than the S&P 500. The robust methods discussed in Section 2 are applied to find the "optimal" weights for each industry.

For every estimator, we use the following portfolio rebalancing strategy: estimate the industry sector weights using the most recent 100 daily returns and rebalance the portfolio weights every five trading days (a week). Since there are 2600 trading days in the data, there are 500 rebalances in total. In practice, there are transaction costs when we change the weights of each asset using updated information. So we will compare the results both without considering transaction costs and with 5 cents for each $100 bought or sold. We apply a target return constraint and convexity constraint to all estimates:

$$\boldsymbol{w}'\boldsymbol{\mu} = \mu_p \,, \qquad \boldsymbol{w}'\boldsymbol{e} = 1 \,. \tag{3.1}$$

The resulting stream of ex-post portfolio returns is collected for each estimator/target return combination. We calculate the following statistics of the ex-post returns of each estimator/target return combination:

Mean: the sample mean of weekly ex-post returns;

STD: the sample standard deviation of weekly ex-post returns;

Information Ratio: $IR = mean/STD \times \sqrt{52}$, where the standardization by $\sqrt{52}$ makes the information ratio an annual estimate assuming 260 trading days per year;

α**-VaR** for $\alpha = 5\%$ and 1%: the loss at the α-quantile of the weekly ex-post return;

MaxDD: the maximum drawdown, which is the maximum loss in a week;

CRet: cumulative return;

Turnover: weekly asset turnover, defined as the mean of the absolute weight changes $\left(\sum_{i=1}^{51} |w_{t,i} - w_{t-1,i}|\right)$ for 500 updates;

Cret_cost: cumulative return with transaction costs;

IRcost: Information ratio with transaction costs.

Except for the market model, which uses market weights and the corresponding market returns, a range of target expected annual portfolio returns from 10% to 20% are used for portfolio construction. Table 1 shows the summarized results for annual expected return $\mu_p = 15\%$ for V (mean-variance optimization with simple mean and covariance matrix), FAST-MCD, I2D-Winsor, F2D-Winsor, V1 models and market index. More extensive tables are in Zhou [31].

Table 1: Performance of V, FAST-MCD, I2D-Winsor, F2D-Winsor, V1 models and Market index for $\mu_p = 15\%$. For FAST-MCD, I2D-Winsor and F2D-Winsor, the median instead of the mean of the returns is used as the expected return of each asset.

$\mu_p = 15\%$	V	FAST-MCD	I2D-Winsor	F2D-Winsor	V1	Market
mean	0.065%	0.096%	0.156%	0.155%	0.198%	0.160%
STD	1.962%	2.025%	1.948%	2.007%	2.431%	2.343%
IR	0.239	0.341	0.578	0.558	0.589	0.491
VaR(0.05)	3.06%	3.10%	3.10%	3.23%	3.71%	4.06%
VaR(0.01)	5.78%	6.33%	5.80%	5.52%	6.65%	5.28%
MaxDD	−7.48%	−8.57%	−9.39%	−9.40%	−8.35%	−10.01%
Cret	1.256	1.457	1.983	1.965	2.328	1.935
Cret_cost	0.845	0.888	1.801	1.803	2.252	1.923
IRcost	−0.054	−0.013	0.507	0.497	0.569	0.487
Turnover	1.59	1.99	0.39	0.35	0.13	0.02

Both the pair-wise Winsorization methods and the penalization method yield significantly better results than mean-variance optimization with the simple mean and covariance matrix as inputs. The V method has significant asset turnover (159%) and as a result the IRcost — the most popular performance measure — is negative after the transaction costs are taken into consideration. In contrast, I2D-Winsor, F2D-Winsor and V1 methods have much lower turnovers (0.39, 0.35 and 0.13 respectively) and yield an IRcost of 0.507, 0.497 and 0.569 respectively, which are much higher than the V method. All these methods also beat the market in VaR (5%), MaxDD and IRcost, which clearly shows their value in active portfolio management.

The benefit of FAST-MCD is modest compared with the V method and it is inferior to the market. The reason most likely lies in the strict assumptions of the MCD approaches. Although both MCD methods and pair-wise robust estimators are designed to eliminate the effects of outliers, MCD models use a restrictive contamination model assuming complete dependence of outliers for different assets. Basically MCD models assume that each row of returns, $\boldsymbol{R}_t$, is either from the core distribution F_0 or outlier generating distribution H. The data are from the following mixed model:

$$F = (1-\varepsilon)F_0 + \varepsilon H\ , \qquad 0<\varepsilon<\frac{1}{2}\ , \tag{3.2}$$

where F is the mixed model; F_0 is a multivariate normal distribution; H is an arbitrary multivariate distribution that generates outliers.

Such a contamination model is rather restrictive for our application. By looking at N-dimensional outliers, the models assume that all asset returns for any given day are either from a core distribution F_0 or outlier generating distribution H. This assumption is only true if the market is the only factor that determines asset returns or there are high correlations between different assets' returns. In practice, the market return by itself only explains a small percentage of the variance of asset returns. Industrial factors and idiosyncratic risk have been shown to explain the majority of the return variances. The pair-wise models [1, 2] use a much more flexible mixed model for data:

$$\boldsymbol{R}_t = (\boldsymbol{I}-\boldsymbol{B})\boldsymbol{Y}_t + \boldsymbol{B}\boldsymbol{Z}_t\ , \tag{3.3}$$

with $\boldsymbol{B} = \operatorname{diag}\big([B_1\ B_2\ \cdots\ B_N]\big)$, $\boldsymbol{Y}_t$ multivariate normal, $\boldsymbol{Z}_t$ an arbitrary random vector, and the B_i, Bernoulli random variables with success probability ε_i. We can assume any format for the correlation matrix matrix of $(B_1, B_2, \ldots, B_N)$. MCD models assume complete dependence $B_1 = B_2 = \cdots = B_N$, while pair-wise models often assume independent B_i and B_j, $i \neq j$, or independently evaluate the correlation for each pair of B_i and B_j. As a result, pair-wise robust estimators offer more flexibility to calculate the covariances. Once the positive semidefiniteness property of the covariance matrix is guaranteed through transformation, they provide far better results than FAST-MCD.

As shown in Table 2, pair-wise Winsorization methods are also faster than the FAST-MCD method (10 hours) for the same data set. The sampling process of FAST-MCD is much faster than the original MCD method, but the C-steps still require extensive computation. Between the two pair-wise Winsorization methods, F2D-Huber (35 minutes) is faster because it eliminates the repeated iteration step in I2D-Winsor (3 hours), while I2D-Winsor is likely to yield a more robust estimation of the covariance and indeed gives slightly better results than F2D-Huber in our study. It is also worth noting that the estimated covariance matrix often slightly underestimates the real covariance, so the estimation is biased. Yet it is believed that for the constant $c = 5.99$ (the 95% quantile of the χ^2_2 distribution) that we chose, the bias would be small. Furthermore, the asset weights depend on the relative size of the covariance, so the impact of bias on our problem is even smaller.

Table 2: Run Time for 500 Rebalancings.
All programs were run on a computer
with 3 GHz CPU and 3 GB of RAM.

Time	V	FAST-MCD	I2D-Winsor	F2D-winsor	V1	Market
500 Rebalances	40 sec	10 hr	3 hr	35 min	4 hr	$<$ 4 sec

Penalization methods are more computationally intensive than pair-wise Winsorization methods. The addition of the penalty term extends the dimension of the optimization problems and increases the number of constraints. The cross-validation of each penalty coefficient λ increases the computation further by ~ 25 fold. Unlike robust estimation of the mean and covariance matrix, which only need to calculate the parameters once for all μ_p, the optimization problem needs to be performed for every μ_p. As a result, the run times of penalization methods are often longer.

Though computationally intensive, the V1 method using the market index as the prior carries great advantages. It yields the best information ratio with or without transaction costs. Because of the L1 penalty term, most asset weights are mainly restricted to the market weight, which dramatically reduces the asset turnover compared with pair-wise Winsorization methods. Penalization methods are especially valuable when the number of assets is of the same order of magnitude as the number of observations (in our study, $T = 2N$), since the covariance matrix is often ill-conditioned.

We also compared our methods with some of the factor-based models, e.g., CAPM model, Principal Component Analysis model, Shrinkage model ([16]) and multivariate GARCH models (e.g., Constant Conditional Correlation GARCH and Dynamic Conditional Correlation GARCH [28, 29]). The results [31] show that both pair-wise Winsorization methods and penalization methods perform better than these traditional approaches.

4. CONCLUSION

The implementation of the mean-variance portfolio optimization is limited in practice by difficulties in estimating model inputs, expected returns and the covariance matrices of different assets, and the sensitivity of asset weights assigned to these inputs. Traditionally, sample means and covariance matrices from historical data were used, which are subject to large estimation errors.

This paper investigates some of the recently developed robust statistical methods such as FAST-MCD, Iterative 2-D Winsorization, Fast 2-D Winsorization and penalization methods. These methods prove to be valuable tools in improving risk-adjusted portfolio performance and reducing asset turnover. Results also show that the V1 penalization method outperforms the 2-D Winsorization methods. However, they achieve this at the cost of significantly higher computational complexity. The computational problem may be overcome by the recently developed LARS algorithm [9]. LARS greatly speeds up computations for LASSO since all solutions for all λ can be found in about the same time as one-least-squares regression, which removes the need for a grid search on λ. If the LARS algorithm can be successfully applied to penalized portfolio optimization, then penalization methods can be used to allocate weights for 700 individual stocks directly instead of 51 sector index funds. This is work in progress.

ACKNOWLEDGMENTS

This research was supported, in part, by the Singapore-MIT Alliance and the MIT Center for Computational Research in Economics and Management Science.

REFERENCES

[1] Alqallaf, F.A.; Konis, K.P.; Martin, R.D. and Zamar, R.H. (2002). Scalable robust covariance and correlation estimates for data mining, *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Alberta, 14–23.

[2] Alqallaf, F.A.; Van Aelst, S.; Yohai, V.J. and Zamar, R.H. (2005). *A model for contamination of multivariate data*, Working Paper, University of British Columbia.

[3] Balvers, R.J.; Cosimano, T.F. and McDonald, B. (1990). Predicting stock returns in an efficient market, *Journal of Finance*, **45**(4), 1109–1128.

[4] Brunelli, R. and Messelodi, S. (1995). Robust estimation of correlation with applications to computer vision, *Pattern Recognition*, **28**(6), 833–841.

[5] Chilson, J.; Ng, R.; Wagner, A. and Zamar, R. (2004). Parallel computation of high dimensional robust correlation and covariance matrices, *Proceedings of the ACM SIGKDD*, 533–538.

[6] DeMiguel, V. and Nogales, F.J. (2006). *Portfolio selection with robust estimation of risk*, Working Paper, London Business School.

[7] Dickenson, J. (1974). Some statistical aspects of portfolio analysis, *Statistician*, **23**(1), 5–16.

[8] Dueck, A. and Lohr, S. (2005). Robust estimation of multivariate covariance components, *Pattern Recognition*, **61**, 162–169.

[9] Efron, B.; Hastie, T. and Johnstone, I. (2004). Least angle regression, *Annals of Statistics*, **32**(2), 407–499.

[10] Jorion, P. (1976). International portfolio diversification with estimation risk, *Journal of Business*, **58**(3), 259–278.

[11] Khan, J.A.; Van Aelst, S. and Zamar, R.H. (2002). *Robust linear model selection based on least angle regression*, Working Paper, Statistics Department, University of British Columbia.

[12] Klein, R.W. and Bawa, V.S. (1976). Effect of estimation risk on optimal portfolio choice, *Journal of Financial Economics*, **3**, 215–231.

[13] Knez, P.J. and Ready, M.J. (1997). On the robustness of size and book-to-market in cross-sectional regressions, *Journal of Finance*, **52**(4), 1355–1382.

[14] Lauprete, G.J. (2001). *Portfolio risk minimization under departures from normality*, MIT PhD Thesis.

[15] Lauprete, G.; Samarov, A. and Welsch, R. (2002). Robust portfolio optimization, *Metrika*, **55**, 139–149. Also appeared in "Developments in Robust Statistics" (R. Dutter, P. Filzmoser, U. Gather and P. Rouseeuw, Eds.), Physica-Verlag, Heidelberg, 235–245.

[16] Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices, *Journal of Multivariate Analysis*, **88**(2), 365–411.

[17] Markowitz, H. (1952). Portfolio selection, *Journal of Finance*, **7**, 77–97.

[18] Markowitz, H. (1959). *Portfolio Selection*, Blackwell, Cambridge.

[19] Maronna, R.A. (1976). Robust M-estimators of multivariate location and scatter, *The Annals of Statistics*, **4**, 51–67.

[20] Maronna, R.A.; Martin, R.D. and Yohai, V.J. (2006). *Robust Statistics: Theory and Methods*, Wiley, New Jersey.

[21] Maronna, R.A. and Zamar, R.H. (2002). Robust estimates of location and dispersion for high-dimensional datasets, *Technometrics*, **44**(4), 307–317.

[22] Martin, R.D. and Simin, T.T. (2003). Outlier-resistant estimates of beta, *Financial Analysts Journal*, **59**(5), 56–69.

[23] Merton, R. (1972). An analytical derivation of the efficient portfolio frontier, *Journal of Financial and Quantitative Analysis*, **7**, 1851–1872.

[24] Michaud, R. (1989). The Markowitz optimization enigma: Is optimized optimal? *Financial Analysts Journal*, **45**, 31–42.

[25] Rousseeuw, P.J. and Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator, *Technometrics*, **41**(3), 212–223.

[26] Scherer, B. and Martin, R.D. (2005). *Introduction to Modern Portfolio Optimization with NUOPT and S-Plus*, Springer, New York.

[27] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society, Series B*, **58**(1), 267–288.

[28] Tse, Y.K. (2000). A test for constant correlations in a multivariate GARCH model, *Journal of Econometrics*, **98**(1), 107–127.

[29] Tse, Y.K. and Tsui, A.K.C. (2002). A multivariate generalized autoregressive conditional heteroscedasticity model with time-varying correlations, *Journal of Business*, **20**(3), 351–362.

[30] Vaz de Melo, B. and Câmara, R.P. (2003). Robust modeling of multivariate financial data, *COPPEAD Working Paper Series*, No. 355, Graduate School of Business, Federal University at Rio de Janeiro, Brazil.

[31] Zhou, X. (2006). *Application of robust statistics to asset allocation models*, MIT MS Thesis.

REVSTAT – Statistical Journal
Volume 5, Number 1, March 2007, 115–136

PENALIZED TRIMMED SQUARES AND A MODIFICATION OF SUPPORT VECTORS FOR UNMASKING OUTLIERS IN LINEAR REGRESSION *

Authors: G. Zioutas
– General Department, Faculty of Technology,
Aristotle University of Thessaloniki, Greece
zioutas@eng.auth.gr

A. Avramidis
– General Department, Faculty of Technology,
Aristotle University of Thessaloniki, Greece

L. Pitsoulis
– General Department, Faculty of Technology,
Aristotle University of Thessaloniki, Greece

Abstract:

- We consider the problem of identifying multiple outliers in linear regression models. We propose a penalized trimmed squares (PTS) estimator, where penalty costs for discarding outliers are inserted into the loss function. We propose suitable penalties for unmasking the multiple high-leverage outliers. The robust procedure is formulated as a Quadratic Mixed Integer Programming (QMIP) problem, computationally suitable for small sample data. The computational load and the effectiveness of the new procedure are improved by using the idea of ϵ-*insensitive* loss function from support vector machines regression. The small errors are ignored, and the mathematical formula gains the sparseness property. The good performance of the PTS estimator allows identification of multiple outliers avoiding masking effects.

Key-Words:

- *robust regression; mixed integer programming; penalty method; least trimmed squares; identifying outliers; support vector machines.*

AMS Subject Classification:

- 62F35, 62J99, 90C99.

*This research has been partially funded by the Greek Ministry of Education under the program Pythagoras II.

1. INTRODUCTION

In linear regression models data often contain outliers and bad influential observations. It is important to identify these observations and eliminate them from the data set. If the data are contaminated with a single or few outliers the problem of identifying such observations is not difficult. However, in most cases data sets contain more outliers or a group of masking outliers and the problem of identifying such cases becomes more difficult, due to masking effects.

The approaches to outlier identification can be separated into two categories: direct approaches and indirect approaches using residuals from the robust fit. Among famous direct approaches, Hadi and Simonoff [9] presented a procedure where it is attempted to separate the data into a set of "clean" data points (of size $k = (n+p-1)/2$) and a set of points that contain the potential outliers. The potential outliers are then tested to see how extreme they are relative to the clean subset, using an appropriate diagnostic measure like the adjusted residual, or Cook distance. Atkinson [1] proposed an identification method of multiple outliers by using a simple forward search starting from initial random subsets. The procedure requires again that at least one of the subsets does not contain high-leverage outliers. Peña and Yohai [14] proposed a successful fast procedure for detecting group of outliers in many situations, where due to masking effects the usual diagnostics procedures fail. However, they do not claim that their proposal keeps breakdown point of the original estimates. Their procedure has two stages; in the first stage high-leverage points eliminated from the data set irrespective of bad or good leverage points. Although in the second stage the efficiency is improved by testing again the potential outliers, some precision may be lost from the first stage. Generally, the key to the success of the above procedures is to obtain a clean initial subset of data. An indirect approach to outlier identification is through a robust regression estimate. If a robust estimate is relatively unaffected from outliers, then the residuals from the robust fit should be used to flag the outliers. A famous estimator that preserves high breakdown point (HBP) is the least trimmed squares LTS estimator of Rousseeuw and Leroy [16], that minimize the sum of the k, (coverage $k \geq [(n+p-1)/2]$) smallest squared residuals. But is well known that the LTS loses efficiency. Some better proposals obtain high breakdown points and simultaneously improve the efficiency of the LTS estimator. Among them are the S estimators of Rousseeuw and Yohai [18], the MM estimators of Yohai [24] Simpson, Ruppert, and Carroll [20] and Coakley and Hettmansperger [7], which combine good asymptotic efficiency under the normal linear model with HBP. These estimators, uses a less efficient high-breakdown method as an initial estimate, and then uses an M estimation strategy based on the redescending ψ function. Although they have achieved good asymptotic properties, may have low finite-sample efficiencies if the design contains high leverage points. Morgenthaler [12] and Stefanski [21] argue that no estimator with a breakdown point greater than $1/n$, can have high finite-sample efficiency in the presence of extreme leverage points. All these improvements to LTS achieve high breakdown point, improve the efficiency and have the bounded influence property. However, these estimators are based mainly on the initial LTS regression coefficient value. In practice, their performance depends heavily on the precision of the initial coefficient estimates. Sometimes, in data contaminated by high-leverage outliers, a bad initial coefficient value does not lead to a good final robust estimation. Moreover, the LTS method requires the coverage k or equivalently the number $n-k$ of the most likely outliers that produces the largest reduction in the residual sum of square when deleted. Unfortunately, this knowledge of k is typically unknown, Gentleman and Wilk [8].

In this article we propose a different approach penalized trimmed squares PTS, which does not require presetting the number $n-k$ of outliers to delete from the data set. The new estimator PTS is defined by minimizing a convex objective function (*loss function*), which is the sum of squared residuals and penalty costs for discarding bad observations. The robust estimate is obtained by the unique optimum solution of the convex mathematical formula called QMIP. The PTS estimator is very sensitive to the penalties defined a priori. In fact, these penalty costs are a function of the robust scale σ and leverage of the design points provided by the LTS and minimum covariance determinant MCD of Rousseeuw and Van Driessen [17]. In particular, these penalties in the loss function regulate the robustness and the efficiency of the estimator. The main purpose of the presented paper is first to construct a regression estimator that has high breakdown point combined with good efficiency. For this purpose appropriate penalties for high-leverage observations are developed so as to unmask the multiple outliers and delete bad high-leverage outliers whereas keeping all of good high-leverage points, if possible, in the data sample, otherwise most of them. Second, to improve the computation time by bringing together the PTS loss function and the idea of ϵ-insensitive loss function from support vector machines, Vapnik [23]. The support vectors have the advantage to reduce the complexity, as usually not all observations but only the support vectors contribute to the predictions, see Christmann [4]. Residuals within the interval $(-\epsilon, \epsilon)$ are ignored in the loss function, and those points outside the so-called ϵ-tube define the regression line. The mathematical programming formula gains the sparseness property and as a result the computation time is significantly reduced. Besides, the effectiveness of the robust regression method is improved, since noisy training data are ignored. For the support vector machines, Suykens et al. [22] and Christmann and Steinwart [5], have emphasized among other properties and the advantage of being robust. Both of the new estimators PTS and ϵ-insensitive PTS have shown robustness against all type of outliers reasonable high breakdown point and well efficiency. The PTS formula has the advantage to remove the outliers and it suffers little from masking effects. Generally, the proposed estimator has the ability to handle a group of outliers. This is shown by means of Examples and Monte Carlo Study. For small datasets and when the computation time is not a problem, we recommend as robust regression procedure the PTS. For moderate data sets the ϵ-insensitive PTS procedure is faster and successful.

In Section 2, we start from the LTS objective function and afterwards the PTS procedure is described. Moreover, the masking problem is described and a suitable penalty function is searched. A mathematical programming formula QMIP is developed in Section 3, for obtaining a PTS estimate. In Section 4, a support vector machines technique is developed with the new ϵ-insensitive loss function. Some benchmark examples are studied in Section 5. The performance of the new estimators PTS and IPTS are tested using Monte-Carlo simulation study in Section 6. Finally, conclusions and future research are addressed in Section 7.

2. TRIMMED SQUARES REGRESSION

We consider the linear regression model with p independent variables

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} \ ,$$

where $\boldsymbol{y}$ is the $n\times 1$ vector of the response variable $\boldsymbol{y} = (y_1, y_2, ..., y_n)^T$, $\boldsymbol{X}$ is a full rank $n\times p$ matrix of the $p\times 1$ vectors of explanatory variables, $\boldsymbol{x}_i = (\boldsymbol{x}_{i,1}, \boldsymbol{x}_{i,2}, ..., \boldsymbol{x}_{i,p})$, for $i = 1, 2, ..., n$, $\boldsymbol{\beta}$ is a $p\times 1$ vector of unknown parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)^T$, and $\boldsymbol{u}$ is a $n\times 1$ vector $\boldsymbol{u} = (u_1, u_2, ..., u_n)^T$ of iid random errors with expectation zero and variance σ^2. We observe a sample $(y_i, x_{i,1}, x_{i,2}, ..., x_{i,p})$, for $i = 1, 2, ..., n$, and construct an estimator for the unknown parameters $\boldsymbol{\beta}$. The Least Squares Estimator is defined by minimizing the squared error loss function

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \boldsymbol{u}_i^2 \ .$$

Unfortunately, points that are far from the predicted line (outliers) are overemphasized. Least Squares Estimators are very sensitive to outliers. We wish to construct a robust estimator for the parameter $\boldsymbol{\beta}$, in the sense that the influence of any observation $(\boldsymbol{x}_i, y_i)$ on the sample estimator is bounded.

Rousseeuw and Leroy [16], introduced the Least Trimmed Squares LTS estimator, which fits the best subset of k observations, removing the rest $n-k$ observations. The LTS estimator is defined by minimizing:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{k} u_i^2 \ , \qquad (2.1)$$

$$\text{s.t.} \quad u_{(1)}^2 < u_{(2)}^2 < u_{(3)}^2 < ... < u_{(k)}^2 \ ,$$

where k is the coverage, $k > n/2$ chosen a priori, to maximize the so called breakdown point, $k = (n+p-1)/2$. The estimator has high breakdown point but loses efficiency, since $n-k$ observations have to be removed from the sample even they are not outliers. In real applications the coverage k is unknown. The exact computation of LTS is difficult. Given coverage k, we have to find the best set from all combinations (n, k). The exact algorithm for LTS is a combinatory one, and is suitable for small data sets, i.e. $n < 50$. Fast probabilistic algorithms have been developed for larger samples. In the following proposed robust procedures we consider only exact solutions.

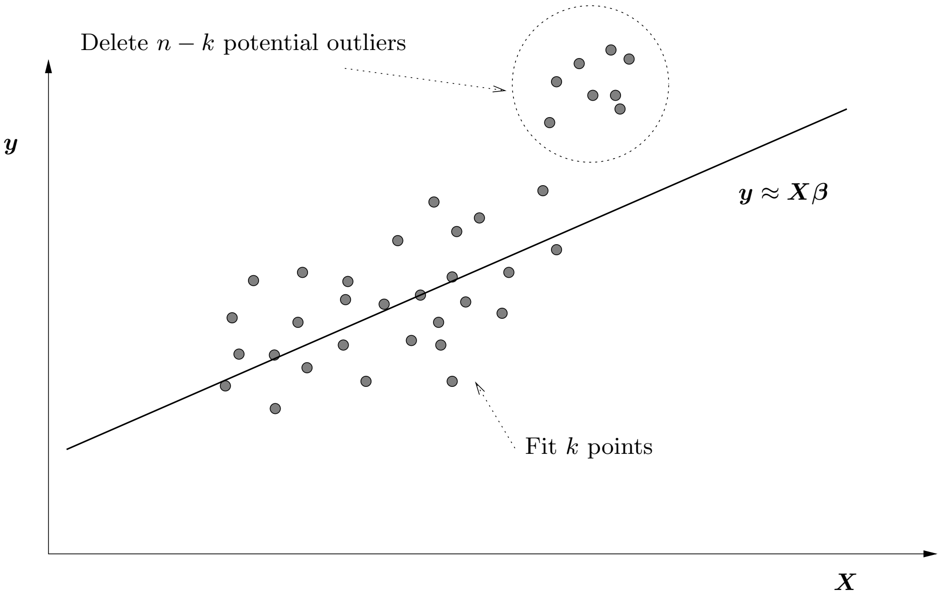
Delete $n-k$ potential outliers
$\boldsymbol{y}$
$\boldsymbol{y} \approx \boldsymbol{X\beta}$
Fit k points
$\boldsymbol{X}$

Figure 1: LTS fitting with coverage k. (In practice the coverage k is unknown).

A problem with the LTS method is that the size $n-k$ of the outlier subset is rarely known. We propose a new approach that does not require presetting the number $n-k$ of outliers to delete from the data set. The basic idea is to insert fixed penalty costs into the loss function for possible deletion. Thus, only observations that produce reduction larger than their penalty costs are deleted from the data set. The penalty costs are defined a priori, in the following section the definition of the penalized trimmed squares estimator PTS is formalized and suitable penalties for multiple high-leverage outliers are proposed. In this work, the PTS estimator is defined over those k observations out of n with the largest maximum likelihood estimation (MLE) fit. We consider as most likely outliers the subset of the observations that produces significant reduction in the residual sum of square when deleted. The proposed PTS estimator minimizes the total sum of squared residuals which is split into two parts; the sum of the k squared residuals in the clean data and the sum of the penalties for deleting the rest $n-k$

observations,

$$\min_{\boldsymbol{\beta},k}\Big(S_k(\boldsymbol{\beta})+S_{n-k}(\boldsymbol{\beta})\Big)\ ,$$

$$\text{or equivalently}\quad \min_{\boldsymbol{\beta},k}\left(\sum_{i=1}^{k}u_i^2+(n-k)\times(c\sigma)^2\right)\ , \tag{2.2}$$

where, $(c\sigma)^2$ can be interpreted as a *penalty* cost for deleting an observation, σ is a robust residual scale, taken from LTS, and c is a cut-off parameter. The estimator performance is very sensitive to the penalties defined a priori, which regulate the robustness and the efficiency of the estimator. The choice of the robust scale σ plays an important role in the coverage of the PTS estimator. If we wish to obtain an initial clean subset from the PTS estimator (coverage 51%), we choose as scale σ the square root of the minimum mean squared residuals resulted from LTS with the same coverage. Alternatively, in order to delete only the bad outliers, we could get the normalized robust scale σ from the LTS estimator. The minimization problem (2.2) is convex, as it will be proved in Section 3, therefore a global minimum exists. Given that the LTS estimate for coverage k converges to the unique optimum solution of (2.1), the following proposition is useful.

Proposition 2.1. *If the PTS estimator for given penalty $(c\sigma)^2$ converges to the solution $(\boldsymbol{\beta}_{PTS},k)$, then for the same coverage k the LTS estimator yields the equal estimate $\boldsymbol{\beta}_{LTS}=\boldsymbol{\beta}_{PTS}$.*

Proof: For given penalty $(c\sigma)^2$, the PTS is defined by solving the minimization problem (2.2), and the resulted global minimum is

$$S_{k,PTS}\ =\ S_k(\boldsymbol{\beta}_{PTS})+(n-k)\times(c\sigma)^2\ .$$

From the resulted coverage k of the PTS solution, the LTS leads to a unique minimum $S_k(\boldsymbol{\beta}_{LTS})$. Increasing this sum by a constant $(n-k)\times(c\sigma)^2$ yields the unique global minimum sum $S_k(\boldsymbol{\beta}_{LTS})+(n-k)\times(c\sigma)^2$, which is the same with $S_{k,PTS}$, since both are global minimum. Therefore, both estimates $\boldsymbol{\beta}_{PTS}$ and $\boldsymbol{\beta}_{LTS}$ coincide. □

As a consequence of Proposition 2.1, the PTS estimator can be considered as high breakdown estimator, for small penalty cost $(c\sigma)^2$. For instance, asymptotically under Gaussian conditions, minimizing (2.2) with penalty cost of $c\approx 0.7$, the solution of (2.2) converges to the LTS estimator with high breakdown point $\approx 49\%$. Increasing the parameter c, we obtain better efficiency with reasonable robustness. We have found that for $c=3$, the PTS estimator works well for the catastrophic outliers and this value has been used in the simulation and the examples. Moreover, the PTS estimate is the OLS estimate of the "clean" data subset k. PTS can be approached equivalently by solving the problem

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho_{c\sigma}(u_i) \ ,$$

$$\rho_{c\sigma}(u_i) \;=\; \begin{cases} u_i^2 & \text{for } |u_i| < c\sigma\,\sqrt{1-h_i}\,, \\ (c\sigma)^2 & \text{for } |u_i| \geq c\sigma\,\sqrt{1-h_i}\ , \end{cases} \tag{2.3}$$

where the leverages h_i are introduced in the following paragraph. The PTS loss function is simple, for large residual u_i the sum of squared residuals is less rapidly increasing. An interpretation of constant penalizing for big residuals is that the observation $(\boldsymbol{x}_i, y_i)$ does not influence further the regression fitting and can be considered as a deleted one.

As it is known from robust literature, Atkinson and Riani [2], a transformation of residuals that has been useful for outlier diagnostics, is the square of adjusted residual, $\frac{u_i^2}{1-h_i}$, where h_i $(0 < h_i < 1)$ measures the leverage of the i^{th} observation, $h_i = \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i$. The **general principle** of PTS estimator (2.3) is to delete an observation if its reduction in the sum of squared errors, $S_k(\boldsymbol{\beta})$, is larger than the penalty cost $\frac{u_i^2}{1-h_i} > (c\sigma)^2$. In the solution of the minimization problem (2.3), every residual in the clean data subset has an upper bound $|u_i| < c\sigma(\sqrt{1-h_i})$. However, as the number of the observations to be deleted increases, there is a combinatorial explosion of the number of deleted subsets to be considered, which can lead to difficulties. Besides, as it is known the leverage value h_i can be distorted by the presence of collection of points, which individually have small leverage values but collectively forms a high leverage group. Peña and Yohai [14] point out that the individual leverage h_i of each point might be small, whereas the final residual u_i may appear very close to 0, and this is a masking problem.

2.1. Masking problem and choice of penalties

For y-outliers and even for few $\boldsymbol{x}$-outliers the PTS estimator has successful performance. Unfortunately, masking problem arises when there is a group of high leverage points in the same direction. In a set of identical high leverage outliers, the leverage of each outlier is masked; the h_i might be small (Peña and Yohai [13]), $h_i \ll 1$. Deleting a masked leverage point, the reduction in the sum of squared residuals may be small $\frac{u_i^2}{1-h_i} \ll (c\sigma)^2$. In order to eliminate the distortion of the masking problem appropriate penalties for high-leverage observations are searched in this work to unmask the multiple outliers and delete bad high-leverage outliers. Most methods for multiple outlier detection as Hadi and Simonoff [9], Peña and Yohai [14], seek to divide the data into two parts, a larger "clean data" part and the outliers. The clean data are then used for the estimation of useful parameters. In the PTS procedure we follow a similar principle, we propose to down-weight the penalties using information from:

1) The initial leverage of each data point $(\boldsymbol{x}_i, y_i)$, $h_i = \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i$.

2) The leverage of each point $(\boldsymbol{x}_i, y_i)$ as it joins the clean data subset taken from MCD with coverage k (Rousseeuw and Van Driessen [17]), is $h_i^* = \boldsymbol{x}_i^T(\boldsymbol{X}_{k+1}^T\boldsymbol{X}_{k+1})^{-1}\boldsymbol{x}_i$, which can be considered as the leverage at the clean data set of coverage k. From robust literature, it is expected that $h_i^* \geq h_i$ for the potential x_i-outliers, i.e. for points not included in $\boldsymbol{X}_k$. For the remaining points, which are included in $\boldsymbol{X}_k$, we take $h_i^* = h_i$.

In a bounded influence estimate we wish for every data point $(\boldsymbol{x}_i, y_i)$, $|u_i| \leq c\sigma\sqrt{1-h_i^*}$. This can be obtained by weighting the penalty as $\frac{1-h_i^*}{1-h_i}(c\sigma)^2$. Applying the proposed robust function (2.3) to the initial data set

$$\rho_{(1-h_i^*)(1-h_i)c\sigma}(u_i) = \begin{cases} u_i^2 & \text{for } |u_i| < c\sigma\,\frac{\sqrt{1-h_i^*}}{\sqrt{1-h_i}}\,\sqrt{1-h_i} = c\sigma\sqrt{1-h_i^*}\,, \\[2ex] \dfrac{\sqrt{1-h_i^*}}{\sqrt{1-h_i}}(c\sigma)^2 & \text{for } |u_i| \geq c\sigma\frac{\sqrt{1-h_i^*}}{\sqrt{1-h_i}}\sqrt{1-h_i} = c\sigma\sqrt{1-h_i^*}\,. \end{cases}$$

The above argument leads to the choice of penalty down-weighting with

$$w_i = \min\left\{1,\ \frac{\sqrt{1-h_i^*}}{\sqrt{1-h_i}}\right\}. \tag{2.4}$$

Therefore, the deleting penalties become $(c_i\sigma)^2$, where $c_i = c\,w_i$. For minimizing the penalty loss function in (2.2), a quadratic mixed integer programming formula is used as it is developed in the next paragraph.

3. QMIP FORMULA FOR THE PTS

The new estimator PTS is defined from the solution of the problem (2.2) or (2.3). In order to minimize the penalty loss function in a robust regression, Zioutas and Avramidis [25] proposed a quadratic mixed integer programming formula, called QMIP:

$$
\begin{aligned}
(3.1) \qquad & \min_{\boldsymbol{\beta},u_i,s_i,\delta_i} \; \sum_{i=1}^{n} \left(u_i^2 + \delta_i (c\,w_i\,\sigma)^2\right) \;, \\
& \text{s.t.} \quad \boldsymbol{x}_i^T \boldsymbol{\beta} + u_i \;\geq\; y_i - s_i \\
& \qquad\;\; \boldsymbol{x}_i^T \boldsymbol{\beta} - u_i \;\leq\; y_i + s_i \\
& \qquad\;\; s_i \leq \delta_i M \\
& \qquad\;\; \delta_i \colon \;\text{zero-one variable} \\
& \qquad\;\; u_i, s_i \geq 0 \quad \text{for} \;\; i = 1,...,n \;,
\end{aligned}
$$

where, s is the pulling distance for moving an outlier towards the regression line, $\boldsymbol{\delta}$ is a zero-one decision vector, to indicate which observations must be removed and M is an upper limit of the residuals u_i, $i = 1,...,n$. Given any fixed $\boldsymbol{\delta} \in \{0,1\}^n$ from the 2^n possible ones, and using matrix notation we have the following mixed integer quadratic problem:

$$
\begin{aligned}
& \min_{\boldsymbol{\beta}} \;\; \boldsymbol{u}^T \boldsymbol{u} + \boldsymbol{\delta}^T \boldsymbol{p} \;, \\
& \text{s.t.} \quad \boldsymbol{X\beta} + \boldsymbol{u} \;\geq\; \boldsymbol{y} - \boldsymbol{s} \\
& \qquad\;\; \boldsymbol{X\beta} - \boldsymbol{u} \;\leq\; \boldsymbol{y} + \boldsymbol{s} \\
& \qquad\;\; \boldsymbol{s} \leq \boldsymbol{\delta} M \\
& \qquad\;\; \boldsymbol{u}, \boldsymbol{s} \geq \boldsymbol{0} \;,
\end{aligned}
$$

where, $\boldsymbol{p} = \left((c\,w_1\,\sigma)^2, (c\,w_2\,\sigma)^2, ..., (c\,w_n\,\sigma)^2\right)^T$, $\boldsymbol{u} = (u_1,...,u_n)^T$, $\boldsymbol{s} = (s_1,...,s_n)^T$, $\boldsymbol{y} = (y_1,...,y_n)^T$, $\boldsymbol{\beta} = (\beta_1,...,\beta_n)^T$ and the matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n]^T$. This problem has linear constraints and a convex quadratic objective function, since the Hessian of $\boldsymbol{u}^T\boldsymbol{u}$ has nonnegative eigenvalues (and it is therefore positive semi-definite). Therefore we have a convex program, which will have a unique global optimum solution according to the Karush–Kuhn–Tucker optimality conditions [3]. Considering that there is a finite number of possible $\boldsymbol{\delta}$, we can conclude that a global optimum solution to the problem exist. Hence, the quadratic mixed integer programming formula (3.1) is convex; therefore, a unique global optimum solution can be obtained for the given data, which is an estimate of the PTS.

In the present work, the solution of the QMIP formula obtained by the Fort/QMIP algorithm, Mitra et al. [11]. Computationally, the PTS estimation is suitable for small number of observations, $n < 50$, otherwise it could be extremely intensive. In the next paragraph we propose an ϵ-insensitive PTS procedure where the QMIP formula gains sparseness and it becomes computationally reasonable even for larger data sets.

4. SUPPORT VECTORS TOLERANT REGRESSION

4.1. ϵ-Insensitive loss function

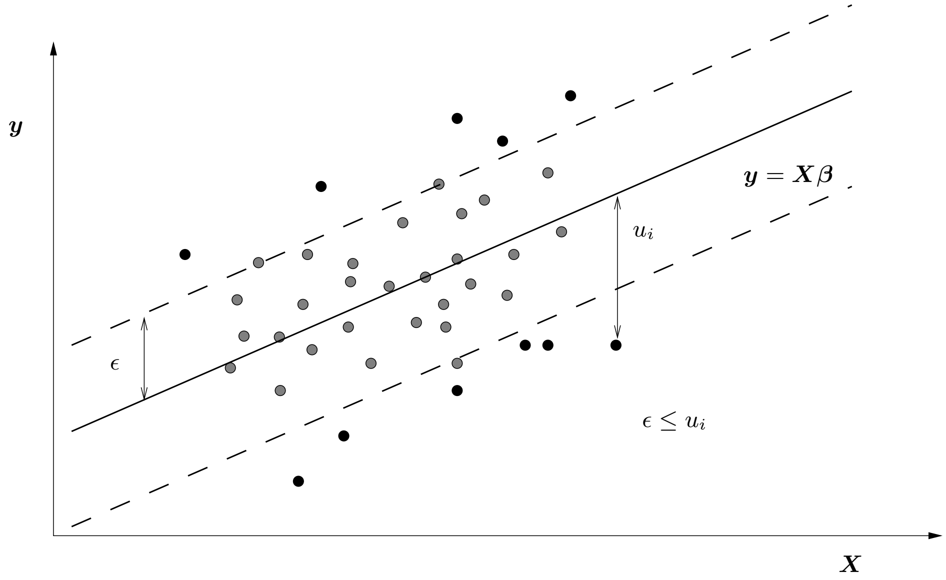

Figure 2: ϵ-insensitive tolerant regression. Only the points outside the tube enter the stochastic term. Points close to actual regression have ϵ loss.

In order, to improve the computation time we use the idea of ϵ-insensitive loss function from support vector machines, proposed by Vapnik [23]. In the ϵ-insensitive loss function small errors are not penalized and it is attempted to fit a tube with radius ϵ to the data, by ignoring (tolerating) small errors, $u < \epsilon$,

$$|y - f(x)|_\epsilon = |y - \boldsymbol{x}^T\boldsymbol{\beta}|_\epsilon = \max\bigl(0, |y - \boldsymbol{x}^T\boldsymbol{\beta}| - \epsilon\bigr) . \tag{4.1}$$

Small errors (below some $\epsilon > 0$) are not penalized in the loss function. The accuracy parameter ϵ controls the number of points outside the tube with radius ϵ. The Support Vectors Regression (SVR) based on the ϵ-insensitive loss function has the advantage to offer sparseness of the solution, Vapnik [23] and Schölkopf and Smola [19]. Christmann and Steinwart [5], [6] proved that kernel methods including SVR have good robustness properties for classification and regression problems if these kernel methods use a bounded and universal kernel and a loss function with bounded first derivative.

We adapt the support vectors technique to our approach modifying the ϵ-insensitive loss function in a squared form, and all the errors smaller than ϵ are penalized with a constant value ϵ^2. Thus, the proposed ϵ-insensitive loss function becomes

$$(y-f(x))^2_\epsilon \;=\; (y-\boldsymbol{x}^T\boldsymbol{\beta})^2_\epsilon \;=\; \max\bigl[\epsilon^2,\,(y-\boldsymbol{x}^T\boldsymbol{\beta})^2\bigr]\ , \tag{4.2}$$

where, the accuracy parameter ϵ controls the number of points outside the tube, and trades off a potential loss in prediction accuracy with gain of sparseness property and faster solutions.

We bring together, the loss functions of the new ϵ-insensitive and the Penalized Trimmed Squares. Thus, a new estimator called IPTS can yield by solving the problem

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho_{\epsilon,c_i\sigma}(u_i)\ ,$$

$$\rho_{\epsilon,c_i\sigma}(u_i) \;=\; \begin{cases} \epsilon^2 & \text{for } |u_i|\leq\epsilon\,,\\ u_i^2 & \text{for } \epsilon<|u_i|<c_i\sigma\sqrt{1-h_i}\,,\\ (c_i\sigma)^2 & \text{for } |u_i|\geq c_i\sigma\sqrt{1-h_i}\ , \end{cases} \tag{4.3}$$

where $c_i\sigma=\max\{\epsilon,\,c_i\sigma\}$. Under Gaussian conditions good efficiency could be obtained for $\epsilon=0.612\,\sigma$, Schölkopf and Smola [19]. From our empirical results $\epsilon=\sigma$ was a good choice for faster computation and efficiency. The minimization of the loss function (4.3) is equivalent to the following constraint optimization problem QMIP

$$\begin{aligned} &\min_{\boldsymbol{\beta},u_i,s_i,\delta_i} \;\sum_{i=1}^{n} \bigl(u_i^2+\delta_i(c\,w_i\sigma)^2\bigr)\ ,\\ &\quad\text{s.t.}\quad \boldsymbol{x}_i^T\boldsymbol{\beta}+u_i \;\geq\; y_i-s_i\\ &\qquad\qquad \boldsymbol{x}_i^T\boldsymbol{\beta}-u_i \;\leq\; y_i+s_i\\ &\qquad\qquad u_i\geq\epsilon\\ &\qquad\qquad s_i\leq\delta_i M\\ &\qquad\qquad \delta_i:\ \text{zero-one variable}\\ &\qquad\qquad u_i,s_i\geq 0 \quad \text{for } i=1,...,n\ , \end{aligned} \tag{4.4}$$

where $c\,w_i\sigma=\max\{\epsilon,\,c\,w_i\sigma\}$, δ_i is a zero-one decision variable, to indicate which observations must be deleted. The IPTS formula is convex, see Section 3, therefore a unique optimum solution can be found and the IPTS is estimated. The tolerance constraint of the above formula leads to sparsity. It should be noted that due to the third constraint any residual smaller than ϵ penalizes the objective function with ϵ^2. A final note must be made regarding the sparseness of the above formula (4.4). All points inside the ϵ-tube do not contribute to the solution: we could remove any one of them, and still obtain the same solution.

The new mathematical programming formula is still convex, see Section 3, and therefore the unique global optimum solution of the convex problem (4.4) yields an estimation of IPTS. In the same solution those $\delta_i = 1$ flag the deleted outliers. This way of identifying outliers with the IPTS, guarantees faster numerical solvability.

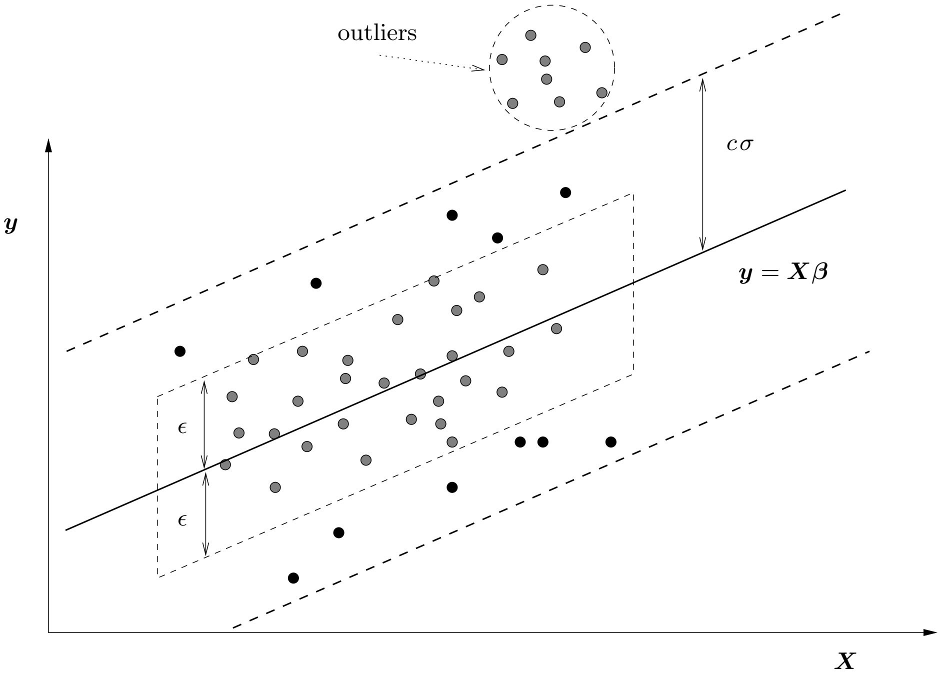

Figure 3: IPTS regression. Appropriate emphasis is given on medium residuals (risk part). De-emphasize small or big errors.

4.2. The Algorithm of IPTS Procedure for large data sets

The parameter ϵ can be useful for the desired accuracy and sparseness. In present case, however, our main goal is the identification of the outliers and faster computation, therefore larger values for the parameter ϵ could be used. Besides, as the size of the data set increases, it would be reasonable to increase the sparseness of the mathematical formula (4.4) in order to reduce the computational time. It should be noted that small changes in the parameter ϵ might increase the sparseness without affecting the correct identification of the outliers. However, as the radius ϵ increases, efficiency of the IPTS estimator may be lost. Therefore, for large data sets, we propose an algorithm of the IPTS procedure which is described briefly as follows:

- Step 1. Estimate the robust scale σ and leverage h_i^*, and determine the penalty costs $(c_i\sigma)^2$.
- Step 2. Solve the QMIP formula for the IPTS estimator.
- Step 3. Remove the detected outliers from the data i.e. points for which $\delta_i = 1$ in Step 2.
- Step 4. Estimate OLS on the clean data set. This is the final IPTS estimator.

Following these steps we obtain the IPTS estimator, which shows good performance as it is illustrated via Examples from literature and Monte Carlo Study in the next sections. More steps could improve further the IPTS estimator by reincluding deleted observations similar to Hadi and Simonoff [9]. However, this is not the goal of the present work.

5. EXAMPLES

The PTS and IPTS procedures have the advantage to remove the outliers and suffers less from masking effects. This is shown by means of real examples or artificial data sets encountered in the literature. The first four data sets, discussed by Rousseeuw and Leroy [16], have become standard "benchmark" data sets for detecting outliers in regression. The high breakdown estimators like LMS, LTS, the MM or its improved versions and the identification procedures of Hadi and Simonoff [9] correctly identify the outliers for these four data sets. Both of our proposals PTS and IPTS identify the true outliers correctly as significantly outlying. Further, the proposed procedures in this article have been tested with many other examples of Rousseeuw and Leroy [16]; in all cases we got good results.

Telephone Data. We start with the data, which relate the number of telephone calls in Belgium to the variable year, for 24 years. Cases 15−20 are unusually high; cases 14 and 21 are marginal. The outliers draw the OLS regression line upwards, masking the true outliers, while swamping in the clean cases 2−24 as too low. The MM estimator is similar to the other high breakdown estimators and correctly flags the outliers. Also, our estimators the PTS and IPTS correctly identify the true outliers.

The Stars Data. This set consists of 47 measurements of the logarithm of effective temperature at the surface of a star and the logarithm of the light intensity of the star. Although there is a direct relationship between the two variables for most of the stars, the four red giants (cases 11, 20, 30 and 34) have low temperature with high light intensity, and a scatter plot shows them as clear outliers and leverage points. The OLS- and M-estimate lines are very similar, being drawn toward the outliers are masked. The bounded influence estimator is less sensitive to the outliers than are the OLS and M estimators, having (small) positive slope, but the outliers are still masked. The high breakdown estimators LTS and MM find the true relationship if the efficiency level is set lower than the typical 95% (for efficiencies up to 80−90%). Considering stronger efficiency the MM estimator fails for this data. Application of the PTS and IPTS procedure both flags correctly the outliers.

Modified Wood Gravity Data. We next analyze the five predictors data set, based on real data but modified by Rousseeuw [15] to contain outliers at cases 4, 6, 8 and 19. All of the identification methods discussed above, as well, the OLS, M, and bounded influence estimates, fail to identify the outliers. The MM estimator is successful for this data, with the true outliers having large residuals. The proposed PTS and IPTS estimators are also successful.

Hawkins, Bradu and Kass Data. The data generated by Hawkins et al. [10] for illustrating the merits of a robust technique. This artificial data set offers the advantage that at least the position of the good or bad leverage points is known. The Hawkins, Bradu and Kass data consists of 75 observations in four dimensions. The first ten observations is a group of identical bad leverage points, the next four points are good leverage while the remaining are good data. The problem in this case is to fit a hyperplane to the observed data. Plotting the regression residuals from the model obtained from the standard OLS estimator, the bad high-leverage point data are masked and do not show up from the residual plot. Some robust methods not only fail to identify the outliers, but they also swamp in the good cases 11−14. The MM estimate is $Y = -0.9525 + 0.1492\,X_1 + 0.1968\,X_2 + 0.1793\,X_3$, which means that the true outliers are masked, whereas cases 11−14 are swamped in. Less efficient versions of the MM (up to 80%) give results similar to LTS and correctly flag the outliers. The LTS estimate is $Y = -0.524 + 0.2723\,X_1 + 0.0552\,X_2 - 0.1876\,X_3$, and correctly flags the outliers. An initial estimate of robust design weights reveals the first 14 points of this data set as high leverage points. Application of the PTS and IPTS to these data, starting with robust scale estimate about $\sigma = 0.61$ from the LTS and down-weighting the penalty cost with weights w_i from (2.4), rejects only the first 10 points as outliers, which are known as the bad leverage points. More specifically, the IPTS estimate gives $Y = -0.6599 + 0.2393X_1 + 0.0598X_2 - 0.1026X_3$, and its computation time is much faster than the PTS procedure.

New Artificial Data. These data have been created by Hadi and Simonoff [9], in order to illustrate the performance of various robust methods in outlier identification. The two predictors were originally created as uniform $(0, 15)$ and were then transformed to have a correlation of 0.5. The depended variable was then created to be consistent with the model $y = x_1 + x_2 + u$ with $u \sim N(0, 1)$. The first 3 cases (1–3) were contaminated to have predictor values around $(15, 15)$, and to satisfy $y = x_1 + x_2 + 4$. Scatterplots or diagnostics have failed to detect the outliers. Many identification methods fail to identify the three outliers. Some bounded influence estimates have largest absolute residual at the clean case 17, indicating potential swamping. The LMS regression line in cases 6, 11, 13, 17 and 24 yields larger absolute LMS residual values than the true outliers. The more efficient high breakdown methods like LTS, MM do identify the three outliers as the most outlying cases in the sample, but the residuals are to small to be considered significantly outliers. In contrast, robust methods proposed by Hadi and Simonoff [9], PTS estimator and IPTS identify correct the clean set 4−25, with each of the cases 1−3 having residuals greater than 3.78.

6. MONTE CARLO RESULTS

In this section we perform Monte Carlo experiments to evaluate the performance of our robust procedure and compare it with the well-known methods discussed in this article. To carry out one simulation run, we proceeded as follows. The distributions of independent variables and errors and the values of parameters are given. The observations y_i, were obtained following the regression model second degree $p = 2$, $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$, where the coefficient values are $\beta_1 = 1.20$, $\beta_2 = -0.80$ and a zero constant term $\beta_0 = 0.0$. We prefer the Gauss distribution for the iid error term $u \sim N(0, \sigma^2 = 16^2)$, while x_{1i} and x_{2i} are iid values drawn also from normal distributions $N(\mu = 20, \sigma^2 = 6^2)$ and $N(\mu = 30, \sigma^2 = 8^2)$ respectively. We consider that the sample may contain three types of outliers, regression outliers ("bad" high-leverage points), "good" high-leverage points, and response outliers (y-outliers). An extra value is drawn from the uniform distribution $U(a = 80, b = 220)$ and for the regression outlier is added to x_{1i} or x_{2i}, for the "good" leverage point is added to x_{1i} or x_{2i} but the value of the dependent variable y_i follows their contamination, according to the above regression model, for the response outlier is added to y_i. All simulation results are based on 100 replications enough to obtain a relative error $< 10\%$ with a reasonable confidence level of at least 90% for all the simulation estimates. The robust scale estimate σ from LTS with coverage $k = 28$ is used throughout the simulation study. We report the results only of the available well-known robust high breakdown methods. The methods examined are, therefore, five different types of robust estimators: the LTS estimator with coverage $k = [(n+p-1)/2]$, the MM and S1S estimators using in both initially the LTS regression estimate, the proposed PTS estimator solving the QMIP in (3.1), the proposed IPTS estimator solving the QMIP in (4.4). We run all of the computer programs on a 1200 Mhz Athlon AMD Processor. The computations for the robust estimators LTS and MM were carried out using the S-Plus package, while S1S estimator has been computed by the S1S algorithm given in Coakley and Hettmansperger [7]. The simplex iterations for the QMIP solution were carried out on the same machine using the solver FortMP/QMIP-Fortran Code provided by CARISMA, Brunel University, U.K., 2003.

All of the following conclusions were supported by careful examination of the individual estimates. Tables 1, 2, 3 and 4 display results concerning the performance of the four robust estimators corresponding to the following cases: Table 1, based on data contaminated by "bad" and "good" high leverage points. Table 2, based on data contaminated only by "good" leverage points. Table 3, based on data contaminated by "bad" high leverage outliers. Table 4, based on data contaminated by "bad" high leverage outliers (heavier contamination).

Table 1: x-outliers 6, "good" leverage points 4, y-outliers 6, $n = 50$. True: $\beta_0 = 0.0$, $\beta_1 = 1.20$, $\beta_2 = -0.80$.

Estimator	LTS	MM	S1S	PTS	IPTS$_{\epsilon=0.8\sigma}$
Mean estimate of β_0	-0.67	1.82	8.54	0.03	$-$**1.12**
Mean estimate of β_1	1.01	0.98	0.96	1.13	**1.21**
Mean estimate of β_2	-0.68	-0.75	-0.75	-0.81	$-$**0.80**
Mean absolute error of $\hat{\beta}_0$	7.76	5.96	9.53	3.89	**2.82**
Mean absolute error of $\hat{\beta}_1$	0.34	0.27	0.34	0.14	**0.05**
Mean absolute error of $\hat{\beta}_2$	0.15	0.09	0.08	0.07	**0.06**
Mean square error of $\hat{\boldsymbol{\beta}}$	98.91	71.53	146.05	25.43	**14.78**
Norm of bias of $\hat{\boldsymbol{\beta}}$	7.78	5.97	9.54	3.90	**2.82**
Trace of covariance	98.41	68.18	73.05	25.42	**13.54**
Mean square fitting error (true value σ^2=256)	353	314	344	275	**263**
Computation Time (secs)				11	**3**

Table 1 presents the measures of the performance criteria for the four estimators in the presence of bad and good high leverage outliers. Taking account all the performance criteria, the PTS and IPTS outperform the other estimators. In this Table, we see that IPTS outperform the PTS estimator and the IPTS procedure is faster, as it was expected. As far as the computation time of MM, LTS and S1S concern, these are not shown in Tables 1, 2, 3 and 4. This is these estimates results from probabilistic solutions. As it has been mentioned in the previous sections, the PTS and IPTS estimates are the exact solution of QMIP formulas. Therefore, the computation time between probabilistic and exact solutions is not comparable. Not surprisingly, most of the methods are more effective in the case of clean data. For the simulation conducted over clean data contaminated only by "good" high leverage points, Table 2, the IPTS estimator outperforms the other estimators. The performance of PTS, MM, S1S and LTS was reasonable well with PTS much better. Of course, one can improve the efficiency of the robust estimates, but at the cost of losing robustness and outlier detection.

Table 2: "good" leverage points 6, $n = 50$.
True: $\beta_0 = 0.0$, $\beta_1 = 1.20$, $\beta_2 = -0.80$.

Estimator	LTS	MM	S1S	PTS	IPTS$_{\epsilon=0.8\sigma}$
Mean estimate of β_0	0.53	-1.16	-2.26	-1.67	$-$**1.26**
Mean estimate of β_1	1.17	1.20	1.20	1.21	**1.21**
Mean estimate of β_2	-0.77	-0.75	-0.91	-0.75	$-$**0.77**
Mean absolute error of $\hat{\beta}_0$	7.55	3.02	3.66	2.88	**2.79**
Mean absolute error of $\hat{\beta}_1$	0.08	0.04	0.09	0.04	**0.03**
Mean absolute error of $\hat{\beta}_2$	0.10	0.07	0.10	0.07	**0.06**
Mean square error of $\hat{\boldsymbol{\beta}}$	76.93	18.02	22.88	15.80	**14.91**
Norm of bias of $\hat{\boldsymbol{\beta}}$	7.55	3.03	3.66	2.88	**2.79**
Trace of covariance	76.65	16.67	17.81	13.02	**13.33**
Mean square fitting error (true value σ^2=256)	308	266	268	263	**262**
Computation Time (secs)				9	**2**

In case of only bad high leverage contamination, shown in Table 3, the penalized trimmed squares approach has shown remarkable improvement in both robustness and efficiency, with IPTS the best. As a final conclusion of Tables $1, 2, 3$ and taking account all the performance criteria, the IPTS procedure improves reasonable the performance of the PTS. Also, the IPTS procedure is faster.

Table 3: "bad" leverage points 6, $n = 50$.
True: $\beta_0 = 0.0$, $\beta_1 = 1.20$, $\beta_2 = -0.80$.

Estimator	LTS	MM	S1S	PTS	IPTS$_{\epsilon=0.8\sigma}$
Mean estimate of β_0	4.95	1.04	3.06	0.91	**0.02**
Mean estimate of β_1	0.87	1.04	0.81	1.10	**1.15**
Mean estimate of β_2	-0.77	-0.74	-0.82	-0.76	$-$**0.76**
Mean absolute error of $\hat{\beta}_0$	11.42	5.46	6.94	4.11	**3.92**
Mean absolute error of $\hat{\beta}_1$	0.44	0.22	0.42	0.17	**0.13**
Mean absolute error of $\hat{\beta}_2$	0.22	0.12	0.17	0.10	**0.10**
Mean square error of $\hat{\boldsymbol{\beta}}$	229.69	48.59	103.16	27.24	**22.61**
Norm of bias of $\hat{\boldsymbol{\beta}}$	11.45	5.47	6.99	4.13	**3.93**
Trace of covariance	205.12	47.48	93.67	26.41	**22.60**
Mean square fitting error (true value σ^2=256)	378	298	327	282	**274**
Computation Time (secs)				9	**2.9**

The most fruitful result concerning the IPTS procedure is presented in Table 4. Data are heavy contaminated by bad high leverage outliers. A masking problem arises affecting the performance of the other robust estimators. The IPTS procedure with $\epsilon = 1.5\,\sigma$ has improved significantly the performance criteria and the computation load as well.

Table 4: "bad" leverage points 10, y-outliers 6, $n = 50$.
True: $\beta_0 = 0.0$, $\beta_1 = 1.20$, $\beta_2 = -0.80$.

Estimator	LTS	MM	S1S	PTS	IPTS$_{\epsilon=1.5\sigma}$
Mean estimate of β_0	0.03	-0.97	6.39	-1.14	$-$**1.69**
Mean estimate of β_1	0.77	0.76	0.79	1.15	**1.16**
Mean estimate of β_2	-0.57	-0.51	-0.52	-0.74	$-$**0.74**
Mean absolute error of $\hat{\beta}_0$	9.46	7.42	12.48	5.12	**4.37**
Mean absolute error of $\hat{\beta}_1$	0.56	0.54	0.65	0.21	**0.18**
Mean absolute error of $\hat{\beta}_2$	0.28	0.31	0.30	0.10	**0.09**
Mean square error of $\hat{\boldsymbol{\beta}}$	128.07	87.84	202.22	57.56	**30.50**
Norm of bias of $\hat{\boldsymbol{\beta}}$	9.50	7.51	12.51	5.15	**4.25**
Trace of covariance	127.84	86.63	161.21	56.25	**27.63**
Mean square fitting error (true value σ^2=256)	456	432	490	293	**277**
Computation Time (secs)				13	**0.5**

For large data sets, we could increase the radius ϵ in order to earn computation time, and following the algorithm of subsection 4.2, we obtain reasonable efficiency. In Tables 5 and 6, the success in outlier detection is obvious in large data sets as also the reduction of the computation time of the IPTS estimator as we increase the tube radius.

Table 5: Large artificial data set, 500 points in $\mathbb{R}^2$ including 120 outliers.

Estimator	LTS	PTS	IPTS$_{\epsilon=1.5\sigma}$	IPTS$_{\epsilon=2.0\sigma}$	IPTS$_{\epsilon=2.5\sigma}$
Deleting outlier success	95%	95%	95%	95%	95%
Computation time (sec.)	3800	3800	2500	681	**21**

Table 6: Hawkins et al. [10] artificial data, 75 points in $\mathbb{R}^3$ including 10 outliers.

Estimator	LTS	PTS	IPTS$_{\epsilon=1.5\sigma}$
Deleting outlier success	100%	100%	100%
Computation time (sec.)	255	255	**1.4**

7. CONCLUSIONS AND FUTURE WORK

The PTS estimate procedure based on robust residual scale and leverage from the LTS and MCD respectively, can be used successfully in regression problems. Through benchmark Examples and Monte Carlo simulation the proposed estimators have shown robustness against all type of outliers. The robust estimates presented in this article give directly a useful diagnostic tool to identify multiple outliers. The penalized procedure has the advantage to remove the catastrophic outliers and it does not suffer from masking problems. Generally, the proposed estimator PTS has the ability to handle effectively a group of outliers. The new estimator PTS is obtained through a convex quadratic mixed integer programming formula (QMIP). The computational effort to solve this formula is heavy. Following a modification of ϵ-insensitive technique from Support Vector Machines we have improved significantly the computational time and the effectiveness of the proposed estimator. However, the computational load of the IPTS estimator is still heavy for large data sets ($n > 100$), since the IPTS procedure is based on Quadratic Mixed Integer Programming which is partly a combinatorial problem. Based on the above optimum criteria and results, we conclude that the PTS estimator outperforms in many circumstances and is reasonable for both regression and response outliers. Therefore, it is accessed that for small sample data the added computational complexity is worth the potential benefits. Further improvements in the penalized procedure are a subject of ongoing research; for example, determine possible better choices of the penalties and continue the method in a second stage to reconsider the outliers, following one step MM-type procedure. Concerning the computation effort, further research is needed to improve the computational time for large size sample data by determining possible better choice of the ϵ-insensitive size for the IPTS procedure or implementing probabilistic techniques, similar to LTS or others known from robust literature. As a final remark, since the number of outliers in a medium sample data is not known, we recommend the use of the PTS or IPTS procedure.

REFERENCES

[1] Atkinson, A. (1994). Fast very robust methods for the detection of multiple outliers, *Journal of the American Statistical Association*, **89**, 1329–1339.

[2] Atkinson, A. and Riani, M. (2000). *Robust Diagnostic Regression Analysis*, John Wiley, Berlin.

[3] Bazaraa, M.; Shevali, H. and Shelty, C. (1993). *Nonlinear Programming: Theory and Algorithms*, John Wiley, New York.

[4] Christmann, A. (2004). *On properties of support vector machines for pattern recognition in finite samples.* In "Statistics for Industry and Technology, Theory and Applications of Recent Robust Methods", Birkhäuser Verlag, Basel, 49–58.

[5] Christmann, A. and Steinwart, I. (2004). On robustness properties of convex risk minimization methods for pattern recognition, *Journal of Machine Learning Research*, **5**, 1007–1034.

[6] Christmann, A. and Steinwart, I. (2005). *Consistency and robustness of kernel based regression*, technical report, University of Dortmund, SFB-475, TR-01/05, submitted.

[7] Coakley, C.W. and Hettmansperger, T.P. (1993). A bounded influence, high breakdown, efficient regression estimator, *Journal of the American Statistical Association*, **88**, 872–880.

[8] Gentleman, J.F. and Wilk, M.B. (1975). Detecting outliers, II, Supplementing the direct analysis of residuals, *Biometrics*, **31**, 387–410.

[9] Hadi, A.S. and Simonoff, J.S. (1993). Procedures for the identification of multiple outliers in linear models, *Journal of the American Statistical Association*, **88**, 1264–1272.

[10] Hawkins, D.M.; Bradu, D. and Kass, G.V. (1984). Location of several outliers in multiple regression data using elemental sets, *Technometrics*, **26**, 197–208.

[11] Mitra, G.; Guertler, M. and Ellison, F. (2003). *Algorithms for the solution of large-scale quadratic mixed integer programming (QMIP) models.* In "International Symposium in Mathematical Programming".

[12] Morgenthaler, S. (1989). Comment on Yohai and Zamar, *Journal of the American Statistical Association*, **84**, 636.

[13] Peña, D. and Yohai, V.J. (1995). The detection of influential subsets in linear regression using an influence matrix, *Journal of the Royal Statistical Society, Series B (Methodological)*, **57**(1), 145–156.

[14] Peña, D. and Yohai, V.J. (1999). A procedure for robust estimation and diagnostic in regression, *Journal of the American Statistical Association*, **94**, 174–188.

[15] Rousseeuw, P.J. (1984). Least median of squares regression, *Journal of the American Statistical Association*, **79**, 871–880.

[16] Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection*, John Wiley, New York.

[17] Rousseeuw, P.J. and Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator, *Technometrics*, **41**, 212–223.

[18] Rousseeuw, P.J. and Yohai, V.J. (1984). *Robust regression by means of S-estimators.* In "Robust and Nonlinear Time Series Analyses" (J. Franke, W. Hardle and R.D. Martin, Eds.), Springer Verlag, 256–272.

[19] Schölkopf, B. and Smola, A. (2000). *Learning with Kernels*, MIT Press.

[20] Simpson, D.J.; Ruppert, D. and Carroll, R.J. (1992). On one step GM estimates and stability of inferences in linear regression, *Journal of the American Statistical Association*, **87**, 439–450.

[21] Stefanski, L.A. (1991). A note on high-breakdown estimators, *Statistics and Probability Letters*, **11**, 353–358.

[22] Suykens, J.A.K.; Brabanter, J.; Lukas, L. and Vandewalle, J. (2002). Weighted least squares support vector machines: Robustness and sparse approximation, *Neurocomputing*, **48**, 85–105.

[23] Vapnik, V.N. (1998). *Statistical Learning Theory*, John Wiley, New York.

[24] Yohai, V.J. (1987). High breakdown-point and high efficiency robust estimates for regression, *Annals of Statistics*, **15**, 642–656.

[25] Zioutas, G. and Avramidis, A. (2005). Deleting outliers in robust regression with mixed integer programming, *Acta Mathematicae Applicatae Sinica*, **21**, 323–334.

REVSTAT – STATISTICAL JOURNAL

Background

Statistical Institute of Portugal (INE), well aware of how vital a statistical culture is in understanding most phenomena in the present-day world, and of its responsibility in disseminating statistical knowledge, started the publication of the scientific statistical journal *Revista de Estatística*, in Portuguese, publishing three times a year papers containing original research results, and application studies, namely in the economic, social and demographic fields.

In 1998 it was decided to publish papers also in English. This step has been taken to achieve a larger diffusion, and to encourage foreign contributors to submit their work.

At the time, the Editorial Board was mainly composed by Portuguese university professors, being now composed by national and international university professors, and this has been the first step aimed at changing the character of *Revista de Estatística* from a national to an international scientific journal.

In 2001, the *Revista de Estatística* published three volumes special issue containing extended abstracts of the invited contributed papers presented at the 23rd European Meeting of Statisticians.

The name of the Journal has been changed to REVSTAT – STATISTICAL JOURNAL, published in English, with a prestigious international editorial board, hoping to become one more place where scientists may feel proud of publishing their research results.

The editorial policy will focus on publishing research articles at the highest level in the domains of Probability and Statistics with emphasis on the originality and importance of the research.

All research articles will be refereed by at least two persons, one from the Editorial Board and another, external.

The only working language allowed will be English.

Three volumes are scheduled for publication, one in March, one in June and the other in November.

On average, four articles will be published per issue.

Aims and Scope

The aim of REVSTAT is to publish articles of high scientific content, in English, developing innovative statistical scientific methods and introducing original research, grounded in substantive problems.

REVSTAT covers all branches of Probability and Statistics. Surveys of important areas of research in the field are also welcome.

Abstract/indexed in

REVSTAT is expected to be abstracted/indexed at least in *Current Index to Statistics*, *Mathematical Reviews*, *Statistical Theory and Method Abstracts*, and *Zentralblatt für Mathematic*.

Instructions to Authors, special-issue editors and publishers

Papers may be submitted in two different ways:

— By sending a paper copy to the Executive Editor and one copy to one of the two Editors or Associate Editors whose opinion the author(s) would like to be taken into account, together with a postscript or a PDF file of the paper to the e-mail: revstat@fc.ul.pt.

— By sending a paper copy to the Executive Editor, together with a postscript or a PDF file of the paper to the e-mail: revstat@fc.ul.pt.

Submission of a paper means that it contains original work that has not been nor is about to be published elsewhere in any form.

Submitted manuscripts (text, tables and figures) should be typed only in black, on one side, in double spacing, with a left margin of at least 3 cm and not have more than 30 pages.

The first page should include the name, affiliation and address of the author(s) and a short abstract with the maximum of 100 words, followed by the key words up to the limit of 6, and the AMS 2000 subject classification.

Authors are obliged to write the final version of accepted papers using LaTeX, in the REVSTAT style.

This style (REVSTAT.sty), and examples file (REVSTAT.tex), which may be download to *PC Windows System* (Zip format), *Mackintosh, Linux* and *Solaris Systems* (StuffIt format), and *Mackintosh System* (BinHex Format), are available in the REVSTAT link of the National Statistical Institute's Website: http://www.ine.pt/revstat.html

Additional information for the authors may be obtained in the above link.

Accepted papers

Authors of accepted papers are requested to provide the LaTeX files and also a postscript (PS) or an acrobat (PDF) file of the paper to the Secretary of REVSTAT: liliana.martins@ine.pt.

Such e-mail message should include the author(s)'s name, mentioning that it has been accepted by REVSTAT.

The authors should also mention if encapsulated postscript figure files were included, and submit electronics figures separately in .tiff, .gif, .eps or .ps format. Figures must be a minimum of 300 dpi.

Also send always the final paper version to:

Maria José Carrillho
Executive Editor, REVSTAT – STATISTICAL JOURNAL
Instituto Nacional de Estatística
Av. António José de Almeida
1000-043 LISBOA
PORTUGAL

Copyright and Reprints

Upon acceptance of an article, the author(s) will be asked to transfer copyright of the article to the publisher, the INE, in order to ensure the widest possible dissemination of information, namely through the National Statistical Institute's Website (http://www.ine.pt).

After assigning the transfer copyright form, authors may use their own material in other publications provided that the REVSTAT is acknowledged as the original place of publication. The Executive Editor of the Journal must be notified in writing in advance.

Authors of articles published in the REVSTAT will be entitled to one free copy of the respective issue of the Journal and twenty-five reprints of the paper are provided free. Additional reprints may be ordered at expenses of the author(s), and prior to publication.