# Robust Estimation and Inference under Cellwise and Casewise Contamination

by

Andy Chun Yin Leung

B.Sc., The University of British Columbia, 2011

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

The Faculty of Graduate and Postdoctoral Studies

(Statistics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

December 2016

# Abstract

Cellwise outliers are likely to occur together with casewise outliers in datasets of relatively large dimension. Recent work has shown that traditional high breakdown point procedures may fail when applied to such datasets. In this thesis, we consider this problem when the goal is to (1) estimate multivariate location and scatter matrix and (2) estimate regression coefficients and confidence intervals for inference, which both are cornerstones in multivariate data analysis.

To address the first problem, we propose a two-step procedure to deal with casewise and cellwise outliers, which generally proceeds as follows: first, it uses a filter to identify cellwise outliers and replace them by missing values; then, it applies a robust estimator to the incomplete data to down-weight casewise outliers. We show that the two-step procedure is consistent under the central model provided the filter is appropriately chosen.

The proposed two-step procedure for estimating location and scatter matrix is then applied in regression for the case of continuous covariates by simply adding a third step, which computes robust regression coefficients from the estimated robust multivariate location and scatter matrix obtained in the second step. We show that the three-step estimator is consistent and asymptotically normal at the central model, for the case of continuous covariates. Finally, the estimator is extended to handle both continuous and dummy covariates.

Extensive simulation results and real data examples show that the proposed methods can handle both cellwise and casewise outliers similarly well.

# Preface

This dissertation was prepared under the supervision of Professor Ruben Zamar and it is mainly based on the three papers coauthored with the supervisor and other collaborators. Two of the papers have been published and a third one is submitted for publication.

Chapter 2 is mostly based on the published discussion paper in TEST, "Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination" (Agostinelli et al., 2015). The problem addressed in this paper, the proposed two-step procedures and the theoretical developments resulted from a broad discussion among the authors. Moreover, the author of this dissertation developed a bivariate filter version of the procedure which is included in the third paper. The author also designed and conducted the empirical study and prepared the first version of the manuscript, as well as the rejoinder, followed by the revisions proposed by the coauthors. Finally, the author implemented the proposed procedure, as well as its workhorse, the generalized S-estimator (GSE) (Danilov et al., 2012), that is used in the second step. This resulted in an `R` package, `GSE` (Leung et al., 2015), available on CRAN (R Core Team, 2015).

Chapter 3 follows up on the discussion and rejoinder of the discussion paper in TEST (Agostinelli et al., 2015). Professor Ricardo Maronna in his discussion suggested replacing the second step of the two-step procedure with a GSE of Danilov et al. (2012) with a Rocke-type $\rho$ function. This way the procedure could achieve robustness in higher dimensions. The author of the dissertation investigated the suggestion and extended the GSE to its Rocke-type counterpart. The author invented a new fast subsampling procedure for computing a needed initial estimator. The author used the same empirical study design as presented in Chapter 2 and prepared all the empirical results. The author also wrote the first version of this manuscript followed by the revisions proposed by the supervisor. The work has been submitted for publication. Finally, all the relevant procedures proposed in this chapter are made available in the above mentioned `R` package, `GSE`.

Chapter 4 is based on the published paper in Computational Statistics & Data Analysis, "Robust regression estimation and inference in the presence of cellwise and casewise contamination" (Leung et al., 2016). The problem addressed in this paper, the proposed procedures and the theoretical developments resulted from a broad discussion among the authors. In particular, the author came up with the key idea to skip the filtering step if cellwise contamination is not suspected, which considerably helps the theoretical analysis of the approach. The first version of the manuscript, including the asymptotic results and the empirical study, were all prepared by the author, followed by the revisions proposed by the supervisor. Finally, the author implemented the three-step approach in the `R` package `robreg3S` (Leung et al., 2015), also available on CRAN.

In addition to the aforementioned contributions, the supervisor made several suggestions regarding the presentation of material in this dissertation, relevant literature, and motivation of the research. He also checked all the proofs and made some changes in the writing to improve the flow of ideas in the dissertation and papers.

# Table of Contents

**Abstract** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . ii

**Preface** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . iii

**Table of Contents** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . v

**List of Tables** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . viii

**List of Figures** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . x

**Dedication** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . xiii

**1 Introduction** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1

1.1 Background . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1

1.2 Real data examples . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

1.2.1 Geochemical data . . . . . . . . . . . . . . . . . . . . . . . . . 3

1.2.2 Micro-cap stock returns data . . . . . . . . . . . . . . . . . . 6

1.3 Contributions and outline of the thesis . . . . . . . . . . . . . . . . . 9

**2 Robust Estimation of Multivariate Location and Scatter in the Presence of Cellwise and Casewise Contamination in Moderate Dimension** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12

2.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12

2.2 Global-robust estimation under THCM and ICM . . . . . . . . . . . . 16

2.2.1 Step I: Filtering cellwise outliers . . . . . . . . . . . . . . . . . 16

2.2.2 Step II: Dealing with casewise outliers . . . . . . . . . . . . . 21

2.3 Consistency of GSE on filtered data . . . 24

2.4 Simulations . . . 25

2.5 Real data example: geochemical data revisit . . . 33

2.6 Discussion . . . 36

2.6.1 Controlling the robustness and efficiency for large $p$ . . . 37

2.6.2 Computational issue . . . 37

2.6.3 Remarks on handling large and flat data sets . . . 38

2.7 Conclusions . . . 39

**3 Robust Estimation of Multivariate Location and Scatter in the Presence of Cellwise and Casewise Contamination in High Dimension** 41

3.1 Introduction . . . 41

3.2 Generalized Rocke S-estimators . . . 42

3.3 Computing issues . . . 45

3.3.1 Initial estimator . . . 45

3.3.2 Another computing issue . . . 48

3.4 Two-step estimation and extended simulations . . . 48

3.5 Real data example: small-cap stock returns data . . . 52

3.6 Conclusions . . . 56

**4 Robust Regression Estimation and Inference in the Presence of Cellwise and Casewise Contamination** . . . 58

4.1 Introduction . . . 58

4.2 Consistent filter for general continuous data . . . 60

4.3 Three-step regression . . . 62

4.3.1 The estimator . . . 62

4.3.2 Models with continuous and dummy covariates . . . 65

4.4 Asymptotic properties of three-step regression . . . 67

4.5 Simulations . . . 69

4.5.1 Models with continuous covariates . . . 70

4.5.2 Models with continuous and dummy covariates . . . 75

4.6 Real data example: Boston Housing data . . . . . . . . . . . . . . . . 77
4.7 Conclusions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 80

**5 Conclusions** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 81

**Bibliography** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 83

## Appendices

**A Supplementary material for Chapter 2** . . . . . . . . . . . . . . . . . . 89
A.1 Additional tables from the simulation study in Section 2.4 . . . . . . 89
A.2 Proofs of propositions and theorem . . . . . . . . . . . . . . . . . . . 91

**B Supplementary material for Chapter 3** . . . . . . . . . . . . . . . . . . 101
B.1 Efficiency of GRE and tuning parameter $\alpha$ in Rocke-$\rho$ function . . . 101
B.2 Performance comparison between GSE and GRE . . . . . . . . . . . 102

**C Supplementary material for Chapter 4** . . . . . . . . . . . . . . . . . . 104
C.1 Additional figures from the simulation study in Section 4.5 . . . . . . 104
C.2 Investigation on the performance on non-normal covariates . . . . . . 105
C.3 Supplementary material for the Boston housing data analysis . . . . 107
C.4 Proofs of lemmas and theorems . . . . . . . . . . . . . . . . . . . . . 107

# List of Tables

2.1 Maximum average LRT distances under cellwise contamination. The sample size is $n = 10p$. . . . . . . . . . . . . . . . . . . . . . . . . . . 31

2.2 Maximum average LRT distances under casewise contamination. The sample size is $n = 10p$. . . . . . . . . . . . . . . . . . . . . . . . . . . . 32

2.3 Finite sample efficiency for first order autoregressive correlations, AR1($\rho$), with $\rho = 0.9$. The sample size is $n = 10p$. . . . . . . . . . . . . . . . . . 33

2.4 Average “CPU time” – in seconds of a 2.8 GHz Intel Xeon – evaluated using the `R` command, `system.time`. The sample size is $n = 10p$. . . 33

3.1 Maximum average LRT distances under cellwise contamination. The sample size is $n = 10p$. . . . . . . . . . . . . . . . . . . . . . . . . . . 49

3.2 Maximum average LRT distances under casewise contamination. The sample size is $n = 10p$. . . . . . . . . . . . . . . . . . . . . . . . . . . 50

3.3 Finite sample efficiency for random correlations. The sample size is $n = 10p$. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 51

3.4 Average “CPU time” – in seconds of a 2.8 GHz Intel Xeon – evaluated using the `R` command, `system.time`. The sample size is $n = 10p$. . . 52

4.1 Maximum $\overline{MSE}$ in all the considered scenarios for models with continuous covariates. . . . . . . . . . . . . . . . . . . . . . . . . . . . 71

4.2 Average lengths of confidence intervals for clean data and for cellwise and casewise contamination. . . . . . . . . . . . . . . . . . . . . . . . . 74

4.3 Maximum $\overline{MSE}$ in all the considered scenarios for models with continuous and dummy covariates. . . . . . . . . . . . . . . . . . . . . . 76

4.4 Estimates and p-values of the regression coefficients for the original Boston Housing data. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 78
4.5 Pairwise squared norm distances between the estimates for the original Boston housing data. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 78
4.6 Estimates and p-values of the regression coefficients for the imputed Boston Housing data. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 79
4.7 Pairwise squared norm distances between the estimates for the imputed Boston housing data. . . . . . . . . . . . . . . . . . . . . . . . . 79

A.1 Maximum average LRT distances under cellwise contamination. The sample size is $n = 5p$. . . . . . . . . . . . . . . . . . . . . . . . . . . 89
A.2 Maximum average LRT distances under casewise contamination. The sample size is $n = 5p$. . . . . . . . . . . . . . . . . . . . . . . . . . . . 90
A.3 Finite sample efficiency for first order autoregressive correlations, AR1($\rho$), with $\rho = 0.9$. The sample size is $n = 5p$. . . . . . . . . . . . . . . . . 90

B.1 Finite sample efficiency and maximum average LRT distances for GRE-C with various values of $\alpha$. The sample size is $n = 10p$. . . . . . 101
B.2 Maximum average LRT distances. The sample size is $n = 10p$. . . . . 102
B.3 Finite sample efficiency. The sample size is $n = 10p$. . . . . . . . . . . 103

C.1 $\overline{MSE}$ for clean data and cellwise contaminated data. . . . . . . . . . 106
C.2 Description of the variables in the Boston Housing data . . . . . . . . 107

# List of Figures

1.1 Illustration of (a) casewise contamination and (b) cellwise contamination in a data matrix. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2

1.2 (a) Histograms and (b) quantile–quantile plots of the compound content in the geochmical data. . . . . . . . . . . . . . . . . . . . . . . . . . 4

1.3 Squared Mahalanobis distances of the samples in the geochemical data based on estimates calculated on (a) the original data and (b) the cleaned data. Samples that contain one or more flagged components (large cellwise outliers) are in green. . . . . . . . . . . . . . . . . . . . . 6

1.4 (a) Histograms and (b) quantile–quantile plots of the weekly returns in the micro-cap asset returns data. . . . . . . . . . . . . . . . . . . . . 7

1.5 Squared Mahalanobis distances of the weekly observations in the micro-cap asset returns data based on estimates calculated on (a) the original data and (b) the cleaned data. Large distances are truncated for better visualization. Observations that contain one flagged components (large cellwise outliers) are in blue and those contain at least two flagged components are in green. The weeks corresponding to the 2008 financial crisis are enclosed by the vertical dashed lines. . . . . . 8

1.6 Results of UBF-GRE-C applied to the original data. . . . . . . . . . . 9

2.1 Average LRT distances for various contamination values, $k$, under 10% cellwise contamination. The dimension is $p = 20$ and sample size is $n = 10p$. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 31

2.2 Squared Mahalanobis distances of the samples in the geochemical data based on different estimates. Large distances are truncated for better visualization. Samples that contain one or more flagged components (large cellwise outliers) are in green. . . . . . . . . . . . . . . . . . . 34

2.3 Univariate scatterplots of the 10 components in the geochemical data. Each component is standardized by its median and MAD. Points are randomly jittered to avoid much overlapping. The points flagged by the univariate filter are in blue, and those additionally flagged by the bivariate filter are in orange. . . . . . . . . . . . . . . . . . . . . . . 35

2.4 Pairwise scatterplots of the geochemical data for $V2$ versus $V8$, $V2$ versus $V3$, and $V3$ versus $V9$. Points with components flagged by the univariate filter are in blue. Points with components additionally flagged by the bivariate filter are in orange. . . . . . . . . . . . . . . . 36

3.1 Weight functions of the Tukey bisquare and the Rocke for $p = 40$. Chi-square density functions are also plotted in blue for comparison. All the functions are scaled so that their maximum is 1 to facilitate comparison. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 44

3.2 Average LRT distance behaviors of UBF-GSE and UBF-GRE-C for random correlations under 10% casewise contamination. The dimension is $p = 30$ and the sample size is $n = 10p$. . . . . . . . . . . . . . 51

3.3 Normal quantile–quantile plots of weekly returns. Weekly returns that are three MAD away from the coordinatewise-median are shown in green. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 53

3.4 Squared Mahalanobis distances of the weekly observations in the small-cap asset returns data based on the MLE, the Rocke, the UF-GSE, and the UBF-GSE estimates. Weeks that contain one or more asset returns with values three MAD away from the coordinatewise-median are in green. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 54

3.5 Pairwise scatterplots of the asset returns data for WTS versus HTLD, HTLD versus WSBC, and WSBC versus SUR. Points with components flagged by the univariate filter are in blue. Points with components additionally flagged by the bivariate filter are in orange. . . . . 55

3.6 Squared Mahalanobis distances of the weekly observations in the small-cap asset returns data based on the UBF-GSE and the UBF-GRE-C estimates. Weeks that contain one or more asset returns with values three MAD away from the coordinatewise-median are in green. . . . 56

4.1 $\overline{MSE}$ for various cellwise and casewise contamination values, $k$, for models with continuous covariates. The sample size is $n = 300$. . . . . 72

4.2 $\overline{CR}$ for clean data and for cellwise and casewise contaminated data of various sample size, $n$. . . . . . . . . . . . . . . . . . . . . . . . . . . 74

4.3 $\overline{MSE}$ for various cellwise and casewise contamination values, $k$, for models with continuous and dummy covariates. The sample size is $n = 300$. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 76

B.1 Average LRT distances for various contamination sizes, $k$, for random correlations under 10% casewise contamination. The dimension is $p = 30$ and the sample size is $n = 10p$. . . . . . . . . . . . . . . . . . 103

C.1 $\overline{MSE}$ for various cellwise and casewise contamination values, $k$, for models with $p = 15$ continuous covariates. The sample size is $n = 150$. For details see Section 4.5.1 in the chapter. . . . . . . . . . . . . . . 104

C.2 $\overline{MSE}$ for various cellwise and casewise contamination values, $k$, for models with $p_x = 12$ continuous and $p_d = 3$ dummy covariates. The sample size is $n = 150$. For details see Section 4.5.2 in the chapter. . . 105

*I dedicate this to my mother.*

# Chapter 1

# Introduction

## 1.1 Background

Statistical estimations and inferences are generally based on some assumptions about the underlying situation (e.g., model distributions of the data). In practice, these assumptions may not always be fulfilled. For instance, observations may deviate from specified model distribution or may contain gross errors. All such differently behaving observations are generally called contaminated observations. It is well known that some of the most common statistical procedures are extremely sensitive to the presence of contaminated observations, and therefore, many robust alternatives have been proposed.

Most of the traditional robust procedures are based on the assumption that the majority of the observations or cases in the data follow a specified model distribution, while only a minority follow an arbitrary, unspecified distribution. This assumption usually refers to casewise contamination or rowwise contamination, whose name comes from representing an observed data set in a matrix where rows are cases and columns are variables (see Figure 1.1). Traditional robust procedures then flag and down-weight contaminated cases entirely, and they have been shown to work well under these assumption.

However, the casewise contamination model does not always hold for real data because observations may only be partially contaminated. This type of contamination often appears as single outlying cells in a data matrix and can be modeled as independent cellwise contamination (see Figure 1.1). Under the cellwise contamination model, the traditional practice of down-weighting entire cases is no longer appropriate because it could entail a serious loss of information. Nonetheless, it still works

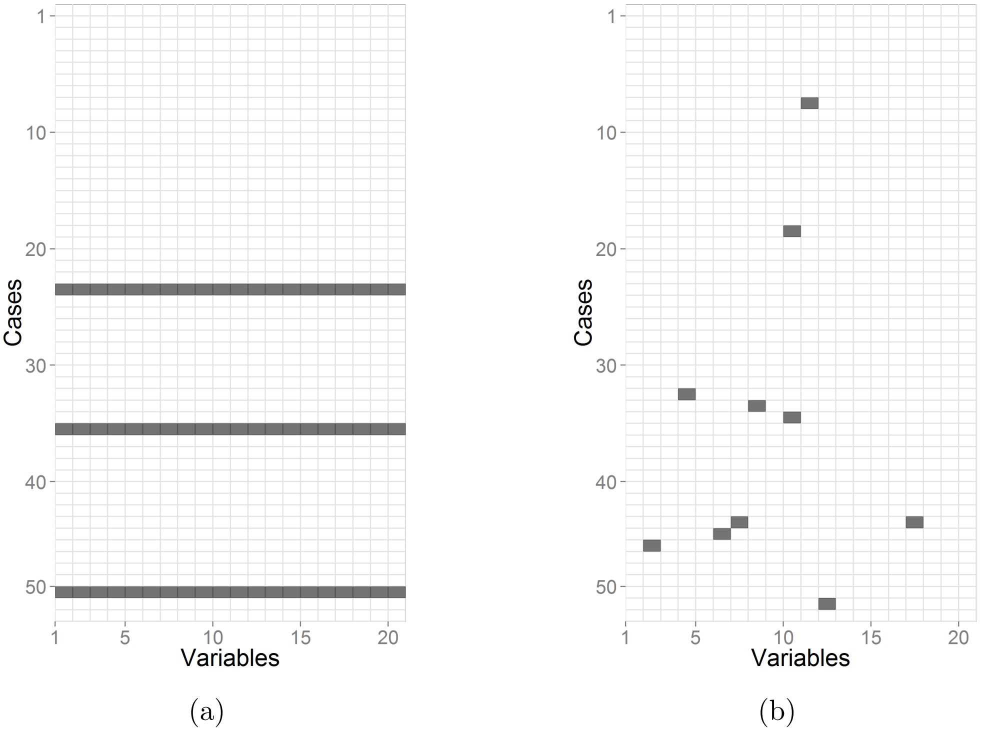


Figure 1.1: Illustration of (a) casewise contamination and (b) cellwise contamination in a data matrix.

under the circumstance that the fraction of cases affected by cellwise contamination is small, which is usually the case for small data sets (small number of variables).

Datasets in recent years often contain a large number of variables, and as such, they could suffer from a large fraction of cellwise contaminated cases. For example, if the proportion of cellwise contamination is $\varepsilon = 0.05$ and the dimension is $p = 10$, then the probability $\overline{\varepsilon}$ that at least one component of a case is contaminated is

$$\overline{\varepsilon} = 1 - (1 - \varepsilon)^p = 0.40;$$

if $\varepsilon = 0.05$ and $p = 20$, then $\overline{\varepsilon} = 0.64$; and if $\varepsilon = 0.05$ and $p = 30$, then $\overline{\varepsilon} = 0.79$. This phenomenon is referred to as the propagation of (cellwise) outliers, challenging the fundamental assumption required by the traditional robust procedures.

In real life situations, data sets may even contain both casewise and cellwise outliers, further complicating the problem. The following two examples provide evidence of the occurrence of cellwise and casewise outliers in real data and illustrate the following fact: Traditional casewise-robust procedures are not sufficient for dealing with these two types of outliers simultaneously. As a result, they fail to provide a good fit to the bulk of the data and miss out real outliers.

## 1.2 Real data examples

### 1.2.1 Geochemical data

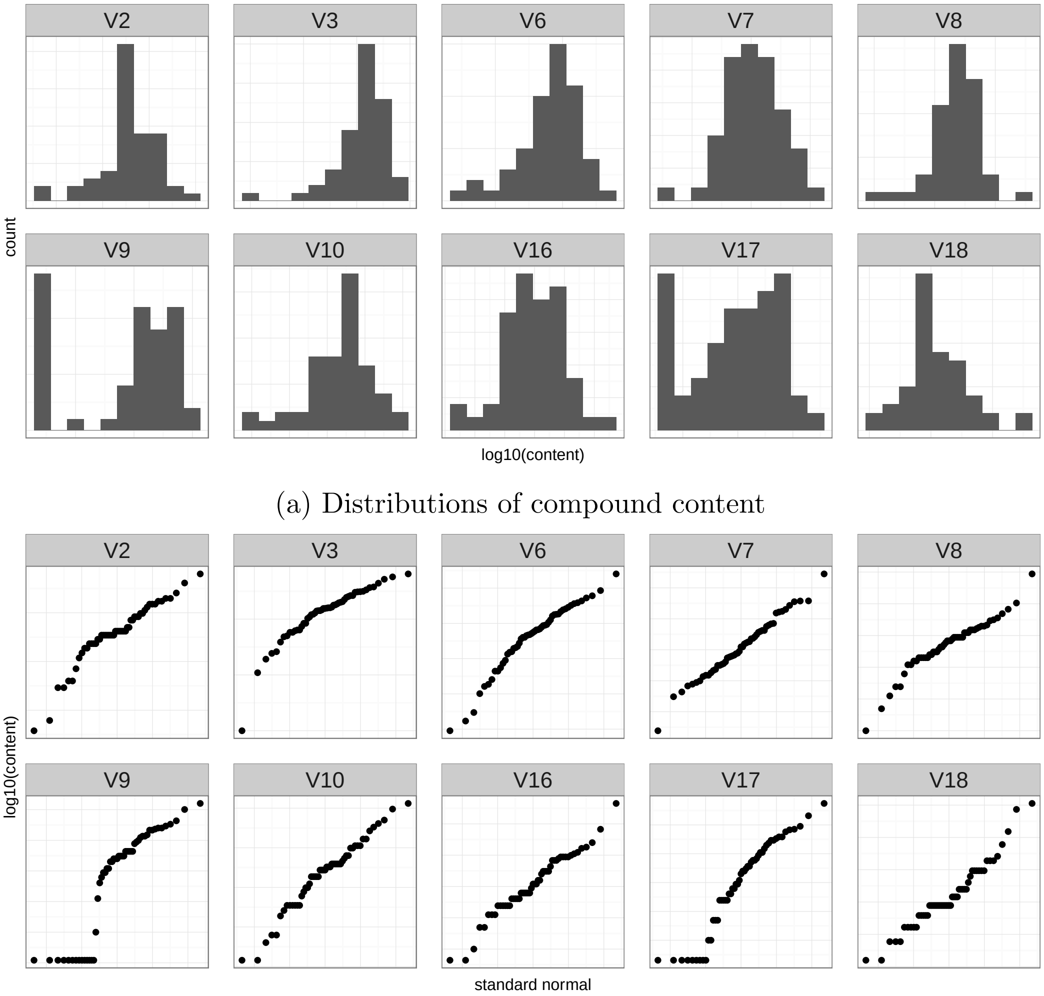
Consider the geochemical data in Smith et al. (1984). The data contain content measure (in parts per million) for 20 chemical compounds in 53 samples of rocks in Western Australia. In this example, we focus on a subset of 10 chemical compounds with the most suspected cellwise contamination. As the original data are skewed, we apply a log transformation to the data to make them more symmetric.

Figure 1.2 presents the distribution of the content measure for the 10 compounds in histograms and normal quantile–quantile plots. We notice a relatively large number of outliers in compound *V9* and *V17*, and we suspect a few outliers in the other compounds.

Consider the following outlier detection rule. Denote a sample of content measure by $X_1, \ldots, X_n$. Consider a pair of location and dispersion estimators $T_{0n}$ and $S_{0n}$. Here we use the median and the median absolute deviance for $T_{0n}$ and $S_{0n}$, respectively. Suppose the content measures have normal distributions. We flag a content measure if

$$X_i < T_{0,n} - 2.81 \cdot S_{0,n} \quad \text{or} \quad X_i > T_{0,n} + 2.81 \cdot S_{0,n}, \tag{1.1}$$

so that approximately only 0.5% of the measures would be flagged, assuming that $T_{0,n}$ and $S_{0,n}$ are close to true parameter values. Applying this rule, we find in total 5.1% of cellwise outliers that propagates to 41.5% of the cases, which is close to the



(a) Distributions of compound content

(b) Normal quantile–quantile plots of compound content

Figure 1.2: (a) Histograms and (b) quantile–quantile plots of the compound content in the geochmical data.

maximum contamination that traditional robust procedures can handle (i.e., 50%).

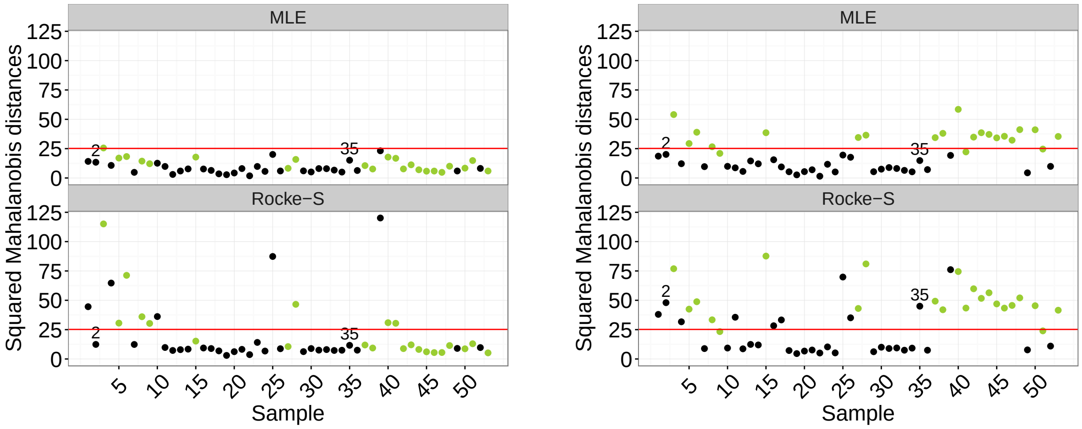
Let's investigate the effect of this propagation of cellwise outliers on traditional robust estimates. A common way for detecting outlying cases in high dimension is to use the squared Mahalanobis distance (MD):

$$d(\boldsymbol{x}, \hat{\boldsymbol{m}}, \hat{\boldsymbol{C}}) = (\boldsymbol{x} - \hat{\boldsymbol{m}})^t \hat{\boldsymbol{C}}^{-1} (\boldsymbol{x} - \hat{\boldsymbol{m}}),$$

where $\boldsymbol{x}$ is a $p$-dimensional data vector, $\hat{\boldsymbol{m}}$ is a multivariate location estimate, and $\hat{\boldsymbol{C}}$ is a covariance matrix estimate. Under the assumption that the data are normally distributed and that $\hat{\boldsymbol{m}}$ and $\hat{\boldsymbol{C}}$ are close to the true parameter values $\boldsymbol{m}$ and $\boldsymbol{C}$, the distance $d(\boldsymbol{x}, \boldsymbol{m}, \boldsymbol{C})$ has an approximate chi-squared distribution with $p$ degrees of freedom. With reasonably large sample sizes, the estimates will be close to their true values provided they are robust and consistent. Therefore, it is common practice to compare the squared Mahalanobis distances with a high percentile (such as the 99.5-th percentile) of a chi-squared with $p$ degrees of freedom. Points exceeding this threshold are flagged as possible outliers.

We consider the MLE approach (sample mean and covariance matrix) and a state-of-the-art robust approach (against casewise contamination), the Rocke-S-estimator (Maronna & Yohai, 2015). Figure 1.3a shows the squared Mahalanobis distances of the samples based on the two estimates. Cases that contain one or more flagged components are shown in green in the figure. Clearly, the MLE estimates are much affected by the outliers in the data; the corresponding MD's flag no green cases. The robust estimates are also affected by the propagation of cellwise outliers; the corresponding MD's fail to flag more than half of the suspected green cases.

Next, we replace the flagged cells by their coordinate medians in an attempt to control the effect of outliers propagation. Then, we re-estimate the multivariate location and covariance matrix using these "cleaned" data. The resulting squared Mahalanobis distances of the observations are given in Figure 1.3b. The robust estimates now recognize most of the cellwise contaminated observations as outlying observations. In addition, they unmask two new outlying observations (samples 2 and 35). These two observations do not have any outlying cellwise components and



(a) Original estimates

(b) Cleaned estimates

Figure 1.3: Squared Mahalanobis distances of the samples in the geochemical data based on estimates calculated on (a) the original data and (b) the cleaned data. Samples that contain one or more flagged components (large cellwise outliers) are in green.

they did not appear outlying based on the original robust estimates (see Figure 1.3a).

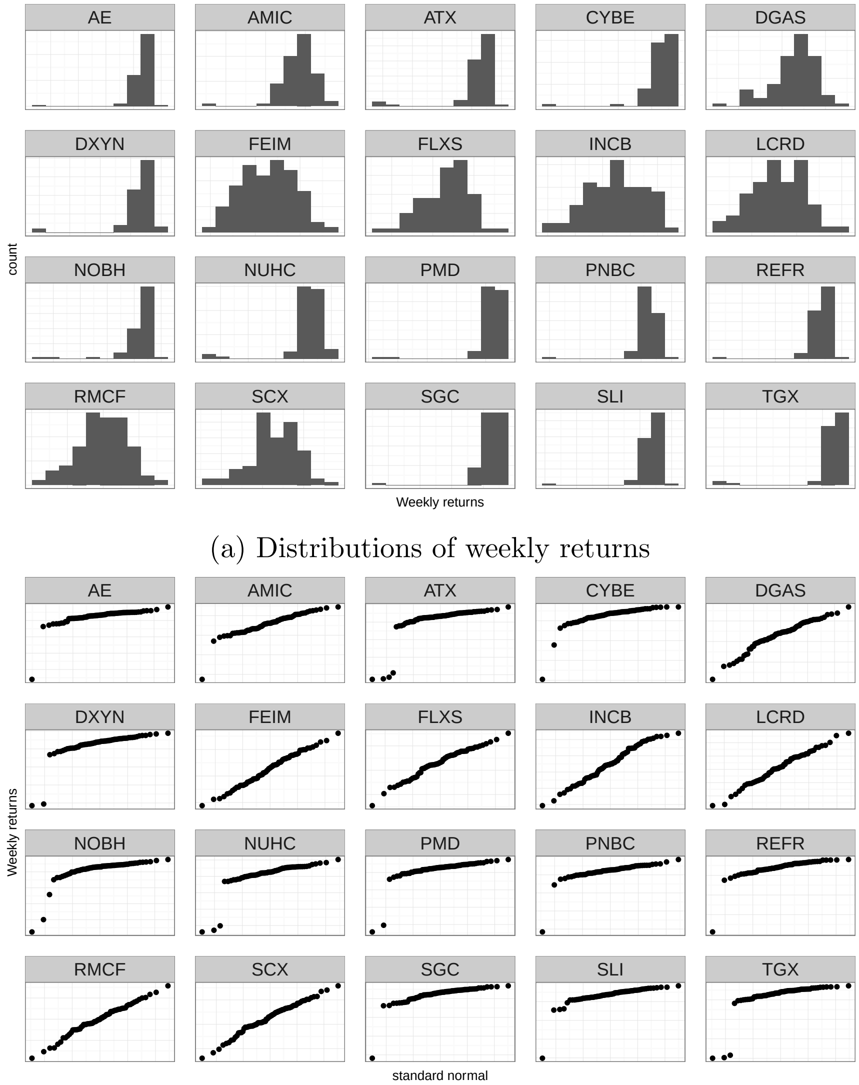
It is clear that cellwise data-preprocessing followed by some casewise-robust procedure are indeed necessary for capturing the full extent of contamination in these data (cellwise and casewise).

### 1.2.2 Micro-cap stock returns data

These data contain the weekly returns from 01/08/2008 to 12/28/2010 ($n = 730$ weeks) in a portfolio of 20 micro-cap stocks ($p = 20$) in Martin (2013).

Figure 1.4 shows the distributions and normal QQ-plots of the 20 micro-cap stocks returns in the portfolio. Overall, the returns do not seem to deviate much from normal, but they do contain a few outliers. Applying the outlier detection rule in (1.1) to these data, we identify 7.2% of outlying returns in the portfolio, propagating to 55.4% of the cases. Most of the weeks with outlying returns correspond to the 2008 financial crisis (from late-2008 to mid-2009).

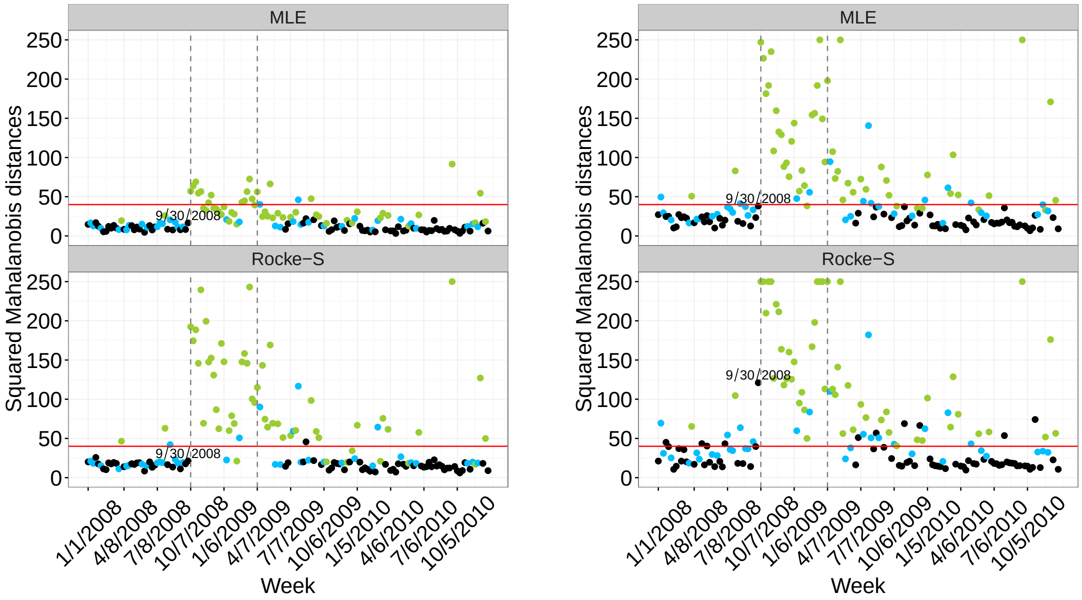
Figure 1.5a shows the squared Mahalanobis distances of the weekly observations



(a) Distributions of weekly returns

(b) Normal quantile–quantile plots of weekly returns

Figure 1.4: (a) Histograms and (b) quantile–quantile plots of the weekly returns in the micro-cap asset returns data.



(a) Original Data

(b) Cleaned Data

Figure 1.5: Squared Mahalanobis distances of the weekly observations in the microcap asset returns data based on estimates calculated on (a) the original data and (b) the cleaned data. Large distances are truncated for better visualization. Observations that contain one flagged components (large cellwise outliers) are in blue and those contain at least two flagged components are in green. The weeks corresponding to the 2008 financial crisis are enclosed by the vertical dashed lines.

based on the MLE and the Rocke-S estimates calculated on the original data. A total of 35 weeks (22.3% of the cases) contain one flagged component and are shown in blue in the figures. Another 52 weeks (33.1% of the cases) contain two or more flagged components and are shown in green. The financial crisis is the period between the two vertical dashed lines. Notice that all the weeks in the financial crisis are either blue or green.

As expected, the MLE estimates are adversely affected by the outliers and consider many of the weeks during the crisis as normal weeks, contradicting intuition. They fail to flag 35 green and 32 blue cases. The casewise-robust estimate is also upset by the propagation of cellwise outliers and misses 7 green and 29 blue cases, two of them during the financial crisis. The propagation of cellwise outliers has distorted

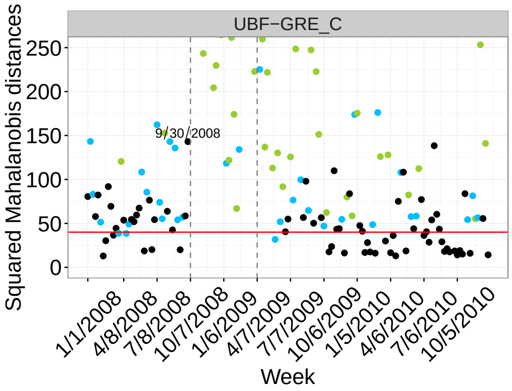


Figure 1.6: Results of UBF-GRE-C applied to the original data.

the robust estimates and corresponding MD's.

To try to control the effect of propagation of cellwise outliers, we repeat the analysis but this time replacing the flagged cells by their coordinate medians. Figure 1.5b shows the squared Mahalanobis distances based on the new estimates calculated on the cleaned data. The MLE and the robust estimates now flag all the weeks during the crisis as outliers, no longer contradicting intuition. Also, interestingly, the robust estimates now unmask additional outlying weeks (e.g., Week 09/30/2008) that are casewise outliers masked in the original analysis.

Figure 1.6 shows the results from more sophisticated approach developed in Chapter 3, called UBF-GRE-C. This procedure is able to flag all the green weeks, all but 3 of the blue weeks and 40 new casewise outlying weeks which were masked by the propagation of cellwise outliers in previous analysis.

## 1.3 Contributions and outline of the thesis

From the examples, we see that casewise and cellwise outliers could co-exist in real data. We also see from these examples that traditional robust procedures are inadequate to provide reliable estimates and to detect outliers in the presence of cellwise-outliers propagation. To address these problems, we propose and study new robust methods for dealing with cellwise and casewise contamination in this thesis. Our contributions are listed below.

- We provide several real data examples with convincing evidence of simultaneous occurrence of cellwise and casewise contamination.

- We study the problem of robust estimation of multivariate location and scatter matrices in the presence of cellwise and casewise contamination. These quantities are of great importance as they are cornerstones in multivariate data analysis. We propose a new procedure and show that this procedure can efficiently deal with cellwise outliers. Moreover, we show that the proposed procedure can deal with casewise outliers for datasets of moderate dimension ($p \leq 15$, say), performing similarly to traditional robust methods. Furthermore, we show that, under no contamination, the procedure is consistent and highly efficient. Part of the procedure is published as a discussion paper in TEST (see Agostinelli et al., 2015).

- We revisit the same problem above but for higher dimensional data ($p > 15$, say). We improve the proposed procedure in TEST in two different aspects: robustness under casewise contamination and computation speed. We equip the procedure with a new robust estimator for incomplete data that can simultaneously attain high robustness and reasonable efficiency for moderate to large dimension. We also develop a new fast subsampling method for computing initial estimates for the procedure when the dimension is large. This work is submitted for publication.

- We study the classic problem of multiple linear regression in the same paradigm. We propose a new procedure for estimating regression coefficients that can handle cellwise and casewise outliers similarly well. We prove that the procedure is consistent and asymptotically normal for the case of continuous covariates. This allows for statistical inference, at least for large sample sizes. Furthermore, the procedure is extended to handle both continuous and dummy covariates using an iterative algorithm in estimation. To the best of our knowledge, the proposed procedure is the first robust regression methods that can achieve robust estimation and inference in the presence of cellwise and casewise contamination, and can deal with numerical and dummy covariates. The procedure is published as a methodology paper in Computational Statistics and Data Analysis (CSDA) (see Leung et al., 2016).

- Finally, we develop two `R` packages for robust analysis in the presence of cellwise and casewise contamination. The first `R` package, `GSE` (Leung et al., 2015), implements our proposed procedure for estimating multivariate location and scatter (Agostinelli et al., 2015). The package also implements several robust multivariate location and scatter estimators for incomplete data that are heavily used by our procedure such as the generalized S-estimator (Danilov, 2010). The second `R` package, `robreg3S` (Leung et al., 2015), implements our proposed regression estimator for robust estimation and inference in multiple linear regression (see Leung et al., 2016). The two `R` packages are freely available on CRAN (R Core Team, 2015).

The rest of the thesis is organized as follows. Chapter 2 and 3 is dedicated to our work on robust estimation of multivariate location and scatter and Chapter 4 is on robust linear regression analysis. The thesis concludes in Chapter 5, where some of the challenges that remain to be solved and the directions we foresee for future work are presented.

# Chapter 2

# Robust Estimation of Multivariate Location and Scatter in the Presence of Cellwise and Casewise Contamination in Moderate Dimension

## 2.1 Introduction

Outliers are a common problem for data analysts because they may have a big detrimental effect on estimation, inference and prediction. In robust statistics it is generally assumed that a relatively small fraction of cases may be contaminated, however, it has also been recently noticed that the majority of the cases (and even all of them) could be partially contaminated for moderate and high-dimensional data. The problem of interest in this chapter is robust estimation of multivariate location and scatter matrix in consideration of the latter case. The estimation of these parameters is a corner stone in many applications such as principal component analysis, factor analysis, and multiple linear regression.

### Classical contamination model

To fix ideas, suppose that a multivariate data set is organized in a table with rows as cases and columns as variables, that is, $\mathbb{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^t$, with $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})$.

The vast majority of procedures for robust analysis of multivariate data are based on the classical Tukey–Huber contamination model (THCM), or sometimes called casewise contamination model, where a small fraction of rows in the data table may be contaminated. In THCM, the contamination mechanism is modeled as a mixture of two distributions: one corresponding to the nominal model and the other corresponding to the outliers. More precisely, THCM considers the following family of distributions:

$$\mathscr{H}_\epsilon = \{H = (1-\epsilon)H_0 + \epsilon\widetilde{H} : \widetilde{H} \text{ is any distribution on } \mathbb{R}^p\} \tag{2.1}$$

where $H_0$ is a central parametric distribution such as the multivariate normal $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\widetilde{H}$ is an unspecified outlier generating distribution. We then assume a case follows a distribution from the above family, that is $\boldsymbol{X}_i \sim H$ where $H \in \mathscr{H}_\epsilon$. The key feature of this model is that when $\epsilon$ is small we have $\boldsymbol{X}_i \sim H_0$ most of the time, therefore, detection and down-weighting of outlying cases make sense and work well in practice. High breakdown point affine equivariant estimators such as MVE (Rousseeuw, 1985), MCD (Rousseeuw, 1985), S (Davies, 1987), MM (Tatsuoka & Tyler, 2000) and Stahel–Donoho estimators (Stahel, 1981; Donoho, 1982) proceed in this general way.

## Independent contamination model

In many applications, however, the contamination mechanism may be different in that individual components (or cells) in $\mathbb{X}$ are independently contaminated. For instance, in the case of high-dimensional data, variables could be gathered separately and therefore, exposed to contamination independently. The cellwise contamination mechanism may in principle seem rather harmless, but in fact it has far reaching consequences including the possible breakdown of classical high breakdown point estimators.

The new contamination framework, called independent contamination model (ICM), was presented and formalized in Alqallaf et al. (2009). In the ICM framework, we consider a different family of distribution:

$$\mathscr{I}_\epsilon = \{H : H \text{ is the distribution of } \boldsymbol{X} = (\boldsymbol{I} - \boldsymbol{B}_\epsilon)\boldsymbol{X}_0 + \boldsymbol{B}_\epsilon \widetilde{\boldsymbol{X}}\}, \tag{2.2}$$

where $\boldsymbol{X}_0 \sim H_0$, $\widetilde{\boldsymbol{X}} \sim \widetilde{H}$, and $\boldsymbol{B}_\epsilon = \text{diag}(B_1, \ldots, B_p)$, where the $B_j$ are independent $Bin(1, \epsilon)$. In other words, each component of $\boldsymbol{X}$ has a probability $\epsilon$ of being independently contaminated. Furthermore, the probability $\overline{\epsilon}$ that at least one component of $\boldsymbol{X}$ is contaminated is now

$$\overline{\epsilon} = 1 - (1 - \epsilon)^p.$$

This implies that even if $\epsilon$ is small, $\overline{\epsilon}$ could be large for large $p$, and could exceed the 0.5 breakdown point of highly robust affine equivariant estimators under THCM. For example, if $\epsilon = 0.1$ and $p = 10$, then $\overline{\epsilon} = 0.65$; if $\epsilon = 0.05$ and $p = 20$, then $\overline{\epsilon} = 0.64$ and if $\epsilon = 0.01$ and $p = 100$, then $\overline{\epsilon} = 0.63$.

Alqallaf et al. (2009) showed that for this type of contamination, the breakdown point of all the traditional 0.5 breakdown point and affine equivariant location estimators is $1 - 0.5^{1/p} \to 0$ as $p \to \infty$. It can be shown that the same holds for robust and affine equivariant scatter estimators. Hence, we have a new manifestation of the *curse of dimensionality*: when $p$ is large, traditional robust estimators break down for a rather small fraction of independent contamination.

To remedy this problem, some researchers have proposed to Winsorize potential outliers for each variable separately. For instance, Alqallaf et al. (2002) revisited Huberized Pairwise Covariance (Huber, 1981), which is constructed using transformed correlation coefficients calculated separately on Huberized data as basic building blocks. *Huberization* is a form of Winsorization. Although pairwise robust estimators show some robustness under ICM, they cannot deal with casewise outliers from THCM, as well as finely shaped multivariate data. Another approach to deal with ICM outliers was proposed in Van Aelst et al. (2012). They modified the Stahel–Donoho (SD) estimator (Stahel, 1981; Donoho, 1982) by calculating the SD-outlyingness measure and weights on Huberized data instead of the raw data. In our simulation study, this estimator performs very well under THCM, but is not sufficiently robust under ICM.

An alternative approach consists of replacing cellwise outliers by NA's, which we call the approach *filtering*, like how oversize particles can be separated in filtration. The use of filtering to fend against cellwise contamination has been suggested by various authors (e.g., Danilov, 2010; Van Aelst et al., 2012; Farcomeni, 2014a). In particular, Danilov (2010) has compared maximum likelihood estimates of covariance matrix computed for various pre-processed data, and empirically found that filtered data yield the most robust estimates against large cellwise contamination. Farcomeni (2014a) proposed an interesting idea to estimate location and scatter matrix by optimizing some maximum likelihood over the parameters of interest, as well as the filtering set with a fixed size (the filtering operation was called *snipping* in the paper). The original method of Farcomeni (2014a) was for clustering multivariate data where each cluster has an unknown location and scatter matrix, but it can be easily adapted to our problem by fixing the number of clusters to one. In our simulation study, the estimators of Danilov (2010) and Farcomeni (2014a) perform very well under ICM, but neither is sufficiently robust under THCM.

The main goal of this chapter is to emphasize the need for a new generation of global–robust estimators that can simultaneously deal with outliers from ICM and THCM, as well as to to define new robust estimators that can deal with them.

In Section 2.2, we introduce a global–robust estimator of multivariate location and scatter. In Section 2.3, we show that our estimation procedure is strongly consistent. That is, the multivariate location estimator converges a.s. to the true location and the scatter matrix estimator converges a.s. to a scalar multiple of the true scatter matrix, for a general elliptical distribution. Moreover, for a normal distribution the scalar factor is equal to one. In Section 2.4, we report the result of a simulation study. In Section 2.5, we analyze two real data sets using the proposed and several competing estimators. In Section 2.6, we discuss several main points raised by the discussants of the original paper of this chapter. Finally, we conclude in Section 2.7. We also provide some additional numerical results and all the proofs in Appendix A.

## 2.2 Global-robust estimation under THCM and ICM

Our approach for global–robust estimation under THCM and ICM is to first flag univariate outlying cells in the data table and to replace them by NA's. In the second step we then apply a procedure that is robust against casewise outliers. Two-step procedures like this were relatively unpopular in the robustness field because of the potential lack of desirable statistical properties for the final estimate (such as consistency and efficiency) and also because it had not been convincingly shown that the final estimate is robust under THCM and ICM. These two limitations are overcome in our procedure by the use of an adaptive filter (Gervini & Yohai, 2002) in the first step and a generalized S-estimator (GSE) (Danilov et al., 2012) in the second step.

More precisely, our procedure has two major steps:

**Step I.** *Filtering large cellwise outliers.* We flag cellwise outliers and replace them by NA's. This step prevents cellwise contaminated cases from having large robust Mahalanobis distances in the second step. See Section 2.2.1 for further details.

**Step II.** *Dealing with casewise outliers.* We apply GSE, to the filtered data coming from Step I. Notice that GSE has been specifically designed to deal with incomplete multivariate data with casewise outliers. See Section 2.2.2 for further details.

Full account of these steps is provided in the remaining of this section.

### 2.2.1 Step I: Filtering cellwise outliers

Consider a random sample of $\mathbb{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^t$, where $\boldsymbol{X}_i$ follows a distribution from $\mathscr{I}_\epsilon$ in (2.2). The filtering we present consists of two parts: a part that aims at detecting large cellwise outliers by looking at marginals, and another part that aims at detecting moderate cellwise outliers by incorporating information about the correlation structure of the data.

## Univariate filtering

Let $X_1, \ldots, X_n$ be a random (univariate) sample of observations. Consider a pair of initial location and dispersion estimators, $T_{0n}$ and $S_{0n}$. Common choices for $T_{0n}$ and $S_{0n}$ that are also adopted in this chapter are the median and median absolute deviation (MAD). Denote the standardized sample by $Z_i = (X_i - T_{0n})/S_{0n}$. Let $F$ be a chosen reference distribution for $Z_i$. An ideal choice for a reference distribution would be $F_0$, the actual distribution of $(X_i - \mu_0)/\sigma_0$. Unfortunately, $F_0$ is unknown in practice. Thus, we use the standard normal distribution, $F = \Phi$, as an approximation. A normalizing transformation could be applied if the marginal data do not seem normal from standard diagnostic tools such as normal quantile-quantile plots.

Instead of a fixed cutoff value for $Z_i$, we introduce an adaptive cutoff (Gervini & Yohai, 2002) which is asymptotically "correct", meaning that for clean data the fraction of flagged outliers tends to zero as the sample size $n$ tends to infinity. The adaptive cutoff values are defined as follows. Let $F_n^+$ be the empirical distribution function for the absolute standardized value, that is,

$$F_n^+(t) = \frac{1}{n} \sum_{i=1}^{n} I(|Z_i| \leq t).$$

The proportion of flagged outliers is defined by

$$\begin{aligned} d_n &= \sup_{t \geq \eta} \left\{ F^+(t) - F_n^+(t) \right\}^+ \\ &= \max_{i > i_0} \left\{ F^+(|Z|_{(i)}) - \frac{(i-1)}{n} \right\}^+, \end{aligned} \tag{2.3}$$

where in general $\{a\}^+$ represents the positive part of $a$ and $F^+$ is the distribution of $|Z|$ when $Z \sim F$. Here, $|Z|_{(i)}$ is the order statistics of $|Z_i|$, $i_0 = \max\{i : |Z|_{(i)} < \eta\}$, and $\eta = (F^+)^{-1}(\alpha)$ is a large quantile of $F^+$. We use $\alpha = 0.95$ for univariate filtering as the aim is to detect large outliers, but other choices could be considered. Then, we flag $\lfloor nd_n \rfloor$ observations with the largest standardized value as cellwise outliers and replace them by NA's (here, $\lfloor a \rfloor$ is the largest integer less than or equal to $a$). Finally, the resulting adaptive cutoff value for $Z_i$ is

$$t_n = \min \left\{ t : F_n^+(t) \geq 1 - d_n \right\}, \tag{2.4}$$

that is, $t_n = Z_{(i_n)}$ with $i_n = n - \lfloor nd_n \rfloor$. Equivalently, we flag $X_i$ if $|Z_i| \geq t_n$.

The following proposition states that even when the actual distribution is unknown, asymptotically, the univariate filter will not flag outliers when the tail of the chosen reference distribution is heavier (or equal) than the tail of the actual distribution. We call this property *consistency* throughout this thesis.

**Proposition 2.1.** *Consider a random variable $X \sim F_0$ with $F_0$ continuous. Also, consider a pair of location and dispersion estimators, $T_{0n}$ and $S_{0n}$, such that $T_{0n} \to \mu_0 \in \mathbb{R}$ and $S_{0n} \to \sigma_0 > 0$ a.s. [$F_0$]. Let $F_0^+(t) = P_{F_0}(|\frac{X-\mu_0}{\sigma_0}| \leq t)$. If the reference distribution $F^+$ satisfies the inequality*

$$\max_{t \geq \eta} \left\{ F^+(t) - F_0^+(t) \right\} \leq 0, \tag{2.5}$$

*then*

$$\frac{n_0}{n} \to 0 \ a.s.,$$

*where*

$$n_0 = \lfloor nd_n \rfloor.$$

**Proof:** See Section A.2 in the Appendix.

## Bivariate filtering

As pointed out by Rousseeuw & Van den Bossche (2015), to filter the univariate outliers based solely on their value may be too limiting as no correlation with other variables is taken into account. A moderately contaminated cell may pass the filter when viewed marginally, but it may be flagged as an outlier when viewed together with other components, especially for highly correlated data.

Let $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$, with $\boldsymbol{X}_i = (X_{i1}, X_{i2})^t$, be a random sample of bivariate observations. Consider also a pair of initial location and scatter estimators,

$$\boldsymbol{T}_{0n} = \begin{pmatrix} T_{0n,1} \\ T_{0n,2} \end{pmatrix} \quad \text{and} \quad \boldsymbol{C}_{0n} = \begin{pmatrix} C_{0n,11} & C_{0n,12} \\ C_{0n,21} & C_{0n,22} \end{pmatrix}.$$

Similar to the univariate case we use the coordinate-wise median and the bivariate Gnanadesikan-Kettenring estimator with MAD scale (Gnanadesikan & Kettenring, 1972) for $\boldsymbol{T}_{0n}$ and $\boldsymbol{C}_{0n}$, respectively. More precisely, the initial scatter estimators are defined by

$$C_{0n,jk} = \frac{1}{4}\left(\text{MAD}(\{X_{ij} + X_{ik}\})^2 - \text{MAD}(\{X_{ij} - X_{ik}\})^2\right),$$

where $\text{MAD}(\{Y_i\})$ denotes the MAD of $Y_1, \ldots, Y_n$. Note that $C_{0n,jj} = \text{MAD}(\{X_j\})^2$, which agrees with our choice of the coordinate-wise dispersion estimators. Now, denote the pairwise (squared) Mahalanobis distances by $D_i = (\boldsymbol{X}_i - \boldsymbol{T}_{0n})^t \boldsymbol{C}_{0n}^{-1}(\boldsymbol{X}_i - \boldsymbol{T}_{0n})$. Let $G_n$ be the empirical distribution for pairwise Mahalanobis distances,

$$G_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(D_i \leq t).$$

Finally, we filter outlying points $\boldsymbol{X}_i$ by comparing $G_n(t)$ with $G(t)$, where $G$ is a chosen reference distribution. In this thesis, we use the chi-squared distribution with two degrees of freedom, $G = \chi^2_2$. The proportion of flagged bivariate outliers is defined by

$$\begin{aligned} d_n &= \sup_{t \geq \eta} \{G(t) - G_n(t)\}^+ \\ &= \max_{i > i_0} \left\{ G(D_{(i)}) - \frac{(i-1)}{n} \right\}^+. \end{aligned} \tag{2.6}$$

Here, $\eta = G^{-1}(\alpha)$, and we use $\alpha = 0.85$ for bivariate filtering since we now aim for moderate outliers, but other choices of $\alpha$ can be considered. Then, we flag $\lfloor n d_n \rfloor$

observations with the largest pairwise Mahalanobis distances as outlying bivariate points. The resulting adaptive cutoff value for the distances can be defined in the same way as in (2.4). Finally, the following proposition states the consistency property of the bivariate filter.

**Proposition 2.2.** *Consider a random vector* $\boldsymbol{X} = (X_1, X_2)^t \sim H_0$*. Also, consider a pair of bivariate location and scatter estimators,* $\boldsymbol{T}_{0n}$ *and* $\boldsymbol{C}_{0n}$*, such that* $\boldsymbol{T}_{0n} \to \boldsymbol{\mu}_0 \in \mathbb{R}^2$ *and* $\boldsymbol{C}_{0n} \to \boldsymbol{\Sigma}_0 \in \mathrm{PDS}(2)$ *a.s. [*$H_0$*]* ($\mathrm{PDS}(q)$ *is the set of all positive definite symmetric matrices of size* $q$*). Let* $G_0(t) = P_{H_0}((\boldsymbol{X} - \boldsymbol{\mu}_0)^t \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{X} - \boldsymbol{\mu}_0) \leq t)$ *and suppose that* $G_0$ *is continuous. If the reference distribution* $G$ *satisfies:*

$$\max_{t \geq \eta} \{G(t) - G_0(t)\} \leq 0, \tag{2.7}$$

*then*

$$\frac{n_0}{n} \to 0 \ a.s.,$$

*where*

$$n_0 = \lfloor n d_n \rfloor.$$

**Proof:** See Section A.2 in the Appendix.

## Combining the univariate and bivariate filtering

We first apply the univariate filter to each variable in $\mathbb{X}$ separately using the initial location and dispersion estimators, $\boldsymbol{T}_{0n} = (T_{0n,1}, \ldots, T_{0n,p})$ and $\boldsymbol{S}_{0n} = (S_{0n,1}, \ldots, S_{0n,p})$. Let $\mathbb{U}$ be the resulting auxiliary matrix of zeros and ones with zeros indicating the filtered entries in $\mathbb{X}$. We next iterate over all pairs of variables in $\mathbb{X}$ to identify outlying bivariate points which helps filtering the moderately contaminated cells.

Fix a pair of variables, $(X_{ij}, X_{ik})$ and set $\boldsymbol{X}_i^{(jk)} = (X_{ij}, X_{ik})$. Let $\boldsymbol{C}_{0n}^{(jk)}$ be an initial pairwise scatter matrix estimator for this pair of variables. We calculate the pairwise Mahalanobis distances $D_i^{(jk)} = (\boldsymbol{X}_i^{(jk)} - \boldsymbol{T}_{0n}^{(jk)})^t(\boldsymbol{C}_{0n}^{(jk)})^{-1}(\boldsymbol{X}_i^{(jk)} - \boldsymbol{T}_{0n}^{(jk)})$ and perform the bivariate filtering on the pairwise distances with no flagged components from the univariate filtering: $\{D_i^{(jk)} : U_{ij} = 1, U_{ik} = 1\}$. We apply this procedure to all pairs of variables $1 \leq j < k \leq p$. Let

$$J = \left\{ (i, j, k) : D_i^{(jk)} \text{ is flagged as bivariate outlier} \right\},$$

be the set of triplets which identify the pairs of cells flagged by the bivariate filter in rows $i = 1, ..., n$. It remains to determine which cells $(i, j)$ in row $i$ are to be flagged as cellwise outliers. For each cell $(i, j)$ in the data table, $i = 1, \ldots, n$ and $j = 1, \ldots, p$, we count the number of flagged pairs in the $i$-th row where cell $(i, j)$ is involved:

$$m_{ij} = \# \left\{ k : (i, j, k) \in J \right\}.$$

Cells with large $m_{ij}$ are likely to correspond to univariate outliers. Suppose that observation $X_{ij}$ is not contaminated by cellwise contamination. Then $m_{ij}$ approximately follows the binomial distribution, $Bin(\sum_{k \neq j} U_{ik}, \delta)$, under ICM, where $\delta$ is the overall proportion of cellwise outliers that were not detected by the univariate filter. We flag observation $X_{ij}$ if

$$m_{ij} > c_{ij},$$

where $c_{ij}$ is the 0.99-quantile of $Bin(\sum_{k \neq j} U_{ik}, \delta)$. In practice we obtained good results (in both simulation and real data applications) using the conservative choice $\delta = 0.10$, which is adopted in this thesis.

### 2.2.2 Step II: Dealing with casewise outliers

This second step introduces robustness against casewise outliers that went undetected in Step I. Data that emerged from Step I have *holes* (i.e., NA's) that correspond to potentially contaminated cells. To estimate the multivariate location and scatter matrix from that data, we use a recently developed estimator called GSE, briefly reviewed below.

Let $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^t$, $1 \leq i \leq n$ be $p$-dimensional i.i.d. random vectors that follow a distribution in an elliptical family $\mathcal{E}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ with density

$$f_{\boldsymbol{X}}(\boldsymbol{x}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) = \frac{1}{|\boldsymbol{\Sigma}_0|} f_0(D(\boldsymbol{x}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)) \tag{2.8}$$

where $|A|$ is the determinant of $A$, $f_0$ is non-increasing and strictly decreasing at 0, and

$$D(\boldsymbol{x}, \boldsymbol{m}, \boldsymbol{C}) = (\boldsymbol{x} - \boldsymbol{m})^t \boldsymbol{C}^{-1} (\boldsymbol{x} - \boldsymbol{m})$$

is the squared Mahalanobis distance. We also use the normalized squared Mahalanobis distances

$$D^*(\boldsymbol{x}, \boldsymbol{m}, \boldsymbol{C}) = D(\boldsymbol{x}, \boldsymbol{m}, \boldsymbol{C}^*),$$

where $\boldsymbol{C}^* = \boldsymbol{C}/|\boldsymbol{C}|^{1/p}$, so $|\boldsymbol{C}^*| = 1$.

Let $\mathbb{U}$ be the auxiliary matrix of zeros and ones, with zeros indicating the corresponding missing entries. Let $p_i = p(\boldsymbol{U}_i) = \sum_{j=1}^{p} U_{ij}$ be the actual dimension of the observed part of $\boldsymbol{X}_i$. Given a $p$-dimensional vector of zeros and ones $\boldsymbol{u}$, a $p$-dimensional vector $\boldsymbol{m}$ and a $p \times p$ matrix $\boldsymbol{A}$, we denote by $\boldsymbol{m}^{(\boldsymbol{u})}$ and $\boldsymbol{A}^{(\boldsymbol{u})}$ the sub-vector of $\boldsymbol{m}$ and the sub-matrix of $\boldsymbol{A}$, respectively, with columns and rows corresponding to the positive entries in $\boldsymbol{u}$.

Let $\boldsymbol{\Omega}_{0n}$ be a $p \times p$ positive definite initial estimator for $\boldsymbol{\Sigma}_0$. Given the location vector $\boldsymbol{\mu} \in \mathbb{R}^p$ and a $p \times p$ positive definite matrix $\boldsymbol{\Sigma}$, we define the generalized M-scale, $s_{GS}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Omega}_{0n}, \mathbb{X}, \mathbb{U})$, as the solution in $s$ to the following equation:

$$\sum_{i=1}^{n} c_{p(\boldsymbol{U}_i)} \rho \left( \frac{D^*\left(\boldsymbol{X}_i^{(\boldsymbol{U}_i)}, \boldsymbol{\mu}^{(\boldsymbol{U}_i)}, \boldsymbol{\Sigma}^{(\boldsymbol{U}_i)}\right)}{s \, c_{p(\boldsymbol{U}_i)} \left|\boldsymbol{\Omega}_{0n}^{(\boldsymbol{U}_i)}\right|^{1/p(\boldsymbol{U}_i)}} \right) = b \sum_{i=1}^{n} c_{p(\boldsymbol{U}_i)} \tag{2.9}$$

where $\rho(t)$ is an even, non-decreasing in $|t|$ and bounded loss function. The tuning constants $c_k$, $1 \leq k \leq p$, are chosen such that

$$E_\Phi \left( \rho \left( \frac{||\boldsymbol{X}||^2}{c_k} \right) \right) = b, \quad \boldsymbol{X} \sim N_k(\boldsymbol{0}, \boldsymbol{I}), \tag{2.10}$$

to ensure consistency under the multivariate normal. We consider the Tukey's bisquare rho function, $\rho(u) = \min(1, 1-(1-u)^3)$, and $b = 0.5$ throughout this chapter.

The inclusion of $\boldsymbol{\Omega}_{0n}$ in (2.9) is needed to re-normalize the distances $D^*$ to achieve robustness. A heuristic argument for the inclusion of $\boldsymbol{\Omega}_{0n}$ is as follows. Suppose that $\boldsymbol{T}_n \approx \boldsymbol{\mu}_0$ and $\boldsymbol{S}_n \approx \boldsymbol{\Omega}_{0n} \approx \boldsymbol{\Sigma}_0$. Then, given $\boldsymbol{U} = \boldsymbol{u}$,

$$\frac{D^*(\boldsymbol{X}^{(\boldsymbol{u})}, \boldsymbol{T}_n^{(\boldsymbol{u})}, \boldsymbol{S}_n^{(\boldsymbol{u})})}{c_{p(\boldsymbol{u})} \left|\boldsymbol{\Omega}_{0n}^{(\boldsymbol{u})}\right|^{1/p(\boldsymbol{u})}} \approx \frac{D^*(\boldsymbol{X}^{(\boldsymbol{u})}, \boldsymbol{\mu}_0^{(\boldsymbol{u})}, \boldsymbol{\Sigma}_0^{(\boldsymbol{u})})}{c_{p(\boldsymbol{u})} \left|\boldsymbol{\Sigma}_0^{(\boldsymbol{u})}\right|^{1/p(\boldsymbol{u})}} \sim \frac{||\boldsymbol{Y}^{(\boldsymbol{u})}||^2}{c_{p(\boldsymbol{u})}}$$

where $\boldsymbol{Y}^{(\boldsymbol{u})}$ is a $p(\boldsymbol{u})$ dimensional random vector with an elliptical distribution. Hence, $||\boldsymbol{Y}^{(\boldsymbol{u})}||^2/c_{p(\boldsymbol{u})}$ has M-scale of 1 for the given $\rho$ function if $\boldsymbol{Y}$ is normal, and large Mahalanobis distances can be down-weighted accordingly. Here, we use extended minimum volume ellipsoid (EMVE) for $\boldsymbol{\Omega}_{0n}$ as suggested in Danilov et al. (2012).

A generalized S-estimator is then defined by

$$(\boldsymbol{T}_{GS}, \boldsymbol{C}_{GS}) = \arg\min_{\boldsymbol{\mu}, \boldsymbol{\Sigma}} s_{GS}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Omega}_{0n}, \mathbb{X}, \mathbb{U}) \tag{2.11}$$

subject to the constraint

$$s_{GS}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}, \mathbb{X}, \mathbb{U}) = 1. \tag{2.12}$$

Under mild regularity assumptions, in the case of elliptical data with $\boldsymbol{U}_i$ independent of $\boldsymbol{X}_i$ (missing completely at random assumption) any solution to (2.11) is a consistent estimator for the shape of the scatter matrix. Moreover, in the case of normal data, any solution to (2.11) satisfying (2.12) is consistent in shape and size for the *true* covariance matrix. Proofs of these claims, as well as the formulas and the derivations of the estimating equation for GSE, can be found in Danilov et al. (2012).

Finally, our two-step location and scatter estimator is defined by

$$\begin{aligned} \boldsymbol{T}_{2S} &= \boldsymbol{T}_{GS}(\mathbb{X}, \mathbb{U}_n) \\ \boldsymbol{C}_{2S} &= \boldsymbol{C}_{GS}(\mathbb{X}, \mathbb{U}_n) \end{aligned} \tag{2.13}$$

where $\mathbb{U}_n$ is an estimated matrix of zeros and ones with zeros indicated filtered entries in the data table $\mathbb{X}$.

## 2.3 Consistency of GSE on filtered data

The missing data created in Step I are not missing at random because the missing data indicator, $\mathbb{U}$, depends on the original data $\mathbb{X}$ (univariate outliers are declared missing). Therefore, the consistency of our two-step estimator cannot be directly derived from Danilov et al. (2012). However, as shown in Theorem 2.1 below, our procedure is consistent at the central model provided the fraction of missing data converges to zero. We need the following assumptions:

**Assumption 2.1.** *The function $\rho$ is (i) non-decreasing in $|t|$, (ii) strictly increasing at 0, (iii) continuous, and (iv) $\rho(0) = 0$ and (v) $\lim_{v\to\infty} \rho(v) = 1$ (e.g., Tukey's bisquare rho function).*

**Assumption 2.2.** *The random vector $\boldsymbol{X}$ follows a distribution, $H_0$, in the elliptical family defined by (2.8).*

**Assumption 2.3.** *Let $H_0$ be the distribution of $\boldsymbol{X}$ and denote $\sigma(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ the solution in $\sigma$ to the following equation*

$$E_{H_0}\left(\rho\left(\frac{D(\boldsymbol{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{c_p \sigma}\right)\right) = b,$$

*and consider the minimization problem,*

$$\min_{|\boldsymbol{\Sigma}|=1} \sigma(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{2.14}$$

*We assume that (2.14) has a unique solution, $(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_{00})$, where $\boldsymbol{\mu}_0 \in \mathbb{R}^p$ and $\boldsymbol{\Sigma}_{00} \in$ PDS$(p)$. We also put $\sigma_0 = \sigma(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_{00})$.*

**Assumption 2.4.** *The proportion of fully observed entries,*

$$q_n = \#\{i, 1 \le i \le n : p_i = p(\boldsymbol{U}_{n,i}) = p\}/n,$$

*tends to one a.s. as $n$ tends to infinity. Recall that $\boldsymbol{U}_{n,i}$ is the indicator vector of non-filtered entries in $\boldsymbol{X}_i$.*

**Remark 2.1.** *Davies (1987) showed that Assumption 2.2 implies Assumption 2.3 with $\boldsymbol{\Sigma}_{00} = \boldsymbol{\Sigma}_0/|\boldsymbol{\Sigma}_0|$.*

**Remark 2.2.** *By Proposition 2.1 and 2.2, the procedure described in Step I satisfies Assumption 2.4, provided that the marginal distributions for the distribution that generated the data satisfy equation (2.5) and (2.7).*

**Theorem 2.1.** *Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be a random sample from $H_0$ and $\boldsymbol{U}_{n,1}, \ldots, \boldsymbol{U}_{n,n}$ be as described in Section 2.2.2. Suppose Assumptions 2.1–2.4 hold. Let $(\boldsymbol{T}_{GS}, \boldsymbol{C}_{GS})$ be the GSE defined by (2.11)–(2.13). Then,*

*(i) $\boldsymbol{T}_{GS} \to \boldsymbol{\mu}_0$ a.s. and*

*(ii) $\boldsymbol{C}_{GS} \to \sigma_0 \boldsymbol{\Sigma}_{00}$ a.s..*

*(iii) When $\boldsymbol{X} \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, we have $\sigma_0 \boldsymbol{\Sigma}_{00} = \boldsymbol{\Sigma}_0$.*

**Proof:** See Section A.2 in the Appendix.

## 2.4 Simulations

We conduct a simulation study in `R` (R Core Team, 2015) to compare various estimators from different generations of robust estimators of multivariate location and scatter:

(a) MCD, the fast minimum covariance determinant proposed by Rousseeuw & Van Driessen (1999) (see also Section 6.7.5 in Maronna et al., 2006). MCD is implemented in the `R` package `rrcov`, function `CovMcd`;

(b) MVE-S, the estimator proposed by Maronna et al. (2006, Section 6.7.5). It is an S-estimator with bisquare $\rho$ function that uses as initial value of the iterative algorithm, an MVE estimator. The MVE estimator is computed by subsampling with concentration step. The number of subsample in MVE is 500. Once the estimator of location and covariance corresponding to one subsample are computed, the concentration step consists in computing the sample mean and sample covariance of the [n/2] observations with smallest Mahalanobis distance. MVE-S is implemented in the `R` package `rrcov`, function `CovSest`, option `method="bisquare"` (Todorov & Filzmoser, 2009);

(c) Rocke-S (or shortened to Rocke), the estimator recently promoted by Maronna & Yohai (2015). It is an S-estimator with a non-monotonic weight function (Rocke, 1996). The only difference to the original proposal is that the estimator uses the KSD estimator (Peña & Prieto, 2001) as initial value for the iterative algorithm. The KSD estimator is computed by finding directions that maximize or minimize the kurtosis of the respective projections, as well as random "specific" directions aimed at detecting casewise outliers. The KSD estimator is implemented in a `MATLAB` code kindly provided by the author. The initial estimate can then be used to calculate Rocke-S, which is implemented in the `R` package `rrcov`, function `CovSest`, option `method="rocke"`.

(d) HSD, Stahel–Donoho estimator with Huberized outlyingness proposed by Van Aelst et al. (2012). We use a `MATLAB` code kindly provided by the authors. The number of subsamples used in HSD is $200p$;

(e) SnipEM (or shortened to Snip), the procedure proposed in Farcomeni (2014a). This method requires an initial specification of the position of the snipped cells in the form of a binary data table. We compared (using simulation) several possible choices for this initial set including: (a) snipping the largest 10% of the absolute standardized values for each variable; (b) snipping the largest 15% of the absolute standardized values for each variable; and (c) snipping the standardized values that are more than 1.5 times the interquartile range less the first quartile or more than 1.5 times the interquartile range plus the third quartile, for each variable. We only report the results from case (b) as it yields the best performances. SnipEM is implemented in the `R` package `snipEM`, function `snipEM`, default option (Farcomeni & Leung, 2014);

(f) DetMCDScore (or shortened to DMCDSc), the procedure proposed in the comment by Rousseeuw & Van den Bossche (2015). The DetMCDScore is calculated by applying deterministic MCD (DetMCD) on the normal scores (copula) of the data. A similar approach was also proposed in Öllerer & Croux (2015). The DetMCDScore is very computationally efficient and has been shown to deal with cellwise and casewise outliers adequately. DetMCD is implemented in the `R` package `DetMCD`, function `DetMCD`, default option (Kaveh, 2015);

(g) UF-GSE and UBF-GSE, the proposed two-step approach. The first step applies either the univariate filter only (shortened to UF) or the combination of univariate and bivariate filter (shortened to UBF) to the data. The second step then calculates GSE for the incomplete data, starting from the EMVE estimator. The EMVE estimator is computed by subsampling with concentration step. The number of subsamples used in EMVE is 500. The two-step procedure is available as the `TSGS` function, option `alpha=c(0.95, 0)` (UF) and option `alpha=c(0.95, 0.85)` (UBF), in the `R` package `GSE` (Leung et al., 2015).

The tuning parameters for the high breakdown point estimators MVE-S, Rocke-S, and MCD are chosen to attain 0.5 breakdown point under THCM.

We consider clean and contaminated samples from a $N_p(\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0})$ distribution with dimension $p = 5, 10, 15, 20$ and sample size $n = 10p$. We have also considered $n = 5p$, but the results are generally similar and are provided in Section A.1 in the Appendix. The simulation mechanisms are described below.

Since the contamination models and the estimators considered in our simulation study are location and scale equivariant, we can assume without loss of generality that the mean, $\boldsymbol{\mu}_0$, is equal to $\mathbf{0}$ and the variances in $\text{diag}(\boldsymbol{\Sigma}_0)$ are all equal to $\mathbf{1}$. That is, $\boldsymbol{\Sigma}_0$ is a correlation matrix.

Since the cellwise contamination model and the estimators are not affine-equivariant, we consider the two different approaches to introduce correlation structures: random correlation and first order autoregressive correlation (AR1).

### Random Correlation

For each sample in our simulation, we create a different random correlation matrix with condition number, which is defined as the largest eigenvalue of a correlation matrix divided by the smallest eigenvalue, fixed at $CN = 100$. Correlation matrices with high condition number are generally less favorable for non-affine equivariant estimators as extensively explored by Alqallaf (2003) and Danilov (2010). We use the following procedure to obtain random correlations with a fixed condition number $CN$:

1. For a fixed condition number $CN$, we first obtain a diagonal matrix $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_p)$, $[\lambda_1 < \lambda_2 < \cdots < \lambda_p]$ with smallest eigenvalue $\lambda_1 = 1$ and largest eigenvalue $\lambda_p = \text{CN}$. The remaining eigenvalues $\lambda_2, \ldots, \lambda_{p-1}$ are $p - 2$ sorted independent random variables with a uniform distribution in the interval $(1, \text{CN})$.

2. We first generate a random $p \times p$ matrix $\boldsymbol{Y}$, which elements are independent standard normal random variables. Then, we form the symmetric matrix $\boldsymbol{Y}^t\boldsymbol{Y} = \boldsymbol{QVQ}^t$ to obtain a random orthogonal matrix $\boldsymbol{Q}$ via eigendecomposition.

3. Using the results of 1 and 2 above, we construct the random covariance matrix by $\boldsymbol{\Sigma}_0 = \boldsymbol{Q\Lambda Q^t}$. Notice that the condition number of $\boldsymbol{\Sigma}_0$ is equal to the desired $CN$.

4. Convert the covariance matrix $\boldsymbol{\Sigma}_0$ into the correlation matrix $\boldsymbol{R}_0$ as follows:

$$\boldsymbol{R}_0 = \boldsymbol{W}^{-1/2}\boldsymbol{\Sigma}_0\boldsymbol{W}^{-1/2}$$

   where

$$\boldsymbol{W} = \text{diag}(\sigma_1, \ldots, \sigma_p)$$

   and $\sigma_1, \ldots, \sigma_p$ are the standard deviations in the covariance matrix $\boldsymbol{\Sigma}_0$.

5. After the conversion to correlation matrix in Step 4 above, the condition number of $\boldsymbol{R}_0$ is no longer necessarily equal to $CN$. To remedy this problem, we

consider the eigenvalue diagonalization of $\boldsymbol{R}_0$

$$\boldsymbol{R}_0 = \boldsymbol{Q_0 \Lambda_0 Q_0^t}. \tag{2.15}$$

where

$$\boldsymbol{\Lambda}_0 = \text{diag}(\lambda_1^{R_0}, \ldots, \lambda_p^{R_0}), \qquad \lambda_1^{R_0} < \lambda_2^{R_0} < \cdots < \lambda_p^{R_0}.$$

is the diagonal matrix formed using the eigenvalues of $\boldsymbol{R}_0$. We now re-establish the desired condition number $CN$ by redefining

$$\lambda_p^{R_0} = \text{CN} \times \lambda_1^{R_0}$$

and using the modified eigenvalues in (2.15).

6. Repeat 4 and 5 until the condition number of $\boldsymbol{R}_0$ is within a tolerance level (or until we reach some maximum iterations). In our simulation study, we set the tolerance for the difference in $CN$ at $10^{-5}$ and the maximum iterations to be 100. However, convergence was reached after a few iteration in all the cases.

### First Order Autoregressive Correlation

The random correlation structure generally has small correlations, especially with increasing $p$. For example, for $p = 10$, the maximum correlation values have an average of 0.49, and for $p = 50$, the average maximum is 0.28. So, we consider also a different correlation structure with higher correlations, in which the correlation matrix has entries

$$\Sigma_{0,jk} = \rho^{|j-k|},$$

with $\rho = 0.9$. This correlation is also known as the first order autoregressive correlation (AR1).

### Contamination Scenarios

We then consider the following scenarios:

- Clean data: No further changes are done to the data.
- Cellwise contamination: We randomly replace a $\epsilon$ of the cells in the data matrix by $X_{ij}^{cont} \sim N(k, 0.1^2)$, where $k = 1, 2, \ldots, 10$.
- Casewise contamination: We randomly replace a $\epsilon$ of the cases in the data matrix by $\boldsymbol{X}_i^{cont} \sim 0.5N(c\boldsymbol{v}, 0.1^2\boldsymbol{I}) + 0.5N(-c\boldsymbol{v}, 0.1^2\boldsymbol{I})$, where $c = \sqrt{k(\chi^2)_p^{-1}(0.99)}$ and $k = 1, 2, \ldots, 10$ and $\boldsymbol{v}$ is the eigenvector corresponding to the smallest eigenvalue of $\boldsymbol{\Sigma}_0$ with length such that $(\boldsymbol{v} - \boldsymbol{\mu}_0)^t \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{v} - \boldsymbol{\mu}_0) = 1$. Experiments show that the placement of outliers in this way is the least favorable for the proposed estimator.

We consider $\epsilon = 0.05, 0.10$ for cellwise contamination, and $\epsilon = 0.10, 0.20$ for casewise contamination. The number of replicates in our simulation study is $N = 500$.

The performance of a given scatter estimator $\boldsymbol{\Sigma}_n$ is measured by the Kullback–Leibler divergence between two Gaussian distribution with the same mean and covariances $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_0$:

$$D(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_0) = \text{trace}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{-1}) - \log(|\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{-1}|) - p.$$

This divergence also appears in the likelihood ratio test statistics for testing the null hypothesis that a multivariate normal distribution has covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$. We call this divergence measure the likelihood ratio test distance (LRT). Then, the performance of an estimator $\boldsymbol{\Sigma}_n$ is summarized by

$$\overline{D}(\boldsymbol{\Sigma}_n, \boldsymbol{\Sigma}_0) = \frac{1}{N} \sum_{i=1}^{N} D(\hat{\boldsymbol{\Sigma}}_{n,i}, \boldsymbol{\Sigma}_0)$$

where $\hat{\boldsymbol{\Sigma}}_{n,i}$ is the estimate at the $i$-th replication. Finally, the maximum average LRT distances over all considered contamination values, $k$, is also calculated.

Table 2.1 shows the maximum average LRT distances among the considered contamination sizes for the cellwise contamination setting for $n = 10p$. UBF-GSE and UF-GSE perform similarly when correlations are small because the bivariate filter

Table 2.1: Maximum average LRT distances under cellwise contamination. The sample size is $n = 10p$.

| Corr. | $p$ | $\epsilon$ | MCD | MVE-S | Rocke | HSD | Snip | DMCDSc | UF-GSE | UBF-GSE |
|---|---|---|---|---|---|---|---|---|---|---|
| Random | 5 | 0.05 | 1.2 | 0.6 | 0.8 | 1.7 | 10.0 | 1.7 | 0.8 | 1.0 |
| | | 0.10 | 2.8 | 6.6 | 19.1 | 8.5 | 57.8 | 6.8 | 3.8 | 3.7 |
| | 10 | 0.05 | 2.6 | 11.1 | 3.6 | 11.4 | 7.6 | 3.7 | 4.7 | 4.6 |
| | | 0.10 | 99.1 | 150.0 | 260.6 | 61.5 | 11.0 | 15.2 | 16.3 | 16.0 |
| | 15 | 0.05 | 21.6 | 65.2 | 46.3 | 31.8 | 12.0 | 6.7 | 8.5 | 8.4 |
| | | 0.10 | 168.2 | 198.2 | 202.8 | 155.0 | 15.4 | 21.3 | 21.0 | 21.1 |
| | 20 | 0.05 | 53.6 | 99.7 | 190.4 | 58.5 | 15.5 | 9.6 | 11.1 | 11.3 |
| | | 0.10 | 216.3 | 240.0 | 737.5 | 253.1 | 18.6 | 25.8 | 24.3 | 24.4 |
| AR1(0.9) | 5 | 0.05 | 1.1 | 0.6 | 0.8 | 0.6 | 7.7 | 1.1 | 0.7 | 0.9 |
| | | 0.10 | 4.0 | 9.7 | 21.2 | 1.8 | 33.9 | 3.9 | 2.1 | 1.6 |
| | 10 | 0.05 | 2.3 | 13.8 | 3.8 | 2.8 | 7.2 | 3.4 | 2.1 | 1.2 |
| | | 0.10 | 166.9 | 205.9 | 629.0 | 20.6 | 14.8 | 14.1 | 11.0 | 2.7 |
| | 15 | 0.05 | 60.8 | 104.4 | 111.3 | 12.3 | 9.7 | 7.8 | 5.1 | 1.8 |
| | | 0.10 | 328.7 | 381.2 | 412.8 | 103.0 | 14.3 | 26.0 | 21.6 | 6.6 |
| | 20 | 0.05 | 140.1 | 208.3 | 690.4 | 31.4 | 14.4 | 12.9 | 9.3 | 2.7 |
| | | 0.10 | 479.3 | 526.9 | 1677.4 | 274.0 | 20.4 | 38.1 | 34.1 | 14.5 |

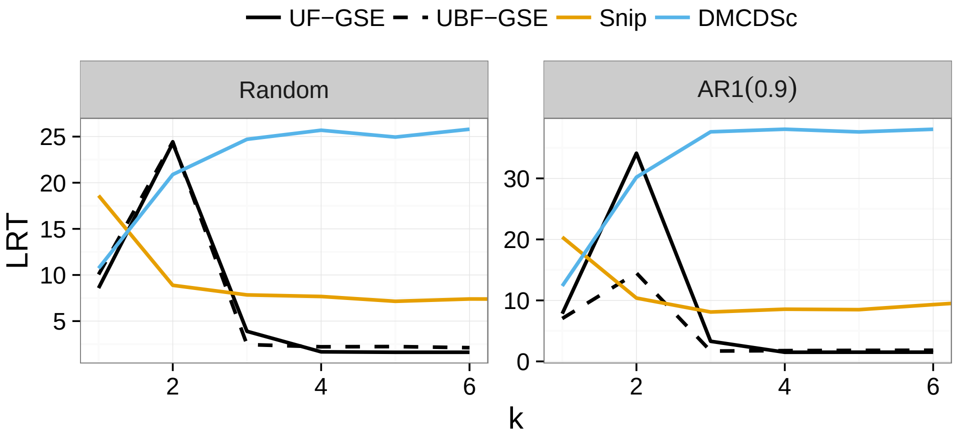


Figure 2.1: Average LRT distances for various contamination values, $k$, under 10% cellwise contamination. The dimension is $p = 20$ and sample size is $n = 10p$.

Table 2.2: Maximum average LRT distances under casewise contamination. The sample size is $n = 10p$.

| Corr. | $p$ | $\epsilon$ | MCD | MVE-S | Rocke | HSD | Snip | DMCDSc | UF-GSE | UBF-GSE |
|---|---|---|---|---|---|---|---|---|---|---|
| Random | 5 | 0.10 | 1.5 | 0.9 | 3.5 | 1.3 | 10.6 | 1.9 | 2.4 | 3.8 |
| | | 0.20 | 18.3 | 6.4 | 27.9 | 5.2 | 33.3 | 16.5 | 20.0 | 31.3 |
| | 10 | 0.10 | 7.7 | 4.2 | 8.9 | 3.9 | 32.7 | 5.1 | 9.9 | 18.5 |
| | | 0.20 | 113.7 | 38.2 | 4.3 | 21.9 | 62.3 | 61.2 | 80.7 | 97.2 |
| | 15 | 0.10 | 31.2 | 7.4 | 5.3 | 8.2 | 48.5 | 16.4 | 19.0 | 32.5 |
| | | 0.20 | 145.6 | 102.7 | 4.4 | 49.1 | 85.4 | 81.4 | 116.2 | 135.3 |
| | 20 | 0.10 | 64.8 | 10.4 | 5.0 | 13.5 | 61.3 | 37.8 | 30.3 | 52.4 |
| | | 0.20 | 174.6 | 142.6 | 4.7 | 92.4 | 107.4 | 99.9 | 148.4 | 175.3 |
| AR1(0.9) | 5 | 0.10 | 1.3 | 1.0 | 1.8 | 0.9 | 7.0 | 1.4 | 1.2 | 1.5 |
| | | 0.20 | 12.2 | 6.3 | 32.4 | 2.5 | 18.3 | 5.2 | 7.2 | 7.8 |
| | 10 | 0.10 | 5.8 | 3.5 | 4.0 | 1.7 | 20.2 | 2.9 | 3.8 | 4.4 |
| | | 0.20 | 101.6 | 37.8 | 15.7 | 8.8 | 45.6 | 31.9 | 52.5 | 52.3 |
| | 15 | 0.10 | 21.9 | 6.9 | 3.6 | 3.0 | 29.9 | 6.1 | 7.3 | 8.0 |
| | | 0.20 | 133.8 | 99.6 | 14.9 | 17.3 | 68.3 | 58.7 | 100.7 | 102.5 |
| | 20 | 0.10 | 61.2 | 9.6 | 3.2 | 4.4 | 42.9 | 15.7 | 13.5 | 15.4 |
| | | 0.20 | 165.6 | 128.9 | 16.0 | 32.7 | 85.6 | 85.9 | 129.9 | 132.9 |

is not sufficient enough to filter moderate cellwise outliers (e.g., $k = 2$). However, UBF-GSE outperforms UF-GSE when correlations are high because the bivariate filter can filter moderate cellwise outliers. See, for example, Figure 2.1 that shows the average LRT distance behaviors for different contamination sizes, $k$, for $p = 20$.

Table 2.2 shows the maximum average LRT distances among the considered contamination sizes for the casewise contamination setting for $n = 10p$. UF-GSE and UBF-GSE have an acceptable performance for moderate dimensions (e.g., $p \leq 10$), comparable with that of MVE-S, but neither UF-GSE, nor UBF-GSE, nor MVE-S perform very well for higher dimensions (e.g., $p \geq 15$) and unsatisfactorily for higher contamination level (see Section 2.6 for further discussion).

Table 2.3 shows the finite sample relative efficiency under clean samples with AR1(0.9) correlation for the considered robust estimates, taking the MLE average LRT distances as the baseline. The results for the random correlation are very similar and not shown here. UF-GSE and UBF-GSE, like MVE-S, shows increasing

Table 2.3: Finite sample efficiency for first order autoregressive correlations, AR1($\rho$), with $\rho = 0.9$. The sample size is $n = 10p$.

| $p$ | MCD | MVE-S | Rocke | HSD | Snip | DMCDSc | UF-GSE | UBF-GSE |
|---|---|---|---|---|---|---|---|---|
| 5 | 0.40 | 0.74 | 0.45 | 0.40 | 0.15 | 0.34 | 0.63 | 0.52 |
| 10 | 0.51 | 0.90 | 0.50 | 0.70 | 0.26 | 0.43 | 0.82 | 0.71 |
| 15 | 0.62 | 0.94 | 0.54 | 0.85 | 0.13 | 0.50 | 0.87 | 0.79 |
| 20 | 0.66 | 0.96 | 0.56 | 0.90 | 0.14 | 0.55 | 0.91 | 0.85 |

Table 2.4: Average "CPU time" – in seconds of a 2.8 GHz Intel Xeon – evaluated using the `R` command, `system.time`. The sample size is $n = 10p$.

| $p$ | UF-GSE | UBF-GSE |
|---|---|---|
| 10 | 0.7 | 1.1 |
| 20 | 7.7 | 11.0 |
| 30 | 34.5 | 45.6 |
| 40 | 120.5 | 144.9 |
| 50 | 278.4 | 338.0 |

efficiency when increasing $p$. Results for larger sample sizes, not reported here, show an identical pattern, except for MCD which efficiency increases with the sample size.

The computing times for our estimator for various dimensions and $n = 10p$ are averaged over all replications and are shown in Table 2.4. Comparatively longer computing times for the two-step procedures arise for higher dimensions because GSE becomes more computationally intensive for higher dimensions and for higher fractions of cases affected by filtered cells (see Section 2.6 for further discussion).

## 2.5 Real data example: geochemical data revisit

In Chapter 1, we presented the geochemical data from Smith et al. (1984) and highlighted the presence of cellwise and casewise outliers in these data. We showed there that traditional robust procedures provide poor estimates and fail to identify real outliers. In this section, we revisit this example. Our purpose here is twofold: first, to show that the two-step procedures can provide good estimates and identify outliers

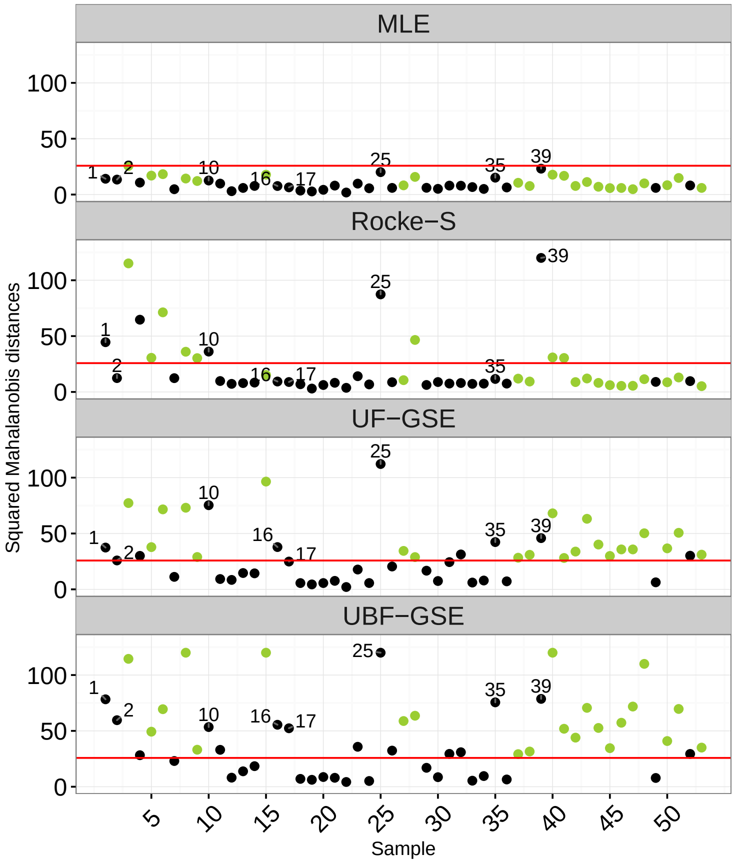


Figure 2.2: Squared Mahalanobis distances of the samples in the geochemical data based on different estimates. Large distances are truncated for better visualization. Samples that contain one or more flagged components (large cellwise outliers) are in green.

that were missed by traditional procedures; and second, to show that the bivariate filter version (UBF-GSE) can also identify moderate cellwise outliers, and as such, yield further improved results.

As mentioned in Chapter 1, the geochemical data give the content measure for 10 chemical compounds in 53 samples of rocks. A log transformation was applied to the data to make them more symmetric. From the quantile–quantile plots the data appear fairly normal for the most part, but outliers are observed.

We now compute the UF-GSE and the UBF-GSE estimates for these data, as well

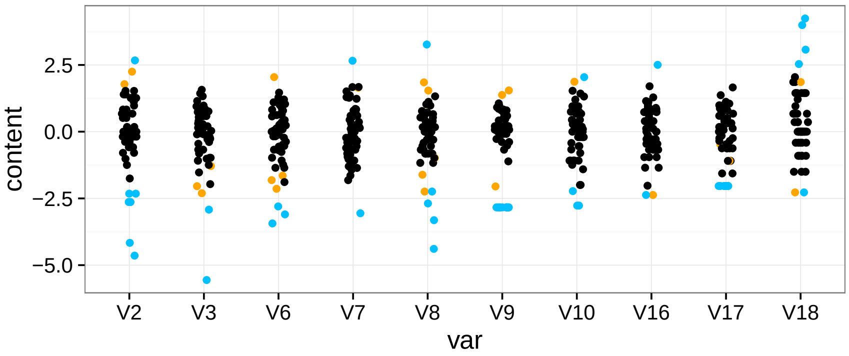


Figure 2.3: Univariate scatterplots of the 10 components in the geochemical data. Each component is standardized by its median and MAD. Points are randomly jittered to avoid much overlapping. The points flagged by the univariate filter are in blue, and those additionally flagged by the bivariate filter are in orange.

as the Rocke-S estimates for comparison. Figure 2.2 shows the squared Mahalanobis distances (MD) of the 53 samples based on the different estimates. Samples that contain at least one component that lies more than three MAD away from the median (i.e., the large cellwise outliers) are shown in green in the figure. There are in total 41.5% of samples flagged as such. Notice that all the cases with large cellwise outliers are also flagged MD outliers by UBF-GSE and UF-GSE. But this is not the case for Rocke-S. This is so because UBF-GSE and UF-GSE makes a more efficient use of the clean part of cases affected by cellwise outliers. Notice that Rocke-S must assign a final weight to each case by looking at the whole case, even in situations when there is only a single outlying component. As a result, for these data, Rocke-S fails to provide a good fit and produces unreliable Mahalanobis distances and final weights. In addition to the samples with large cellwise outliers, eight new cases (samples 1, 2, 10, 16, 17, 25, 35, 39) are flagged as outliers by UBF-GSE, with estimated full Mahalanobis distances exceeding the 99.99% chi-square cutoff. UF-GSE also flags most of these cases, but misses samples 2 and 17.

UF-GSE and UBF-GSE are both equally resistant against large cellwise outliers but UF-GSE is less resistant against moderate cellwise outliers, which are present in these data. Figure 2.3 depicts univariate scatterplots for each component (stan-

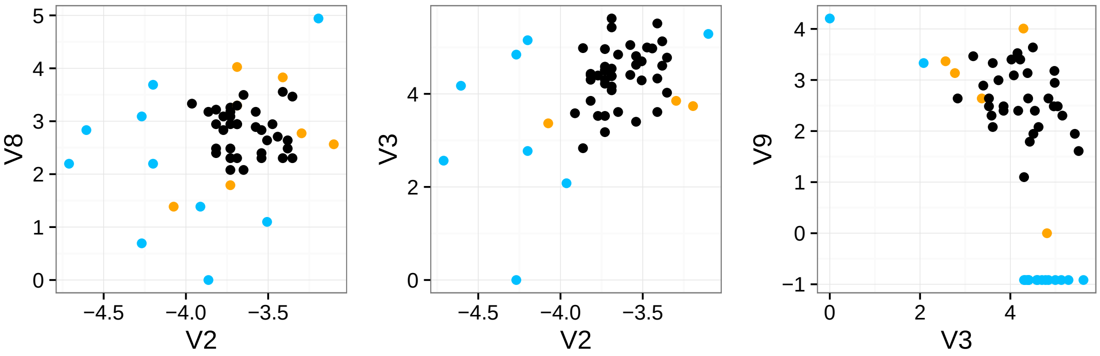


Figure 2.4: Pairwise scatterplots of the geochemical data for $V2$ versus $V8$, $V2$ versus $V3$, and $V3$ versus $V9$. Points with components flagged by the univariate filter are in blue. Points with components additionally flagged by the bivariate filter are in orange.

dardized and with random jitter) in the data. The points flagged by the univariate filter are in blue, and those flagged by the bivariate filter are in orange. Additionally, in Figure 2.4 bivariate scatterplots are shown for $V2$ versus $V8$, $V2$ versus $V3$, and $V3$ versus $V9$, where some correlations are observed. From these figures, we see that the bivariate filter has identified some additional cellwise outliers that are not-so-large marginally but become more visible when viewed together with other correlated components (the orange points in Figure 2.3). These moderate cellwise outliers account for 5.1% of the cells in the dataset and propagate to 28.3% of the cases. The final median weight assigned to these cases by UF-GSE and UBF-GSE are 0.36 and 0.70, respectively. By filtering the mild outliers UBF-GSE is able to make a more efficient use of the clean components in 28.3% of the cases.

## 2.6 Discussion

In the results section, we have found that the two-step approach performs overall the best under ICM, but not so well under THCM for $p > 10$. The computing times of the two-step procedures for higher dimensions are rather long, making the procedures less appealing for real time use. These points were also raised by the discussants in the published paper connected with this chapter, along with a remark on handling large and flat data sets. For the rest of the section, we discuss these main points individually.

### 2.6.1 Controlling the robustness and efficiency for large $p$

Maronna (2015) made a thoughtful remark regarding the loss of robustness of UF-GSE–and in general, S-estimators with a fixed loss function $\rho$–when $p$ is large. The Gaussian efficiency of UF-GSE, as well as UBF-GSE, systematically increases to one as $p$ increases, but this gain in efficiency comes at the expense of a decrease in robustness. Hence, we agree that for large $p$ we need to modify the GSE step to avoid the lack of robustness of S-estimators with fixed loss function. A possibility could be to use a well-calibrated MM-estimator of multivariate location and scatter (Tatsuoka & Tyler, 2000) after adapting it for handling data with missing values. The resulting generalized MM-estimator would then gain robustness for $p$ large. Another possibility could be to replace the bisquare rho function in GSE by a Rocke-type loss function, which changes with dimension in order to preserve the robustness of the estimator (Rocke, 1996). A comparison of MM-estimators and Rocke type estimators with S-estimators based on a bisquare rho-function for complete data and casewise contamination can be found in Maronna & Yohai (2015). Further work on this topic can be found in Chapter 3.

### 2.6.2 Computational issue

Several authors (Croux & Öllerer, 2015; Maronna, 2015; Rousseeuw & Van den Bossche, 2015; Van Aelst, 2015; Welsch, 2015) commented on the high computational cost of the two-step procedures and the need for faster alternatives.

The first step of the two-step procedure (filter) is fast, but the second step is slow due to the computation of the generalized S-estimator (GSE) (Danilov et al., 2012). We notice that GSE first resamples the filtered data to compute an initial estimate and then iterates until convergence a sequence of robust imputation and estimation steps. These iterations can be computationally intensive and time consuming when a large fraction of the data has been filtered. However, the main computational burden in GSE comes from the computation of the initial robust estimator (extended minimum volume ellipsoid, EMVE) which is needed to achieve high robustness against casewise outliers.

EMVE introduced by Danilov et al. (2012) is a generalized version for incomplete data of the MVE (Rousseeuw, 1985). The computation of EMVE consists of a combination of subsampling and concentration steps. Once the estimators of location and scatter for a given subsample are obtained, the concentration step consists of computing the Gaussian MLE via the classical EM algorithm on the half of the observations with the smallest Mahalanobis distances. The concentration steps are time consuming, especially when there is a large number of filtered cells. This problem is aggravated by the required large number of subsamples, especially when $p$ is large. Therefore, there is a need for finding a fast and fully robust initial estimate for GSE. Further investigation on choices of initial estimator are done in Chapter 3.

### 2.6.3 Remarks on handling large and flat data sets

Croux & Öllerer (2015) gave an extensive and detailed discussion of the performance of UF-GSE in the case of large and flat data sets (large $p$ and relatively small $n$).

Our numerical experiments confirm that the two-step procedure does not handle well large and flat data sets (e.g., $n < 5p$). In fact, when $n \leq 2p$, the generalized S-estimator fails to exist, likewise S-estimator and all classical robust estimators with breakdown point 1/2. When much data are filtered and the fraction of complete data is small, the iterations in GSE may fail to converge. In this case, GSE produces a nearly singular covariance matrix. This situation is more likely to occur for datasets with relatively small $n$ compared to $p$. Danilov et al. (2012) provided a sufficient condition for the existence of GSE: the proportion of complete observations in general position to be larger than $1/2 + (p + 1)/n$. Numerical results have shown that GSE performs well for some smaller proportions of complete observations. However, no theoretical results are available for these cases.

To overcome the lack of convergence of the two-step procedure for large and flat data sets, we may partially impute the filtered cells to ensure a fraction $1/2 + (p+1)/n$ of complete observations. More precisely, the procedure is to first filter outliers, then randomly select observations and impute the filtered cells using coordinate-wise medians, and finally estimate the location and scatter using GSE. Although this procedure is rather ad hoc, our initial numerical experiments suggests that it may work for $n \geq 5p$.

## 2.7 Conclusions

Affine equivariance, a proven asset for achieving THCM robustness, becomes a hindrance under ICM because of outliers propagation.

We advocate the practical and theoretical importance of ICM and point to the perils and drawbacks of relying solely on the THCM paradigm. ICM promotes a less aggressive cellwise down-weighting of outliers and becomes an essential tool for modeling contamination in moderate and high dimensional data.

We introduce a non-affine equivariant, two-step procedure to achieve robustness under ICM and THCM. The first step applies a filter to the data, aim to reduce the impact of outliers propagation and to overcome the curse of dimensionality posed by ICM. The second step then applies the generalized S-estimator of Danilov et al. (2012) to the incomplete data from the first step, aiming to achieve robustness under THCM. A univariate filtering and a combination of univariate and bivariate filtering are proposed in the first step, resulting in two versions of the two-step procedure: UF-GSE and UBF-GSE.

The two-step procedures (UF-GSE and UBF-GSE) exhibits high robustness against large cellwise outliers from ICM, but UF-GSE is not resistant against moderate cellwise outliers. In this case, UBF-GSE exhibits higher robustness than UF-GSE when the correlations between the uncontaminated variables are high. However, this gain in robustness comes at the expense of a decrease in robustness under THCM. Therefore, we recommend UBF-GSE when the correlations are seemingly high, but UF-GSE otherwise. Overall, the two-step procedures exhibits satisfactory robustness against casewise outliers from THCM for low to moderate dimensional data (e.g.,

$p \leq 10$), but starts losing robustness with increasing $p$ (e.g., $p > 10$).

Finally, UF-GSE and UBF-GSE both are not yet capable of handling high dimensional data for the following reasons. (1) The high computational cost of the current initial estimator (EMVE) required by the procedure, making it infeasible for clock-time computation. (2) The generalized S-estimators employed in the second step is incapable of dealing with casewise outliers when $p$ is large. Further work on these two topics are presented in the next chapter.

# Chapter 3

# Robust Estimation of Multivariate Location and Scatter in the Presence of Cellwise and Casewise Contamination in High Dimension

## 3.1 Introduction

In this chapter, we continue our work on the problem of robust estimation of multivariate location and scatter matrix under cellwise and casewise contamination.

Most traditional robust estimators assume that the majority of cases is totally free of contamination. Any case that deviates from the model distribution is fully flagged as an outlier. In situations that only a small number of components of a case are contaminated, down-weighting the whole case may not be appropriate and can cause a huge loss of information, especially when the dimension is large. When data contain both cellwise and casewise outliers, the problem becomes even more difficult. For this, we proposed two-step procedures called UF-GSE and UBF-GSE in Chapter 2. The first step applies either a univariate filter (UF) or a combination of univariate and bivariate filter (UBF) to the data matrix $\mathbb{X}$ and sets the flagged cells to missing values, NA's. The second step then applies the generalized S-estimator (GSE) of Danilov et al. (2012) to this incomplete data set. The two-step procedures are shown to be simultaneously robust against cellwise and casewise outliers for low dimensional data (e.g., $p \leq 10$). Unfortunately, the procedures do not scale well for higher dimensions due to the high computational cost of its initial estimator EMVE.

Furthermore, the procedures loses robustness against casewise outliers for higher dimensional data (e.g., $p > 10$).

One goal of this chapter is to improve the robustness of UF-GSE and UBF-GSE in high dimension. For that, we introduce a new robust estimator called *Generalized Rocke S-estimator* or *GRE* to replace GSE in the second step. The resulting procedures are called *UF-GRE* and *UBF-GRE*. In his discussion of Agostinelli et al. (2015), Maronna (2015) made a thoughtful remark regarding the loss of robustness of UF-GSE–and in general, S-estimators with a fixed loss function $\rho$–when $p$ is large. S-estimators with a fixed $\rho$ uncontrollably gain efficiency and lose their robustness for large $p$ (Rocke, 1996). Such curse of dimensionality has also been confirmed for UF-GSE and UBF-GSE, which use a GSE with a fixed $\rho$ in its second step.

Another goal of this chapter is to reduce the high computational cost of the two-step approach in high dimension. The first step of filtering is generally fast, but the second step is slow due to the computation of the extended minimum volume ellipsoid (EMVE), used as initial estimate by the generalized S-estimators. Subsampling is the standard way to compute EMVE, but it requires an impractically large number of subsamples, making the initial estimation extremely slow. To address this computational issue, we introduce a new subsampling procedure based on clustering for computing EMVE. The new initial estimator is called EMVE-C.

The rest of the chapter is organized as follows. In Section 3.2, we introduce the GRE. In Section 3.3, we describe the computational issues of the proposed estimators regarding the initial estimation and introduce EMVE-C to serve this capacity. In Section 3.4 and 3.5, we compare the two-step approaches equipped with GSE and GRE through an extended simulation study of that in Chapter 2, as well as through a real data example. Finally, we conclude in Section 3.6. We also provide additional simulation results and other supplementary material in Appendix B.

## 3.2 Generalized Rocke S-estimators

Rocke (1996) showed that if the weight function $W(x) = \rho'(x)/x$ in S-estimators is non-increasing, the efficiency of the estimators tends to one when $p \to \infty$. However, this gain in efficiency is paid for by a decrease in robustness. Not surprisingly, the same phenomenon has been observed for generalized S-estimators in simulation studies. Therefore, there is a need for new generalized S-estimators with controllable efficiency/robustness trade off.

Rocke (1996) proposed that the $\rho$ function used to compute S-estimators should change with the dimension to prevent loss of robustness in higher dimensions. The Rocke-$\rho$ function is constructed based on the fact that for large $p$ the scaled squared Mahalanobis distances for normal data

$$\frac{D(\boldsymbol{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\sigma} \approx \frac{Z}{p} \quad \text{with} \quad Z \sim \chi_p^2,$$

and hence that $D/\sigma$ are increasingly concentrated around one. So, to have a high enough, but not too high, efficiency, we should give a high weight to the values of $D/\sigma$ near one and down-weight the cases where $D/\sigma$ is far from one.

Let

$$\gamma = \min\left(\frac{\chi^2(1-\alpha)}{p} - 1, 1\right), \tag{3.1}$$

where $\chi^2(\beta)$ is the $\beta$-quantile of $\chi_p^2$. In this chapter, we use a conventional choice of $\alpha = 0.05$ that gives a satisfactory efficiency of the estimator, but we have also explored smaller values of $\alpha$ (see Section B.1 in the Appendix). Maronna et al. (2006) proposed a modification of the Rocke-$\rho$ function, namely

$$\rho(u) = \begin{cases} 0 & \text{for} \quad 0 \leq u \leq 1-\gamma \\ \left(\frac{u-1}{4\gamma}\right)\left[3 - \left(\frac{u-1}{\gamma}\right)^2\right] + \frac{1}{2} & \text{for} \quad 1-\gamma < u < 1+\gamma \\ 1 & \text{for} \quad u \geq 1+\gamma \end{cases} \tag{3.2}$$

which has as derivative the desired weight function that vanishes for $u \notin [1-\gamma, 1+\gamma]$

$$W(u) = \frac{3}{4\gamma}\left[1 - \left(\frac{u-1}{\gamma}\right)^2\right] I(1-\gamma \leq u \leq 1+\gamma).$$

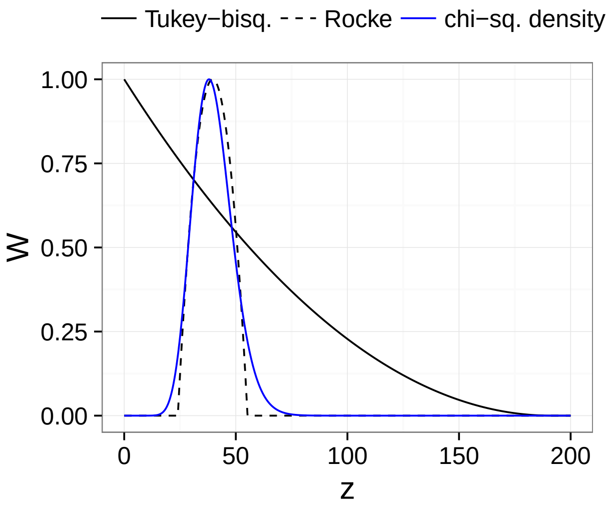


Figure 3.1: Weight functions of the Tukey bisquare and the Rocke for $p = 40$. Chi-square density functions are also plotted in blue for comparison. All the functions are scaled so that their maximum is 1 to facilitate comparison.

Figure 3.1 compares the Rocke-weight function, $W_{Rocke}(z/c_p)$, and the Tukey-bisquare weight function, $W_{Tukey}(z/c_p)$, for $p = 40$, where $c_p$ as defined in (2.10). The chi-square density function is also plotted in blue for comparison. When $p$ is large the tail of the Tukey-bisquare weight function greatly deviates from the tail of the chi-square density function and inappropriately assigns high weights to large distances. On the other hand, the Rocke-weight function can resemble the shape of the chi-square density function and is capable of assigning low weights to large distances.

Finally, we define the generalized Rocke S-estimators or GRE by (2.11) and (2.12) with the $\rho$-function in (2.9) replaced by the modified Rocke-$\rho$ function in (3.2). We compared GRE with GSE via simulation and found that GRE has a substantial better performance in dealing with casewise outliers when $p$ is large (e.g., $p > 10$). Results from this simulation study are provided in Section B.2 in the Appendix.

## 3.3 Computing issues

The generalized S-estimators described above are computed via iterative reweighted means and covariances, starting from an initial estimate. We now discuss some computing issues associated with this iterative procedure.

### 3.3.1 Initial estimator

For the initial estimate, the extended minimum volume ellipsoid (EMVE) has been used, as suggested by Danilov et al. (2012). The EMVE is computed with a large number of subsamples ($> 500$) to increase the chance that at least one clean subsample is obtained. Let $\varepsilon$ be the proportion of contamination in the data and $m$ be the subsample size. The probability of having at least one clean subsample of size $m$ out of $M$ subsamples is

$$q = 1 - \left[ 1 - \binom{n \cdot (1-\varepsilon)}{m} \Big/ \binom{n}{m} \right]^M . \tag{3.3}$$

For large $p$, the number of subsamples $M$ required for a large $q$, say $q = 0.99$, can be impractically large, dramatically slowing down the computation. For example, suppose $m = p$, $n = 10p$, and $\varepsilon = 0.50$. If $p = 10$, then $M = 7758$; if $p = 30$, then $M = 2.48 \times 10^{10}$; and if $p = 50$, then $M = 4.15 \times 10^{16}$. Therefore, there is a need for a faster and more reliable starting point for large $p$.

#### Cluster-Based Subsampling

Next, we introduce a cluster-based algorithm for faster and more reliable subsampling for the computation of EMVE. The EMVE computed with the cluster-based subsampling is called called EMVE-C throughout the chapter.

High-dimensional data have several interesting geometrical properties as described in Hall et al. (2005). One such property that motivated the Rocke-$\rho$ function, as well as the following algorithm, is that for large $p$ the $p$-variate standard normal distribution $N_p(\mathbf{0}, \boldsymbol{I})$ is concentrated "near" the spherical shell with radius $\sqrt{p}$. So, if outliers have a slightly different covariance structure from clean data, they would appear geometrically different. Therefore, we could apply a clustering algorithm to first separate the outliers from the clean data. Subsampling from a big cluster, which in principle is composed of mostly clean cases, should be more reliable and require fewer number of subsamples.

Given a data matrix $\mathbb{X}$, let $\mathbb{U}$ be the auxiliary matrix of zeros and ones, with zeros indicating the missing entries in $\mathbb{X}$. The following steps describe our clustering-based subsampling:

1. Standardize the data $\mathbb{X}$ with some initial location and dispersion estimator $T_{0j}$ and $S_{0j}$. Common choices for $T_{0j}$ and $S_{0j}$ that are also adopted in this chapter are the coordinate-wise median and median absolute deviance (MAD). Denote the standardized data by $\mathbb{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)^t$, where $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{ip})^t$ and $Z_{ij} = (X_{ij} - T_{0j})/S_{0j}$.

2. Compute a simple robust correlation matrix estimate $\boldsymbol{R} = (R_{jk})$. Here, we use the Gnanadesikan-Kettenring estimator (Gnanadesikan & Kettenring, 1972), where
$$R_{ij} = \frac{1}{4}(S_{0jk+}^2 - S_{0jk-}^2),$$
and where $S_{0jk+}$ is the dispersion estimate for $\{Z_{ij} + Z_{ik} | U_{ij} = 1, U_{ik} = 1\}$ and $S_{0jk-}$ the estimate for $\{Z_{ij} - Z_{ik} | U_{ij} = 1, U_{ik} = 1\}$. We use $Q_n$ (Rousseeuw & Croux, 1993) for the dispersion estimate.

3. Compute the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$ and eigenvectors $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p$ of the correlation matrix estimate
$$\boldsymbol{R} = \boldsymbol{E}\boldsymbol{\Lambda}\boldsymbol{E}^t,$$
where $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$ and $\boldsymbol{E} = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p)$. Let $p_+$ be the largest dimension such that $\lambda_j > 0$ for $j = 1, \ldots, p_+$. Retain only the eigenvectors $\boldsymbol{E}_0 = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{p_+})$ with a positive eigenvalue.

4. Complete the standardized data $\mathbb{Z}$ by replacing each missing entry, as indicated by $\mathbb{U}$, by zero. Then, project the data onto the basis eigenvectors $\tilde{\boldsymbol{Z}} = \boldsymbol{Z}\boldsymbol{E}_0$, and then standardize the columns of $\tilde{\boldsymbol{Z}}$, or so called principal components,

using coordinate-wise median and MAD of $\tilde{\boldsymbol{Z}}$.

5. Search for a "clean" cluster $C$ in the standardized $\tilde{\boldsymbol{Z}}$ using a hierarchical clustering framework by doing the following. First, compute the dissimilarity matrix for the principal components using the Euclidean metric. Then, apply classical hierarchical clustering (with any linkage of choice). A common choice is the Ward's linkage, which is adopted in this chapter. Finally, define the "clean" cluster by the smallest sub-cluster $C$ with a size at least $n/2$. This can be obtained by cutting the clustering tree at various heights from the top until all the clusters have size less than $n/2$.

6. Take a subsample of size $n_0$ from $C$.

With good clustering results, we can draw fewer subsamples, and equally important, we can use a larger subsample size. The current default choices in GSE are $M = 500$ subsamples of size $n_0 = (p+1)/(1-\alpha_{mis})$ as suggested in Danilov et al. (2012), where $\alpha_{mis}$ is the fraction of missing data ($\alpha_{mis}$ = number of missing entries $/(np)$). For the new clustering-based subsampling, we choose $M = 50$ and $n_0 = 2(p+1)/(1-\alpha_{mis})$ in this chapter, but other choices of $M$ and $n_0$ can be considered. However, we found that choosing a too large subsample size could result in contaminated subsamples with outliers that went undetected.

In principle, this procedure could be time-consuming because the number of operations required by hierarchical clustering is of order $n^3$. As an alternative, one may bypass the hierarchical clustering step and sample directly from the data points with the smallest Euclidean distances to the origin calculated from $\tilde{\boldsymbol{Z}}$. This is because the Euclidean distances, in principle, should approximate the Mahalanobis distances to the mean of the original data. However, our simulations show that the hierarchical clustering step is essential for the excellent performance of the estimates, and that this step entails only a small increase in computational time, even for $n = 1000$. For much larger $n$, when computational time becomes a serious concern, we can always perform the clustering procedure on a randomly chosen smaller fraction of the data to keep the computational speed, which should be sufficient for finding a reliable initial estimate.

### 3.3.2 Another computing issue

There is no formal proof that the recursive algorithm decreases the objective function at each iteration for the case of generalized S-estimators with a monotonic weight function (Danilov et al., 2012). This also the case for generalized S-estimators with a non-monotonic weight function. For Rocke estimators with complete data, Maronna et al. (2006, see Section 9.6.3) described an algorithm that ensures attaining a local minimum. We have adapted this algorithm for the generalized counterparts. Although we cannot provide a formal proof, we have seen so far in our experiments that the descending property of the recursive algorithms always holds.

## 3.4 Two-step estimation and extended simulations

The proposed two-step approach for global–robust estimation under cellwise and casewise contamination is to first flag outlying cells in the data table and to replace them by NA's using either univariate filtering only (shortened to UF) or univariate and bivariate filtering (shortened to UBF). In the second step, the generalized S-estimator is then applied to this incomplete data. Our new version of this is to replace GSE in the second step by GRE-C (i.e., GRE starting from EMVE-C). We call the new two-step procedure UF-GRE-C and UBF-GRE-C. The new procedures with GRE in the second step are available as the `TSGS` function, option `method="rocke"`, in the `R` package `GSE` (Leung et al., 2015).

We now conduct the same simulation study in Chapter 2 comparing UF-GRE-C and UBF-GRE-C with UF-GSE and UBF-GSE, as well as several other robust estimators that were shown to be as competitive under cellwise (Snip) or casewise contamination (Rocke and HSD) or both (DMCDSc). In addition, we consider higher dimensions, $p = 30, 40, 50$. To show the influence of cellwise outliers propagation on casewise robust estimates in higher dimensions, we also consider three levels of cellwise contamination, $\varepsilon = 0.02, 0.05, 0.10$. Finally, the number of replications in this simulation study is $N = 500$.

Table 3.1: Maximum average LRT distances under cellwise contamination. The sample size is $n = 10p$.

| Corr. | $p$ | $\epsilon$ | Rocke | HSD | Snip | DMCDSc | UF-GSE | UBF-GSE | UF-GRE-C | UBF-GRE-C |
|---|---|---|---|---|---|---|---|---|---|---|
| Random | 10 | 0.02 | 1.2 | 2.3 | 6.9 | 1.6 | 1.2 | 1.4 | 1.3 | 1.4 |
| | | 0.05 | 3.6 | 11.2 | 7.5 | 3.2 | 4.5 | 4.4 | 2.2 | 2.5 |
| | | 0.10 | 256.0 | 59.9 | 10.8 | 14.7 | 16.2 | 15.9 | 11.8 | 11.5 |
| | 20 | 0.02 | 2.7 | 10.6 | 13.9 | 2.6 | 4.0 | 4.4 | 2.9 | 3.0 |
| | | 0.05 | 187.2 | 57.1 | 15.5 | 9.3 | 11.0 | 11.1 | 8.0 | 8.2 |
| | | 0.10 | 725.8 | 251.5 | 18.4 | 24.7 | 24.0 | 24.1 | 21.5 | 21.6 |
| | 30 | 0.02 | 23.1 | 22.6 | 18.5 | 4.4 | 5.8 | 6.3 | 5.4 | 5.9 |
| | | 0.05 | 380.5 | 123.1 | 20.8 | 13.7 | 14.2 | 14.8 | 12.3 | 13.4 |
| | | 0.10 | 938.5 | 503.0 | 23.6 | 33.2 | 30.3 | 30.4 | 28.2 | 28.5 |
| | 40 | 0.02 | 121.3 | 38.9 | 22.6 | 6.0 | 7.3 | 8.0 | 9.4 | 10.9 |
| | | 0.05 | 584.1 | 212.4 | 25.8 | 17.9 | 16.6 | 17.4 | 18.4 | 19.9 |
| | | 0.10 | 1104.3 | 744.2 | 30.0 | 39.5 | 35.3 | 35.3 | 37.5 | 37.7 |
| | 50 | 0.02 | 192.8 | 58.7 | 27.1 | 8.1 | 9.1 | 10.0 | 12.5 | 12.9 |
| | | 0.05 | 618.1 | 298.7 | 29.7 | 20.7 | 19.6 | 20.6 | 22.7 | 23.6 |
| | | 0.10 | 1251.8 | 1002.2 | 32.0 | 46.7 | 43.1 | 43.2 | 45.7 | 46.3 |
| AR1(0.9) | 10 | 0.02 | 1.2 | 0.9 | 4.9 | 1.5 | 0.9 | 0.9 | 1.2 | 1.3 |
| | | 0.05 | 2.6 | 2.8 | 7.0 | 3.1 | 2.1 | 1.1 | 1.7 | 1.4 |
| | | 0.10 | 627.4 | 20.3 | 13.8 | 13.3 | 10.9 | 2.6 | 10.4 | 2.5 |
| | 20 | 0.02 | 2.5 | 3.9 | 10.5 | 2.6 | 2.1 | 1.5 | 2.2 | 2.1 |
| | | 0.05 | 690.6 | 31.3 | 14.3 | 12.3 | 9.3 | 2.7 | 7.6 | 2.8 |
| | | 0.10 | 1679.3 | 273.5 | 20.4 | 36.4 | 34.1 | 14.4 | 32.1 | 8.8 |
| | 30 | 0.02 | 71.1 | 10.7 | 13.9 | 5.4 | 4.0 | 2.3 | 3.9 | 3.4 |
| | | 0.05 | 1190.1 | 103.3 | 19.8 | 22.6 | 20.3 | 6.2 | 18.1 | 5.5 |
| | | 0.10 | 3440.3 | 916.9 | 29.0 | 63.4 | 59.7 | 40.5 | 56.0 | 27.7 |
| | 40 | 0.02 | 222.1 | 22.7 | 16.2 | 8.9 | 6.7 | 3.5 | 6.5 | 5.7 |
| | | 0.05 | 1785.5 | 259.9 | 23.7 | 34.8 | 31.4 | 14.0 | 29.7 | 12.4 |
| | | 0.10 | 5712.8 | 1966.5 | 35.9 | 90.4 | 84.9 | 74.4 | 81.5 | 52.9 |
| | 50 | 0.02 | 628.1 | 43.3 | 18.9 | 12.8 | 9.7 | 4.9 | 9.7 | 6.4 |
| | | 0.05 | 4271.7 | 534.5 | 28.9 | 46.5 | 42.8 | 22.6 | 40.8 | 20.4 |
| | | 0.10 | 4129.1 | 2998.6 | 44.7 | 119.5 | 111.8 | 104.6 | 106.3 | 75.3 |

Table 3.2: Maximum average LRT distances under casewise contamination. The sample size is $n = 10p$.

| Corr. | $p$ | $\epsilon$ | Rocke | HSD | Snip | DMCDSc | UF- | UBF- | UF- | UBF- |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | GSE | GSE | GRE-C | GRE-C |
| Random | 10 | 0.10 | 2.8 | 3.9 | 44.4 | 4.9 | 9.7 | 18.5 | 11.0 | 19.1 |
| | | 0.20 | 4.7 | 21.8 | 110.3 | 123.6 | 91.8 | 146.8 | 30.1 | 53.0 |
| | 20 | 0.10 | 3.4 | 13.4 | 76.9 | 37.8 | 29.7 | 50.1 | 11.5 | 20.9 |
| | | 0.20 | 5.6 | 95.9 | 166.5 | 187.6 | 291.8 | 311.4 | 22.0 | 49.3 |
| | 30 | 0.10 | 4.3 | 26.1 | 82.3 | 118.6 | 75.3 | 101.3 | 12.8 | 21.8 |
| | | 0.20 | 7.4 | 297.7 | 220.9 | 268.4 | 415.5 | 445.2 | 21.7 | 47.6 |
| | 40 | 0.10 | 5.2 | 46.3 | 101.6 | 130.6 | 140.2 | 168.8 | 18.6 | 29.5 |
| | | 0.20 | 9.1 | 547.4 | 186.2 | 340.1 | 534.1 | 579.9 | 22.7 | 52.3 |
| | 50 | 0.10 | 5.9 | 80.0 | 121.9 | 139.5 | 258.1 | 228.8 | 27.5 | 43.4 |
| | | 0.20 | 10.0 | 682.4 | 224.3 | 407.7 | 650.1 | 710.9 | 24.2 | 64.8 |
| AR1(0.9) | 10 | 0.10 | 2.8 | 1.7 | 20.2 | 2.9 | 3.7 | 4.3 | 3.1 | 3.6 |
| | | 0.20 | 4.8 | 8.7 | 49.7 | 29.7 | 50.8 | 50.1 | 7.2 | 8.4 |
| | 20 | 0.10 | 2.8 | 4.7 | 43.8 | 14.8 | 12.9 | 14.9 | 3.5 | 4.3 |
| | | 0.20 | 5.3 | 35.3 | 113.0 | 87.6 | 260.5 | 193.9 | 7.3 | 10.5 |
| | 30 | 0.10 | 3.4 | 8.9 | 66.1 | 32.2 | 31.3 | 37.7 | 4.1 | 5.1 |
| | | 0.20 | 8.2 | 155.5 | 144.8 | 122.9 | 372.7 | 365.1 | 8.4 | 13.3 |
| | 40 | 0.10 | 4.3 | 15.6 | 83.7 | 49.2 | 69.1 | 75.5 | 6.4 | 7.3 |
| | | 0.20 | 9.2 | 430.3 | 151.9 | 209.3 | 477.6 | 479.7 | 10.0 | 17.4 |
| | 50 | 0.10 | 5.1 | 26.5 | 103.3 | 64.4 | 148.2 | 160.1 | 7.6 | 8.1 |
| | | 0.20 | 11.1 | 538.3 | 188.5 | 276.0 | 581.6 | 585.0 | 11.0 | 21.2 |

We report the results for $n = 10p$ only since the results for $n = 5p$ are similar. Table 3.1 and Table 3.2 show the maximum average LRT distances under cellwise and casewise contamination, respectively. In general, UF-GSE and UBF-GSE perform similarly as UF-GRE-C and UBF-GRE-C, respectively, under cellwise contamination. However, UF-GRE-C and UBF-GRE-C substantially outperforms UF-GSE and UBF-GSE under casewise contamination. The Rocke $\rho$ function used in GRE in the second step is capable of giving smaller weights to points that are at moderate-to-large distances from the main mass of points; see, for example, Figure 3.2 that shows the average LRT distance behaviors of UBF-GSE and UBF-GRE-C for dimension $p = 30$ and AR1(0.9) correlated data under 10% casewise contamination.

Table 3.3 shows the finite sample relative efficiency under clean samples with

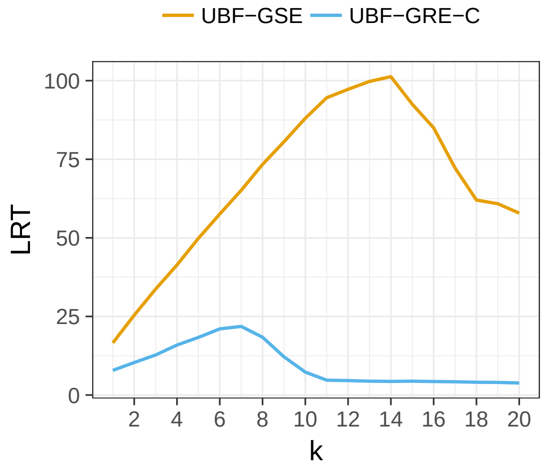


Figure 3.2: Average LRT distance behaviors of UBF-GSE and UBF-GRE-C for random correlations under 10% casewise contamination. The dimension is $p = 30$ and the sample size is $n = 10p$.

Table 3.3: Finite sample efficiency for random correlations. The sample size is $n = 10p$.

| $p$ | Rocke | HSD | Snip | DMCDSc | UF-GSE | UBF-GSE | UF-GRE-C | UBF-GRE-C |
|---|---|---|---|---|---|---|---|---|
| 10 | 0.50 | 0.73 | 0.12 | 0.41 | 0.75 | 0.66 | 0.53 | 0.48 |
| 20 | 0.57 | 0.92 | 0.09 | 0.56 | 0.83 | 0.73 | 0.59 | 0.55 |
| 30 | 0.58 | 0.93 | 0.10 | 0.63 | 0.87 | 0.79 | 0.49 | 0.44 |
| 40 | 0.60 | 0.94 | 0.10 | 0.68 | 0.89 | 0.83 | 0.39 | 0.36 |
| 50 | 0.60 | 0.94 | 0.11 | 0.70 | 0.91 | 0.84 | 0.48 | 0.49 |

random correlation for the considered robust estimates, taking the MLE average LRT distances as the baseline. The results for the AR1(0.9) correlation are very similar and not shown here. As expected, UF-GSE and UBF-GSE show an increasing efficiency as $p$ increases while UF-GRE-C and UBF-GRE-C have lower efficiency. Improvements can be achieved by using smaller $\alpha$ in the Rocke $\rho$ function with some trade-off in robustness. Results from this experiment are provided in Section B.1 in the Appendix.

Finally, we compare the computing times of the two-step procedures. Table

Table 3.4: Average "CPU time" – in seconds of a 2.8 GHz Intel Xeon – evaluated using the R command, `system.time`. The sample size is $n = 10p$.

| $p$ | UF-<br>GSE | UBF-<br>GSE | UF-<br>GRE-C | UBF-<br>GRE-C |
|---|---|---|---|---|
| 10 | 0.7 | 1.1 | 0.1 | 0.2 |
| 20 | 7.7 | 11.0 | 1.2 | 1.7 |
| 30 | 34.5 | 45.6 | 5.4 | 6.3 |
| 40 | 120.5 | 144.9 | 14.5 | 16.9 |
| 50 | 278.4 | 338.0 | 33.0 | 37.0 |

3.4 shows the average computing times over all contamination settings for various dimensions and for $n = 10p$. The computing times for the two-step procedure have been substantially improved with the implementation of the faster initial estimator, EMVE-C.

## 3.5 Real data example: small-cap stock returns data

In this section, we consider the weekly returns from 01/08/2008 to 12/28/2010 ($n = 730$ weeks) for a portfolio of 20 small-cap stocks ($p = 20$) from Martin (2013). The data set is different from the micro-cap stock returns data set in Chapter 1. It contains more correlated stock returns and contains relatively more moderate cellwise outliers as we will show next.

The purpose of this example is fourfold: first, to show that the classical MLE and traditional robust procedures perform poorly on data affected by propagation of cellwise outliers; second, to show that the two-step procedures can provide better estimates by filtering large outliers; third, that the bivariate-filter version of the two-step procedure provides even better estimates by flagging additional moderate cellwise outliers; and fourth, that UBF-GRE-C can more effectively down-weight some high-dimensional casewise outliers than UBF-GSE, for this 20-dimensional dataset. Therefore, UBF-GRE-C provides the best results for the dataset.

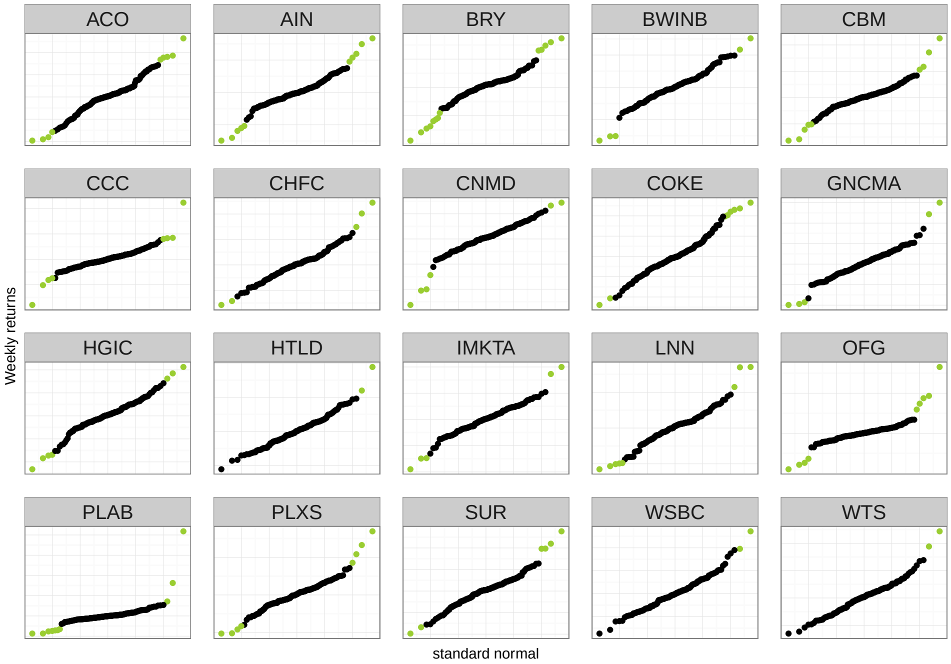


Figure 3.3: Normal quantile–quantile plots of weekly returns. Weekly returns that are three MAD away from the coordinatewise-median are shown in green.

Figure 3.3 shows the normal QQ-plots of the 20 small-cap stocks returns in the portfolio. The bulk of the returns in all stocks seem roughly normal, but large outliers are clearly present for most of these stocks. Stocks with returns lying more than three mads away from the coordinatewise-median (i.e., the large outliers) are shown in green in the figure. There is a total of 4.8% large cellwise outliers that propagate to 40.1% of the cases. Over 75% of these weeks correspond to the 2008 financial crisis.

First, we compute the MLE and the Rocke-S estimates, as well as the UF-GSE and the UBF-GSE estimates that were proposed in Chapter 2, for these data. Figure 3.4 shows the squared Mahalanobis distances of the 157 weekly observations based on the estimates. Weeks that contain large cellwise outliers (asset returns with values three MAD away from the coordinatewise-median) are in green. From the figure, we see that the MLE and the Rocke-S estimates have failed to identify many of

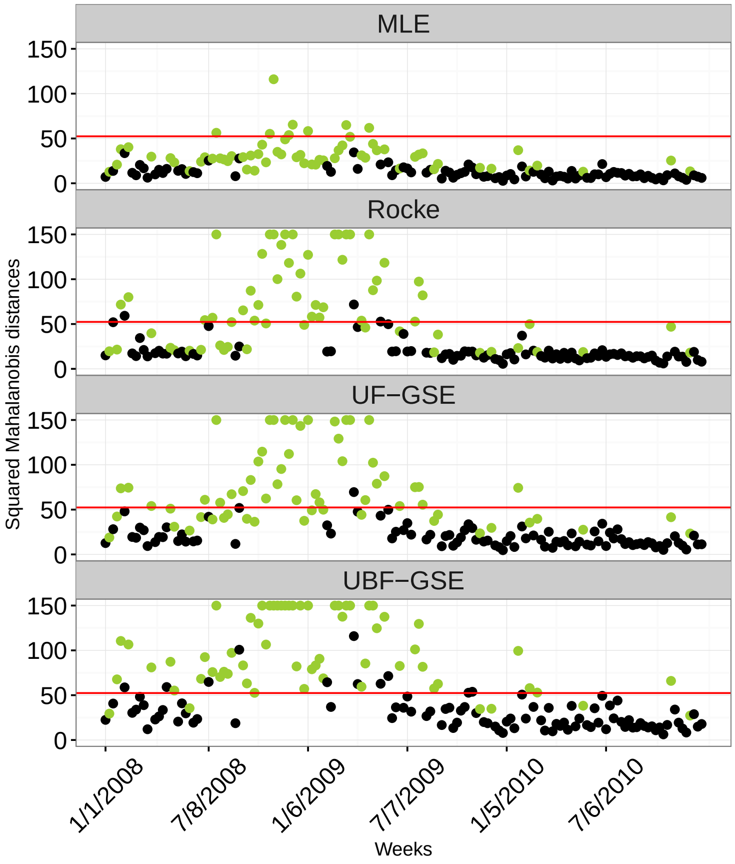


Figure 3.4: Squared Mahalanobis distances of the weekly observations in the small-cap asset returns data based on the MLE, the Rocke, the UF-GSE, and the UBF-GSE estimates. Weeks that contain one or more asset returns with values three MAD away from the coordinatewise-median are in green.

those weeks as MD outliers (i.e., failed to flag these weeks as having estimated full Mahalanobis distance exceeding the 99.99% quantile chi-squared distribution with 20 degrees of freedom). The MLE misses all but seven of the 59 green cases. The Rocke-S estimate does slightly better but still misses one third of the green cases. This is because it is severely affected by the large cellwise outliers that propagate to 40.1% of the cases. The UF-GSE estimate also does a relatively poor job. On the other hand, the UBF-GSE estimate successfully flaggs all but five of the 59 green cases.

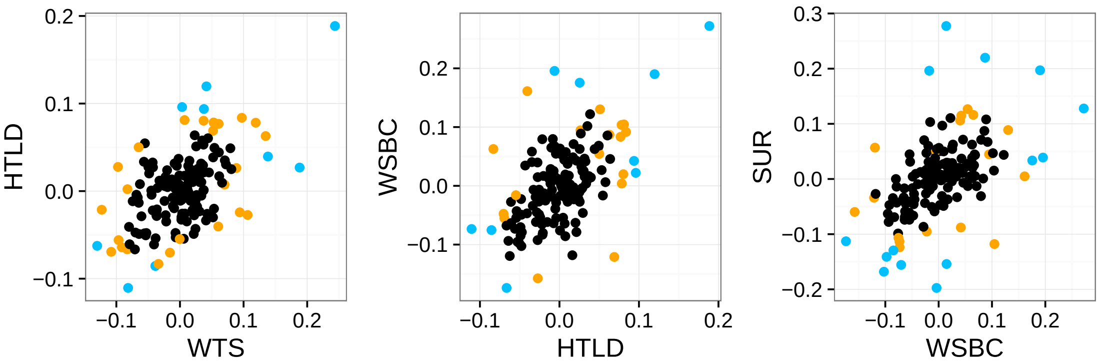


Figure 3.5: Pairwise scatterplots of the asset returns data for WTS versus HTLD, HTLD versus WSBC, and WSBC versus SUR. Points with components flagged by the univariate filter are in blue. Points with components additionally flagged by the bivariate filter are in orange.

Figure 3.5 shows the pairwise scatterplots for WTS versus HTLD, HTLD versus WSBC, and WSBC versus SUR with the results from the univariate and the bivariate filter. The points flagged by the univariate filter are in blue, and those flagged by the bivariate filter are in orange. We see that the bivariate filter has identified some additional cellwise outliers that are not-so-large marginally but become more visible when viewed together with other correlated components. These moderate cellwise outliers account for 6.9% of the cells in the data and propagate to 56.7% of the cases. The final median weight assigned to these cases by UF-GSE and UBF-GSE are 0.50 and 0.65, respectively. By filtering the moderate cellwise outliers, UBF-GSE makes a more effective use of the clean part of these partly contaminated data points (i.e., the 56.7% of the cases).

Figure 3.6 shows the squared Mahalanobis distances produced by UBF-GRE-C and UBF-GSE, for comparison. Here, we see that UBF-GRE-C has missed only 3 of the 59 green cases, while UBF-GSE has missed 6 of the 59. UBF-GRE-C has also clearly flagged weeks 36, 59, and 66 (with final weights 0.6, 0.0, and 0.0, respectively) as casewise outliers. In contrast, UBF-GSE gives final weights 0.8, 0.5, and 0.5 to these cases. As shown in the simulations, UBF-GSE has difficulty down-weighting some high dimensional outlying cases on datasets of high dimension.

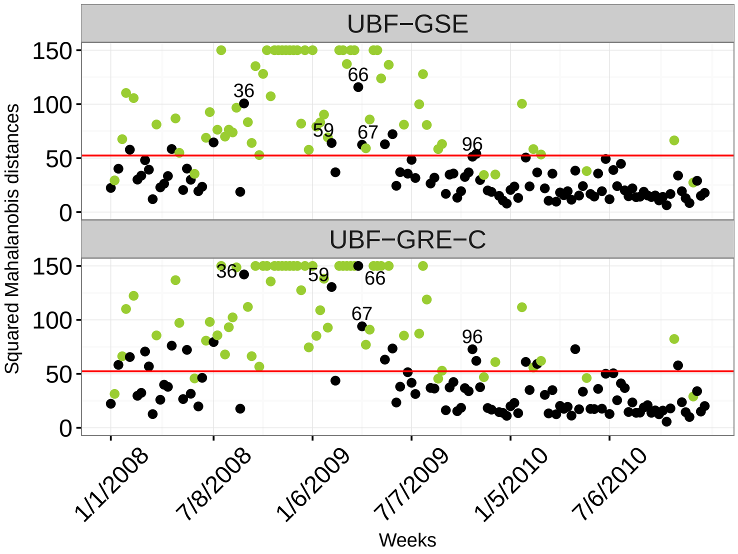


Figure 3.6: Squared Mahalanobis distances of the weekly observations in the small-cap asset returns data based on the UBF-GSE and the UBF-GRE-C estimates. Weeks that contain one or more asset returns with values three MAD away from the coordinatewise-median are in green.

In this example, UBF-GRE-C makes the most effective use of the clean part of the data and has the best outlier detecting performance among the considered estimates.

## 3.6 Conclusions

In this chapter, we overcome two serious limitations of GSE and UF-/UBF-GSE in higher dimensions ($p \geq 20$). First, these estimators show an incontrollable increase in Gaussian efficiency, which is paid off by a serious decrease in robustness, for larger $p$. Second, the initial estimator (extended minimum volume ellipsoids, EMVE) used by GSE and UF-/UBF-GSE does not scale well in higher dimensions because it requires an impractically large number of subsamples to achieve a high breakdown point in larger dimensions.

To achieve a controllable efficiency/robustness trade off in higher dimensions, we equip GSE and UF-/UBF-GSE with a Rocke type loss function. To overcome the high computational cost of the EMVE, we introduce a clustering-based subsampling procedure. We show via simulation studies that, in higher dimensions, estimators using the proposed subsampling with only 50 subsamples can achieve equivalent or even better performance than the usual uniform subsampling with 500 subsamples.

# Chapter 4

# Robust Regression Estimation and Inference in the Presence of Cellwise and Casewise Contamination

## 4.1 Introduction

In this chapter, we study another classic but fundamental problem, *linear regression estimation and inference*, under the same contamination paradigm as in the previous chapters.

The vast majority of procedures for robust linear regression are based on the classical Tukey–Huber contamination model (THCM) in which a relatively small fraction of cases may be contaminated. High breakdown point affine equivariant estimators such as least trimmed squares (Rousseeuw, 1984), S-regression (Rousseeuw & Yohai, 1984) and MM-regression (Yohai, 1985) proceed by down-weighting outlying cases, which makes sense and works well in practice, under THCM. However, in some applications, the contamination mechanism may be different in that random cells in a data table (with rows as cases and columns as variables) are independently contaminated. In this paradigm, a small fraction of random cellwise outliers could propagate to a relatively large fraction of cases, breaking down classical high breakdown point affine equivariant estimators (see Alqallaf et al., 2009). Since cellwise and casewise outliers may co-exist in some applications, our goal in this chapter is to develop a method for robust regression estimation and inference that can deal with both cellwise and casewise outliers.

There is a vast literature on robust regression for casewise outliers, but only a scant literature for cellwise outliers and none for both types of outliers in the regression context. Recently, Öllerer et al. (2015) combined the ideas of coordinate descent algorithm (called the shooting algorithm in Fu, 1998) and simple S-regression (Rousseeuw & Yohai, 1984) to propose an estimator called the shooting S. The shooting S-estimator assigns individual weight to each cell in the data table to handle cellwise outliers in the regression context. The shooting S-estimator is robust against cellwise outliers and vertical response outliers.

In this chapter, we propose a three-step regression estimator which combines the ideas of filtering cellwise outliers and robust regression via covariance matrix estimate (Maronna & Morgenthaler, 1986; Croux et al., 2003), namely 3S-regression estimator. By filtering, here we mean detecting outliers and replacing them by missing values as defined in Chapter 2. Our estimator proceeds as follows: first, it uses a univariate filter to detect and eliminate extreme cellwise outliers in order to control the effect of outliers propagation; second, it applies a robust estimator of multivariate location and scatter to the filtered data to down-weight casewise outliers; third, it computes robust regression coefficients from the estimates obtained in the second step. With the choice of a filter that has simultaneous good sensitivity (is capable of filtering outliers) and good specificity (can preserve all or most of the clean data), the resulting estimator can be resilient to both cellwise and casewise outliers; furthermore, it attains consistency and asymptotic normality for clean data. In this regards, we propose a new filter that is consistent under some assumptions on the tails of the covariates distributions. By consistent filter, we mean a filter that asymptotically can preserve all the data when they are clean.

The rest of the chapter is organized as follows. In Section 4.2, we introduce a new family of consistent filters. In Section 4.3, we introduce 3S-regression. In Section 4.4, we show some asymptotic properties of 3S-regression. In Section 4.5, we evaluate the performance of 3S-regression in an extensive simulation study. In Section 4.6, we analyze two real data sets with cellwise and casewise outliers. In Section 4.7, we conclude with some remarks. Finally, we also provide all the proofs, additional simulation results, and other related material in Appendix C.

## 4.2 Consistent filter for general continuous data

Filtering is a method for preprocessing data in order to control the effect of potential cellwise outliers. In this chapter, we pre-process the data by flagging outliers and replacing them by missing values, NAs, similar to what was proposed in Chapter 2, as well as in other applications (see e.g., Danilov, 2010; Farcomeni, 2014b,c).

Consistent filters are ones that do not filter good data points asymptotically. Gervini & Yohai (2002) introduced a consistent filter for normal residuals in regression estimation to achieve a fully-efficient robust regression estimator. Consistent filters are desirable because their good asymptotic properties are shared by the following-up estimation procedure. In this chapter, we introduce a new family of consistent filters for univariate data that are sufficiently general in regression application.

Consider a random variable $X$ with a continuous distribution function $G(x)$. We define the scaled upper and lower tail distributions of $G(x)$ as follows:

$$\begin{aligned} F^u(t) &= P_G\left(\frac{X-\eta^u}{\text{med}(X-\eta^u|X>\eta^u)} \le t|X>\eta^u\right) \quad \text{and} \\ F^l(t) &= P_G\left(\frac{\eta^l-X}{\text{med}(\eta^l-X|X<\eta^l)} \le t|X<\eta^l\right). \end{aligned} \tag{4.1}$$

Here, med stands for median, $\eta^u = G^{-1}(1-\alpha)$, $\eta^l = G^{-1}(\alpha)$, and $0 < \alpha < 0.5$. We use $\alpha = 0.20$, but other choices could be considered. To simplify the notation, we set $s^u = \text{med}(X-\eta^u|X>\eta^u)$ and $s^l = \text{med}(\eta^l - X|X<\eta^l)$. Alternatively, a combined tails approach could be used for symmetric distributions as in Gervini & Yohai (2002).

Let $\{X_1, \dots, X_n\}$ be a random sample from $G$, and let $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ be the corresponding order statistics. Consistent estimators for $(\eta^u, s^u, \eta^l, s^l)$ are given by

$$\eta_n^u = G_n^{-1}(1-\alpha), \quad s_n^u = \mathrm{med}(\{X_i - \eta_n^u | X_i > \eta_n^u\}),$$
$$\eta_n^l = G_n^{-1}(\alpha), \quad s_n^l = \mathrm{med}(\{\eta_n^l - X_i | X_i < \eta_n^l\}),$$

where $G_n^{-1}(a) = X_{(\lceil na \rceil)}$, $0 < a < 1$, is the empirical quantile and $\mathrm{med}(\{Y_1, \dots, Y_m\}) = Y_{(\lceil m/2 \rceil)}$ is the sample median (see Lemma C.1 in Section C.4 in the Appendix for a proof of the consistency for $s_n^u$ and $s_n^l$). The empirical distribution functions for the scaled upper and lower tails in (4.1) are now given by

$$F_n^u(t) = \frac{\sum_{i=1}^n I(0 < (X_i - \eta_n^u)/s_n^u \le t)}{\sum_{i=1}^n I(X_i > \eta_n^u)} \quad \text{and}$$
$$F_n^l(t) = \frac{\sum_{i=1}^n I(0 < (\eta_n^l - X_i)/s_n^l \le t)}{\sum_{i=1}^n I(X_i < \eta_n^l)}.$$

Upper and lower tails outliers can be flagged by comparing the empirical distribution functions for the scaled tails with their expected distributions. We assume that aside from contamination, $F^u$ and $F^l$ decay exponentially fast or faster. Let $\{a\}^+ = \max(0, a)$ denote the positive part of $a$. Then, we define the proportions of flagged upper and lower tails outliers by

$$d_n^u = \sup_{t \ge t_0} \{F_0(t) - F_n^u(t)\}^+ \quad \text{and} \quad d_n^l = \sup_{t \ge t_0} \left\{F_0(t) - F_n^l(t)\right\}^+,$$

where $F_0(t) = 1 - \exp(-\log(2)t)$ and $t_0 = 1/\log(2)$. When $X - \eta^u | X > \eta^u$ is exponentially distributed with a rate of $\lambda^u > 0$, the standardized tail $(X - \eta^u)/s^u | X > \eta^u$ have exponential distribution with a rate of $\log(2)$, leading to our choice of $F_0(t)$ and $t_0$. Finally, we flag $\lfloor n^u d_n^u \rfloor$ of the most extreme points in the upper tail and flag $\lfloor n^l d_n^l \rfloor$ of the most extreme points in the lower tail, where $n^u$ and $n^l$ are the number of observations in $\{X_i | X_i > \eta_n^u\}$ and $\{X_i | X_i < \eta_n^l\}$, respectively. Equivalently, setting

$$t_n^u = \min \{t : F_n^u(t) \ge 1 - d_n^u\} \quad \text{and} \quad t_n^l = \min \left\{t : F_n^l(t) \ge 1 - d_n^l\right\},$$

we filter $X_i$'s with $X_i < \eta_n^l - s_n^l t_n^l$ or $X_i > \eta_n^u + s_n^u t_n^u$.

We tried several heavy tail models for $F_0(t)$ including Pareto distributions with different tail indexes, and we found that the chosen exponential model strikes a good balance between the robustness and consistency of the filtering procedure.

Theorem 4.1 below shows that our filter is consistent under the following assumption on the tails of $G(x)$.

**Assumption 4.1.** *$G(x)$ is continuous, and $F^u(t)$ and $F^l(t)$ satisfy the following:*

$$F_0(t) - F^u(t) \leq 0, \quad t \geq t_0 \quad \textit{and} \quad F_0(t) - F^l(t) \leq 0, \quad t \geq t_0.$$

**Theorem 4.1.** *Suppose that Assumption 4.1 holds for $G(x)$. Then, $d_n^u \to 0$ a.s. and $d_n^l \to 0$ a.s.*

**Proof:** See Section C.4 in the Appendix.

In practice, the distributions $F^u(t)$ and $F^l(t)$ are unknown. To allow for some flexibility, Assumption 4.1 does not completely specify $F^u(t)$ and $F^l(t)$, but it only requires that their upper tails are as heavy as or lighter than the upper tail of $F_0(t)$.

## 4.3 Three-step regression

### 4.3.1 The estimator

Consider the model

$$Y_i = \alpha + \boldsymbol{X}_i^t \boldsymbol{\beta} + \varepsilon_i \tag{4.2}$$

for $i = 1, \ldots, n$, where the error terms $\varepsilon_i$ are i.i.d. and independent of the covariates $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^t$. The least squares (LS) estimator $(\alpha_{LS}, \boldsymbol{\beta}_{LS}^t)$ are defined as the minimizers of the sum squares of residuals,

$$(\alpha_{LS}, \boldsymbol{\beta}_{LS}^t) = \underset{(\alpha, \boldsymbol{\beta}^t) \in \mathbb{R}^{(p+1)}}{\arg\min} \sum_{i=1}^n (Y_i - \alpha - \boldsymbol{X}_i^t \boldsymbol{\beta})^2.$$

The solution to this problem is explicit:

$$\begin{aligned}\boldsymbol{\beta}_{LS} &= \boldsymbol{S}_{n,xx}^{-1}\boldsymbol{S}_{n,xy},\\ \alpha_{LS} &= \boldsymbol{m}_{n,y} - \boldsymbol{m}_{n,x}^{t}\boldsymbol{\beta}_{LS}.\end{aligned} \tag{4.3}$$

Here, $\boldsymbol{S}_{n,xx}, \boldsymbol{S}_{n,xy}, m_{n,y}$, and $\boldsymbol{m}_{n,x}$ are the components of the empirical covariance matrix and mean:

$$\boldsymbol{S}_n = \begin{pmatrix} \boldsymbol{S}_{n,xx} & \boldsymbol{S}_{n,xy} \\ \boldsymbol{S}_{n,yx} & S_{n,yy} \end{pmatrix} \quad \text{and} \quad \boldsymbol{m}_n = \begin{pmatrix} \boldsymbol{m}_{n,x} \\ m_{n,y} \end{pmatrix} \tag{4.4}$$

for the joint data $\{\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_n\}$ with $\boldsymbol{Z}_i = (\boldsymbol{X}_i^t, Y_i)^t$.

Several authors (see Maronna & Morgenthaler, 1986; Croux et al., 2003) proposed to achieve robust regression and inference for casewise outliers by robustifying the components in (4.3). Croux et al. (2003) replaced the empirical covariance matrix and mean by the multivariate S-estimator (Davies, 1987). We will refer to this approach as two-step regression (2S-regression). Croux et al. (2003) have shown that under mild assumptions (including symmetry of $\varepsilon_i$ and independence of $\varepsilon_i$ and $\boldsymbol{X}_i$) 2S-regression is Fisher consistent and asymptotically normal even if the S-estimators of multivariate location and scatter themselves are not consistent. Furthermore, 2S-regression is resilient to all kinds of outliers, that is, vertical outliers, bad leverage points, and good leverage points. Note that down-weighting good leverage points could lead to some efficiency loss, but it may also prevent the underestimation of the variance of the estimator, which could be problematic for inferential purposes (see for example, Ruppert & Simpson, 1990).

To deal with casewise and cellwise outliers, we propose to use a generalized S-estimator that uses the consistent filter described in Section 4.2. The estimator is similar to that in Agostinelli et al. (2015), but with the filter which is consistent for a broader range of distributions. This generality is needed in the regression setting.

Our proposed globally robust regression estimator, called 3S-regression, is given by:

$$\begin{aligned}\boldsymbol{\beta}_{3S} &= \boldsymbol{C}_{2S,xx}^{-1}\boldsymbol{C}_{2S,xy}\\ \alpha_{3S} &= T_{2S,y} - \boldsymbol{T}_{2S,m}^{t}\boldsymbol{\beta}_{3S}.\end{aligned} \tag{4.5}$$

Here, $(\boldsymbol{T}_{2S}, \boldsymbol{C}_{2S})$ is a two-step generalized S-estimator computed as follows:

Step 1. Filter extreme cellwise outliers to prevent cellwise contaminated cases from having large robust Mahalanobis distances in Step 2, and

Step 2. Down-weight casewise outliers by applying generalized S-estimator (GSE) for multivariate location and scatter (Danilov et al., 2012) to the filtered data from Step 1. The GSE is a generalization of the S-estimator for incomplete data that are missing completely at random (MCAR). Since the independent contamination model (ICM) assumes that cells are outlying completely at random, the MCAR assumption is fulfilled if the ICM model holds.

More precisely, consider a set of covariates $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$. We perform univariate filtering as described in Section 4.2 on each variable, $\{X_{1j}, \ldots, X_{nj}\}$, $j = 1, \ldots, p$. Let $\{\boldsymbol{U}_{n,1}, \ldots, \boldsymbol{U}_{n,n}\}$ be the resulting auxiliary vectors of zeros and ones with zeros indicating the filtered entry in $\boldsymbol{X}_i$. More precisely, $\boldsymbol{U}_{n,i} = (U_{n,i1}, \ldots, U_{n,ip})^t$, where

$$U_{n,ij} = I(\eta_{j,n}^{l} - s_{j,n}^{l} t_{j,n}^{l} \le X_i \le \eta_{j,n}^{u} + s_{j,n}^{u} t_{j,n}^{u}).$$

The goal of the filter is to prevent propagation of cellwise outliers. If the fraction of cases with at least one flagged cell is very small (below 1%, say) then propagation of cellwise outliers is not an issue and the filter can be safely turned off. The procedure that turns the filter off when the fraction of affected cases is below a given small threshold, $\xi$, is considerably simpler to analyze from the asymptotic point of view. Moreover, it retains all the robustness properties derived from the filter. Let $n_0 = \#\{1 \le i \le n : \boldsymbol{U}_{n,i} = \boldsymbol{1}\}$ be the number of complete observations after filtering. We set

$$\boldsymbol{U}_{n,i}^{*} = \boldsymbol{1} I\left(\frac{n-n_0}{n} \le \xi\right) + \boldsymbol{U}_{n,i} I\left(\frac{n-n_0}{n} > \xi\right), \quad i = 1, \ldots, n, \tag{4.6}$$

with $\xi$ equal to some small threshold. In this chapter, we use $\xi = 0.01$.

Finally, let $\mathbb{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)^t$ and $\mathbb{U}_n = ((\boldsymbol{U}^*_{n,1}, \ldots, \boldsymbol{U}^*_{n,n})^t, \boldsymbol{1})$. The two-step generalized S-estimator can now be defined as

$$\begin{aligned} \boldsymbol{T}_{2S} &= \boldsymbol{T}_{GS}(\mathbb{Z}, \mathbb{U}_n), \\ \boldsymbol{C}_{2S} &= \boldsymbol{C}_{GS}(\mathbb{Z}, \mathbb{U}_n), \end{aligned} \tag{4.7}$$

where $\boldsymbol{T}_{GS}$ and $\boldsymbol{C}_{GS}$ are robust multivariate location and scatter generalized S-estimator for incomplete data, $(\mathbb{Z}, \mathbb{U})$, with Tukey's bisquare rho function $\rho_B(t) = \min(1, 1-(1-t)^3)$ and 50% breakdown point (see Chapter 2 for full definition). Note that when $\mathbb{U} = (\boldsymbol{1}, \ldots, \boldsymbol{1})$ (i.e., when the input data is complete), the generalized S-estimator reduces to S-estimator (Danilov et al., 2012).

Alternatively, the second step can be replaced by GRE as introduced in Chapter 3. In the chapter, GSE was shown to lose robustness against casewise outliers for higher dimensional data. GRE has been proposed to remedy this problem. So, in the case of large $p$ in $\mathbb{X}$, GRE may be a more appropriate choice for the second step than GSE. However, in this chapter, we generally focus on smaller to moderate dimensional data (e.g., $p \leq 15$) and hence, GSE should be sufficient.

### 4.3.2 Models with continuous and dummy covariates

For models with continuous and dummy covariates, the direct computation of 3S-regression is likely to fail because the subsampling algorithm (needed to compute the generalized S-estimator) is likely to yield collinear subsamples. In this case, we endow 3S-regression with an iterative algorithm similar to that in Maronna & Yohai (2000) to deal with continuous and dummy covariates.

Consider now the following model:

$$Y_i = \alpha + \boldsymbol{X}_i^t \boldsymbol{\beta}_x + \boldsymbol{D}_i^t \boldsymbol{\beta}_d + \varepsilon_i \tag{4.8}$$

for $i = 1, \ldots, n$ where $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip_x})^t$ is a $p_x$ dimensional vector of continuous covariates and $\boldsymbol{D}_i = (D_{i1}, \ldots, D_{ip_d})^t$ is a $p_d$ dimensional vector of dummy covariates.

Set $\mathbb{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^t$, $\mathbb{D} = (\boldsymbol{D}_1, \ldots, \boldsymbol{D}_n)^t$, and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^t$. We assume that the columns in $\mathbb{X}$ and $\mathbb{D}$ are linearly independent.

We modify the alternating M- and S-regression approach proposed by Maronna & Yohai (2000). Our algorithm uses 3S-regression to estimate the coefficients of the continuous covariates and regression M-estimators with Huber's rho function $\rho_H(t) = \min(1, t^2/2)$ (Huber & Ronchetti, 2009) to estimate the coefficients of the dummy covariates. More specifically, the algorithm works as follows:

$$\begin{aligned} (\hat{\alpha}^{(k)}, \hat{\boldsymbol{\beta}}_x^{(k)}) &= g(\mathbb{X}, \boldsymbol{Y} - \mathbb{D}\hat{\boldsymbol{\beta}}_d^{(k-1)}), \\ \hat{\boldsymbol{\beta}}_d^{(k)} &= M(\mathbb{D}, \boldsymbol{Y} - \hat{\alpha}^{(k)} - \hat{\mathbb{X}}\hat{\boldsymbol{\beta}}_x^{(k)}), \quad \text{for} \quad k = 1, \ldots, K, \end{aligned} \tag{4.9}$$

where $g(\mathbb{X}, \boldsymbol{Y})$ denotes the operation of 3S-regression for a response vector $(\boldsymbol{Y}, \mathbb{X})$ as defined in (4.5) and $M(\mathbb{D}, \boldsymbol{Y})$ denotes the operation of regression $M$-estimator with no intercept for $(\boldsymbol{Y}, \mathbb{D})$. We let $\hat{\mathbb{X}}$ be the imputed $\mathbb{X}$ with the filtered entries imputed by the best linear predictor using $\hat{\boldsymbol{T}}^{(k)}$ and $\hat{\boldsymbol{C}}^{(k)}$, the generalized S-estimates at the $k$-th iteration as defined in (4.7). We use $\hat{\mathbb{X}}$ instead of $\mathbb{X}$ to control the effect of propagation of cellwise outliers.

As in Maronna & Yohai (2000), to calculate the initial estimates, $(\hat{\alpha}^{(0)}, \hat{\boldsymbol{\beta}}_x^{(0)}, \hat{\boldsymbol{\beta}}_d^{(0)})$, we first remove the effect of $\boldsymbol{D}_i$ from the continuous covariates and the response variable. Let

$$\overline{\boldsymbol{Y}} = \boldsymbol{Y} - \mathbb{D}\boldsymbol{t} \quad \text{and} \quad \overline{\mathbb{X}} = \mathbb{X} - \mathbb{D}\mathbb{T},$$

where $\boldsymbol{t} = M(\mathbb{D}, \boldsymbol{Y})$ and $\mathbb{T}$ is a $p_d \times p_x$-matrix with the $j$-th column as $\boldsymbol{T}_j = M(\mathbb{D}, (X_{1j}, \ldots, X_{nj})^t)$. Now, the initial estimates are defined by

$$\begin{aligned} (\hat{\alpha}^{(0)}, \hat{\boldsymbol{\beta}}_x^{(0)t}) &= g(\overline{\mathbb{X}}, \overline{\boldsymbol{Y}}), \\ \hat{\boldsymbol{\beta}}_d^{(0)} &= M(\mathbb{D}, \boldsymbol{Y} - \hat{\alpha}^{(0)} - \hat{\mathbb{X}}\hat{\boldsymbol{\beta}}_x^{(0)}). \end{aligned}$$

Finally, the procedure in (4.9) is iterated until convergence or until it reaches a maximum of $K = 20$ iterations. We choose $K = 20$ because our simulation has shown that the procedure usually converges for $K < 20$, provided good initial estimates are used.

## 4.4 Asymptotic properties of three-step regression

Theorem 4.2 establishes the equivalence between 3S-regression and 2S-regression (Croux et al., 2003) for the case of continuous covariates. Let $(\alpha_{3S}, \boldsymbol{\beta}_{3S}^t)$ be the 3S-regression estimator and $(\alpha_{2S}, \boldsymbol{\beta}_{2S}^t)$ be the 2S-regression estimator based on the sample $\{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n\}$, where $\boldsymbol{Z}_i = (\boldsymbol{X}_i^t, Y_i)$. Let $G(\boldsymbol{x})$ and $G_j(x)$ be the distribution functions for $\boldsymbol{X}_i$ for $X_{ij}$ respectively.

**Theorem 4.2.** *Suppose that Assumption 4.1 holds for each $G_j$, $j = 1, \ldots, p$. Then, with probability one, for sufficiently large $n$, $\alpha_{3S} = \alpha_{2S}$ and $\boldsymbol{\beta}_{3S} = \boldsymbol{\beta}_{2S}$.*

**Proof:** See Section C.4 in the Appendix.

Since 3S-regression becomes 2S-regression for sufficiently large $n$, 3S-regression inherits the established asymptotic properties of 2S-regression. Corollary 4.3 states the strong consistency and asymptotic normality of 3S-regression. The corollary requires the following regularity assumptions that are needed for deriving the consistency and asymptotic normality of 2S-regression (see Croux et al., 2003).

**Assumption 4.2.** *Let $F_\varepsilon$ be the distribution of the error term $\varepsilon_i$ in (4.2). The distribution $F_\varepsilon$ has a positive, symmetric and unimodal density $f_\varepsilon$.*

**Assumption 4.3.** *For all $\boldsymbol{v} \in \mathbb{R}^p$ and $\delta \in \mathbb{R}$, $P_G(\boldsymbol{X}_i^t \boldsymbol{v} = \delta) < 1/2$.*

**Corollary 4.3.** *Suppose that Assumption 4.1 holds for each $G_j$, $j = 1, \ldots, p$, and Assumption 4.2–4.3 hold. Denote $\boldsymbol{\theta}_{3S} = (\alpha_{3S}, \boldsymbol{\beta}_{3S}^t)^t$ and $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}^t)^t$. Then,*

*(a) $\theta_{3S} \to \theta$ a.s..*

*(b) Let $H$ be the distribution of $(\boldsymbol{X}^t, Y)$ and let $(\boldsymbol{T}_H, \boldsymbol{C}_H)$ be the S-estimator functional (see Lopuhaä, 1989). We use the same partition outlined in (4.4) for $(\boldsymbol{T}_H, \boldsymbol{C}_H)$. Set $\tilde{\boldsymbol{X}} = (1, \boldsymbol{X}^t)^t$. Then,*

$$\sqrt{n}(\boldsymbol{\theta}_{3S} - \boldsymbol{\theta}) \to_d N(\boldsymbol{0}, ASV(H)),$$

*where*

$$ASV(H) = C(H)^{-1} D(H) C(H)^{-1},$$

*and where*

$$\begin{aligned}
C(H) &= E_H\left\{ w(d_H(\boldsymbol{Z}))\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^t \right\} + \frac{2}{\sigma_\varepsilon^2(H)} E_H\left\{ w'(d_H(\boldsymbol{Z}))(Y - \tilde{\boldsymbol{X}}^t\boldsymbol{\theta})^2\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^t \right\},\\
D(H) &= E_H\left\{ w^2(d_H(\boldsymbol{Z}))(Y - \tilde{\boldsymbol{X}}^t\boldsymbol{\theta})^2\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^t \right\},\\
\sigma_\varepsilon(H) &= \sqrt{C_{H,yy} - \boldsymbol{\beta}^t\boldsymbol{C}_{H,xx}\boldsymbol{\beta}},\\
d_H(\boldsymbol{Z}) &= (\boldsymbol{Z} - \boldsymbol{T}_H)^t\boldsymbol{C}_H^{-1}(\boldsymbol{Z} - \boldsymbol{T}_H),\\
w(t) &= \rho_B'(t).
\end{aligned}$$

*Here, $\rho_B(t)$ is the Tukey's bisquare rho function.*

**Remark 4.1.** *Croux et al. (2003) proved the Fisher consistency of 2S-regression, but the strong consistency also follows from that and Theorem 3.2 in Lopuhaä (1989).*

The asymptotic covariance matrix needed for inference can be estimated in the following natural way. Let $(\hat{\boldsymbol{m}}, \hat{\boldsymbol{S}})$ be the generalized S-estimate and $(\hat{\alpha}_{3S}, \hat{\boldsymbol{\beta}}^t_{3S})$ be the 3S-regression estimate. Then, replace $\boldsymbol{Z}_i = (\boldsymbol{X}_i^t, Y_i)$ by $\hat{\boldsymbol{Z}}_i = (\hat{\boldsymbol{X}}^t_i, Y_i)$ and $\tilde{\boldsymbol{X}}_i = (1, \boldsymbol{X}_i^t)^t$ by $\hat{\tilde{\boldsymbol{X}}}_i = (1, \hat{\boldsymbol{X}}^t_i)^t$, where $\hat{\boldsymbol{X}}_i$ is the best linear prediction of $\boldsymbol{X}_i$ (which is possibly incomplete due to filter) using $(\hat{\boldsymbol{T}}, \hat{\boldsymbol{C}})$. The identified cellwise outliers in $\boldsymbol{X}_i$ are filtered and imputed in order to avoid the effect of propagation of outliers on the asymptotic covariance matrix estimation. Now,

$$\widehat{ASV(H)} = \widehat{C(H)}^{-1}\widehat{D(H)}\widehat{C(H)}^{-1},$$

where

$$\begin{aligned}
\widehat{C(H)} &= \frac{1}{n}\sum_{i=1}^{n}\left\{w(d_n(\hat{\boldsymbol{Z}}_i)) + \frac{2}{\hat{\sigma}^2_{\varepsilon,n}}w'(d_n(\hat{\boldsymbol{Z}}_i))\hat{r}_i^2\right\}\hat{\tilde{\boldsymbol{X}}}_i\hat{\tilde{\boldsymbol{X}}}_i^t,\\
\widehat{D(H)} &= \frac{1}{n}\sum_{i=1}^{n} w^2(d_n(\hat{\boldsymbol{Z}}_i))\hat{r}_i^2\hat{\tilde{\boldsymbol{X}}}_i\hat{\tilde{\boldsymbol{X}}}_i^t,\\
\hat{\sigma}_{\varepsilon,n} &= \sqrt{\hat{s}_{yy} - \hat{\boldsymbol{\beta}}_{3S}^t\hat{\boldsymbol{C}}_{xx}\hat{\boldsymbol{\beta}}_{3S}},\\
d_n(\hat{\boldsymbol{Z}}_i) &= (\hat{\boldsymbol{Z}}_i - \hat{\boldsymbol{T}})^t\hat{\boldsymbol{C}}^{-1}(\hat{\boldsymbol{Z}}_i - \hat{\boldsymbol{T}}),\\
\hat{r}_i &= Y_i - \hat{\tilde{\boldsymbol{X}}}_i^t\hat{\boldsymbol{\theta}}_{3S}.
\end{aligned}$$

Although the asymptotic covariance matrix formula is valid under clean data, we shall show in Section 4.5 that our proposed inference remains approximately valid in the presence of a moderate fraction of cellwise and casewise outliers.

In the case of continuous and dummy covariates, Maronna & Yohai (2000) derived asymptotic results for the alternating regression M- and S-estimates. However, there is no proof of asymptotic results when regression S-estimators are replaced by 2S-regression. The study of the asymptotic properties of the alternating M- and 2S-regression is worth of future research.

## 4.5 Simulations

We carried out extensive simulation studies in R (R Core Team, 2015) to investigate the performance of 3S-regression by comparing it with least square (LS) and two robust alternatives:

(i) 2S-regression as in Croux et al. (2003). The location and scatter S-estimator with bisquare $\rho$ function and 50% breakdown point is computed by an iterative algorithm that uses an initial MVE estimator. The MVE estimator is computed by subsampling with a concentration step. This procedure is implemented in the `R` package `rrcov`, function `CovSest`, option `method="bisquare"` (Todorov & Filzmoser, 2009); and

(ii) Shooting S-estimator introduced in Öllerer et al. (2015) with bisquare $\rho$ function and 20% breakdown point (for each simple regression) as suggested by the authors to attain a good trade-off between robustness and efficiency. The `R` code is available at `http://feb.kuleuven.be/Viktoria.Oellerer/software`.

The generalized S-estimates needed by 3S-regression are computed using the `R` package `GSE`, function `GSE` with default options (Leung et al., 2015). The regression M-estimates needed by the alternating M- and 3S-regression are computed using the `R` package `MASS`, function `rlm`, option `method="M"` (Venables & Ripley, 2002). Finally, the proposed procedures are implemented in the R package `robreg3S`, which is freely available on CRAN (the Comprehensive R Archive Network, R Core Team, 2015).

### 4.5.1 Models with continuous covariates

We consider the regression model in (4.2) with $p = 15$ and $n = 150, 300, 500, 1000$. The random covariates $\boldsymbol{X}_i$, $i = 1, \ldots, n$, are generated from multivariate normal distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We set $\boldsymbol{\mu} = \mathbf{0}$ and $\Sigma_{jj} = 1$ for $j = 1, \ldots, p$ without loss of generality because GSE in the second step of 3S-regression is location and scale equivariant. To address the fact that 3S-regression and the shooting S-estimator are not affine-equivariant, we consider the random correlation structure for $\boldsymbol{\Sigma}$ as described in Agostinelli et al. (2015). We fix the condition number of the random correlation matrix at 100 to mimic the practical situation for data sets of similar dimensions. Furthermore, to address the fact that the two estimators are not regression equivariant, we randomly generate $\boldsymbol{\beta}$ as $\boldsymbol{\beta} = R\boldsymbol{b}$, where $\boldsymbol{b}$ has a uniform distribution on the unit spherical surface and $R$ is set to 10. We set $\alpha = 0$ because GSE is location equivariant. The response variable $Y_i$ is given by $Y_i = \boldsymbol{X}_i^t\boldsymbol{\beta} + \varepsilon_i$, where $\varepsilon_i$ are independent (also independent of $\boldsymbol{X}_i$'s) identically normally distributed with mean 0 and $\sigma = 0.5$. Finally, we consider the following scenarios:

- Clean data: No further changes are done to the data;
- Cellwise contamination: Randomly replace a fraction $\epsilon$ of the cells in the covariates by outliers $X_{ij}^{cont} = E(X_{ij}) + k \times SD(X_{ij})$ and $\epsilon$ proportion of the

Table 4.1: Maximum $\overline{MSE}$ in all the considered scenarios for models with continuous covariates.

| | Clean | | 1% Cellwise | | 5% Cellwise | | Casewise | |
|---|---|---|---|---|---|---|---|---|
| $n =$ | 150 | 300 | 150 | 300 | 150 | 300 | 150 | 300 |
| 3S | 0.012 | 0.005 | 0.039 | 0.020 | 0.902 | 0.797 | 0.223 | 0.143 |
| ShootS | 0.034 | 0.017 | 0.134 | 0.080 | 1.129 | 0.912 | 1.570 | 1.460 |
| 2S | 0.010 | 0.004 | 0.025 | 0.014 | 3.364 | 3.041 | 0.109 | 0.122 |
| LS | 0.009 | 0.004 | 2.723 | 2.440 | 4.812 | 4.732 | 8.286 | 8.182 |

responses by outliers $Y_{ij}^{cont} = E(Y_{ij}) + k \times SD(\varepsilon_i)$, where $k = 1, 2, \ldots, 10$;

- Casewise contamination: Randomly replace a fraction $\epsilon$ of the cases by leverage outliers $(\boldsymbol{X}_i^{cont\,t}, Y_i^{cont})$, where $\boldsymbol{X}_i^{cont} = c\boldsymbol{v}$ and $Y_i^{cont} = \boldsymbol{X}_i^{cont\,t}\boldsymbol{\beta} + \varepsilon_i^{cont}$ with $\varepsilon_i^{cont} \sim N(k, \sigma^2)$, where $k = 1, 2, \ldots, 15$. Here, $\boldsymbol{v}$ is the eigenvector corresponding to the smallest eigenvalue of $\boldsymbol{\Sigma}$ with length such that $(\boldsymbol{v}-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{v}-\boldsymbol{\mu}) = 1$. Monte Carlo experiments show that the placement of outliers in this direction, $\boldsymbol{v}$, is the least favorable for our estimator. We repeat the simulation study in Agostinelli et al. (2015) for dimension 16 and observe that $c = 8$ is the least favorable value for the performance of the scatter estimator.

We consider $\epsilon = 0.01, 0.05$ for cellwise contamination, and $\epsilon = 0.10$ for casewise contamination. The number of replicates for each setting is $N = 1000$.

### Coefficient estimation performance

We examine the effect of cellwise and casewise outliers on the bias of the estimated coefficients. We evaluate the bias using the Monte Carlo mean squared error (MSE):

$$\overline{MSE} = \frac{1}{N}\sum_{m=1}^{N}\frac{1}{p}\sum_{j=1}^{p}(\hat{\beta}_{n,j}^{(m)} - \beta_j^{(m)})^2$$

where $\hat{\beta}_{n,j}^{(m)}$ is the estimate for $\beta_j^{(m)}$ at the $m$-th simulation run.

Table 4.1 shows the $\overline{MSE}$ for clean data and the maximum $\overline{MSE}$ for all the cellwise and casewise contamination settings for $n = 150, 300$. Figure 4.1 shows the

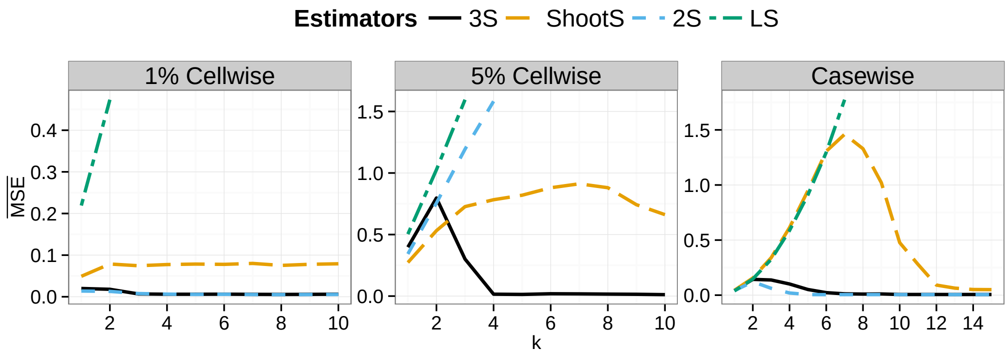


Figure 4.1: $\overline{MSE}$ for various cellwise and casewise contamination values, $k$, for models with continuous covariates. The sample size is $n = 300$.

curves of $\overline{MSE}$ for various cellwise and casewise contamination values for $n = 300$. The results for $n = 150$ are similar and the corresponding figure is shown in Section C.1 in the Appendix.

In the cellwise contamination setting, 3S-regression is highly robust against moderate and large cellwise outliers ($k \geq 3$), but less robust against inliers ($k \leq 2$). Notice that inliers also affect the performance of the shooting S-estimator but to a lesser extent. Since the filter does not flag inliers, 3S-regression and 2S-regression perform similarly in the presence of inliers (see the central panel of Figure 4.1). The shooting S-estimator is highly robust against large outliers, but less so against moderate cellwise outliers. As expected, 2S-regression breaks down in the case of $\epsilon = 0.05$, when the propagation of large cellwise outliers is expected to affect more than 50% of the cases.

In the casewise contamination setting, 2S-regression has the best performance, as expected. 3S-regression also performs fairly well in this setting. The shooting S-estimator performs less satisfactorily in this case.

We have also considered other simulation settings and observed similar results (not shown here). In particular, we considered $p = 5$ with $n = 50, 100$ and $p = 25$ with $n = 250, 500$ under the same set of scenarios (clean data, cellwise contamination, and casewise contamination). Moreover, we studied the performance of 3S-regression for larger casewise contamination levels up to 20%. 3S-regression maintains its competitive performance, outperforming Shooting S and not falling too far behind 2S-regression, which is expected to win in these situations.

### Performance of confidence intervals

We then assess the performance of confidence intervals for the regression coefficients based on the asymptotic covariance matrix as described in Section 4.4. Intervals that have a coverage close to the nominal value, while being relatively short, are desirable.

The $100(1-\tau)\%$ confidence interval (CI) of 3S-regression has the form:

$$CI(\hat{\beta}_{n,j}) = \left[\hat{\beta}_{n,j} - \Phi^{-1}(1-\tau/2)\sqrt{\widehat{ASV}(\hat{\beta}_{n,j})/n},\ \hat{\beta}_{n,j} + \Phi^{-1}(1-\tau/2)\sqrt{\widehat{ASV}(\hat{\beta}_{n,j})/n}\right],$$

for $j = 0, 1, \ldots, p$, where $\hat{\beta}_{n,0} = \hat{\alpha}_n$. We consider $\tau = 0.05$ here. We evaluate the performance of CI using the Monte Carlo mean coverage rate (CR):

$$\overline{CR} = \frac{1}{N}\sum_{m=1}^{N}\frac{1}{p}\sum_{j=1}^{p} I(\beta_j^{(m)} \in CI(\hat{\beta}_{n,j}^{(m)})),$$

and the Monte Carlo mean CI lengths:

$$\overline{CIL} = \frac{1}{N}\sum_{m=1}^{N}\frac{1}{p}\sum_{j=1}^{p} 2\Phi^{-1}(1-\tau/2)\sqrt{\widehat{ASV}(\hat{\beta}_{n,j})/n}.$$

Figure 4.2 shows the $\overline{CR}$ in the case of clean data, 5% cellwise contamination ($k = 5$), and 10% casewise contamination ($k = 3$) simulation, with different sample sizes $n = 150, 300, 500, 1000$. The nominal value of 95% is indicated by the horizontal line in the figure.

For clean data, the coverage rates of all the intervals reach the nominal level when the sample size grows, as expected. For data with casewise outliers, 2S-regression yields the best coverage rate, which is closest to the nominal level. However, 3S-regression has an acceptable performance, comparable with that of 2S-regression.

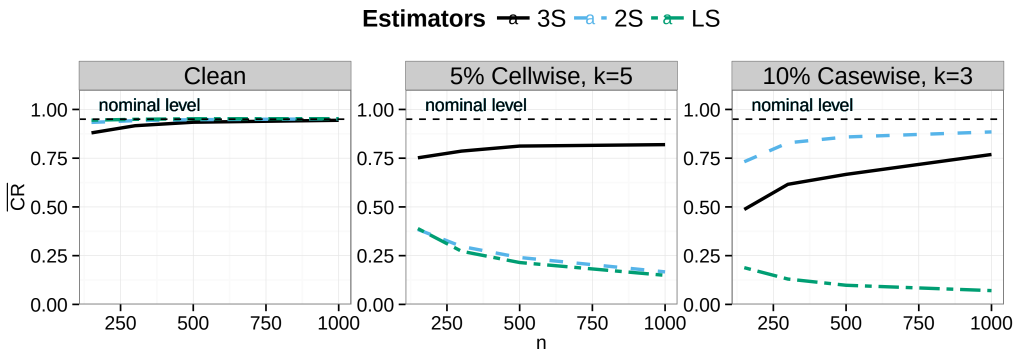


Figure 4.2: $\overline{CR}$ for clean data and for cellwise and casewise contaminated data of various sample size, $n$.

Table 4.2: Average lengths of confidence intervals for clean data and for cellwise and casewise contamination.

| | Clean | | 1% Cell., $k = 5$ | | 5% Cell., $k = 5$ | | 10% Case., $k = 3$ | |
|---|---|---|---|---|---|---|---|---|
| Size ($n$) | 3S | 2S | 3S | 2S | 3S | 2S | 3S | 2S |
| 150 | 0.341 | 0.352 | 0.355 | 0.402 | 0.450 | 1.519 | 0.329 | 0.355 |
| 300 | 0.242 | 0.247 | 0.244 | 0.275 | 0.294 | 1.148 | 0.239 | 0.253 |
| 500 | 0.187 | 0.189 | 0.190 | 0.212 | 0.222 | 0.912 | 0.189 | 0.197 |
| 1000 | 0.133 | 0.133 | 0.134 | 0.150 | 0.155 | 0.662 | 0.137 | 0.140 |

For data with cellwise outliers, 3S-regression yields intervals with a coverage rate relatively closer to the nominal value than LS and 2S-regression.

Furthermore, the length of the intervals obtained from 3S regression is comparable to that LS for clean data and that of 2S-regression for clean data and data with casewise outliers. For data with cellwise outliers, 3S-regression yields intervals with lengths relatively closer to the case of clean data. Table 4.2 shows the average lengths of the confidence intervals obtained from 3S- and 2S-regression in the case of clean data, 1% cellwise contamination ($k = 5$), 5% cellwise contamination ($k = 5$), and 10% casewise contamination ($k = 3$) simulation, with different sample sizes $n = 150, 300, 500, 1000$. The results of LS are not included here.

In general, 3S-regression yields slightly shorter intervals than 2S-regression in all scenarios because the asymptotic variance is calculated on the data with the filtered cells imputed instead of the complete data. On the other hand, 2S-regression tends to yield longer intervals in the cellwise contamination model, even when the propagation of outliers is below the 0.5 breakdown point under THCM, for example, when $\varepsilon = 0.01$. This maybe because 2S-regression loses a significant amount of clean data for estimation when it down-weights cases with outlying components.

### 4.5.2 Models with continuous and dummy covariates

We now conduct a simulation study to assess the performance of our procedure when the model includes continuous and dummy covariates. We consider the regression model in (4.8) with $p_x = 12$, $p_d = 3$, and $n = 150, 300$. The random covariates $(\boldsymbol{X}_i, \boldsymbol{D}_i)$, $i = 1, \ldots, n$, are first generated from multivariate normal distribution $N_p(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is the randomly generated correlation matrix with a fixed condition number of 100. Then, we dichotomize $D_{ij}$ at $\Phi^{-1}(\pi_j)$ where $\pi_j = \frac{1}{4}, \frac{1}{3}, \frac{1}{2}$ for $j = 1, 2, 3$, respectively. Finally, the rest of data are generated in the same way as described in Section 4.5.1.

In the simulation study, we consider the following scenarios:

- Clean data: No further changes are done to the data;
- Cellwise contamination: Randomly replace a $\epsilon$ fraction of the cells in $\mathbb{X}$ by outliers $X_{ij}^{cont} = E(X_{ij}) + k \times SD(X_{ij})$ and $\epsilon$ proportion of the responses by outliers $Y_{ij}^{cont} = E(Y_{ij}) + k \times SD(\varepsilon_i)$, where $k = 1, 2, \ldots, 10$;
- Casewise contamination: Let $\boldsymbol{\Sigma}_x$ be the sub-matrix of $\boldsymbol{\Sigma}$ with rows and columns corresponding to the continuous covariates. Randomly replace a $\epsilon$ fraction of the cases in $\mathbb{X}$ by leverage outliers $\boldsymbol{X}_i^{cont} = c\boldsymbol{v}$, where $\boldsymbol{v}$ is the eigenvector corresponding to the smallest eigenvalue of $\boldsymbol{\Sigma}_x$ with length such that $(\boldsymbol{v} - \boldsymbol{\mu}_x)^t \boldsymbol{\Sigma}^{-1} (\boldsymbol{v} - \boldsymbol{\mu}_x) = 1$. In this case, the number of continuous variables is 13 (instead of 16) and the corresponding least favorable casewise contamination size is found to be $c = 7$ (instead of 8) using the same procedure

Table 4.3: Maximum $\overline{MSE}$ in all the considered scenarios for models with continuous and dummy covariates.

| | Clean | | 1% Cellwise | | 5% Cellwise | | Casewise | |
|---|---|---|---|---|---|---|---|---|
| $n =$ | 150 | 300 | 150 | 300 | 150 | 300 | 150 | 300 |
| 3S | 0.010 | 0.004 | 0.018 | 0.008 | 0.636 | 0.507 | 0.090 | 0.071 |
| ShootS | 0.012 | 0.005 | 0.026 | 0.015 | 0.746 | 0.468 | 0.450 | 0.387 |
| 2S | 0.008 | 0.003 | 0.014 | 0.007 | 1.894 | 1.341 | 0.060 | 0.054 |
| LS | 0.007 | 0.003 | 2.785 | 2.532 | 5.162 | 4.981 | 1.332 | 1.322 |

as in Section 4.5.1. Finally, we replace the corresponding response value by $Y_i^{cont} = \boldsymbol{X}_i^{cont\,t}\boldsymbol{\beta}_x + \boldsymbol{D}_i^t\boldsymbol{\beta}_d + \varepsilon_i^{cont}$ with $\varepsilon_i^{cont} \sim N(k, \sigma^2)$, where $k = 1, 2, \ldots, 10$.

Again, we consider $\epsilon = 0.01, 0.05$ for cellwise contamination, and $\epsilon = 0.10$ for casewise contamination. The number of replicates for each setting is $N = 1000$.

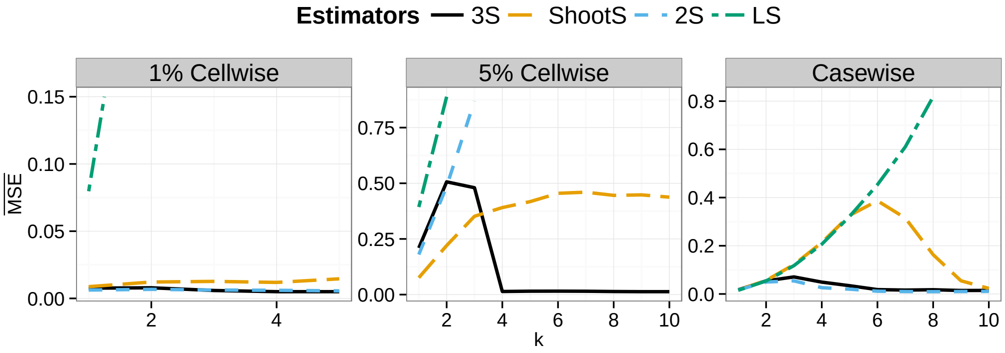


Figure 4.3: $\overline{MSE}$ for various cellwise and casewise contamination values, $k$, for models with continuous and dummy covariates. The sample size is $n = 300$.

Table 4.3 shows the $\overline{MSE}$ for clean data and the maximum $\overline{MSE}$ for all the cellwise and casewise contamination settings for $n = 150, 300$. Figure 4.3 shows the curves of $\overline{MSE}$ for various cellwise and casewise contamination values for $n = 300$. The results for $n = 150$ are similar and the corresponding figure is shown in Section C.1 in the Appendix. Overall, 3S-regression remains competitive in the case of continuous and dummy covariates.

We also consider the case of non-normal covariates. The covariates are generated from several asymmetric distributions, and the data are contaminated in a similar fashion. The performance of 3S-regression in the case of non-normal covariates is similar to the performance in the case of normal covariates. Results are available in Section C.2 in the Appendix.

## 4.6 Real data example: Boston Housing data

We illustrate the effect of cellwise outlier propagation on classical robust estimators using the Boston Housing data. The data, available at the UCI repository (Bache & Lichman, 2013), was collected from 506 census tracts in the Boston Standard Statistical Metropolitan Area in the 1970s on 14 different features. We consider the nine quantitative variables that were extensively studied (e.g., see in Öllerer et al., 2015). The variables are listed and described in Table C.2 in the Appendix. There is no missing data. The original objective of the study in Harrison & Rubinfeld (1978) was to analyze the association between the median housing values (medv) in Boston and the residents' willingness to pay for clean air, as well as the association between medv and those variables on the list.

We fit the following model using 3S-regression, the shooting S-estimator, 2S-regression and the LS estimator:

$$\begin{aligned} log(medv) &= \alpha + \beta_1\, log(crim) + \beta_2\, nox^2 + \beta_3\, rm^2 + \beta_{x,4}\, age \\ &\qquad + \beta_5\, log(dis) + \beta_6\, tax + \beta_7\, ptratio + \beta_8\, black + \beta_9\, log(lstat) + \varepsilon. \end{aligned}$$

The regression coefficient estimates and their P-values are given in Table 4.4. In particular, we observe that the regression coefficients for the covariates *age* and *black* are very different under 3S and 2S-regression. Moreover, *age* is significant under 2S-regression but highly non-significant under 3S-regression. 2S-regression is somewhat inefficient because it throws away a substantial amount of clean data due to the propagation of cellwise outliers. It fully down-weights 16.4% of the cases in the dataset (cases that receive a zero weight by the multivariate S-estimator). Slightly

Table 4.4: Estimates and p-values of the regression coefficients for the original Boston Housing data.

| Variable | 3S | | ShootS | | 2S | | LS | |
|---|---|---|---|---|---|---|---|---|
| | Coeff. | P-Val. | Coeff. | P-Val. | Coeff. | P-Val. | Coeff. | P-Val. |
| log(lstat) | -0.243 | <0.001 | -0.266 | - | -0.153 | <0.001 | -0.395 | <0.001 |
| $\mathrm{rm}^2$ | 0.015 | <0.001 | 0.013 | - | 0.018 | <0.001 | 0.007 | <0.001 |
| tax | -0.051 | <0.001 | -0.021 | - | -0.046 | <0.001 | -0.028 | 0.006 |
| log(dis) | -0.125 | <0.001 | -0.157 | - | -0.126 | <0.001 | -0.139 | <0.001 |
| ptratio | -0.026 | <0.001 | -0.027 | - | -0.025 | <0.001 | -0.029 | <0.001 |
| $\mathrm{nox}^2$ | -0.578 | 0.013 | -0.463 | - | -0.445 | 0.023 | -0.451 | <0.001 |
| age | -0.023 | 0.645 | -0.040 | - | -0.152 | 0.001 | 0.050 | 0.391 |
| black | -0.726 | 0.398 | 0.787 | - | -0.007 | 0.993 | 0.500 | <0.001 |
| log(crim) | -0.006 | 0.513 | 0.004 | - | 0.005 | 0.527 | -0.002 | 0.813 |

Table 4.5: Pairwise squared norm distances between the estimates for the original Boston housing data.

| | 3S | ShootS | 2S | LS |
|---|---|---|---|---|
| 3S | - | 1.389 | 3.145 | 6.725 |
| ShootS | | - | 4.312 | 4.661 |
| 2S | | | - | 16.614 |
| LS | | | | - |

more than half of these cases (8.7%) are affected by the propagation of cellwise outliers mainly in the covariates $nox^2$ and $black$ (1.3% of the cells in the dataset are flagged by the consistent filter). After filtering, these cases have relatively small partial Mahalanobis distances, indicating they are close to the bulk of the data for the remaining variables.

We further compare the four estimators by computing their squared norm distances, $n \times \sum_{j=1}^{p}(\widehat{\beta}_{n,j,A} - \widehat{\beta}_{n,j,B})^2 \times MAD(\{X_{1j}, \dots, X_{nj}\})^2$ (see Öllerer et al., 2015), where $MAD$ is the median absolute deviation. Table 4.5 shows the squared norm distances for the considered estimators. Overall, the three robust estimators are very different from LS. As expected, 3S-regression and shooting S are closer to each other than they are to 2S-regression. We next illustrate that the observed differences between the three robust estimators are indeed mostly caused by the propagation of cellwise outliers in the Boston housing data.

Table 4.6: Estimates and p-values of the regression coefficients for the imputed Boston Housing data.

| Variable | 3S | | ShootS | | 2S | | LS | |
|---|---|---|---|---|---|---|---|---|
| | Coeff. | P-Val. | Coeff. | P-Val. | Coeff. | P-Val. | Coeff. | P-Val. |
| log(lstat) | -0.243 | 0.000 | -0.264 | - | -0.227 | <0.001 | -0.385 | <0.001 |
| $\text{rm}^2$ | 0.015 | 0.000 | 0.013 | - | 0.014 | <0.001 | 0.009 | <0.001 |
| tax | -0.051 | 0.000 | -0.030 | - | -0.047 | <0.001 | -0.032 | 0.002 |
| log(dis) | -0.125 | 0.000 | -0.161 | - | -0.129 | <0.001 | -0.144 | <0.001 |
| ptratio | -0.026 | 0.000 | -0.028 | - | -0.025 | <0.001 | -0.027 | <0.001 |
| $\text{nox}^2$ | -0.578 | 0.013 | -0.522 | - | -0.619 | 0.010 | -0.479 | <0.001 |
| age | -0.023 | 0.645 | -0.037 | - | -0.037 | 0.471 | 0.051 | 0.386 |
| black | -0.726 | 0.398 | 0.371 | - | -0.882 | 0.376 | -0.206 | 0.519 |
| log(crim) | -0.006 | 0.513 | -0.001 | - | -0.012 | 0.233 | -0.012 | 0.213 |

Table 4.7: Pairwise squared norm distances between the estimates for the imputed Boston housing data.

| | 3S | ShootS | 2S | LS |
|---|---|---|---|---|
| 3S | - | 0.862 | 0.172 | 5.486 |
| ShootS | | - | 1.158 | 3.992 |
| 2S | | | - | 6.366 |
| LS | | | | - |

Recall that half of the cases fully down-weighted by 2S-regression have entries flagged as cellwise outliers. We replace these flagged cells by their best linear predictions (using the 3S-regression estimate) and then, refit the model with the four considered estimators. The resulting coefficient estimates and their P-values are given in Table 4.6. Notice that the covariate *age* is no longer significant under 2S-regression. Moreover, Table 4.7 shows the norm distances between all the estimates calculated from such imputed data. Now, 2S-regression is considerably closer to the cellwise robust estimators, and it no longer fully down-weights the cases formerly affected by cellwise outliers (the median weight of these cases is now 0.64, closer to the overall median weight, 0.69). The LS estimator remains different from the robust estimators, possibly due to the existence of casewise outliers in the data. MM-regression (Yohai, 1985) behaves similarly to 2S-regression in this example.

## 4.7 Conclusions

High breakdown point affine equivariant robust estimators are neither efficient nor robust in the independent cellwise contamination model (ICM). By efficiency here we mean the ability to use the clean part of the data. In fact, classical robust estimators are inefficient under ICM because they may down-weight an entire row with a single component being contaminated. Therefore, they may lose some useful information contained in the data. Furthermore, the classical high breakdown point affine equivariant robust estimators may break down under ICM. A small fraction of cellwise outliers could propagate, affecting a large proportion of cases. For instance, the probability $\overline{\epsilon}$ that at least one component of a case is contaminated is $\overline{\epsilon} = 1 - (1 - \epsilon)^p$, where $\epsilon$ is the proportion of independent cellwise outliers. This implies that even if $\epsilon$ is small, $\overline{\epsilon}$ could be large for large $p$, and could exceed the 0.5 breakdown point under THCM. For example, if $\epsilon = 0.1$ and $p = 10$, then $\overline{\epsilon} = 0.65$; and if $\epsilon = 0.05$ and $p = 20$, then $\overline{\epsilon} = 0.64$.

To overcome these deficiencies of the classical robust estimators, we introduce a three-step regression estimator that can deal with cellwise and casewise outliers. The first step of our estimator is aimed at reducing the impact of outliers propagation posed by ICM. The second step is aimed at achieving robustness under THCM. As a result, the robust regression estimate from the third step is shown to be efficient (in terms of data usage) and robust under ICM and THCM. We also prove that our estimator is consistent and asymptotically normal at the central regression model distribution. Finally, we extend our estimator to models with continuous and dummy covariates and provide an algorithm to compute the regression coefficients.

# Chapter 5

# Conclusions

In this thesis, two important aspects of robust analysis are studied:

- Multivariate location and scatter matrix under cellwise and casewise contamination;
- Regression analysis under cellwise and casewise contamination.

The following is an outline of the main results, some limitations that the proposed methods have, and the directions we foresee for future work.

In Chapter 2 and 3, a two-step procedure for estimating multivariate location and scatter matrix is proposed. Four estimators are derived from the procedure:

- UF-GSE for less correlated data in moderate dimensions ($p \leq 15$);
- UBF-GSE for more correlated data in moderate dimensions ($p \leq 15$);
- UF-GRE for less correlated data in high dimensions ($p > 15$); and
- UBF-GRE for more correlated data in high dimensions ($p > 15$).

Simulation results have shown that these estimators provide fairly high resistance against cellwise and casewise outliers, when comparing with the best performing robust estimators in their settings. However, the two-step procedure still has some limitations. For sample sizes $2p < n \leq 5p$, the estimator may encounter convergence problems and result in close-to-singular estimates. The problems may be remedied by using graphical lasso (GLASSO; Friedman et al., 2008) to make close-to-singular (or even singular) estimates better conditioned. However, for sample sizes $n \leq 2p$, the estimators fail to exist.

There is a recently proposed cellwise robust estimator for high dimensional data with small $n/p$ called the GGQ estimator by Öllerer & Croux (2015), which in spirit coincides with the work of Tarr et al. (2016). The GGQ estimator is defined by a procedure of calculating correlations pairwise using normal scores (also known as the Gaussian rank correlation), then applying GLASSO to the pairwise scatter matrix. Because the correlation estimation is done pairwise, it can handle data with $n < p$. In our preliminary study, we have found that GGQ exhibits fairly high robustness against cellwise outliers, but is not so robust against casewise outliers. We believe that pairwise estimation itself is not sufficient to deal with casewise outliers and finely structured high dimensional data. Hence, there is still a need for further research on high dimensional estimation in the presence of cellwise and casewise contamination when $n/p$ is small.

In Chapter 4, a three-step procedure for robust regression with continuous covariates is proposed. The procedure coincides with the classical two-step procedure for robust regression of Croux et al. (2003), when the first step of filtering is removed. The method is extended to handle both continuous and dummy covariates. Simulation results and example have shown that the procedure handles both cellwise outliers and casewise outliers similarly well. Asymptotic results are provided for the case of continuous covariates, but no results are available for the case of continuous and dummy covariates, which is a major limitation of the procedure. Interestingly, there are also no asymptotic results for the classical procedure of Croux et al. (2003), in this setting. Hence, we believe the study of the asymptotic properties of the extended procedure for regression with continuous and dummy covariates is a worthwhile project for future research.

Due to the novelty of the topic of cellwise and casewise contamination, we hope that in general further research will follow this thesis.

# Bibliography

Agostinelli, C., Leung, A., Yohai, V. J., & Zamar, R. H. (2015). Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. *TEST*, *24*(3), 441–461. iii, 10, 11, 42, 63, 70, 71

Alqallaf, F. (2003). *A New Contamination Model for Robust Estimation with Large High-Dimensional Data Sets*. PhD thesis, University of British Columbia. 28

Alqallaf, F., Van Aelst, S., Yohai, V. J., & Zamar, R. H. (2009). Propagation of outliers in multivariate data. *Ann Statist*, *37*(1), 311–331. 13, 14, 58

Alqallaf, F. A., Konis, K. P., Martin, R. D., & Zamar, R. H. (2002). Scalable robust covariance and correlation estimates for data mining. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, (pp. 14–23). 14

Bache, K. & Lichman, M. (2013). UCI machine learning repository. `http://archive.ics.uci.edu/ml`. 77

Croux, C. & Öllerer, V. (2015). Comments on: Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. *TEST*, *24*(3), 462–466. 37, 38

Croux, C., van Aelst, S., & Dehon, C. (2003). Bounded influence regression using high breakdown scatter matrices. *55*, 265–285. 59, 63, 67, 68, 69, 82

Danilov, M. (2010). *Robust Estimation of Multivariate Scatter under Non-Affine Equivarint Scenarios*. PhD thesis, University of British Columbia. 11, 15, 28, 60

Danilov, M., Yohai, V. J., & Zamar, R. H. (2012). Robust estimation of multivariate location and scatter in the presence of missing data. *J Amer Statist Assoc*, *107*, 1178–1186. iii, 16, 23, 24, 37, 38, 39, 41, 45, 47, 48, 64, 65, 97, 112

Davies, P. (1987). Asymptotic behaviour of S-estimators of multivariate location parameters and dispersion matrices. *Ann Statist*, *15*, 1269–1292. 13, 25, 63

Donoho, D. L. (1982). *Breakdown Properties of Multivariate Location Estimators.* PhD thesis, Harvard University. 13, 14

Farcomeni, A. (2014a). Robust constrained clustering in presence of entry-wise outliers. *Technometrics*, *56*, 102–111. 15, 26

Farcomeni, A. (2014b). Robust constrained clustering in presence of entry-wise outliers. *Technometrics*, *56*, 102–111. 60

Farcomeni, A. (2014c). Snipping for robust K-means clustering under component-wise contamination. *Stat Comp*, *24*, 909–917. 60

Farcomeni, A. & Leung, A. (2014). *snipEM: Snipping methods for robust estimation and clustering.* R package version 1.0. 27

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*(3), 432–441. 81

Fu, W. (1998). Penalized regressions: The bridge versus the lasso. *J Comput Graph Statist*, *7*(3), 397–416. 59

Gervini, D. & Yohai, V. J. (2002). A class of robust and fully efficient regression estimators. *Ann Statist*, *30*(2), 583–616. 16, 17, 60

Gnanadesikan, R. & Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, *28*, 81–124. 19, 46

Hall, P., Marron, J., & Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *J R Stat Soc Ser B Stat Methodol*, *67*, 427–444. 45

Harrison, D. & Rubinfeld, D. L. (1978). Hedonic prices and the demand for clean air. *J Environ Econ Manage*, *5*, 81–102. 77

Huber, P. J. (1981). *Robust Statistics*. New York: John Wiley & Sons. 14

Huber, P. J. & Ronchetti, E. M. (2009). *Robust Statistics (2nd edition)*. New Jersey: John Wiley & Sons. 66

Kaveh, V. (2015). *DetMCD: DetMCD Algorithm (Robust and Deterministic Estimation of Location and Scatter)*. R package version 0.0.2. 27

Leung, A., Danilov, M., Yohai, V., & Zamar, R. (2015). *GSE: Robust Estimation in the Presence of Cellwise and Casewise Contamination and Missing Data*. R package version 3.2.3. iii, 11, 27, 48, 70

Leung, A., Zhang, H., & Zamar, R. (2015). *robreg3S: Three-Step Regression and Inference for Cellwise and Casewise Contamination*. R package version 0.4. iv, 11

Leung, A., Zhang, H., & Zamar, R. (2016). Robust regression estimation and inference in the presence of cellwise and casewise contamination. *Comput Statist Data Anal*, *99*, 1–11. iv, 11

Lopuhaä, H. P. (1989). On the relation between S-estimators and M-estimators of multivariate location and covariance. *Ann Statist*, *17*, 1662–1683. 67, 68

Maronna, R. A. (2015). Comments on: Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. *TEST*, *24*(3), 471–472. 37, 42

Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. Chichister: John Wiley & Sons. 25, 43, 48

Maronna, R. A. & Morgenthaler, S. (1986). Robust regression through robust covariance matrices. *Comm Statist Theory Methods*, *15*, 1347–1365. 59, 63

Maronna, R. A. & Yohai, V. J. (2000). Robust regression with both continuous and categorical predictors. *J Statist Plann Inference*, *89*, 197–214. 65, 66, 69

Maronna, R. A. & Yohai, V. J. (2015). Robust and efficient estimation of high dimensional scatter and location. *arXiv:1504.03389 [math.ST]*. 5, 26, 37

Martin, R. (2013). Robust covariances: Common risk versus specific risk outliers. Presented at the 2013 R-Finance Conference, Chicago, IL, `www.rinfinance.com/agenda/2013/talk/DougMartin.pdf`, visited 2016-08-24. 6, 52

Öllerer, V., Alfons, A., & Croux, C. (2015). The shooting S-estimator for robust regression. 59, 69, 77, 78

Öllerer, V. & Croux, C. (2015). Robust high-dimensional precision matrix estimation. In K. Nordhausen & S. Taskinen (Eds.), *Modern Nonparametric, Robust and Multivariate Methods: Festschrift in Honour of Hannu Oja* (pp. 325–350). Springer. 27, 82

Peña, D. & Prieto, F. J. (2001). Multivariate outlier detection and robust covariance matrix estimation. *Technometrics*, *43*, 286–310. 26

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. iii, 11, 25, 69, 70

Rocke, D. M. (1996). Robustness properties of S-estimators of multivariate location and shape in high dimension. *Ann Statist*, *24*, 1327–1345. 26, 37, 42, 43, 102

Rousseeuw, P. (1984). Least median of squares regression. *J Amer Statist Assoc*, *79*, 871–880. 58

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. In W. Grossmann, G. Pflug, I. Vincze, & W. Wertz (Eds.), *Mathematical statistics and applications*, volume B (pp. 256–272). Dordrecht: Reidel Publishing Company. 13, 38

Rousseeuw, P. J. & Croux, C. (1993). Alternatives to the median absolute deviation. *J Amer Statist Assoc*, *88*, 1273–1283. 46

Rousseeuw, P. J. & Van den Bossche, W. (2015). Comments on: Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. *TEST*, *24*(3), 473–477. 18, 27, 37

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

Rousseeuw, P. J. & Yohai, V. J. (1984). Robust regression by means of S-estimators. In J. Franke, W. Härdle, & D. Martin (Eds.), *Robust and Nonlinear Time Series*, volume 26 of *Lecture Notes in Statistics* (pp. 256–272). New York, US: Springer. 58, 59

Ruppert, D. & Simpson, D. (1990). Unmasking multivariate outliers and leverage points: Comment. *J Amer Statist Assoc*, *85*, 644–646. 63

Smith, R. E., Campbell, N. A., & Lichfield, A. (1984). Multivariate statistical techniques applied to pisolitic laterite geochemistry at Golden Grove, Western Australia. *J Geochem Explor*, *22*, 193–216. 3, 33

Stahel, W. A. (1981). Breakdown of covariance estimators. Technical Report 31, Fachgruppe für Statistik, ETH Zürich, Switzerland. 13, 14

Tarr, G., Müller, S., & Weber, N. (2016). Robust estimation of precision matrices under cellwise contamination. *Comput Statist Data Anal*, *93*, 404–420. 82

Tatsuoka, K. S. & Tyler, D. E. (2000). On the uniqueness of S-functionals and M-functionals under nonelliptical distributions. *Ann Statist*, *28*, 1219–1243. 13, 37

Todorov, V. & Filzmoser, P. (2009). An object-oriented framework for robust multivariate analysis. *Journal of Statistical Software*, *32*(3), 1–47. 26, 69

Van Aelst, S. (2015). Comments on: Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. *TEST*, *24*(3), 478–481. 37

Van Aelst, S., Vandervieren, E., & Willems, G. (2012). A Stahel-Donoho estimator based on Huberized outlyingness. *Comput Statist Data Anal*, *56*, 531–542. 14, 15, 26

Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth ed.). New York: Springer. ISBN 0-387-95457-0. 70

Welsch, R. (2015). Comments on: Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. *TEST*, *24*(3), 482–483. 37

Yohai, V. J. (1985). High breakdown point and high efficiency robust estimates for regression. Technical Report 66, Department of Statistics, University of Washington. available at `http://www.stat.washington.edu/research/reports/1985/tr066.pdf`, visited 2016-08-24. 58, 79, 97

# Appendix A

# Supplementary material for Chapter 2

## A.1 Additional tables from the simulation study in Section 2.4

Table A.1: Maximum average LRT distances under cellwise contamination. The sample size is $n = 5p$.

| Corr. | $p$ | $\epsilon$ | MCD | MVE-S | Rocke | HSD | Snip | DMCDSc | UF-GSE | UBF-GSE |
|---|---|---|---|---|---|---|---|---|---|---|
| Random | 5 | 0.05 | 4.1 | 1.5 | 4.1 | 2.3 | 52.8 | 3.4 | 3.5 | 5.8 |
| | | 0.10 | 14.3 | 12.8 | 63.9 | 9.0 | 154.2 | 8.7 | 7.9 | 10.3 |
| | 10 | 0.05 | 9.7 | 13.4 | 32.0 | 11.7 | 18.7 | 5.8 | 5.6 | 5.8 |
| | | 0.10 | 142.9 | 134.4 | 156.3 | 56.9 | 66.4 | 16.4 | 16.5 | 16.1 |
| | 15 | 0.05 | 41.2 | 60.9 | 62.9 | 30.8 | 13.0 | 9.0 | 9.2 | 9.6 |
| | | 0.10 | 198.5 | 198.4 | 202.6 | 134.7 | 26.9 | 21.3 | 21.5 | 21.5 |
| | 20 | 0.05 | 72.9 | 94.8 | 90.7 | 55.9 | 15.4 | 11.2 | 12.2 | 12.4 |
| | | 0.10 | 240.9 | 240.7 | 242.1 | 243.4 | 19.8 | 26.1 | 25.1 | 25.1 |
| AR1(0.9) | 5 | 0.05 | 3.8 | 1.5 | 3.7 | 1.5 | 39.6 | 2.3 | 3.2 | 5.9 |
| | | 0.10 | 21.3 | 15.6 | 44.5 | 2.6 | 117.6 | 5.3 | 4.6 | 9.2 |
| | 10 | 0.05 | 15.4 | 19.8 | 44.6 | 3.7 | 11.9 | 5.1 | 3.6 | 2.7 |
| | | 0.10 | 219.1 | 187.4 | 220.6 | 19.8 | 90.7 | 15.1 | 11.8 | 5.3 |
| | 15 | 0.05 | 96.3 | 99.9 | 134.2 | 12.2 | 12.5 | 9.7 | 6.8 | 3.3 |
| | | 0.10 | 367.3 | 369.6 | 387.6 | 83.9 | 55.4 | 26.8 | 21.8 | 8.8 |
| | 20 | 0.05 | 174.4 | 197.3 | 235.9 | 28.5 | 19.1 | 14.9 | 10.9 | 4.4 |
| | | 0.10 | 518.5 | 526.5 | 557.8 | 260.8 | 34.3 | 39.8 | 33.7 | 15.3 |

Table A.2: Maximum average LRT distances under casewise contamination. The sample size is $n = 5p$.

| Corr. | $p$ | $\epsilon$ | MCD | MVE-S | Rocke | HSD | Snip | DMCDSc | UF-GSE | UBF-GSE |
|---|---|---|---|---|---|---|---|---|---|---|
| Random | 5 | 0.10 | 4.1 | 1.5 | 3.0 | 1.7 | 8.2 | 3.4 | 4.7 | 9.7 |
| | | 0.20 | 32.3 | 12.1 | 30.5 | 7.2 | 29.9 | 26.3 | 35.9 | 54.3 |
| | 10 | 0.10 | 11.1 | 3.9 | 4.7 | 5.1 | 27.8 | 9.0 | 15.9 | 28.9 |
| | | 0.20 | 128.2 | 71.0 | 16.6 | 28.1 | 57.2 | 62.9 | 86.6 | 102.7 |
| | 15 | 0.10 | 28.7 | 7.6 | 4.9 | 8.8 | 41.3 | 26.3 | 33.1 | 47.9 |
| | | 0.20 | 146.5 | 109.3 | 20.5 | 64.1 | 81.5 | 86.9 | 122.1 | 143.2 |
| | 20 | 0.10 | 76.6 | 17.8 | 6.7 | 16.0 | 57.6 | 49.2 | 59.7 | 67.5 |
| | | 0.20 | 167.7 | 141.9 | 19.0 | 111.0 | 103.4 | 110.1 | 154.4 | 183.3 |
| AR1(0.9) | 5 | 0.10 | 3.9 | 1.4 | 1.9 | 1.6 | 8.9 | 2.5 | 2.2 | 3.8 |
| | | 0.20 | 18.7 | 8.0 | 28.4 | 3.7 | 22.6 | 6.9 | 11.2 | 13.3 |
| | 10 | 0.10 | 9.4 | 3.9 | 3.3 | 2.7 | 19.1 | 4.8 | 4.9 | 5.9 |
| | | 0.20 | 122.7 | 59.2 | 29.1 | 11.6 | 44.8 | 31.0 | 69.0 | 57.2 |
| | 15 | 0.10 | 19.3 | 7.1 | 4.3 | 3.9 | 29.8 | 8.5 | 10.1 | 13.1 |
| | | 0.20 | 139.8 | 98.0 | 32.6 | 22.8 | 64.0 | 69.5 | 100.2 | 103.4 |
| | 20 | 0.10 | 72.3 | 16.3 | 5.6 | 6.3 | 48.7 | 16.6 | 34.9 | 37.3 |
| | | 0.20 | 161.5 | 127.2 | 23.6 | 48.5 | 87.5 | 108.5 | 129.3 | 133.8 |

Table A.3: Finite sample efficiency for first order autoregressive correlations, AR1($\rho$), with $\rho = 0.9$. The sample size is $n = 5p$.

| $p$ | MCD | MVE-S | Rocke | HSD | Snip | DMCDSc | UF-GSE | UBF-GSE |
|---|---|---|---|---|---|---|---|---|
| 5 | 0.22 | 0.62 | 0.50 | 0.50 | 0.21 | 0.30 | 0.45 | 0.29 |
| 10 | 0.32 | 0.86 | 0.55 | 0.71 | 0.08 | 0.39 | 0.74 | 0.61 |
| 15 | 0.42 | 0.94 | 0.55 | 0.83 | 0.28 | 0.44 | 0.84 | 0.71 |
| 20 | 0.48 | 0.96 | 0.56 | 0.88 | 0.17 | 0.48 | 0.87 | 0.77 |

## A.2 Proofs of propositions and theorem

### Proof of Proposition 2.1

*Proof.* Without loss of generality, set $\mu_0 = 0$ and $\sigma_0 = 1$. Let $Z_{0i} = \frac{X_i - \mu_0}{\sigma_0} = X_i$ and $Z_i = \frac{X_i - T_{0n}}{S_{0n}}$. Denote the empirical distributions of $Z_{01}, \dots, Z_{0n}$ and $Z_1, \dots, Z_n$ by

$$F_{0n}^{+}(t) = \frac{1}{n}\sum_{i=1}^{n} I\left(|Z_{0i}| \leq t\right) \quad \text{and} \quad F_{n}^{+}(t) = \frac{1}{n}\sum_{i=1}^{n} I\left(|Z_{i}| \leq t\right).$$

By assumption, with probability one, there exists $n_1$ such that $n \geq n_1$ implies $0 < 1 - \delta \leq S_{0n} \leq 1 + \delta$ and $-\delta \leq T_{0n} \leq \delta$, and we have

$$\begin{aligned}
F_n^{+}(t) &= \frac{1}{n}\sum_{i=1}^{n} I\left(-t \leq Z_i \leq t\right) = \frac{1}{n}\sum_{i=1}^{n} I\left(-t \leq \frac{X_i - T_{0n}}{S_{0n}} \leq t\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} I\left(-tS_{0n} + T_{0n} \leq X_i \leq tS_{0n} + T_{0n}\right) \\
&\geq \frac{1}{n}\sum_{i=1}^{n} I\left(-t(1-\delta) + T_{0n} \leq X_i \leq t(1-\delta) + T_{0n}\right) \\
&\geq \frac{1}{n}\sum_{i=1}^{n} I\left(-t(1-\delta) + \delta \leq X_i \leq t(1-\delta) - \delta\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} I\left(|X_i| \leq t(1-\delta) - \delta\right) = F_{0n}^{+}(t(1-\delta) - \delta).
\end{aligned}$$

Now, by the Glivenko–Cantelli Theorem, with probability one there exists $n_2$ such that $n \geq n_2$ implies that $\sup_t |F_{0n}^{+}(t) - F_0^{+}(t)| \leq \varepsilon/2$. Also, by the uniform continuity of $F_0^{+}$, given $\varepsilon > 0$, there exists $\delta > 0$ such that $|F_0^{+}(t(1-\delta) - \delta) - F_0^{+}(t)| \leq \varepsilon/2$.

Finally, note that

$$\begin{aligned}
F_n^{+}(t) &\geq F_{0n}^{+}(t(1-\delta) - \delta) \\
&= \left(F_{0n}^{+}(t(1-\delta) - \delta) - F_0^{+}(t(1-\delta) - \delta)\right) \\
&\qquad + \left(F_0^{+}(t(1-\delta) - \delta) - F_0^{+}(t)\right) + \left(F_0^{+}(t) - F^{+}(t)\right) + F^{+}(t).
\end{aligned}$$

Let $n_3 = \max(n_1, n_2)$, then $n \geq n_3$ imply

$$\begin{aligned}\sup_{t>\eta}(F^+(t) - F_n^+(t)) &\leq \sup_{t>\eta}\left|F_0^+(t(1-\delta)-\delta) - F_{0n}^+(t(1-\delta)-\delta)\right| \\ &\quad + \sup_{t>\eta}\left|F_0^+(t) - F_0^+(t(1-\delta)-\delta)\right| + \sup_{t>\eta}(F^+(t) - F_0^+(t)) \\ &\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} + 0 = \varepsilon.\end{aligned}$$

This implies that $n_0/n \to 0$ a.s.. □

## Proof of Proposition 2.2

We need the following lemma for the proof.

**Lemma A.1.** *Consider a sample of p-dimensional random vectors $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$. Also, consider a pair of multivariate location and scatter estimators $\boldsymbol{T}_{0n}$ and $\boldsymbol{C}_{0n}$. Suppose that $\boldsymbol{T}_{0n} \to \boldsymbol{\mu}_0$ and $\boldsymbol{C}_{0n} \to \boldsymbol{\Sigma}_0$ a.s.. Let $D_i = (\boldsymbol{X}_i - \boldsymbol{T}_{0n})^t \boldsymbol{C}_{0n}^{-1}(\boldsymbol{X}_i - \boldsymbol{T}_{0n})$ and $D_i = (\boldsymbol{X}_i - \boldsymbol{\mu}_0)^t \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{X}_i - \boldsymbol{\mu}_0)$. Given $K < \infty$. For all $i = 1, \dots, n$, if $D_{0i} \leq K$ , then:*

$$D_i \to D_{0i} \quad a.s..$$

*Proof of Lemma A.1.* Note that

$$\begin{aligned}|D_i - D_{0i}| &= |(\boldsymbol{X}_i - \boldsymbol{T}_{0n})^t \boldsymbol{C}_{0n}^{-1}(\boldsymbol{X}_i - \boldsymbol{T}_{0n}) - (\boldsymbol{X}_i - \boldsymbol{\mu}_0)^t \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{X}_i - \boldsymbol{\mu}_0)| \\ &= |((\boldsymbol{X}_i - \boldsymbol{\mu}_0) + (\boldsymbol{\mu}_0 - \boldsymbol{T}_{0n}))^t(\boldsymbol{\Sigma}_0^{-1} + (\boldsymbol{C}_{0n}^{-1} - \boldsymbol{\Sigma}_0^{-1}))((\boldsymbol{X}_i - \boldsymbol{\mu}_0) + (\boldsymbol{\mu}_0 - \boldsymbol{T}_{0n})) \\ &\qquad - (\boldsymbol{X}_i - \boldsymbol{\mu}_0)^t \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{X}_i - \boldsymbol{\mu}_0)| \\ &\leq |(\boldsymbol{\mu}_0 - \boldsymbol{T}_{0n})^t \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{T}_{0n})| + |(\boldsymbol{\mu}_0 - \boldsymbol{T}_{0n})^t(\boldsymbol{C}_{0n}^{-1} - \boldsymbol{\Sigma}_0^{-1})(\boldsymbol{\mu}_0 - \boldsymbol{T}_{0n})| \\ &\quad + |2(\boldsymbol{X}_i - \boldsymbol{\mu}_0)^t \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{T}_{0n})| + |2(\boldsymbol{X}_i - \boldsymbol{\mu}_0)^t(\boldsymbol{C}_{0n}^{-1} - \boldsymbol{\Sigma}_0^{-1})(\boldsymbol{\mu}_0 - \boldsymbol{T}_{0n})| \\ &\qquad + |(\boldsymbol{X}_i - \boldsymbol{\mu}_0)^t(\boldsymbol{C}_{0n}^{-1} - \boldsymbol{\Sigma}_0^{-1})(\boldsymbol{X}_i - \boldsymbol{\mu}_0)| \\ &= A_n + B_n + C_n + D_n + E_n.\end{aligned}$$

By assumption, there exists $n_1$ such that for $n \geq n_1$ implies $A_n \leq \varepsilon/5$ and $B_n \leq \varepsilon/5$.

Next, note that

$$\begin{aligned}|(\boldsymbol{X}_i-\boldsymbol{\mu}_0)^t\boldsymbol{\Sigma}_0^{-1/2}\boldsymbol{y}| &= |\boldsymbol{y}^t\boldsymbol{\Sigma}_0^{-1/2}(\boldsymbol{X}_i-\boldsymbol{\mu}_0)|\\ &\le ||\boldsymbol{y}||||\boldsymbol{\Sigma}_0^{-1/2}(\boldsymbol{X}_i-\boldsymbol{\mu}_0)|| = ||\boldsymbol{y}||\sqrt{(\boldsymbol{X}_i-\boldsymbol{\mu}_0)^t\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{X}_i-\boldsymbol{\mu}_0)} \le ||\boldsymbol{y}||\sqrt{K}.\end{aligned}$$

So, there exists $n_2$ such that $n \ge n_2$ implies

$$\begin{aligned}C_n &= |2(\boldsymbol{X}_i-\boldsymbol{\mu}_0)^t\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}_0-\boldsymbol{T}_{0n})|\\ &= |2(\boldsymbol{X}_i-\boldsymbol{\mu}_0)^t\boldsymbol{\Sigma}_0^{-1/2}\boldsymbol{\Sigma}_0^{-1/2}(\boldsymbol{\mu}_0-\boldsymbol{T}_{0n})|\\ &\le 2||\boldsymbol{\Sigma}_0^{-1/2}(\boldsymbol{\mu}_0-\boldsymbol{T}_{0n})||\sqrt{K}\\ &\le \varepsilon/5.\end{aligned}$$

Similarly, there exists $n_3$ such that $n \ge n_3$ implies

$$\begin{aligned}D_n &= |2(\boldsymbol{X}_i-\boldsymbol{\mu}_0)^t(\boldsymbol{C}_{0n}^{-1}-\boldsymbol{\Sigma}_0^{-1})(\boldsymbol{\mu}_0-\boldsymbol{T}_{0n})|\\ &= |2(\boldsymbol{X}_i-\boldsymbol{\mu}_0)^t\boldsymbol{\Sigma}_0^{-1/2}\boldsymbol{\Sigma}_0^{1/2}(\boldsymbol{C}_{0n}^{-1}-\boldsymbol{\Sigma}_0^{-1})(\boldsymbol{\mu}_0-\boldsymbol{T}_{0n})|\\ &\le 2||\boldsymbol{\Sigma}_0^{1/2}(\boldsymbol{C}_{0n}^{-1}-\boldsymbol{\Sigma}_0^{-1})(\boldsymbol{\mu}_0-\boldsymbol{T}_{0n})||\sqrt{K}\\ &\le \varepsilon/5.\end{aligned}$$

Also, there exists $n_4$ such that $n \ge n_4$ implies

$$\begin{aligned}E_n &= |(\boldsymbol{X}_i-\boldsymbol{\mu}_0)^t(\boldsymbol{C}_{0n}^{-1}-\boldsymbol{\Sigma}_0^{-1})(\boldsymbol{X}_i-\boldsymbol{\mu}_0)|\\ &= |(\boldsymbol{X}_i-\boldsymbol{\mu}_0)^t\boldsymbol{\Sigma}_0^{-1/2}\boldsymbol{\Sigma}_0^{1/2}(\boldsymbol{C}_{0n}^{-1}-\boldsymbol{\Sigma}_0^{-1})(\boldsymbol{X}_i-\boldsymbol{\mu}_0)|\\ &\le ||\boldsymbol{\Sigma}_0^{1/2}(\boldsymbol{C}_{0n}^{-1}-\boldsymbol{\Sigma}_0^{-1})(\boldsymbol{X}_i-\boldsymbol{\mu}_0)||\sqrt{K}\\ &\le ||(\boldsymbol{C}_{0n}^{-1}-\boldsymbol{\Sigma}_0^{-1})||\ ||\boldsymbol{\Sigma}_0^{1/2}(\boldsymbol{X}_i-\boldsymbol{\mu}_0)||\sqrt{K}\\ &\le ||(\boldsymbol{C}_{0n}^{-1}-\boldsymbol{\Sigma}_0^{-1})||K\\ &\le \varepsilon/5.\end{aligned}$$

Finally, let $n_5 = \max\{n_1, n_2, n_3, n_4\}$, then for all $i$, $n \geq n_5$ implies

$$|D_i - D_{0i}| \leq \varepsilon/5 + \varepsilon/5 + \varepsilon/5 + \varepsilon/5 + \varepsilon/5 = \varepsilon.$$

□

*Proof of Proposition 2.2.* Let $D_{0i} = (\boldsymbol{X}_i - \boldsymbol{\mu}_0)^t \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{X}_i - \boldsymbol{\mu}_0)$ and $D_i = (\boldsymbol{X}_i - \boldsymbol{T}_{0n})^t \boldsymbol{C}_{0n}^{-1} (\boldsymbol{X}_i - \boldsymbol{T}_{0n})$. Denote the empirical distributions of $D_{01}, \ldots, D_{0n}$ and $D_1, \ldots, D_n$ by

$$G_{0n}(t) = \frac{1}{n} \sum_{i=1}^{n} I\left(D_{0i} \leq t\right) \quad \text{and} \quad G_n(t) = \frac{1}{n} \sum_{i=1}^{n} I\left(D_i \leq t\right).$$

Note that

$$\begin{aligned} |G_n(t) - G_{0n}(t)| &= \left| \frac{1}{n} \sum_{i=1}^{n} I\left(D_i \leq t\right) - \frac{1}{n} \sum_{i=1}^{n} I\left(D_{0i} \leq t\right) \right| \\ &= \left| \frac{1}{n} \sum_{i=1}^{n} I\left(D_i \leq t\right) I(D_{0i} > K) + \frac{1}{n} \sum_{i=1}^{n} I\left(D_i \leq t\right) I(D_{0i} \leq K) \right. \\ &\quad \left. - \frac{1}{n} \sum_{i=1}^{n} I\left(D_{0i} \leq t\right) I(D_{0i} > K) - \frac{1}{n} \sum_{i=1}^{n} I\left(D_{0i} \leq t\right) I(D_{0i} \leq K) \right| \\ &\leq \left| \frac{1}{n} \sum_{i=1}^{n} I\left(D_i \leq t\right) I(D_{0i} > K) - \frac{1}{n} \sum_{i=1}^{n} I\left(D_{0i} \leq t\right) I(D_{0i} > K) \right| \\ &\quad + \left| \frac{1}{n} \sum_{i=1}^{n} I\left(D_i \leq t\right) I(D_{0i} \leq K) - \frac{1}{n} \sum_{i=1}^{n} I\left(D_{0i} \leq t\right) I(D_{0i} \leq K) \right| \\ &= |A_n| + |B_n|. \end{aligned}$$

We will show that $|A_n| \to 0$ and $|B_n| \to 0$ a.s..

Choose a large $K$ such that $P_{G_0}(D_0 > K) \leq \varepsilon/8$. By law of large numbers, there exists $n_1$ such that for $n \geq n_1$ implies $|\frac{1}{n} \sum_{i=1}^{n} I(D_{0i} > K) - P_{G_0}(D_0 > K)| \leq \varepsilon/8$

and

$$\begin{aligned}
|A_n| &= \left| \frac{1}{n} \sum_{i=1}^{n} [I\left(D_i \leq t\right) - I\left(D_{0i} \leq t\right)] I(D_{0i} > K) \right| \\
&\leq \frac{1}{n} \sum_{i=1}^{n} \left| I\left(D_i \leq t\right) - I\left(D_{0i} \leq t\right) \right| I(D_{0i} > K) \\
&\leq \frac{1}{n} \sum_{i=1}^{n} I(D_{0i} > K) \\
&\leq P_{G_0}(D_0 > K) + \varepsilon/8 \\
&\leq \varepsilon/8 + \varepsilon/8 = \varepsilon/4.
\end{aligned}$$

By assumption, we have from Lemma A.1 that $D_i \to D_{0i}$ a.s. for all $i$ where $D_{0i} \leq K$. Let $E_i = D_i - D_{0i}$. So, with probability 1, there exists $n_2$ such that $n \geq n_2$ implies that $-\delta \leq E_i \leq \delta$ for all $i$. Then,

$$\begin{aligned}
B_n &= \frac{1}{n} \sum_{i=1}^{n} [I\left(D_i \leq t\right) - I\left(D_{0i} \leq t\right)] I(D_{0i} \leq K) \\
&= \frac{1}{n} \sum_{i: D_{0i} \leq K} [I\left(D_i \leq t\right) - I\left(D_{0i} \leq t\right)] \\
&= \frac{1}{n} \sum_{i: D_{0i} \leq K} [I\left(D_{0i} \leq t - E_i\right) - I\left(D_{0i} \leq t\right)] \\
&\leq \frac{1}{n} \sum_{i: D_{0i} \leq K} [I\left(D_{0i} \leq t + \delta\right) - I\left(D_{0i} \leq t\right)] \\
&\leq \frac{1}{n} \sum_{i=1}^{n} [I\left(D_{0i} \leq t + \delta\right) - I\left(D_{0i} \leq t\right)].
\end{aligned}$$

Also,

$$\begin{aligned}
B_n &= \frac{1}{n} \sum_{i:D_{0i} \leq K} \left[ I\left(D_{0i} \leq t - E_i\right) - I\left(D_{0i} \leq t\right) \right] \\
&\geq \frac{1}{n} \sum_{i:D_{0i} \leq K} \left[ I\left(D_{0i} \leq t - \delta\right) - I\left(D_{0i} \leq t\right) \right] \\
&\geq \frac{1}{n} \sum_{i=1}^{n} \left[ I\left(D_{0i} \leq t - \delta\right) - I\left(D_{0i} \leq t\right) \right]
\end{aligned}$$

Now, by the Gilvenko–Cantelli Theorem, with probability one there exists $n_3$ such that $n \geq n_3$ implies that $\sup_t |\frac{1}{n}\sum_{i=1}^{n} I\left(D_{0i} \leq t + \delta\right) - G_0(t+\delta)| \leq \varepsilon/16$, $\sup_t |\frac{1}{n}\sum_{i=1}^{n} I\left(D_{0i} \leq t - \delta\right) - G_0(t-\delta)| \leq \varepsilon/16$, and $\sup_t |\frac{1}{n}\sum_{i=1}^{n} I\left(D_{0i} \leq t\right) - G_0(t)| \leq \varepsilon/16$. Also, by the uniform continuity of $G_0$, there exists $\delta > 0$ such that $|G_0(t+\delta) - G_0(t)| \leq \varepsilon/8$ and $|G_0(t-\delta) - G_0(t)| \leq \varepsilon/8$. Together,

$$\begin{gathered}
\frac{1}{n}\sum_{i=1}^{n} I\left(D_{0i} \leq t - \delta\right) - I\left(D_{0i} \leq t\right) \leq B_n \leq \frac{1}{n}\sum_{i=1}^{n} I\left(D_{0i} \leq t + \delta\right) - I\left(D_{0i} \leq t\right) \\
G_0(t-\delta) - \varepsilon/16 - G_0(t) - \varepsilon/16 \leq B_n \leq G_0(t+\delta) + \varepsilon/16 - G_0(t) + \varepsilon/16 \\
(G_0(t-\delta) - G(t)) - \varepsilon/8 \leq B_n \leq (G_0(t+\delta) - G_0(t)) + \varepsilon/8 \\
-\varepsilon/8 - \varepsilon/8 = -\varepsilon/4 \leq B_n \leq \varepsilon/8 + \varepsilon/8 = \varepsilon/4.
\end{gathered}$$

Finally, note that

$$G(t) - G_n(t) = (G(t) - G_0(t)) + (G_0(t) - G_{0n}(t)) + (G_{0n}(t) - G_n(t)).$$

Let $n_4 = \max\{n_1, n_2, n_3\}$, then $n \geq n_4$ implies

$$\begin{aligned}
\sup_{t>\eta}(G(t) - G_n(t)) &\leq \sup_{t>\eta}(G(t) - G_0(t)) + \sup_{t>\eta}(G_0(t) - G_{0n}(t)) + \sup_{t>\eta}(G_{0n}(t) - G_n(t)) \\
&\leq (\varepsilon/4 + \varepsilon/4) + \varepsilon/16 + 0 \leq \varepsilon.
\end{aligned}$$

□

## Proof of Theorem *2.1*

We need the following Lemma proved in Yohai (1985).

**Lemma A.2.** *Let $\{\boldsymbol{Z}_i\}$ be i.i.d. random vectors taking values in $\mathbb{R}^k$, with common distribution $Q$. Let $f : \mathbb{R}^k \times \mathbb{R}^h \to \mathbb{R}$ be a continuous function and assume that for some $\delta > 0$ we have that*

$$E_Q\left[\sup_{||\lambda-\lambda_0||\le\delta} |f(\boldsymbol{Z},\lambda)|\right] < \infty.$$

*Then, if $\hat{\lambda}_n \to \lambda_0$ a.s., we have*

$$\frac{1}{n}\sum_{1=1}^{n} f(\boldsymbol{Z}_i,\hat{\lambda}_n) \to E_Q\left[f(\boldsymbol{Z},\lambda_0)\right] \ \ a.s..$$

*Proof of Theorem 2.1.* Define

$$(\hat{\boldsymbol{\mu}}_{GS},\widetilde{\boldsymbol{\Sigma}}_{GS}) = \arg\min_{\boldsymbol{\mu},|\boldsymbol{\Sigma}|=1} s_{GS}(\boldsymbol{\mu},\boldsymbol{\Sigma},\hat{\boldsymbol{\Omega}}). \tag{A.1}$$

We drop out $\mathbb{X}$ and $\mathbb{U}$ in the argument to simplify the notation. Since $s_{GS}(\boldsymbol{\mu},\lambda\boldsymbol{\Sigma},\hat{\boldsymbol{\Omega}}) = s_{GS}(\boldsymbol{\mu},\boldsymbol{\Sigma},\hat{\boldsymbol{\Omega}})$, to prove Theorem 2.1 it is enough to show

**(a)**

$$(\hat{\boldsymbol{\mu}}_{Gs},\ \widetilde{\boldsymbol{\Sigma}}_{GS}) \to (\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_{00}) \text{ a.s.,} \qquad \text{and} \tag{A.2}$$

**(b)**

$$s_{GS}(\hat{\boldsymbol{\mu}}_{GS},\widetilde{\boldsymbol{\Sigma}}_{GS},\widetilde{\boldsymbol{\Sigma}}_{GS}) \to \sigma_0 \text{ a.s..} \tag{A.3}$$

Note that since we have

$$E_{H_0}\left(\rho\left(\frac{d\left(\boldsymbol{X},\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0\right)}{\sigma_0 c_p}\right)\right) = b,$$

then part (i) of Lemma 6 in the Supplemental Material of Danilov et al. (2012)

implies that given $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\varliminf_{n\to\infty} \inf_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in C_\varepsilon^C, |\boldsymbol{\Sigma}|=1} \frac{1}{n}\sum_{i=1}^{n} c_p \rho\left(\frac{d\left(\boldsymbol{X}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)}{\sigma_0 c_p\left(1+\delta\right)}\right) > (b+\delta)c_p, \tag{A.4}$$

where $C_\varepsilon$ is a neighborhood of $(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_{00})$ of radius $\varepsilon$ and if $A$ is a set, then $A^C$ denotes its complement. In addition, by part (iii) of the same Lemma we have for any $\delta > 0$,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} c_p \rho\left(\frac{d\left(\boldsymbol{X}_i, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_{00}\right)}{\sigma_0 c_p\left(1+\delta\right)}\right) < b\, c_p. \tag{A.5}$$

Let

$$Q_i(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = c_p \rho\left(\frac{d\left(\boldsymbol{X}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)}{\sigma_0 c_p(1+\delta)}\right)$$

and

$$Q_i^{(\boldsymbol{U})}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = c_{p(\boldsymbol{U}_i)} \rho\left(\frac{d^*\left(\boldsymbol{X}_i^{(\boldsymbol{U}_i)}, \boldsymbol{\mu}^{(\boldsymbol{U}_i)}, \boldsymbol{\Sigma}^{(\boldsymbol{U}_i)}\right)}{S\, c_{p(\boldsymbol{U}_i)} \left|\hat{\boldsymbol{\Omega}}^{(\boldsymbol{U}_i)}\right|^{1/p(\boldsymbol{U}_i)}}\right),$$

Now, if $|\boldsymbol{\Sigma}| = 1$ and $S = \sigma_0(1+\delta)/|\hat{\boldsymbol{\Omega}}|^{1/p}$, we have

$$\frac{1}{n}\sum_{i=1}^{n} Q_i^{(\boldsymbol{U})}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{n}\sum_{p_i=p} Q_i(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \frac{1}{n}\sum_{p_i\neq p} Q_i^{(\boldsymbol{U})}(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{A.6}$$

We also have

$$\frac{1}{n}\sum_{p_i\neq p} Q_i^{(\boldsymbol{U})}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \leq c_p(1-t_n) \tag{A.7}$$

and, therefore, by Assumption 2.4 we have

$$\lim_{n\to\infty} \sup_{\boldsymbol{\mu}, |\boldsymbol{\Sigma}|=1} \frac{1}{n}\sum_{p_i\neq p} Q_i^{(\boldsymbol{U})}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = 0 \text{ a.s.}. \tag{A.8}$$

Similarly, we can prove that

$$\lim_{n\to\infty} \sup_{\boldsymbol{\mu},|\boldsymbol{\Sigma}|=1} \frac{1}{n} \sum_{p_i \neq p} Q_i(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = 0 \text{ a.s.} \tag{A.9}$$

and

$$c_p - \frac{1}{n} \sum_{i=1}^{n} c_{p(\boldsymbol{U}_i)} \to 0, \text{ a.s..} \tag{A.10}$$

Then, from (A.4) and (A.6)–(A.10) we get

$$\varliminf_{n\to\infty} \inf_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in C_\varepsilon^C, |\boldsymbol{\Sigma}|=1} \frac{1}{n} \sum_{i=1}^{n} Q_i^{(\boldsymbol{U})}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) > (b+\delta) \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} c_{p(\boldsymbol{U}_i)} = (b+\delta) c_p \text{ a.s..} \tag{A.11}$$

Using similar arguments, from (A.5) we can prove

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} Q_i^{(\boldsymbol{U})}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_{00}) < b \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} c_{p(\boldsymbol{U}_i)} = b\, c_p \text{ a.s..} \tag{A.12}$$

Equations (A.11)–(A.12) imply that

$$\varliminf_{n\to\infty} \inf_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in C_\varepsilon^C, |\boldsymbol{\Sigma}|=1} s_{GS}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \hat{\boldsymbol{\Omega}}) > S \text{ a.s.}$$

and

$$\lim_{n\to\infty} s_{GS}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_{00}, \hat{\boldsymbol{\Omega}}) < S \text{ a.s..}$$

Therefore, with probability one there exists $n_0$ such that for $n > n_0$ we have $(\hat{\boldsymbol{\mu}}_{GS}, \widetilde{\boldsymbol{\Sigma}}_{GS}) \in C_\varepsilon^C$. Then, $(\hat{\boldsymbol{\mu}}_{GS}, \widetilde{\boldsymbol{\Sigma}}_{GS}) \to (\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_{00})$ a.s. proving (a).

Let

$$P_i(\boldsymbol{\mu}, \boldsymbol{\Sigma}, s) = c_p \rho \left( \frac{d\left(\boldsymbol{X}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)}{c_p \; s} \right)$$

and

$$P_i^{(\boldsymbol{U})}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, s) = c_{p(\boldsymbol{U}_i)} \rho \left( \frac{d\left(\boldsymbol{X}_i^{(\boldsymbol{U}_i)}, \boldsymbol{\mu}^{(\boldsymbol{U}_i)}, \boldsymbol{\Sigma}^{(\boldsymbol{U}_i)}\right)}{c_{p(\boldsymbol{U}_i)} \; s} \right).$$

Since $|\widetilde{\boldsymbol{\Sigma}}_{GS}| = 1$, we have that $s_{GS}(\hat{\boldsymbol{\mu}}_{GS}, \widetilde{\boldsymbol{\Sigma}}_{GS}, \widetilde{\boldsymbol{\Sigma}}_{GS})$ is the solution in $s$ in the following equation

$$\frac{1}{n}\sum_{i=1}^{n} P_i^{(\boldsymbol{U})}(\hat{\boldsymbol{\mu}}_{GS}, \widetilde{\boldsymbol{\Sigma}}_{GS}, s) = \frac{b}{n}\sum_{i=1}^{n} c_{p(\boldsymbol{U}_i)}. \tag{A.13}$$

Then, to prove (A.3) it is enough to show that for all $\varepsilon > 0$

$$\begin{aligned} &\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} P_i^{(\boldsymbol{U})}(\hat{\boldsymbol{\mu}}_{GS}, \widetilde{\boldsymbol{\Sigma}}_{GS}, \sigma_0+\varepsilon) < b\,c_p \text{ a.s.} \quad \text{and} \\ &\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} P_i^{(\boldsymbol{U})}(\hat{\boldsymbol{\mu}}_{GS}, \widetilde{\boldsymbol{\Sigma}}_{GS}, \sigma_0-\varepsilon) > b\,c_p \text{ a.s.} \end{aligned} \tag{A.14}$$

Using Assumption 2.4, to prove (A.14) it is enough to show

$$\begin{aligned} &\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} P_i(\hat{\boldsymbol{\mu}}_{GS}, \widetilde{\boldsymbol{\Sigma}}_{GS}, \sigma_0+\varepsilon) < b\,c_p \text{ a.s.} \quad \text{and} \\ &\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} P_i(\hat{\boldsymbol{\mu}}_{GS}, \widetilde{\boldsymbol{\Sigma}}_{GS}, \sigma_0-\varepsilon) > b\,c_p \text{ a.s.} \end{aligned} \tag{A.15}$$

It is immediate that

$$E\left(\rho\left(\frac{d\left(\boldsymbol{X}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right)}{c_p\left(\sigma_0+\varepsilon\right)}\right)\right) < E\left(\rho\left(\frac{d\left(\boldsymbol{X}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right)}{c_p\,\sigma_0}\right)\right) = b$$

and

$$E\left(\rho\left(\frac{d\left(\boldsymbol{X}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right)}{c_p\left(\sigma_0-\varepsilon\right)}\right)\right) > E\left(\rho\left(\frac{d\left(\boldsymbol{X}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right)}{c_p\,\sigma_0}\right)\right) = b.$$

Then Equation (A.15) follows from Lemma A.2 and part (a). This proves (b).

□

# Appendix B

# Supplementary material for Chapter 3

## B.1 Efficiency of GRE and tuning parameter $\alpha$ in Rocke-$\rho$ function

The tuning parameter $\alpha$ in the Rocke-$\rho$ function in $\gamma$ in (3.1) is chosen small to control the efficiency. In this chapter, we used the conventional choice $\alpha = 0.05$, as seen to achieve reasonable efficiency while achieving high robustness. Here, we explore the performance of GRE-C with smaller values of $\alpha$. We repeat the simulation study as in Section 3.4 for $p = 10, 30, 50$ and $n = 10p$. The number of replicates is $N = 30$. Table B.1 reports the finite sample efficiency and maximum average LRT distances under 20% casewise contamination. In general, higher efficiency can be achieved using smaller values of $\alpha$, but with the cost of some loss in robustness.

Table B.1: Finite sample efficiency and maximum average LRT distances for GRE-C with various values of $\alpha$. The sample size is $n = 10p$.

| $p$ | Efficiency, clean data | | | Max LRT, 20% casewise | | |
|---|---|---|---|---|---|---|
| | $\alpha = 0.05$ | $\alpha = 0.01$ | $\alpha = 0.001$ | $\alpha = 0.05$ | $\alpha = 0.01$ | $\alpha = 0.001$ |
| 10 | 0.54 | 0.67 | 0.67 | 33.1 | 32.1 | 32.1 |
| 30 | 0.58 | 0.85 | 0.95 | 16.0 | 20.2 | 28.7 |
| 50 | 0.55 | 0.58 | 0.93 | 27.1 | 28.1 | 47.7 |

Table B.2: Maximum average LRT distances. The sample size is $n = 10p$.

| $p$ | $\epsilon$ | EMVE | GSE | EMVE-C | GRE-C |
|---|---|---|---|---|---|
| 10 | 0.10 | 8.7 | 4.6 | 17.3 | 10.9 |
|  | 0.20 | 81.4 | 84.8 | 43.4 | 36.1 |
| 20 | 0.10 | 20.8 | 24.1 | 9.2 | 8.1 |
|  | 0.20 | 123.0 | 156.8 | 13.1 | 14.9 |
| 30 | 0.10 | 31.2 | 54.8 | 13.4 | 9.4 |
|  | 0.20 | 299.1 | 223.2 | 24.3 | 16.0 |
| 40 | 0.10 | 77.5 | 80.7 | 21.9 | 12.2 |
|  | 0.20 | 511.8 | 287.9 | 43.2 | 17.1 |
| 50 | 0.10 | 172.5 | 125.1 | 29.4 | 16.5 |
|  | 0.20 | 644.3 | 349.8 | 60.2 | 26.3 |

## B.2 Performance comparison between GSE and GRE

We conduct a simulation study to compare the standalone performances of the second steps (i.e. the estimation step) in the two-step S-estimators: GRE-C starting from EMVE-C versus GSE starting from EMVE.

We consider clean and casewise contaminated samples from a $N_p(\boldsymbol{\mu}_\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{0})$ distribution with $p = 10, 20, \ldots, 50$ and $n = 10p$. The simulation mechanisms are the same as that of Chapter 2, but in addition, 5% of the cells in the generated samples are randomly selected and assigned a missing value. The number of replicates is $N = 500$.

Table B.2 shows the maximum average LRT distances from the true correlation matrices among the considered contamination sizes and, for brevity, shows only the values for random correlations. EMVE is capable of dealing small fraction of outliers with 500 subsamples, but breaks down when the fraction gets larger, and brings down the performance of GSE. EMVE-C with more refined subsampling procedure and larger subsample sizes shows better performance than EMVE, even for relatively larger fraction of outliers. Overall, GRE performs better than GSE. The Rocke $\rho$ function used in GRE is capable of giving smaller weights to points that are moderate-to-large distances from the main mass of points (Rocke, 1996); see, for example,

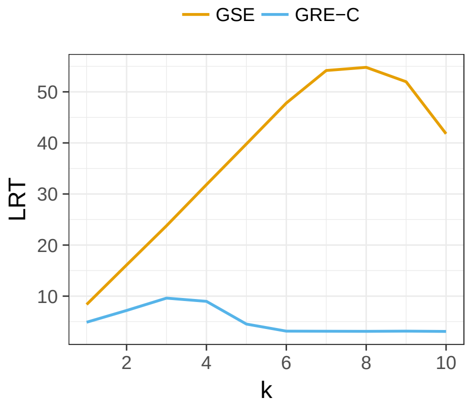


Figure B.1: Average LRT distances for various contamination sizes, $k$, for random correlations under 10% casewise contamination. The dimension is $p = 30$ and the sample size is $n = 10p$.

Table B.3: Finite sample efficiency. The sample size is $n = 10p$.

| $p$ | EMVE | GSE | EMVE-C | GRE-C |
|---|---|---|---|---|
| 10 | 0.24 | 0.89 | 0.26 | 0.54 |
| 20 | 0.30 | 0.95 | 0.30 | 0.59 |
| 30 | 0.34 | 0.98 | 0.33 | 0.58 |
| 40 | 0.35 | 0.98 | 0.34 | 0.47 |
| 50 | 0.37 | 0.99 | 0.35 | 0.48 |

Figure B.1 that shows the average LRT distance behaviors for 10% contamination for dimension 30 and sample size 300 data. In the figure, we see that GRE outperforms GSE for moderate sizes contamination points, as expected.

Table B.3 shows the finite sample relative efficiency under clean samples, taking the classical EM estimator as the baseline. As expected, GSE shows an increasing efficiency as $p$ increases. GRE, overall, has lower efficiency.

# Appendix C

# Supplementary material for Chapter 4

## C.1 Additional figures from the simulation study in Section 4.5

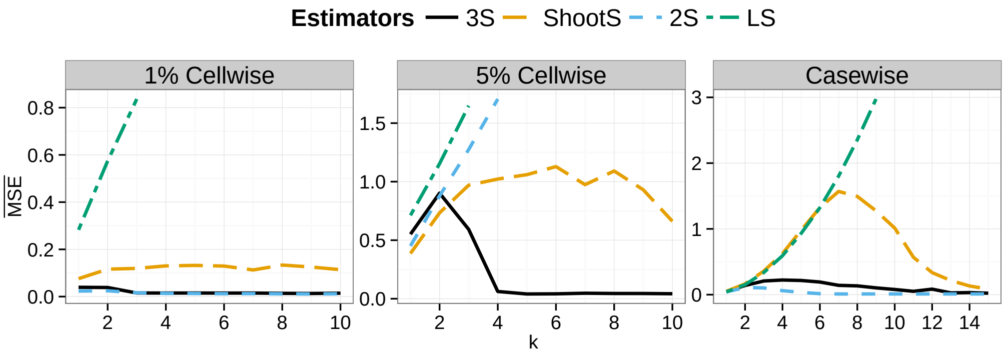


Figure C.1: $\overline{MSE}$ for various cellwise and casewise contamination values, $k$, for models with $p = 15$ continuous covariates. The sample size is $n = 150$. For details see Section 4.5.1 in the chapter.

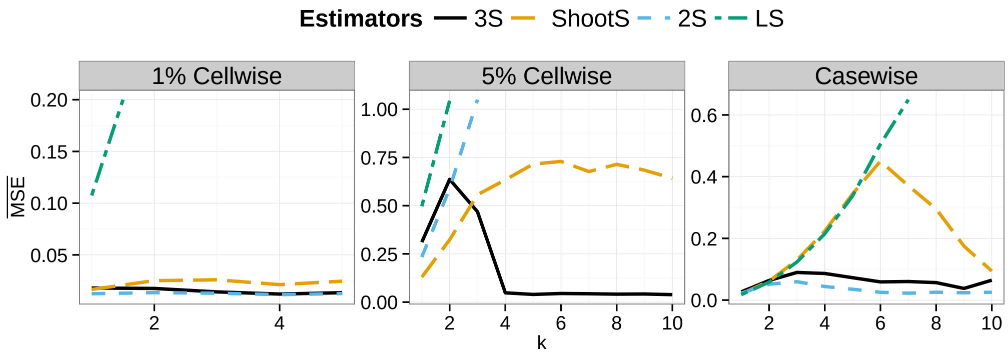


Figure C.2: $\overline{MSE}$ for various cellwise and casewise contamination values, $k$, for models with $p_x = 12$ continuous and $p_d = 3$ dummy covariates. The sample size is $n = 150$. For details see Section 4.5.2 in the chapter.

## C.2 Investigation on the performance on non-normal covariates

Here, we conduct a modest simulation study to compare the performance of 3S-regression, the shooting S-estimator, 2S-regression and the LS estimator for data with non-normal covariates.

We consider the same regression model with $p = 15$ and $n = 300$ as in Section 4.5, but the covariates are generated from a non-normal distribution as follows. The random covariates $\boldsymbol{X}_i$, $i = 1, \ldots, n$, are first generated from multivariate normal distribution $N_p(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is the randomly generated correlation matrix with a fix condition number of 100. Then, we transform the variables by doing the following:

$$(X_{i1}, X_{i2}, \ldots, X_{ip}) \leftarrow (G_1^{-1}(\Phi(X_{i1})), G_2^{-1}(\Phi(X_{i2})), \ldots, G_p^{-1}(\Phi(X_{ip}))),$$

where $\Phi(x)$ is the standard normal. We set $G_j$ as $N(0,1)$ for $j = 1, 2, 3$, $\chi^2(20)$ for $j = 4, 5, 6$, $F(90, 10)$ for $j = 7, 8, 9$, $\chi^2(1)$ for $j = 10, 11, 12$, and Pareto$(1, 3)$ for $j = 13, 14, 15$.

In the simulation study, we consider the following scenarios:

- Clean data: No further changes are done to the data;

- Cellwise contamination: Randomly replace $\epsilon = 0.05$ fraction of the cells in the covariates by outliers $X_{ij}^{cont} = k \times G_j^{-1}(0.999)$ and $\epsilon$ proportion of the responses by outliers $Y_{ij}^{cont} = E(Y_{ij}) + k \times SD(\varepsilon_i)$. We present the results for $k = 1, 5, 10$, but for larger values of $k$ we obtain similar results.

The number of replicates for each setting is $N = 1000$.

The performance of the estimator in terms of $\overline{MSE}$ are summarized in Table C.1. The performance of 3S-regression is comparable to that of LS and 2S-regression for clean data and outperforms the shooting S, LS and 2S-regression for cellwise-contaminated data, even under some deviations from the assumptions on the tail distributions of the covariates.

Table C.1: $\overline{MSE}$ for clean data and cellwise contaminated data.

| Estimators | Clean | Cellwise | | |
|---|---|---|---|---|
| | | $k = 2$ | $k = 5$ | $k = 10$ |
| 3S | 0.007 | 0.014 | 0.013 | 0.015 |
| ShootS | 0.254 | 0.839 | 1.048 | 0.882 |
| 2S | 0.003 | 4.102 | 3.851 | 4.057 |
| LS | 0.001 | 4.311 | 6.438 | 6.588 |

## C.3 Supplementary material for the Boston housing data analysis

Table C.2: Description of the variables in the Boston Housing data

| Variables | Description |
|---|---|
| medv (response) | corrected median value of owner-occupied homes in USD 1000's |
| crim | per capita crime rate by town |
| nox | nitric oxides concentration (parts per 10 million) |
| rm | average number of rooms per dwelling |
| age | proportion of owner-occupied units built prior to 1940 |
| dis | weighted distances to five Boston employment centers |
| tax | full-value property-tax rate per USD 1,000,000 |
| ptratio | pupil-teacher ratio by town |
| black | $(B - 0.63)^2$ where B is the proportion of blacks by town |
| lstat | percentage of lower status of the population |

## C.4 Proofs of lemmas and theorems

### Proof of Theorem 4.1

We need to following lemma in the proof.

**Lemma C.1.** *Let $X, X_1, \dots, X_n$ be independent with a continuous distribution function $G(x)$. Given $0 < \alpha < 1$, let $\eta = G^{-1}(1-\alpha)$ and $s = med(X - \eta | X > \eta)$. Now, consider the following estimator: $\eta_n = G_n^{-1}(1-\alpha)$ and $s_n = med(\{X_i - \eta_n | X_i > \eta_n\})$. Then, $s_n \to s$ a.s.*

*Proof.* Without loss of generality, assume that $X_1 < X_2 < \cdots < X_n$. So, $\eta_n = G_n^{-1}(1-\alpha) = X_{\lceil n(1-\alpha)\rceil}$, and

$$\#\{X_i | X_i > \eta_n\} = n - \lceil n(1-\alpha)\rceil = n - (n + \lceil -n\alpha\rceil) = \lfloor n\alpha \rfloor.$$

Then, $X_k = \text{med}(\{X_i | X_i > \eta_n\})$ where

$$\begin{aligned} k &= \lceil n(1-\alpha) \rceil + \left\lceil \frac{\lfloor n\alpha \rfloor}{2} \right\rceil \\ &= n - \lfloor n\alpha \rfloor + \left\lceil \frac{\lfloor n\alpha \rfloor}{2} \right\rceil \\ &= n - \left\lfloor \frac{\lfloor n\alpha \rfloor}{2} \right\rfloor = n - \left\lfloor \frac{n\alpha}{2} \right\rfloor . \end{aligned}$$

In other words, $\text{med}(\{X_i | X_i > \eta_n\}) = X_{\lceil n(1-\alpha/2) \rceil} = G_n^{-1}(1-\alpha/2)$, and

$$s_n = G_n^{-1}(1-\alpha/2) - \eta_n.$$

Therefore, $s_n \to s$ a.s., where $s = G^{-1}(1-\alpha/2) - \eta$.

□

## Proof of Theorem 4.1

*Proof.* Without loss of generality, we consider only the upper tail. Also, to simplify the notation, we drop out the $G$ in the probability and the $u$ that was used to distinguish between the notations for upper tail and lower tail.

Define $F(t)$ and $F_n(t)$ by

$$F(t) = \frac{P(0 < (X-\eta)/s \le t)}{P(X > \eta)} \quad \text{and} \quad F_n(t) = \frac{\frac{1}{n}\sum_{i=1}^{n} I(0 < (X_i - \eta_n)/s_n \le t)}{\frac{1}{n}\sum_{i=1}^{n} I(X_i > \eta_n)}.$$

Let $F_0(t) = 1 - e^{-t}$. It is sufficient to prove that for every $\epsilon > 0$ there exists $N$ such that for all $n \ge N$,

$$\sup_{t \ge t_0} \{F_0(t) - F_n(t)\}^+ < \epsilon.$$

Note that

$$\begin{aligned}
|F_0(t) - F_n(t)| \leq& \left| F_0(t) - \frac{P(0 < (X-\eta)/s \leq t)}{P(X \geq \eta)} \right| \\
&+ \left| \frac{P(0 < (X-\eta)/s \leq t)}{P(X > \eta)} - \frac{P(0 < (X-\eta_n)/s_n \leq t)}{P(X > \eta)} \right| \\
&+ \left| \frac{P(0 < (X-\eta_n)/s_n \leq t)}{P(X > \eta)} - \frac{P(0 < (X-\eta_n)/s_n \leq t)}{P(X > \eta_n)} \right| \\
&+ \left| \frac{P(0 < (X-\eta_n)/s_n \leq t)}{P(X > \eta_n)} - \frac{\frac{1}{n}\sum_{i=1}^n I(0 < (X_i-\eta_n)/s_n \leq t)}{P(X > \eta_n)} \right| \\
&+ \left| \frac{\frac{1}{n}\sum_{i=1}^n I(0 < (X_i-\eta_n)/s_n \leq t)}{P(X > \eta_n)} - \frac{\frac{1}{n}\sum_{i=1}^n I(0 < (X_i-\eta_n)/s_n \leq t)}{\frac{1}{n}\sum_{i=1}^n I(X_i > \eta_n)} \right| \\
&= A + B + C + D + E.
\end{aligned}$$

By Assumption 4.1, $A = 0$.

Note that

$$\begin{aligned}
B &= \frac{1}{\alpha}|P(0 < (X-\eta)/s \leq t) - P(0 < (X-\eta_n)/s_n \leq t)| \\
&= \frac{1}{\alpha}|[G(st+\eta) - G(s_n t + \eta_n)] - [G(\eta) - G(\eta_n)]|.
\end{aligned}$$

Next, we show that $\sup_t |G(st+\eta) - G(s_n t+\eta_n)| < \varepsilon\alpha/4$ and $|G(\eta) - G(\eta_n)| < \varepsilon\alpha/4$.

Given a small $\delta_0 > 0$ such that $s - \delta_0 > c$ and $\eta - \delta_0 > c$ for $c > 0$. Choose a large $K > 0$ such that for $K_{\delta_0} = (s-\delta_0)K + \eta - \delta_0$, $G(K_{\delta_0}) > 1 - \frac{\varepsilon\alpha}{4}$. First, consider $t > K$. Since $\delta_0 > 0$, we have $st + \eta > (s-\delta_0)K + (\eta - \delta_0) = K_{\delta_0}$, and therefore, $G(st+\eta) > G(K_{\delta_0})$. Also, by Lemma C.1, $s_n \to s$ a.s. and $\eta_n \to \eta$ a.s.. So, there exists $N_0$ such that $|s_n - s| < \delta_0$ and $|\eta_n - \eta| < \delta_0$ for all $n \geq N_0$. So, we have $s_n > s - \delta_0$ and $\eta_n > \eta - \delta_0$, which implies $s_n t + \eta_n > (s-\delta_0)K + (\eta-\delta_0) = K_{\delta_0}$ and $G(s_n t + \eta_n) > G(K_{\delta_0})$. Therefore,

$$\sup_{t>K}\{G(st+\eta) - G(s_n t + \eta_n)\} \leq \frac{\varepsilon\alpha}{4}.$$

Now, consider $t \leq K$. We have $|(st+\eta) - (s_n t + \eta_n)| \leq t|s - s_n| + |\eta - \eta_n| <$

$K\delta_0 + \delta_0 < \delta_1$. Now by the uniform continuity of $G$, given $\varepsilon > 0$, there exists $N_1$ such that for $n \geq N_1$, $|(st+\eta) - (s_n t + \eta_n)| < \delta$, and therefore, $|G(st+\eta) - G(s_n t + \eta_n)| < \frac{\varepsilon\alpha}{4}$. Similarly, there exists $N_2$ such that for $n \geq N_2$, $|\eta - \eta_n| < \delta$, and therefore, $|G(\eta) - G(\eta_n)| < \frac{\varepsilon\alpha}{4}$.

So, with probability one, take $N = \max\{N_0, N_1, N_2\}$ such that for $n \geq N$, it implies that $|(st+\eta) - (s_n t + \eta_n)| < \delta$ and $|\eta - \eta_n| < \delta$. Then, we have

$$\begin{aligned} B &\leq \frac{1}{\alpha}\left[\sup_t |G(st+\eta) - G(s_n t + \eta_n)| + |G(\eta) - G(\eta_n)|\right] \\ &\leq \frac{1}{\alpha}(\frac{\varepsilon\alpha}{4} + \frac{\varepsilon\alpha}{4}) = \frac{\varepsilon}{2}. \end{aligned}$$

Next, we have

$$\begin{aligned} C &\leq \frac{|G(s_n t + \eta_n) - G(\eta_n)|}{(1-G(\eta))(1-G(\eta_n))}|G(\eta) - G(\eta_n)| \\ &\leq \frac{1}{1-G(\eta)}|G(\eta) - G(\eta_n)| \leq \frac{1}{\alpha}\frac{\varepsilon\alpha}{4} = \frac{\varepsilon}{4}. \end{aligned}$$

By the Gilvenko–Cantelli Theorem, with probability one, we can show that there exists $N_3$ such that for $n \geq N_3$, $\sup_t |P(0 < (X - \eta_n)/s_n \leq t) - \frac{1}{n}\sum_{i=1}^n I(0 < (X - \eta_n)/s_n \leq t)| < \frac{\varepsilon\alpha}{16}$. Note that for large enough $n$, we have $P(X > \eta_n) > \frac{\alpha}{2}$. So,

$$D = \left| \frac{P(0 < (X-\eta_n)/s_n \leq t)}{P(X > \eta_n)} - \frac{\frac{1}{n}\sum_{i=1}^n I(0 < (X_i - \eta_n)/s_n \leq t)}{P(X > \eta_n)} \right| \leq \frac{2}{\alpha}\frac{\varepsilon\alpha}{16} = \frac{\varepsilon}{8}.$$

Next, by the Gilvanko–Cantelli Theorem again, there exists $N_4$ such that for $n \geq N_4$, $|P(X > \eta_n) - \frac{1}{n}\sum_{i=1}^n I(X_i > \eta_n)| < \sup_t |P(X > t) - \frac{1}{n}\sum_{i=1}^n I(X_i > t)| < \frac{\varepsilon\alpha}{16}$.

Then, we have

$$\begin{aligned}
E &\leq (\frac{1}{n}\sum_{i=1}^{n} I(0 < (X_i - \eta_n)/s_n \leq t))\frac{\left|P(X > \eta_n) - \frac{1}{n}\sum_{i=1}^{n} I(X_i > \eta_n)\right|}{P(X > \eta_n)\,(\frac{1}{n}\sum_{i=1}^{n} I(X_i > \eta_n))} \\
&\leq (\frac{1}{n}\sum_{i=1}^{n} I(X_i > \eta_n))\frac{\left|P(X > \eta_n) - \frac{1}{n}\sum_{i=1}^{n} I(X_i > \eta_n)\right|}{P(X > \eta_n)\,(\frac{1}{n}\sum_{i=1}^{n} I(X_i > \eta_n))} \\
&\leq \frac{2}{\alpha}\frac{\varepsilon\alpha}{16} = \frac{\varepsilon}{8}.
\end{aligned}$$

Finally, take $N = \max\{N_0, N_1, N_2, N_3, N_4\}$, we have

$$\sup_t\{F(t) - \hat{F}_n(t)\} \leq A + B + C + D + E \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{4} + \frac{\varepsilon}{8} + \frac{\varepsilon}{8} = \varepsilon.$$

□

## Proof of Theorem 4.2

*Proof.* Let $(\boldsymbol{U}_{n,1}, \ldots, \boldsymbol{U}_{n,n})^t$ be the matrix of zeros and ones with zero corresponding to a filtered component in $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^t$, and $n_0$ be the number of complete observations after the filter step. Now, let $C_j = \{i, 1 \leq i \leq n : U_{n,ij} = 0\}$ and $C = \cup_{j=1}^{p} C_j$. So, $C_j$ is the set of indices of filtered values for variable $j$, and $C$ is the set of indices of incomplete observations. By Boole's inequality,

$$n - n_0 = \#C \leq \sum_{j=1}^{p} \#C_j.$$

Let $\xi > 0$ be as described in Section 4.3. Now, for each variable $\{X_{1j}, \ldots, X_{nj}\}$, $j = 1, \ldots, p$, apply Theorem 4.1 to obtain $N_j$ such that, with probability one,

$$\#C_j \leq n\xi/p, \quad \text{for} \quad n \geq N_j.$$

Set $N = \max\{N_1, \dots, N_p\}$. Hence, with probability one,

$$n - n_0 \leq \sum_{j=1}^{p} \#C_j \leq \sum_{j=1}^{p} n\xi/p = n\xi,$$

for $n \geq N$, or equivalently,

$$\frac{n_0}{n} \geq 1 - \xi.$$

Therefore, $\boldsymbol{U}^*_{n,i} = (1, \dots, 1)^t$ according to (4.6), and $\mathbb{U}_n = \mathbb{I}$, where $\mathbb{I}$ has every entry equal to 1. In other words, for $n \geq N$, the GSE in Section 4.3 becomes

$$\begin{aligned} \boldsymbol{T}_{2S} &= \boldsymbol{T}_{GS}(\mathbb{Z}, \mathbb{I}) \\ \boldsymbol{C}_{2S} &= \boldsymbol{C}_{GS}(\mathbb{Z}, \mathbb{I}). \end{aligned}$$

Since GSE on complete data reduces to the regular S-estimator (Danilov et al., 2012), this implies that 3S-regression reduces to S-regression for $n \geq N$. □