Robust methods based on shrinkage

by

Elisa Cabana Garceran del Vall

Thesis submitted in partial compliance with the requirements for the degree of Doctor of Philosophy in Mathematical Engineering

Universidad Carlos III de Madrid

Advisor:
Rosa Elvira Lillo Rodríguez

September, 2019

To my family and friends,

"I have no special talent,
I am only passionately curious."

— Albert Einstein

Acknowledgements

These are certainly the most difficult pages from this thesis I have to write. Not because I don't know how to thank the people involved in this Ph.D. stage of my life, but because I think that I can write a whole book about it and it is hard to summarize all my gratefulness in a few pages. This has been quite a journey. It feels this is not enough, but I will try to do my best in expressing my most sincere appreciation.

First, I would like to thank the director of this piece of work, Rosa Elvira Lillo, my "work mother", who adopted me and taught me to be pragmatic, proactive, efficient and consistent. I even felt like a robust estimator. Henry Laniado contributed also to this thesis. He was the first person I asked about research in my early stages, and since he was working together with Rosa in this topic they both captured me to work with them. I am grateful for all of their advices and for encouraging me to join them.

Next, I have to thank my real mother María Eugenia for always being there, despite the $7432km$ that have separated us the last six years. Her support is what has sustained me in my worst moments. My dad is now closer than it was most of my life, for that I am truly grateful. I would like to take advantage of these lines to say what I never say out loud but I have always thought: thank you both for giving me life and for all your help so I could get here and fulfill my dreams when the situation was difficult.

When I came to Spain to do the master, it was my first trip out of the bubble that Cuba is. I had no idea I would spend so many years in this university. I had no clue that I was capable of research, creating new methods, or teaching Statistics to freshman students. Thanks to the Department, I had the possibility to share my knowledge and meet new people in several congresses and for that, I am really grateful. I also had difficulties on the way, but I consider that they made me stronger. Now all those challenges and struggles, all the never give up and I cannot surrender, are the true victory. But the most significant of my experience these years, is the friendship of my closest friends. Vero, my sister in soul, who has

been a huge support from the beginning. Alba, MJ, Rubén and Mañas, my dearest friends, without them the two years of the master would never have been the same; I will never forget all of our *shafiness* and our adventures. Fer, the one who has been by my side these last three years, holding my cries, my cravings, my fears, my worst moments, and even then, he has managed to help me create my best version of myself. I have infinite gratitude towards him and his family, for making me feel like home.

Finally, I want to acknowledge the financial support received from the Spanish Ministry of Economy and Competitiveness ECO2015-66593-P and the UC3M PIF pre-doctoral scholarship.

Published and submitted contents

Published contents:

- E. Cabana, Rosa E. Lillo, H. Laniado. Multivariate outlier detection based on a robust Mahalanobis distance with shrinkage estimators. UC3M Working Papers Statistics and Econometrics, 27-10, 2017. https://e-archivo.uc3m.es/handle/10016/24613
  – Co–author.
  – It is partially included in Chapters 2 and 3 of the thesis.
  – The material from this source included in the thesis is not indicated by typographical means or references.

- E. Cabana, Rosa E. Lillo, H. Laniado. Shrinkage reweighted regression. UC3M Working papers Statistics and Econometrics, 19-08, 2019. https://e-archivo.uc3m.es/handle/10016/28500
  – Co–author.
  – It is partially included in Chapters 4 and 5 of the thesis.
  – The material from this source included in the thesis is not indicated by typographical means or references.

- E. Cabana, Rosa E. Lillo, H. Laniado. Multivariate outlier detection based on a robust Mahalanobis distance with shrinkage estimators. ArXiv, 2019. https://arxiv.org/abs/1904.02596
  – Co–author.
  – It is partially included in Chapters 2 and 3 of the thesis.
  – The material from this source included in the thesis is not indicated by typographical means or references.

- E. Cabana, Rosa E. Lillo, H. Laniado. Robust regression based on shrinkage estimators. ArXiv, 2019. https://arxiv.org/abs/1905.02962
  – Co–author.
  – It is partially included in Chapters 4 and 5 of the thesis.
  – The material from this source included in the thesis is not indicated by typographical means or references.

Contents submitted for publication:

- E. Cabana, Rosa E. Lillo, H. Laniado. Multivariate outlier detection based on a robust Mahalanobis distance with shrinkage estimators. 2019. Statistical Papers.
  – Co–author.
  – It is partially included in Chapters 2 and 3 of the thesis.
  – The material from this source included in the thesis is not indicated by typographical means or references.

- E. Cabana, Rosa E. Lillo, H. Laniado. Robust regression based on shrinkage with application to Living Environment Deprivation. 2019. Stochastic Environmental Research and Risk Assessment.
  – Co–author.
  – It is partially included in Chapters 4 and 5 of the thesis.
  – The material from this source included in the thesis is not indicated by typographical means or references.

Abstract

In this thesis, robust methods based on the notion of *shrinkage* are proposed for outlier detection and robust regression. A collection of robust Mahalanobis distances is proposed for multivariate outlier detection. The robust intensity and scaling factors, needed to define the shrinkage of the robust estimators used in the distances, are optimally estimated. Some properties are investigated, such as the affine equivariance and the breakdown value. The performance of the proposal is illustrated through the comparison to other robust techniques from the literature, in a simulation study and with a real example of breast cancer data. The robust alternatives are also reviewed, highlighting their advantages and disadvantages. The behavior when the underlying distribution is heavy-tailed or skewed, shows the appropriateness of the proposed method when we deviate from the common assumption of normality. The resulting high true positive rates and low false positive rates in the vast majority of cases, as well as the significantly smaller computational time show the advantages of the proposal.

On the other hand, a robust estimator is proposed for the parameters that characterize the linear regression problem. It is also based on the notion of shrinkages. A thorough simulation study is conducted to investigate the efficiency with Normal and heavy-tailed errors, the robustness under contamination, the computational times, the affine equivariance and breakdown value of the regression estimator. It is compared to the classical Ordinary Least Squares (OLS) approach and the robust alternatives from the literature, which are also briefly reviewed in the thesis. Two classical data-sets often used in the literature and a real socio-economic data-set about the Living Environment Deprivation (LED) of areas in Liverpool (UK), are studied. The results from the simulations and the real data examples show the advantages of the proposed robust estimator in regression. Also, with the LED data-set it is also shown that the proposed robust regression method has improved performance than machine learning techniques previously used for this data, with the advantage of interpretability.

Furthermore, an adaptive threshold, that depends on the sample size and the dimension of the data, is introduced for the proposed robust Mahalanobis distance

based on shrinkage estimators. The cut-off is different than the classical choice of the 0.975 chi-square quantile providing a more accurate method to detect multivariate outliers. A simulation study is done to check the performance improvement of the new cut-off against the classical. The adjusted quantile shows improved performance, even when the underlying distribution is heavy-tailed or skewed. The method is illustrated using the LED data-set, and the results demonstrate the additional advantages of the adaptive threshold for the regression problem.

Resumen

En esta tesis, se proponen métodos robustos basados en la noción de shrinkage para la detección de atípicos y la regresión robusta. Se propone una colección de distancias de Mahalanobis robustas para la detección de outliers multivariantes. Los factores de intensidad y escala, necesarios para definir el shrinkage de los estimadores robustos utilizados en las distancias, se estiman de manera óptima. Se investigan algunas propiedades como la equivarianza afín y el breakdown value (valor de ruptura). El desempeño de la propuesta se ilustra mediante la comparación con otras técnicas robustas de la literatura, en un estudio de simulación y con un ejemplo real de datos de cáncer de mama. Las alternativas robustas también se revisan, destacando sus ventajas y desventajas. El comportamiento cuando la distribución subyacente es de cola pesada o asimétrica, muestra lo apropiado que es el método propuesto cuando nos apartamos de la suposición común de normalidad. Las altas tasas de verdaderos positivos y las bajas tasas de falsos positivos, en la gran mayoría de los casos, así como el tiempo de cómputo significativamente menor, muestran las ventajas de la propuesta.

Por otro lado, se introduce un estimador robusto para los parámetros que caracterizan la regresión lineal. También se basa en la noción de shrinkage. Se lleva a cabo un estudio de simulación exhaustivo para investigar la eficiencia con errores Normales y de cola pesada, la robustez bajo contaminación, los tiempos de cómputo, la equivarianza afín y el valor de ruptura del estimador de regresión. Se compara con el método Mínimos Cuadrados Ordinarios (OLS) clásico y las alternativas sólidas de la literatura, que también se revisan brevemente en la tesis. Se estudian dos conjuntos de datos clásicos que se utilizan a menudo en la literatura y un conjunto de datos socioeconómicos reales sobre la privación del entorno vital (LED) de las áreas de Liverpool (Reino Unido). Los resultados de las simulaciones y los ejemplos de datos reales muestran las ventajas del estimador robusto propuesto para regresión. Además, con el conjunto de datos LED también se muestra que el método de regresión robusta propuesto presenta mejoras con respecto a las técnicas de aprendizaje automático utilizadas anteriormente para estos datos, con la ventaja de la interpretabilidad.

Además, se introduce un recorte adaptativo, que depende del tamaño de la muestra y la dimensión de los datos, para la distancia robusta de Mahalanobis propuesta, basada en estimadores shrinkage. El valor de recorte es diferente a la opción clásica del cuantil 0.975 de la chi-cuadrado, y proporciona un método más preciso para detectar valores atípicos multivariados. Se realiza un estudio de simulación para verificar el rendimiento del nuevo punto de corte respecto al clásico. El cuantil ajustado muestra un desempeño mejorado, incluso cuando la distribución subyacente es de cola pesada o asimétrica. El método se ilustra utilizando el conjunto de datos LED y los resultados demuestran las ventajas adicionales del recorte adaptativo para el problema de regresión.

Contents

List of figures

List of tables

CHAPTER 1

Introduction

The detection of outliers in multivariate data is an important task in Statistics since that kind of observations can distort any statistical procedure. In data mining and machine learning contexts, many standard techniques such as principal component analysis and linear discriminant analysis are inherently susceptible to atypical observations (Tarr et al. [2016]). The task of detecting multivariate outliers can be useful in various fields (Vargas N [2003], Brettschneider et al. [2008], Hubert et al. [2008], Hubert and Debruyne [2010], Perrotta and Torti [2010] and Choi et al. [2016]). However, nowadays, there are several real situations from the outlier detection field, in which the data contains a large number of variables. For example, in neuroimaging, data almost surely contains rare observations due to problems like acquisition, pre-processing artifacts, or inter-subject variability. Functional Magnetic Resonance Imaging (fMRI) is a concrete example. In the analysis of fMRI data, even small movements of the head of the patients, or even the subject's heartbeat and breathing, may produce large artifacts in the signals and noise directly in the data (Wager et al. [2005], Lazar [2008], Lindquist [2008], Monti [2011], Poline and Brett [2012]). High-dimensional data are increasingly encountered in other applications of statistics, e.g., in biological and financial studies (Chen et al. [2010] and Zeng et al. [2015]), and also geochemical data (Reimann and Filzmoser [2000], Templ et al. [2008]), which practically always contains outliers.

The definition of outlier is not unique, but they are generally defined as observations resulting from a secondary process, which differs from the background distribution. This kind of data does not need to be especially high or low concerning all values of the variables in the data-set. Thus, this is the reason why the task of identifying multivariate outliers with the classical univariate methods commonly fail. In the multivariate case, there must be considered both the distance of an observation from the centroid of the data, and the shape of the data. The covariance matrix characterizes the shape of multivariate observations, and the Mahalanobis distance (MD) (see Mahalanobis [1936]) is a well-known measure which takes it into account.

The classical Mahalanobis distance is defined for every $p-$dimensional observation $\mathbf{x}_i$ of the multivariate sample $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$, as:

$$MD(\mathbf{x}_i) = \left((\mathbf{x}_i - \hat{\boldsymbol{\mu}}) \, \hat{\Sigma}^{-1} \, (\mathbf{x}_i - \hat{\boldsymbol{\mu}})^t \right)^{1/2},$$

where $\hat{\boldsymbol{\mu}}$ is the estimated multivariate location (sample mean) and $\hat{\Sigma}$ is the estimated covariance matrix (sample covariance matrix).

The problem with this definition is that the classical estimates of location and covariance matrix are often highly influenced by the presence of outliers (Rousseeuw et al. [1986], Rousseeuw and Van Zomeren [1990]). This means that a single extreme observation or groups of observations, departing from the main data structure can have a high influence on the distance measure. Two problems can arise, there might be outliers with not a large MD value, which is called a *masking problem*, and not all observations with large MD values are necessarily outliers, which is called *swamping problem* (Hadi [1992]). The problems of masking and swamping arise due to the influence of outliers on classical location and scatter estimates (sample mean and sample covariance matrix), which implies that the estimated distance will not be robust. The solution is to consider robust estimators to obtain a robust Mahalanobis distance (RMD):

$$RMD(\mathbf{x}_i) = \left((\mathbf{x}_i - \hat{\boldsymbol{\mu}}_R) \, \hat{\Sigma}_R^{-1} \, (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_R)^t \right)^{1/2}, \tag{1.1}$$

where $\hat{\boldsymbol{\mu}}_R$ and $\hat{\Sigma}_R$ are robust estimators of centrality and covariance matrix, respectively.

For multivariate normally distributed data, the distribution of the classical squared Mahalanobis distance, MD^2, is known (Gnanadesikan and Kettenring [1972]) to be chi-squared with p (the dimension of the data) degrees of freedom, i.e., χ^2_p. Then, the adopted rule for identifying the outliers is selecting the threshold as the 0.975 quantile of the χ^2_p. However, the squared RMD does not necessarily follow a chi-squared distribution when the data are not Gaussian distributed. Thus, determining exact cut-off values for outlying distances continues to be a difficult problem and has found much attention because no universally applicable method has been proposed. Despite this fact, the $\chi^2_{p;0.975}$ quantile is often considered as the threshold for recognizing outliers in the robust distance case, but this approach may have some drawbacks. Evidence of this behavior is now well documented even in moderately large samples, especially when the number of dimensions increases (Becker and Gather [1999], Hardin and Rocke [2005], Cerioli et al. [2009] and Riani et al. [2008]).

On the other hand, one special case in the multivariate space is the linear regression problem, which is widely used in numerous fields. Consider the linear regression model:

$$y_i = \alpha + \mathbf{x}_i^t \boldsymbol{\beta} + \epsilon_i \,,$$

for $i = 1, ..., n$, where n is the sample size, α is the unknown intercept, $\boldsymbol{\beta}$ is the unknown $(p \times 1)$ vector of regression parameters, the error terms ϵ_i are i.i.d and they are also independent from the p-dimensional carriers $\mathbf{x}_i$ (often also called regressor or explanatory variables).

The classical approach to estimate the parameters of the model is the Ordinary Least Squares (OLS) estimator of Gauss and Legendre, which minimizes the sum of squared residuals:

$$\hat{\boldsymbol{\beta}}_{OLS} = \underset{\boldsymbol{\beta}}{argmin} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^t \boldsymbol{\beta})^2 \,. \tag{1.2}$$

However, OLS estimator is not robust to the presence of outliers. The efficiency and breakdown point (bdp) are two traditionally used criteria to compare the existing robust methodologies. Since OLS has the smallest variance among unbiased estimates when the errors are normally distributed, and there are no outliers, in this scenario, OLS has maximum efficiency. Thus, the *relative efficiency* of the robust estimate compared to OLS when the error distribution is exactly Normal, and the data is clean, is often considered as a measure to study the performance of the methods and to compare them with each other. The bdp measures the proportion of outliers an estimate can tolerate. Usually, the definition of *finite sample bdp* is used (Donoho and Huber [1983]). Given any sample $\mathbf{z} = (\mathbf{z}_1, ..., \mathbf{z}_n)$, with $\mathbf{z}_i = (\mathbf{x}_i, y_i)$, where $\mathbf{x}_i$ is of dimension $1 \times p$, for all $i = 1, ..., n$, denote by $T(\mathbf{z})$ an estimate of the parameter $\boldsymbol{\beta}$. Let $\widetilde{\mathbf{z}}$ be the corrupted sample where any q of the original points of $\mathbf{z}$ are replaced by arbitrary outliers. Then the finite sample bdp γ^* is defined as:

$$\gamma^*(T, \mathbf{z}) = \underset{1 \le q \le n}{min} \{ \frac{q}{n} : \underset{\widetilde{z}}{sup} \, ||T(\widetilde{\mathbf{z}}) - T(\mathbf{z})|| = \infty \} \,,$$

where $|| \cdot ||$ is the Euclidean norm. The asymptotic bdp is understood as the limit of the finite sample bdp when n goes to infinity. Intuitively, the maximum possible asymptotic bdp is $1/2$ because if more than half of the observations are contaminated, it is not possible to distinguish between the background data and the contamination (Leroy and Rousseeuw [1987]). OLS has a finite sample bdp of $1/n$, i.e., the occurrence of even a single outlier can affect the results drastically. Therefore, its asymptotic bdp is 0.

OLS estimator can be alternatively expressed as follows. Denote the joint variable of the response and carriers as $\mathbf{z} = (\mathbf{x}, \mathbf{y})$. Denote the location of $\mathbf{z}$ by $\boldsymbol{\mu}$ and the scatter matrix by Σ. Partitioning $\boldsymbol{\mu}$ and Σ yields the notation:

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_x \\ \mu_y \end{pmatrix} \quad , \quad \Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix} .$$

Traditionally they are estimated by the empirical mean $\hat{\boldsymbol{\mu}}$ and the empirical covariance matrix $\hat{\Sigma}$. OLS estimators of $\boldsymbol{\beta}$ and the intercept α can be written as functions of the components of $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$, namely

$$\hat{\boldsymbol{\beta}} = \hat{\Sigma}_{xx}^{-1} \hat{\Sigma}_{xy}, \qquad \hat{\alpha} = \hat{\mu}_y - \hat{\boldsymbol{\beta}}^t \hat{\boldsymbol{\mu}}_x \,. \tag{1.3}$$

The drawback is that the classical sample estimators (sample mean and sample covariance matrix) are sensitive to the presence of outliers. Through all these past three decades there have been different approaches attempting the robustification of the procedure of finding the regression parameters, by either changing the sum of squares criteria in the definition of the OLS estimator from Equation 1.2 or using robust estimators in the analogous definition from Equation 1.3. Although no consensus establishes which method is recommended in practical situations. The diversity of data makes the estimation problem extremely difficult, because not all available methods work well for high dimension, high sample size, not all are sufficiently resistant to the presence of anomalous values, and are computationally feasible at the same time.

In summary, there are three main issues when we are dealing with multivariate data:

1. Robust outlier detection needs to be done.
2. A robust regression method is crucial in case of regression problems.
3. An accurate threshold needs to be used for the robust Mahalanobis distance.

In this thesis, we propose a solution to each of those issues. The approach is going to be based on a notion, frequently used in finance and portfolio optimization, known as *shrinkage*. It is widely used in those fields because its good performance even for large dimension p and small sample size n problems (see Couillet and McKay [2014], Chen et al. [2011] and Steland [2018]). Here, we focus on data with $n > p$. The shrinkage estimator relies on the fact that "shrinking" an estimator $\hat{E}$ of a parameter, towards a target estimator $\hat{T}$, would help to reduce the estimation error because although the shrinkage target is usually biased, it also contains less variance than the estimator $\hat{E}$. Therefore, under general conditions, there exists a *shrinkage intensity* η, so the resulting shrinkage estimator would contain less estimation error than $\hat{E}$ (James and Stein [1992]).

$$\hat{E}_{Sh} = (1 - \eta)\hat{E} + \eta\hat{T}\,. \tag{1.4}$$

The main advantage of using a shrinkage estimator is to obtain a trade-off between bias and variance. This approach can be applied to estimate both the location and dispersion. In the case of covariance matrices, the shrinkage has the additional advantage that it provides a positive definite and well-conditioned estimate, which is of crucial importance whenever we have to invert that estimate to use it in the definition of a Mahalanobis distance.

The contributions of this thesis to solve the previous list of issues are:

1. A robust outlier detection method is proposed, which uses the definition of a robust Mahalanobis distance based on the notion of shrinkage (RMD-S).
2. For linear regression, a robust approach is proposed, based on the idea of using robust estimators based on shrinkage in Equation 1.3 and weighting the observations using RMD-S, which gives place to a robust shrinkage reweighted (SR) regression estimator.

3. An adjusted quantile, which can be estimated adaptively from the data, is proposed as the threshold for the robust Mahalanobis distance based on shrinkage, giving place to a more accurate method of outlier detection: RMD-SAQ, and it can be used in the weighting step of method SR which gives an alternative method: SR-AQ.

On the other hand, we also contribute to the analysis of real data-sets of great importance.

4. One of them is an outlier detection study on the Breast Cancer Wisconsin (Diagnostic) Data-Set (WDBC), containing features from a digitized image of a breast mass.

5. The other is the robust regression study of the Living Environment Deprivation (LED) index. This measure allows studying the urban quality of life, an essential matter for environmental research, citizens, and political actions.

1.1 Structure of the thesis

The structure of the thesis is the following. First, in Chapter 2, a review of the most popular robust estimators in the literature for the definition of robust Mahalanobis distances is presented. Their properties and their drawbacks are analyzed. The reviewed methods are the Minimum Covariance Determinant (MCD) estimator, which is based on the computation of the ellipsoid with the smallest covariance determinant that would encompass at least half of the data points. The adjusted MCD ($Adj\ MCD$), which uses an adjusted quantile, instead of the classical quantile, for the RMD based on MCD. *Kurtosis* method, based on the analysis of the projections of the sample points onto a certain set of directions obtained by maximizing and minimizing the kurtosis coefficient of the projections, and some random directions generated by a stratified sampling scheme. The Orthogonalized Gnanadesikan-Kettenring (OGK) estimator and the Comedian method (COM).

Then, in Chapter 3, a collection of robust Mahalanobis distances based on the notion of shrinkage are proposed for the outlier detection problem. The approaches are studied through simulations and compared to the robust alternatives from the literature. Simulations were performed with Normal data, and also with heavy-tailed and skewed distributed data, to study the case in which we deviate from the normality assumption. From these studies, the proposed RMD with the best performance is selected, and called RMD-S. Some properties are studied for RMD-S: the affine equivariance, the breakdown value, and the performance under correlations. It is shown that the proposed procedure has an advantageous behavior in all the simulation results, especially when dimension increases. Finally, a real data-set example about the Breast Cancer Wisconsin (Diagnostic) Data, illustrates that the proposed method works well in practice and requires reasonable computational times, even for large problems.

Chapter 4 summarizes the state-of-the-art about robust regression in the literature, their properties, their advantages, and disadvantages. The reviewed methods

are: M-estimation, MM-estimation, Generalized M-estimation (GM), R-estimate, S-estimation, Generalized S-estimation (GS), Least Absolute Deviation (LAD) regression, Least Median of Squares (LMS) regression, Least Trimmed Squares (LTS) regression, Covariance approach and the "robust and efficient weighted least square" estimator (REWLSE).

Chapter 5 introduces the proposed robust regression approach called shrinkage reweighted (SR) regression estimator. The performance of SR is compared to the classical OLS and the other existing robust alternative methods. The advantages of using the shrinkage are shown in the simulation study. SR approach yields competitive results compared to the alternatives from the literature for the regression problem, even in high dimension, heavy-tailed distributed errors, large contamination or transformed data. Furthermore, SR is quite stable computationally. Finally, the results with the real data-set examples bear out with the conclusions from the simulation study. Especially with the Living Environment Deprivation (LED) index example, where SR approach provides an improvement with respect to classical OLS and machine learning techniques RF and GBR while maintaining the advantage of interpretability.

In Chapter 6, an adjusted quantile is proposed as the threshold for the robust distance RMD-S introduced in Chapter 3, because the latter uses the classical chi-squared quantile as the cut-off value for detecting outliers in multivariate data. The adaptive approach RMD-SAQ was studied by means of simulations that show the efficiency improvement, even when the underlying distribution is heavy-tailed or skewed, evidencing the advantages of the adjusted quantile even when we deviate from the common assumption of normality. On the other hand, the overall improvement in performance is reflected in the rest of simulation scenarios, when the adaptive threshold is considered. Finally, the LED index example is studied to investigate if the estimated model can be improved with the introduction of the adjusted quantile, which is referred to as method SR-AQ. In summary, the use of the adaptive threshold provides advantages in robust outlier detection and robust regression.

Finally, Chapter 7 provides general conclusions and the proposed continuity of the research lines for future work.

CHAPTER 2

Robust outlier detection

In multivariate data, the presence of outliers is of crucial importance. The robust Mahalanobis distance (RMD) is commonly used for detecting multivariate outliers, because the classical version uses the sample estimators, which are sensitive to the presence of atypical values. The definition for an RMD is not unique because several robust estimators of location and covariance matrix from the literature can be used to define it (Equation 1.1). In this chapter, a review is made of some of the most used robust estimators for this task.

2.1 Minimum Covariance Determinant (MCD)

The MCD estimator was proposed by Rousseeuw [1985], and it consists on determining the subset H of observations of size h which minimizes the determinant of the sample covariance matrix, computed from only these h points. The choice of h determines the robustness of the estimator, in fact, it is a compromise between robustness and efficiency. The breakdown value of the MCD estimator is $(n-h)/n$ approximately. Thus, $h = 0.75n$ gives a breakdown value of approximately 25%. Once this subset of size h is found, it is possible to estimate the centrality ($\hat{\boldsymbol{\mu}}_{MCD}$) and the covariance matrix ($\hat{\Sigma}_{MCD}$), based only upon that subset, and they will be robust estimates.

$$
\begin{aligned}
H = \{&\text{set of } h \text{ points} : |\hat{\Sigma}_H| \leq |\hat{\Sigma}_K|, \\
&\text{for all subsets K s.t. } \#K = h\} \\
\hat{\boldsymbol{\mu}}_{MCD} &= \frac{1}{h}\sum_{i\in H}\mathbf{x}_i \\
\hat{\Sigma}_{MCD} &= \frac{1}{h}\sum_{i\in H}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{MCD})(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{MCD})^t\,,
\end{aligned}
$$

where $|A|$ denotes the determinant of the matrix A, and $\#K$ denotes the cardinality of the subset K.

Using the MCD robust estimators in the definition of the Mahalanobis distance gives place to a robust measure.

$$RMD_{MCD}(\mathbf{x}_i) = \left((\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{MCD}) \, \hat{\Sigma}^{-1}_{MCD} \, (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{MCD})^t \right)^{1/2} . \tag{2.1}$$

The rule in this approach for detecting outliers is usually based on the classical threshold $c = \chi^2_{p;0.975}$, i.e., the 0.975 quantile of the χ^2 distribution with p degrees of freedom. When the distance of an observation $\mathbf{x}_k$ is higher than the cut-off, $RMD_{MCD}(\mathbf{x}_k) > c$, the observation is declared as an outlier.

Figure 2.1 shows an example of the difference between considering robust and non-robust Mahalanobis distance to detect outliers. The observations are from the Hertzsprung-Russell Diagram of the Star Cluster CYG OB1 (Leroy and Rousseeuw [1987]). It contains 47 stars data. In the figure, there are the 97.5% tolerance ellipses corresponding to the classical and a robust Mahalanobis distance. It is obvious how the presence of outliers influences the ellipsoid corresponding to the threshold for the classical Mahalanobis distance, masking the outliers 7, 9 and 14. Meanwhile, the robust distance correctly detects the atypical observations.

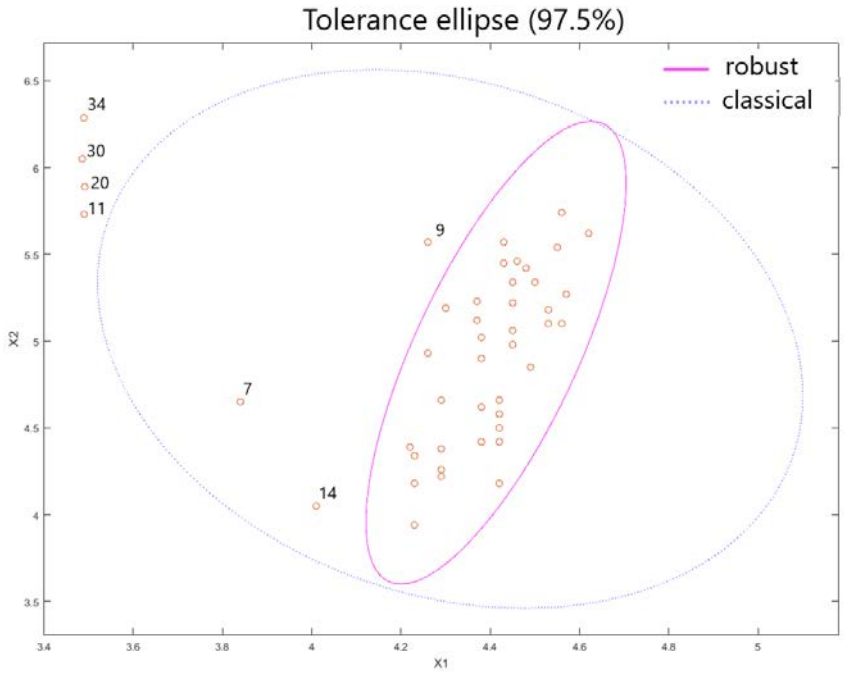

Figure 2.1: Star data with 97.5% tolerance ellipses corresponding to MD and RMD.

The procedure to find the MCD estimates required naive subsampling for minimizing the objective function, but an improvement much more effective, the Fast-MCD, was introduced by Rousseeuw and Driessen [1999] and the code is available in Matlab (Verboven and Hubert [2005]). Unfortunately, Fast-MCD still requires substantial running times for large dimension p, because the number of candidate solutions grows exponentially with the dimension of the sample and, as a consequence, the procedure becomes computationally expensive for even moderately sized problems.

2.2 Adjusted Minimum Covariance Determinant (Adj MCD)

For the RMD based on MCD estimators, the classical quantile $\chi^2_{p;0.975}$ is often used as the threshold to detect outliers. The problem is that fixing this threshold value is rather subjective in the robust distance case because there is no demonstration of the true distribution of the squared robust Mahalanobis distance. Furthermore, there is no reason why this fixed threshold should be appropriate for every data-set. The cut-off value should be adjusted to the sample size (Reimann et al. [2005]). On the other hand, if the data is clean and the observations come from a single multivariate Normal distribution, there are no outliers, no observations coming from a different distribution, there are only extremes. In this case, the threshold should be infinity.

Since the squared RMD does not necessarily follow a chi-squared distribution, the problem of the selection of the cut-off value continues to be of crucial importance, because there is no consensus. Filzmoser et al. [2005] proposed to use an adjusted quantile, instead of the classical choice. The adjusted threshold is estimated adaptively from the data, but their proposal is defined for a specific robust Mahalanobis distance, the one based on the MCD estimator.

The idea was based on measuring the difference between the empirical distribution of the squared robust distances and the distribution considered in theory, the chi-squared. Consider a sample $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$ of dimension p. Let $G(u)$ be the distribution function of χ^2_p and let $G_n(u)$ denote the empirical distribution function of the squared robust Mahalanobis distance $RMD_{MCD}(\mathbf{x}_i)$ from Equation 2.1.

For multivariate normally distributed samples, G_n converges to G. Therefore, the next step is to compare the tails of G_n and G in order to detect outliers. The maximum possible positive difference between the two distributions is defined as $p_n(\delta)$, where $\delta = \chi^2_{p;0.98}$ is the quantile that define the tails. In case of clean multivariate normally distributed background data, the threshold should be infinity and no observation should be declared as an outlier. In this case, observations with a large RMD should be seen as extremes of the distribution. Therefore, it is necessary to consider a critical value p_{crit}, which will help to distinguish between outliers and extremes, if the departure in the tails between G_n and G is higher enough. The author derived the equation for the critical value p_{crit} by simulations and defined a measure of outliers in the sample as:

$$\alpha_n(\delta) = \begin{cases} 0, & \text{if } p_n(\delta) \leq p_{crit} \\ p_n(\delta), & \text{if } p_n(\delta) > p_{crit} \end{cases} .$$

Then, in case of no contamination, i.e., no outliers, the maximum difference between the empirical and the distribution considered in theory, the chi-squared, should not be greater than the p_{crit} value. On the other hand, when the difference (in the tail) between the two distributions is big enough (greater than the p_{crit} value), then the $p_n(\delta)$ should be selected as the α value for calculating the threshold $c_n(\delta)$, which is determined as:

$$c_n(\delta) = G_n^{-1}(1 - \alpha_n(\delta)) \,. \tag{2.2}$$

Let us call this method Adj MCD. The idea of Filzmoser is an improved manner of estimating the threshold adaptively from the data. This procedure can be applied to any robust distance, other than the robust Mahalanobis distance based on the MCD estimator, the one used by Filzmoser. The only difference is to properly estimate the equations for the p_{crit} value based on the distance measure selected. The advantage is that the cut-off is adaptively estimated from the data and it improves the false positive rates, while maintaining the same true positive rates, except in some cases where the true positive rates can also be slightly declining.

2.3 Kurtosis

Another approach is the one proposed by Peña and Prieto [2001] and Peña and Prieto [2007], which is based on the idea that high or low values of the kurtosis coefficient suggest the presence of outliers. The authors take the projections of the sample points onto the set of directions obtained by maximizing and minimizing the Kurtosis coefficient, and they also consider a set of random directions generated by a stratified sampling scheme. The authors proposed to project the "n" cloud of points in $\mathbb{R}^p$ over two new $p-$dimensional spaces: the first one obtained with the maximum kurtosis orthogonal direction, and the second one with the minimum kurtosis orthogonal direction, and also over a set of random directions. After obtaining the whole set of directions, the next step is to determine a "measure of outlyingness" for each observation (actually for their univariate projections $z_i^{(j)}$) as:

$$r_i = max_{1 \leq j \leq d} \frac{|z_i^{(j)} - median(\mathbf{z}^{(j)})|}{MAD(\mathbf{z}^{(j)})},$$

where d is the total number of directions in which the data are projected, the univariate projections are $\mathbf{z}^{(j)} = (z_1^{(j)}, ..., z_n^{(j)})$, $median$ is the univariate median and MAD denotes the Median Absolute Deviation (Gauss [1816], Rousseeuw and Croux [1993], Leys et al. [2013]), which is a robust measure of the variability of a univariate sample and it is defined as the median of the absolute deviations from the data's median:

$$MAD(\mathbf{z}^{(j)}) = median\left(\left| z_i^{(j)} - median(\mathbf{z}^{(j)}) \right| \right) .$$

With the above measure r_i, a given observation is considered as an outlier if the condition r_i being greater than a certain cut-off value holds. If the condition holds for some i, a new sample composed of all observations whose r_i is less than the cut-off value is formed, and the procedure is applied again to the reduced sample. This is repeated until either no additional observations satisfy that their r_i is greater than the cut-off value, or the number of remaining observations is less than $\lfloor (n+p+1)/2 \rfloor$. Finally, a Mahalanobis distance is computed for all observations labeled as outliers in the preceding steps, using the mean and the covariance estimator based upon the remaining observations. Let U be the set of observations not labeled as outliers by the method, then the estimates of location $\hat{\boldsymbol{\mu}}_K$ and covariance matrix $\hat{\Sigma}_K$ (where the

subscript K stands as a notation for "Kurtosis"), based upon this subset U defines a robust Mahalanobis distance as:

$$RMD_K(\mathbf{x}_i) = \left((\mathbf{x}_i - \hat{\boldsymbol{\mu}}_K) \, \hat{\Sigma}_K^{-1} \, (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_K)^t \right)^{1/2} .$$

The final step is using this Mahalanobis distance to recover observations "mislabeled" as outliers, i.e., if the observation $i \notin U$ has $RMD_K(\mathbf{x}_i) < \chi^2_{p;0.99}$, then $\mathbf{x}_i$ is included in U. The process is repeated until no more such observations are found or U becomes the set of all observations.

This method is a powerful approach for robust estimation and outlier detection. However, when the dimension p of the sample space increases, the method worsens its performance, and in the presence of correlation between the variables, the method loses power (Marcano and Fermin [2013]). On the other hand, it is not very efficient computationally because of the optimization problem associated with the computation of the directions.

2.4 Orthogonalized Gnanadesikan-Kettenring (OGK)

Maronna and Zamar [2002] proposed the Orthogonalized Gnanadesikan-Kettenring (OGK) estimator. It was the result of applying a general method to a pairwise robust scatter matrix that may be non-positive definite, in order to obtain a positive-definite matrix. The method was applied to the robust covariance estimator from Gnanadesikan and Kettenring [1972], which calculated a robust covariance estimate for two variables X and Y based on the following identity.

$$cov(X, Y) = \frac{1}{4} \left(\sigma(X + Y)^2 - \sigma(X - Y)^2 \right) .$$

where σ is a robust estimate of the standard deviation. The drawback is that these pairwise estimates will not necessarily be positive definite. So, Maronna and Zamar [2002] propose an eigen-decomposition based procedure to obtain positive-definiteness. The variables in an eigenvector space are orthogonal, which means the covariances are zero and it is sufficient to obtain robust variance estimates of the data projected onto each eigenvector direction. In OGK procedure, the eigenvalues are replaced with these robust variances, and the eigenvector transformation is applied in reverse to yield a positive semi-definite robust covariance matrix. OGK estimate is scale invariant of the original data matrix is robustly scaled, i.e., each component is divided by its robust variance. The authors stated that the procedure could be iterated, although it is not always better. They also find that using weighted estimates may improve the performance, in which case the observations are weighted according to their robust distances:

$$RMD_{OGK}(\mathbf{x}_i) = \left((\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{OGK}) \, \hat{\Sigma}_{OGK}^{-1} \, (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{OGK})^t \right)^{1/2} ,$$

with $\hat{\boldsymbol{\mu}}_{OGK}$ and $\hat{\Sigma}_{OGK}$ the robust OGK estimates. They use hard rejection weights, of the form $I(RMD_{OGK} < c)$, where $I(\cdot)$ is the indicator function and c is the threshold value, which results from:

$$c = \frac{\chi^2_{p;\beta}\ med(RMD_{OGK_1}, ..., RMD_{OGK_n})}{\chi^2_{p;0.5}}\,,$$

where *med* denotes the median, and $\chi^2_{p;\beta}$ is the β-quantile of the χ^2_p distribution. The observations have full weight unless their robust distance is greater than c, in which case they will have zero weight.

2.5 Comedian

Sajesh and Srinivasan [2012] proposed a method, called the Comedian method to detect outliers from multivariate data based on the *comedian matrix* estimator from Falk [1997], which is also a robust estimate of scatter but it can be non-positive semi-definite. With the Comedian method, a positive definite scatter matrix can be obtained. The idea is based on the concept of comedian between two random variables X and Y, which is defined as:

$$COM(X, Y) = med((X - med(X))(Y - med(Y)))\,. \tag{2.3}$$

The *comedian* is a robust measure of dependence between X and Y. Based on the median concept as a robust measure of location, there is a robust measure of dispersion for a random variable X, which is the *Median Absolute Deviation (MAD)* from the data's median:

$$MAD(X) = median(|X - median(X)|)\,.$$

A comedian matrix can be defined based on a multivariate version of (2.3). Let $\mathbf{x} = \{\mathbf{x}_1, ..., \mathbf{x}_p\}$ be the $n \times p$ data matrix with n being the sample size and p the number of variables. Then the comedian matrix is defined as:

$$COM(\mathbf{x}) = (\ COM(\mathbf{x}_j, \mathbf{x}_t)\ ) \quad j, t = 1, ..., p\,. \tag{2.4}$$

Sajesh and Srinivasan [2012] also defined the *correlation median* matrix, based on the comedian matrix:

$$\delta(X) = DCOM(X)D^t\,,$$

where D is a diagonal matrix with diagonal elements $1/MAD(\mathbf{x}_i)$, $i = 1, ..., p$. Then, they adopted some transformations based on the eigenspace of the correlation median matrix and projections of the data, to overcome the non-positive semi-definiteness of the comedian matrix and to obtain robust estimates for location $\hat{\boldsymbol{\mu}}_{COM}$ and scatter $\hat{\Sigma}_{COM}$. The authors claim that the estimates can be improved through an iterative process, by replacing in the first step δ by the estimated $\hat{\Sigma}_{COM}$ and repeat the other steps. For the outlier detection problem, a robust Mahalanobis distance can be defined.

$$RMD_{COM}(\mathbf{x}_i) = \left((\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{COM})\,\hat{\Sigma}^{-1}_{COM}\,(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{COM})^t\right)^{1/2}\,.$$

They defined the threshold value to detect outliers as

$$c = 1.4826 \frac{\chi^2_{p;0.95}\ med(RMD_{COM_1}, ..., RMD_{COM_n})}{\chi^2_{p;0.5}}.$$

Then, if any $RMD_{COM}(\mathbf{x}_i) > c$, the corresponding observation $\mathbf{x}_i$ is labeled as an outlier. By using this cut-off value and the robust Mahalanobis distance, a weight function can be defined and robust estimates for location and scatter can be obtained. The authors proved that these estimates are positive definite and approximately affine equivariant. They also study the breakdown value through simulations and the method showed good performance. Another conclusion of their work was that the efficiency of the method increases with the increase in dimension p, as examined through various numerical studies.

2.6 Summary

Through this chapter, a review of some of the most used robust estimators of location and covariance matrix is done. Rousseeuw [1985] proposed the MCD estimator which has good properties but becomes computationally expensive for even moderately sized problems. On the other hand, Filzmoser et al. [2005] proposed to use an adjusted quantile for this particular RMD definition (Adj MCD), estimated adaptively from the data. In general, the advantages over MCD with classical quantile, are that it holds the same properties but the false positive rate gets decreased, especially when there are no outliers in the data-set and the observations are generated from a Normal distribution. Peña and Prieto [2001] and Peña and Prieto [2007] proposed the Kurtosis approach based on the idea that maximizing and minimizing the kurtosis coefficient is an indicator of the presence of outliers. It is a powerful procedure for robust estimation and outlier detection. However, it has some drawbacks when the dimension p of the sample space grows, it is not very efficient computationally and in the presence of correlation between the variables, the method loses power. Maronna and Zamar [2002] proposed the OGK estimator, applying a general method to the pairwise robust scatter matrix from Gnanadesikan and Kettenring [1972]. With this procedure, a positive-definite scatter matrix can be obtained, which is of great importance when it is used in the Mahalanobis distance since inversion of the covariance matrix is done. Sajesh and Srinivasan [2012] proposed the Comedian method (COM) to detect outliers from multivariate data based on the comedian matrix estimator from Falk [1997]. Stated by their authors, OGK and Comedian method seems to have good performance for high dimension and good properties like high efficiency and approximate affine equivariance.

CHAPTER 3

Robust outlier detection based on shrinkage

In this section, a collection of RMD's is proposed for outlier detection, especially in high dimension. They are based on considering different combinations of robust estimators of location and covariance matrix. Two basic options are considered for the location parameter: a component-wise median and the L_1 multivariate median (Gower [1974], Brown [1983], Dodge [1987], Small [1990]). The notion of *shrinkage* (Ledoit and Wolf [2003a], Ledoit and Wolf [2003b], Ledoit and Wolf [2004], DeMiguel et al. [2013]) described in Chapter 1, is considered. Recall the shrinkage definition, from Equation 1.4, which is based on the fact that shrinking a sample estimator towards a target estimator would help to reduce the estimation error. The shrinkage can be applied to both location and dispersion estimates, obtaining different combinations to define robust Mahalanobis distances. In the case of covariance matrices, the shrinkage provides positive definite and well-conditioned estimates, which is an additional advantage when the matrix needs to be inverted in the definition of an RMD. As for the covariance matrix, the proposed estimates consists on a shrinkage estimator over special cases of *comedian matrices* (Hall and Welsh [1985], Falk [1997]), as the sample estimator to base the shrinking on. The comedian matrix is a robust estimator of scatter, and its definition is stated in Equation 2.4, Section 2.5, Chapter 2, where it is also described in terms of its properties. The special cases of comedian matrices that are proposed are based upon a location parameter, which will be estimated using the robust estimator of centrality (or its shrinkage), in a way that an RMD can be obtained with meaningful combinations of both location and covariance matrix estimators. In this chapter, we analyze the best option for shrinking the location and the scale. Through a simulation study, the satisfactory practical performance is shown, especially when the dimension of the problem grows. The computational cost is studied by both simulations and a real data-set example.

3.1 Location parameter

Let $\mathbf{x} = \{\mathbf{x}_1, ..., \mathbf{x}_p\}$ be the $n \times p$ data matrix with n being the sample size and p the number of variables. Based on the fact that the *median* is a better choice in terms

of robustness, we start by considering as a location estimator the *component-wise median*:

$$\hat{\boldsymbol{\mu}}_{CCM} = (median(\mathbf{x}_1), ..., median(\mathbf{x}_p)), \tag{3.1}$$

where *median* denotes the univariate median and $(\mathbf{x}_j) = (x_{1j}, ..., x_{nj})^T$ for all $j = 1, ..., p$ is the j-th column of $\mathbf{x}$.

Another option is to consider a multivariate median $\hat{\boldsymbol{\mu}}_{MM}$ called $L_1-median$ which is a robust and highly efficient estimator of central tendency (Lopuhaa and Rousseeuw [1991], Vardi and Zhang [2002], Hubert [2011]). It is defined as:

$$\hat{\boldsymbol{\mu}}_{MM} = argmin_{\mathbf{x}_m,\ m\in\{1,...,n\}} \frac{1}{n}\sum_{i=1}^{n} ||\mathbf{x}_m - \mathbf{x}_i||_1 \,. \tag{3.2}$$

DeMiguel et al. [2013] proposed a shrinkage estimator over the sample mean, towards a scaled vector of ones as the target. In the same way we propose to study shrinkage estimators for both (3.1) and (3.2). Consider $\nu_{\boldsymbol{\mu}}\mathbf{e}$ as the target estimator, where $\mathbf{e}$ is the $p-$dimensional vector of ones, and consider $\hat{\boldsymbol{\mu}}_{CCM}$ as the sample estimator $\hat{E}$. Then, the shrinkage estimator over the component-wise median is:

$$\hat{\boldsymbol{\mu}}_{Sh(CCM)} = (1-\eta)\hat{\boldsymbol{\mu}}_{CCM} + \eta\nu_{\boldsymbol{\mu}}\mathbf{e}\,.$$

The scaling factor $\nu_{\boldsymbol{\mu}}$ and the intensity η should minimize the expected quadratic loss, that is:

$$\begin{array}{cc} \min_{\nu_{\boldsymbol{\mu}},\eta} & E\left[\left\|\hat{\boldsymbol{\mu}}_{Sh(CCM)} - \boldsymbol{\mu}\right\|_2^2\right] \\ \text{s.t.} & \hat{\boldsymbol{\mu}}_{Sh(CCM)} = (1-\eta)\hat{\boldsymbol{\mu}}_{CCM} + \eta\nu_{\boldsymbol{\mu}}\mathbf{e}, \end{array} \tag{3.3}$$

where $\|\mathbf{x}\|_2^2 = \sum_{j=1}^{p} x_j^2$.

Proposition 1 *The solution of the problem in (3.3) is:*

$$\hat{\nu}_{\boldsymbol{\mu}} = \frac{\hat{\boldsymbol{\mu}}_{CCM}\mathbf{e}}{p}, \quad \hat{\eta} = \frac{E\left[\|\hat{\boldsymbol{\mu}}_{CCM} - \boldsymbol{\mu}\|_2^2\right]}{E\left[\|\hat{\boldsymbol{\mu}}_{CCM} - \hat{\nu}_{\boldsymbol{\mu}}\mathbf{e}\|_2^2\right]}\,. \tag{3.4}$$

See the proof in Appendix A.1. Note that the denominator in the above expression (3.4) is estimable, but the numerator is not straightforward because $\boldsymbol{\mu}$ is unknown. Then, it is necessary to provide another expression for the numerator. Chu [1955] investigated the distribution for the sample median estimator and obtained the following result about the variance in presence of normality. Fix j, for $j \in \{1, ..., p\}$:

$$\sigma^2_{\hat{\boldsymbol{\mu}}_{CCM_j}} = Var(\hat{\boldsymbol{\mu}}_{CCM_j}) = \frac{\pi}{2n}\sigma^2_{\mathbf{x_j}}\,.$$

Therefore, the numerator in the expression (3.4) for determining the $\hat{\eta}$ in Proposition 1 is:

$$E\left[\|\hat{\boldsymbol{\mu}}_{CCM}-\boldsymbol{\mu}\|_2^2\right]=E\left[\sum_{j=1}^{p}(\hat{\boldsymbol{\mu}}_{CCMj}-\boldsymbol{\mu}_j)^2\right]$$
$$=\sum_{j=1}^{p}\sigma^2_{\hat{\boldsymbol{\mu}}_{CCMj}}=\frac{\pi}{2n}\sum_{j=1}^{p}\sigma^2_{\mathbf{x}_j}\,. \tag{3.5}$$

We need to estimate $\sigma^2_{\mathbf{x}_j}$ robustly, and we will do so as explained in the next Section 3.2, with property (3.10). The estimate for $\hat{\eta}$ in expression (3.4) can be calculated as stated in Equation 3.11 in next Section 3.2.

On the other hand, consider $\nu_{\boldsymbol{\mu}}\mathbf{e}$ again as the target estimator and consider $\hat{\boldsymbol{\mu}}_{MM}$ as the sample estimator. Then, the shrinkage estimator over the multivariate L_1-median is:

$$\hat{\boldsymbol{\mu}}_{Sh(MM)}=(1-\eta)\hat{\boldsymbol{\mu}}_{MM}+\eta\nu_{\boldsymbol{\mu}}\mathbf{e}\,.$$

The scaling factor $\nu_{\boldsymbol{\mu}}$ and the intensity η should minimize the expected quadratic loss:

$$\begin{array}{cc}\min_{\nu_{\boldsymbol{\mu}},\eta} & E\left[\left\|\hat{\boldsymbol{\mu}}_{Sh(MM)}-\boldsymbol{\mu}\right\|_2^2\right]\\ \text{s.t.} & \hat{\boldsymbol{\mu}}_{Sh(MM)}=(1-\eta)\hat{\boldsymbol{\mu}}_{MM}+\eta\nu_{\boldsymbol{\mu}}\mathbf{e},\end{array} \tag{3.6}$$

where $\|\mathbf{x}\|_2^2=\sum_{j=1}^{p}x_j^2$.

Proposition 2 *The solution of the problem in (3.6) is:*

$$\hat{\nu}_{\boldsymbol{\mu}}=\frac{\hat{\boldsymbol{\mu}}_{MM}\mathbf{e}}{p},\quad \hat{\eta}=\frac{E\left[\|\hat{\boldsymbol{\mu}}_{MM}-\boldsymbol{\mu}\|^2\right]}{E\left[\|\hat{\boldsymbol{\mu}}_{MM}-\hat{\nu}_{\boldsymbol{\mu}}\mathbf{e}\|^2\right]}\,. \tag{3.7}$$

The proof is Appendix A.2. As in the previous case, the denominator in the η expression (3.7) can be described as:

$$E\left[\|\hat{\boldsymbol{\mu}}_{MM}-\boldsymbol{\mu}\|_2^2\right]=E\left[\sum_{j=1}^{p}(\hat{\boldsymbol{\mu}}_{MMj}-\boldsymbol{\mu}_j)^2\right]=\sum_{j=1}^{p}\sigma^2_{\hat{\boldsymbol{\mu}}_{MMj}}\,.$$

Bose and Chaudhuri [1993], Bose [1995] and Möttönen et al. [2010] investigated the asymptotic distribution for the L_1-median. In page 184, section 3, from Möttönen et al. [2010], the authors describe the necessity of the following two assumptions, for $\mathbf{x}$ a $p-$variate random vector with cdf F, density function f and $p>1$:

- (C1) The $p-$variate density function of $\mathbf{x}$ is continuous and bounded.
- (C2) The spatial median of the distribution of $\mathbf{x}$ is zero and unique.

According to Theorem 2, page 185, section 3 in Möttönen et al. [2010], under assumptions C1 and C2, $\sqrt{n}\hat{\boldsymbol{\mu}}_{MM} \rightarrow_d N_p(\mathbf{0}, A^{-1}BA^{-1})$, where $\hat{\boldsymbol{\mu}}_{MM}$ is the observed spatial median, and A and B are the following:

$$A(\mathbf{x}) = \frac{1}{||\mathbf{x}||}\left[I_p - \frac{\mathbf{x}\mathbf{x}^t}{||\mathbf{x}||^2}\right] \qquad B(\mathbf{x}) = \frac{\mathbf{x}\mathbf{x}^t}{||\mathbf{x}||^2}\,.$$

In section 4, page 185 from Möttönen et al. [2010], the authors also provide an estimation for the asymptotic covariance matrix $A^{-1}BA^{-1}$ of the spatial median. They are assuming the true value $\hat{\boldsymbol{\mu}}_{MM} = \mathbf{0}$ is zero (condition C2). Then they write $\hat{A} = ave\{A(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{MM})\}$ and $\hat{B} = ave\{B(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{MM})\}$ and prove that under C1 and C2: $\hat{A} \rightarrow_P A$ and $\hat{B} \rightarrow_P B$, which means that the estimators converge in probability to the population values A and B, respectively. This result is Theorem 3, section 4, page 185 from Möttönen et al. [2010]. According to the authors (stated in page 186), Theorems 2 and 3 suggest that the distribution of $\hat{\boldsymbol{\mu}}_{MM}$ can be approximated by $N_p\left(\boldsymbol{\mu}, \frac{1}{n}\hat{A}^{-1}\hat{B}\hat{A}^{-1}\right)$, where $\hat{A}(\mathbf{x}_i) = \frac{1}{||\mathbf{x}_i||_2}\left(I_p - \frac{\mathbf{x}_i\mathbf{x}_i^t}{||\mathbf{x}_i||_2^2}\right)$ and $\hat{B}(\mathbf{x}_i) = \frac{\mathbf{x}_i\mathbf{x}_i^t}{||\mathbf{x}_i||_2^2}$, with $\mathbf{x}_i \in \mathbb{R}^p$, for each $i = 1, ..., n$.

The asymptotic result is also given in page 9-11 of Becker et al. [2014] as well as the estimate for the approximate covariance matrix on page 11. The assumptions in that paper are analogous, but it can be seen that C2 assumption about the spatial median being zero is not necessary, only that it is unique and the density function f is bounded and continuous at μ (Section 1.4, page 9 from Becker et al. [2014]). The difference is that when approximating the covariance matrix, the data should be centered around the estimated spatial median.

The numerator $E\left[\|\hat{\boldsymbol{\mu}}_{MM} - \boldsymbol{\mu}\|^2\right]$ from the expression (3.7) can be approximated with $trace\left(\frac{1}{n}\hat{A}^{-1}\hat{B}\hat{A}^{-1}\right)$, as suggested by Möttönen et al. [2010]. Then the estimators for $\hat{\nu}$ and $\hat{\eta}$ in Equation 3.7 would be estimated as:

$$\hat{\nu}_{\boldsymbol{\mu}} = \frac{\hat{\boldsymbol{\mu}}_{MM}\mathbf{e}}{p} \quad \text{and} \quad \hat{\eta} = \frac{trace\left(\frac{1}{n}\hat{A}^{-1}\hat{B}\hat{A}^{-1}\right)}{\|\hat{\boldsymbol{\mu}}_{MM} - \hat{\nu}_{\boldsymbol{\mu}}\mathbf{e}\|^2}\,.$$

3.2 Dispersion parameter

Based on the median concept, which is a robust measure of location, there is a robust measure of dispersion for a random variable X, which is the *Median Absolute Deviation (MAD)* from the data's median:

$$MAD(X) = median(|X - median(X)|)\,.$$

Falk [1997] showed the following relation, assuming normality, between the MAD and the standard deviation σ_X:

$$MAD(X) = \sigma_X \Phi^{-1}(3/4)\,, \tag{3.8}$$

where Φ denotes the standard normal cdf. Taking the square in (3.8) we obtain a relation between the variance σ_X^2 and $MAD^2(X)$:

$$\sigma_X^2 = 2.198 \cdot MAD^2(X)\,.$$

Extending the idea of the MAD, a robust measure of dependence between two random variables X and Y is the *comedian* (Falk [1997]):

$$COM\,(X,Y) = med\,((X - med\,(X))\,(Y - med\,(Y)))\ . \tag{3.9}$$

The comedian generalizes the MAD because $COM(X,X) = MAD^2(X)$, and also has the highest possible breakdown point (Falk [1997]). An important fact is that the comedian parallels the covariance, but the latter requires the existence of the first two moments of the two random variables, whereas the comedian always exists. Other known properties of the comedian are that it is symmetric, location invariant and scale equivariant. Furthermore, Hall and Welsh [1985] discussed the strong consistency and asymptotic normality of the MAD, and Falk [1997] established similar results for the comedian.

Finally, let us recall that a *comedian matrix* can be defined based on a multivariate version of (3.9). Let $\mathbf{x} = \{\mathbf{x}_1, ..., \mathbf{x}_p\}$ be the $n \times p$ data matrix with n being the sample size and p the number of variables. Then the comedian matrix is defined as:

$$COM(\mathbf{x}) = (COM(\mathbf{x}_j, \mathbf{x}_t))\,, \quad j,t = 1, ..., p\,.$$

Note that from relation described in (3.8), one can consider the adjusted comedian:

$$\hat{S}_{CCM} = 2.198 \cdot COM(\mathbf{x})\,.$$

Recall the previous Section 3.1 in which we needed to provide a robust estimator for $\sigma^2_{\mathbf{x}_j}$, for each $j = 1, ..., p$ (Equation 3.5) note that, because of the relation in (3.8):

$$\begin{aligned} trace(\hat{S}_{CCM}) &= \sum_{j=1}^{p} 2.198 \cdot COM(\mathbf{x}_j, \mathbf{x}_j) \\ &= \sum_{j=1}^{p} 2.198 \cdot MAD^2(\mathbf{x}_j) = \sum_{j=1}^{p} \sigma^2_{\mathbf{x}_j}\,. \end{aligned} \tag{3.10}$$

Thus, when considering a shrinkage estimator of the component-wise median, in order to estimate the variance of $\hat{\boldsymbol{\mu}}_{CCM}$ needed in the expression (3.4) for the shrinkage intensity $\hat{\eta}$, and according to the relation (3.5), we propose to estimate $\sum_{j=1}^{p} \sigma^2_{\mathbf{x}_j}$ using the $trace(\hat{S}_{CCM})$. Therefore, the estimates for $\hat{\nu}_{\boldsymbol{\mu}}$ and $\hat{\eta}$ in expression (3.4) can be calculated as:

$$\hat{\nu}_{\boldsymbol{\mu}} = \frac{\hat{\boldsymbol{\mu}}_{CCM}\mathbf{e}}{p} \quad \text{and} \quad \hat{\eta} = \frac{(\pi/2n)trace(\hat{S}_{CCM})}{\|\hat{\boldsymbol{\mu}}_{CCM} - \hat{\nu}_{\boldsymbol{\mu}}\mathbf{e}\|_2^2}\,. \tag{3.11}$$

Note that $\hat{S}_{CCM}$ is a robust alternative for the covariance matrix, but in general, it is not positive (semi-) definite (see Falk [1997]). Since we need this property for

inverting the covariance matrix in a Mahalanobis distance, we propose a shrinkage over $\hat{S}_{CCM}$, because of its advantage of always providing a positive definite and well-conditioned matrix. Therefore, if a shrinkage estimator is considered for the dispersion parameter:

$$\hat{\Sigma}_{Sh} = (1-\eta)\hat{E} + \eta\hat{T}, \tag{3.12}$$

we propose to use in (3.12), the estimator $\hat{E} = \hat{S}_{CCM}$.

Several choices for the shrinkage target $\hat{T}$ have been proposed in the literature. For example, Ledoit and Wolf [2003b] proposed a weighted average of the sample covariance matrix and a single-index covariance matrix. Ledoit and Wolf [2003a] proposed selecting the shrinkage target as a "constant correlation matrix", whose correlations are set equal to the average of all sample correlations. Finally, Ledoit and Wolf [2004] proposed to use a multiple of the identity matrix as the shrinkage target. The authors proved that the resulting shrinkage covariance matrix is well-conditioned, even if the sample covariance matrix is not. There is also another approach introduced by DeMiguel et al. [2013]. The authors proposed a shrinkage estimator both for the covariance matrix and its inverse. The estimators were constructed as a convex combination of the sample covariance matrix or its inverse, respectively, and a scaled shrinkage target, which they consider the scaled identity matrix as Ledoit and Wolf [2004]. Therefore, we propose to use as shrinkage target $\hat{T} = \nu_\Sigma I$. Thus (3.12) results in:

$$\hat{\Sigma}_{Sh(CCM)} = (1-\eta)\hat{S}_{CCM} + \eta\nu_\Sigma I\,.$$

Finally, the scaling parameter ν_Σ and the shrinkage intensity parameter η needs to be estimated. They both are chosen to minimize the expected quadratic loss as in Ledoit and Wolf [2004]:

$$\begin{array}{ll} \min_{\nu_\Sigma,\eta} & E\left[\left\|\hat{\Sigma}_{Sh} - \Sigma\right\|^2\right] \\ \text{s.t.} & \hat{\Sigma}_{Sh} = (1-\eta)\hat{S}_{CCM} + \eta\nu_\Sigma I, \end{array} \tag{3.13}$$

where $\|A\|^2 = trace(AA^T)/p$.

Proposition 3 *The solution of the problem (3.13) is:*

$$\hat{\nu}_\Sigma = trace(\hat{S}_{CCM})/p, \quad \hat{\eta} = \frac{E\left[\left\|\hat{S}_{CCM} - \Sigma\right\|^2\right]}{E\left[\left\|\hat{S}_{CCM} - \nu_\Sigma I\right\|^2\right]}\,.$$

The proof can be found in Appendix A.3. In practice, we propose to estimate the numerator of the expression for η as Ledoit and Wolf [2003a], Ledoit and Wolf [2003b] and Ledoit and Wolf [2004], but considering $\hat{S}_{CCM}$ instead of the sample covariance matrix, as the estimator of Σ.

Note that the comedian matrix depends on centered data considering the componentwise median $\hat{\boldsymbol{\mu}}_{CCM}$. A special case of comedian matrix can be defined if the data

are centered using a different location estimator. We propose to center the data using the other location estimators described in Section 3.1, i.e., the multivariate L_1-median $\hat{\boldsymbol{\mu}}_{MM}$, and the shrinkage estimators $\hat{\boldsymbol{\mu}}_{Sh(CCM)}$ and $\hat{\boldsymbol{\mu}}_{Sh(MM)}$. We will consider shrinkages over those special comedian matrices.

1. $\hat{\Sigma}_{Sh(MM)} = (1-\eta)\hat{S}_{MM} + \eta\nu_\Sigma I$, with for $j,t = 1,...,p$:
   $\hat{S}_{MM} = 2.198 \cdot COM_{MM}(\mathbf{x}) = 2.198 \cdot (med((\mathbf{x}_j - (\hat{\boldsymbol{\mu}}_{MM})_j)(\mathbf{x}_t - (\hat{\boldsymbol{\mu}}_{MM})_t))\,.$
2. $\hat{\Sigma}_{Sh(Sh(CCM))} = (1-\eta)\hat{S}_{Sh(CCM)} + \eta\nu_\Sigma I$, with for $j,t = 1,...,p$:
   $\hat{S}_{Sh(CCM)} = 2.198 \cdot COM_{Sh(CCM)}(\mathbf{x}) = 2.198 \cdot (med((\mathbf{x}_j - (\hat{\boldsymbol{\mu}}_{Sh(CCM)})_j)(\mathbf{x}_t - (\hat{\boldsymbol{\mu}}_{Sh(CCM)})_t))\,.$
3. $\hat{\Sigma}_{Sh(Sh(MM))} = (1-\eta)\hat{S}_{Sh(MM)} + \eta\nu_\Sigma I$, with for $j,t = 1,...,p$:
   $\hat{S}_{Sh(MM)} = 2.198 \cdot COM_{Sh(MM)}(\mathbf{x}) = 2.198 \cdot (med((\mathbf{x}_j - (\hat{\boldsymbol{\mu}}_{Sh(MM)})_j)(\mathbf{x}_t - (\hat{\boldsymbol{\mu}}_{Sh(MM)})_t))\,.$

The optimal expression for the parameters η and ν_Σ in the above cases is analogous to the Proposition 3, but considering in each case the sample estimator as the corresponding special comedian matrix.

3.3 Proposed Robust Mahalanobis Distances

A robust Mahalanobis distance can be defined for each of the following six possible combinations for the location and the dispersion estimators (see Table 3.1). Note that they are meaningful combinations because the shrinkage estimator of dispersion is made upon a special comedian matrix closely based on the location estimator jointly considered for defining the *RMD*.

Table 3.1: Combinations of location and dispersion

Name	RMDv1	RMDv2	RMDv3	RMDv4	RMDv5	RMDv6
$\hat{\mu}_R$	$\hat{\mu}_{CCM}$	$\hat{\mu}_{Sh(CCM)}$	$\hat{\mu}_{Sh(CCM)}$	$\hat{\mu}_{MM}$	$\hat{\mu}_{Sh(MM)}$	$\hat{\mu}_{Sh(MM)}$
$\hat{\Sigma}_R$	$\hat{\Sigma}_{Sh(CCM)}$	$\hat{\Sigma}_{Sh(CCM)}$	$\hat{\Sigma}_{Sh(Sh(CCM))}$	$\hat{\Sigma}_{Sh(MM)}$	$\hat{\Sigma}_{Sh(MM)}$	$\hat{\Sigma}_{Sh(Sh(MM))}$

For all our proposed combinations, the threshold considered to detect the outliers is the $\chi^2_{p;0.975}$ quantile, because it is the cut-off used in the literature for most of the robust distances. Although, since explained in the Introduction of the thesis, it does not necessarily have to be that one.

3.4 Simulation results

3.4.1 Normal distribution

A simulation study is performed considering a $p-$dimensional random variable X following a contaminated multivariate normal distribution given as a mixture of Normals of the form $(1-\alpha)N(\mathbf{0}, I) + \alpha N(\delta\mathbf{e}, \lambda I)$, where $\mathbf{e}$ denotes the $p-$dimensional

vector of ones. This model is analogous to the one used by Rousseeuw and Driessen [1999], Peña and Prieto [2001], Filzmoser et al. [2005], Peña and Prieto [2007], Maronna and Zamar [2002] and Sajesh and Srinivasan [2012]. This experiment has been conducted for different values of the sample-space dimension $p = 5, 10, 30, 50$, and the chosen sample size in relation to the dimension was $n = 100, 100, 500, 1000$, respectively. The contamination levels were $\alpha = 0, 0.1, 0.2, 0.3$, the distance of the outliers $\delta = 5$ and 10 and the concentration of the contamination $\lambda = 0.1$ and 1. For each set of values, 100 random sample repetitions have been generated.

For the methods mentioned in previous sections some measures are studied: the true positive rate (TPR) and the false positive rate (FPR). If we call NO the real number of not outlying observations and TO the real number of outliers, then:

$$TPR = \frac{TP}{TO} \quad \text{and} \quad FPR = \frac{FP}{NO},$$

where TP means true positives and are the outliers correctly identified by the method, while FP means false positives and are the observations incorrectly detected as outliers by the method. The TPR is also equal to $1 - FNR = 1 - \frac{FN}{TO}$, where FN means False Negatives and are the outliers that the approach fails to identify as such.

Then the two measures TPR and FPR are selected to study the performance of the methods. The method *MCD* refers to the *RMD* based on the *MCD* estimator and with the classical threshold, the method *Adj MCD* refers to the latter distance considering the adjusted quantile of Filzmoser et al. [2005], the method *Kurtosis* refers to the Peña and Prieto [2007] approach, the method *OGK* refers to the Orthogonalized Gnanadesikan-Kettenring method proposed by Maronna and Zamar [2002] and *COM* is the Comedian method proposed by Sajesh and Srinivasan [2012]. We have also presented the results for the collection *RMDv1-RMDv6* proposed in Table 3.1. All simulations were performed in Matlab.

Appendix B.1 contains the tables corresponding to all simulation scenarios with Normal data. Here we show only the most significant and representative results. Nevertheless, the tables show overall outcomes. For example, *Adj MCD*, actually improves *MCD* with respect to the FPR, lowering it, and in most cases maintaining the same TPR. Although, in other cases, it also slightly lowers the TPR. On the other hand, the FPR in case of no contamination is sufficiently low for all methods, but our proposed collection shows the lowest values especially in high dimension, actually here the best performance is observed for *RMDv6*. With certain percent of contamination, the worst behavior of our proposed methods is when the dimension is low and the highest percentage of outliers are considered to be near the center of the data. This matter can be seen in Table 3.2, which corresponds to the TPR. When the outliers are near the center of the data ($\delta = 5$), in case of low dimension ($p = 5$), with 30% of outliers, *Kurtosis* has better performance. This happens also with $p = 10$, but in all other cases *MCD*, *Adj MCD*, *Kurtosis* and *OGK* are the ones with the worst behavior about the TPR. Meanwhile, *COM* is a good alternative, but the overall best performance is made by *RMDv6* especially in high dimension and even with large contamination.

Table 3.2: True positive rates, with Normal distribution.

$\delta = 5$		$\lambda = 0.1$										
p	α	MCD	Adj MCD	Kurtosis	OGK	COM	RMDv1	RMDv2	RMDv3	RMDv4	RMDv5	RMDv6
5	0.1	1	1	0.9000	1	1	1	1	1	1	1	1
	0.2	0.8700	0.8700	0.5100	0.9500	0.9941	1	1	1	1	1	1
	0.3	0.0600	0.0600	0.9800	0.1500	0.5719	0.8766	0.8782	0.8782	0.9146	0.9090	0.9130
10	0.1	0.9900	0.9900	0.8600	1	1	1	1	1	1	1	1
	0.2	0.2800	0.2800	0.4600	0.9416	1	1	1	1	1	1	1
	0.3	0	0	0.9900	0.1612	0.7205	0.8774	0.8747	0.8750	0.9711	0.9672	0.9711
30	0.1	0.1900	0.1900	1	1	1	1	1	1	1	1	1
	0.2	0	0	0.1000	1	1	1	1	1	1	1	1
	0.3	0	0	0.6100	0.0100	0.9407	0.5308	0.5275	0.5286	0.9990	0.9988	0.9991
50	0.1	0	0	1	1	1	1	1	1	1	1	1
	0.2	0	0	0	1	1	1	1	1	1	1	1
	0.3	0	0	0	0	0.9839	0.5021	0.5000	0.5000	0.9939	0.9932	0.9942

Another situation is when outliers are far from the center of the data, i.e., $\delta = 10$. This scenario is shown in Table 3.3.

Table 3.3: True positive rates, with Normal distribution.

$\delta = 10$		$\lambda = 1$										
p	α	MCD	Adj MCD	Kurtosis	OGK	COM	RMDv1	RMDv2	RMDv3	RMDv4	RMDv5	RMDv6
5	0.1	1	1	1	1	1	1	1	1	1	1	1
	0.2	0.8480	0.8465	0.9900	1	1	1	1	1	1	1	1
	0.3	0.2190	0.1976	0.9307	0.9591	0.9991	1	1	1	1	1	1
10	0.1	1	1	0.9800	1	1	1	1	1	1	1	1
	0.2	0.8623	0.8548	0.6558	1	1	1	1	1	1	1	1
	0.3	0.2280	0.2046	0.4618	0.9911	1	1	1	1	1	1	1
30	0.1	1	1	0.8919	1	1	1	1	1	1	1	1
	0.2	0.4879	0.4654	0.0125	1	1	1	1	1	1	1	1
	0.3	0.0810	0.0509	0.1087	1	1	1	1	1	1	1	1
50	0.1	1	1	0.6017	1	1	1	1	1	1	1	1
	0.2	0.2695	0.2348	0.0017	1	1	1	1	1	1	1	1
	0.3	0.0643	0.0378	0.0006	1	1	1	1	1	1	1	1

It is clear from Table 3.3 that when outliers are far from the center, our proposed methods lead to the best performance, achieving 100% of TPR, for all dimension and percentage of contamination considered. *OGK* and *COM* are good alternatives in case of high dimension $p = 30, 50$. Other tables about the TPR can be found in Appendix B.1, as well as the FPR tables, which show that in the vast majority of cases our proposed distances have an FPR value equal to zero and when not, a value very close to zero, which is what is desirable.

3.4.2 t_3-distribution

In order to check the behavior of the methods when the distribution deviates from normality, a simulation study is performed considering a $p-$dimensional random variable X following a contaminated multivariate t-distribution with 3 degrees of freedom of the form $(1-\alpha)T_3(\mathbf{0}, I) + \alpha T_3(\delta\mathbf{e}, \lambda I)$. The first parameter of the notation of $T_3(\cdot,\cdot)$ refers to the mean and the second one to the covariance matrix. The parameters for the contamination are the same considered above and the same

measures TPR and FPR are studied. All the results can be found in Appendix B.2. It should be noted the unsatisfactory behavior of the alternative methods with respect to the TPR especially in high dimension or with large contamination level, meanwhile, in most cases we attain a 100% TPR. Concerning the FPR value, all methods show non-zero FPR values, and the best performance is shown by *COM* and our proposed methods.

3.4.3 Exponential distribution

We considered also a $p-$dimensional random variable X following a contaminated multivariate exponential distribution given as a mixture $(1-\alpha)Exp(\mathbf{0})+\alpha Exp(\delta\mathbf{e})$. The parameter of the notation $Exp(\cdot)$ refers to the mean. This case is analogous to the previous ones, with the difference that only the schemes associated with the distance of the outliers are considered. Tables in Appendix B.3 show all the results and it can be seen that our proposed methods achieve 100% of TPR in the majority of cases. The highest value of TPR is also achieved by $Kurtosis$, OGK and COM, when dimension is high. When dimension is low, the TPR is high in most situations for all $Kurtosis$, OGK, COM and $RMDv1-RMDv6$, but in the majority of cases our proposed method's TPR is higher. MCD and $AdjMCD$ decreases their TPR value with the increase of dimension or contamination level. With respect to the FPR value of $Kurtosis$ and OGK their FPR is high in most cases. COM, MCD and $AdjMCD$ have low FPR values. $RMDv1-RMDv6$ also have low FPR values in the majority of cases, except in some cases when the level of contamination is the lowest. On the other hand, in case of no contamination, MCD and $AdjMCD$ show more or less the same FPR value than our proposed methods, while the other alternatives $Kurtosis$, OGK and COM show higher values than them. Considering both the TPR and the FPR, the best overall performance is showed by *RMDv6*.

3.4.4 Summary and selection of one of our proposed distances

In the simulation study, for each contamination scheme we have also calculated a measure called F-score (Goutte and Gaussier [2010], Sokolova et al. [2006], Powers [2011]), often used in Engineering, which is a measure of a test's accuracy. Its expression is F-score$= 2PR/(P+R)$, where P is called precision and R is known as the recall. The precision P is the number of correctly detected outliers divided by the total number of detected outliers, and the recall R is the number of correctly detected outliers divided by the real total number of outliers. The recall coincides with the TPR.

$$P=\frac{TP}{TP+FP} \quad \text{and} \quad R=\frac{TP}{TP+FN}\,.$$

This measure provides a trade-off between the two desired outcomes: a high rate of correctly identified outliers and a low rate of observations mislabel as outliers. The results are not included for avoiding large extension, but the method with the overall classification between the top 3 best positions ranking with respect to the

F-score, is method *RMDv6*.

It is clear the out-performance of our proposed methods with Gaussian data, especially in high dimension and even when we deviate from the normality assumption, for example when considering heavy-tailed and skewed distributions like the multivariate t_3-distribution and the multivariate exponential distribution. From all of our six proposed robust distances, the one that shows the best results in the vast majority of cases is *RMDv6*. Thus, we decided to select it as the best one in the matter of performance, and from now on we will refer to it as *RMD-S*. The definition is the following:

$$RMD\text{-}S(\mathbf{x}_i) = \left(\left(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{Sh(MM)} \right) \hat{\Sigma}^{-1}_{Sh(Sh(MM))} \left(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{Sh(MM)} \right)^t \right)^{1/2} ,$$

where the location estimator is:

$$\hat{\boldsymbol{\mu}}_{Sh(MM)} = (1-\eta)\hat{\boldsymbol{\mu}}_{MM} + \eta\nu_{\boldsymbol{\mu}}\mathbf{e} \,, \tag{3.14}$$

and the covariance estimator is:

$$\hat{\Sigma}_{Sh(Sh(MM))} = (1-\eta)\hat{S}_{Sh(MM)} + \eta\nu_{\Sigma} I \,, \tag{3.15}$$

with, for $j, t = 1, ..., p$:

$$\hat{S}_{Sh(MM)} = 2.198 \cdot COM_{Sh(MM)}(\mathbf{x}) = 2.198 \cdot (med((\mathbf{x}_j - (\hat{\boldsymbol{\mu}}_{Sh(MM)})_j)(\mathbf{x}_t - (\hat{\boldsymbol{\mu}}_{Sh(MM)})_t)) \,.$$

3.5 Properties of the estimator

In this section, some properties like the behavior under correlated data, the affine equivariance, the breakdown value, and the computational times are studied.

3.5.1 Correlation and affine equivariance

Consider $X = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ and a pair of multivariate location and covariance estimators (m, S). In general, these estimators are called affine equivariant if for any nonsingular matrix A it holds that:

$$m_A = m(X_A) = Am(X) \quad \text{and} \quad S_A = S(X_A) = S(X)A^t \,. \tag{3.16}$$

The affine transformation of X is $X_A = \{A\mathbf{x}_1, ..., A\mathbf{x}_n\}$. Affine equivariance implies that the estimator transforms well under any nonsingular reparametrization of the space of the $\mathbf{x}_i$. The data might for instance be rotated, translated or rescaled (for example through a change of the measurement units).

The method *RMD-S* is ultimately based on not affine equivariant estimators which are the L_1-median (Lopuhaa and Rousseeuw [1991]) and the comedian matrix. However, the L_1-median is orthogonal equivariant, i.e., it satisfies Equation (3.16) with A any orthogonal matrix ($A^t = A^{-1}$). This implies that the L_1-median

transforms appropriately under all transformations that preserve Euclidean distances (such as translations, rotations and reflections). About the comedian matrix, which always exists, it is symmetric, location invariant and scale equivariant (Falk [1997]), i.e., $COM(X, \mathbf{a}Y + \mathbf{b}) = \mathbf{a}COM(X, Y) = \mathbf{a}COM(Y, X)$. Since the proposed method is not affine equivariant, it is important to investigate the behavior under correlated data. Devlin et al. [1981] used a correlation matrix P for generating Monte Carlo data from different distributions of moderate dimension $p = 6$. The matrix has the form:

$$P = \begin{bmatrix} P_1 & 0 \\ 0 & P_2 \end{bmatrix}$$

$$P_1 = \begin{bmatrix} 1 & 0.95 & 0.3 \\ 0.95 & 1 & 0.1 \\ 0.3 & 0.1 & 1 \end{bmatrix}, \quad P_2 = \begin{bmatrix} 1 & -0.499 & -0.499 \\ -0.499 & 1 & -0.499 \\ -0.499 & -0.499 & 1 \end{bmatrix}.$$

The reason for the selection of the matrix P is because the dimension is large enough to study multivariate estimators and the range of correlation values is large. This way the differences in the abilities of the methods to detect outliers from highly correlated data can be observed. For the simulations, $n = 100$ observations were generated from a mixture of Normals $(1-\alpha)N(\mathbf{0}, P) + \alpha N(5\mathbf{e}, P)$. The contamination level $\alpha = 10\%, 20\%, 30\%$.

Table 3.4 shows that the TPR and FPR of *MCD*, *Adj MCD*, *Kurtosis* and *OGK* are worse than that of our proposal. On the other hand, *COM* shows more or less the same behavior in case of 10% and 20% of contamination, and slightly worse than our proposal when the contamination level increases to 30%. The methods *MCD*, *Adj MCD* and *Kurtosis* are affine equivariant, while *OGK* and *COM* are not. Hence, the proposed procedure *RMD-S* is more efficient than other affine and not affine equivariant methods in case of correlated data-sets. Also the FPR is very low even in this case of presence of correlation.

Table 3.4: Simulation results for correlated data.

	MCD		Adj MCD		Kurtosis		OGK		COM		RMD-S	
α	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
0.1	1	0.0397	1	0.0226	1	0.0371	1	0.0736	1	0.0025	1	0.0128
0.2	0.8659	0.0127	0.8565	0.0062	0.8771	0.0453	0.9792	0.0533	1	0.0011	1	0.0013
0.3	0.1504	0.0762	0.1238	0.0614	0.8186	0.0443	0.4780	0.0460	0.8302	0.0001	0.9274	0

Affine equivariance of the estimators is equivalent to say that the robust Mahalanobis distance is affine invariant:

$$RMD(\mathbf{Ax_i}, \mathbf{m_A}) = (\mathbf{Ax_i} - \mathbf{m_A})\mathbf{S_A}^{-1}(\mathbf{Ax_i} - \mathbf{m_A})^t = RMD(\mathbf{x_i}, \mathbf{m}).$$

Maronna and Zamar [2002] and Sajesh and Srinivasan [2012] proposed to investigate the lack of equivariance with transformed data, by simulations. We study the same for our proposal. They propose to generate random matrices as $A = TD$, where T is a random orthogonal matrix and $D = diag(u_1, ..., u_p)$, where the u_j's

are independent and uniformly distributed in $(0, 1)$. Then, the proposed simulations consist on affinely transform each generated data matrix X in each repetition, by applying the random matrix of transformation A to X, in order to obtain X_A. The contamination scheme consist in data generated from a mixture of Normals $(1-\alpha)N(\mathbf{0}, I) + \alpha N(\delta\mathbf{e}, \lambda I)$. The dimension $p = 5, 10, 30, 50$, with sample size $n = 100, 100, 500, 1000$ respectively, the contamination level $\alpha = 10, 20, 30\%$, the distance of the outliers $\delta = 5$ and 10, and the concentration of the contamination $\lambda = 0.1$ and 1. Table 3.5 and 3.7 show the obtained results about the TPR and FPR. As it can be observed, even under affine transformations, *RMD-S* is able to detect all the outliers, except for a few cases (in bold type) that corresponds to large contamination level (30%) in case of outliers close to the center of the distribution. However, it can be noted that these cases improve in performance when dimension increases.

Table 3.5: True posive rates and false positive rates of RMD-S for transformed data, $\lambda = 0.1$.

		$\delta = 5$		$\delta = 10$	
p	α	TPR	FPR	TPR	FPR
5	0.1	1	0.0454	1	0.0455
	0.2	1	0.0155	1	0.0165
	0.3	**0.9709**	0.0034	1	0.0023
10	0.1	1	0.0328	1	0.0252
	0.2	1	0.0088	1	0.0062
	0.3	**0.9844**	0.0023	1	0.0009
30	0.1	1	0.0089	1	0.0074
	0.2	1	0.0006	1	0.0003
	0.3	1	0	1	0
50	0.1	1	0.0008	1	0.0004
	0.2	1	0.0002	1	0.0001
	0.3	1	0	1	0

Table 3.6: True posive rates and false positive rates of RMD-S for transformed data, $\lambda = 1$.

		$\delta = 5$		$\delta = 10$	
p	α	TPR	FPR	TPR	FPR
5	0.1	1	0.0451	1	0.0400
	0.2	1	0.0189	1	0.0120
	0.3	**0.9344**	0.0039	1	0.0046
10	0.1	1	0.0282	1	0.0279
	0.2	1	0.0113	1	0.0071
	0.3	**0.9872**	0.0020	1	0.0010
30	0.1	1	0.0093	1	0.0072
	0.2	1	0.0006	1	0.0004
	0.3	1	0	1	0
50	0.1	1	0.0009	1	0.0002
	0.2	1	0.0002	1	0.0001
	0.3	1	0	1	0

3.5.2 Breakdown value

For an estimator, the maximum proportion of outliers that it can safely tolerate is known as the breakdown value. For an outlier detection method, the breakdown value can be defined as the maximum m^* outliers that the procedure can successfully detect, so that if the data is contaminated with m outliers and $m > m^*$ the method will fail to identify most of the true outliers and it will falsely detect many inliers, reducing the true positive rate drastically, and inflating the false positive rate (Sajesh and Srinivasan [2012]). Thus, it is necessary to use the true positive and false positive rates for studying the breakdown value of the outlier detection procedure. Analogously as in Sajesh and Srinivasan [2012], n observations were generated from a $p-$dimensional $N(\mathbf{0}, I)$ and two forms of contamination are considered: α percent symmetric, for which the ith observation is multiplied by $100i$, and α percent asymmetric, for which the ith observation is replaced by $(100i)\mathbf{e}$, $i = 1, ..., n\alpha$, where $\mathbf{e} = (1, ..., 1)$. In the first case the outliers are symmetrically distributed, and asymmetrically in the second case. The dimensions considered are $p = 10, 30, 50, 80, 100$ and the sample size $n = 1000$. The contamination level $\alpha = 10, 20, 30, 40, 45\%$. Table 3.7 gives the resulting TPR and FPR for both forms of contamination.

Table 3.7: Simulation results for breakdown value.

$n = 1000$		Symmetric		Asymmetric	
p	α	TPR	FPR	TPR	FPR
10	0.1	1	0.0055	1	0.0047
	0.2	1	0.0001	1	0.0002
	0.3	1	0	1	0
	0.4	1	0	1	0
	0.45	1	0	1	0
30	0.1	1	0.0002	1	0.0002
	0.2	1	0	1	0
	0.3	1	0	1	0
	0.4	1	0	1	0
	0.45	1	0	1	0
50	0.1	1	0	1	0
	0.2	1	0	1	0
	0.3	1	0	1	0
	0.4	1	0	1	0
	0.45	1	0	1	0
80	0.1	1	0	1	0
	0.2	1	0	1	0
	0.3	1	0	1	0
	0.4	1	0	1	0
	0.45	1	0	1	0
100	0.1	1	0	1	0
	0.2	1	0	1	0
	0.3	1	0	1	0
	0.4	1	0	1	0
	0.45	1	0	1	0

It can be seen that the TPR is not affected and the FPR is zero for most cases

or it is very reduced and near zero, then RMD-S can successfully detect the outliers even when there is large contamination and even in high dimension, without falsely detect many inliers.

3.5.3 Computational times

Table 3.8 show the resulting computational times in seconds for the Normal case when outliers are close to the center of the data and they are concentrated.

Table 3.8: Computational times with Normal data, $\delta = 5$ and $\lambda = 0.1$.

p	α	MCD	Adj MCD	Kurtosis	OGK	COM	RMD-S
5	0.1	1.0951	0.7670	0.1880	0.1087	0.0228	0.0096
	0.2	0.7619	0.7910	0.0499	0.0203	0.0088	0.0085
	0.3	0.7605	0.8304	0.0266	0.0196	0.0089	0.0074
	Mean	**0.8725**	**0.7961**	**0.0882**	**0.0495**	**0.0135**	**0.0085**
10	0.1	1.3184	0.9970	0.2191	0.1626	0.0247	0.0200
	0.2	1.0329	1.0477	0.1358	0.0793	0.0120	0.0118
	0.3	0.9685	1.0641	0.0482	0.0865	0.0128	0.0108
	Mean	**1.1066**	**1.0363**	**0.1344**	**0.1095**	**0.0165**	**0.0142**
30	0.1	6.2387	6.0934	0.7154	0.8969	0.2000	0.2206
	0.2	5.8676	6.3999	1.4635	0.8158	0.1687	0.1804
	0.3	5.9453	7.0405	1.6572	0.8407	0.1669	0.1674
	Mean	**6.0172**	**6.5113**	**1.2787**	**0.8511**	**0.1785**	**0.1895**
50	0.1	7.3521	7.2307	2.2497	1.2854	0.2174	0.2053
	0.2	7.2501	7.2337	2.2778	1.2678	0.2166	0.2018
	0.3	7.2479	7.2376	2.3753	1.2774	0.2169	0.2099
	Mean	**7.2834**	**7.2340**	**2.3009**	**1.2769**	**0.2169**	**0.2057**

The other tables can be founded in Appendix B.1. The experiment is carried out on a PC with a 3.40 GHz Intel Core i7 processor with 32GB RAM. On average, the fastest methods are *COM* and *RMD-S* with very similar computational times. Compared to the *MCD* and its adjusted version *Adj MCD*, the latters are much more slower than our proposal. Depending on the dimension of the data, *MCD* and *Adj MCD* are between 31-93 and 34-102 times slower than *RMD-S*, respectively. *Kurtosis* and *OGK* are not as slower as *MCD* and *Adj MCD*, but they show worse computational times than *COM* and *RMD-S*. *Kurtosis* and *OGK* are between 6-11 and 4-8 times slower than our proposal. Thus, *RMD-S* shows competitive computational times as well as *COM*.

3.6 Real data-set example

The proposed *RMD* is applied to a real data-set to evaluate its performance. The following data-set was taken from the *UCI Knowledge Discovery in Databases Archive* (Bay [1999]). Specifically, we have chosen the *Breast Cancer Wisconsin (Diagnostic) Data-Set* (WDBC). Features are computed from a digitized image of a fine needle aspirate of a breast mass. They describe 30 characteristics of the cell nuclei present in the image, for 569 samples, from which 357 are benign and 212 malign.

We propose to study only the 357 benign data. In Maronna and Zamar [2002] the authors analyzed several data-sets but they only show the results of four of them. Specifically, in section 4.5 page 314 the authors mention the data we used and they specify that the dimension was $p = 30$ and the sample size $n = 357$, which means that they selected only the 357 observations corresponding to benign data. This is the same that they do with the study of Ionospheric data (section 4.3, page 312), since the classification of the observations is previously known and it makes sense to study only one of the two groups because they come from a different distribution and the observations from the "bad" group are almost half of the entire data-set. The data is available at UCI repository. The archive has 32 columns but the first two are (1) the ID number and (2) Diagnosis (good or bad), which leaves us with 30 features. Therefore, this example has dimension $p = 30$ and sample size $n = 357$. We applied each method for detecting outliers and we retained the results, along with the computational times.

In order to interpret the outcome, we show the standardized data (after the detection) only for better visualization aim. We have also plotted the multivariate L_1 median and a kind of "multivariate boxplot", which is based on the idea from Sun and Genton [2011] method, but for finite dimensional. What the "box" would be is constructed sorting the data according to their L_1 depth value. The corresponding Q_1 and Q_3 "quartiles" delimiting the "box" are in fact the minimum and maximum values for each coordinate taking only into account the 50% of the most central data. Thus, the "fences" can be constructed with the same approach $F_1 = Q_1 - 1.5RI$ and $F_2 = Q_3 + 1.5RI$, where the "interquartile range" is $RI = Q_3 - Q_1$. Then, we can look for each method's result how many detected outliers are inside the "fences" for all their coordinates, and how many are outside the "fences". Figure 3.1 shows the data in blue color plotted in parallel coordinates (Inselberg and Dimsdale [1990], Wegman [1990], Inselberg [2009]), the "box" delimiting the 50% of most central data in yellow color, the "fences" in red and the multivariate L_1-median in cyan.

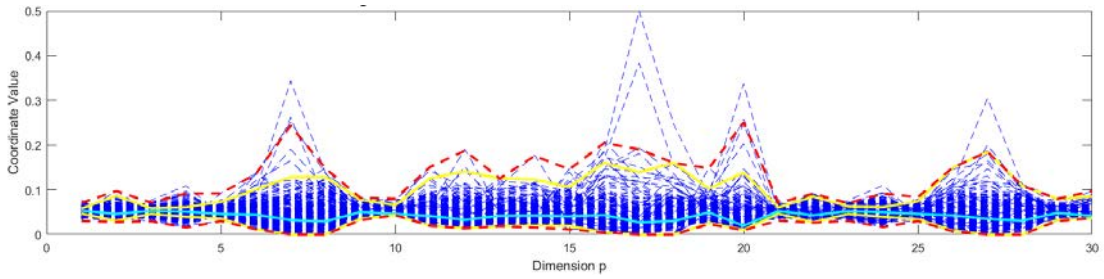

Figure 3.1: Standardized data with the "multivariate boxplot".

Table 3.9 shows the detected outliers by each method. Outside the "fences" there are 3 or 4 for all the methods. Also, the method *Kurtosis* detected 162 outliers out of the 357 data. More or less like *OGK*, which detected 148. Furthermore, our method *RMD-S* is the one that labels less amount of data as outliers.

Table 3.9: Detected outliers inside and outside the fences.

Method	**Inside**	**Outside**	**Total**
MCD	72	4	76
Adj MCD	64	4	68
Kurtosis	158	4	162
OGK	144	4	148
COM	59	4	63
RMD-S	25	3	28

Table 3.10 shows how many outliers belong to the 50% of the most central data, i.e., the observation that fall inside the "box" of the multivariate box-plot.

Table 3.10: Detected outliers inside the "box" with the 50% of the most central data.

Method	**Inside**	**Total**
MCD	29	76
Adj MCD	27	68
Kurtosis	65	162
OGK	58	148
COM	20	63
RMD-S	7	28

We can investigate the shape of the detected outliers that are inside the "multivariate box", in order to see if they are similar or near to the median, or if they have a distinct shape. The motivation is that in case of real data we do not know the true outliers. Thus, we propose to study the shape of these observations in parallel coordinates (similar as in Maronna and Zamar [2002] with Ionospheric data, where they studied the shape of each observation's sequence of coordinates). Then, since the methods detected a large number of observations as outliers, the multivariate boxplot is used to study the shape of the ones that are closest to the multivariate median, i.e., the ones belonging to the "box" of the multivariate boxplot. Figure 3.2 shows the shape of some of the outliers detected by the alternative methods that belong to the 50% of the most central data. The figures to see all the observations detected by the alternative methods can be found in Appendix C. In cyan color is the multivariate median, in yellow color the "box" and in blue color the detected outlier. The title of each subplot represents the index of the observation. The three subplots from the first column correspond to observations 236 (detected by MCD, AdjMCD, KUR and OGK), 155 (detected by KUR) and 212 (detected by KUR and OGK). The next three from the second column of subplots correspond to observations 254 (detected by MCD, AdjMCD, KUR and OGK), 182 (detected by KUR and OGK) and 234 (detected by MCD, AdjMCD, KUR, OGK and COM). The observation's sequence of coordinates can be considered similar to the multivariate median.

The general outcomes are that Adj MCD detected the same outliers as MCD except for the observations 266 and 332 which shape can be considered near the median. This makes sense since with the adjusted quantile the false positive rate decreases. Kurtosis and OGK detected a lot of observations as outliers and some

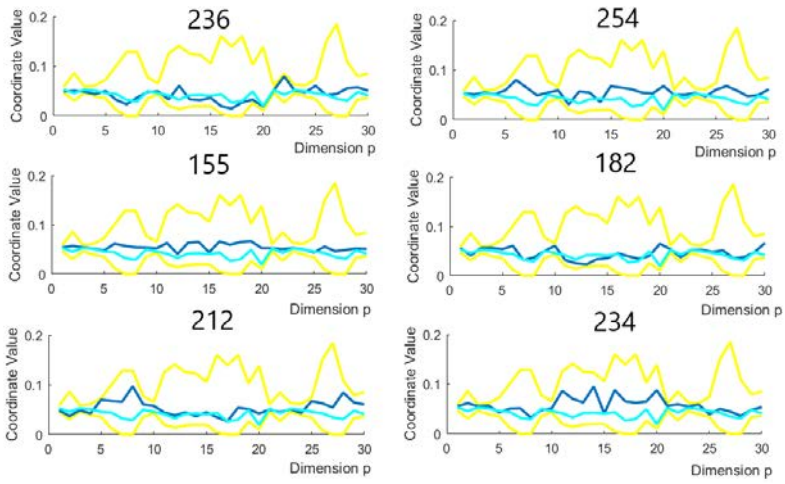

Figure 3.2: Some of the alternative methods detected outliers belonging to the 50% of the most central data.

of the ones inside the "box" are very similar to the multivariate median in parallel coordinates. Comedian method's detected outliers also have some observations similar to the median. In summary, for all of the alternative methods there seems to be some outliers having a shape very alike to the multivariate median or close to it for all the values of its components, leading us to think that maybe the alternative methods are detecting too many observations as outliers, in other words, the false positive rate is inflated. However, in Figure 3.3, we can see that all outliers detected by *RMD-S*, belonging to the "box", are quite different than the multivariate median, in fact, they might be "shape outliers". For a final argument, we can say that all the outliers inside the "box", detected by method *RMD-S*, are recognized by the alternative methods, so this also makes us think that our proposed method detects just enough.

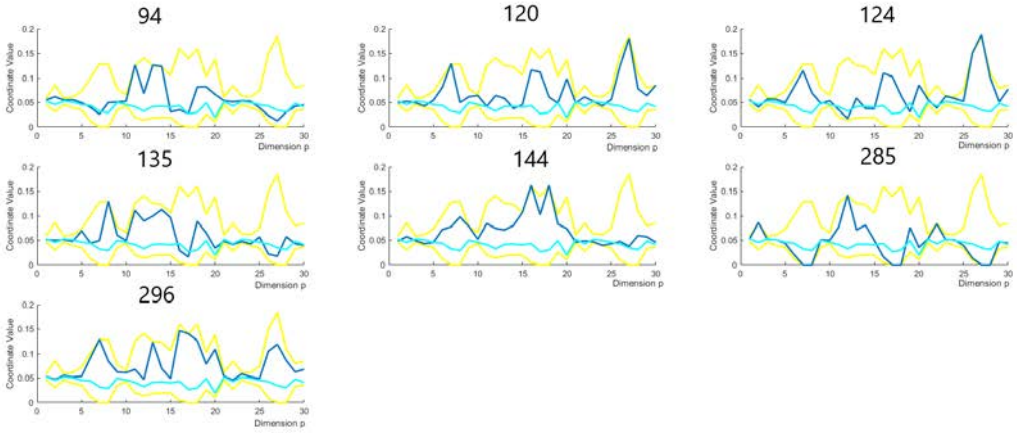

Figure 3.3: RMD-S detected outliers that belong to the 50% of the most central data.

Table 3.11 shows the computational times for each method in the task of detecting outliers with this example of a real data-set. The results demonstrate that the

alternative methods are much slower than our proposal, except for *COM* which has a similar computational time.

Table 3.11: Computational times for each method with the WDBC data-set.

Method	MCD	Adj MCD	Kurtosis	OGK	COM	RMD-S
Times in sec.	12.155	12.3378	6.3077	3.5325	0.3534	0.3299

3.7 Summary

Correct detection of outliers in the multivariate case is well-known to be a crucial task for thorough data analysis. In order to reach that goal properly, it is necessary to consider the shape of the data and its structure in the multivariate space. That is the reason why the Mahalanobis distance approach is frequently used for the task of identifying the outliers. Various robust Mahalanobis distances can be defined according to the selected robust location and dispersion estimators. A collection of different combinations of robust location and covariance matrix estimators based on the notion of shrinkage is proposed, in order to define with each combination a robust Mahalanobis distance for the outlier detection problem. The performance of the proposed RMD's and the others from the literature is shown through a simulation study. It can be concluded that the alternative methods increase their FPR and decrease the TPR in the presence of contamination, especially in high dimension. The proposed RMD's have the ability to discover outliers with high TPR and low FPR in the vast majority of cases in the simulations, with Gaussian data and with skewed or heavy-tailed distributions. RMD-S is the most competitive version, as the simulation results showed. That is the reason why it is selected and some properties are investigated. The behavior under correlated and transformed data shows that RMD-S is approximately affine equivariant. With highly contaminated data it is shown that the approach has high breakdown value even in high dimension. There is also evidence of its reasonable computational time. The behavior is studied with a real data-set example and it shows that the proposed method works well in practice and require reasonable computational times, even for large problems.

CHAPTER 4

Robust regression

Linear regression problems are widely used. The model is defined as:

$$y_i = \alpha + \mathbf{x}_i^t \boldsymbol{\beta} + \epsilon_i \,, \tag{4.1}$$

for $i = 1, ..., n$, where n is the sample size, α is the unknown intercept, $\boldsymbol{\beta}$ is the unknown $(p \times 1)$ vector of regression parameters, and the error terms ϵ_i are i.i.d and they are also independent from the p-dimensional carriers $\mathbf{x}_i$ (often also called regressor or explanatory variables).

Classical ordinary least squares (OLS) regression consists on minimizing the sum of the squared residuals (Equation 1.2). But in spite of its mathematical beauty and computational simplicity, the OLS estimator lacks of robustness, since this approach is not robust to the presence of outliers in the data. In the literature, several authors have proposed robust versions of this estimator, for example, by replacing the sum of squares of the residuals by other function of the residuals. The alternative approach is to use robust estimators of location and covariance matrix in the analogous definition for the OLS regression estimator described previously in Equation 1.3.

4.1 Least Absolute Deviation (LAD) regression

A first proposal of a robust estimate came from Edgeworth [1887] who proposed to replace the squared residuals by the absolute values of them:

$$\hat{\boldsymbol{\beta}}_{LAD} = \underset{\boldsymbol{\beta}}{argmin} \sum_{i=1}^{n} |y_i - \mathbf{x}_i^t \boldsymbol{\beta}| \tag{4.2}$$

This was also called L_1 estimate because of the use of the L_1 norm. It was more resistant than OLS against outlying values in the response variable y, but still couldn't resist outlying values in the carriers. These kinds of outliers are called "leverage points", which have a large effect on the fit. Thus, the bdp of the LAD

estimate is $1/n$.

4.2 M-estimator

The next idea was made by Huber [1964] (also see Huber [1973] and Huber [1981]) which proposed to "huberize" the residuals replacing the least-square criterion by a function $\rho(\cdot)$ of the residuals, where this function had to be symmetric with a unique minimum at zero. It was called $M-$estimator.

$$\hat{\boldsymbol{\beta}}_M = \underset{\boldsymbol{\beta}}{argmin} \sum_{i=1}^{n} \rho\left(\frac{y_i - \mathbf{x}_i^t \boldsymbol{\beta}}{\hat{\sigma}}\right) \tag{4.3}$$

The function $\rho(\cdot)$ is a robust loss function of the residuals and $\hat{\sigma}$ is an error's scale estimate. Since this method did not yield estimators with the invariant property with respect to an increase of the error scale (Rousseeuw [1984]), Huber proposed to estimate the scale parameter simultaneously, making use of a function $\psi(\cdot)$ which is the derivative of ρ. This function was called the influence function. With a minimax procedure, such $M-$estimators were more efficient than LAD at a central model with Gaussian errors. However, the bdp of both of them tend to 0 (since it was $1/n$), because of the possibility of leverage points (Maronna et al. [2006]). Besides, the method implies one first decision: which loss function should be used. It is usually used the Huber's loss function or the Tukey's bisquare function, but there are no rules for which should be selected when we are dealing with real data. Furthermore, they depend on a constant c, which determines the efficiency of the estimator. The authors give their recommendation for the constant to achieve 95% approximately, but this might be a problem as well in practice.

4.3 R-estimator

Another proposal was made by Jaeckel [1972] which consisted on minimizing the sum of some scores of the ranked residuals:

$$\sum_{i=1}^{n} a_n(R_i) r_i \tag{4.4}$$

where R_i represents the rank of the ith residual r_i and $a_n(\cdot)$ is a monotone score function that satisfies:

$$\sum_{i=1}^{n} a_n(i) = 0 \tag{4.5}$$

The problem is that the optimal choice of the score function is unclear, and the bdp is $1/n$.

4.4 Generalized M-estimator

Due to the vulnerability of M-estimators, generalized M-estimators (also called GM-estimators) were proposed, which consisted on the idea to bound the influence of outlying carriers making use of some weight function. This way, the problem of recognizing leverage points was solved because they were downweighted. Some authors developed their methods with this idea (Mallows [1975], Hill [1977], Hampel [1978], Krasker [1980], Krasker and Welsch [1982], Ronchetti and Rousseeuw [1985]). But another problem arose, it cannot distinguish between "good" and "bad" leverage points. And if good leverage points that fall in line with the pattern of the data are down-weighted, this results in a loss of efficiency. On the other hand, they depend on the selection of some constants, which is a nontrivial task in case of real data. Moreover, all $GM-$estimators decrease the bdp with the increase of the dimension p of the carriers.

4.5 Least Median of Squares (LMS) regression

On the other hand, Siegel [1982] proposed a near 50% bdp technique: the Least Median of Squares (LMS), for which the estimates are found by minimizing the median of the squared residuals.

$$\hat{\boldsymbol{\beta}}_{LMS} = \underset{\boldsymbol{\beta}}{argmin}\ Med\{(y_i - \mathbf{x}_i^t\boldsymbol{\beta})^2\} \tag{4.6}$$

However, the procedure had a disadvantage in the order of convergence (see Rousseeuw [1984]), and the ability to provide reasonable estimates when the assumption of Gaussian errors is met was not very good (see Rousseeuw and Croux [1993]).

4.6 Least Trimmed Squares (LTS) regression

Another approach was proposed by Rousseeuw [1985], called Least Trimmed Squares (LTS) estimator and consisted on minimizing the sum of the h (less than and at most n) ordered squared residuals.

$$\hat{\boldsymbol{\beta}}_{LTS} = \underset{\boldsymbol{\beta}}{argmin} \sum_{i=1}^{h} (r_{(i)})^2 \tag{4.7}$$

where $r_{(i)}$ are the ordered squared residuals, and $h = [n(1-\alpha)+1]$ is the proportion of trimming. Usually $h = n/2 + 1$ results in a bdp of 50% and better convergence rate than LMS. The problem is LTS suffers badly in terms of very low efficiency relative to OLS (see Stromberg et al. [2000]). Nevertheless, both LMS and LTS are traditionally used as the initial estimate for some other high bdp and high efficient robust methods.

4.7 S-estimator

Robust regression by means of S-estimator came by hands of Rousseeuw and Yohai [1984]. The method has greater asymptotic efficiency than LTS, although not high enough. It is based on residual scale of M-estimation.

$$\hat{\boldsymbol{\beta}}_S = \underset{\boldsymbol{\beta}}{argmin}\ \hat{\sigma}\,(r_1, ..., r_n) \tag{4.8}$$

where $\hat{\sigma}\,(r_1, ..., r_n)$ is a scale M-estimate. For the biweight scale, S-estimate can attain a high bdp. The matter is that again this approach requires the specification of some constants, the efficiency depends on the values selected, and it is not very high. The computation is not at all scalable, even if there is an iterative procedure or a projection pursuit technique.

4.8 Generalized S-estimator

Croux et al. [1994] proposed the generalized S-estimator (GS-estimator) in an attempt to improve the low efficiency of S-estimators. Again there was a constant to define, which depends on n and p. Overall, the GS-estimator achieves a bdp as high as S-estimator but with higher efficiency.

4.9 MM-estimates

These popular robust estimators were proposed by Yohai [1987] and consisted in three basic steps. The first one is to compute an initial consistent robust estimate of the regression parameters that has high bdp but not necessary high efficiency. The second phase is to use the initial estimator to compute a robust M-estimate of scale of the residuals. The final stage consists on finding an M-estimate of the regression parameters starting at the initial regression estimator. In practice, the typical initial estimators are LMS or S-estimate with Huber or bisquare functions. Playing with the constants necessary for the estimators in the three stages, MM-estimates can attain high efficiency without affecting its bdp. However, the author recognizes in Yohai [1987] that if the constant that handles the efficiency is increased, then the estimates get more sensitive to outliers.

4.10 Covariance approach

Another idea was proposed by Maronna and Morgenthaler [1986] and it was based on covariance estimation. Denote the joint variable as $\mathbf{z} = (\mathbf{x}, \mathbf{y})$, which is of dimension $n \times (p + 1)$, where $\mathbf{x}$ is the $n \times p$ data matrix of the independent variables and y is the $n \times 1$ response variable from the regression problem (Equation 4.1). Denote the location of $\mathbf{z}$ by $\boldsymbol{\mu}$ and the scatter matrix by $\boldsymbol{\Sigma}$. Partitioning $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ yields the notation:

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_x \\ \mu_y \end{pmatrix} \quad , \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \Sigma_{yy} \end{pmatrix} \tag{4.9}$$

Traditionally they are estimated by the empirical mean $\hat{\boldsymbol{\mu}}$ and the empirical covariance matrix $\hat{\boldsymbol{\Sigma}}$. It turns out that the OLS estimators of $\boldsymbol{\beta}$ and the intercept α can be written as functions of the components of $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$, namely

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \hat{\boldsymbol{\Sigma}}_{xx}^{-1} \hat{\boldsymbol{\Sigma}}_{xy} \\ \hat{\alpha} &= \hat{\mu}_y - \hat{\boldsymbol{\beta}}^t \hat{\boldsymbol{\mu}}_x \end{aligned} \tag{4.10}$$

But, since the sample estimators are sensitive to the presence of outliers, we should not use them to estimate the regression parameters. Instead, robust estimators should be used in Equation 4.10. In the literature, there are many proposals of robust location and covariance estimators. For example, the multivariate M-estimators that Maronna and Morgenthaler [1986] considered, or the S-estimator that Croux et al. [2003] considered.

4.11 Robust and efficient weighted least square (REWLSE)

The so-called "robust and efficient weighted least square" estimator (REWLSE) was proposed by Gervini and Yohai [2002]. The authors demonstrate that the method simultaneously achieves maximum bdb and full efficiency under Gaussian errors. The idea is similar to weighted least squares, but the weights are calculated from an initial robust estimator. The weighting scheme is hard rejection (0 or 1), and the cut-off depends on the distribution of the standardized absolute residuals that are computed using the initial robust estimators of regression parameters and scale. Because of the adaptive cut-off, the method is asymptotically equivalent to OLS and hence its full asymptotic efficiency.

4.12 Summary

In summary, all these least squares alternatives have some drawbacks. M-estimation is robust to outliers in the response variable, but it is not resistant to outliers in the explanatory variables (leverage points). Thus, the method has the same bdp as OLS. LAD, R-estimate and GM-estimator suffer the same low bdp. To overcome the lack of resistance, LTS, LMS, S-estimates, MM-estimates, covariance approach with S-estimator and REWLSE are viable alternatives. However, LTS, LMS and S-estimate have low efficiency. GS-estimate improves the efficiency compared to S-estimator but not high enough. MM-estimator, covariance approach with S-estimator and REWLSE estimator seem to be the best alternatives because of their high bdp and high asymptotic efficiency.

It is important to note that even though some mentioned estimators have high bdp, their computation is very challenging, especially in case of large data-sets or high dimension. That is why *approximate* algorithms have to be used for this task. The problem is that this results in worse performance about consistency and bdp, than the exact theoretical estimator would have had. And it gets worse with the increase of the sample size n and/or the dimension p of the samples (Stromberg et al. [2000], Hawkins and Olive [2002]).

Furthermore, with all these methods, there always have to be a decision of which tuning constant choose, which function of the residuals should be used, which first initial estimator use. The problem becomes complicated with all of these decisions in case of real data.

CHAPTER 5

Robust regression based on shrinkage

In Chapter 3, the notion of shrinkage was used to define robust estimators for location and covariance matrix, with the goal of using them to define robust Mahalanobis distances. In summary, from the collection of RMD's, the RMD-S approach was selected as the one with the best performance through the simulation study. In the present chapter, the estimation of the regression parameters in the linear regression model, using these robust estimators based on shrinkage, is proposed.

5.1 Shrinkage reweighted regression estimator

Denote the joint variable of the response and carriers as $\mathbf{z} = (\mathbf{x}, \mathbf{y})$. Denote the location of $\mathbf{z}$ by $\boldsymbol{\mu}$ and the scatter matrix by Σ. Partitioning $\boldsymbol{\mu}$ and Σ yields the notation:

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_x \\ \mu_y \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}. \tag{5.1}$$

In Chapter 1, it was mentioned that OLS estimator can be expressed equivalently as:

$$\hat{\boldsymbol{\beta}} = \hat{\Sigma}_{xx}^{-1}\hat{\Sigma}_{xy}, \qquad \hat{\alpha} = \hat{\mu}_y - \hat{\boldsymbol{\beta}}^t \hat{\boldsymbol{\mu}}_x \,. \tag{5.2}$$

Robust estimates should be used in Equation 5.2. The shrinkage estimators for the location and covariance matrix of $\mathbf{z}$ are used for this purpose. They are defined in Chapter 3 in Equations 3.14 and 3.15, respectively. Let us denote them as $\hat{\boldsymbol{\mu}}_{Sh}$ and $\hat{\Sigma}_{Sh}$, respectively, for simplicity and let us call them the *initial shrinkage robust estimators* of central tendency and covariance matrix of $\mathbf{z}$. These robust estimators define the RMD-S. For each observation $\mathbf{z}_i$, with $i = 1, ..., n$:

$$d^2(\mathbf{z}_i) = RMD\text{-}S(\mathbf{z}_i) = (\mathbf{z}_i - \hat{\boldsymbol{\mu}}_{Sh})\hat{\Sigma}_{Sh}^{-1}(\mathbf{z}_i - \hat{\boldsymbol{\mu}}_{Sh})^t \,. \tag{5.3}$$

Since in Chapter 3, RMD-S shows to be a robust and advantageous method for outlier detection, a weight function can be defined depending on the robust squared

Mahalanobis distance $w_i = w(d^2(\mathbf{z}_i))$. The second step is to obtain $\hat{\boldsymbol{\mu}}_{Sh}^{SW}$ and $\hat{\Sigma}_{Sh}^{SW}$, the *shrinkage weighted estimator for the mean and covariance matrix:*

$$\hat{\boldsymbol{\mu}}_{Sh}^{SW} = \frac{\sum_{i=1}^{n} w_i \mathbf{z}_i}{\sum_{i=1}^{n} w_i}, \quad \hat{\Sigma}_{Sh}^{SW} = \frac{\sum_{i=1}^{n} w_i (\mathbf{z}_i - \hat{\boldsymbol{\mu}}_{Sh}^{SW})(\mathbf{z}_i - \hat{\boldsymbol{\mu}}_{Sh}^{SW})^t}{\sum_{i=1}^{n} w_i} . \tag{5.4}$$

Based on $\hat{\boldsymbol{\mu}}_{Sh}^{SW}$ and $\hat{\Sigma}_{Sh}^{SW}$ we can obtain $\hat{\boldsymbol{\beta}}^{SW}$ and $\hat{\alpha}^{SW}$ which are initial estimates for the regression parameters. Let us call them *shrinkage weighted (SW) regression estimators*:

$$\hat{\boldsymbol{\beta}}^{SW} = (\hat{\Sigma}_{Sh}^{SW})_{xx}^{-1} (\hat{\Sigma}_{Sh}^{SW})_{xy}, \quad \hat{\alpha}^{SW} = (\hat{\boldsymbol{\mu}}_{Sh}^{SW})_y - (\hat{\boldsymbol{\beta}}^{SW})^t (\hat{\boldsymbol{\mu}}_{Sh}^{SW})_x . \tag{5.5}$$

The SW estimate of the residual's scale is:

$$\hat{\sigma}^{SW} = (\hat{\Sigma}_{Sh}^{SW})_{yy} - (\hat{\boldsymbol{\beta}}^{SW})^t (\hat{\Sigma}_{Sh}^{1})_{xx} \hat{\boldsymbol{\beta}}^{SW} .$$

The third step is reweighting, taking into consideration the residuals based on the SW regression estimators:

$$r_i^{SW} = y_i - (\hat{\boldsymbol{\beta}}^{SW})^t \mathbf{x}_i - \hat{\alpha}^{SW} . \tag{5.6}$$

Define the Mahalanobis distance for the SW residuals:

$$d(r_i^{SW}) = ((r_i^{SW})^t (\hat{\sigma}^{SW})^{-1} r_i^{SW})^{1/2} . \tag{5.7}$$

Let $wr_i = w(d^2(r_i^{SW}))$ a weighting function that depends on the Mahalanobis distance of the SW residuals. Note that the weights now depend on the size of the residual distance. This way, the good leverage points, which are observations with large distance in the x-space but small residual distance, are no longer outliers for the regression model because they are not downweighted. This is a necessary step in the regression problem because these types of observations are not outliers in the regression sense, even if they are atypical in the multivariate space.

Now, define $\mathbf{u}_i = (\mathbf{x}_i^t, 1)^t$ and obtain:

$$\hat{\boldsymbol{\varphi}}^{SR} = ((\hat{\boldsymbol{\beta}}^{SR})^t, \hat{\alpha}^{SR})^t = \left(\sum_{i=1}^{n} wr_i \mathbf{u}_i \mathbf{u}_i^t \right)^{-1} \sum_{i=1}^{n} wr_i y_i \mathbf{u}_i . \tag{5.8}$$

Then, $\hat{\boldsymbol{\varphi}}^{SR} = \left(\left(\hat{\boldsymbol{\beta}}^{SR} \right)^t, \hat{\alpha}^{SR} \right)^t$ are the *shrinkage reweighted (SR) regression estimators.*

For the weighting functions the inverse of the squared robust Mahalanobis distance was studied, but the indicator function in both cases (as in Rousseeuw et al. [2004]) had improved performance. The first weight function is $w_i = w(d^2(\mathbf{z}_i)) = I(d^2(\mathbf{z}_i) \leq q_1)$, which assigns weight 1 to the $\mathbf{z}_i$, for $i = 1, ..., n$, with a robust squared Mahalanobis distance less than certain quantile q_1 of the chi-square distribution with

$p+1$ degrees of freedom. The second weighting function is $wr_i = w(d^2(r_i^{SW})) = I(d^2(r_i^{SW}) \leq q_2)$, which assigns weight 1 to the residuals r_i^{SW} with a Mahalanobis distance less than certain quantile q_2 of the chi-square distribution with 1 degree of freedom.

The quantiles:

$$q_1 = \chi^2_{p+1,1-\delta_1} \quad \text{and} \quad q_2 = \chi^2_{1,1-\delta_2}\,, \tag{5.9}$$

depend on the significance levels δ_1 and δ_2, for which 0.025 and 0.01 are chosen, respectively, as in Rousseeuw et al. [2004], because those are the classical choices for the threshold to detect outliers (Leroy and Rousseeuw [1987]).

5.2 Simulation structure

In this section, a simulation study is conducted to investigate the performance of the proposed SR regression estimator and compare it with OLS and some of the previously mentioned robust regression methods: LTS, MM, method S and REWLSE. The simulations were done in Matlab. The *fitlm* function was used for OLS. The *ltsregres* function from LIBRA library (see Verboven and Hubert [2005]) considering the default option for the proportion of trimming which is $h = n/2 + 1$ and the default fraction of outliers the algorithm should resist which is equal to 0.75, was used for LTS. MM was computed with the *MMreg* function from the FSDA toolbox (see Riani et al. [2012]), with default values for the nominal efficiency: 0.95 and the rho function to weight the residuals as the bisquare which uses Tukey's functions. Method S was calculated with the function *SEst* from the Discriminant Analysis Programme toolbox which computes the biweight multivariate S-estimator for location and dispersion (see Ruppert [1992]). REWLSE was computed with the functions that Gervini and Yohai [2002] kindly provided, with hard rejection weights and starting from an initial S-estimator.

Consider the linear regression model in matrix form:

$$\mathbf{y} = \alpha + X\boldsymbol{\beta} + \boldsymbol{\epsilon}\,, \tag{5.10}$$

where X is of size $n \times p$, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^t$ is the unknown $p \times 1$ vector of regression parameters, α the unknown intercept, and the errors $\boldsymbol{\epsilon}$ are i.i.d and independent from the carriers. The independent variables are distributed according to a multivariate standard Gaussian distribution $X \sim N(\mathbf{0}_p, I_p)$, where $\mathbf{0}_p$ is the $p-$dimensional vector of zeros and I_p is the $p-$dimensional identity matrix. The simulation parameters are the following sets of dimension and sample size: $p = 5$ with $n = 20, 30, 50, 100, 1000$, $p = 20$ with $n = 80, 100, 200, 500, 5000$ and $p = 30$ with $n = 100, 150, 300, 500, 5000$. The simulations are repeated $M = 1000$ times and each time the parameter estimates are drawn anew.

Three simulation scenarios are proposed, analogously as the simulation models found in the literature (Maronna and Morgenthaler [1986], Gervini and Yohai [2002], Croux et al. [2003], Rousseeuw et al. [2004], Agulló et al. [2008], Yu and Yao [2017]).

(NE): The response is generated from a standard Normal distribution $N(0, I)$, which corresponds to putting $\boldsymbol{\beta} = \mathbf{0}$ and $\alpha = 0$ when Gaussian errors are considered.

(TE): The response is generated from a t-distribution with 3 d.f, which corresponds to putting $\boldsymbol{\beta} = \mathbf{0}$ and $\alpha = 0$ when t_3-distributed errors are considered.

(NEO): Normal errors as in [NE], but with probability δ the randomly selected observations in the independent variables were generated as $N(\lambda\sqrt{\chi^2_{p,0.99}}, 1.5)$ and the new response as $N(k\sqrt{\chi^2_{1,0.99}}, 1.5)$ where $\lambda, k = 0, 0.5, 1, 1.5, 2, 3, 4, 5, 6, 7, 8, 9, 10$.

For the last simulation scenario [NEO], the levels of contamination considered were $\delta = 10\%, 20\%$. Note that if $\lambda = 0$ and $k > 0$ we obtain *vertical outliers*, if $\lambda > 0$ and $k = 0$ we obtain *good leverage* points and if $\lambda > 0$ and $k > 0$ we obtain *bad leverage* points. On the other hand, large values of λ and k produce extreme outliers, whereas small values produce intermediate outliers (see Croux et al. [2003] and Agulló et al. [2008]).

5.3 Efficiency

It is known that under simulation scheme [NE] the OLS estimator has maximum efficiency. The efficiency for each robust estimator, for finite samples, is calculated relative to OLS, considering the sum of squared deviations from the true coefficients and averaging over all repetitions. Consider the joint vector of regression parameters including the intercept $\boldsymbol{\varphi} = (\boldsymbol{\beta}^t, \alpha)^t$, which has dimension $(p+1) \times 1$. For a certain robust method R, the finite sample efficiency for the joint estimator $\hat{\boldsymbol{\varphi}}_R$ is defined as:

$$\text{Eff} = \frac{1/M \sum_{m=1}^{M} ||\hat{\boldsymbol{\varphi}}_{OLS}^{(m)} - \boldsymbol{\varphi}||_2^2}{1/M \sum_{m=1}^{M} ||\hat{\boldsymbol{\varphi}}_{R}^{(m)} - \boldsymbol{\varphi}||_2^2}. \tag{5.11}$$

Table 5.1 shows the simulated efficiencies relative to OLS, for the joint regression estimator $\hat{\boldsymbol{\varphi}}$ obtained with the proposed approach SR and the other robust regression methods, under simulation scheme [NE]. In each row, the bold letter represents the higher efficiency, and the italic letter represents the lowest efficiency. The results show that for a fixed dimension when the sample size is increased, all methods improve the resulting finite sample efficiency. LTS is the method that behaves poorly even when the sample size increases. S, REWLSE and MM require large samples in order to have efficiencies higher than 90%. The proposed method SR has higher efficiency for every dimension and sample sizes considered.

Table 5.1: Finite sample efficiency in case of Normal errors, scenario [NE]

$p = 5$	n	SR	LTS	S	REWLSE	MM
	20	**0.9182**	0.2352	0.2715	*0.2346*	0.2272
	30	**0.9828**	*0.3486*	0.4292	0.5026	0.4915
	50	**0.9833**	*0.5061*	0.5070	0.5129	0.5047
	100	**0.9839**	*0.5870*	0.7051	0.7441	0.7192
	1000	**0.9859**	*0.7816*	0.8691	0.9570	0.9159
$p = 20$	80	**0.9852**	0.3763	0.6786	*0.2809*	0.2963
	100	**0.9956**	*0.3973*	0.7966	0.5028	0.4955
	200	**0.9900**	*0.4971*	0.8630	0.5811	0.8015
	500	**0.9951**	*0.6163*	0.8719	0.8737	0.8393
	5000	**0.9981**	*0.6822*	0.9461	0.9611	0.9068
$p = 30$	100	**0.9900**	0.4458	0.5068	0.3622	*0.2978*
	150	**0.9927**	0.4699	0.5155	*0.4347*	0.5532
	300	**0.9933**	*0.5110*	0.5187	0.7524	0.5770
	500	**0.9970**	*0.6467*	0.8660	0.8479	0.8486
	5000	**0.9980**	*0.6504*	0.9646	0.9863	0.9781

In the simulation scenario [TE], OLS is not a maximum efficient estimator, due to the heavy-tailed errors. Therefore, Table 5.2 shows the mean squared errors (MSE) instead. The results show that, for all methods, a large sample size translates into a decrease of the MSE, but method SR outperformed, in general, the other alternatives.

Table 5.2: MSE in case of $t-$student distributed errors, scenario [TE]

$p = 5$	n	SR	LTS	S	REWLSE	MM
	20	**0.1499**	0.2980	0.3634	*0.4892*	0.3193
	30	**0.0579**	0.0745	0.0662	*0.1074*	0.0713
	50	**0.0304**	0.0479	0.0409	*0.0548*	0.0322
	100	**0.0114**	0.0125	*0.0150*	0.0115	0.0116
	1000	**0.0012**	0.0016	0.0015	*0.0017*	0.0014
$p = 20$	80	**0.0244**	0.0443	0.0293	*0.1218*	0.0881
	100	**0.0126**	0.0376	0.0228	*0.0720*	0.0364
	200	**0.0107**	0.0108	0.0114	0.0117	*0.0118*
	500	**0.0033**	*0.0039*	0.0036	*0.0039*	0.0034
	5000	**0.0003**	*0.0004*	*0.0004*	**0.0003**	**0.0003**
$p = 30$	100	**0.0202**	0.0637	0.0375	*0.1767*	0.0855
	150	**0.0110**	0.0208	0.0157	*0.0328*	0.0240
	300	**0.0052**	0.0067	0.0074	*0.0075*	0.0055
	500	**0.0032**	*0.0040*	0.0038	0.0039	0.0033
	5000	**0.0003**	*0.0005*	*0.0005*	**0.0003**	**0.0003**

5.4 Robustness

Simulations to study the robustness are carried out, considering the third simulation scheme [NEO]. The most significant results are those consisting on dimensions $p = 5, 30$ with sample sizes $n = 100, 500$, respectively. The two statistical criteria used to compare the estimators from the different approaches were the squared Bias and the MSE for the estimated parameter vector $\hat{\boldsymbol{\beta}}$ and for the estimated intercept $\hat{\alpha}$ averaging over all M simulation runs (see Gervini and Yohai [2002], Croux et al. [2003], Rousseeuw et al. [2004]). The following figures show, for each value of λ, the maximal value of MSE or Bias, obtained over all possible values of k.

$$\begin{aligned} MMSE_{\lambda}(\cdot) &= max_{k\in\{0,\ldots,10\}} MSE_{\lambda,k}(\cdot) \\ MBias_{\lambda}(\cdot) &= max_{k\in\{0,\ldots,10\}} Bias_{\lambda,k}(\cdot)\,, \end{aligned} \tag{5.12}$$

for each $\lambda \in \{0, ..., 10\}$. Figure 5.1 shows the $MMSE(\hat{\boldsymbol{\beta}})$, in case of low dimension $p = 5$ with sample size $n = 100$ and when the data is contaminated with a level of 10%. OLS shows high MSE when the data contains atypical observations, especially for vertical outliers and bad leverage observations associated with the first values of λ.

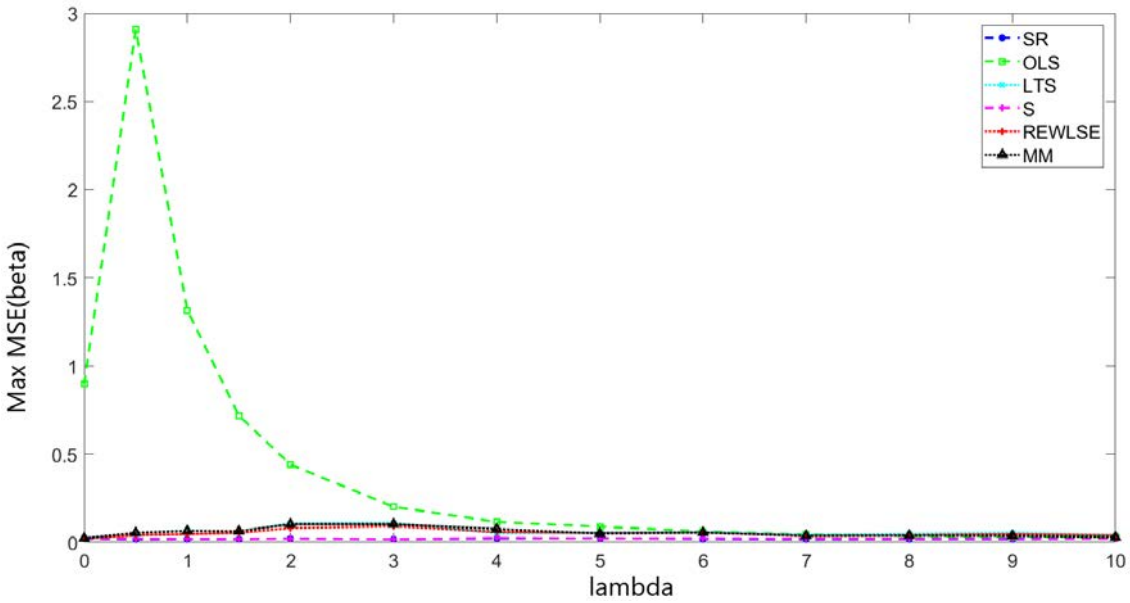

Figure 5.1: $MMSE(\hat{\boldsymbol{\beta}})$ with $p = 5$, $n = 100$, $\delta = 10\%$.

If the previous image is zoomed, Figure 5.2, it can be seen that for vertical outliers, i.e., $\lambda = 0$, all robust methods have similar MSE, but for the remaining values of λ, the smallest errors correspond to the proposed method SR and method S.

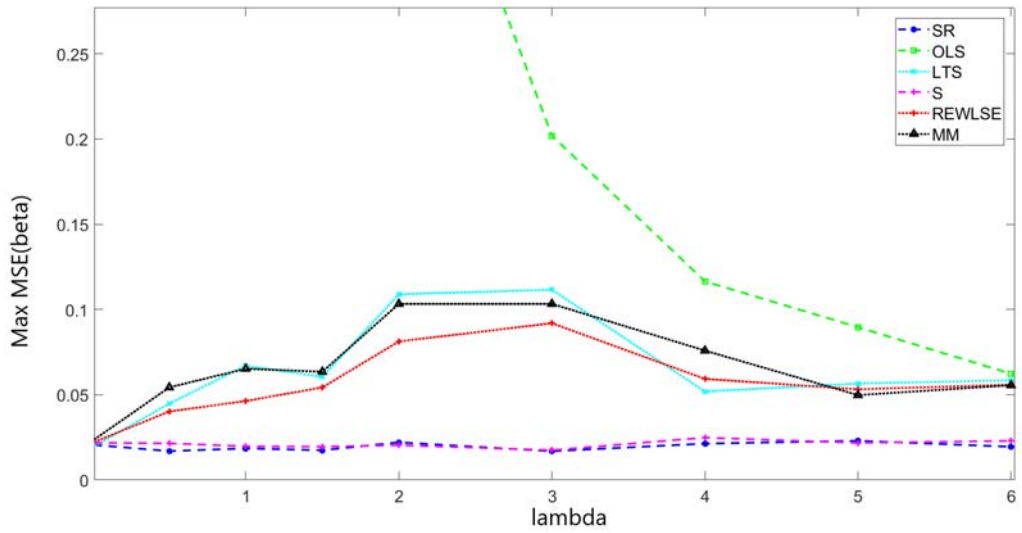

Figure 5.2: (Zoom) $MMSE(\hat{\boldsymbol{\beta}})$ with $p = 5$, $n = 100$, $\delta = 10\%$.

For the MSE of $\hat{\alpha}$, and for the Bias of both $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$, similar conclusions are obtained. In order to see these results from a different perspective, the error measures are summarized in a single graph for each dimension, sample size and contamination level. Figure 5.3 corresponds to $p = 5$, $n = 100$ and $\delta = 10\%$. Each line represents a method. In the x-axis each number from 1 to 4 represents the maximum error measures: 1-MMMSE($\hat{\boldsymbol{\beta}}$), 2-MMMSE($\hat{\alpha}$), 3-MMBias($\hat{\boldsymbol{\beta}}$) and 4-MMBias($\hat{\alpha}$), over all possible values of λ.

$$\begin{aligned} MMMSE(\cdot) &= max_{\lambda\in\{0,\ldots,10\}} MMSE_{\lambda}(\cdot) \\ MMBias(\cdot) &= max_{\lambda\in\{0,\ldots,10\}} MBias_{\lambda}(\cdot)\,, \end{aligned} \tag{5.13}$$

for each $\lambda \in \{0, ..., 10\}$.

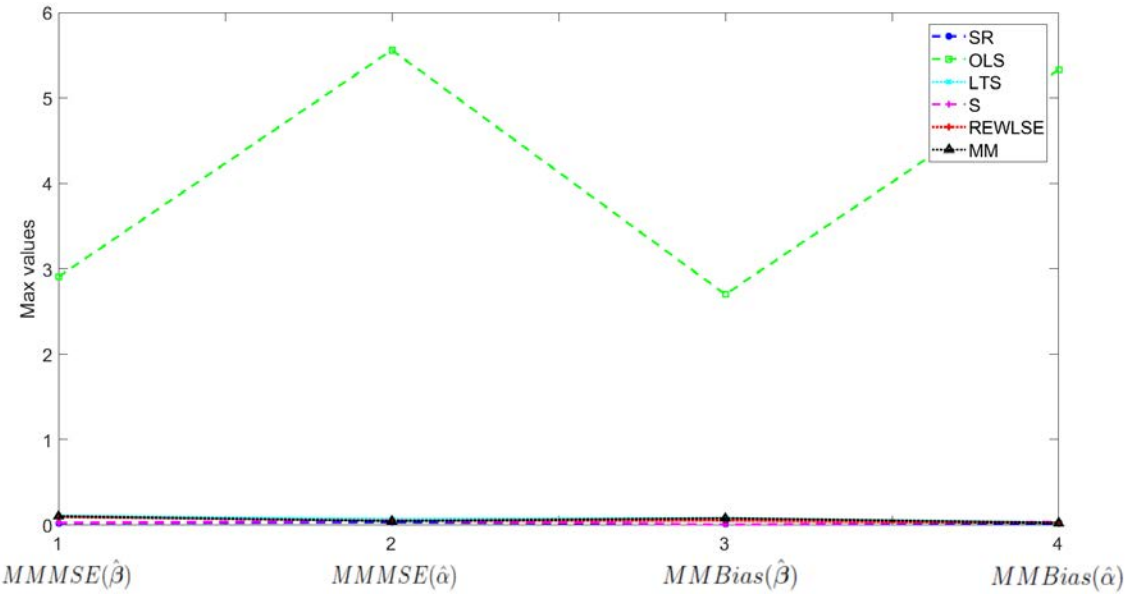

Figure 5.3: $MMMSE$ and $MMBias$, with $p = 5$, $n = 100$ and $\delta = 10\%$.

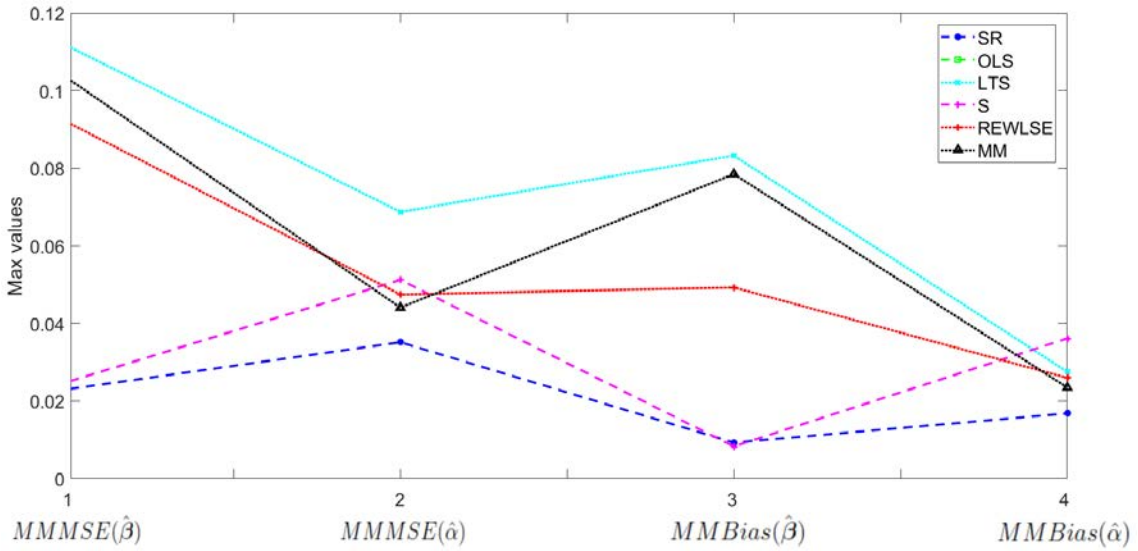

Figure 5.4: (Zoom) $MMMSE$ and $MMBias$, with $p = 5$ and $\delta = 10\%$.

Figure 5.4 is a zoom of the previous Figure 5.3. We can see in Figure 5.4 that in the majority of cases the proposed method SR has the lowest maximum MSE or Bias, except for one case in which method S has slightly lower maximum Bias($\hat{\boldsymbol{\beta}}$), but this happens only under low level of contamination.

When the contamination level δ increases to 20%, method S worsens its performance as it can be seen in Figure 5.5.

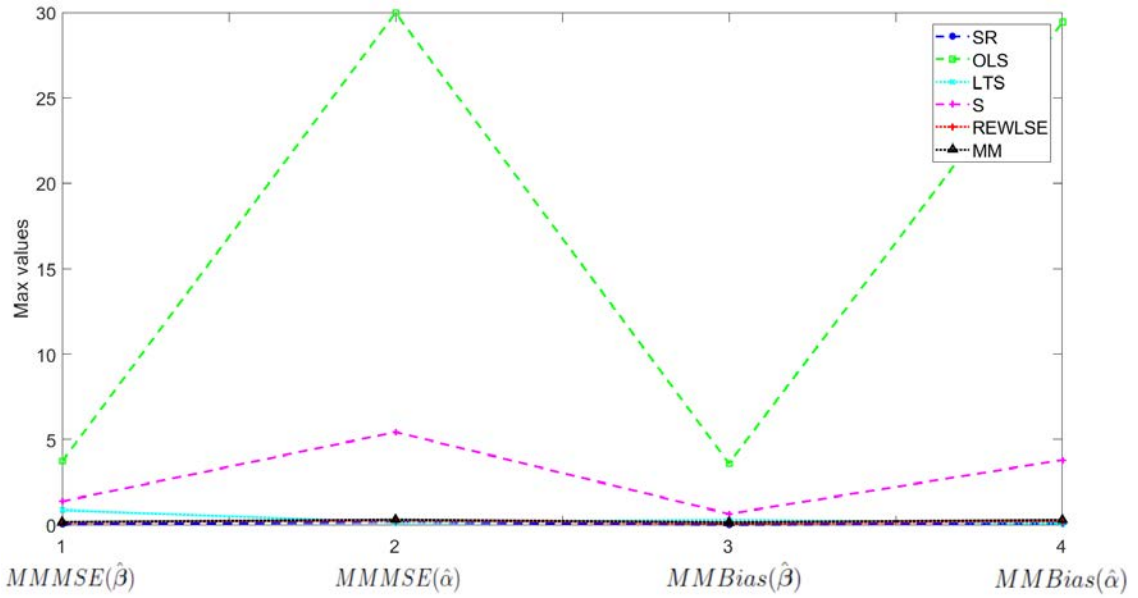

Figure 5.5: $MMMSE$ and $MMBias$, with $p = 5$ and $\delta = 20\%$.

Zoomed Figure 5.6 shows that, in case of higher contamination level, SR is the overall best performance method taking into account that although MSE($\hat{\alpha}$) and Bias($\hat{\alpha}$) are slightly lower for LTS, the MSE and Bias of the $\hat{\boldsymbol{\beta}}$ for LTS is much higher than SR, REWLSE and even MM estimator.

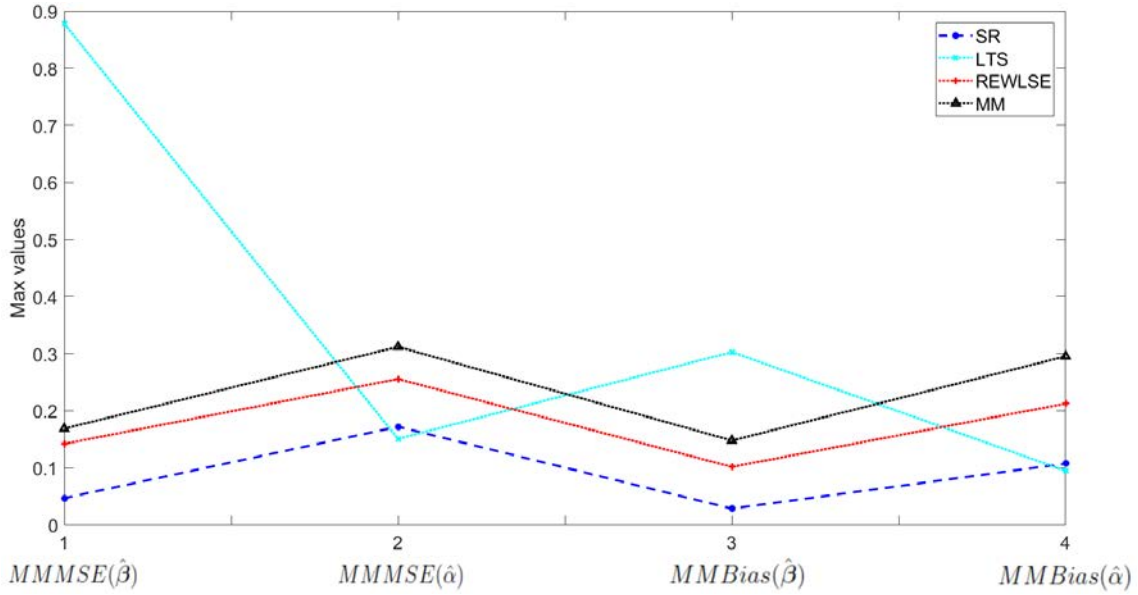

Figure 5.6: (Zoom) $MMMSE$ and $MMBias$, with $p = 5$ and $\delta = 20\%$.

Figure 5.7 shows that when the dimension is increased to $p = 30$, and the contamination is $\delta = 10\%$, the most affected methods are OLS and S. Method SR is the one that has the lowest maximum value for the MSE and Bias of both $\boldsymbol{\beta}$ and α.

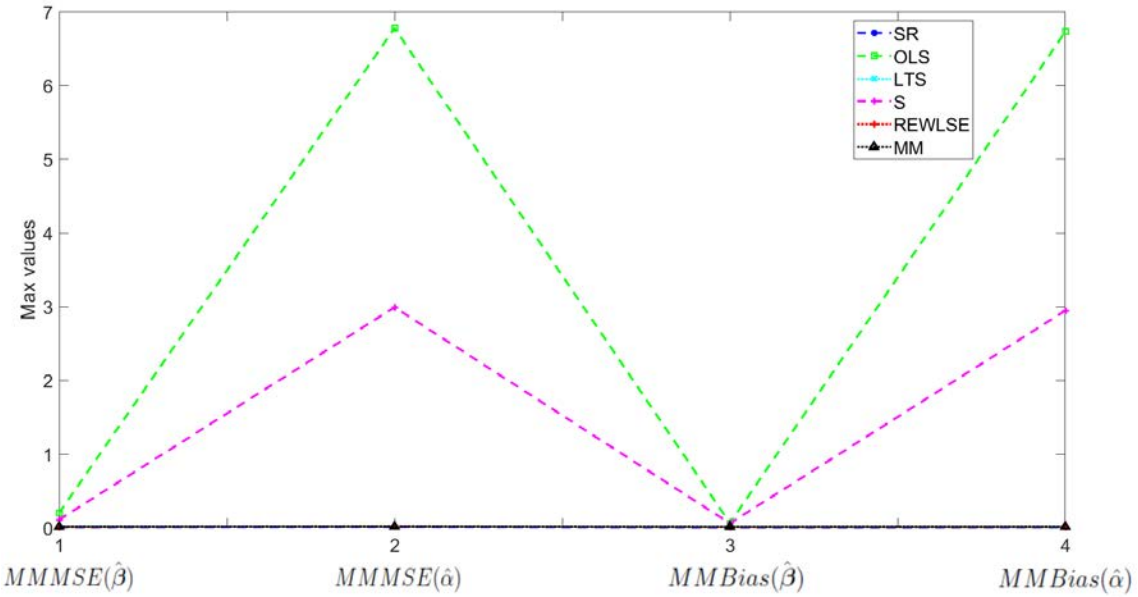

Figure 5.7: $MMMSE$ and $MMBias$, with $p = 30$ and $\delta = 10\%$.

Figure 5.8 is a zoom of Figure 5.7 so we can see the four methods with lowest errors. A similar situation happens in case of $\delta = 20\%$ of contamination.

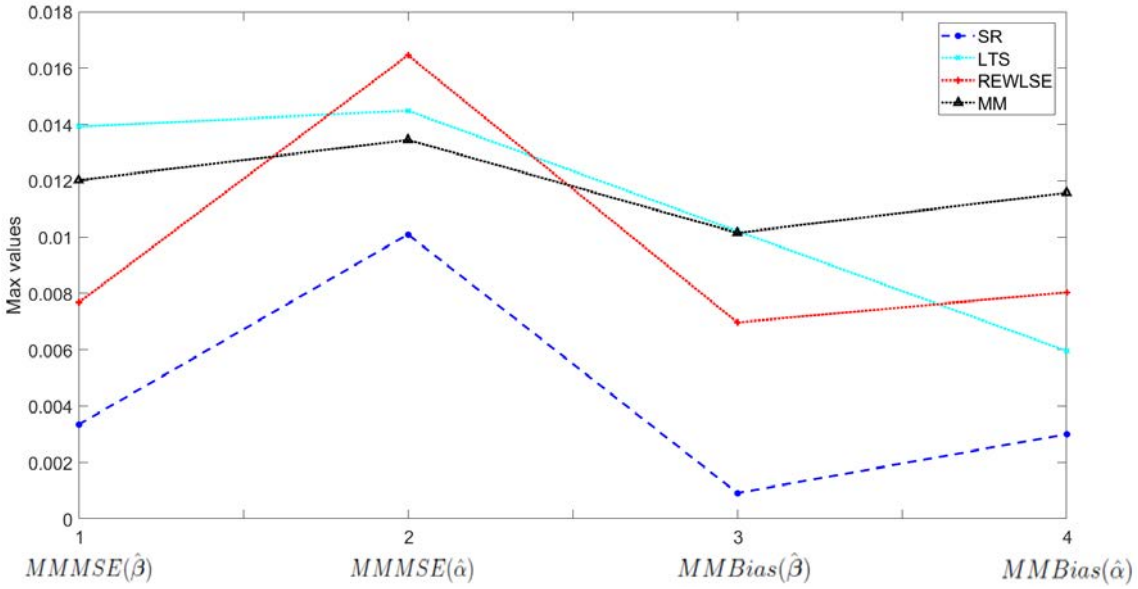

Figure 5.8: (Zoom) $MMMSE$ and $MMBias$, with $p = 30$ and $\delta = 10\%$.

Appendix D contains the Tables with the numerical results, showing for each method the maximum (across λ and k) MSE and Bias for both $\hat{\boldsymbol{\beta}}$ and $\hat{\alpha}$ for each combination of the dimension p and the contamination level δ. In bold letter are the lowest error and in italic letter are the highest error after OLS.

5.4.1 Computational times

The computational times in seconds for each method in the simulation scenario [NEO] are also measured. The study was performed in a PC with a 3.40 GHz Intel Core i7 processor with 32GB RAM. The results are averaged for 10% and 20% of contamination since they were similar. OLS is the fastest one because of its simplicity. Following OLS, the proposed method SR is the second fastest method because it does not relies on iterative algorithms to calculate the estimations. The other robust alternatives are between 3 and 9 times slower than our proposal SR for low dimension, and between 3 and 12 times slower for higher dimension.

Table 5.3: Computational times with Normal distribution $p = 5$ and $n = 100$

α	SR	OLS	LTS	S	REWLSE	MM
0.1	0.0206	0.0126	0.0989	0.0515	0.0572	0.1816
0.2	0.0200	0.0102	0.0966	0.0514	0.0545	0.1862

Table 5.4: Computational times with Normal distribution $p = 30$ and $n = 500$

α	SR	OLS	LTS	S	REWLSE	MM
0.1	0.1246	0.0120	0.4350	0.3825	0.3967	1.5263
0.2	0.1209	0.0104	0.4102	0.3820	0.4192	1.5456

5.5 Equivariance properties

The initial shrinkage robust estimators $\hat{\boldsymbol{\mu}}_{Sh}$ and $\hat{\Sigma}_{Sh}$ are approximately affine equivariant, as it was studied in Chapter 3. This means that the equivariance property cannot be demonstrated analytically because only part of the property holds, but it can be studied by means of simulations (as in Maronna and Zamar [2002] and Sajesh and Srinivasan [2012]). Then, the distance defined in Equation 5.3 and used in the weights for the SW estimators of mean and covariance matrix (Equation 5.4) remains approximately invariant under affine transformations. Since the weights considered are hard rejection depending on an approximately invariant robust distance, the estimators $\hat{\boldsymbol{\mu}}_{Sh}^{SR}$ and $\hat{\Sigma}_{Sh}^{SR}$ should hold the property.

Thus, we propose to study the equivariance property on the parameter estimators, denoted as:

$$\hat{\boldsymbol{\varphi}}^{SR} = \left(\left(\hat{\boldsymbol{\beta}}^{SR} \right)^t, \hat{\alpha}^{SR} \right)^t .$$

Affine equivariance in regression can be split in the three following properties (Rousseeuw et al. [2004] and Maronna and Morgenthaler [1986]):

1. **Regression equivariance:** If a linear function of the explanatory variables is added to the response, then the coefficients of this linear function are also added to the estimators.

2. **y-equivariance:** If the response variable is transformed linearly, then the estimators transform correctly.

   Property (1) and (2) can be seen together as:

   $$\hat{\boldsymbol{\varphi}}^{SR}(X, \mathbf{y}c + X\mathbf{g} + v) = \hat{\boldsymbol{\varphi}}^{SR}(X, \mathbf{y})c + (\mathbf{g}^t, v)^t , \tag{5.14}$$

   where $c \in \mathbb{R}$ is any non-zero constant, $\mathbf{g}$ is any $p \times 1$ vector and $v \in \mathbb{R}$ is any constant. This means that, keeping the same X, and transforming the response as $\mathbf{y}c + X\mathbf{g} + v$, the resulting transformed estimators are: $\hat{\boldsymbol{\beta}}_{new}^{SR} = c(\hat{\boldsymbol{\beta}}^{SR}) + \mathbf{g}$ and $\hat{\alpha}_{new}^{SR} = c\hat{\alpha}^{SR} + v$.

3. **x-equivariance:** Also called *carrier equivariance*. It says that if the explanatory variables are transformed linearly (coordinate system transformation), then the estimators transform correctly.

   $$\hat{\boldsymbol{\varphi}}^{SR}(XA, \mathbf{y}) = ((\hat{\boldsymbol{\beta}}^{SR})^t (A^{-1})^t, \hat{\alpha}^{SR})^t . \tag{5.15}$$

   This means that if the carriers are transformed as XA with any non-singular $p \times p$ matrix A, the resulting estimators are: $\hat{\boldsymbol{\beta}}_{new}^{SR} = A^{-1}\hat{\boldsymbol{\beta}}^{SR}$ and the intercept should remain the same $\hat{\alpha}_{new}^{SR} = \hat{\alpha}^{SR}$.

Exploring all possible transformations is infeasible, that is the reason why Maronna and Zamar [2002] and Sajesh and Srinivasan [2012] proposed to generate the random matrices A for the **x-equivariance** as $A = TD$, where T is a random orthogonal matrix and $D = diag(u_1, ..., u_p)$, where the u_j's are independent and uniformly distributed in $(0, 1)$, for all $j = 1, ..., p$. Then, each generated data matrix X in each repetition, is transformed with a random transformation A. We propose to generate A as the authors indicated, and following this idea generate randomly for each repetition the the non-zero c, the $\mathbf{g}$ and the v for the equivariance properties.

The MSE of the proposed method SR is studied when the transformations described above are made to the simulated data-set. Consider the simulation scenario [NE] for Normal data without outliers ($\delta = 0\%$) and scenario [NEO] when there is $\delta = 10\%, 20\%$ of contamination, to see the impact of the presence of outliers. The vector of regression parameters $\hat{\boldsymbol{\varphi}}^{SR}$ is estimated with the untransformed data and saved. After that, the data is transformed according to Equation 5.14 for the regression and **y**-equivariance and according to Equation 5.15 for the **x**-equivariance. Next, the method SR is applied to the transformed data and the resulting $\hat{\boldsymbol{\varphi}}^{SR}_{new}$ are saved. The MSE is calculated between the obtained $\hat{\boldsymbol{\varphi}}^{SR}_{new}$ and what it should be obtained if the equivariance properties hold. Table 5.5 shows for each λ, the resulting $MMSE_\lambda(\hat{\boldsymbol{\varphi}}^{SR}_{new})$ in case of regression and **y**-equivariance. For vertical outliers, i.e., when $\lambda = 0$, the error increases with the increase in dimension and contamination level, a fact that is influenced mostly by the error of the intercept. Nevertheless, for the rest of the cases, the maximum possible error is low.

Table 5.5: $MMSE_\lambda(\hat{\boldsymbol{\varphi}}^{SR}_{new})$ for regression and **y**-equivariance

	$p = 5$			$p = 30$		
λ	$\delta = 0\%$	$\delta = 10\%$	$\delta = 20\%$	$\delta = 0\%$	$\delta = 10\%$	$\delta = 20\%$
0	0.01205	0.04173	0.12625	0.00006	0.26366	0.30312
0.5	0.00567	0.01994	0.03135	0.00009	0.00267	0.00085
1	0.00645	0.01206	0.00876	0.00005	0.00272	0.00066
1.5	0.00615	0.00924	0.00373	0.00009	0.00428	0.00046
2	0.00686	0.00822	0.00384	0.00008	0.00156	0.00037
3	0.01718	0.00521	0.00454	0.00008	0.00215	0.00057
4	0.00726	0.00905	0.00756	0.00008	0.00298	0.00068
5	0.00863	0.01228	0.00737	0.00007	0.00208	0.00063
6	0.00586	0.01305	0.00677	0.00004	0.00166	0.00034
7	0.00822	0.00934	0.00550	0.00003	0.00265	0.00044
8	0.00707	0.01955	0.00628	0.00007	0.00227	0.00056
9	0.00545	0.00948	0.01328	0.00002	0.00306	0.00077
10	0.00676	0.02298	0.00686	0.00009	0.00409	0.00037

Table 5.6 shows the results for the **x**-equivariance. In this case, both for vertical outliers and leverage points, the error remains low. Thus, since the errors are mostly controlled, the proposed robust regression estimator is approximately regression, y- and x-equivariant.

Table 5.6: $MMSE_\lambda(\hat{\boldsymbol{\varphi}}_{new}^{SR})$ for **x**-equivariance

	$p=5$			$p=30$		
λ	$\delta=0\%$	$\delta=10\%$	$\delta=20\%$	$\delta=0\%$	$\delta=10\%$	$\delta=20\%$
0	0.00206	0.00421	0.01874	0.00005	0.01324	0.09468
0.5	0.00162	0.00456	0.01310	0.00003	0.00026	0.00008
1	0.00178	0.00348	0.00493	0.00003	0.00030	0.00003
1.5	0.00153	0.00392	0.00132	0.00004	0.00012	0.00006
2	0.00198	0.00320	0.00234	0.00003	0.00034	0.00003
3	0.00144	0.00293	0.00208	0.00003	0.00016	0.00002
4	0.00177	0.00329	0.00359	0.00005	0.00026	0.00005
5	0.00194	0.00339	0.00182	0.00003	0.00020	0.00001
6	0.00173	0.00481	0.00205	0.00005	0.00016	0.00002
7	0.00214	0.00329	0.00184	0.00002	0.00012	0.00002
8	0.00186	0.00415	0.00177	0.00004	0.00013	0.00002
9	0.00242	0.00356	0.00188	0.00004	0.00016	0.00001
10	0.00193	0.00287	0.00250	0.00003	0.00011	0.00001

5.6 Breakdown property

The bdp measures the maximum proportion of outliers that the estimator can safely tolerate. The highest possible value for the bdp is 50%. The empirical breakdown value can be examined through simulations, as in Sajesh and Srinivasan [2012], considering high contamination levels. Although these situations are not that relevant in practice because low levels of contamination should be expected, we propose to study if the error and the bias are controlled in these scenarios in order to see the performance of the proposed SR estimator. For this, [NEO] contamination scheme is used, but considering higher levels of contaminations $\delta = 30\%, 40\%, 45\%$. Table 5.7 shows the resulting MMMSE and MMBias for $\hat{\boldsymbol{\varphi}}_{new}^{SR}$ in the low dimension $p = 5$ case.

Table 5.7: MMMSE and MMBias, $p = 5$

	$\delta=30\%$		$\delta=40\%$		$\delta=45\%$	
Method	MMMSE	MMBias	MMMSE	MMBias	MMMSE	MMBias
OLS	6.9013	5.9143	7.5851	6.3344	7.6727	6.3215
SR	**0.1216**	**0.1160**	**0.2733**	**0.1343**	**0.3314**	**0.1301**
LTS	6.0032	5.6686	6.6864	6.3431	6.9428	6.4081
S	*6.0679*	*5.7893*	*7.2842*	*7.0237*	*7.2403*	*6.7814*
REWLSE	0.3251	0.2994	1.0797	0.7422	1.7883	1.0121
MM	0.5190	0.4884	1.4912	1.1475	3.6681	2.6982

Table 5.8 shows the results for higher dimension $p = 30$. The bold letter represents lower error or bias and the italic letter represents the highest measures after OLS, which is the method with worse results. LTS, S and MM have high error and bias for both low and high dimension, especially with the increase of the contamination level. REWLSE is competitive with SR in high dimension, but in low

dimension, REWLSE shows higher errors. The MSE and Bias of SR remain low, especially in high dimension and even with large contamination in the data, compared with the other robust methods supposedly having a high bdp. As discussed in Yu and Yao [2017] where the authors review and compare some robust regression approaches, the issue here is that although LTS, S and MM have high bdp, the computation is very challenging (Hawkins and Olive [2002] and Stromberg et al. [2000]). That is why resampling algorithms are used to obtain a number of subsets and then compute the robust regression estimate from some initial estimates. However, the high breakdown property usually requires that the number of elementary sets goes to infinity. For example, Hawkins and Olive [2002] proved that LTS computed with fast-LTS algorithm had zero bdp. In order to compute these estimators with high bdp, one should consider all possible elemental sets. SR approach shows high resistance to large contamination even in high dimension, which can be translated in high empirical bdp.

Table 5.8: MMMSE and MMBias, $p = 30$

	$\delta = 30\%$		$\delta = 40\%$		$\delta = 45\%$	
Method	MMMSE	MMBias	MMMSE	MMBias	MMMSE	MMBias
OLS	1.2970	1.0677	1.3839	1.0666	1.2738	1.0701
SR	**0.0131**	**0.0025**	**0.0642**	**0.0182**	**0.1138**	**0.0232**
LTS	*0.6640*	*0.1567*	*1.0824*	*0.2211*	*0.9589*	*0.1980*
S	0.2660	0.0678	0.3764	0.0665	0.3042	0.0749
REWLSE	0.0218	0.0034	0.0977	0.0310	0.2184	0.0630
MM	0.0732	0.0677	0.2274	0.0668	0.4012	0.0675

5.7 Real data-set examples

In this section, we study two known data-sets, very often used in the literature, to illustrate the performance of the proposed robust regression method comparing to the other robust alternatives. And also a socioeconomic and environmental related data-set that explains the Living Environment Deprivation of areas of Liverpool through remote-sensed data obtained from Google Earth technologies (Arribas-Bel et al. [2017]).

5.7.1 Star data

The first example is the star data-set, and it is reported in Leroy and Rousseeuw [1987], and based on Humphreys [1978] and De Grève and Vanbeveren [1980]. It has become a benchmark for robust regression methodologies. It consists on $n = 47$ observations corresponding to 47 stars of the CYG OB1 cluster in the direction of Cygnus. There is only one carrier x, which is the logarithm of the effective temperature at the surface of the star. The response variable y is the logarithm of its light intensity. There is a positive linear relationship between the response and the explanatory variables, except for four red giant stars (observations 11, 20, 30

and 34) which are outliers because they have low temperatures and a high output of light (the four observations on the upper left corner in Figure 5.9).

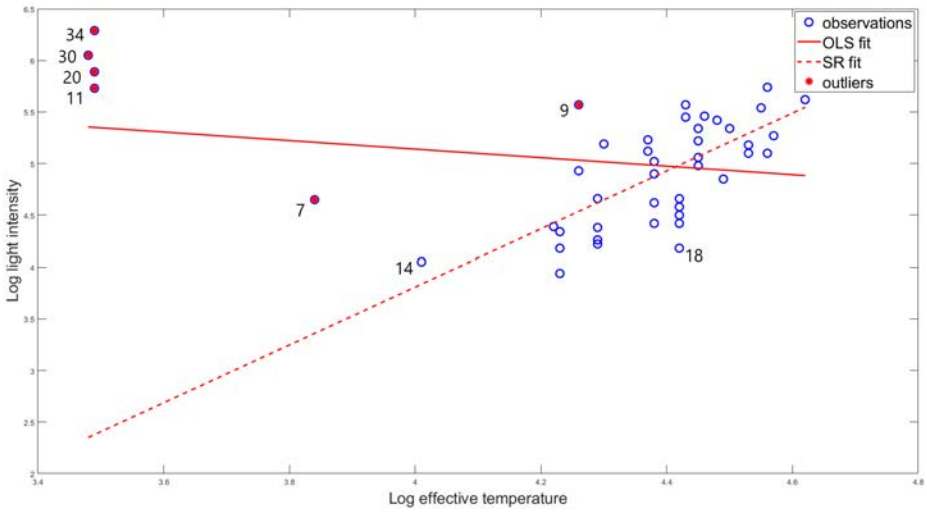

Figure 5.9: Star data-set with OLS and SR regression fit.

These giant stars represent a different population. They are bad leverage points because they influence the OLS regression line due to the poor estimation of the parameters. Figure 5.9 shows how the four giant stars pull the OLS line towards them. Observations 7 and 9 are intermediate outliers. And finally, in the multivariate sense, observation 14 is often detected as an outlier, but in the regression sense, it is a good leverage point because it follows the same linear pattern than the bulk data. Robust regression fit made by the proposed method SR detected the giant stars 11, 20, 30, 34 and the intermediate outliers 7 and 9.

Table 5.9 summarizes all method's estimation of the intercept and slope, and the outliers detected by the robust techniques. Note that OLS estimates are entirely changed, they have even different signs. SR and REWLSE correctly detect the regression outliers, method S detects the good leverage point, observation 14, as an outlier. LTS identifies observation 18 as atypical when it is not. In Figure 5.9, it can be seen that observation 18 is an example of the swamping effect problem. On the other hand, the MM approach only detects as outliers the giant stars (masking effect).

Table 5.9: Estimation of intercept and slope and detected outliers with star data.

Method	$\hat{\alpha}$	$\hat{\beta}$	Detected outliers
OLS	6.7935	-0.4133	
SR	-7.4035	2.9028	7 9 11 20 30 34
LTS	-8.5001	3.0462	7 9 11 18 20 30 34
S	-10.5034	3.4994	7 9 11 14 20 30 34
REWLSE	-7.5001	3.0462	7 9 11 20 30 34
MM	-5.1234	2.2879	11 20 30 34

The R^2 values for the linear regression models fitted by each method are summarized in Table 5.10. OLS's coefficient of determination is low, while that of the robust methods is high, except for the MM approach, which is lower than the rest.

Table 5.10: R^2 for each method with stars data-set.

Method	OLS	SR	LTS	S	REWLSE	MM
R^2	0.0443	0.7113	0.7006	0.7035	0.7095	0.5578

5.7.2 Hawkins-Bradu-Kass data

HBK data-set was artificially created by Hawkins et al. [1984] and it was also used in Leroy and Rousseeuw [1987], and many others. It contains $p = 3$ explanatory variables and a response variable. The first 14 observations are leverage points: 1-10 of bad type and 11-14 of good type. Thus, only observations 1-10 are outliers in the regression sense. Table 5.11 shows the estimation by all methods for the three parameters, and it can be seen that OLS is highly influenced by the presence of these leverage points. Also, the parameters estimated by S method are different than that of the other robust approaches, and the reason for this is that all robust methods correctly detect the true outliers, except for method S, which also includes the good leverage points 11-14.

Table 5.11: Estimation of the parameters and detected outliers with HBK data.

Method	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	Detected outliers
OLS	-0.3875	0.2392	-0.3345	0.3833	
SR	-0.1800	0.0836	0.0396	-0.0518	1 2 3 4 5 6 7 8 9 10
LTS	-0.1805	0.0814	0.0399	-0.0517	1 2 3 4 5 6 7 8 9 10
S	-0.0174	0.0957	0.0041	-0.1286	1 2 3 4 5 6 7 8 9 10 11 12 13 14
REWLSE	-0.1805	0.0814	0.0399	-0.0517	1 2 3 4 5 6 7 8 9 10
MM	-0.1913	0.0860	0.0412	-0.0541	1 2 3 4 5 6 7 8 9 10

The adjusted R^2 values are summarized in Table 5.12. Here, all robust methods, except S, have high and similar R^2.

Table 5.12: Adjusted R^2 for each method with HBK data-set.

Method	OLS	SR	LTS	S	REWLSE	MM
R^2	0.5850	0.9818	0.9816	0.9002	0.9817	0.9811

5.7.3 Living Environment Deprivation data

In Arribas-Bel et al. [2017], the authors studied the Living Environment Deprivation (LED) index. This measure allows studying quantitatively the concept of quality of the local environment, known also as urban quality of life, which is a qualitative concept. This is an essential matter for environmental research, citizens and politics. This kind of indices can be explained through remote sensing data, i.e.,

information collected without making physical contact, for example, from satellite technologies. The authors in Arribas-Bel et al. [2017] proposed to model the LED index of Liverpool (UK) based on four sets of explanatory variables extracted from a very high spatial resolution (VHR) image downloaded from Google Earth. The four groups are called: land cover (LC), spectral (SP), texture (TX) and structure features (ST). See Arribas-Bel et al. [2017] for more detailed description of the features. The authors first propose to explain the LED index with a linear combination of the four sets of variables. The linear regression model is the following:

$$LED = \alpha + \boldsymbol{\beta} LC + \boldsymbol{\gamma} SP + \boldsymbol{\delta} TX + \boldsymbol{\zeta} ST + \epsilon \,. \tag{5.16}$$

There are 35 explanatory variables, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\delta}$ and $\boldsymbol{\zeta}$ are vectors, containing the parameters for each carrier, and ϵ is an error term assumed to be i.i.d., following a Gaussian distribution. The classical approach to estimate the regression parameters is using Ordinary Least Squares (OLS). The problem here is that the way of acquisition of the data, which is obtaining features from processing images from satellite technology, may imply the presence of atypical observations that could invalidate the results. Therefore, robust methodologies need to be used.

On the other hand, the large number of variables derived from the Google Earth image, particularly those of spectral, texture and structure types, are substantially correlated (Figure 5.10).

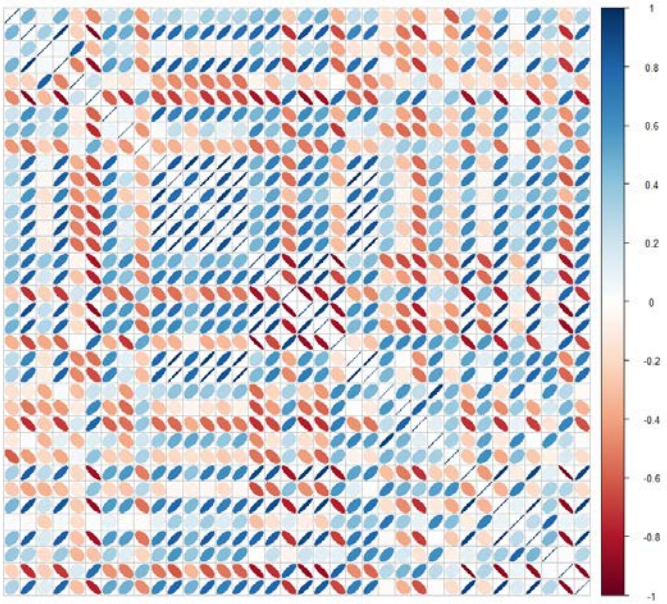

Figure 5.10: Correlation matrix for LED index data-set.

The multicollinearity issue violates another assumption for using OLS to estimate the parameters of the model. The authors propose to use a dimensionality-reduction step to preserve as much of the variation contained in the entire data-set while avoiding collinearity. They performed a principal components analysis (pca) (Jolliffe [2011], Ballabio [2015]) on all the spectral, texture and structure variables, which makes a total of 27 variables, and after the analysis, they propose to use

only the first four components because they accounted for 90% of the total variance. Other methods for data containing columns of uninformative variables in the regression problem have been proposed in the literature as well (Hoffmann et al. [2015], Li et al. [2018], Wang et al. [2019]). The four extracted components were used as regressors, together with the three land cover variables that prove most relevant: water, shadow, and vegetation. They came up with this result about the relevance by using another approach, but from machine learning area, which is the random forest (RF), since one of the main objectives of the paper was to study the potential of modern machine learning techniques: RF and gradient boost regressor (GBR), in the estimation of socioeconomic indices with remote-sensing data. Focusing on the classical OLS regression, the authors obtained that the third and fourth components were significant, as well as the proportion of an area occupied by water and vegetation.

We propose to study if the results can be improved by using robust regression methods. Let us apply the proposed SR approach and compare it with LTS, S, REWLSE and MM. The raw data, kindly provided by the authors was pre-processed the same way as they propose, by applying pca to the last 27 explanatory variables and join the first four components with the three land cover variables: water, shadow and vegetation, which makes a total of 7 explanatory variables. Table 5.13 shows the adjusted R^2 of the models estimated by each method.

Table 5.13: R^2 with (pca transformed) LED index data-set.

Method	OLS	SR	LTS	S	REWLSE	MM
R^2	0.5059	**0.6716**	0.6287	0.6031	0.5904	0.6166

Variables PC3, PC4, water and vegetation resulted significant in the model obtained by the methods. The percentage of variability explained by the robust methods shows the advantage of robust regression. The R^2 of SR is higher than that of the other approaches, although not as high as one would wish. The authors compare the results from OLS with the application of the two machine learning approaches. RF showed an $R^2 = 0.9354$ and GBR an $R^2 = 0.8320$. They were interested in finding the best possible model with the ability to capture the highest possible proportion of the variation inherent in the data. But the problem here is the drawback both machine learning methods have in terms of interpretability. Also, as the authors point out, RF and GBR suffer from the issue of overfitting.

That is why they propose a cross-validation (CV) study. It consisted on dividing the data into two groups, one to train the model, and the other one to test its predictive performance. The 5-fold CV was used and the procedure was repeated 250 times, to obtain the scores for the R^2, as in Arribas-Bel et al. [2017]. The scores for the MSE of the response are also saved. Table 5.14 shows the median cross-validated R^2 obtained by the authors for RF and GBR together with the one we obtained for method SR. The results show that SR is more robust to overfitting since the R^2 is reduced slightly, while that of RF and GBR are significantly reduced. Between the three values, SR has the highest median cross-validated R^2. On the

other hand, for method SR, the median absolute deviation from the data's median (MAD) of these scores is 0.0145 which is low, meaning that the uncertainty is under control.

Table 5.14: Median cross-validated R^2 with (pca transformed) LED index data-set.

Method	SR	RF	GBR
R^2	**0.6704**	0.54	0.50

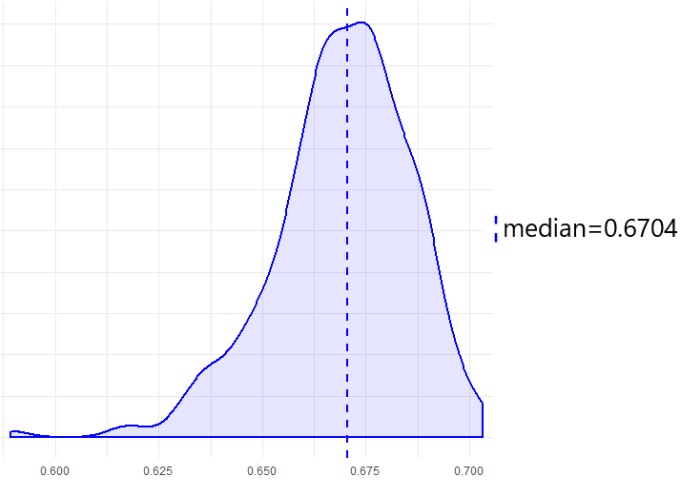

Figure 5.11: Cross-validated R^2 and median values (dashed line), with pca.

Figure 5.11 shows the distribution of the cross-validated scores for the R^2 obtained with method SR and the median value in a dashed line. Figure 5.12 shows the results for the MSE. The median of the cross-validated MSE is equal to 2.6260 and the MAD is 0.1199 which are also low values.

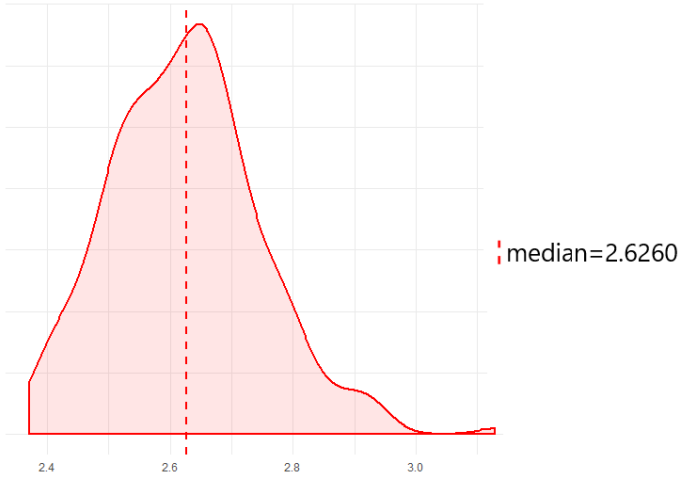

Figure 5.12: Cross-validated MSE and median values (dashed line), with pca.

Since it was mentioned before, the same pca transformation the authors proposed for the data was made for this research. Now, we propose another transformation

that improves the performance according to the results: sparse pca (spca) (Zou et al. [2006], Gajjar et al. [2017]), which has advantages in case of high correlated variables since it is a kind of variable selection transformation. The spca was made over the 27 variables of the three last groups and the first 10 components were selected since they account for 92.04% of the total variance. These 10 components and the three most relevant land cover variables: water, shadow and vegetation were used to estimate the model.

Figure 5.13 shows the distribution of the cross-validated R^2 and the median value in a dashed line obtained with SR, which is 0.8530. The MAD of these scores increases to 0.0346 but it is still a low value.

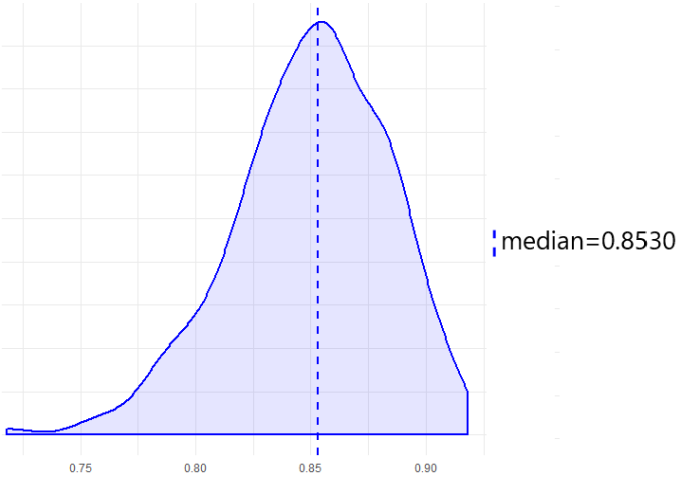

Figure 5.13: Cross-validated R^2

Figure 5.14 shows the distribution for the MSE. The median MSE reduces to 0.7244 and the MAD reduces to 0.0177.

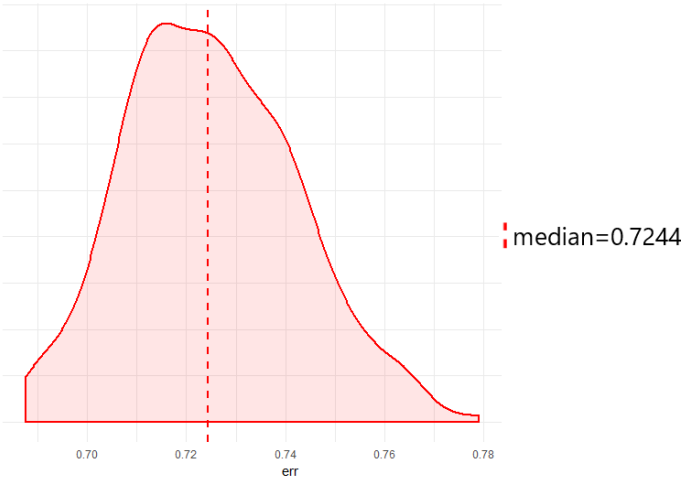

Figure 5.14: Cross-validated MSE

Table 5.15 shows that the median cross-validated R^2 is higher than that obtained with pca transformation but also higher than the obtained with both machine learning techniques, reported in Arribas-Bel et al. [2017].

Table 5.15: Median cross-validated R^2.

Method	SR spca	SR pca	RF	GBR
R^2	**0.8530**	0.6704	0.54	0.50

The uncertainty of the obtained R^2 is slightly higher with spca transformation, compared to that with the pca transformation. But Figure 5.15 shows that the distributions of the R^2 scores are quite separated, and the gain is evident because of the increase in the median value.

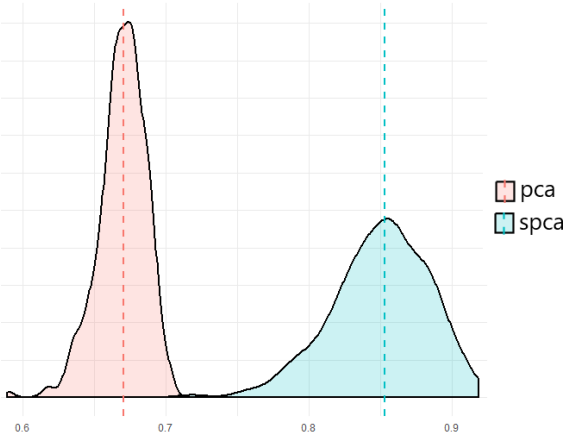

Figure 5.15: Cross-validated R^2 and median values (dashed line), for both pca and spca.

Finally, Table 5.16 contains the estimated coefficients, the p-values and the R^2 estimated by SR with spca transformation using the complete data-set, which is competitive with respect to the R^2 of RF and GBR reported in Arribas-Bel et al. [2017]. As the results point out, the same land cover variables as in the paper remained significant and with the same negative sign, meaning that larger proportions of water and vegetation are associated with smaller deprivation.

Table 5.16: Results for the model estimated by SR with spca transformation and the R^2 for RF and GBR.

	coefficient	**p-value**	**RF**	**GBR**
constant	0.27191	2.03E-05		
water	-1.42641	2.00E-16		
vegetation	-0.44513	2.00E-05		
SPC2	-0.04409	4.51E-03		
SPC3	0.13215	1.52E-06		
SPC4	0.32566	1.03E-15		
SPC5	-0.26745	2.35E-11		
SPC7	-0.13735	2.24E-03		
SPC8	0.19544	1.64E-03		
R^2	0.86820		0.9354	0.8320

5.8 Summary

The proposed SR approach is compared to classical OLS and other existing robust regression methods. The robust alternatives have some drawbacks and their performance depends on decisions that, in case of real data, increase the difficulty of robustly estimate the regression parameters. On the other hand, not all available methods have good behavior in case of large data-sets, high dimension, not all are scalable in terms of computational time, and sufficiently resistant to the presence of outliers. The proposal is to use robust estimators based on *shrinkage* in the alternative definition for the OLS estimators based on estimates of location and covariance of the joint vector of response and independent variables. The approach passes through a pair of weighting steps depending on robust Mahalanobis distances, which results in the shrinkage reweighted (SR) regression estimator. The advantages of using the shrinkage are shown in the simulation study and some conclusions can be noted. SR approach yielded competitive results compared to the alternative robust methods from the literature for the regression problem, even in high dimension, heavy-tailed distributed errors, large contamination or transformed data. Furthermore, SR is quite stable computationally since it involves contributions from all the observations instead of sub-sample iterations from the data. Finally, the results with the real data-set examples bear out with the conclusions from the simulation study. This is shown especially with the LED index data where the SR approach provides an improvement of the cross-validated R^2 and MSE with respect to classical OLS and machine learning techniques RF and GBR while maintaining the advantage of interpretability.

CHAPTER 6

Adjusted quantile

The rule for identifying outliers in multivariate data when the robust Mahalanobis distance is used consists on a threshold value. If the squared RMD of the observation in question exceeds that cut-off value, it is considered as an outlier because it is far from the center of the underlying distribution. The squared classical MD (with sample mean and sample covariance matrix estimators) has a chi-squared distribution with p degrees of freedom, where p is the dimension of the data. This does not need to be true for the robust measure when other robust estimators are used. Although, the classical threshold value $\chi^2_{p;0.975}$, used for the classical MD, is often used in the case of RMD.

The problem is that this assumption has some drawbacks. For example, if the data is clean and comes from a multivariate Normal distribution, no outliers should be detected. Filzmoser et al. [2005] proposed to use an adjusted quantile, instead of the classical choice, estimated adaptively from the data, depending on n and p, which is based on the difference between the χ^2 distribution and the empirical distribution of the squared RMD. The authors proposed the adjusted cut-off for a specific robust Mahalanobis distance, the one based on the Minimum Covariance Determinant (MCD) robust estimator proposed by Rousseeuw [1985]. This method was introduced in Chapter 2 as Adj MCD.

In this chapter, we propose to find an adjusted quantile as the adaptive threshold, following the idea from Filzmoser et al. [2005], adapted to the RMD-S introduced in Chapter 3. A simulation study is done to check the performance improvement of the new cutoff against the classical. The behavior when the underlying distribution is heavy-tailed or skewed shows the appropriateness of the method when we deviate from the common assumption of normality. The approach is illustrated using the Living Environment Deprivation (LED) example introduced in Chapter 5.

6.1 Estimating the adjusted threshold

Let us recall the estimators that defined the RMD-S. In Chapter 3, RMD-S, a robust distance based on shrinkage, is introduced. The shrinkage estimator $\hat{\boldsymbol{\mu}}_{Sh}$ was proposed as a robust estimator of central tendency.

$$\hat{\boldsymbol{\mu}}_{Sh} = (1-\eta)\hat{\boldsymbol{\mu}}_{MM} + \eta\nu_{\boldsymbol{\mu}}\mathbf{e}\,.$$

where $\hat{\boldsymbol{\mu}}_{MM}$ is the multivariate $L_1-median$ which is a robust and highly efficient estimator of location. The scaling factor $\nu_{\boldsymbol{\mu}}$ and the intensity η are obtained minimizing the expected quadratic loss. The solution can be found in Proposition 2 from Chapter 3. On the other hand, an adjusted special comedian matrix $\hat{S}_{Sh}$, based on the classical definition of comedian from Falk [1997], was proposed:

$$\hat{S}_{Sh} = 2.198 \cdot (median((\mathbf{x}_j - (\hat{\boldsymbol{\mu}}_{Sh})_j)(\mathbf{x}_t - (\hat{\boldsymbol{\mu}}_{Sh})_t))\,,$$

for $j,t = 1,...,p$, and with it, a shrinkage estimator for the covariance matrix can be obtained:

$$\hat{\Sigma}_{Sh} = (1-\eta)\hat{S}_{Sh} + \eta\nu_{\Sigma}I\,.$$

The idea came from the fact that the comedian matrix is a robust alternative for the covariance matrix, but in general, it is not positive (semi-)definite, and with the shrinkage approach applied to the comedian, a robust and well-conditioned estimate is obtained. The optimal expression for the parameters η and ν_{Σ} is described in Proposition 3 from Chapter 3.

These robust estimators of location $\hat{\boldsymbol{\mu}}_{Sh}$ and covariance matrix $\hat{\Sigma}_{Sh}$ based on shrinkage are used to define an RMD, called RMD-S, and it was proved through a simulation study that the proposal has advantages compared to other existing methods for robust multivariate outlier detection. But since the threshold used to declare observations as outliers was the classical choice, we propose in this section to find an adaptive quantile following the idea from Filzmoser et al. [2005] adapted to the RMD-S.

Denote by $G_n(u)$ the empirical distribution function of the squared robust Mahalanobis distances $RMD\text{-}S_i^2$, and by $G(u)$ the distribution function of χ^2_p, the distribution used in theory. It is known that for multivariate normally distributed samples, G_n converges to G. Then, G_n and G can be compared in the tails to detect outliers. The tails will be defined by $\delta = \chi^2_{p;1-\alpha}$ and α can be for example 0.02. Then, the difference between the empirical and the chi-squared distribution functions, in the tail, is:

$$p_n(\delta) = sup_{u\geq\delta}(G(u) - G_n(u))^+\,, \tag{6.1}$$

where $+$ means positive differences. The measure $p_n(\delta)$ is not used as a measure of outliers, because the cut-off should be infinity in case of clean multivariate normally distributed data. In this case, no observation should be declared as atypical, and observations with a large RMD should be seen as extremes of the distribution.

To distinguish between the two cases, a critical value p_{crit} is introduced. If the difference between the two distributions in the tails is lower than the critical value, the measure of outliers should be considered as zero. If the difference is higher than the critical value, it can be considered as a measure of outliers:

$$\alpha_n(\delta) = \begin{cases} 0, & \text{if} \quad p_n(\delta) \leq p_{crit} \\ p_n(\delta), & \text{if} \quad p_n(\delta) > p_{crit} \end{cases} .$$

With this condition, the threshold value for detecting outliers with the robust Mahalanobis distance is:

$$c_n(\delta) = G_n^{-1}(1 - \alpha_n(\delta)) .$$

The critical value p_{crit} should be adjusted to data, depending on the dimension and the sample size. It can be obtained by simulations, which are analogous as the study of Filzmoser et al. [2005] for Adj MCD. For different sample sizes n and dimensions p, data from a multivariate Normal distribution are simulated. Next step is to apply Equation 6.1 for computing the value $p_n(\delta)$ for a fixed δ. The fixed value considered is the same as in Filzmoser et al. [2005], $\delta = \chi^2_{p;0.98}$. For every combination for the value of n and p, this is repeated 100 times.

The results are how the differences between the chi-squared and the empirical distributions, $G(u) - G_n(u)$, should be if the data are sampled from multivariate Normal distributions. From these results, the 95% percentile of the 100 simulated values is selected, and these percentiles are shown for $p = 2, 4, 6, 8, 10$ by different symbols in Figure 6.1. The x-axis is transformed by the inverse of $\sqrt{n}$.

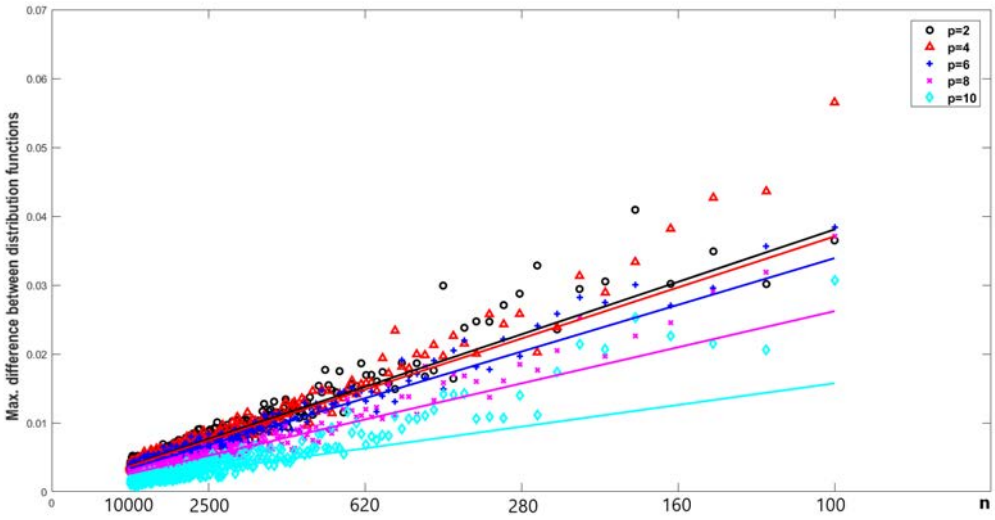

Figure 6.1: Simulated $p_n(\delta)$ for multivariate Normal distributions with different sample sizes ($x-$axis) and dimensions $p \leq 10$.

In Figure 6.1 it can be seen that the points lie on a line, at least for higher sample sizes. The lines are estimated by LTS regression because the less precise simulation results for smaller sample sizes should have less influence. Also, the lines should have zero intercept because for n tending to infinity the difference between

the empirical and the chi-squared distribution is zero. The slopes of the different lines estimated by LTS from Figure 6.1 are shown in Figure 6.2.

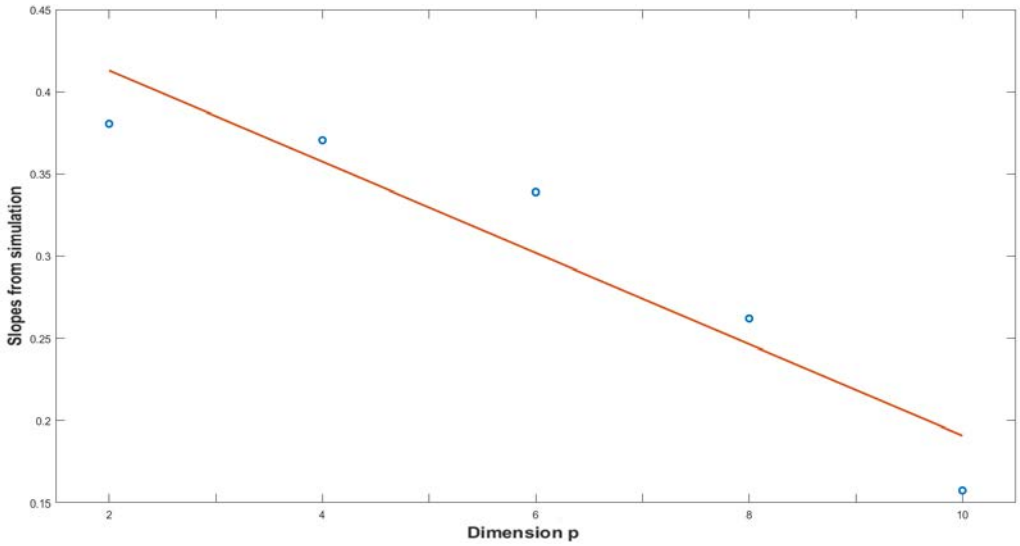

Figure 6.2: Slopes of lines from Figure 6.1 plotted against dimension p.

The resulting points can again be approximated by a straight line, which allows the definition of the critical value as a function of n and p:

$$p_{crit}(\delta, n, p) = \frac{0.4686 - 0.0278p}{\sqrt{n}} \quad \text{for} \quad p \leq 10\,.$$

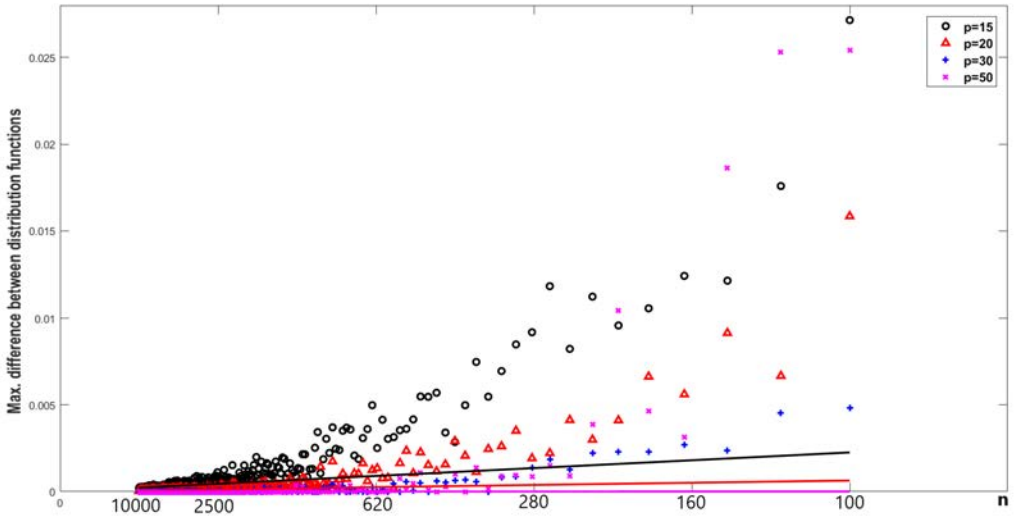

Figure 6.3: Simulated $p_n(\delta)$ for multivariate Normal distributions with different sample sizes ($x-$axis) and dimensions $p > 10$.

For larger dimension ($p > 10$) the same procedure can be applied. The 95% percentiles of 100 simulated values for different sample sizes and dimensions are shown in Figure 6.3. The linear dependency becomes worse for high dimension and low

sample size.

The estimated slopes from the LTS regression lines are shown in Figure 6.4.

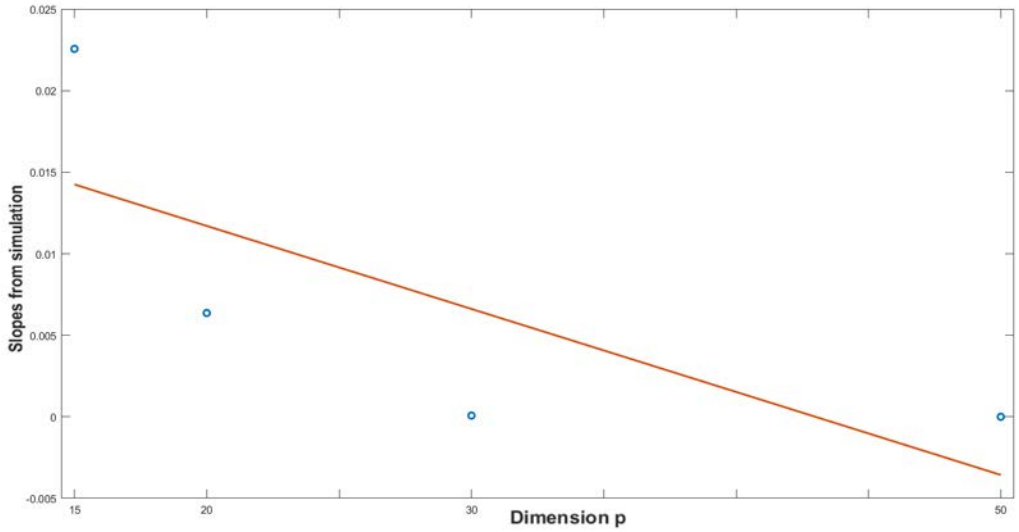

Figure 6.4: Slopes of lines from Figure 6.3 plotted against dimension p.

The resulting formula for the definition of the critical value as a function of n and p, for higher dimension is:

$$p_{crit}(\delta, n, p) = \frac{0.0219 - 0.0005p}{\sqrt{n}} \quad \text{for} \quad p > 10\,.$$

6.2 Simulations

In this section, a simulation study is performed in order to investigate the behavior of the RMD-S with the new adaptive threshold with respect to the true positive rate (TPR) and false positive rate (FPR), in different scenarios. From now on let us refer the RMD-S with the adjusted quantile as RMD-SAQ.

6.2.1 Normal distribution

Consider a $p-$dimensional random variable X following a contaminated multivariate Normal distribution given as a mixture of Normals of the form $(1-\alpha)N(\mathbf{0}, I) + \alpha N(\delta\mathbf{e}, \lambda I)$, where $\mathbf{e}$ denotes the $p-$dimensional vector of ones. The dimensions $p = 5, 10, 30, 50$, and the sample sizes $n = 100, 100, 500, 10000$, respectively. The contamination levels $\alpha = 0, 0.1, 0.2, 0.3$, the distance of the outliers $\delta = 5$ and 10 and the concentration of the contamination $\lambda = 0.1$ and 1. For each set of values, 100 random repetitions are generated.

The method is considered improved if the FPR decreases and the TPR increases or remains the same. Table 6.1 shows the results for the FPR when there is no contamination. As it can be seen, for all dimensions the FPR decreases when the

Table 6.1: FPR for Normal data with $\alpha = 0\%$.

		RMD-S	RMD-SAQ
$p = 5$	$n = 100$	0.003162	0.000092
$p = 10$	$n = 100$	0.001601	0.000014
$p = 30$	$n = 500$	0.000012	0.000001
$p = 50$	$n = 1000$	0.000011	0.000001

adjusted quantile is used. This traduces in an efficiency improvement with method RMD-SAQ.

Table 6.2 shows the $F-$scores in case of contamination.

Table 6.2: $F-$scores in case of Normal data.

$p = 5$	$\delta = 5$, $\lambda = 0.1$	RMD-S	RMD-SAQ	$\delta = 10$, $\lambda = 0.1$	RMD-S	RMD-SAQ
	0.1	0.9672	0.9937	0.1	0.9633	0.9832
	0.2	0.9909	0.9929	0.2	0.9986	1
	0.3	0.8881	0.8795	0.3	1	1
	$\delta = 5$, $\lambda = 1$	RMD-S	RMD-SAQ	$\delta = 10$, $\lambda = 1$	RMD-S	RMD-SAQ
	0.1	0.9664	0.9903	0.1	0.9658	0.9926
	0.2	0.9986	0.9998	0.2	0.9968	0.9997
	0.3	0.9443	0.9305	0.3	1	1
$p = 10$	$\delta = 5$, $\lambda = 0.1$	RMD-S	RMD-SAQ	$\delta = 10$, $\lambda = 0.1$	RMD-S	RMD-SAQ
	0.1	0.9840	0.9972	0.1	0.9870	0.9966
	0.2	1	1	0.2	1	1
	0.3	0.9595	0.9538	0.3	1	1
	$\delta = 5$, $\lambda = 1$	RMD-S	RMD-SAQ	$\delta = 10$, $\lambda = 1$	RMD-S	RMD-SAQ
	0.1	0.9882	0.9887	0.1	0.9878	0.9907
	0.2	0.9994	0.9996	0.2	0.9993	1
	0.3	0.9744	0.9680	0.3	1	1
$p = 30$	$\delta = 5$, $\lambda = 0.1$	RMD-S	RMD-SAQ	$\delta = 10$, $\lambda = 0.1$	RMD-S	RMD-SAQ
	0.1	0.9996	1	0.1	0.9993	1
	0.2	1	1	0.2	1	1
	0.3	0.9880	0.9889	0.3	1	1
	$\delta = 5$, $\lambda = 1$	RMD-S	RMD-SAQ	$\delta = 10$, $\lambda = 1$	RMD-S	RMD-SAQ
	0.1	0.9996	1	0.1	0.9996	1
	0.2	1	1	0.2	1	1
	0.3	0.9998	0.9999	0.3	1	1
$p = 50$	$\delta = 5$, $\lambda = 0.1$	RMD-S	RMD-SAQ	$\delta = 10$, $\lambda = 0.1$	RMD-S	RMD-SAQ
	0.1	0.9997	1	0.1	0.9995	1
	0.2	1	1	0.2	1	1
	0.3	0.9898	0.9925	0.3	1	1
	$\delta = 5$, $\lambda = 1$	RMD-S	RMD-SAQ	$\delta = 10$, $\lambda = 1$	RMD-S	RMD-SAQ
	0.1	0.9998	1	0.1	0.9997	1
	0.2	1	1	0.2	1	1
	0.3	0.9999	1	0.3	1	1

The motivation for the $F-$score measure is that in case of contamination, the FPR value always reduces when the adjusted quantile is considered and in the vast majority of cases the TPR increases or remains the same, except in a few cases where the TPR slightly decreases. The $F-$score is considered to measure the trade-off between the TPR and the FPR. Its expression is $F-\text{score}= 2PR/(P+R)$, where P is called precision and R is known as the recall. It was described previously, in Chapter 3, but let us remember that the precision P is the number of correctly detected outliers divided by the total number of detected outliers, and the recall R is the number of correctly detected outliers divided by the real total number of outliers. Thus, this measure provides a balance between the two desired outcomes: a high rate of correctly identified outliers and a low rate of observations mislabel as outliers.

Table 6.2 shows that the adjusted quantile improves the performance in the majority of cases, except when the dimension is low or moderate ($p = 5$ and 10), the atypical observations are near the center of the data ($\delta = 5$) and there is a high level of contamination ($\alpha = 30\%$), in which cases the $F-$score slightly decreases. When the outliers are more separated from the background data ($\delta = 10$), RMD-SAQ is more or equally accurate than RMD-S for all dimensions and all contamination levels. In the high dimension case ($p = 30$ and 50) the use of the adjusted quantile exhibits clear advantages over the classical one, even if the outliers are near the center of the data and even with high contamination.

6.2.2 t_3-distribution

Let us study the performance of the methods when the distribution deviates from normality, considering a $p-$dimensional random variable X following a contaminated multivariate t-distribution with 3 degrees of freedom of the form $(1-\alpha)T_3(\mathbf{0}, I) + \alpha T_3(\delta\mathbf{e}, \lambda I)$. The first parameter of the notation of $T_3(\cdot,\cdot)$ refers to the mean and the second one to the covariance matrix. The parameters for the contamination are the same considered above and the same measures TPR and FPR are studied. For the case when the contamination level is 0%, Table 6.3 shows the FPR values. The fact that the data are not Normal, but heavy-tailed, influences the FPR values. Nevertheless, for every dimension, method RMD-SAQ decreases the FPR, improving the performance.

Table 6.3: FPR for t_3-distributed data with $\alpha = 0\%$.

		RMD-S	RMD-SAQ
$p = 5$	$n = 100$	0.183703	0.166825
$p = 10$	$n = 100$	0.127602	0.108975
$p = 30$	$n = 500$	0.114004	0.072873
$p = 50$	$n = 1000$	0.071239	0.056921

As in the previous case when the data came from a multivariate Normal distribution, for the remaining contamination scenarios, we show in Table 6.4 the $F-$score as the trade-off between the TPR and the FPR. But in general, the $F-$score value gets influenced mainly due to the high levels of the FPR for contamination level

$\alpha = 10\%$, in which case for RMD-S is around 0.12. In any case, the distance with the adjusted quantile improves the FPR because it always reduces its value, although for $\alpha = 10\%$ contamination level, the FPR remains around 0.10. This fact reflect in the $F-$scores results, in the rows for $\alpha = 10\%$ in Table 6.4.

Table 6.4: $F-$scores in case of t_3-distributed data.

$p = 5$	$\delta = 5$, $\lambda = 0.1$	RMD-S	RMD-SAQ	$\delta = 10$, $\lambda = 0.1$	RMD-S	RMD-SAQ
	0.1	0.7378	0.7731	0.1	0.7328	0.7719
	0.2	0.8692	0.8916	0.2	0.8757	0.9003
	0.3	0.8512	0.8536	0.3	0.9534	0.9623
	$\delta = 5$, $\lambda = 1$	RMD-S	RMD-SAQ	$\delta = 10$, $\lambda = 1$	RMD-S	RMD-SAQ
	0.1	0.7329	0.7686	0.1	0.7426	0.7785
	0.2	0.8600	0.8824	0.2	0.8776	0.9003
	0.3	0.8020	0.7976	0.3	0.9644	0.9743
$p = 10$	$\delta = 5$, $\lambda = 0.1$	RMD-S	RMD-SAQ	$\delta = 10$, $\lambda = 0.1$	RMD-S	RMD-SAQ
	0.1	0.7287	0.7628	0.1	0.7728	0.8067
	0.2	0.8528	0.8757	0.2	0.8703	0.8968
	0.3	0.8286	0.8336	0.3	0.9587	0.9699
	$\delta = 5$, $\lambda = 1$	RMD-S	RMD-SAQ	$\delta = 10$, $\lambda = 1$	RMD-S	RMD-SAQ
	0.1	0.7206	0.7549	0.1	0.7062	0.7407
	0.2	0.8655	0.8894	0.2	0.8576	0.8812
	0.3	0.9064	0.9069	0.3	0.9575	0.9684
$p = 30$	$\delta = 5$, $\lambda = 0.1$	RMD-S	RMD-SAQ	$\delta = 10$, $\lambda = 0.1$	RMD-S	RMD-SAQ
	0.1	0.7907	0.7933	0.1	0.7815	0.7844
	0.2	0.8635	0.8962	0.2	0.8806	0.9125
	0.3	0.8802	0.8889	0.3	0.9684	0.9789
	$\delta = 5$, $\lambda = 1$	RMD-S	RMD-SAQ	$\delta = 10$, $\lambda = 1$	RMD-S	RMD-SAQ
	0.1	0.7836	0.7849	0.1	0.7852	0.7974
	0.2	0.8545	0.8887	0.2	0.8695	0.9032
	0.3	0.9392	0.9396	0.3	0.9702	0.9811
$p = 50$	$\delta = 5$, $\lambda = 0.1$	RMD-S	RMD-SAQ	$\delta = 10$, $\lambda = 0.1$	RMD-S	RMD-SAQ
	0.1	0.8104	0.8237	0.1	0.8099	0.8121
	0.2	0.8836	0.9169	0.2	0.8909	0.9311
	0.3	0.8968	0.9089	0.3	0.9713	0.9849
	$\delta = 5$, $\lambda = 1$	RMD-S	RMD-SAQ	$\delta = 10$, $\lambda = 1$	RMD-S	RMD-SAQ
	0.1	0.8091	0.8125	0.1	0.8151	0.8272
	0.2	0.8814	0.9078	0.2	0.8991	0.9362
	0.3	0.9497	0.9599	0.3	0.9790	0.9912

On the other hand, the use of the adjusted quantile, instead of the classical one, improves the performance according to the $F-$score in the majority of cases. One case is the exception: when the dimension is low ($p = 5$), the atypical observations are near the center of the data ($\delta = 5$), the concentration is one ($\lambda = 1$) and the level of contamination is high ($\alpha = 30\%$), in which the $F-$score slightly decreases from 0.8020 to 0.7976. In all remaining cases, the improvement is reflected.

6.2.3 Exponential distribution

We considered also a $p-$dimensional random variable X following a contaminated multivariate exponential distribution given as a mixture $(1-\alpha)Exp(\mathbf{0})+\alpha Exp(\delta\mathbf{e})$. The parameter of the notation $Exp(\cdot)$ refers to the mean. This case is analogous to the previous ones, with the difference that only the schemes associated with the distance of the outliers are considered. Table 6.5 shows the FPR values when there is no contamination.

Table 6.5: FPR for exponential distributed data with $\alpha = 0\%$.

		RMD-S	RMD-SAQ
$p = 5$	$n = 100$	0.112803	0.102572
$p = 10$	$n = 100$	0.136090	0.125437
$p = 30$	$n = 500$	0.184643	0.156330
$p = 50$	$n = 500$	0.117781	0.108962

With the multivariate exponential distribution, the TPR values of RMD-S are high in the majority of cases, and it gets improved or remains the same in the vast majority of cases, by means of the adjusted quantile. On the other hand, the FPR values of RMD-S are low and decrease even more with RMD-SAQ. Table 6.6 shows the $F-$score values in case of contamination.

Table 6.6: $F-$scores in case of exponential distributed data.

$p = 5$	$\delta = 5$	RMD-S	RMD-SAQ	$\delta = 10$	RMD-S	RMD-SAQ
	0.1	0.8555	0.8583	0.1	0.8549	0.8577
	0.2	0.8785	0.8809	0.2	0.8506	0.8734
	0.3	0.8899	0.8985	0.3	0.9560	0.9667
$p = 10$	$\delta = 5$	RMD-S	RMD-SAQ	$\delta = 10$	RMD-S	RMD-SAQ
	0.1	0.8503	0.8528	0.1	0.8505	0.8531
	0.2	0.8937	0.9186	0.2	0.9087	0.9350
	0.3	0.9120	0.9293	0.3	0.9426	0.9573
$p = 30$	$\delta = 5$	RMD-S	RMD-SAQ	$\delta = 10$	RMD-S	RMD-SAQ
	0.1	0.8448	0.8474	0.1	0.8477	0.8507
	0.2	0.8948	0.9297	0.2	0.9593	0.9905
	0.3	0.9460	0.9694	0.3	0.9798	0.9919
$p = 50$	$\delta = 5$	RMD-S	RMD-SAQ	$\delta = 10$	RMD-S	RMD-SAQ
	0.1	0.8601	0.8790	0.1	0.8707	0.8946
	0.2	0.9149	0.9417	0.2	0.9643	0.9980
	0.3	0.9667	0.9789	0.3	0.9819	0.9988

As it can be seen, the robust distance with the adjusted quantile has advantages with asymmetric distributions as well, since the performance according to the $F-$score in all the simulation scenarios is improved.

6.3 Real data-set example

In Chapter 5 the Living Environment Deprivation (LED) index of Liverpool (UK) was introduced as an example to study the robust regression methods and the proposed Shrinkage Reweighted (SR) regression estimator. The linear regression model is the following:

$$LED = \alpha + \boldsymbol{\beta} LC + \boldsymbol{\gamma} SP + \boldsymbol{\delta} TX + \boldsymbol{\zeta} ST + \epsilon\,.$$

Arribas-Bel et al. [2017] propose to use a principal component analysis (pca) on the data to eliminate the multicollinearity problem. In Chapter 5, the proposed robust regression approach based on shrinkage makes use of the robust Mahalanobis distance from Chapter 3, the RMD-S with the classical quantiles from the chi-squared distribution. In Chapter 5, the LED index data was studied and good results are obtained with SR method. Another transformation that improves the performance is proposed there: sparse pca (spca), which has advantages in case of highly correlated variables since it is a kind of variable selection transformation. The spca was made over the 27 variables of the three last groups and the first 10 components were selected since they account for 92.04% of the total variance. These 10 components and the three most relevant land cover variables: water, shadow and vegetation were used to estimate the model.

The authors from Arribas-Bel et al. [2017] compared the results from OLS with the application of the two machine learning approaches Random Forest (RF) and Gradient Boost Regressor (GBR). The problem is both machine learning methods lacks of interpretability. Table 6.7 shows the R^2 obtained for method SR with pca and spca transformations, together with the R^2 obtained by Arribas-Bel et al. [2017] for RF and GBR with pca transformed data.

Table 6.7: R^2 measures.

Method	SR spca	SR pca	RF	GBR
R^2	**0.8682**	0.6716	0.9354	0.8320

To avoid overfitting, a cross-validation (CV) study is proposed both in Arribas-Bel et al. [2017] and in Chapter 5. The 5-fold CV was used and the procedure was repeated 250 times, to obtain the scores for the R^2, as well as the MSE of the response. In Chapter 5, the CV study is performed with method SR and the transformed data with both pca and spca. SR showed to be more robust to overfitting than RF and GBR. Between the four methods, SR with spca transformation had the highest median cross-validated R^2, and the MAD of those scores was low, which means that the uncertainty is under control.

In this section, we propose to study if the results can be improved by using RMD-SAQ, i.e., the adjusted quantile instead of the classical quantile q_1 from Equation 5.9, in the robust Mahalanobis distance based on shrinkage RMD-S, used in the robust SR regression approach proposed in Chapter 5. With spca transformation, we estimated the model using SR approach with RMD-SAQ, which we call from now

on method SR-AQ (Shrinkage reweighted regression estimator with the adjusted quantile). We also performed a cross-validation study, with the same characteristics as in Arribas-Bel et al. [2017] and Chapter 5, to investigate if there can be an improvement with respect to the R^2 and the MSE by using SR-AQ. Table 6.8 shows the resulting cross-validated R^2 of SR-AQ and SR with classical quantile, both with spca transformed data, and also RF and GBR, both with pca transformed data. SR-AQ provides an improvement because the median cross-validated R^2 increases.

Table 6.8: Median cross-validated R^2.

Method	SR-AQ	SR	RF	GBR
R^2	**0.9135**	0.8530	0.54	0.50

Figure 6.5 shows the distribution of the cross-validated R^2 and the median value in a dashed line obtained with SR. The MAD of those scores also reduces from 0.0346 (SR) to 0.0142 (SR-AQ).

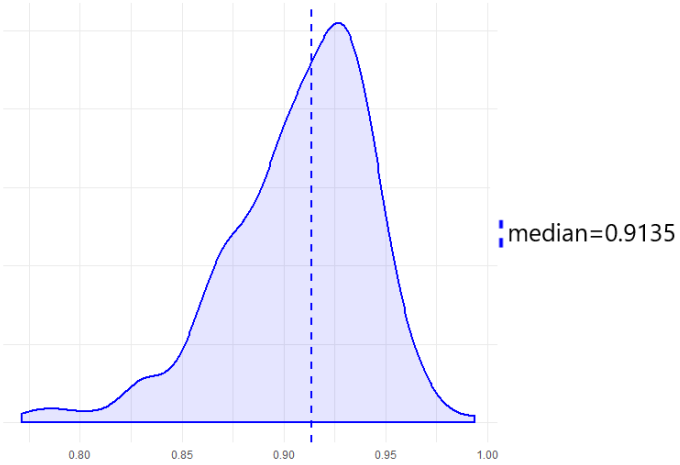

Figure 6.5: Cross-validated R^2.

Figure 6.6 shows the distribution for the cross-validated MSE, with method SR-AQ considering spca transformation. The median cross-validated MSE for SR obtained in Chapter 5 was of 0.7244 (with MAD 0.0177) and the one we obtained for SR-AQ results to be 0.4138 (with MAD 0.0150). Then, the error reduces with controlled uncertainty. In summary, the cross-validation showed that RF and GBR are sensitive to overfitting, reducing largely their R^2, while SR (with classical threshold) and SR-AQ (with adjusted quantile) are more robust and showed higher median cross-validated R^2, with an improvement on behave of SR-AQ, with low values of MSE.

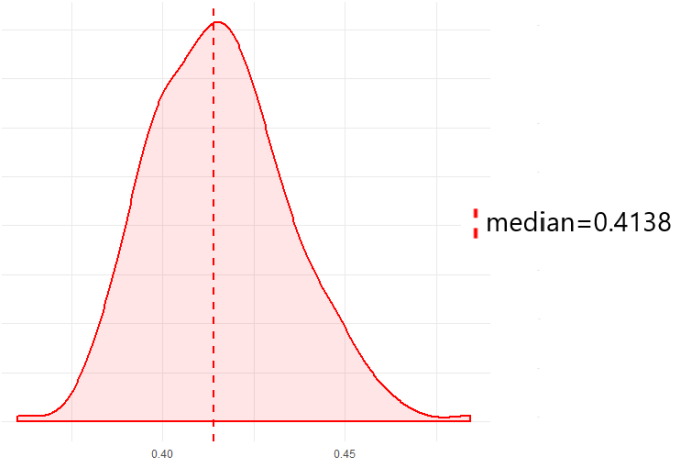

Figure 6.6: Cross-validated MSE.

Table 6.9 contains the results of the model estimated by SR-AQ. The p-values show that the same land cover variables remained significant and with the same negative sign, meaning that larger proportions of water and vegetation are associated with smaller deprivation. The R^2 estimated with SR-AQ is higher than SR, so the method provides an advantage compared to the classical quantile as threshold. It is also competitive with respect to the R^2 of RF and GBR reported in Arribas-Bel et al. [2017].

Table 6.9: Results for the model estimated by SR with spca transformation and the R^2 for RF and GBR.

	SR-AQ		**SR**		**RF**	**GBR**
	coefficient	**p-value**	**coefficient**	**p-value**		
constant	0.35191	4.61E-04	0.27191	2.03E-05		
water	-1.12131	3.12E-10	-1.42641	2.00E-16		
vegetation	-0.64527	1.45E-08	-0.44513	2.00E-05		
SPC2	-0.01406	4.21E-05	-0.04409	4.51E-03		
SPC3	0.19925	2.62E-07	0.13215	1.52E-06		
SPC4	0.56633	4.09E-09	0.32566	1.03E-15		
SPC5	-0.41745	1.39E-10	-0.26745	2.35E-11		
SPC7	-0.31453	2.69E-12	-0.13735	2.24E-03		
SPC8	0.15433	3.90E-08	0.19544	1.64E-03		
R^2	0.9094		0.8682		0.9354	0.8320

6.4 Summary

In this chapter, an adjusted quantile is proposed as the threshold for method RMD-S presented in Chapter 3 in which it was used the classical chi-squared quantile as the cut-off value for detecting outliers in multivariate data. RMD-SAQ was studied by means of simulations, and the results showed that for all dimensions the FPR decreases with the adjusted cut-off. This fact evidence the efficiency improvement,

even when the underlying distribution is heavy-tailed or skewed, which evidence the advantages of the adjusted quantile even when we deviate from the common assumption of normality. On the other hand, the overall improvement in performance is reflected by the high $F-$score measures in the rest of simulation scenarios, when the adaptive threshold is considered.

Finally, a real data-set example is studied to investigate if the estimated model can be improved with the introduction of the adjusted quantile. In Arribas-Bel et al. [2017] the authors explain the Living Environment Deprivation (LED) index of areas of Liverpool (UK), through remote sensing data. They studied the linear regression model with the classical OLS approach and two machine learning techniques: Random Forest (RF) and Gradient Boost Regressor (GBR). In Chapter 5, this data is studied with the proposed robust regression approach, the Shrinkage Reweighted (SR) regression estimator, which is based on RMD-S, the robust Mahalanobis distance based on shrinkage, for which we propose the adjusted quantile. Then, it was logical to analyze if the model could be improved by the introduction of the adaptive threshold method RMD-SAQ in the SR robust regression approach, which we called method SR-AQ. The results with the real example were that the median cross-validated R^2 of the resulting model based on spca transformed data, estimated with SR-AQ, increases a 7% with respect to SR. On the other hand, the median cross-validated MSE with SR-AQ decreases a 43% with respect to SR. Then, the use of the adjusted quantile provides advantages in robust outlier detection and robust regression.

CHAPTER 7

Conclusions and Future research

This thesis addresses the problem of the presence of outliers in multivariate data and regression. In general, it is known that the classical sample estimators, i.e., the sample mean and the sample covariance matrix are sensitive to the presence of outliers. The same happens in regression with Ordinary Least Squares estimator because a single atypical observation can highly distort the results. A thorough review of the existing robust approaches in the literature is presented in two of the chapters of the thesis to introduce the background of the problem. There are several methods for robust estimation of location, covariance and regression parameters to overcome the influence of outliers. Although, no consensus establishes which technique is recommended in practical situations. Furthermore, not all available methods work well for high dimension, high sample size, not all are sufficiently resistant to the presence of anomalous values, and are computationally feasible at the same time. The decision of which method should be used in practice is a difficult task because of the trade-off between efficiency and breakdown value. Finally, the need for an adaptive threshold to detect outliers with a robust Mahalanobis distance is essential, since the classical choice of the chi-squared quantile is not completely accurate.

The first contribution of the thesis is a robust approach, RMD-S, to detect outliers in multivariate data, based on the robust Mahalanobis distance. The robust estimators to define the proposed distance are based on the notion of shrinkage. The shrinkage has the advantages of reducing the estimation error, obtaining a trade-off between bias and variance. In the case of covariance matrices, the shrinkage has the additional advantage that it provides a positive definite and well-conditioned estimate, which is of crucial importance because the inverse of the covariance matrix is needed in the definition of the Mahalanobis distance. The performance of RMD-S and the other alternatives from the literature is shown through a simulation study measuring the True Positive Rates (TPR) and the False Positive Rates (FPR). The alternative methods decrease their TPR and increase their FPR in most cases when the data is contaminated, especially in high dimension. While RMD-S has the ability to discover outliers with high TPR and low FPR in the vast majority of cases

in the simulations, with Gaussian data and with skewed or heavy-tailed distributions. On the other hand, some properties of RMD-S are investigated. Correlated and transformed data show that RMD-S is approximately affine equivariant. Highly contaminated data show that the approach has a high breakdown value even in high dimension. The method has a competitive computational time. A real data-set about outlier detection with benign breast cancer data shows that the proposed method works well in practice and require reasonable computational times, even for large problems.

The second contribution of the thesis is a robust regression approach using robust estimators based on shrinkage in the alternative definition for the OLS estimator. The method is based on weighting the observations using RMD-S, which gives place to a robust shrinkage reweighted (SR) regression estimator. The performance of the proposed SR approach is compared by simulations to the classical OLS and other existing robust regression methods. SR approach yielded competitive results compared to the alternative robust methods, even in high dimension, heavy-tailed distributed errors, large contamination or transformed data. Computationally, SR results to be stable with low computational times. A couple of typical real data-set examples are introduced, together with a high dimensional example about the robust regression to explain the LED index based on remote sensing data. The SR approach provides an improvement in the real example, especially with the LED index data. The cross-validated R^2 and MSE with respect to classical OLS and machine learning techniques RF and GBR, are improved with method SR, that also has the advantage of interpretability over the machine learning approaches.

The third contribution of the thesis is the adjusted quantile as the adaptive threshold for RMD-S, giving place to a more accurate cut-off for outlier detection. The distance with the adjusted threshold is denoted as RMD-SAQ. The simulation study showed that for all dimensions, the FPR decreases with method RMD-SAQ with respect to RMS-S with classical quantile. Thus, there is efficiency improvement, even when we deviate from the common assumption of normality with heavy-tailed or skewed distributions. The high $F-$score measures in the simulations with contamination evidence the advantages of using the adjusted quantile instead of the classical threshold. On the other hand, since the method SR for robust regression weights observations based on RMD-S, the adjusted threshold can be used to improve the performance of method SR, which gives the alternative: SR-AQ. The real data-set example of the LED index data is studied to investigate if the estimated model can be improved with SR-AQ. The median cross-validated R^2 of the resulting model increases and the median cross-validated MSE with method SR-AQ decreases, with respect to the results with SR. Therefore, the adjusted quantile has advantages in both robust outlier detection and robust regression.

7.1 Future work

In relation to Chapter 3, it remains to be examined some theoretical properties of the proposed robust estimators of location and covariance matrix that defines the

robust Mahalanobis distance RMD-S, such as asymptotic distribution, consistency, and others. In the definition of shrinkage, other target estimators, different than the scaled identity matrix, could also be studied to see how it affects the results. Furthermore, since the shrinkage can be defined for an scenario where the dimension is much higher than the sample size $p >> n$, this situation could be investigated because of its high importance in practice, for example in Genetics. An additional possibility to further investigate is to consider multivariate depth measures instead of the multivariate median, as it is known that depth is a robust measure for location (Tukey [1975], Liu et al. [1990], Serfling [2002], Chen et al. [2008], Agostinelli and Romanazzi [2011], Paindaveine and Van Bever [2013]), and see if the outlier detection approach can be improved with these kind of estimators.

In relation to Chapter 5, it could also be an interesting matter to study whether the use of the different definitions of depth in the literature could improve the performance of the approach. In the proposed robust regression method SR, other weights different than hard rejection, could also be investigated, for example, weights depending on a depth measure. Additionally, other quantities could also be considered as the cut-off measures, different than the classical choices proposed in Chapter 5 and the adjusted quantile proposed in Chapter 6. Finally, the multivariate regression problem, in which the response has more than one variable, could also be investigated with the proposed method SR or the other possible variations proposed for future work. In this case, the robust alternatives studied in Chapter 4 cannot be used, but there are other approaches proposed in the literature for this problem, such as multivariate Least Trimmed Squares regression (Agulló et al. [2008]) or the the LR-weighted MCD regression (Rousseeuw et al. [2004]).

On the other hand, the first goal of the early work of this thesis was to proposed a robust outlier detection method for functional Magnetic Resonance Imaging (fMRI) data. Through the years of research it was decided to develop the approach for general multivariate data and general regression. But we are currently working on the application to fMRI problem. The aim of fMRI data analysis is to determine which regions of the brain are either activated or inactivated with respect to an experimental design. In order to do this, one must consider a partition of the whole brain, consisting of a set of very small cuboid elements called voxels, each of one representing a million of brain cells. After the patient is subjected to some type of stimulus (auditory, visual, mechanical, etc), the result of the entire procedure is an image of the brain, as it can be seen in Figure 7.1.

Those colored spots in the image above, designate the activated zones in the brain that were related to the experiment. Note that they are actually clusters of voxels, perhaps hundreds of them. And the rest of the area, designated by a gray color, represents the non activated zones, i.e. the areas that did not have relation at all with the experiment. The exact size of a voxel may vary, although they typically represent a volume of $27mm^3$ (a cube with $3mm$ length sides), meaning that the partition of the brain will consist in a set of 20.000 up to 100.000 voxels. This leads to the statistical problem of how to manage this kind of data (Lazar [2008], Budde [2012]). Also, small movements of the head, and even heartbeat and breathing can

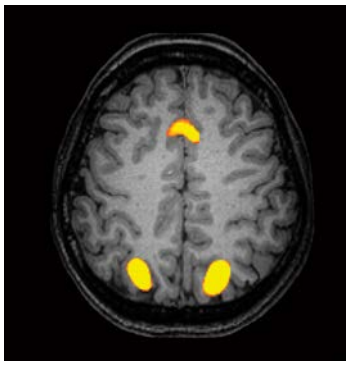

Figure 7.1: fMRI scan.

induce pulsatile motion in the brain, which creates physiological noise artifacts in the signals. Various kinds of acquisition artifacts may be present in fMRI data, that can influence an individual subject's regression parameter estimates dramatically. Complexity and massive amount of this kind of data, and the presence of different types of noises and atypical observations, makes the fMRI data analysis a challenging one, that demands robust and computationally efficient statistical analysis methods. For further research, it remains to be examined the parallelization of the algorithms, to gain in computational efficiency.

The Mahalanobis distance is used in many other Statistical analysis, a fact that provides additional lines of future research to investigate the performance of the proposed robust Mahalanobis distance for these tasks. For example, Hotelling's t-squared statistic distribution (Hotelling et al. [1931]), is used in multivariate statistics as a generalization of the Student's t-distribution, which is used in the univariate case when the scatter in unknown. Hotelling's t-squared statistic is defined as:

$$t^2 = (\overline{\mathbf{x}} - \boldsymbol{\mu})\Sigma_{\overline{x}}^{-1}(\overline{\mathbf{x}} - \boldsymbol{\mu})^t$$

It can also be defined for two samples in hypothesis tests for the differences between the multivariate means of different populations. More robust and powerful tests can be found in the literature, such as the interpoint distance based tests which can be applied also when the dimension p is comparable with, or even larger than, the sample size n (Marozzi [2015], Marozzi [2016]), or the test based on pseudo-Mahalanobis ranks (Hallin et al. [2002]). Thus, our proposed distance may be of interest in this analysis.

Furthermore, a distance measure is needed in several machine learning algorithms. For example, k-means and k-nearest neighbor (kNN) classifier, need a suitable distance metric to identify neighboring data points. The Euclidean distance is commonly used, but it assumes that each feature of the observations is equally important and independent from the others. In practice, these assumptions may not be always satisfied, especially in high dimension. The Mahalanobis distance has been used for clustering and classification algorithms, such as kNN, discriminant analysis, and many others (Xing et al. [2003], Xiang et al. [2008], Weinberger and Saul [2009], Zhang et al. [2011], Morozova et al. [2013]). Then, it is another field in which the performance of the proposed distance can be studied.

Appendices

APPENDIX A

Proofs from Chapter 3

Here are the proofs of the Propositions from Chapter 3.

A.1 Proof of Proposition 1.

The optimization problem is:

$$\begin{array}{cc} \min_{\nu_{\boldsymbol{\mu}},\eta} & E\left[\left\|\hat{\boldsymbol{\mu}}_{Sh(CCM)}-\boldsymbol{\mu}\right\|_2^2\right] \\ \text{s.t.} & \hat{\boldsymbol{\mu}}_{Sh(CCM)}=(1-\eta)\hat{\boldsymbol{\mu}}_{CCM}+\eta\nu_{\boldsymbol{\mu}}\mathbf{e}, \end{array} \tag{A.1}$$

where $\|\mathbf{x}\|_2^2=\sum_{j=1}^p x_j^2$ and the associated inner product is: $\langle x,y\rangle=\sum_{j=1}^p x_j y_j$.

The objective function is equivalent to:

$$\begin{aligned} &E\left[\left\|\hat{\boldsymbol{\mu}}_{Sh(CCM)}-\boldsymbol{\mu}\right\|_2^2\right] \\ &=E\left[\|(1-\eta)\hat{\boldsymbol{\mu}}_{CCM}+\eta\nu_{\boldsymbol{\mu}}\mathbf{e}-\boldsymbol{\mu}\|_2^2\right] \\ &=(1-\eta)^2E\left[\|\hat{\boldsymbol{\mu}}_{CCM}-\boldsymbol{\mu}\|_2^2\right]+\eta^2\left\|\nu_{\boldsymbol{\mu}}\mathbf{e}-\boldsymbol{\mu}\right\|_2^2 \\ &+2E\left[\langle(1-\eta)(\hat{\boldsymbol{\mu}}_{CCM}-\boldsymbol{\mu}),\eta(\nu_{\boldsymbol{\mu}}\mathbf{e}-\boldsymbol{\mu})\rangle\right]. \end{aligned}$$

The latter element in the above expression is equal to zero because $E(\hat{\boldsymbol{\mu}}_{CCM})=\boldsymbol{\mu}$ (see Chu [1995]). Then, the optimization problem (A.1) reduces to minimize:

$$\begin{aligned} E\left[\left\|\hat{\boldsymbol{\mu}}_{Sh(CCM)}-\boldsymbol{\mu}\right\|_2^2\right]&=(1-\eta)^2E\left[\|\hat{\boldsymbol{\mu}}_{CCM}-\boldsymbol{\mu}\|_2^2\right] \\ &+\eta^2\left\|\nu_{\boldsymbol{\mu}}\mathbf{e}-\boldsymbol{\mu}\right\|_2^2. \end{aligned} \tag{A.2}$$

In order to find the optimal $\nu_{\boldsymbol{\mu}}$, it is necessary to minimize only the right element of the above expression.

$$||\nu_{\boldsymbol{\mu}}\mathbf{e}-\boldsymbol{\mu}||_2^2=\nu_{\boldsymbol{\mu}}^2\|\mathbf{e}\|_2^2+\|\boldsymbol{\mu}\|_2^2-2\nu_{\boldsymbol{\mu}}\langle\mathbf{e},\boldsymbol{\mu}\rangle.$$

Then, with respect to the scaling parameter, the first order optimality condition give:

$$0 = 2p\nu_{\boldsymbol{\mu}} - 2\left\langle \mathbf{e}, \boldsymbol{\mu} \right\rangle = 2\left(p\nu_{\boldsymbol{\mu}} - \sum_{j=1}^{p} \boldsymbol{\mu}_j \right) .$$

Thus:

$$\hat{\nu}_{\boldsymbol{\mu}} = \frac{1}{p} \sum_{j=1}^{p} \boldsymbol{\mu}_j \, .$$

Estimating $\boldsymbol{\mu}$ with $\hat{\boldsymbol{\mu}}_{CCM}$, we obtain:

$$\hat{\nu}_{\boldsymbol{\mu}} = \frac{\hat{\boldsymbol{\mu}}_{CCM}\mathbf{e}}{p} \, .$$

In (A.2), with respect to the shrinkage intensity parameter η, the first order optimality condition give:

$$0 = 2(1-\eta)E\left[\|\hat{\boldsymbol{\mu}}_{CCM} - \boldsymbol{\mu}\|_2^2 \right] + 2\eta \, \|\nu_{\boldsymbol{\mu}}\mathbf{e} - \boldsymbol{\mu}\|_2^2 \, .$$

Hence:

$$\hat{\eta} = \frac{E\left[\|\hat{\boldsymbol{\mu}}_{CCM} - \boldsymbol{\mu}\|_2^2 \right]}{E\left[\|\hat{\boldsymbol{\mu}}_{CCM} - \hat{\nu}_{\boldsymbol{\mu}}\mathbf{e}\|_2^2 \right]} \, .$$

A.2 Proof of Proposition 2.

The optimization problem is:

$$\begin{array}{cc} \min_{\nu_{\boldsymbol{\mu}},\eta} & E\left[\left\| \hat{\boldsymbol{\mu}}_{Sh(MM)} - \boldsymbol{\mu} \right\|_2^2 \right] \\ \text{s.t.} & \hat{\boldsymbol{\mu}}_{Sh(MM)} = (1-\eta)\hat{\boldsymbol{\mu}}_{MM} + \eta\nu_{\boldsymbol{\mu}}\mathbf{e}, \end{array} \tag{A.3}$$

where $\|x\|_2^2 = \sum_{j=1}^{p} x_j^2$.

Similarly to the previous demonstration, we can consider the following expression for the objective function:

$$\begin{aligned} E\left[\left\| \hat{\boldsymbol{\mu}}_{Sh(MM)} - \boldsymbol{\mu} \right\|_2^2 \right] &= E\left[\|(1-\eta)\hat{\boldsymbol{\mu}}_{MM} + \eta\nu_{\boldsymbol{\mu}}\mathbf{e} - \boldsymbol{\mu}\|_2^2 \right] \\ &= (1-\eta)^2 E\left[\|\hat{\boldsymbol{\mu}}_{MM} - \boldsymbol{\mu}\|_2^2 \right] + \eta^2 \, \|\nu_{\boldsymbol{\mu}}\mathbf{e} - \boldsymbol{\mu}\|_2^2 \\ &\quad + 2E\left[\langle (1-\eta)(\hat{\boldsymbol{\mu}}_{MM} - \boldsymbol{\mu}), \eta(\nu_{\boldsymbol{\mu}}\mathbf{e} - \boldsymbol{\mu}) \rangle \right] \, . \end{aligned}$$

The expectation of the inner product is equal to zero because Bose and Chaudhuri [1993], Bose [1995], Möttönen et al. [2010], and Becker et al. [2014] investigated the asymptotic distribution for the L_1-median. According to the authors the distribution of $\hat{\boldsymbol{\mu}}_{MM}$ can be approximated by $N_p\left(\boldsymbol{\mu}, \frac{1}{n}\hat{A}^{-1}\hat{B}\hat{A}^{-1} \right)$, where

$\hat{A}(\mathbf{x}_i) = \frac{1}{||\mathbf{x}_i||_2}\left(I_p - \frac{\mathbf{x}_i\mathbf{x}_i^T}{||\mathbf{x}_i||_2^2}\right)$ and $\hat{B}(\mathbf{x}_i) = \frac{\mathbf{x}_i\mathbf{x}_i^T}{||\mathbf{x}_i||_2^2}$, with $\mathbf{x}_i \in \mathbb{R}^p$, for each $i = 1, ..., n$.

Then, the optimization problem (A.3) reduces to minimize:

$$E\left[\left\|\hat{\boldsymbol{\mu}}_{Sh(MM)} - \boldsymbol{\mu}\right\|_2^2\right] = (1-\eta)^2 E\left[\|\hat{\boldsymbol{\mu}}_{MM} - \boldsymbol{\mu}\|_2^2\right] + \eta^2 \|\nu_{\boldsymbol{\mu}}\mathbf{e} - \boldsymbol{\mu}\|_2^2 . \tag{A.4}$$

Then, the optimal parameter $\nu_{\boldsymbol{\mu}}$ can be found minimizing only the right element of the above expression, which is the only one depending on that parameter.

$$||\nu_{\boldsymbol{\mu}}\mathbf{e} - \boldsymbol{\mu}||_2^2 = \nu_{\boldsymbol{\mu}}^2 \|\mathbf{e}\|_2^2 + \|\boldsymbol{\mu}\|_2^2 - 2\nu_{\boldsymbol{\mu}} \langle \mathbf{e}, \boldsymbol{\mu} \rangle .$$

The associated first order optimality condition give:

$$0 = 2p\nu_{\boldsymbol{\mu}} - 2\langle \mathbf{e}, \boldsymbol{\mu} \rangle = 2\left(p\nu_{\boldsymbol{\mu}} - \sum_{j=1}^{p} \boldsymbol{\mu}_j\right) .$$

Therefore:

$$\hat{\nu}_{\boldsymbol{\mu}} = \frac{1}{p}\sum_{j=1}^{p} \boldsymbol{\mu}_j .$$

In practice, we propose to estimate $\boldsymbol{\mu}$ with $\hat{\boldsymbol{\mu}}_{MM}$. Thus:

$$\hat{\nu}_{\boldsymbol{\mu}} = \frac{\hat{\boldsymbol{\mu}}_{MM}\mathbf{e}}{p} .$$

With respect to the shrinkage intensity parameter η, the first order optimality condition associated to (A.4), give:

$$0 = 2(1-\eta)E\left[\|\hat{\boldsymbol{\mu}}_{MM} - \boldsymbol{\mu}\|_2^2\right] + 2\eta \|\nu_{\boldsymbol{\mu}}\mathbf{e} - \boldsymbol{\mu}\|_2^2 .$$

Hence:

$$\hat{\eta} = \frac{E\left[\|\hat{\boldsymbol{\mu}}_{MM} - \boldsymbol{\mu}\|_2^2\right]}{E\left[\|\hat{\boldsymbol{\mu}}_{MM} - \hat{\nu}_{\boldsymbol{\mu}}\mathbf{e}\|_2^2\right]} .$$

A.3 Proof of Proposition 3.

The optimization problem is:

$$\begin{array}{ll} \min_{\nu_\Sigma, \eta} & E\left[\left\|\hat{\Sigma}_{Sh} - \Sigma\right\|^2\right] \\ \text{s.t.} & \hat{\Sigma}_{Sh} = (1-\eta)\hat{S}_{CCM} + \eta\nu_\Sigma I, \end{array} \tag{A.5}$$

where $\|A\|^2 = trace(AA^T)/p$, and the associated inner product is $\langle A_1, A_2 \rangle = trace(A_1A_2^T)/p$.

Analogous to the previous Propositions, the objective function in the above minimization problem (A.5) can be seen as:

$$\begin{aligned}
&E\left[\left\|\hat{\Sigma}_{Sh}-\Sigma\right\|^2\right] \\
&= E\left[\left\|(1-\eta)\hat{S}_{CCM}+\eta\nu_\Sigma I-\Sigma\right\|^2\right] \\
&= (1-\eta)^2 E\left[\left\|\hat{S}_{CCM}-\Sigma\right\|^2\right]+\eta^2\left\|\nu_\Sigma I-\Sigma\right\|^2 \\
&+ 2E\left[\left\langle(1-\eta)(\hat{S}_{CCM}-\Sigma),\eta(\nu_\Sigma I-\Sigma)\right\rangle\right] .
\end{aligned}$$

In this case, note that the latter element in the above expression is equal to zero because $E(\hat{S}_{CCM})=\Sigma$. Hence, the optimization problem (A.5) reduces to minimize the following expression:

$$E\left[\left\|\hat{\Sigma}_{Sh}-\Sigma\right\|^2\right] = (1-\eta)^2 E\left[\left\|\hat{S}_{CCM}-\Sigma\right\|^2\right]+\eta^2\left\|\nu_\Sigma I-\Sigma\right\|^2 . \tag{A.6}$$

The optimal ν_Σ can be obtained by minimizing only the right element of the above expression, because it is the only one depending on that parameter. Also, note that:

$$||\nu_\Sigma I-\Sigma||^2 = \nu_\Sigma^2\left\|I\right\|^2+\left\|\Sigma\right\|^2-2\nu_\Sigma\left\langle I,\Sigma\right\rangle .$$

Then, the first order optimality condition with respect to the scaling parameter, give:

$$\begin{aligned}
0 &= 2\nu_\Sigma - 2\left\langle I,\Sigma\right\rangle . \\
\nu_\Sigma &= \left\langle I,\Sigma\right\rangle = trace(\Sigma I^T)/p .
\end{aligned}$$

Therefore:

$$\hat{\nu}_\Sigma = trace(\Sigma)/p .$$

In practice, we propose to estimate Σ with $\hat{S}_{CCM}$, thus:

$$\hat{\nu}_\Sigma = trace(\hat{S}_{CCM})/p .$$

In (A.6), with respect to the shrinkage intensity parameter η, the first order optimality condition give:

$$\hat{\eta} = \frac{E\left[\left\|\hat{S}_{CCM}-\Sigma\right\|^2\right]}{E\left[\left\|\hat{S}_{CCM}-\hat{\nu}_\Sigma I\right\|^2\right]} .$$

APPENDIX B

Tables from Chapter 3

Here are the tables that were not included in the main text from Chapter 3.

B.1 Normal distribution

Table B.1 shows the false positive rates (FPR) when there is no contamination. The Tables B.2-B.3 show the true positive rates (TPR) and Tables B.4-B.5 the FPR, for each method, corresponding to the simulations with multivariate Normal distribution, for contamination levels $\alpha = 0.1, 0.2, 0.3$, dimension $p = 5, 10, 30, 50$, distance of the outliers $\delta = 5$ and 10, and concentration of the contamination $\lambda = 0.1$ and 1.

Table B.1: False positive rates with Normal distribution $\alpha = 0$.

p	MCD	Adj MCD	Kurtosis	OGK	COM	RMDv1	RMDv2	RMDv3	RMDv4	RMDv5	RMDv6
5	0.06440	0.04640	0.02630	0.08120	0.00480	0.03140	0.02820	0.02770	0.03100	0.02900	0.00316
10	0.11760	0.09830	0.07110	0.09580	0.00217	0.00245	0.00222	0.00215	0.00172	0.00161	0.00160
30	0.06276	0.03922	0.00804	0.09084	0.00008	0.00003	0.00003	0.00003	0.00003	0.00002	0.00001
50	0.05860	0.03460	0.00630	0.08670	0.00005	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001

Table B.2: True positive rates with Normal distribution.

$\delta = 5$		$\lambda = 0.1$										
p	α	MCD	Adj MCD	Kurtosis	OGK	COM	RMDv1	RMDv2	RMDv3	RMDv4	RMDv5	RMDv6
5	0.1	1	1	0.9000	1	1	1	1	1	1	1	1
	0.2	0.8700	0.8700	0.5100	0.9500	0.9941	1	1	1	1	1	1
	0.3	0.0600	0.0600	0.9800	0.1500	0.5719	0.8766	0.8782	0.8782	0.9146	0.9090	0.9130
10	0.1	0.9900	0.9900	0.8600	1	1	1	1	1	1	1	1
	0.2	0.2800	0.2800	0.4600	0.9416	1	1	1	1	1	1	1
	0.3	0	0	0.9900	0.1612	0.7205	0.8774	0.8747	0.8750	0.9711	0.9672	0.9711
30	0.1	0.1900	0.1900	1	1	1	1	1	1	1	1	1
	0.2	0	0	0.1000	1	1	1	1	1	1	1	1
	0.3	0	0	0.6100	0.0100	0.9407	0.5308	0.5275	0.5286	0.9990	0.9988	0.9991
50	0.1	0	0	1	1	1	1	1	1	1	1	1
	0.2	0	0	0	1	1	1	1	1	1	1	1
	0.3	0	0	0	0	0.9839	0.5021	0.5000	0.5000	0.9939	0.9932	0.9942
$\delta = 5$		$\lambda = 1$										
p	α	MCD	Adj MCD	Kurtosis	OGK	COM	RMDv1	RMDv2	RMDv3	RMDv4	RMDv5	RMDv6
5	0.1	1	1	1	1	1	1	1	1	1	1	1
	0.2	0.8578	0.8486	0.9602	0.9654	0.9975	1	1	1	1	1	1
	0.3	0.1955	0.1664	0.9336	0.5792	0.8735	0.8935	0.8947	0.8938	0.8740	0.8698	0.8755
10	0.1	1	1	1	1	1	1	1	1	1	1	1
	0.2	0.9016	0.8935	0.7212	0.9993	1	1	1	1	1	1	1
	0.3	0.2375	0.2080	0.5935	0.6108	0.9505	0.8846	0.8838	0.8838	0.9608	0.9581	0.9621
30	0.1	1	1	0.8816	1	1	1	1	1	1	1	1
	0.2	0.4461	0.4232	0.0154	1	1	1	1	1	1	1	1
	0.3	0.0823	0.0532	0.1483	0.9772	0.9990	0.7142	0.7035	0.7059	1	1	1
50	0.1	0.0901	0.0801	0.6708	1	1	1	1	1	1	1	1
	0.2	0.0801	0.0515	0.0019	1	1	1	1	1	1	1	1
	0.3	0.0671	0.0367	0.0111	0.9087	0.9987	0.5425	0.5339	0.5363	0.9997	0.9997	0.9997

Table B.3: True positive rates with Normal distribution.

$\delta = 10$		$\lambda = 0.1$										
p	α	MCD	Adj MCD	Kurtosis	OGK	COM	RMDv1	RMDv2	RMDv3	RMDv4	RMDv5	RMDv6
5	0.1	1	1	0.9200	1	1	1	1	1	1	1	1
	0.2	0.8200	0.8200	0.6500	1	1	1	1	1	1	1	1
	0.3	0.1000	0.1000	1	0.7400	1	1	1	1	1	1	1
10	0.1	1	1	0.9200	1	1	1	1	1	1	1	1
	0.2	0.7200	0.7200	0.4400	1	1	1	1	1	1	1	1
	0.3	0.0500	0.0500	0.9700	0.7500	1	1	1	1	1	1	1
30	0.1	0.8800	0.8800	1	1	1	1	1	1	1	1	1
	0.2	0	0	0.1200	1	1	1	1	1	1	1	1
	0.3	0	0	0.5400	0.9300	1	1	1	1	1	1	1
50	0.1	0	0	1	1	1	1	1	1	1	1	1
	0.2	0	0	0	1	1	1	1	1	1	1	1
	0.3	0	0	0	0.9787	1	1	1	1	1	1	1
$\delta = 10$		$\lambda = 1$										
p	α	MCD	Adj MCD	Kurtosis	OGK	COM	RMDv1	RMDv2	RMDv3	RMDv4	RMDv5	RMDv6
5	0.1	1	1	1	1	1	1	1	1	1	1	1
	0.2	0.8480	0.8465	0.9900	1	1	1	1	1	1	1	1
	0.3	0.2190	0.1976	0.9307	0.9591	0.9991	1	1	1	1	1	1
10	0.1	1	1	0.9800	1	1	1	1	1	1	1	1
	0.2	0.8623	0.8548	0.6558	1	1	1	1	1	1	1	1
	0.3	0.2280	0.2046	0.4618	0.9911	1	1	1	1	1	1	1
30	0.1	1	1	0.8919	1	1	1	1	1	1	1	1
	0.2	0.4879	0.4654	0.0125	1	1	1	1	1	1	1	1
	0.3	0.0810	0.0509	0.1087	1	1	1	1	1	1	1	1
50	0.1	1	1	0.6017	1	1	1	1	1	1	1	1
	0.2	0.2695	0.2348	0.0017	1	1	1	1	1	1	1	1
	0.3	0.0643	0.0378	0.0006	1	1	1	1	1	1	1	1

Table B.4: False positive rates with Normal distribution.

$\delta = 5$		$\lambda = 0.1$										
p	α	MCD	Adj MCD	Kurtosis	OGK	COM	RMDv1	RMDv2	RMDv3	RMDv4	RMDv5	RMDv6
5	0.1	0.0327	0.0184	0.0336	0.0640	0.0040	0.0080	0.0070	0.0073	0.0069	0.0067	0.0076
	0.2	0.0265	0.0171	0.0512	0.0600	0.0028	0.0015	0.0015	0.0012	0.0013	0.0013	0.0013
	0.3	0.1504	0.1247	0.0373	0.1641	0.0015	0	0	0	0	0	0
10	0.1	0.0735	0.0529	0.1167	0.0823	0.0027	0.0055	0.0057	0.0048	0.0025	0.0024	0.0027
	0.2	0.1566	0.1330	0.2803	0.0667	0.0023	0.0002	0.0002	0.0001	0.0001	0.0001	0.0001
	0.3	0.2485	0.2206	0.0767	0.2502	0.0010	0	0	0	0	0	0
30	0.1	0.0767	0.0507	0.0078	0.0699	5E-05	0.0007	0.0006	0.0006	0.0001	4.4E-05	0.0001
	0.2	0.1079	0.0784	0.0565	0.0552	3E-05	0	0	0	0	0	0
	0.3	0.1491	0.1150	0.0547	0.5030	3E-05	0	0	0	0	0	0
50	0.1	0.0679	0.0411	0.0010	0.0688	0.0003	0.0001	0.0001	0.0001	0	0	0
	0.2	0.0916	0.0609	0.0045	0.0529	0	0	0	0	0	0	0
	0.3	0.1362	0.1013	0.0174	0.6761	0	0	0	0	0	0	0
$\delta = 5$		$\lambda = 1$										
p	α	MCD	Adj MCD	Kurtosis	OGK	COM	RMDv1	RMDv2	RMDv3	RMDv4	RMDv5	RMDv6
5	0.1	0.0339	0.0198	0.0341	0.0636	0.0033	0.0084	0.0073	0.0069	0.0075	0.0071	0.0077
	0.2	0.0110	0.0055	0.0347	0.0507	0.0017	0.0014	0.0012	0.0012	0.0010	0.0009	0.0012
	0.3	0.0374	0.0260	0.0379	0.0383	0.0001	0	0	0	0	0	0
10	0.1	0.0704	0.0505	0.0917	0.0842	0.0023	0.0047	0.0042	0.0041	0.0027	0.0022	0.0029
	0.2	0.0270	0.0171	0.0860	0.0610	0.0010	0.0002	0.0002	0.0002	0.0001	0.0001	0.0001
	0.3	0.0851	0.0706	0.0782	0.0457	0.0003	0	0	0	0	0	0
30	0.1	0.0347	0.0115	0.0081	0.0713	0.0001	0.0008	0.0007	0.0007	0.0001	0.0001	0.0001
	0.2	0.0357	0.0198	0.0066	0.0544	0	0	0	0	0	0	0
	0.3	0.0552	0.0343	0.0077	0.0366	0	0	0	0	0	0	0
50	0.1	0.0273	0.0046	0.0006	0.0710	0.0002	0.0001	0.0001	0.0001	0	0	0
	0.2	0.0466	0.0259	0.0008	0.0540	0	0	0	0	0	0	0
	0.3	0.0491	0.0277	0.0009	0.0361	0	0	0	0	0	0	0

Table B.5: False positive rates with Normal distribution.

$\delta = 10$		$\lambda = 0.1$										
p	α	MCD	Adj MCD	Kurtosis	OGK	COM	RMDv1	RMDv2	RMDv3	RMDv4	RMDv5	RMDv6
5	0.1	0.0309	0.0154	0.0301	0.0681	0.0022	0.0090	0.0076	0.0066	0.0066	0.0058	0.0069
	0.2	0.0337	0.0248	0.0414	0.0485	0.0036	0.0005	0.0004	0.0004	0.0011	0.0008	0.0007
	0.3	0.1434	0.1183	0.0309	0.0603	0.0031	0	0	0	0	0	0
10	0.1	0.0665	0.0464	0.1147	0.0772	0.0017	0.0041	0.0047	0.0042	0.0034	0.0030	0.0035
	0.2	0.0809	0.0647	0.2827	0.0606	0.0022	0.0002	0.0001	0.0001	0	0	0
	0.3	0.2353	0.2088	0.0934	0.0895	0.0008	0	0	0	0	0	0
30	0.1	0.0415	0.0174	0.0081	0.0715	4E-05	0.0011	0.0010	0.0011	0.0001	0.0001	0.0001
	0.2	0.1072	0.0776	0.0481	0.0561	3E-05	0	0	0	0	0	0
	0.3	0.1500	0.1163	0.0578	0.0623	0.0001	0	0	0	0	0	0
50	0.1	0.0698	0.0426	0.0006	0.0704	0.0002	0.0003	0.0002	0.0003	0	0	0
	0.2	0.0954	0.0642	0.0032	0.0534	0.0001	0	0	0	0	0	0
	0.3	0.1257	0.0913	0.0091	0.0400	0	0	0	0	0	0	0
$\delta = 10$		$\lambda = 1$										
p	α	MCD	Adj MCD	Kurtosis	OGK	COM	RMDv1	RMDv2	RMDv3	RMDv4	RMDv5	RMDv6
5	0.1	0.0332	0.0174	0.0297	0.0682	0.0021	0.0076	0.0068	0.0069	0.0058	0.0052	0.0062
	0.2	0.0136	0.0058	0.0352	0.0535	0.0023	0.0011	0.0012	0.0011	0.0008	0.0008	0.0007
	0.3	0.0385	0.0289	0.0372	0.0397	0.0007	0	0	0	0.0001	0	0
10	0.1	0.0668	0.0469	0.0907	0.0771	0.0020	0.0036	0.0036	0.0031	0.0025	0.0021	0.0026
	0.2	0.0278	0.0170	0.0930	0.0629	0.0009	0.0002	0.0001	0.0002	0	0	0
	0.3	0.0856	0.0694	0.0685	0.0440	0.0004	0	0	0	0	0	0
30	0.1	0.0351	0.0122	0.0077	0.0735	4E-05	0.0009	0.0008	0.0009	0.0001	2.21E-05	0.0001
	0.2	0.0334	0.0181	0.0053	0.0535	0	0	0	0	0	0	0
	0.3	0.0577	0.0368	0.0085	0.0388	0	0	0	0	0	0	0
50	0.1	0.0249	0.0033	0.0003	0.0717	0.0007	0	0	0	0	0	0
	0.2	0.0377	0.0204	0.0002	0.0552	0.0005	0	0	0	0	0	0
	0.3	0.0493	0.0264	0.0002	0.0373	0	0	0	0	0	0	0

Table B.6: Computational times with Normal data $\delta = 5$ and $\lambda = 1$.

p	α	MCD	Adj MCD	Kurtosis	OGK	COM	RMD-S
5	0.1	0.8547	0.8078	0.0959	0.0181	0.0070	0.0029
	0.2	1.2146	0.7129	0.0763	0.0186	0.0061	0.0034
	0.3	1.0064	0.7544	0.0612	0.0176	0.0063	0.0025
	Mean	**1.0252**	**0.7584**	**0.0778**	**0.0181**	**0.0065**	**0.0030**
10	0.1	1.0090	1.1250	0.1592	0.0793	0.0113	0.0047
	0.2	1.0135	1.0448	0.1679	0.0623	0.0100	0.0025
	0.3	1.0335	1.0595	0.1515	0.0612	0.0091	0.0009
	Mean	**1.0187**	**1.0765**	**0.1595**	**0.0676**	**0.0101**	**0.0027**
30	0.1	6.5263	6.2629	0.4788	0.9623	0.2530	0.2752
	0.2	6.1268	6.2031	0.5737	0.8317	0.1700	0.2139
	0.3	5.9068	6.1767	0.5034	0.9472	0.1791	0.2101
	Mean	**6.1866**	**6.2142**	**0.5186**	**0.9137**	**0.2007**	**0.2331**
50	0.1	7.3298	7.2712	2.3726	1.2543	0.2172	0.2122
	0.2	7.2441	7.2303	2.3827	1.2277	0.2165	0.2066
	0.3	7.2472	7.2544	2.5319	1.2322	0.2167	0.2105
	Mean	**7.2737**	**7.2520**	**2.4291**	**1.2381**	**0.2168**	**0.2098**

Table B.7: Computational times with Normal data $\delta = 10$ and $\lambda = 0.1$.

p	α	MCD	Adj MCD	Kurtosis	OGK	COM	RMD-S
5	0.1	0.7704	0.7425	0.1090	0.0195	0.0107	0.0026
	0.2	0.6878	0.6534	0.0573	0.0195	0.0079	0.0033
	0.3	0.6990	0.7434	0.0294	0.0268	0.0072	0.0054
	Mean	**0.7191**	**0.7131**	**0.0652**	**0.0219**	**0.0086**	**0.0038**
10	0.1	1.0631	1.1115	0.1297	0.0824	0.0101	0.0066
	0.2	1.1940	0.9625	0.1179	0.0719	0.0080	0.0041
	0.3	1.0673	0.9902	0.0756	0.0663	0.0097	0.0047
	Mean	**1.1081**	**1.0214**	**0.1077**	**0.0735**	**0.0093**	**0.0051**
30	0.1	6.0350	6.0913	0.7347	0.8205	0.1701	0.1697
	0.2	6.4506	6.1627	0.7385	0.7491	0.1837	0.1572
	0.3	6.2114	6.1146	1.1310	0.7342	0.1714	0.1281
	Mean	**6.2324**	**6.1229**	**0.8681**	**0.7679**	**0.1750**	**0.1517**
50	0.1	7.3524	7.2659	2.3297	1.3001	0.2170	0.2167
	0.2	7.2749	7.2592	2.4093	1.2846	0.2166	0.2132
	0.3	7.2559	7.2307	2.4729	1.2732	0.2166	0.2091
	Mean	**7.2944**	**7.2519**	**2.4040**	**1.2860**	**0.2167**	**0.2130**

Table B.8: Computational times with Normal data $\delta = 10$ and $\lambda = 1$.

p	α	MCD	Adj MCD	Kurtosis	OGK	COM	RMD-S
5	0.1	0.7956	0.8301	0.0814	0.0176	0.0078	0.0037
	0.2	0.7939	0.7684	0.0836	0.0189	0.0124	0.0019
	0.3	0.8614	0.7207	0.0662	0.0183	0.0049	0.0014
	Mean	**0.8170**	**0.7731**	**0.0770**	**0.0183**	**0.0084**	**0.0023**
10	0.1	0.9990	1.0609	0.1350	0.0634	0.0117	0.0047
	0.2	1.0917	1.1028	0.1613	0.0682	0.0093	0.0049
	0.3	1.0111	1.1860	0.1610	0.0766	0.0089	0.0025
	Mean	**1.0340**	**1.1166**	**0.1524**	**0.0694**	**0.0100**	**0.0040**
30	0.1	5.7161	5.6622	0.5731	0.7563	0.1693	0.1220
	0.2	5.6394	5.6993	0.5821	0.7191	0.1654	0.1083
	0.3	5.7104	5.7975	0.9465	0.7193	0.1561	0.1138
	Mean	**5.6886**	**5.7197**	**0.7006**	**0.7316**	**0.1636**	**0.1147**
50	0.1	7.3088	7.2330	2.2537	1.2833	0.2170	0.2081
	0.2	7.2340	7.2334	2.2005	1.2644	0.2165	0.2055
	0.3	7.2644	7.2349	2.2755	1.2685	0.2166	0.2055
	Mean	**7.2691**	**7.2338**	**2.2432**	**1.2720**	**0.2167**	**0.2063**

B.2 Multivariate Student-t distribution with 3 degrees of freedom

Table B.9 shows the false positive rates (FPR) when there is no contamination. Tables B.10-B.11 show the true positive rates (TPR) and Tables B.12-B.13 the FPR, for each method, corresponding to the simulations with multivariate Student-t distribution with 3 degrees of freedom, for contamination levels $\alpha = 0.1, 0.2, 0.3$, dimension $p = 5, 10, 30, 50$, distance of the outliers $\delta = 5$ and 10, and concentration of the contamination $\lambda = 0.1$ and 1.

Table B.9: False positive rates with Student-t distribution with 3 d.f, $\alpha = 0$.

p	MCD	Adj MCD	Kurtosis	OGK	COM	RMDv1	RMDv2	RMDv3	RMDv4	RMDv5	RMDv6
5	0.12860	0.11060	0.17670	0.21320	0.07730	0.20290	0.20070	0.20030	0.17950	0.17660	0.18370
10	0.17560	0.15640	0.31610	0.24330	0.08410	0.19040	0.19550	0.19570	0.11780	0.11220	0.12760
30	0.15820	0.13530	0.20380	0.28760	0.08270	0.16660	0.16430	0.16490	0.11220	0.11130	0.11400
50	0.15280	0.12920	0.12620	0.30900	0.07660	0.07152	0.07143	0.07156	0.07113	0.07108	0.07124

Table B.10: True positive rates with Student-t distribution with 3 d.f.

$\delta = 5$		$\lambda = 0.1$										
p	α	MCD	Adj MCD	Kurtosis	OGK	COM	RMDv1	RMDv2	RMDv3	RMDv4	RMDv5	RMDv6
5	0.1	0.9300	0.9300	0.4900	1	1	1	1	1	1	1	1
	0.2	0.1800	0.1800	0.2902	0.5300	0.6872	1	1	1	0.9927	0.9915	0.9915
	0.3	0.0007	0.0007	0.6890	0.0007	0.0411	0.6573	0.6446	0.6441	0.6878	0.6758	0.7072
10	0.1	0.5107	0.5107	0.3407	0.9800	1	1	1	1	1	1	1
	0.2	0.0009	0.0009	0.1809	0.3709	0.6576	1	1	1	0.9997	0.9997	1
	0.3	0.0100	0.0100	0.6200	0.0203	0.0400	0.6500	0.6525	0.6522	0.8021	0.7947	0.8137
30	0.1	0.0003	0.0003	0.2000	1	1	1	1	1	1	1	1
	0.2	0.0004	0.0004	0.1703	0.1905	0.6055	1	1	1	1	1	1
	0.3	0	0	1	0.0002	0	0.1689	0.1638	0.1649	0.7437	0.7417	0.7628
50	0.1	0	0	0.4000	1	1	1	1	1	1	1	1
	0.2	0	0	0.4000	0	0.9004	1	1	1	1	1	1
	0.3	0	0	0.7900	0.0003	0	0.4200	0.4210	0.4180	0.8937	0.8938	0.8939
$\delta = 5$		$\lambda = 1$										
p	α	MCD	Adj MCD	Kurtosis	OGK	COM	RMDv1	RMDv2	RMDv3	RMDv4	RMDv5	RMDv6
5	0.1	1	1	1	1	1	1	1	1	1	1	1
	0.2	0.7445	0.7263	0.8974	0.9177	0.9155	0.9970	0.9952	0.9952	0.9922	0.9919	0.9924
	0.3	0.1844	0.1591	0.8421	0.4498	0.4732	0.8442	0.8485	0.8457	0.8411	0.8360	0.8377
10	0.1	0.9826	0.9819	0.9492	1	1	1	1	1	1	1	1
	0.2	0.5344	0.5237	0.5320	0.9341	0.9629	1	1	1	1	0.9996	0.9996
	0.3	0.1864	0.1637	0.6358	0.3756	0.4901	0.8602	0.8571	0.8575	0.9126	0.9063	0.9126
30	0.1	0.9923	0.9917	0.4932	1	1	1	1	1	1	1	1
	0.2	0.1823	0.1581	0.2148	1	1	1	1	1	1	1	1
	0.3	0.1658	0.1403	0.5935	0.4170	0.8556	0.6581	0.6485	0.6501	0.9670	0.9664	0.9689
50	0.1	0.7572	0.7464	0.1536	1	1	1	1	1	1	1	1
	0.2	0.1718	0.1407	0.1394	1	1	1	1	1	1	1	1
	0.3	0.1656	0.1401	0.5530	0.4156	0.9948	0.6716	0.6626	0.6644	0.9989	0.9839	0.9991

Table B.11: True positive rates with Student-t distribution with 3 d.f.

$\delta = 10$		$\lambda = 0.1$										
p	α	MCD	Adj MCD	Kurtosis	OGK	COM	RMDv1	RMDv2	RMDv3	RMDv4	RMDv5	RMDv6
5	0.1	0.9900	0.9900	0.6600	1	1	1	1	1	1	1	1
	0.2	0.5400	0.5400	0.4700	0.9600	1	1	1	1	1	1	1
	0.3	0.0400	0.0400	0.8985	0.3803	0.7211	0.9900	0.9900	0.9900	1	1	0.9900
10	0.1	0.9300	0.9300	0.4900	1	1	1	1	1	1	1	1
	0.2	0.1000	0.1000	0.2200	0.9300	1	1	1	1	1	1	1
	0.3	0	0	0.7804	0.2200	0.7709	1	1	1	1	1	1
30	0.1	0.0200	0.0200	0.2594	1	1	1	1	1	1	1	1
	0.2	0	0	0.3098	1	1	1	1	1	1	1	1
	0.3	0.0001	0.0001	1	0.0003	0.7862	1	1	1	1	1	1
50	0.1	0	0	0.1000	1	1	1	1	1	1	1	1
	0.2	0.0004	0.0004	0.2004	1	1	1	1	1	1	1	1
	0.3	0	0	1	0.0003	0.8174	1	1	1	1	1	1
$\delta = 10$		$\lambda = 1$										
p	α	MCD	Adj MCD	Kurtosis	OGK	COM	RMDv1	RMDv2	RMDv3	RMDv4	RMDv5	RMDv6
5	0.1	1	1	1	1	1	1	1	1	1	1	1
	0.2	0.8890	0.8859	0.9580	1	1	1	1	1	1	1	1
	0.3	0.2324	0.2127	0.9470	0.7598	0.9791	0.9998	0.9998	0.9998	1	1	1
10	0.1	1	1	0.9830	1	1	1	1	1	1	1	1
	0.2	0.7050	0.6964	0.5019	1	1	1	1	1	1	1	1
	0.3	0.2563	0.2314	0.6768	0.8453	0.9914	1	1	1	1	1	1
30	0.1	1	1	0.5614	1	1	1	1	1	1	1	1
	0.2	0.2407	0.2166	0.2203	1	1	1	1	1	1	1	1
	0.3	0.1640	0.1394	0.7130	0.9624	1	1	1	1	1	1	1
50	0.1	0.9179	0.9141	0.1338	1	1	1	1	1	1	1	1
	0.2	0.1828	0.1595	0.1487	1	1	1	1	1	1	1	1
	0.3	0.1571	0.1304	0.6445	0.9791	1	1	1	1	1	1	1

Table B.12: False positive rates with Student-t distribution with 3 d.f.

$\delta = 5$		$\lambda = 0.1$										
p	α	MCD	Adj MCD	Kurtosis	OGK	COM	RMDv1	RMDv2	RMDv3	RMDv4	RMDv5	RMDv6
5	0.1	0.0901	0.0741	0.2020	0.1840	0.0641	0.1305	0.1293	0.1289	0.1155	0.1153	0.1208
	0.2	0.1591	0.1381	0.3055	0.2014	0.0494	0.0741	0.0743	0.0748	0.0668	0.0656	0.0694
	0.3	0.2237	0.1972	0.2046	0.4273	0.0654	0.0299	0.0300	0.0300	0.0298	0.0298	0.0310
10	0.1	0.1724	0.1528	0.3689	0.2149	0.0703	0.1876	0.1851	0.1874	0.1289	0.1259	0.1377
	0.2	0.2460	0.2224	0.4709	0.2743	0.0534	0.0896	0.0882	0.0890	0.0643	0.0636	0.0687
	0.3	0.2978	0.2723	0.3106	0.5624	0.0833	0.0411	0.0411	0.0410	0.0375	0.0364	0.0387
30	0.1	0.1775	0.1523	0.3154	0.2626	0.0717	0.3472	0.3450	0.3452	0.1563	0.1557	0.1581
	0.2	0.2116	0.1831	0.4937	0.4232	0.0551	0.1459	0.1454	0.1452	0.0779	0.0777	0.0785
	0.3	0.2562	0.2235	0.1669	0.8151	0.0991	0.0418	0.0418	0.0417	0.0401	0.0400	0.0446
50	0.1	0.1769	0.1506	0.1672	0.2841	0.0653	0.2589	0.2566	0.2574	0.1150	0.1144	0.1144
	0.2	0.2013	0.1720	0.4487	0.5081	0.0639	0.1124	0.1130	0.1127	0.0680	0.0680	0.0680
	0.3	0.2552	0.2213	0.1423	0.9152	0.1091	0.0424	0.0431	0.0430	0.0412	0.0410	0.0364
$\delta = 5$		$\lambda = 1$										
p	α	MCD	Adj MCD	Kurtosis	OGK	COM	RMDv1	RMDv2	RMDv3	RMDv4	RMDv5	RMDv6
5	0.1	0.0809	0.0662	0.2027	0.1903	0.0664	0.1361	0.1333	0.1346	0.1194	0.1183	0.1232
	0.2	0.0619	0.0500	0.1846	0.1613	0.0513	0.0800	0.0799	0.0796	0.0711	0.0702	0.0727
	0.3	0.1098	0.0951	0.1537	0.1340	0.0362	0.0375	0.0370	0.0372	0.0384	0.0380	0.0394
10	0.1	0.1209	0.1032	0.2985	0.2166	0.0769	0.1922	0.1910	0.1927	0.1414	0.1383	0.1497
	0.2	0.1235	0.1073	0.2940	0.1805	0.0538	0.0986	0.0960	0.0962	0.0710	0.0701	0.0754
	0.3	0.1705	0.1523	0.2096	0.1712	0.0360	0.0392	0.0394	0.0397	0.0361	0.0355	0.0375
30	0.1	0.1039	0.0819	0.1931	0.2587	0.0694	0.3466	0.3454	0.3460	0.1547	0.1540	0.1570
	0.2	0.1514	0.1291	0.1888	0.2311	0.0557	0.1524	0.1519	0.1520	0.0781	0.0778	0.0788
	0.3	0.1525	0.1304	0.1681	0.2248	0.0392	0.0445	0.0447	0.0444	0.0364	0.0363	0.0368
50	0.1	0.1154	0.0922	0.1403	0.2909	0.0684	0.2425	0.2411	0.2415	0.1111	0.1109	0.1111
	0.2	0.1568	0.1351	0.1239	0.2575	0.0658	0.1174	0.1179	0.1174	0.0672	0.0672	0.0678
	0.3	0.1523	0.1295	0.1082	0.2502	0.0432	0.0568	0.0562	0.0562	0.0408	0.0408	0.0405

Table B.13: False positive rates with Student-t distribution with 3 d.f.

$\delta = 10$		$\lambda = 0.1$										
p	α	MCD	Adj MCD	Kurtosis	OGK	COM	RMDv1	RMDv2	RMDv3	RMDv4	RMDv5	RMDv6
5	0.1	0.0831	0.0669	0.2094	0.1955	0.0636	0.1378	0.1352	0.1368	0.1234	0.1222	0.1273
	0.2	0.1090	0.0935	0.2544	0.1602	0.0442	0.0693	0.0690	0.0687	0.0629	0.0609	0.0647
	0.3	0.2158	0.1904	0.1578	0.2988	0.0365	0.0330	0.0327	0.0331	0.0294	0.0295	0.0314
10	0.1	0.1272	0.1091	0.3392	0.2118	0.0737	0.1874	0.1851	0.1839	0.1316	0.1285	0.1387
	0.2	0.2362	0.2148	0.4324	0.2050	0.0603	0.1033	0.1023	0.1013	0.0776	0.0761	0.0818
	0.3	0.3045	0.2798	0.2388	0.4548	0.0346	0.0321	0.0320	0.0330	0.0253	0.0250	0.0277
30	0.1	0.1785	0.1533	0.3069	0.2618	0.0702	0.3474	0.3449	0.3457	0.1545	0.1530	0.1572
	0.2	0.2129	0.1841	0.4603	0.2327	0.0529	0.1384	0.1382	0.1379	0.0647	0.0641	0.0659
	0.3	0.2611	0.2282	0.1588	0.7518	0.0391	0.0387	0.0386	0.0387	0.0258	0.0258	0.0264
50	0.1	0.1835	0.1567	0.1968	0.2834	0.0643	0.3222	0.3212	0.3233	0.1381	0.1372	0.1395
	0.2	0.2118	0.1824	0.3996	0.2598	0.0600	0.1305	0.1306	0.1316	0.0649	0.0648	0.0653
	0.3	0.2589	0.2246	0.1458	0.8799	0.0390	0.0408	0.0405	0.0409	0.0236	0.0236	0.0238
$\delta = 10$		$\lambda = 1$										
p	α	MCD	Adj MCD	Kurtosis	OGK	COM	RMDv1	RMDv2	RMDv3	RMDv4	RMDv5	RMDv6
5	0.1	0.0843	0.0660	0.1778	0.1854	0.0636	0.1368	0.1353	0.1361	0.1221	0.1213	0.1264
	0.2	0.0454	0.0343	0.1855	0.1658	0.0506	0.0774	0.0779	0.0788	0.0670	0.0666	0.0684
	0.3	0.0964	0.0818	0.1508	0.1287	0.0264	0.0287	0.0276	0.0274	0.0269	0.0269	0.0272
10	0.1	0.1151	0.0973	0.2771	0.2101	0.0716	0.1870	0.1859	0.1872	0.1325	0.1295	0.1403
	0.2	0.0892	0.0768	0.2694	0.1759	0.0531	0.0896	0.0872	0.0887	0.0641	0.0635	0.0682
	0.3	0.1494	0.1337	0.2122	0.1437	0.0328	0.0371	0.0360	0.0360	0.0305	0.0300	0.0318
30	0.1	0.1027	0.0803	0.1994	0.2635	0.0710	0.3423	0.3417	0.3420	0.1523	0.1509	0.1552
	0.2	0.1451	0.1233	0.1937	0.2297	0.0551	0.1453	0.1453	0.1455	0.0685	0.0679	0.0699
	0.3	0.1501	0.1279	0.1633	0.1997	0.0409	0.0390	0.0388	0.0389	0.0244	0.0243	0.0250
50	0.1	0.1111	0.0885	0.2218	0.2853	0.0680	0.3711	0.3687	0.3690	0.1408	0.1397	0.1432
	0.2	0.1578	0.1340	0.3384	0.2537	0.0578	0.1173	0.1175	0.1170	0.0682	0.0680	0.0687
	0.3	0.1546	0.1329	0.1106	0.4323	0.0431	0.0433	0.0435	0.0433	0.0226	0.0226	0.0230

B.3 Multivariate Exponential distribution

Table B.14 shows the false positive rates (FPR) when there is no contamination. Table B.15 shows the true positive rates (TPR) and Table B.16 the FPR, for each method, corresponding to the simulations with multivariate Exponential distribution, for contamination levels $\alpha = 0.1, 0.2, 0.3$, dimension $p = 5, 10, 30, 50$ and distance of the outliers $\delta = 5$ and 10.

Table B.14: False positive rates with Exponential distribution, $\alpha = 0$.

p	MCD	Adj MCD	Kurtosis	OGK	COM	RMDv1	RMDv2	RMDv3	RMDv4	RMDv5	RMDv6
5	0.15070	0.13310	0.29900	0.26540	0.21340	0.16060	0.15850	0.16090	0.11820	0.11580	0.11280
10	0.18710	0.16890	0.36000	0.26580	0.38930	0.13638	0.15970	0.15910	0.13454	0.13920	0.13609
30	0.14292	0.11952	0.13558	0.27938	0.44708	0.13324	0.13266	0.13246	0.18204	0.18092	0.18464
50	0.13280	0.11900	0.12880	0.27950	0.42860	0.11788	0.11787	0.11787	0.11778	0.11777	0.11778

Table B.15: True positive rates with Exponential distribution.

$\delta = 5$												
p	α	MCD	Adj MCD	Kurtosis	OGK	COM	RMDv1	RMDv2	RMDv3	RMDv4	RMDv5	RMDv6
5	0.1	1	1	1	1	1	1	1	1	1	1	1
	0.2	0.8578	0.8486	0.9602	0.9654	0.9975	1	1	1	0.9993	0.9993	0.9993
	0.3	0.1955	0.1664	0.8336	0.5792	0.8735	0.8935	0.8947	0.8938	0.8740	0.8698	0.8755
10	0.1	0.9983	0.9976	0.9974	0.9983	0.9975	0.9993	0.9993	0.9993	0.9995	0.9995	0.9999
	0.2	0.9671	0.9649	0.9893	0.9984	0.9923	0.9992	0.9992	0.9992	0.9984	0.9984	0.9986
	0.3	0.7821	0.7792	0.9847	0.9887	0.9832	0.9957	0.9958	0.9957	0.9951	0.9951	0.9958
30	0.1	1	1	1	1	1	1	1	1	1	1	1
	0.2	0.9985	0.9985	0.9987	1	1	1	1	1	1	1	1
	0.3	0.7829	0.7550	1	1	1	1	1	1	1	1	1
50	0.1	1	1	1	1	1	1	1	1	1	1	1
	0.2	0.8916	0.8511	1	1	1	1	1	1	1	1	1
	0.3	0.5916	0.5715	1	1	1	1	1	1	1	1	1
$\delta = 10$												
p	α	MCD	Adj MCD	Kurtosis	OGK	COM	RMDv1	RMDv2	RMDv3	RMDv4	RMDv5	RMDv6
5	0.1	0.9991	0.9970	1	1	0.9991	0.9991	0.9991	0.9991	1	1	1
	0.2	0.9770	0.9733	0.9991	0.9991	0.9927	0.9961	0.9961	0.9961	0.9952	0.9952	0.9997
	0.3	0.8185	0.7952	0.9900	0.9966	0.9853	0.9916	0.9919	0.9919	0.9906	0.9903	0.9911
10	0.1	1	1	1	1	1	1	1	1	1	1	1
	0.2	0.9756	0.9756	1	1	1	1	1	1	1	1	1
	0.3	0.7905	0.7902	1	1	1	1	1	1	1	1	1
30	0.1	1	1	1	1	1	1	1	1	1	1	1
	0.2	0.9987	0.9987	1	1	1	1	1	1	1	1	1
	0.3	0.7719	0.7601	1	1	1	1	1	1	1	1	1
50	0.1	1	1	1	1	1	1	1	1	1	1	1
	0.2	0.9711	0.9614	1	1	1	1	1	1	1	1	1
	0.3	0.6711	0.6662	1	1	1	1	1	1	1	1	1

Table B.16: False positive rates with Exponential distribution.

$\delta = 5$												
p	α	MCD	Adj MCD	Kurtosis	OGK	COM	RMDv1	RMDv2	RMDv3	RMDv4	RMDv5	RMDv6
5	0.1	0.0339	0.0198	0.0341	0.0636	0.0033	0.0084	0.0073	0.0069	0.0075	0.0071	0.0077
	0.2	0.0110	0.0055	0.0347	0.0507	0.0017	0.0014	0.0012	0.0012	0.0010	0.0009	0.0012
	0.3	0.0374	0.0260	0.0379	0.0383	0.0001	0	0	0	0	0	0
10	0.1	0.1194	0.1018	0.3285	0.2299	0.0683	0.2380	0.2362	0.2372	0.2302	0.2258	0.2442
	0.2	0.0315	0.0235	0.2634	0.1838	0.0466	0.1176	0.1183	0.1182	0.1338	0.1302	0.1479
	0.3	0.0021	0.0016	0.1890	0.1445	0.0258	0.0575	0.0576	0.0576	0.0722	0.0681	0.0803
30	0.1	0.0887	0.0639	0.1276	0.2371	0.0335	0.4021	0.4020	0.4015	0.2726	0.2697	0.2781
	0.2	0.0239	0.0099	0.1390	0.2022	0.0202	0.1751	0.1750	0.1754	0.1531	0.1504	0.1583
	0.3	0.0001	0	0.1196	0.1734	0.0111	0.0525	0.0528	0.0524	0.0551	0.0541	0.0580
50	0.1	0.0796	0.0541	0.2255	0.2429	0.0401	0.4072	0.4070	0.4076	0.2168	0.2146	0.2213
	0.2	0.0224	0.0089	0.2350	0.2065	0.0431	0.1354	0.1360	0.1355	0.1342	0.1325	0.1371
	0.3	0	0	0.2295	0.1732	0.0380	0.0426	0.0427	0.0430	0.0320	0.0314	0.0328
$\delta = 10$												
p	α	MCD	Adj MCD	Kurtosis	OGK	COM	RMDv1	RMDv2	RMDv3	RMDv4	RMDv5	RMDv6
5	0.1	0.0842	0.0679	0.2846	0.2198	0.0799	0.1567	0.1572	0.1557	0.1659	0.1631	0.1708
	0.2	0.0271	0.0176	0.2408	0.1824	0.0551	0.0743	0.0752	0.0754	0.1015	0.0994	0.1039
	0.3	0.0011	0.0006	0.1969	0.1451	0.0288	0.0302	0.0299	0.0289	0.0460	0.0443	0.0470
10	0.1	0.1148	0.0961	0.3201	0.2284	0.0620	0.2103	0.2121	0.2131	0.2015	0.1963	0.2161
	0.2	0.0357	0.0261	0.2647	0.1841	0.0404	0.0895	0.0906	0.0890	0.0942	0.0912	0.1032
	0.3	0.0028	0.0022	0.1873	0.1474	0.0240	0.0298	0.0293	0.0309	0.0411	0.0388	0.0463
30	0.1	0.0863	0.0621	0.1343	0.2416	0.0334	0.3536	0.3530	0.3536	0.2394	0.2371	0.2450
	0.2	0.0253	0.0116	0.1179	0.2043	0.0213	0.1053	0.1056	0.1058	0.0913	0.0899	0.0944
	0.3	0	0	0.1240	0.1662	0.0108	0.0127	0.0128	0.0130	0.0146	0.0144	0.0156
50	0.1	0.0814	0.0557	0.1262	0.2436	0.0405	0.3465	0.3463	0.3475	0.1063	0.1045	0.1098
	0.2	0.0256	0.0185	0.1235	0.2121	0.0316	0.1002	0.1007	0.1010	0.0895	0.0881	0.0902
	0.3	0	0	0.1238	0.1762	0.0257	0.0054	0.0054	0.0052	0.0091	0.0088	0.0095

APPENDIX C

Figures from Chapter 3

The following are figures corresponding to the real dataset example from Chapter 3.

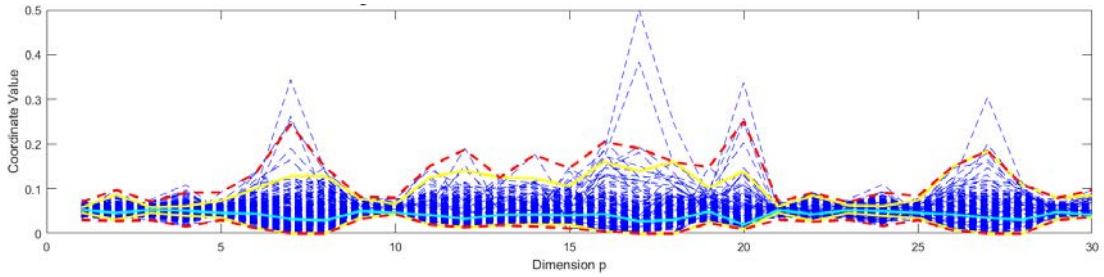

Figure C.1: Standardized data with the "multivariate boxplot".

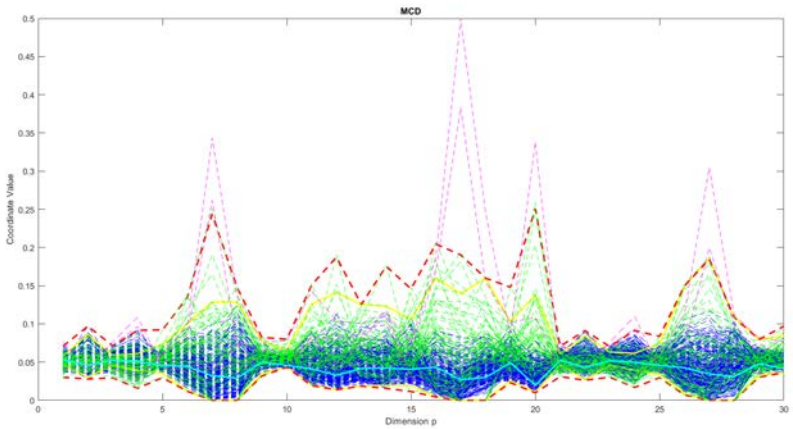

Figure C.2: Detected outliers by MCD.

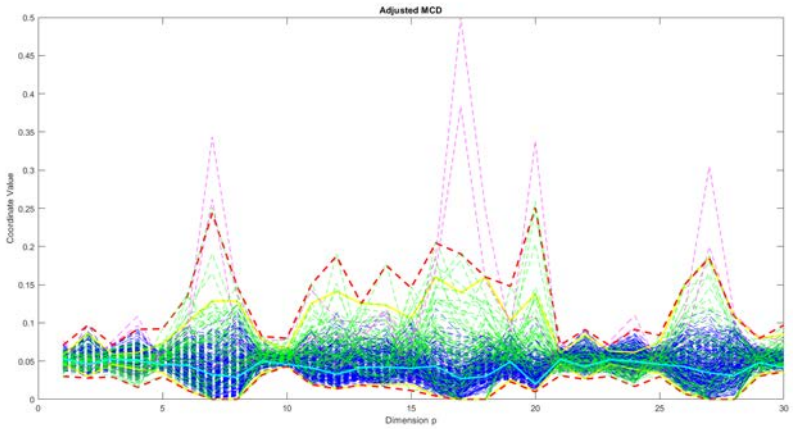

Figure C.3: Detected outliers by Adjusted MCD.

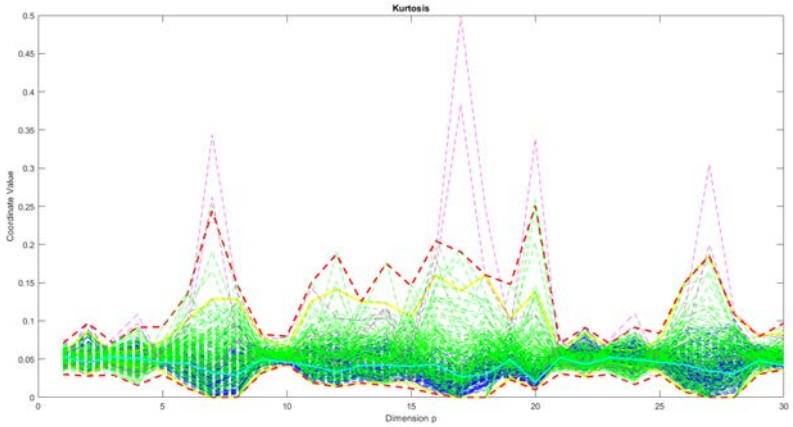

Figure C.4: Detected outliers by Kurtosis.

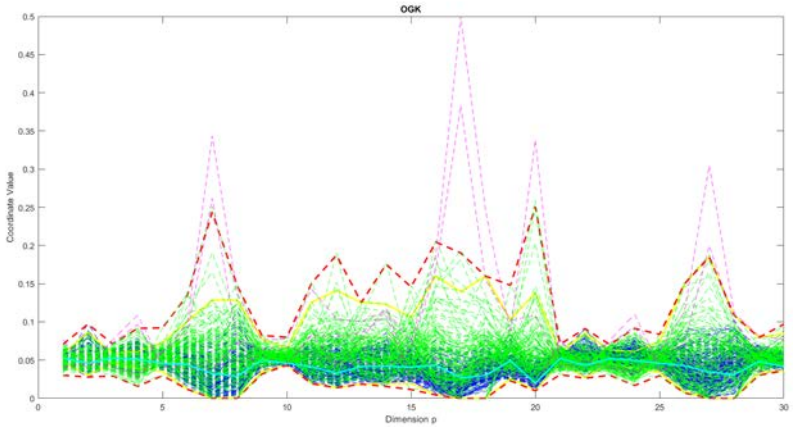

Figure C.5: Detected outliers by OGK.

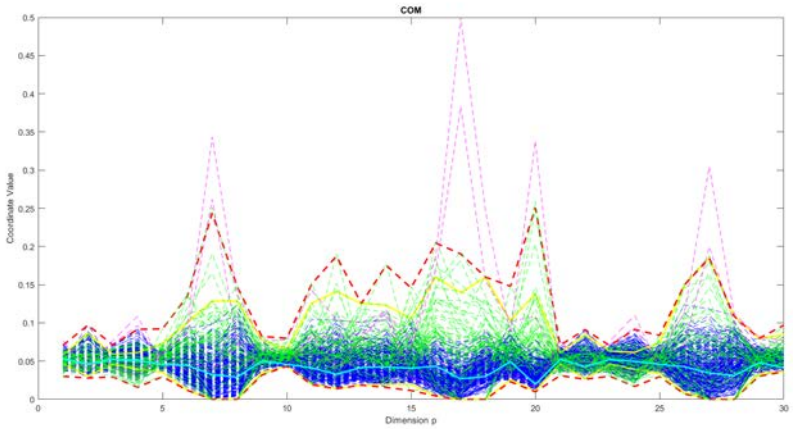

Figure C.6: Detected outliers by COM.

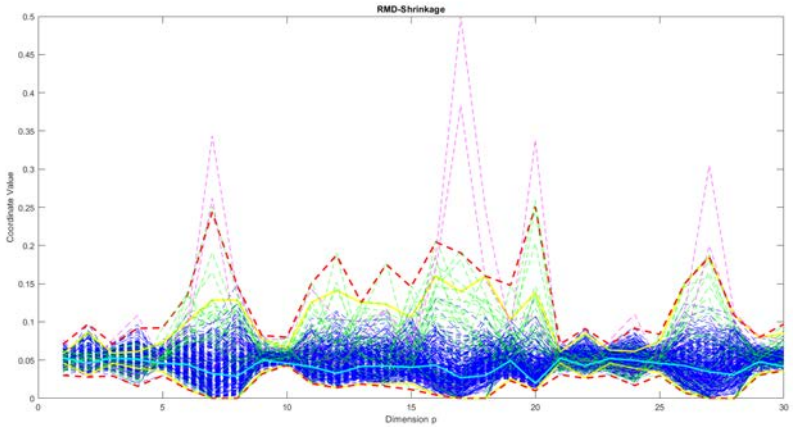

Figure C.7: Detected outliers by RMD-S.

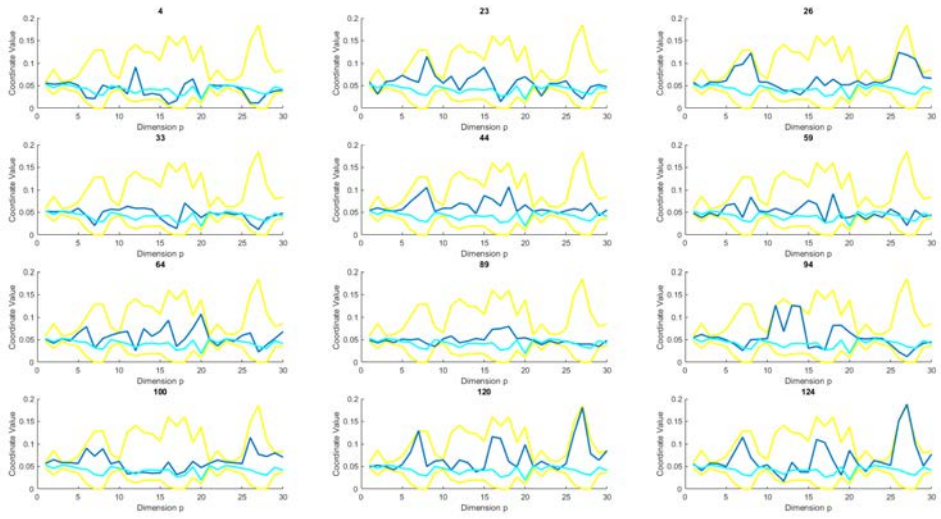

Figure C.8: MCD detected outliers that belong to the 50% of the most central data.

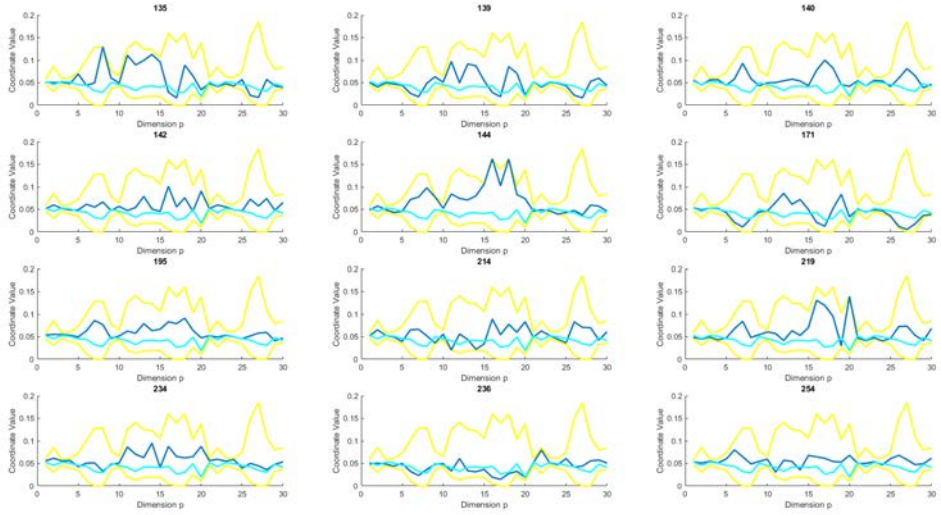

Figure C.9: MCD detected outliers that belong to the 50% of the most central data.

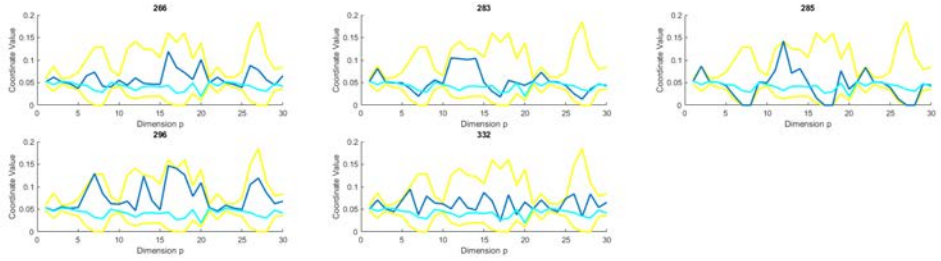

Figure C.10: MCD detected outliers that belong to the 50% of the most central data.

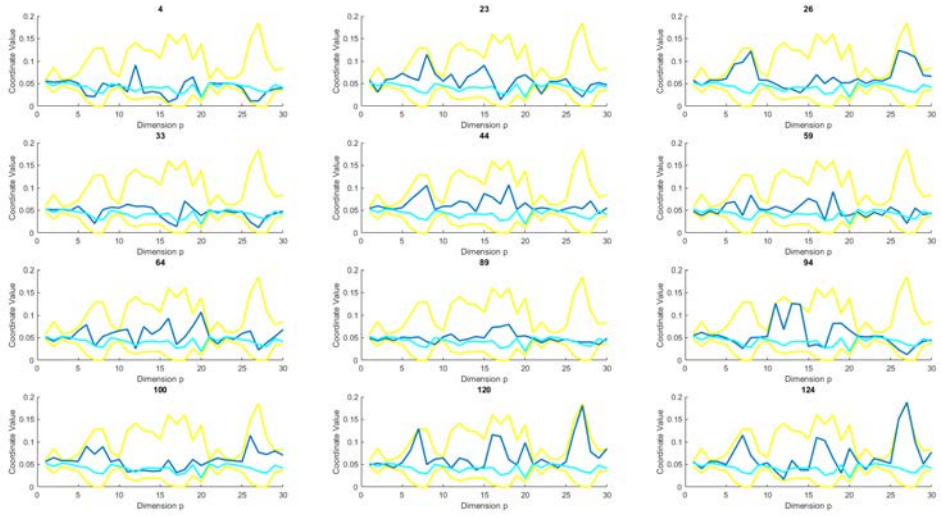

Figure C.11: Adjusted MCD detected outliers that belong to the 50% of the most central data.

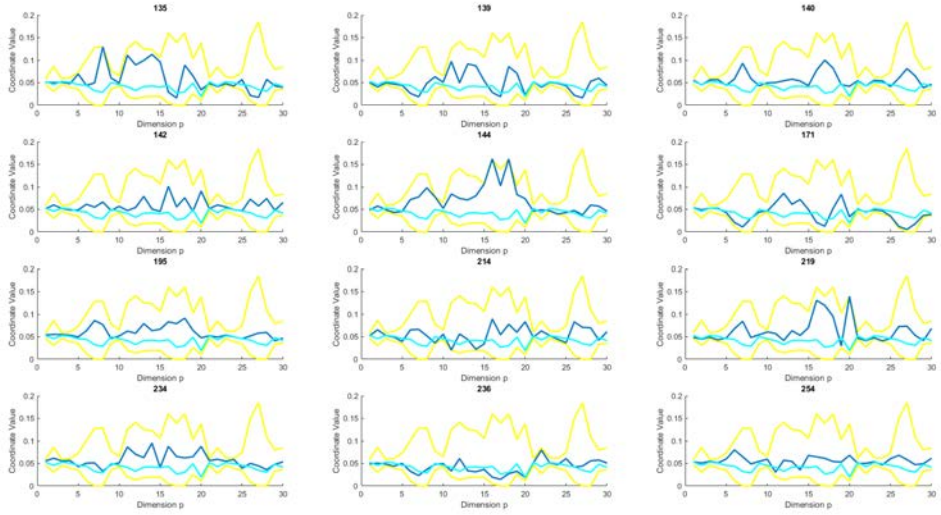

Figure C.12: Adjusted MCD detected outliers that belong to the 50% of the most central data.

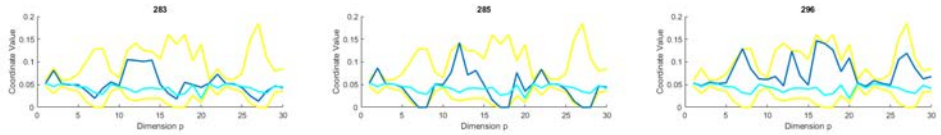

Figure C.13: Adjusted MCD detected outliers that belong to the 50% of the most central data.

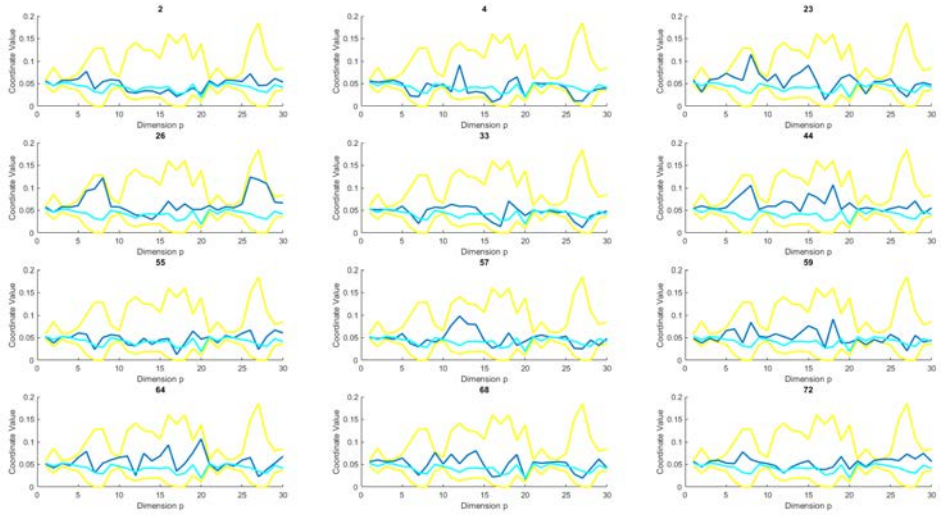

Figure C.14: Kurtosis detected outliers that belong to the 50% of the most central data.

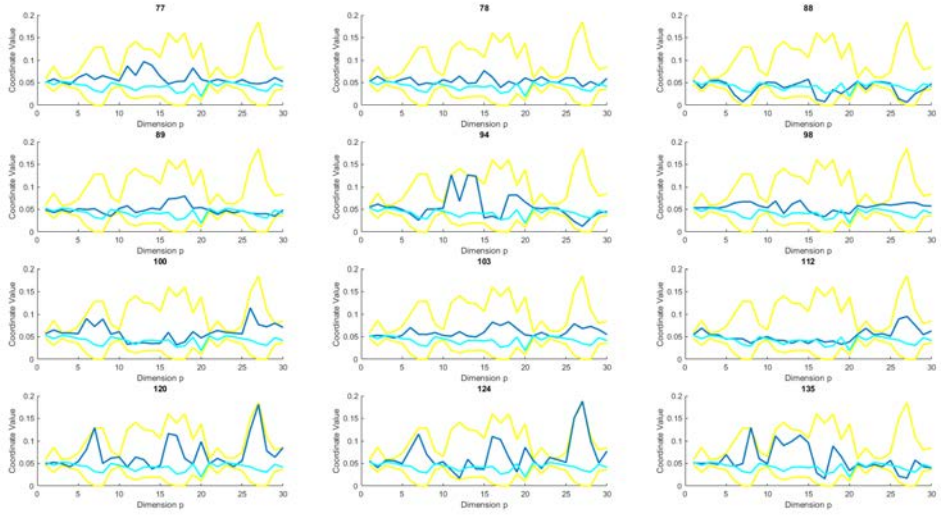

Figure C.15: Kurtosis detected outliers that belong to the 50% of the most central data.

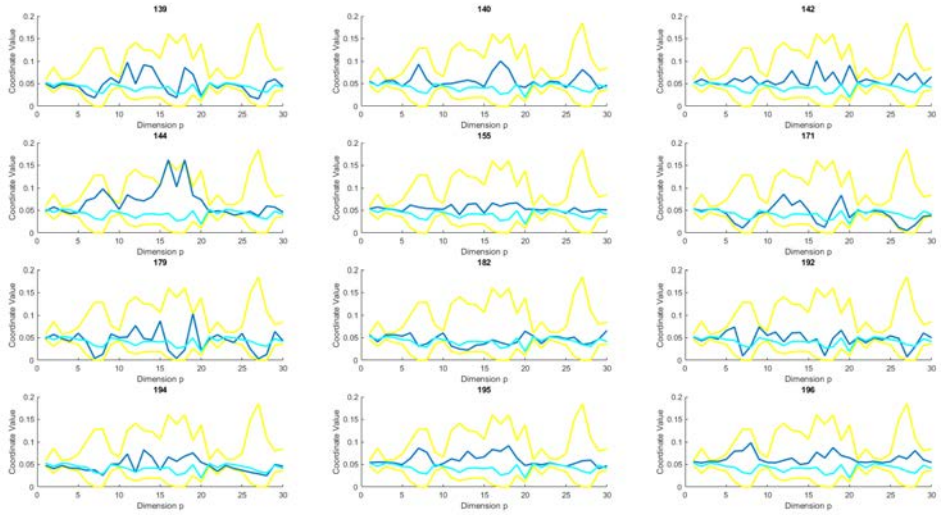

Figure C.16: Kurtosis detected outliers that belong to the 50% of the most central data.

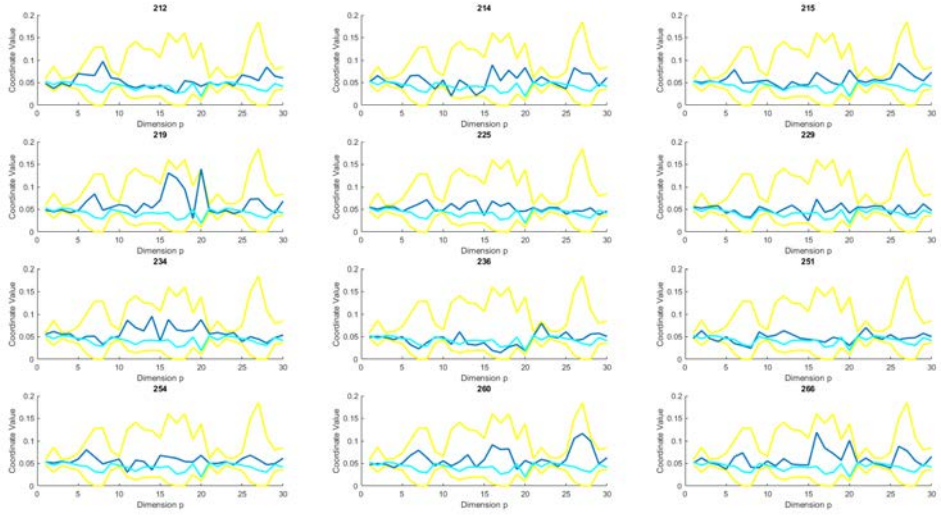

Figure C.17: Kurtosis detected outliers that belong to the 50% of the most central data.

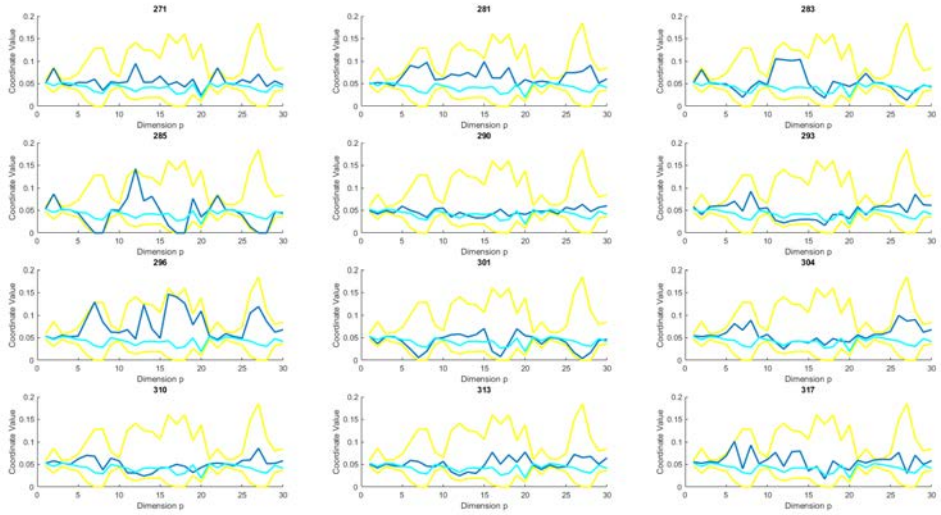

Figure C.18: Kurtosis detected outliers that belong to the 50% of the most central data.

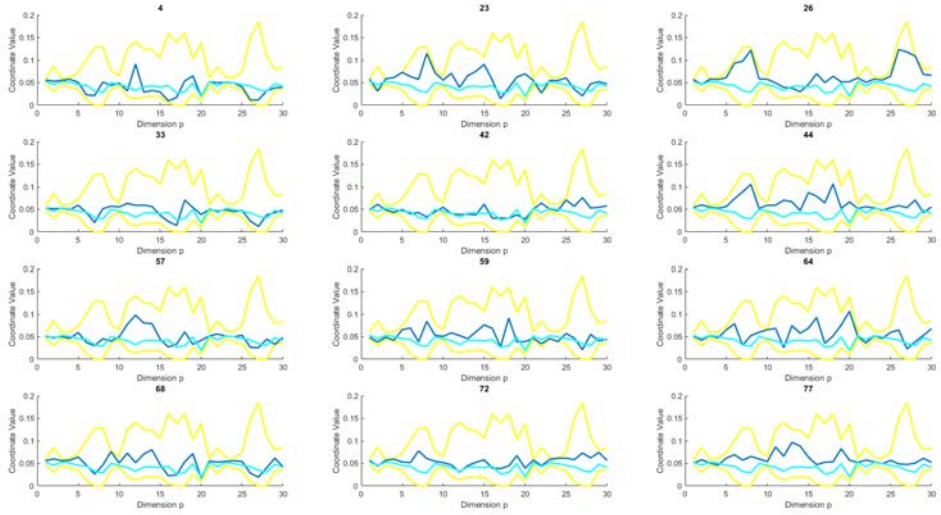

Figure C.19: OGK detected outliers that belong to the 50% of the most central data.

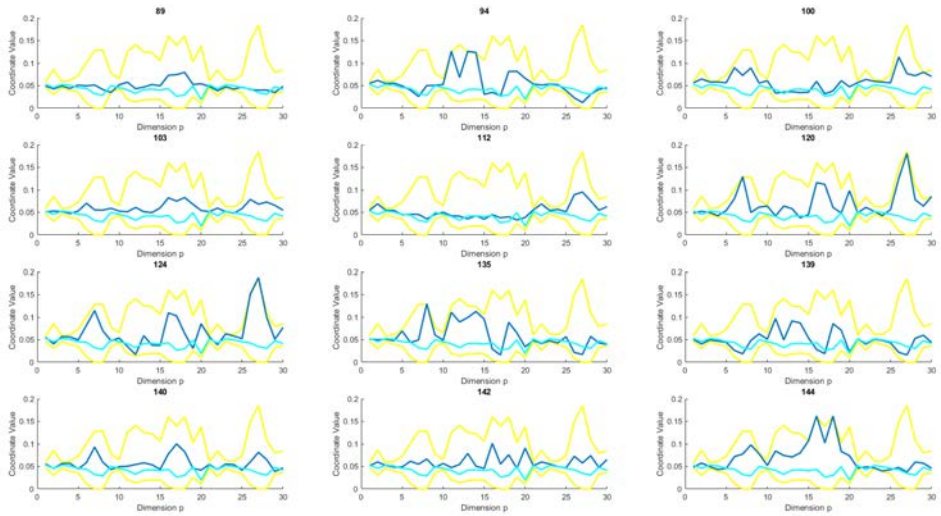

Figure C.20: OGK detected outliers that belong to the 50% of the most central data.

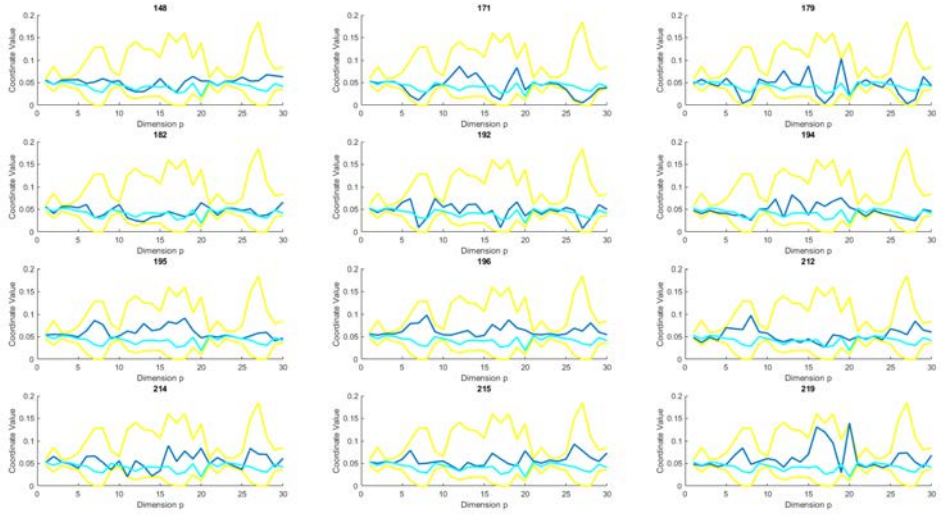

Figure C.21: OGK detected outliers that belong to the 50% of the most central data.

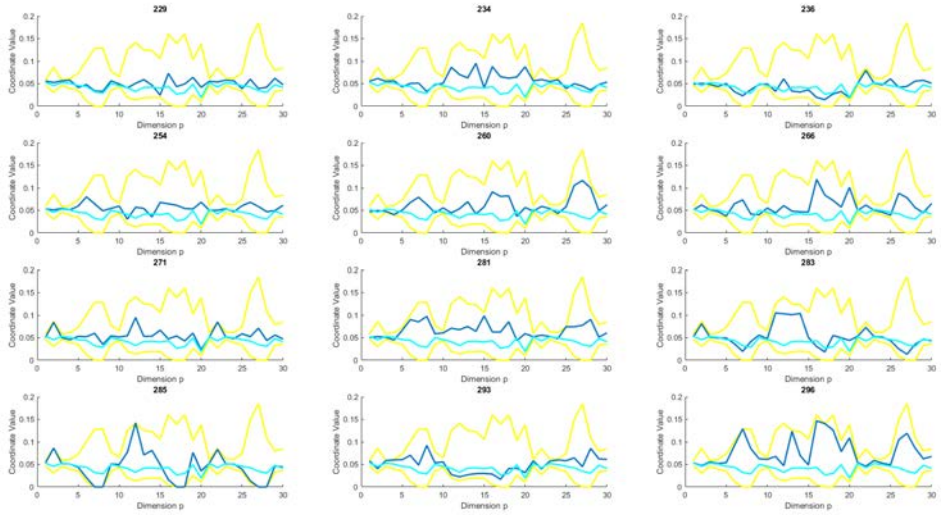

Figure C.22: OGK detected outliers that belong to the 50% of the most central data.

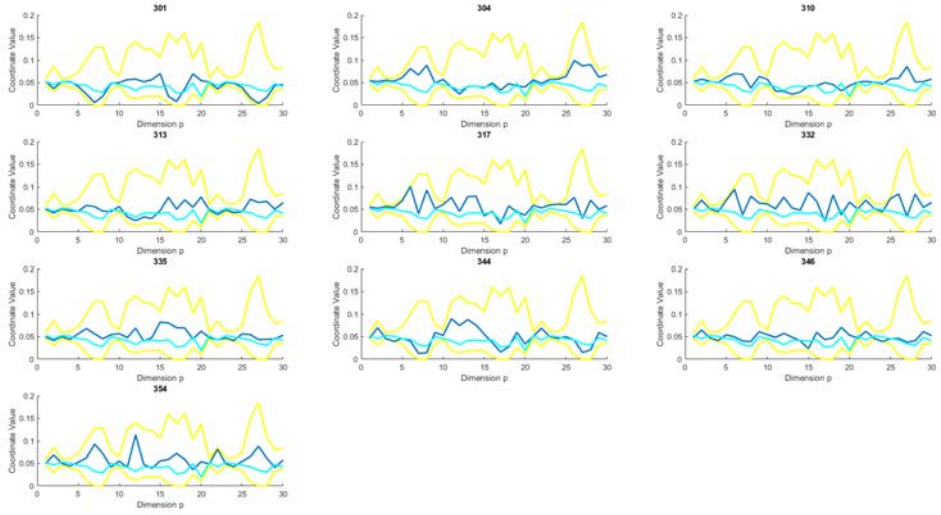

Figure C.23: OGK detected outliers that belong to the 50% of the most central data.

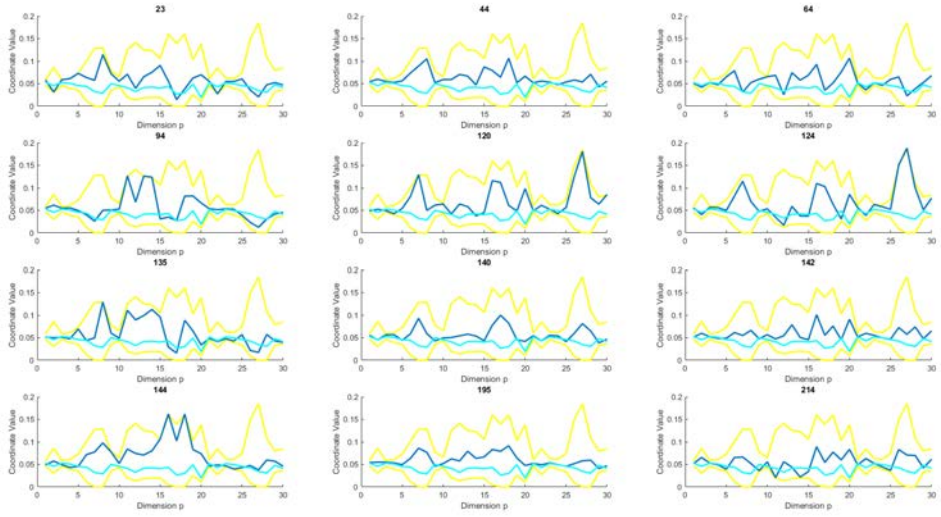

Figure C.24: Comedian detected outliers that belong to the 50% of the most central data.

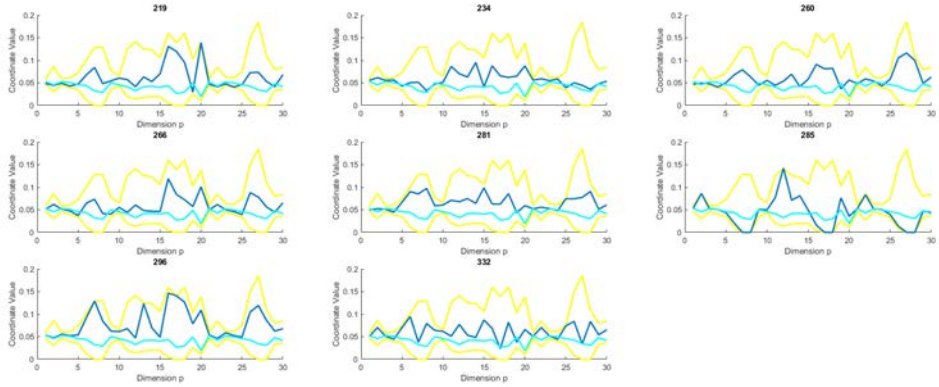

Figure C.25: Comedian detected outliers that belong to the 50% of the most central data.

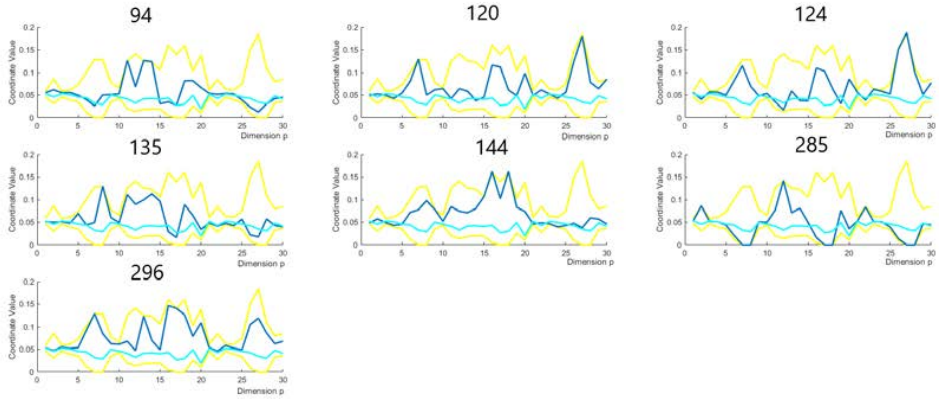

Figure C.26: RMD-S detected outliers that belong to the 50% of the most central data.

APPENDIX D

Tables from Chapter 5

Tables D.1 - D.4 show the numerical results from Chapter 5, in simulation scheme [NEO]. For each method, the maximum (across λ and k) MSE and Bias for both $\hat{\boldsymbol{\beta}}$ and $\hat{\alpha}$ for each combination of the dimension p and the contamination level δ, is showed. In bold letter are the lowest error and in italic letter are the highest error after OLS.

Table D.1: MMMSE and MMBias of $\hat{\boldsymbol{\beta}}$ and $\hat{\alpha}$, for $p = 5$ and $\delta = 10\%$.

Method	MSE($\hat{\boldsymbol{\beta}}$)	MSE($\hat{\alpha}$)	BIAS($\hat{\boldsymbol{\beta}}$)	BIAS($\hat{\alpha}$)
OLS	2.9065	5.5593	2.7004	5.3280
SR	**0.0230**	**0.0351**	0.0093	**0.0168**
LTS	*0.1116*	*0.0688*	*0.0832*	0.0275
S	0.0249	0.0512	**0.0083**	*0.0361*
REWLSE	0.0919	0.0474	0.0493	0.0260
MM	0.1033	0.0441	0.0785	0.0235

Table D.2: MMMSE and MMBias of $\hat{\boldsymbol{\beta}}$ and $\hat{\alpha}$, for $p = 5$ and $\delta = 20\%$.

Method	MSE($\hat{\boldsymbol{\beta}}$)	MSE($\hat{\alpha}$)	BIAS($\hat{\boldsymbol{\beta}}$)	BIAS($\hat{\alpha}$)
OLS	3.7360	29.9723	3.6101	29.4112
SR	**0.0470**	0.1720	**0.0287**	0.1075
LTS	0.8779	**0.1508**	0.3028	**0.0947**
S	*1.3853*	*5.4441*	*0.6577*	*3.8112*
REWLSE	0.1422	0.2556	0.1018	0.2124
MM	0.1688	0.3120	0.1478	0.2954

Table D.3: MMMSE and MMBias of $\hat{\boldsymbol{\beta}}$ and $\hat{\alpha}$, for $p = 30$ and $\delta = 10\%$.

Method	MSE($\hat{\boldsymbol{\beta}}$)	MSE($\hat{\alpha}$)	BIAS($\hat{\boldsymbol{\beta}}$)	BIAS($\hat{\alpha}$)
OLS	0.1995	6.7748	0.0610	6.7250
SR	**0.0033**	**0.0101**	**0.0009**	**0.0030**
LTS	0.0139	0.0145	0.0102	0.0060
S	*0.1079*	*2.9888*	*0.0584*	*2,9439*
REWLSE	0.0077	0.0165	0.0070	0.0080
MM	0.0120	0.0134	0.0101	0.0116

Table D.4: MMMSE and MMBias of $\hat{\boldsymbol{\beta}}$ and $\hat{\alpha}$, for $p = 30$ and $\delta = 20\%$.

Method	MSE($\hat{\boldsymbol{\beta}}$)	MSE($\hat{\alpha}$)	BIAS($\hat{\boldsymbol{\beta}}$)	BIAS($\hat{\alpha}$)
OLS	0.2317	25.5388	0.0639	25.3395
SR	**0.0044**	**0.0596**	**0.0011**	**0.0554**
LTS	0.0450	0.3952	0.0400	0.3677
S	*0.1710*	*15.0446*	*0.0635*	*14.8378*
REWLSE	0.0120	0.0980	0.0017	0.0930
MM	0.0356	0.1994	0.0262	0.1860

Bibliography

C. Agostinelli and M. Romanazzi. Local depth. *Journal of Statistical Planning and Inference*, 141(2):817–830, 2011.

J. Agulló, C. Croux, and S. Van Aelst. The multivariate least-trimmed squares estimator. *Journal of Multivariate Analysis*, 99(3):311–338, 2008.

D. Arribas-Bel, J. E. Patino, and J. C. Duque. Remote sensing-based measurement of Living Environment Deprivation: Improving classical approaches with machine learning. *PLOS ONE*, 12(5):e0176684, 2017.

D. Ballabio. A MATLAB toolbox for Principal Component Analysis and unsupervised exploration of data structure. *Chemometrics and Intelligent Laboratory Systems*, 149:1–9, 2015.

S. D. Bay. The UCI KDD Archive [http://kdd. ics. uci. edu]. Irvine, CA: University of California. *Department of Information and Computer Science*, 404:405, 1999.

C. Becker and U. Gather. The masking breakdown point of multivariate outlier identification rules. *Journal of the American Statistical Association*, 94(447):947–955, 1999.

C. Becker, R. Fried, S. Kuhnt, and U. Gather. *Robustness and complex data structures : festschrift in honour of Ursula Gather*. Springer Science & Business Media, 2014.

A. Bose. Estimating the asymptotic dispersion of theL 1 median. *Annals of the Institute of Statistical Mathematics*, 47(2):267–271, 1995.

A. Bose and P. Chaudhuri. On the dispersion of multivariate median. *Annals of the Institute of Statistical Mathematics*, 45(3):541–550, 1993.

J. Brettschneider, F. Collin, B. M. Bolstad, and T. P. Speed. Quality assessment for short oligonucleotide microarray data. *Technometrics*, 50(3):241–264, 2008.

B. M. Brown. Statistical uses of the spatial median. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 25–30, 1983.

K. K. Budde. A matlab toolbox for fmri data analysis: Detection, estimation and brain connectivity, 2012.

A. Cerioli, M. Riani, and A. C. Atkinson. Controlling the size of multivariate outlier tests with the MCD estimator of scatter. *Statistics and Computing*, 19(3):341–353, 2009.

S. X. Chen, Y.-L. Qin, and Others. A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics*, 38(2):808–835, 2010.

Y. Chen, X. Dang, H. Peng, and H. L. Bart. Outlier detection with the kernelized spatial depth function. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):288–305, 2008.

Y. Chen, A. Wiesel, and A. O. Hero. Robust shrinkage estimation of high-dimensional covariance matrices. *IEEE Transactions on Signal Processing*, 59 (9):4097–4107, 2011.

H. C. Choi, H. P. Edwards, C. H. Sweatman, and V. Obolonkin. Multivariate outlier detection of dairy herd testing data. *ANZIAM Journal*, 57:38–53, 2016.

J. T. Chu. On the distribution of the sample median. *The Annals of Mathematical Statistics*, pages 112–116, 1955.

R. Couillet and M. McKay. Large dimensional analysis and optimization of robust shrinkage covariance matrix estimators. *Journal of Multivariate Analysis*, 131: 99–120, 2014.

C. Croux, P. J. Rousseeuw, and O. Hössjer. Generalized S-Estimators. *Journal of the American Statistical Association*, 89(428):1271, 1994.

C. Croux, S. Van Aelst, and C. Dehon. Bounded influence regression using high breakdown scatter matrices. *Annals of the Institute of Statistical Mathematics*, 55(2):265–285, 2003.

J. P. De Grève and D. Vanbeveren. Close binary systems before and after mass transfer: A comparison of observations and theory. *Astrophysics and Space Science*, 68 (2):433–457, 1980.

V. DeMiguel, A. Martin-Utrera, and F. J. Nogales. Size matters: Optimal calibration of shrinkage estimators for portfolio selection. *Journal of Banking and Finance*, 37(8):3018–3034, 2013.

S. J. Devlin, R. Gnanadesikan, and J. R. Kettenring. Robust estimation of dispersion matrices and principal components. *Journal of the American Statistical Association*, 76(374):354–362, 1981.

Y. Dodge. An introduction to L1-norm based statistical data analysis. *Computational Statistics & Data Analysis*, 5(4):239–253, 1987.

D. L. Donoho and P. J. Huber. The notion of breakdown point. In *A festschrift for Erich L. Lehmann*, volume 157184. 1983.

F. Y. Edgeworth. On observations relating to several quantities. *Hermathena*, 6: 279–285, 1887.

M. Falk. On Mad and Comedians. *Annals of the Institute of Statistical Mathematics*, 49(4):615–644, 1997.

P. Filzmoser, R. G. Garrett, and C. Reimann. Multivariate outlier detection in exploration geochemistry. *Computers and Geosciences*, 31(5):579–587, 2005.

S. Gajjar, M. Kulahci, and A. Palazoglu. Selection of non-zero loadings in sparse principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 162:160–171, 2017.

C. F. Gauss. Bestimmung der genauigkeit der beobachtungen. *Astronomi*, 1:185–197, 1816.

D. Gervini and V. J. Yohai. A class of robust and fully efficient regression estimators. *The Annals of Statistics*, 30(2):583–616, 2002.

R. Gnanadesikan and J. R. Kettenring. Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data. *Biometrics*, 28(1):81–124, 1972.

C. Goutte and E. Gaussier. A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. In *European Conference on Information Retrieval*, pages 345–359, 2010.

J. C. Gower. Algorithm as 78: The mediancentre. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 23(3):466–470, 1974.

A. S. Hadi. Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 761–771, 1992.

P. Hall and A. H. Welsh. Limit theorems for the median deviation. *Annals of the Institute of Statistical Mathematics*, 37(1):27–36, 1985.

M. Hallin, D. Paindaveine, et al. Optimal tests for multivariate location based on interdirections and pseudo-mahalanobis ranks. *The Annals of Statistics*, 30(4): 1103–1133, 2002.

F. R. Hampel. Optimally bounding the gross-error-sensitivity and the influence of position in factor space. In *Proceedings of the ASA Statistical Computing Section, ASA, Washington, DC*, pages 59–64, 1978.

J. Hardin and D. M. Rocke. The distribution of robust distances. *Journal of Computational and Graphical Statistics*, 14(4):928–946, 2005.

D. M. Hawkins and D. J. Olive. Inconsistency of Resampling Algorithms for High-Breakdown Regression Estimators and a New Algorithm. *Journal of the American Statistical Association*, 97(457):136–148, 2002.

D. M. Hawkins, D. Bradu, and G. V. Kass. Location of Several Outliers in Multiple-Regression Data Using Elemental Sets. *Technometrics*, 26(3):197, 1984.

R. W. Hill. *Robust regression when there are outliers in the carriers*. PhD thesis, Harvard University, 1977.

I. Hoffmann, S. Serneels, P. Filzmoser, and C. Croux. Sparse partial robust M regression. *Chemometrics and Intelligent Laboratory Systems*, 149:50–59, 2015.

H. Hotelling et al. The generalization of student's ratio. *The Annals of Mathematical Statistics*, 2(3):360–378, 1931.

P. J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

P. J. Huber. Robust Regression: Asymptotics, Conjectures and Monte Carlo. *The Annals of Statistics*, 1(5):799–821, 1973.

P. J. Huber. Robust statistics. *New York John Wiley and Sons*, 1981.

M. Hubert. *Multivariate Nonparametric Methods with R. An Approach Based on Spatial Signs and Ranks by OJA, H.*, volume 67. 2011.

M. Hubert and M. Debruyne. Minimum covariance determinant. *Wiley interdisciplinary reviews: Computational statistics*, 2(1):36–43, 2010.

M. Hubert, P. J. Rousseeuw, and S. Van Aelst. High-breakdown robust multivariate methods. *Statistical Science*, pages 92–119, 2008.

R. M. Humphreys. Studies of luminous stars in nearby galaxies. I. Supergiants and O stars in the Milky Way. *The Astrophysical Journal Supplement Series*, 38:309, 1978.

A. Inselberg. *Parallel coordinates*. Springer, 2009.

A. Inselberg and B. Dimsdale. Parallel coordinates: a tool for visualizing multi-dimensional geometry. In *Proceedings of the 1st conference on Visualization'90*, pages 361–378. IEEE Computer Society Press, 1990.

L. A. Jaeckel. Estimating Regression Coefficients by Minimizing the Dispersion of the Residuals. *The Annals of Mathematical Statistics*, 43(5):1449–1458, 1972.

W. James and C. Stein. Estimation with Quadratic Loss. In *Breakthroughs in statistics*, pages 443–460. Springer, New York, NY, 1992.

I. Jolliffe. Principal Component Analysis. In *International Encyclopedia of Statistical Science*, pages 1094–1096. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

W. S. Krasker. Estimation in Linear Regression Models with Disparate Data Points. *Econometrica*, 48(6):1333, 1980.

W. S. Krasker and R. E. Welsch. Efficient Bounded-Influence Regression Estimation. *Journal of the American Statistical Association*, 77(379):595–604, 1982.

N. Lazar. *The statistical analysis of functional MRI data*. Springer Science & Business Media, 2008.

O. Ledoit and M. Wolf. Honey, I Shrunk the Sample Covariance Matrix. *Ssrn*, 2003a.

O. Ledoit and M. Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10:603–621, 2003b.

O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88:365–411, 2004.

A. M. Leroy and P. J. Rousseeuw. *Robust regression and outlier detection*. 1987.

C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49:764–766, 2013.

H.-D. Li, Q.-S. Xu, and Y.-Z. Liang. libPLS: An integrated library for partial least squares regression and linear discriminant analysis. *Chemometrics and Intelligent Laboratory Systems*, 176:34–43, 2018.

M. A. Lindquist. The statistical analysis of fMRI data. *Statistical Science*, pages 439–464, 2008.

R. Y. Liu et al. On a notion of data depth based on random simplices. *The Annals of Statistics*, 18(1):405–414, 1990.

H. P. Lopuhaa and P. J. Rousseeuw. Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics*, 19(1):229–248, 1991.

P. C. Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, 2:49–55, 1936.

C. L. Mallows. On some topics in robustness. *Unpublished memorandum, Bell Telephone Laboratories, Murray Hill, NJ*, 1975.

L. Marcano and W. Fermin. Comparación de métodos de detección de datos anómalos multivariantes mediante un estudio de simulación. *SABER. Revista Multidisciplinaria del Consejo de Investigación de la Universidad de Oriente*, 25 (2):192–201, 2013.

R. Maronna and S. Morgenthaler. Robust regression through robust covariances. *Communications in Statistics - Theory and Methods*, 15(4):1347–1365, 1986.

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

R. A. Maronna, R. D. Martin, and V. J. Yohai. *Robust statistics : theory and methods*. John Wiley & Sons, 2006.

M. Marozzi. Multivariate multidistance tests for high-dimensional low sample size case-control studies. *Statistics in medicine*, 34(9):1511–1526, 2015.

M. Marozzi. Multivariate tests based on interpoint distances with application to magnetic resonance imaging. *Statistical methods in medical research*, 25(6):2593–2610, 2016.

M. M. Monti. Statistical analysis of fMRI time-series: a critical review of the GLM approach. *Frontiers in human neuroscience*, 5(28), 2011.

M. Morozova, T. Elizarova, and T. Pleteneva. Discriminant analysis and mahalanobis distance (nir diffuse reflectance spectra) in the assessment of drug's batch-to-batch dispersion and quality threshold establishment. *European Scientific Journal, ESJ*, 9(27), 2013.

J. Möttönen, K. Nordhausen, and H. Oja. Asymptotic theory of the spatial median. pages 182–193. Institute of Mathematical Statistics, 2010.

D. Paindaveine and G. Van Bever. From depth to local depth: a focus on centrality. *Journal of the American Statistical Association*, 108(503):1105–1119, 2013.

D. Peña and F. J. Prieto. Multivariate outlier detection and robust covariance matrix estimation. *Technometrics*, 43(3):286–310, 2001.

D. Peña and F. J. Prieto. Combining Random and Specific Directions for Outlier Detection and Robust Estimation in High-Dimensional Multivariate Data. *Journal of Computational and Graphical Statistics*, 16(1):228–254, 2007.

D. Perrotta and F. Torti. Detecting price outliers in European trade data with the forward search. In *Data Analysis and Classification*, pages 415–423. Springer, 2010.

J.-B. Poline and M. Brett. The general linear model and fMRI: does love last forever? *Neuroimage*, 62(2):871–880, 2012.

D. M. Powers. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2 (1):37–63, 2011.

C. Reimann and P. Filzmoser. Normal and lognormal data distribution in geochemistry: death of a myth. Consequences for the statistical treatment of geochemical and environmental data. *Environmental geology*, 39(9):1001–1014, 2000.

C. Reimann, P. Filzmoser, and R. G. Garrett. Background and threshold: critical comparison of methods of determination. *Science of the Total Environment*, 346 (1):1–16, 2005.

M. Riani, A. Cerioli, and A. C. Atkinson. Fitting mixtures of regression lines with the forward search. *... : Advances in Data Mining, Search, ...*, 19:271, 2008.

M. Riani, D. Perrotta, and F. Torti. FSDA: A MATLAB toolbox for robust analysis and interactive data exploration. *Chemometrics and Intelligent Laboratory Systems*, 116:17–32, 2012.

E. Ronchetti and P. J. Rousseeuw. Change-of-variance sensitivities in regression analysis. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 68 (4):503–519, 1985.

P. J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79(388):871–880, 1984.

P. J. Rousseeuw. Multivariate Estimation with High Breakdown Point. In *Mathematical Statistics and Applications*, volume 8, pages 283–297. 1985.

P. J. Rousseeuw and C. Croux. Alternatives to the Median Absolute Deviation. *Journal of the American Statistical Association*, 88(424):1273, 1993.

P. J. Rousseeuw and K. V. Driessen. A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, 41(3):212–223, 1999.

P. J. Rousseeuw and B. C. Van Zomeren. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical association*, 85(411):633–639, 1990.

P. J. Rousseeuw and V. Yohai. Robust Regression by Means of S-Estimators. pages 256–272. Springer, New York, NY, 1984.

P. J. Rousseeuw, F. R. Hampel, E. M. Ronchetti, and W. A. Stahel. Robust statistics: the approach based on influence functions. *J. Wiley, New York*, 1986.

P. J. Rousseeuw, S. V. Aelst, K. Van Driessen, and J. Agulló. Robust Multivariate Regression. 2004.

D. Ruppert. Computing S Estimators for Regression and Multivariate Location/Dispersion. *Journal of Computational and Graphical Statistics*, 1(3):253, 1992.

T. A. Sajesh and M. R. Srinivasan. Outlier detection for high dimensional data using the Comedian approach. *Journal of Statistical Computation and Simulation*, 82 (5):745–757, 2012.

R. Serfling. A depth function and a scale curve based on spatial quantiles. In *Statistical Data Analysis Based on the L1-Norm and Related Methods*, pages 25–38. Springer, 2002.

A. F. Siegel. Robust Regression Using Repeated Medians. *Biometrika*, 69(1):242, 1982.

C. G. Small. A survey of multidimensional medians. *International Statistical Review/Revue Internationale de Statistique*, pages 263–277, 1990.

M. Sokolova, N. Japkowicz, and S. Szpakowicz. Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. In *Australasian Joint Conference on Artificial Intelligence*, pages 1015–1021. Springer, 2006.

A. Steland. Shrinkage for covariance estimation: asymptotics, confidence intervals, bounds and applications in sensor monitoring and finance. *Statistical Papers*, pages 1–22, 2018.

A. J. Stromberg, O. Hössjer, and D. M. Hawkins. The Least Trimmed Differences Regression Estimator and Alternatives. *Journal of the American Statistical Association*, 95(451):853–864, 2000.

Y. Sun and M. G. Genton. Functional boxplots. *Journal of Computational and Graphical Statistics*, 20(2):316–334, 2011.

G. Tarr, S. Müller, and N. C. Weber. Robust estimation of precision matrices under cellwise contamination. *Computational Statistics & Data Analysis*, 93:404–420, 2016.

M. Templ, P. Filzmoser, and C. Reimann. Cluster analysis applied to regional geochemical data: problems and possibilities. *Applied Geochemistry*, 23(8):2198–2213, 2008.

J. W. Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, volume 2, pages 523–531, 1975.

Y. Vardi and C.-H. Zhang. The multivariate L1-median and associated data depth. *Proceedings of the National Academy of Sciences*, 97(4):1423–1426, 2002.

J. A. Vargas N. Robust estimation in multivariate control charts for individual observations. *Journal of Quality Technology*, 35(4):367–376, 2003.

S. Verboven and M. Hubert. LIBRA: a MATLAB library for robust analysis. *Chemometrics and Intelligent Laboratory Systems*, 75(2):127–136, 2005.

T. D. Wager, M. C. Keller, S. C. Lacey, and J. Jonides. Increased sensitivity in neuroimaging analyses using robust regression. *Neuroimage*, 26(1):99–113, 2005.

H. Wang, J. Gu, S. Wang, and G. Saporta. Spatial partial least squares autoregression: Algorithm and applications. *Chemometrics and Intelligent Laboratory Systems*, 184:123–131, 2019.

E. J. Wegman. Hyperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Association*, 85(411):664–675, 1990.

K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(Feb):207–244, 2009.

S. Xiang, F. Nie, and C. Zhang. Learning a mahalanobis distance metric for data clustering and classification. *Pattern recognition*, 41(12):3600–3612, 2008.

E. P. Xing, M. I. Jordan, S. J. Russell, and A. Y. Ng. Distance metric learning with application to clustering with side-information. In *Advances in neural information processing systems*, pages 521–528, 2003.

V. J. Yohai. High Breakdown-Point and High Efficiency Robust Estimates for Regression. *The Annals of Statistics*, 15(2):642–656, 1987.

C. Yu and W. Yao. Robust linear regression: A review and comparison. *Communications in Statistics - Simulation and Computation*, 46(8):6261–6282, 2017.

Y. Zeng, G. Wang, E. Yang, G. Ji, C. L. Brinkmeyer-Langford, and J. J. Cai. Aberrant gene expression in humans. *PLoS Genet*, 11(1):e1004942, 2015.

Y. Zhang, D. Huang, M. Ji, and F. Xie. Image segmentation using pso and pcm with mahalanobis distance. *Expert Systems with Applications*, 38(7):9036–9040, 2011.

H. Zou, T. Hastie, and R. Tibshirani. Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006.