2023-05-01

Outlier Detection In Multivariate And High-Dimensional Datasets

Yuanhong Wu
University of Texas at El Paso

OUTLIER DETECTION IN MULTIVARIATE AND HIGH-DIMENSIONAL DATASETS

Yuanhong Wu

Master's Program in Statistics and Data Science

APPROVED:

Michael Pokojovy, Ph.D., Chair

Abhijit Mandal, Ph.D.

Vladik Kreinovich, Ph.D.

Stephen L. Crites J, Ph.D.
Dean of the Graduate School

Dedication

To my

FATHER Diwang, MOTHER Jinzai, SISTER Yan

who were always there for me, even on the tough days

OUTLIER DETECTION IN MULTIVARIATE AND HIGH-DIMENSIONAL DATASETS

by

Yuanhong Wu, MS

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Department of Mathematical Sciences

THE UNIVERSITY OF TEXAS AT EL PASO

May 2023

Acknowledgements

First and foremost, I sincerely appreciate the support and guidance provided by my advisor, Dr. Pokojovy. He gave me a lot of assistance regarding my thesis research and consistently encouraged me to believe that completing the thesis is not a daunting challenge. I will forever be grateful to him.

Next, I would extend my gratitude to my friend Jiahao Xu, for providing help for me upon my arrival here. I arrived in the US in 2021 after leaving China. I was not familiar with anyone here. He shared some experiences of living in the United States and explained some guidelines and regulations to make me adapt to the environment faster. And he took me to visit many national parks and saw beautiful landscapes. His help is much more than the things I listed here. With his support, I gradually got used to the American lifestyle and school.

Then I also thank my roomates Roul and Sati. They are from Mexico and also students at UTEP. They are very nice and happy to help me all the time. We are like people living together in a family. Also my thanks to my classmates Kwabena, Noha and Dawa. They helped me a lot with classes and homework. If I didn't understand what the professors were saying, they would explain it to me patiently.

Also I deeply appreciate the absolute love and encouragement provided by my parents and my sister.

Abstract

Accurate detection of outliers is crucial in the field of statistical analysis. Using classical statistical models without considering the presence of outliers in the data can lead to misleading outcomes. There exist a myriad of procedures to detect outliers in statistics. We concentrate on the statistical techniques that can robustly identify outliers in data sets. To this end, we pursue two aims. First, we give an extensive overview of robust statistical methods which are still popular in recent years for outlier detection. We provide the definitions, algorithms and also discuss some important properties for these methods. Second, two real examples are presented to make a comparison between several techniques. Three prevalent methods are selected to illustrate their practical use for outlier detection in both low-dimensional and high-dimensional data.

Table of Contents

Page

Acknowledgements v

Abstract vi

Table of Contents vii

List of Figures viii

Chapter

1 Introduction 1

1.1 Overview of Outlier Detection 1

1.2 Masking and Swamping Effects 4

1.3 Research Objectives 6

2 Robust Statistical Methods for Location and Scatter 8

2.1 Distance-Based Methods 8

2.2 Projection-Based Methods 14

2.3 Other Methods 20

2.4 Measures of Robustness 21

3 Application for a Multivariate Dataset 29

3.1 Data Set *seeds* 29

3.2 Comparisons and Discussion 31

4 Application for a High-Dimensional Dataset 35

4.1 Data Set *characterA* 35

4.2 Comparisons and Discussion 37

5 Conclusions 41

References 42

Vita 49

List of Figures

1.1 Mahalanobis distances versus index number for classical and robust estimators (left), and QQ plots of distances (right). 3

1.2 Masking effect. Mahalanobis distances for the original data (left) and the data deleting six outliers (right). 5

1.3 Swamping effect. Mahalanobis distances for the original data (left) and the data deleting two outliers (right). 6

3.1 Bulk Purity (left), True positive rate (TPR) (middle) and False positive rate (FPR) (right) plots for three methods 31

3.2 Logarithmic Mahalanobis distance plots for $n = 47$ (top) and $n = 48$ (bottom) 32

3.3 Logarithmic Mahalanobis distance plots for $n = 54$ (top) and $n = 55$ (bottom) 32

3.4 Logarithmic Mahalanobis distance plots for $n = 65$ (top) and $n = 66$ (bottom) 33

4.1 X trajectory (left) and Y trajectory (right) of the pen when writing letter a . . 36

4.2 Bulk Purity (left), True positive rate (TPR) (middle) and False positive rate (FPR) (right) plots for three methods 37

4.3 Mahalanobis distance plots for $n = 18$ (top) and $n = 19$ (bottom) 39

4.4 Mahalanobis distance plots for $n = 44$ (top) and $n = 45$ (bottom) 39

4.5 Mahalanobis distance plots for $n = 45$ (top) and $n = 46$ (bottom) 40

Chapter 1

Introduction

1.1 Overview of Outlier Detection

In classical statistics, estimation procedures heavily rely on a number of assumptions. The most widely used model assumption is that the observed data are normally distributed. This assumption has existed in statistics for two centuries. Clearly, it has several reasonable reasons. Many natural phenomena and processes can be presented approximately by normal distribution. And it is also theoretically quite convenient because it allows one to derive explicit formulae for optimal statistical methods such as maximum likelihood and likelihood ratio tests, as well as the sampling distribution of inference quantities such as t-statistics (Maronna et al., 2019). But these explicit or implicit assumptions do not always hold. Since the middle of the 20th century, one has become increasingly aware that some of the most common statistical procedures are excessively sensitive to seemingly minor deviations from the assumptions (Ronchetti and Huber, 2009). It often happens in practice that some observations deviate from the bulk of the data. An assumed normal distributed model can present the large part of the data well, but some data points exhibit a different pattern or no pattern at all. Such data points are called outliers. when there are some outlying observations in the data, classical methods often give quite poor performance. Even the presence of only one outlier can significantly distort the outcomes of a classical statistical method which are optimal with the assumption of normality or linearity. Therefore, alternative robust procedures were developed in the past few decades aiming to deal with deviations from the model and contamination in the data.

The robust approach to classical statistics focuses on deriving methods that produce

reliable parameter estimates, associated tests and confidence intervals. For example, sample median is considered as a good robust alternative to the sample mean when we identify some outliers in the data. In high-dimensional statistics, a major problem is to obtain the robust estimation of multivariate location and scatter. The classical estimators of handling this are the empirical average and the empirical covariance matrix, but they are highly susceptible to even very few anomalous observations. Estimation of covariance matrices is essential in a number of areas of statistical analysis, including dimension reduction by principal component analysis (PCA), classification by linear or quadratic discriminant analysis (LDA and QDA), establishing independence and conditional independence relations in the context of graphical models, and setting confidence intervals on linear functions of the means of the components (Bickel and Levina, 2008).

Robust approaches reduce or remove the effect of outlying observations and allow the "good" data points to primarily determine the result. A standard method to delve into whether a multivariate data set forms a homogeneous group or contains abnormal observations is to calculate the Mahalanobis distances for the data, given by

$$\mathrm{MD}(\boldsymbol{x}_i) = \sqrt{(\boldsymbol{x}_i - \bar{\boldsymbol{x}})' \mathbf{S}^{-1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})}, \qquad i = 1, \ldots, n. \tag{1.1}$$

where $\boldsymbol{x}_i$ is the ith row of $\mathbf{X}_{n \times p}$ data matrix, $\bar{\boldsymbol{x}}$ the sample mean vector, $\mathbf{S}$ the sample covariance matrix of the data and $(\cdot)'$ the transpose of a matrix. Then the squared Mahalanobis distances approximately follow a chi-squared distribution with ν degree of freedom, where ν is the number of variables (Campbell, 1980). The usual cutoff value for Mahalanobis distances is $\sqrt{\chi^2_{\nu,0.975}}$. To illustrate the classical Mahalanobis distance is not reliable when the data involve some outliers, we consider the *wine* data set (Hettich and Bay, 1999). It contains, for each of 59 wines grown in the same region in Italy, the quantities of 13 constituents. We will use two methods – classical and robust – to investigate whether the method of Mahalanobis distance with classical estimation can detect some outlying observations.

The left column of Figure 1.1 presents the plots for the squared distances versus the number of the observations. The horizontal line is the threshold value equal to $\sqrt{\chi^2_{13,0.975}} =$

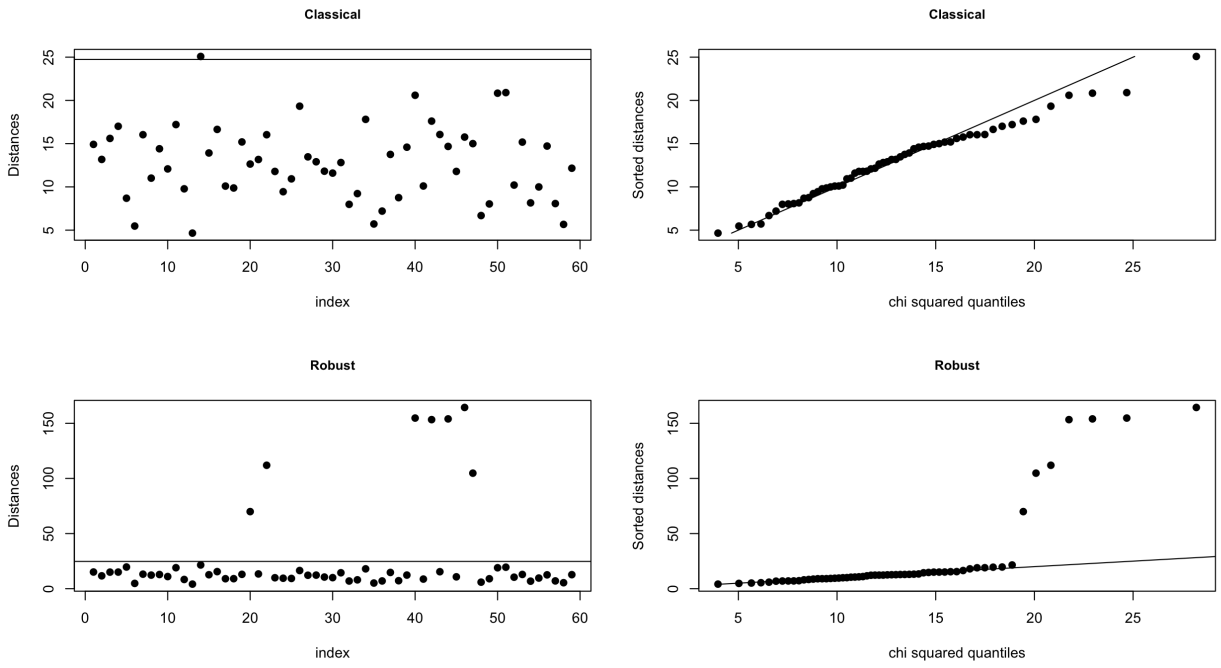

Figure 1.1: Mahalanobis distances versus index number for classical and robust estimators (left), and QQ plots of distances (right).

24.7356. It is evident that classical method fails to identify outliers in *wine* data set. The reason is that outliers may "mask" one another. As stated by Rousseeuw and Hubert (2011), classical methods can be affected by outliers so strongly that the resulting analysis does not allow to detect the deviating observations. This is called the masking effect. But when it comes to robust method, as the bottom-left figure shows, seven outliers stand out clearly. The second column of Figure 1.1 represents the QQ plots with respect to the χ^2_p distribution. In classical QQ plot, no clear outliers stand out. But we can immediately see that the data do not form a homogeneous cloud in robust QQ plot. There are some observations which deviate from the shape of the majority of the data. Therefore, in order to reliably estimate the center and scatter of the data set, robust estimators of location and scatter are needed, which is what we will mainly focus on in next chapter.

1.2 Masking and Swamping Effects

Classical approaches like Mahalanobis distance can be significantly influenced by outlying observations, leading to fail to detect outliers in data sets. Masking and swamping are two main effects which can affect traditional techniques. The following definitions do not include mathematical rigor, but they provide an intuitive understanding of these effects (See Hawkins (1980); Iglewicz and Martinez (1982); Davies and Gather (1993) for more definitions about masking and swamping).

Definition 1. *An outlier masks a second one that is close by if the latter can be considered an outlier by itself, but not if it is considered along with the first one (Ben-Gal, 2005).*

That is to say, as a result of deleting the first outlier, the second instance appears to be an outlier. Generally, masking effect occurs when clusters of discordant points skew the mean and covariance estimates in its direction, resulting in a close distance between the outlying point and the mean.

Definition 2. *An outlier swamps another observation if the latter can be considered outlier only under the presence of the first one (Ben-Gal, 2005).*

Thus, one outlier may behave normal like the large part of data after we delete another outlier. Swamping occurs when a group of outliers skews the mean and covariance matrix in its direction, which leads to a large distance between them and the mean, making them look like outliers.

It is important to note that masking and swamping affect the effectiveness of Mahalanobis distance as an criterion of outliers. Accordingly, the utilization of masking effects could potentially reduce the Mahalanobis distance assigned to an anomalous observation. One possible scenario is that the presence of a few uncommon data points pulls the sample mean and skews the variance in their direction. On the other hand, swamping effect has the potential to augment the Mahalanobis distance of non-outlying observations. This effect occurs when a small group of anomalies draws the sample mean and generates a deviation

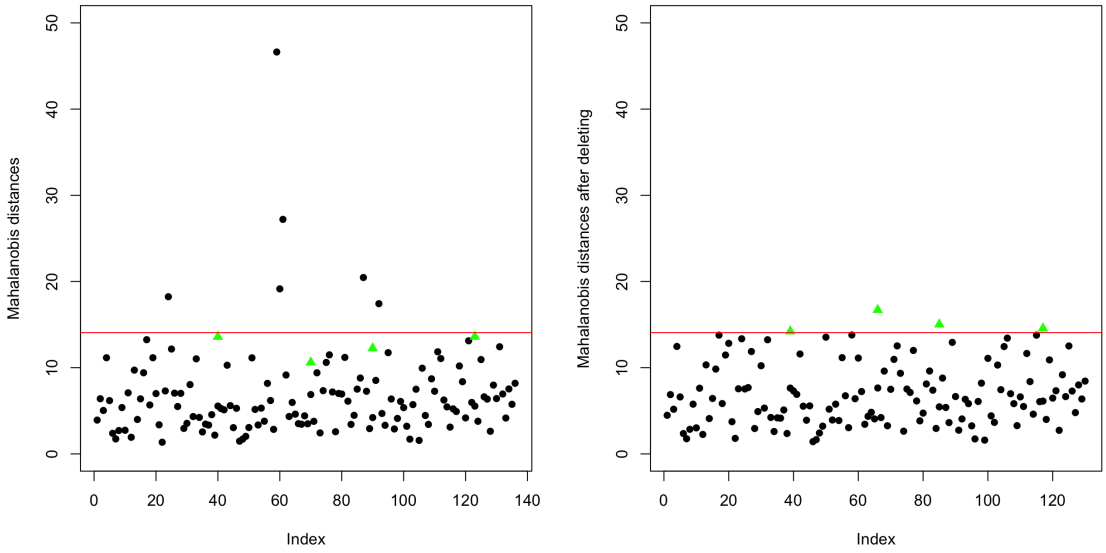

Figure 1.2: Masking effect. Mahalanobis distances for the original data (left) and the data deleting six outliers (right).

in the covariance estimate, thereby diverging from the pattern exhibited by the majority of the data (Penny and Jolliffe, 2001). Figure 1.2 describes masking effect. As is shown on the left of Figure 1.2, there are six outliers whose Mahalanobis distances are above the threshold which is marked by the red line. Upon the elimination of six initial outliers, it is evident that four data points, denoted by green triangles, which were previously deemed non-outliers, have been identified as outlying observations. This can be seen in the picture on the right in Figure 1.2. In this case, we can say four observations with green triangles are masked by the six previous outliers. Figure 1.3 displays swamping effect. two data points with green are swamped by the other two outliers, since the two outliers in green only exist under the presence of the two outliers.

Chiang et al. (2007) documented that the masking and swamping effects are still unavoidable in linear models based on the OLS method and showed that the locations of outliers, their signs of residuals, permutations, and the sum of deviations of all outliers are

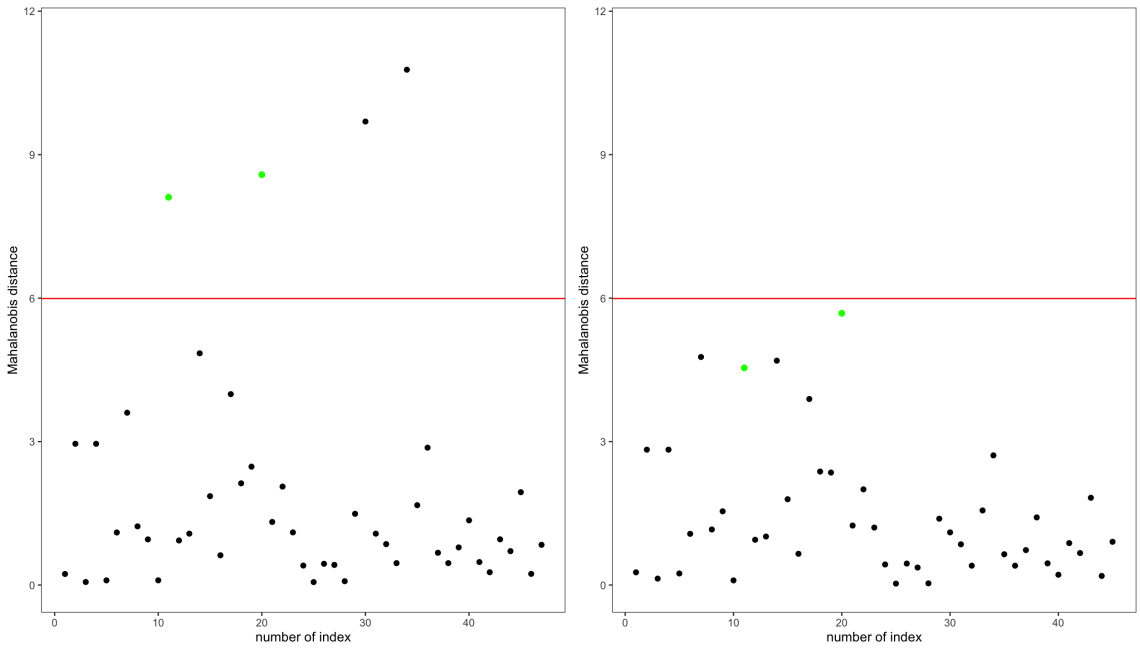

Figure 1.3: Swamping effect. Mahalanobis distances for the original data (left) and the data deleting two outliers (right).

the main factors for the two effects using the planted mean-shift outliers models. Kannan and Manoj (2015) provided an example of masking effect using Grubbs test under the case of outlier detection. Bendre and Kale (1985) compared the performances between the modified Dixon-type test and the Cochran test for exponential samples. They found that Cochran and modified Dixon-type tests do not suffer from the masking effect in the presence of two outliers.

1.3 Research Objectives

Robust statistics have become widespread since Tukey (1960), Hampel (1971), and Huber (1992) fundamental works. There are several methods for outlier detection are based on robust statistics. Like Minimum Covariance Determinant (MCD) estimator, the robust estimator of location and scatter needs to be obtained before the process of identifying

anomalous observations. Therefore, we provided some background of robust statistics. This thesis pursues two aims. One is to offer a broad overview of robust statistical approaches in the field of outlier detection. And the other intends to find a technique with higher efficiency of identifying anomalous data on two scenarios – low-dimensional setting and high-dimensional setting.

This work is organized as follows. The present Chapter 1 describes the background of outlier detection and explores the two significant impacts that render conventional statistical methods insufficient. In Chapter 2, we talk about three types of techniques for outlier detection with details. The end of Chapter 2 is dedicated to the discussion of the several properties of robust estimators. The application for low-dimensional data set is conducted in Chapter 3. Three comparable methods are chosen to compare their performances. Chapter 4 presents the application on high-dimensional data set. Another three methods appropriate for dealing with high-dimensional data set are selected to draw comparisons. And finally, the findings are summarized in Chapter 5.

Chapter 2

Robust Statistical Methods for Location and Scatter

The sample mean and the sample covariance matrix are the cornerstone of the classical multivariate analysis. But they are extremely sensitive to small perturbations in data. In this chapter, we will survey some important robust estimates of multivariate location and scatter. Some of them will be utilized in the following application part.

2.1 Distance-Based Methods

M-Estimators

Affine equivariant M-estimates of multivariate location and scatter were proposed by Maronna in 1976. Maronna (1976) defined multivariate M-estimators as the solutions of the system of estimating equations:

$$\frac{1}{n}\sum_{i=1}^{n} W_1\Big[((\boldsymbol{x}_i-\hat{\boldsymbol{\mu}})'\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}}))^{\frac{1}{2}}\Big](\boldsymbol{x}_i-\hat{\boldsymbol{\mu}})=\mathbf{0}, \tag{2.1}$$

$$\frac{1}{n}\sum_{i=1}^{n} W_2\Big[((\boldsymbol{x}_i-\hat{\boldsymbol{\mu}})'\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}}))^{\frac{1}{2}}\Big](\boldsymbol{x}_i-\hat{\boldsymbol{\mu}})(\boldsymbol{x}_i-\hat{\mu})'=\hat{\boldsymbol{\Sigma}}. \tag{2.2}$$

where W_1, W_2 are weight functions satisfying some general assumptions. For normal distribution $W \equiv 1$, which yields the sample mean and sample covariance matrix for $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$. Affine equivariant M-estimators are a generalization of the maximum likelihood estima-

tors and can be considered as weighted mean and weighted covariance matrix. A primary weakness of these is that the breakdown point is at most $1/(p+1)$, which is relatively low for even a moderately large number of variables (Huber, 1977). Furthermore, Devlin et al. (1981) found that M-estimators in practice could tolerate even fewer outliers than indicated by this upper bound.

GM-Estimators

Generalized M-estimators (GM-estimators) are referred to "bounded influence estimators" which were proposed to overcome the $x-$outlier problem of M-estimators and, therefore, improve the breakdown point. These methods mainly intend to bound the influence of outlying $\boldsymbol{x}_i$ through some weights w_i that assign full influence to observations assumed to come from the majority of the data, but reduced influence to aberrant observations. Using a sensible estimate the iterative technique will continue until the sequence of estimates has converged to within the desired accuracy. The breakdown point of all GM-estimators in general can be no larger than 30%, which is a decreasing function of the dimension p (Maronna et al., 2006). A number of weights have been developed, for example Tukey's biweight (Beaton and Tukey, 1974; Huber, 1992, 1973; Andrews and Hampel, 2015) and Andrew's wave (Campbell, 1980). These weights are not limited to GM-estimators and can be used for all kinds of estimates requiring a weight function.

MVT Estimator

Ellipsoidal multivariate trimming (MVT) was formally proposed by Gnanadesikan and Kettenring (1972). The squared Mahalanobis distance of the observation vector $\boldsymbol{x}_i$ from the current robust estimate of location $\boldsymbol{x}^*$ and scatter matrix $\boldsymbol{C}^*$ are measured for each step of the iterative process. A specified percentage (the trimming percentage) of the most extreme observations (i.e., objects with the largest squared Mahalanobis distance) is temporarily se-

lected (at most 50% of the data) and the remaining observations will be used to calculate $\boldsymbol{x}^*$ and $\boldsymbol{C}^*$ for next iterative step exactly as $\boldsymbol{x}$ and $\boldsymbol{C}$, the sample mean and covariance matrix. The iterative procedure will not terminate until both of the sequences of $\boldsymbol{x}^*$ and $\boldsymbol{C}^*$ converge. Empirically, MVT has been found to converge quickly, usually in two or three steps (Walczak and Massart, 1995). Devlin et al. (1981) claimed that the breakdown point of MVT was the same as its trimming percentage (at most 50%), and does not decrease with the number of variables. However, Donoho (1982) argued that the breakdown point of MVT is at most about $1/p$, thus causing this method less attractive due to its low breakdown point.

MVE Estimator

Another affine equivariant estimator of multivariate location and scatter with high breakdown value is the minimum volume ellipsoid estimator (MVE), which was first introduced by Rousseeuw (1985). The MVE is often used thanks to its high resistance to outliers, making it a reliable tool for outlier detection. It is defined as the smallest ellipsoid covering at least h observations of the dataset $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\} \subseteq \mathbb{R}^p$. The MVE location estimator is the midpoint of the ellipsoid and the MVE scatter estimator is the covariance matrix of all the data points in the ellipsoid. Equivalently, we consider the following optimization problem.

$$(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n) := \underset{\boldsymbol{\mu}, \boldsymbol{\Sigma}}{\operatorname{argmin}} \{\det(\boldsymbol{\Sigma}) | f(i, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \geq h\}, \tag{2.3}$$

$$f(i, \boldsymbol{\mu}, \boldsymbol{\Sigma}) := \{\pi(i); (\boldsymbol{x_i} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x_i} - \boldsymbol{\mu}) \leq c^2\}. \tag{2.4}$$

where $\pi(i)$ is the number of is which satisfies the inequality $(\boldsymbol{x_i} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x_i} - \boldsymbol{\mu}) \leq c^2$. And the parameter c is a constant that determines the magnitude of $\boldsymbol{\Sigma}_n$. Usually, c is selected such that $\boldsymbol{\Sigma}_n$ is a consistent estimator for data coming from a multivariate normal distribution, i.e., $c = \sqrt{\chi^2_{p,\alpha}}$, where $\alpha = h/n$. The parameter h determines the robustness of MVE estimators. h is often set at $h = \lfloor \frac{n+p+1}{2} \rfloor \approx \frac{n}{2}$ (Lopuhaä and Rousseeuw, 1991). The main drawback of MVE is the computational time. Rousseeuw (1985) proposed an

algorithm based on subsampling, called the $(p+1)$-subset algorithm. But this algorithm suffers from inefficiency and high computational complexity, which makes it impractical for use with high-dimensional data sets (Ammann, 1993). In the last few decades, there are a large number of papers working on improved algorithms. Croux and Haesbroeck (1997) proposed an adaption of the $(p+1)-$ subset algorithm whose main difference is to average over several trial values instead of just picking out the optimal one, making a larger finite-sample efficiency. Croux et al. (2002) applied L_1 location adjustment to the MVE, yielding a new estimator which is cheap in computation time, has a low bias curve, and gives more efficient estimates of the multivariate location parameter in the normal case.

MCD Estimator

The minimum covariance determinant (MCD) estimator is one of the most popular procedures to estimate the location and scatter matrix of a multivariate data set. These estimators build a cornerstone in other multivariate statistical methods, such as principal component analysis (Croux and Haesbroeck, 2000), discriminant analysis (Hawkins and McLachlan, 1997), multivariate regression (Rousseeuw et al., 2004), canonical correlations (Taskinen et al., 2006), among others (see Hubert et al. (2008) for a more comprehensive overview). The objective of MCD estimator is to find h observations whose covariance matrix has the smallest determinant, where $n/2 \leq h \leq n$. The parameter h determines the robustness of the MCD estimator. When h is equal to $\lfloor (n+p+1)/2 \rfloor$, the MCD estimator reaches its highest possible breakdown point (Lopuhaä and Rousseeuw, 1991). In definition, it is simple and intuitively attractive, while it has some desirable properties (Butler et al., 1993).

This estimator, however, has a combinatorial optimization problem. Since FASTMCD algorithm was proposed by Rousseeuw and Driessen (1999), the MCD estimator has been widely used and is showing a tendency to replace the MVE estimator. The main feature of FASTMCD algorithm is the C-step which is based on Mahalanobis distance and order statistics. Rousseeuw and Driessen (1999) showed for a fixed numbers of dimension p, the

C-step takes only O(n) time and it must converge within finite steps of iteration. Pokojovy and Jobe (2022) proved the fixed points of C-step iteration are the local minimizers of the covariance determinant instead of some artificial attractor sets, which resolved a long-standing problem in C-step algorithm. In the last decade, a large number of extensions of MCD estimator have emerged. The minimum weighted covariance determinant (MWCD) estimator was introduced by Roelant et al. (2009), which assigns a particular weight to each observation, improving the efficiency of the MCD estimator. Hubert et al. (2012) provided a new algorithm for MCD estimator which is called DetMCD algorithm. In this algorithm, MCD estimator becomes deterministic but not affine equivariant. And it is faster than FASTMCD in practice. The MCD algorithm is not appropriate for high-dimensional data sets. Boudt et al. (2020) regularized the scatter matrix of the MCD estimator such that the covariance matrix of h data points is no longer non-singular when the number of columns of data exceeds the number of rows.

OGK Estimator

Orthogonalized Gnanadesikan-Kettenring (OGK) estimator introduced by Maronna and Zamar (2002), is a robust statistical method used for estimation and analysis of multivariate data. It is a variation of the principal component analysis (PCA) that involves the extraction of orthogonalized components. Unlike the traditional PCA method, OGK estimator reduces the impact of outliers and noisy data in the estimation process. The estimate, first introduced by Gnanadesikan and Kettenring (1972), is mainly based on the equation

$$\mathrm{Cov}(X, Y) = \frac{1}{4}(\sigma(X + Y)^2 - \sigma(X - Y)^2). \tag{2.5}$$

where X, Y are random variables and σ stands for the standard deviation. They applied a robust scale σ to define a "robust covariance matrix". But this covariance matrix is not necessarily positive definite. Maronna and Zamar (2002) modified it and provided a general method to obtain positive-definite and approximately affine-equivariant robust scatter matrix in OGK estimator. And they also demonstrated that the lack of equivariance of OGK

estimator will not cause serious problem. The OGK estimator is composed of two steps. The first step involves the normalization of the data and the calculation of the principal components. In the second step, the principal components are transformed to orthogonalized components using the Gnanadesikan-Kettenring technique. These orthogonalized components are able to resistant to outlying observations compared to the principal components. The OGK estimator is widely used in various disciplines such as finance, economics, and engineering, where multivariate data analysis is critical.

Ledoit-Wolf Estimator

The Ledoit-Wolf estimator is a popular method used in robust statistics for estimating the covariance matrix of a set of variables. It was proposed by Ledoit and Wolf (2004) as an improvement over existing methods such as the Maximum Likelihood Estimator (MLE), which can often be unstable in the presence of outliers. It is not only suitable for small sample size n and large numbers of variables p but at the same time is also completely inexpensive to compute. The Ledoit-Wolf estimator is a shrinkage estimator, which means that it is designed to shrink the estimated covariance matrix towards a more structured target matrix. The target matrix used in this method is the diagonal matrix, which assumes that the variables are uncorrelated. This estimator achieves shrinkage by adapting the sample covariance matrix based on the data. Specifically, it applies a linear transformation to the sample covariance matrix to make it closer to the target matrix. The amount of shrinkage applied is determined by a parameter called the shrinkage intensity which guarantees the estimator has minimum of expected quadratic loss function (Schäfer and Strimmer, 2005).

One of the major advantages of the Ledoit-Wolf estimator is that it can improve the accuracy of covariance matrix estimation in the presence of outliers, by reducing the impact of extreme values. This makes it a popular choice for applications such as portfolio optimization, where accurate covariance matrix estimation is critical.

2.2 Projection-Based Methods

Stahel-Donoho Method

The Stahel–Donoho estimator or "outlyingness-weighted median" was the first location and scatter estimator in high dimension that can integrate affine equivariance with high breakdown points (Donoho, 1982). In Stahel-Donho method, we look for a one-dimensional projection where $\boldsymbol{x}_i$ is most outlying in the way defined as

$$\boldsymbol{\nu_i} = \sup_{||u||=1} \frac{|\boldsymbol{x}_i\boldsymbol{u}' - \text{med}_i(\boldsymbol{x}_i\boldsymbol{u}')|}{\text{med}_k|\boldsymbol{x}_k\boldsymbol{u}' - \text{med}_j(\boldsymbol{x}_j\boldsymbol{u}')|}. \tag{2.6}$$

where $\text{med}_j(\boldsymbol{x}_j\boldsymbol{u}')$ is the median of projections of all observations $\boldsymbol{x}_j$ on the direction of the vector $\boldsymbol{u}$. The location and scatter are then estimated by the weighted mean and the weighted covariance matrix with weights of the form $w(\boldsymbol{\nu})$:

$$\hat{\boldsymbol{\mu}} = \sum_{i=1}^{n} w(\boldsymbol{\nu}_i)\boldsymbol{x}_i / \sum_{i=1}^{n} w(\boldsymbol{\nu}_i), \tag{2.7}$$

$$\hat{\boldsymbol{\Sigma}} = \sum_{i=1}^{n} w(\boldsymbol{\nu}_i)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})' / \sum_{i=1}^{n} w(\boldsymbol{\nu}_i). \tag{2.8}$$

Stahel (1981) considered the asymptotic breakdown point of the estimators. Donoho (1982) derived the finite sample breakdown point for being median (med) and median absolute deviation (MAD), for $\boldsymbol{x}$ in a general position, and for suitable weight function w. The asymptotic behavior of the Stahel-Donoho estimator, however, was a long-standing problem. This hampered this estimator from becoming more popular in reality. Maronna and Yohai (1995) first proved the $\sqrt{n}-$consistency. Establishing the limiting distributions, however, turned out to be extremely challenging. The breakdown properties of the MAD (i.e., the denominator in Equation (2.6) basically determine the breakdown of the Stahel-Donoho method and corresponding modifications have been suggested to obtained further improvements in this respect (Zuo, 2000, 2004).

Kurtosis Method

Peña and Prieto (2001) proposed a method in which kurtosis coefficients are used to obtain directions. The sample points are projected on to a set of $2p$ directions, where p is the dimension of the data. These directions are determined through minimizing and maximizing the kurtosis coefficients of projected points. To decide the outlyingness of observations on any of these $2p$ directions, the maximum distance of observations from the median exceeds a suitable cutoff value. These estimates are then used to compute Mahalanobis distance for entire data and the observations are labeled as outliers whose distances exceed the desired quantile of χ^2 distribution with p degrees of freedom. The performance of kurtosis coefficient directions was not satisfactory for large contamination levels.

Peña and Prieto (2001) examines the influence of outliers on kurtosis values and the effective utilization of this moment coefficient for their detection. They found that in the situation where the outlier model is constructed using a contaminated distribution that belongs to the same family as the original distribution or has heavier tails, it can be anticipated that the kurtosis coefficient of the observed data will exceed that of the original distribution. However, in the context of asymmetry, a significant level of contamination will result in a very small value of the kurtosis coefficient, while a low level of contamination will yield a relatively high value of the coefficient. Hence, it is justifiable to utilize projection directions which either maximize or minimize the kurtosis coefficient of the projected data.

Projection Pursuit MCD Method

Projection Pursuit (PP) MCD algothrim was introduced by Pokojovy and Jobe (2022) under the motivation of Peña and Prieto (2001)'s work. Projection Pursuit MCD method aims to find one "warm start" pair using projection pursuit approach and then continuously improve it with C-step iteration. This method is distinguished from current prominent MCD methods due to its properties. Since PP MCD method does not utilize randomization and it solely relies on summation for statistical measurements, the location and scatter estima-

tors are deterministic and permutation invariant as well. But also the projection pursuit approach in this algorithm is performed in an affine equivariant fashion, which guarantees the nature of affine equivariance for the estimators coming from the PP MCD algorithm. As discussed in the last section, there are two other MCD algorithms which are FASTMCD algorithm and DetMCD algorithm. Now we briefly compare these three algorithms in a general way. FASTMCD uses subsampling method to find a large number of "warm starts", which shows its estimator is affine equivariant. However, DetMCD does not sample to get "warm stars", instead it has six predetermined "warm starts" in which some methods are not affine equivariant. Next we will discuss the core of the PP MCD methodology—projection indices.

The method of Peña and Prieto (2001) was based on large-sample theory. Therefore, their method might be broken in practice for small sample size n and/or large dimension p. Pokojovy and Jobe (2022) in their PP MCD method provided two new projection indices which are not based on large-sample assumptions. The first projection index is related to the ratio of sample variance and raw MCD scatter estimator.

$$Q_{\text{var}}(Z) = \log\left(\frac{s^2}{\hat{\sigma}^2_{\text{MCD}}}\right). \tag{2.9}$$

with $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})$ sample variance and $\hat{\sigma}^2_{\text{MCD}}$ raw MCD scatter estimator in univariate situation. The larger the value of $Q_{\text{var}}(Z)$, the bigger difference between the sample variance and MCD scatter estimator, which shows the direction the data are projected on is more likely to distinguish "good" observations and outliers. That is the reason why the PP step part of the PP MCD algorithm needs to find local maximum of $Q_{\text{var}}(\cdot)$ using the projected gradient method to improve the "potential" direction. And of course, the level of information provided by this measure decreases when the variable Z includes a significant proportion of tightly clustered outlying observations.

The second projection index is obtained from a Gaussian bimodality test.

$$H_0 : z_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_0, \sigma_0^2), \qquad H_1 : z_i \overset{\text{i.i.d.}}{\sim} (1-\varepsilon)\mathcal{N}(\mu_1, \sigma_1^2) + \varepsilon\mathcal{N}(\mu_2, \sigma_2^2). \tag{2.10}$$

where $\varepsilon \in (0, 1)$ and $(\mu_1, \sigma_1^2) \neq (\mu_2, \sigma_2^2)$. Then two variances are computed under H_0 and

H_1:

$$\mathrm{Var}(z_i|H_0) = \sigma_0, \qquad \mathrm{Var}(z_i|H_1) = (1-\varepsilon)^2\sigma_1^2 + \varepsilon^2\sigma_2^2 + \varepsilon(1-\varepsilon)(\mu_1-\mu_2)^2. \tag{2.11}$$

The test statistic is defined as follows

$$Q_{\mathrm{bimod}}(Z) = \frac{s^2}{(1-\hat{\varepsilon}_{\mathrm{ML}})^2\hat{\sigma}_{1,\mathrm{ML}}^2 + \hat{\varepsilon}_{\mathrm{ML}}^2\hat{\sigma}_{2,\mathrm{ML}}^2 + \hat{\varepsilon}_{\mathrm{ML}}(1-\hat{\varepsilon}_{\mathrm{ML}})(\hat{\mu}_{1,\mathrm{ML}}-\hat{\mu}_{2,\mathrm{ML}})^2}. \tag{2.12}$$

where s^2 is again the sample variance and $(\hat{\varepsilon}_{\mathrm{ML}}, \hat{\mu}_{1,\mathrm{ML}}, \hat{\mu}_{2,\mathrm{ML}}, \hat{\sigma}_{1,\mathrm{ML}}, \hat{\sigma}_{2,\mathrm{ML}})$ denotes the maximum likelihood estimators of $(\varepsilon, \mu_1, \mu_2, \sigma_1, \sigma_2)$. Larger value of $Q_{\mathrm{bimod}}(Z)$ provides strong evidence to reject H_0 in favor of H_1, i.e., bimodality, while smaller value tends to accept the null hypothesis, i.e., unimodality.

Even though the test (2.10) looks simple, the challenging part of this test lies in the maximum likelihood estimators. Finding the parameter vector $(\hat{\varepsilon}_{\mathrm{ML}}, \hat{\mu}_{1,\mathrm{ML}}, \hat{\mu}_{2,\mathrm{ML}}, \hat{\sigma}_{1,\mathrm{ML}}, \hat{\sigma}_{2,\mathrm{ML}})$ is a non-convex optimization problem. And no closed-form solution exists. Hence, numerical method needs to be used to find the test statistic. In order to increase the speed of the numerical method, Pokojovy and Jobe (2022) decided to apply Newton's method to maximize the problem instead of the normally used Expectation Maximization approach. $Q_{\mathrm{bimod}}(Z)$ appears to be a sufficient measure for projecting both uni- and bimodal data configurations. The efficacy of $Q_{\mathrm{var}}(Z)$ is generally reliable across various configurations, however, its functionality is susceptible to possible failure in instances where the bimodal mixture components exhibit almost equal sizes but possess highly distinct variances. Under this particular instance, $Q_{\mathrm{var}}(Z)$ displays a tendency to become exponentially large when certain unfavourable projections are employed. Therefore, these two projection indices are jointly used to filter outliers until only h observations are left which is the data set of getting the "warm start" for the following C-steps.

PCOut Method

PCOut algorithm was proposed by Filzmoser et al. (2008), which exhibits remarkable efficiency in situations characterized by a large number of dimensions. The first two letters

with upper case mean principal components decomposition. Filzmoser et al. (2008) employed basic properties of principal components to detect anomalous data points within reduced dimension, thereby resulting in notable computational benefits for data sets with a high degree of dimensions. The PCOut method showed very competitive performance when compared to existing outlier detection methods under high dimensional data sets of gene expression data and geochemical data. The key advantage of this algorithm is its computational speed. As Filzmoser et al. (2008) presented, even though the data have 2000 observations and 2000 variables, it only cost around 2 minutes to identify the outliers. However, other methods displayed an exponential increase in magnitude.

The PCOut algorithm is comprised of two basic parts: one is to identify location outliers and the other is to identify scatter outliers. The method computes the two weights obtained from these two steps, and combines the two weights as the outlyingness measure. If the measure is less than a particular cutoff value, the data point will be classified as outlier. Firstly, each component of the data are scaled using the median and the MAD based on

$$x_{ij}^* = \frac{x_{ij} - \text{med}(x_{1j}, \cdots, x_{nj})}{\text{MAD}(x_{1j}, \cdots, x_{nj})}, \quad j = 1, \cdots, p. \tag{2.13}$$

The scaled data x_{ij}^* can be used to calculate the covariance matrix from which a principal component decomposition is conducted. And next we find the eigenvectors which contribute to at least 99% of the total variance. These eigenvectors consist of the new $p^* \times p^*$ matrix $\boldsymbol{V}$. Thus we obtain the principal components of the data by

$$\boldsymbol{Z} = \boldsymbol{X}^* \boldsymbol{V}. \tag{2.14}$$

We rescale again the principal components based on the median and the MAD by

$$z_{ij}^* = \frac{z_{ij} - \text{med}(z_{1j}, \ldots, z_{nj})}{\text{MAD}(z_{1j}, \ldots, z_{nj})}, \quad j = 1, \ldots, p^*. \tag{2.15}$$

Store the rescaled principal components Z^* for the second part of the method. Then calculate a robust kurtosis measure for each component of Z^* according to

$$w_i^{kurt} = \left| \frac{1}{n} \sum_{i=1}^{n} \frac{(z_{ij}^* - \text{med}(z_{1j}^*, \ldots, z_{nj}^*))^4}{\text{MAD}(z_{1j}^*, \ldots, z_{nj}^*)} \right|, \quad i = 1, \ldots, p^*. \tag{2.16}$$

We assign the relative weights $w_i^{kurt}/\sum_{i=n}^{n} w_i^{kurt}$ to each components in Z^* relying on the kurtosis measure described in Equation (2.16). Then, we find the Mahalanobis distances of the matrix Z^* from the median. And the Mahalanobis distances will be transformed by

$$d_i = \mathrm{RD}_i \frac{\sqrt{\chi^2_{p^*,0.5}}}{\mathrm{med}(\mathrm{RD}_1, \ldots, \mathrm{RD}_n)}, \quad i = 1, \ldots, n. \tag{2.17}$$

According to Filzmoser et al. (2008), the transformed robust distances have the same median as the theoretical distances and bring them closer to $\chi^2_{p^*}$ distribution. Lastly, calculate the weights for all the observations utilizing the translated biweight function.

$$w_{1i} = \begin{cases} 0, & d_i \geq c \\ (1 - (\frac{d_i - M}{c - M})^2)^2, & M \leq d_i \leq c. \\ 1, & d_i \leq M \end{cases} \tag{2.18}$$

where $i = 1, \ldots, n$, $c = \mathrm{med}(d_1, \ldots, d_n) + 2.5 \cdot \mathrm{MAD}(d_1, \ldots, d_n)$, and M is the 1/3 quantile of the distance $\{d_1, \ldots, d_n\}$. The starting point of the second part of PCOut method comes from the principal components space Z^* from the first part. We do not assign the kurtosis weights to the principal components, instead Mahalanobis distances are computed directly from Z^*. Given that the kurtosis weighting scheme did not modify the distribution of these distances and assuming that non-outliers are normally distributed, by applying Equation (2.17) to the robust distances the resulting distribution closely approximates $\chi^2_{p^*}$. Again we use the translated biweight function to obtain the second weight w_{2i}. But the difference is that $M^2 = \chi^2_{p*,0.25}$ and $c^2 = \chi^2_{p^*,0.99}$, where $\chi^2_{p*,0.25}$ is the 25th percentile of χ^2 distribution from the left. Finally, we can combine the two weights w_{1i} and w_{2i} by using the following equation.

$$w_i = \frac{(w_{1i} + s)(w_{2i} + s)}{(1 + s)^2}, \quad i = 1, \ldots, n. \tag{2.19}$$

In general, the value assigned to the scaling constant s is often 0.25. Sometimes, there are instances where numerous non-outliers are given a weight of 0 in just one of the two phases. The purpose of introducing s is to alleviate this issue by setting it not to be 0, which ensures

that the final weight w_i will only be 0 if both phases have a low weight. Points with a weight of less than 0.25 are categorized as outliers.

2.3 Other Methods

Juan and Prieto (2001) introduced a technique for identifying outliers that relies on estimated angles to classify the clustered outliers. The authors explicated that the relationship between the distribution of uncontaminated data, conforming to a uniform distribution, and the reference direction can be modeled as a function of the Beta distribution. Upon obtaining the angles, it was recommended to construct a $Q - Q$ plot in order to assess the presence of outliers utilizing the lack of fit method. Pyke (1965) conducted research on the spacing test as a measure of goodness-of-fit. Subsequently, the distribution of the intervals between each sequentially arranged observation is ascertained. The cutoff value for identifying outliers was computed by utilizing the largest normalized spacing interval obtained from distribution analysis.

A subjective procedure utilizing eigenvalues and eigenvectors was established by Gao et al. (2005), known as the Max-Eigen Difference (MED) technique. The first step of this approach involves identifying the eigenvalues and eigenvectors of the scatter matrix for the initial set of data. The eigenvalues and eigenvectors are determined for the covariance matrices of each dataset after removing the ith observation. Gao et al. (2005) indicated that observations with higher MED values are identified as outliers. Kirschstein et al. (2013) introduced pruning Minimum Spanning Tree (pMST) technique which considers the whole dataset as a network. The pMST methodology initiates by creating a sphere encompassing each observation, which is delineated by a specific radius. The subsample of good observations can be determined for robust estimation purposes by using the count of spheres in the biggest grouping of connected spheres.

2.4 Measures of Robustness

Various concepts of robustness have been taken into consideration for multivariate location and scatter estimators. Hodges Jr (1967) proposed the concept of breakdown point, which is a universal metric for robustness. An attractive and straightforward limited-sample iteration of this concept was presented by Donoho and Huber (1983). Roughly, this finite-sample replacement breakdown point measures the minimum fraction of outliers which will spoil the estimate completely (Lopuhaä and Rousseeuw, 1991). If one estimator is zero breakdown point, it will be treated as non-robust. Equivariance under affine transformations is a fundamental requirement for multivariate estimators in their natural state. The integration of affine equivariance and a high breakdown point is a non-trivial task. Donoho (1982) analyzed several affine equivariant techniques for multivariate estimators, showing that their breakdown point approaches to 0 as the dimension p increases. Stahel (1981) and Donoho (1982) proposed the first estimator of multivariate location and scatter with both affine equivariance and high breakdown value, independently. In this section, we will discuss several measures of robustness of estimators for multivariate location and scatter, such as, affine equivariance breakdown point and influence function.

2.4.1 Affine Equivariance

Affine equivariance is an important property of a robust estimator. This property suggests the estimator will transform well under any affine transformation in the space where the $\boldsymbol{x}_i$'s live. Therefore, even if the data points are subject to rotation, translation, or scaling, it will not impact the outlier detection diagnostic results. Specifically, one estimator of location and scatter is affine equivariant, which means that for any column vector $\boldsymbol{a} \in \mathbb{R}^p$ and nonsingular $p \times p$ matrix $\boldsymbol{A}$ the following system holds that

$$\hat{\boldsymbol{\mu}}(\boldsymbol{X}\boldsymbol{A} + \boldsymbol{a}) = \hat{\boldsymbol{\mu}}(\boldsymbol{X})\boldsymbol{A} + \boldsymbol{a}, \tag{2.20}$$

$$\hat{\boldsymbol{\Sigma}}(\boldsymbol{X}\boldsymbol{A} + \boldsymbol{a}) = \boldsymbol{A}'\hat{\boldsymbol{\Sigma}}(\boldsymbol{X})\boldsymbol{A}. \tag{2.21}$$

where $\boldsymbol{X}$ is $n \times p$ matrix of data, $\boldsymbol{A}'$ is the transpose of matrix $\boldsymbol{A}$. We take MCD estimator as an example. In order to obtain MCD estimator of location and scatter, we need to iterate the C-steps which guarantee the property of affine equivariance. This is based on the fact that the determinant of the scatter matrix for the transformed data is equal to

$$|\boldsymbol{S}(\boldsymbol{X}_H\boldsymbol{A})| = |\boldsymbol{A}'\boldsymbol{S}(\boldsymbol{X}_H)\boldsymbol{A}| = |\boldsymbol{A}|^2|\boldsymbol{S}(\boldsymbol{X}_H)|. \tag{2.22}$$

where $\boldsymbol{S}(\boldsymbol{X}_H)$ is the covariance matrix of data $\boldsymbol{X}_H$, subset H is any subset of $\{1, 2, \cdots, n\}$ of size h, $\boldsymbol{X}_H$ is the corresponding data set for H.

Hence, by transforming a subset of $\boldsymbol{X}_H$ with the smallest determinant, we can obtain a corresponding subset $\boldsymbol{X}_H\boldsymbol{A}$ with the smallest determinant out of all subsets of the transformed dataset. Additionally, the covariance matrix of this subset $\boldsymbol{X}_H\boldsymbol{A}$ is appropriately transformed. The affine equivariance of the raw Minimum Covariance Determinant (MCD) location estimator can be established by the equivariance of the sample mean. Therefore, the MCD estimator of location and scatter from FASTMCD algorithm is affine equivariant. However, DetMCD method does not have this property. The reason is that DetMCD method relies on six initial scatter matrices, which explains its rationale. The optimal data set $\boldsymbol{X}_H$ derived from the transformed data may not be identical to the one obtained from the original data set. Possibly, allowing for some leeway in this property could result in more reliable estimates being accessible or could be applicable to a non-affine equivariant estimator that exhibits superior performance. Nevertheless, the property of being affine equivariant is still beneficial for the estimation of location and scatter in multivariate analysis.

2.4.2 Breakdown Point

The breakdown point is the smallest fraction of data points that can be contaminated by arbitrary values to make the estimate arbitrarily large (Hampel, 1971). Mathematical definition of breakdown value for multivariate location and scatter has been introduced by Lopuhaä and Rousseeuw (1991). Consider a data set matrix $\boldsymbol{X}_{n\times p}$, the breakdown point

$\epsilon^*(\hat{\boldsymbol{\mu}}, \boldsymbol{X})$ for the location estimator $\hat{\boldsymbol{\mu}}$ is defined as:

$$\epsilon^*(\hat{\boldsymbol{\mu}}, \boldsymbol{X}) = \min\left\{\frac{m}{n} : \sup||\hat{\boldsymbol{\mu}}(\boldsymbol{X}_m) - \hat{\boldsymbol{\mu}}(\boldsymbol{X})|| = \infty\right\}. \tag{2.23}$$

where $1 \leq m \leq n$ and data set $\boldsymbol{X}_m$ is derived from the original data set $\boldsymbol{X}$ in which any m observations are replaced by arbitrary values. From another angle, the definition of the breakdown point aims to find the point where the distance between the contaminated sample mean and the clean sample mean approaches to infinity. Even though it may seem like $\epsilon^*(\hat{\boldsymbol{\mu}}, \boldsymbol{X})$ relies on $\boldsymbol{X}$, this is not true for the majority of situations. However, the breakdown point of location estimator $\boldsymbol{\mu}$ depending on $\boldsymbol{X}$ do exist (see Huber (1984)). The definition of breakdown value of scatter matrix $\boldsymbol{\Sigma}$ can be given similarly:

$$\epsilon^*(\hat{\boldsymbol{\Sigma}}, \boldsymbol{X}) = \min\left\{\frac{m}{n} : \sup \max_i|\lambda_i(\hat{\boldsymbol{\Sigma}}(\boldsymbol{X}_m)) - \lambda_i(\hat{\boldsymbol{\Sigma}}(\boldsymbol{X}))| = \infty\right\}. \tag{2.24}$$

where λ_i is the eigenvalue of $\hat{\boldsymbol{\Sigma}}$ and all eigenvalues are sorted as $0 \leq \lambda_p(\hat{\boldsymbol{\Sigma}}) \leq \cdots \leq \lambda_1(\hat{\boldsymbol{\Sigma}})$. From this definition, when the maximum of eigenvalues λ_1 becomes arbitrarily large or the minimum of eigenvalues λ_p becomes arbitrarily close to 0, a scatter estimator will be broken. These two cases are called explosion and implosion respectively. When both or any of them occur, the scatter matrix will be no longer robust. Lopuhaä and Rousseeuw (1991) showed that the breakdown point of any affine equivariant estimator is itself invariant under affine transformations, which can simplify the process of computing the breakdown point of an estimator.

Theorem 1. *Let $\boldsymbol{X}$ be $n \times p$ data matrix, and let $\hat{\boldsymbol{\mu}} \in \mathbb{R}^p$ and $\hat{\boldsymbol{\Sigma}} \in \mathbb{R}^{p\times p}$ be affine equivariant location and covariance estimates based on $\boldsymbol{X}$. Then*

1. $\epsilon^*(\hat{\boldsymbol{\mu}}, \boldsymbol{X}\boldsymbol{A} + \boldsymbol{\nu}) = \epsilon^*(\hat{\boldsymbol{\mu}}, \boldsymbol{X})$
2. $\epsilon^*(\hat{\boldsymbol{\Sigma}}, \boldsymbol{X}\boldsymbol{A} + \boldsymbol{\nu}) = \epsilon^*(\hat{\boldsymbol{\Sigma}}, \boldsymbol{X})$

where $\boldsymbol{\nu}$ is any $n \times p$ constant matrix and $\boldsymbol{A}$ is any nonsingular $p \times p$ matrix.

The proof of this theorem can be found in Lopuhaä and Rousseeuw (1991). This theorem shows that, with affine transformation to the data the breakdown point of affine equivariant estimator will be the same as the one computed from the original data. Davies (1987) showed that when the data matrix $\boldsymbol{X}$ is in general position, i.e., no $p+1$ points are contained in some hyperplane of dimension smaller than p, and if $n \geq p+1$, the breakdown value of any affine equivariant covariance estimator $\hat{\boldsymbol{\Sigma}}$ is at most $\lfloor (n-p+1)/2 \rfloor / n$. This means the covariance part might be broken when one substitute for $\lfloor (n-p+1)/2 \rfloor$ points or more by any other values, no matter what happens to the location estimator. But Rousseeuw (2005) found that for affine equivariant location estimators the upper bound on the breakdown point is also $\lfloor (n-p+1)/2 \rfloor / n$ under natural regularity conditions. Note that the limit $\lim_{n\to\infty} \epsilon_n^* = .5$ which reveals that the maximal breakdown point is 1/2 for affine equivariant estimators.

Rousseeuw (1985) proposed the minimum volume ellipsoid (MVE) estimator and proved it to be affine equivariant with breakdown point $(\lfloor n/2 \rfloor - p + 1)/n$ which is smaller than the upper bound $\lfloor (n-p+1)/2 \rfloor / n$. We will use a theorem about the breakdown point of MVE estimator to end this section. The detailed proof can be found in Lopuhaä and Rousseeuw (1991).

Theorem 2. *Let $\boldsymbol{X}$ be $n \times p$ data matrix in general position, and let $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ be the MVE estimates of location and scatter. Then*

1. *if $p = 1$, then $\epsilon^*(\hat{\boldsymbol{\mu}}, \boldsymbol{X}) = \lfloor (n+1)/2 \rfloor / n$ and $\epsilon^*(\hat{\boldsymbol{\Sigma}}, \boldsymbol{X}) = \lfloor n/2 \rfloor / n$.*

2. *if $p \geq 2$, then $\epsilon^*(\hat{\boldsymbol{\mu}}, \boldsymbol{X}) = \epsilon^*(\hat{\boldsymbol{\Sigma}}, \boldsymbol{X}) = \lfloor (n-p+1)/2 \rfloor / n$.*

2.4.3 Influence Function

The influence function serves as a crucial tool for assessing the robustness of an estimator. It is also used to calculate estimator's asymptotic variances and efficiencies. Hampel (1968, 1974) created the influence function as a means of examining the infinitesimal behavior of a robust estimator. Hampel et al. (1986) provided a description and analysis of it. The defined

standardized effect of an outlier on the estimator was based on its location at the point $\boldsymbol{x}$. Hampel et al. (1986) discussed the minuscule level of steadiness of an estimator. In an ideal situation, a robust estimator should have a bounded influence function.

We take MCD estimator as an example to briefly discuss how to compute influence function for an estimator. First consider the distribution with contamination.

$$F_{\varepsilon,\boldsymbol{x}} = (1-\varepsilon)F + \varepsilon\Delta_{\boldsymbol{x}}. \tag{2.25}$$

where F is any distribution of clean data, ε contamination rate usually set $0 \leq \varepsilon \leq .5$, $\Delta_{\boldsymbol{x}}$ is the cdf of a Dirac measure putting all its mass at $\boldsymbol{x} \in \mathbb{R}^p$. Then we can define the influence function for one parameter of a distribution at point $\boldsymbol{x}$.

$$IF(\boldsymbol{x}, K, F) = \lim_{\varepsilon\to 0} \frac{K(F_{\varepsilon,\boldsymbol{x}}) - K(F)}{\varepsilon} = \left.\frac{\partial K_{\varepsilon,\boldsymbol{x}}}{\partial\varepsilon}\right|_{\varepsilon=0}. \tag{2.26}$$

where function $K(F)$ is to obtain the parameter of the distribution F. As is shown by the definition, the influence function measures the sensitivity of the parameters to tiny amounts of contamination in the distribution.

Now we focus on the elliptically symmetric distribution F with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

$$f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x}) = |\boldsymbol{\Sigma}|^{-1/2} g\Big((\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\Big). \tag{2.27}$$

where $\boldsymbol{x}, \boldsymbol{\mu} \in \mathbb{R}^p$, $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ is any positive definite matrix, $(\cdot)'$ is the transpose of a matrix, $|\cdot|$ is the determinant of a matrix and the derivative of function g is strictly negative and it is assumed to be known.

The MCD estimator can be obtained by the following ellipsoid.

$$A(F_{\boldsymbol{\mu},\boldsymbol{\Sigma}}) = \big\{\boldsymbol{z} \in \mathrm{R} | (\boldsymbol{z}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu}) \leq q_\alpha\big\}. \tag{2.28}$$

where α is the significance level, $q_\alpha = M^{-1}(1-\alpha)$ and $M(t) = P_{F_{\mathbf{0},\boldsymbol{I}}}(\boldsymbol{z}'\boldsymbol{z} \leq t)$. The location estimator $\boldsymbol{\mu}$ and scatter estimator $\boldsymbol{\Sigma}$ can be equal to:

$$\begin{aligned}
\boldsymbol{\mu}_A(F_{\boldsymbol{\mu},\boldsymbol{\Sigma}}) &= \frac{\displaystyle\int_A \boldsymbol{x}\,\mathrm{d}F_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x})}{1-\alpha}, \\
\boldsymbol{\Sigma}_A(F_{\boldsymbol{\mu},\boldsymbol{\Sigma}}) &= c_\alpha \frac{\displaystyle\int_A \big(\boldsymbol{x}-\boldsymbol{\mu}_A(F_{\boldsymbol{\mu},\boldsymbol{\Sigma}})\big)\big(\boldsymbol{x}-\boldsymbol{\mu}_A(F_{\boldsymbol{\mu},\boldsymbol{\Sigma}})\big)'\,\mathrm{d}F_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x})}{1-\alpha}.
\end{aligned} \tag{2.29}$$

The $\boldsymbol{\Sigma}$ part of Equation (2.29) can be simplified as

$$\boldsymbol{\Sigma}_A(F_{\boldsymbol{\mu},\boldsymbol{\Sigma}}) = c_\alpha \left\{ \frac{\int_A \boldsymbol{x}\boldsymbol{x}' \, \mathrm{d}F_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x})}{1-\alpha} - \boldsymbol{\mu}_A(F_{\boldsymbol{\mu},\boldsymbol{\Sigma}})\boldsymbol{\mu}_A(F_{\boldsymbol{\mu},\boldsymbol{\Sigma}})' \right\}. \tag{2.30}$$

As the preceding section discussed, the MCD estimator is affine equivariant. Therefore, we only need to illustrate the process of obtaining the influence function based on the model distribution $F_{\mathbf{0},\boldsymbol{I}_p}$. So from now, we drop the subscripts for simplicity. Therefore, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ can be written in the same way.

$$\boldsymbol{\mu}_A(F) = \frac{\int_{A(F)} \boldsymbol{x} \, \mathrm{d}F(\boldsymbol{x})}{1-\alpha}, \qquad \boldsymbol{\Sigma}_A(F) = c_\alpha \frac{\int_{A(F)} \boldsymbol{x}\boldsymbol{x}' \, \mathrm{d}F(\boldsymbol{x})}{1-\alpha}. \tag{2.31}$$

Now consider the contaminated distribution $F_{\varepsilon,\boldsymbol{x}}$ at point $\boldsymbol{x}_0 \in \mathbb{R}^p$, we rewrite Equation (2.25) as

$$F_{\varepsilon,\boldsymbol{x}_0} = (1-\varepsilon)F + \varepsilon\Delta_{\boldsymbol{x}_0}. \tag{2.32}$$

For the distribution $F_{\varepsilon,\boldsymbol{x}_0}$, the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ will equal:

$$\begin{aligned} \boldsymbol{\mu}_A(F_{\varepsilon,\boldsymbol{x}_0}) &= \frac{\int_A \boldsymbol{x} \, \mathrm{d}F_{\boldsymbol{\varepsilon},\boldsymbol{x}_0}(\boldsymbol{x})}{1-\alpha}, \\ \boldsymbol{\Sigma}_A(F_{\boldsymbol{\varepsilon},\boldsymbol{x}_0}) &= c_\alpha \left\{ \frac{\int_A \boldsymbol{x}\boldsymbol{x}' \, \mathrm{d}F_{\boldsymbol{\varepsilon},\boldsymbol{x}_0}(\boldsymbol{x})}{1-\alpha} - \boldsymbol{\mu}_A(F_{\boldsymbol{\varepsilon},\boldsymbol{x}_0})\boldsymbol{\mu}_A(F_{\boldsymbol{\varepsilon},\boldsymbol{x}_0})' \right\}. \end{aligned} \tag{2.33}$$

Due to Equation (2.26), the influence function of scatter estimator $\boldsymbol{\Sigma}$ at point $\boldsymbol{x}_0$ is as follows.

$$IF(\boldsymbol{x}_0, \boldsymbol{\Sigma}, F_{\varepsilon,\boldsymbol{x}_0}) = \lim_{\varepsilon \to 0} \frac{\boldsymbol{\Sigma}(F_{\varepsilon,\boldsymbol{x}_0}) - \boldsymbol{\Sigma}(F)}{\varepsilon} \tag{2.34}$$

$$= \left. \frac{\partial \boldsymbol{\Sigma}_{\varepsilon,\boldsymbol{x}_0}}{\partial \varepsilon} \right|_{\varepsilon=0}. \tag{2.35}$$

The expression $\boldsymbol{\Sigma}_{\varepsilon,\boldsymbol{x}_0}$ can be simplified further using Equation (2.32). Let $\boldsymbol{\mu}_A(F_{\varepsilon,\boldsymbol{x}_0}) = \boldsymbol{\mu}_\varepsilon$

$$\begin{aligned} \boldsymbol{\Sigma}_{\varepsilon,\boldsymbol{x}_0} &= \boldsymbol{\Sigma}_A(F_{\boldsymbol{\varepsilon},\boldsymbol{x}_0}) \\ &= c_\alpha \left\{ \frac{\int_{A(F_{\boldsymbol{\varepsilon},\boldsymbol{x}_0})} \boldsymbol{x}\boldsymbol{x}' \, \mathrm{d}F_{\boldsymbol{\varepsilon},\boldsymbol{x}_0}(\boldsymbol{x})}{1-\alpha} - \boldsymbol{\mu}_A(F_{\varepsilon,\boldsymbol{x}_0})\boldsymbol{\mu}_A(F_{\varepsilon,\boldsymbol{x}_0})' \right\} \end{aligned} \tag{2.36}$$

$$= c_\alpha \Bigg\{ \frac{1-\varepsilon}{1-\alpha} \int\limits_{A(F_{\varepsilon,\boldsymbol{x}_0})} \boldsymbol{x}\boldsymbol{x}' \,\mathrm{d}F(\boldsymbol{x}) + \frac{\varepsilon}{1-\alpha} \mathbf{I}\Big(\boldsymbol{x}_0 \in A(F_{\varepsilon,\boldsymbol{x}_0})\Big) \boldsymbol{x}_0\boldsymbol{x}_0' - \boldsymbol{\mu}_\varepsilon \boldsymbol{\mu}_\varepsilon' \Bigg\}.$$

Hence, we plug the last equation into Equation (2.35) to calculate the derivative of $\boldsymbol{\Sigma}_{\varepsilon,\boldsymbol{x}_0}$.

$$\begin{aligned} IF(\boldsymbol{x}_0, \boldsymbol{\Sigma}, F_{\varepsilon,\boldsymbol{x}_0}) &= \frac{\partial \boldsymbol{\Sigma}_{\varepsilon,\boldsymbol{x}_0}}{\partial \varepsilon}\Big|_{\varepsilon=0} \\ &= c_\alpha \Bigg\{ -\frac{1}{1-\alpha} \int_{A(F)} \boldsymbol{x}\boldsymbol{x}' \,\mathrm{d}F(\boldsymbol{x}) + \frac{1}{1-\alpha} \frac{\partial}{\partial \varepsilon} \int_{A(F_{\boldsymbol{\varepsilon},\boldsymbol{x}_0})} \boldsymbol{x}\boldsymbol{x}' \,\mathrm{d}F(\boldsymbol{x})\Big|_{\varepsilon=0} \\ &\quad + \frac{1}{1-\alpha} \mathbf{I}\Big(\boldsymbol{x}_0 \in A(F)\Big) \boldsymbol{x}_0\boldsymbol{x}_0' \Bigg\}. \end{aligned} \tag{2.37}$$

Due to the Fisher consistency of $\boldsymbol{\Sigma}$, we find $c_\alpha = \frac{1-\alpha}{\int_{z \le q_\alpha} \boldsymbol{z}\boldsymbol{z}' \,\mathrm{d}F(\boldsymbol{x})}$. And using the fact that $A(F) = \{\boldsymbol{z} \in \mathbb{R}^p | \boldsymbol{z}'\boldsymbol{z} \le q_\alpha\}$ together, we have

$$\begin{aligned} IF(\boldsymbol{x}_0, \boldsymbol{\Sigma}, F_{\varepsilon,\boldsymbol{x}_0}) = -\boldsymbol{I} + \frac{c_\alpha}{1-\alpha} \frac{\partial}{\partial \varepsilon} \int\limits_{A(F_{\varepsilon,\boldsymbol{x}_0})} \boldsymbol{x}\boldsymbol{x}' \,\mathrm{d}F(\boldsymbol{x})\Big|_{\varepsilon=0} \\ + \frac{c_\alpha}{1-\alpha} \mathbf{I}(\|\boldsymbol{x}_0\|^2 \le q_\alpha) \boldsymbol{x}_0\boldsymbol{x}_0'. \end{aligned} \tag{2.38}$$

After that, for the second term of right hand side of Equation (2.38), We make a transformation. $\boldsymbol{x}$ is transformed to $\boldsymbol{y} = \boldsymbol{\Sigma}_\varepsilon^{-1/2}(\boldsymbol{x} - t_\varepsilon)$. Now the domain of the integral becomes a ball under the center at the origin and radius $\sqrt{q_\alpha}$. For convenience, we use polar coordinates to compute the integral part. We omit some lengthy computations and give the final result here. The detailed process of how to calculate the second and third term on the right side of Equation (2.38) can be found in Croux and Haesbroeck (1999).

$$\begin{aligned} IF(\boldsymbol{x}, \boldsymbol{\Sigma}_{ii}, F_{\varepsilon,\boldsymbol{x}}) = \frac{1}{b_1} \Bigg\{ \frac{c_\alpha}{1-\alpha} \boldsymbol{x}_i^2 \mathbf{I}(\|\boldsymbol{x}^2\| \le q_\alpha) + \frac{b_2}{b_1 - pb_2} \frac{c_\alpha}{1-\alpha} \|\boldsymbol{x}^2\| \mathbf{I}(\|\boldsymbol{x}^2\| \le q_\alpha) \\ + \frac{b_1}{b_1 - pb_2} \Bigg[\frac{c_\alpha}{1-\alpha} \frac{q_\alpha}{p} (1 - \alpha - \mathbf{I}(\|\boldsymbol{x}^2\| \le q_\alpha)) - 1 \Bigg] \Bigg\}, \end{aligned} \tag{2.39}$$

$$IF(\boldsymbol{x}, \boldsymbol{\Sigma}_{ij}, F_{\varepsilon,\boldsymbol{x}}) = -\frac{\boldsymbol{x}_i \boldsymbol{x}_j}{2c_3} \mathbf{I}(\|\boldsymbol{x}^2\| \le q_\alpha), \qquad \text{if} \quad i \ne j.$$

where the constants b_1, b_2, c_2, c_3, c_4 are given below.

$$c_2 = \frac{\pi^{p/2}}{\Gamma(p/2+1)} \int_0^{\sqrt{q_\alpha}} r^{p+1} g'(r^2) \,\mathrm{d}r,$$

$$c_3 = \begin{cases} \frac{\pi^{p/2}}{(p+2)\Gamma(p/2+1)} \int_0^{\sqrt{q_\alpha}} r^{p+3} g'(r^2)\,\mathrm{d}r, & \text{if} \quad p \geq 2 \\ 0, & \text{otherwise,} \end{cases}$$

$$c_4 = \frac{3\pi^{p/2}}{(p+2)\Gamma(p/2+1)} \int_0^{\sqrt{q_\alpha}} r^{p+3} g'(r^2)\,\mathrm{d}r,$$

$$b_1 = \frac{c_\alpha(c_3 - c_4)}{1-\alpha},$$

$$b_2 = \frac{1}{2} + \frac{c_\alpha}{1-\alpha}\left[c_3 - \frac{q_\alpha}{p}(c_2 + \frac{1-\alpha}{2})\right].$$

Finally, for location part of MCD estimator, it is simplier. Following the same way of getting $IF(\boldsymbol{x}, \boldsymbol{\Sigma}, F_{\varepsilon,\boldsymbol{x}})$, we get

$$IF(\boldsymbol{x}, \boldsymbol{\mu}, F_{\varepsilon,\boldsymbol{x}}) = \left(-\frac{2}{1-\alpha} \int_{\boldsymbol{z}\boldsymbol{z}' \leq q_\alpha} \boldsymbol{z}\boldsymbol{z}' g'(\boldsymbol{z}'\boldsymbol{z})\,\mathrm{d}\boldsymbol{z} \right)^{-1} \frac{\boldsymbol{x}}{1-\alpha} \mathbf{I}(\|\boldsymbol{x}\|^2 \leq q_\alpha). \tag{2.40}$$

Chapter 3

Application for a Multivariate Dataset

3.1 Data Set *seeds*

In this chapter, a multivariate data set is applied to compare the performances of DetMCD, PP MCD and FASTMCD methodologies. The data set is obtained from Charytanowicz et al. (2010). It consists of three different groups of grains. The examined group comprised kernels belonging to three different varieties of wheat: Kama, Rosa and Canadian, 70 elements each, randomly selected for the experiment (Charytanowicz et al., 2010). A high-quality visualization of the internal kernel structure was observed through the utilization of a soft X-ray methodology. The data was gathered by taking measurements of seven geometric parameters of 210 wheat kernel samples, including area, perimeter, compactness, length and width of the kernel, asymmetry coefficient, and length of the kernel groove. All of the aforementioned parameters were continuous and real-valued.

The last attribute in the original data set explains the type of wheat planting. In this application, the last two types of wheats are selected to perform outlier detection. In order to show how different all the measurements between the two varieties are, we list the summary statistics of all the variables for two varieties in Table 3.1. As is shown, there exists a marked distinction within the two corresponding variables in the two groups. Upon integrating observations from one group to another, the newly amalgamated data set may indicate the presence of outliers, which serves as the foundation for conducting subsequent analyses and evaluations of various methodologies.

Table 3.1: Summary statistics for each variable of two types of wheats

Panel A: Summary statistics for Group 1 (the first 70 observations)						
Area	Perimeter	Compactness	Length	Width	Coefficient	Groove
Min. :10.59	Min. :12.41	Min. :0.3081	Min. :4.899	Min. :2.630	Min. :1.661	Min. :4.745
1st Qu.:11.26	1st Qu.:13.00	1st Qu.:0.3340	1st Qu.:5.136	1st Qu.:2.725	1st Qu.:4.049	1st Qu.:5.002
Median :11.84	Median :13.25	Median :0.3493	Median :5.224	Median :2.834	Median :4.839	Median :5.091
Mean :11.87	Mean :13.25	Mean :0.3494	Mean :5.230	Mean :2.854	Mean :4.788	Mean :5.116
3rd Qu.:12.43	3rd Qu.:13.47	3rd Qu.:0.3618	3rd Qu.:5.324	3rd Qu.:2.967	3rd Qu.:5.467	3rd Qu.:5.229
Max. :13.37	Max. :13.95	Max. :0.3977	Max. :5.541	Max. :3.232	Max. :8.456	Max. :5.491
Panel B: Summary statistics for Group 2 (the last 70 observations)						
Min. :15.38	Min. :14.66	Min. :0.8452	Min. :5.363	Min. :3.231	Min. :1.472	Min. :5.144
1st Qu.:17.33	1st Qu.:15.74	1st Qu.:0.8725	1st Qu.:5.979	1st Qu.:3.554	1st Qu.:2.845	1st Qu.:5.878
Median :18.72	Median :16.21	Median :0.8826	Median :6.149	Median :3.693	Median :3.610	Median :5.981
Mean :18.33	Mean :16.14	Mean :0.8835	Mean :6.148	Mean :3.677	Mean :3.645	Mean :6.021
3rd Qu.:19.14	3rd Qu.:16.56	3rd Qu.:0.8982	3rd Qu.:6.312	3rd Qu.:3.805	3rd Qu.:4.436	3rd Qu.:6.188
Max. :21.18	Max. :17.25	Max. :0.9108	Max. :6.675	Max. :4.033	Max. :6.682	Max. :6.550

3.2 Comparisons and Discussion

As previously stated, the data set *seeds* contains two distinct groups. We consider the first group which are the first 70 observations as "pure" data. The remaining observations in our data set is then outliers. To make comparisons, we initially select bulk purity as a key feature to focus on. In simple terms, bulk purity can be described as the proportion of non-outliers found in the optimal set of the raw estimator. The higher the degree of purity, the better the estimator's quality. The bulk size h is commonly set at $\lfloor (n+p+1)/2 \rfloor$. In a systematic manner, I incorporate the observations of Group 2 into "pure" data set, creating a total of 70 distinct samples with varying sizes ranging from $n = 71$ to 140. Each sample had 70 uncontaminated data points in addition to a number of outliers, with the total amount of outliers being m equivalent to $n - 70$. Since DetMCD method lacks of affine equivance, all the data points are scaled under sphering transformation (Pokojovy and Jobe, 2022).

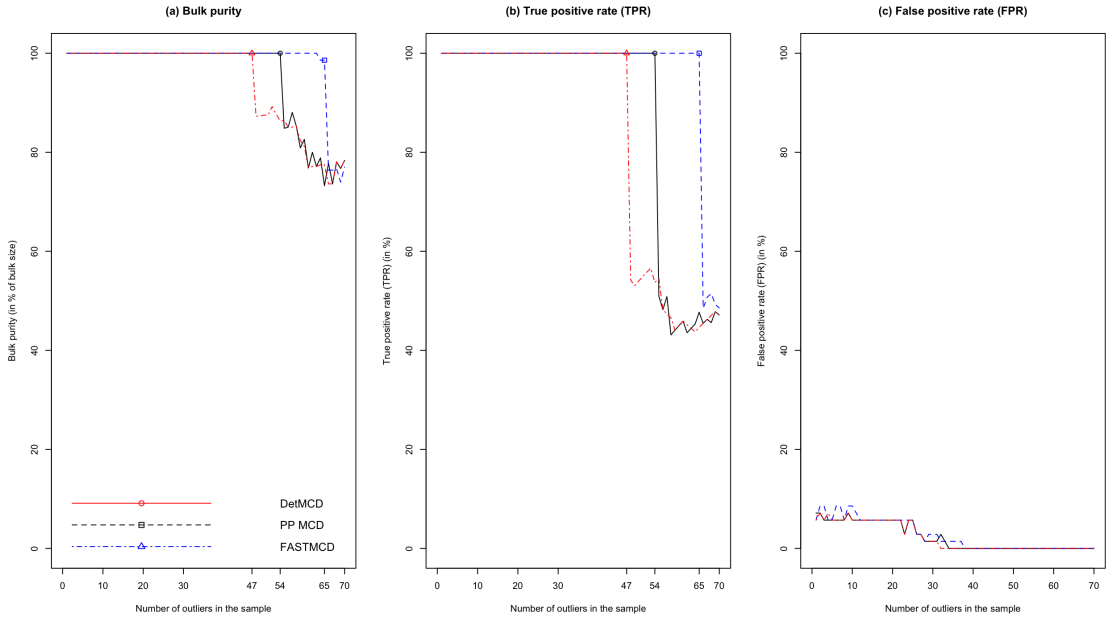

Figure 3.1: Bulk Purity (left), True positive rate (TPR) (middle) and False positive rate (FPR) (right) plots for three methods

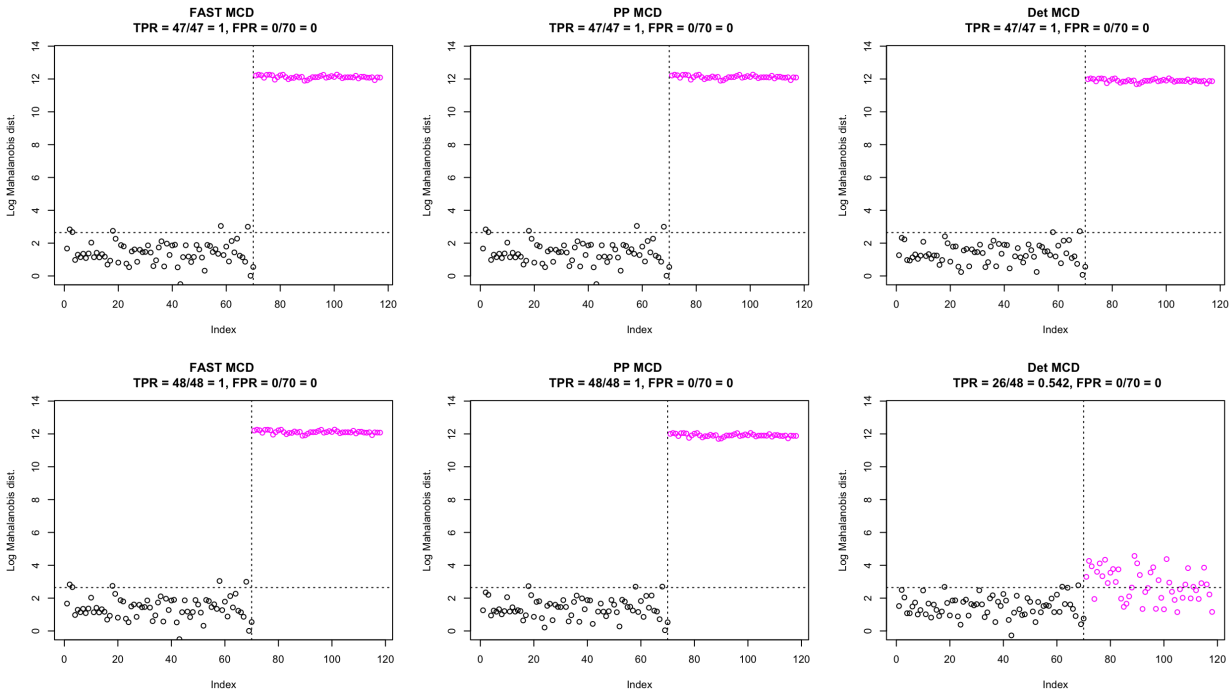

Figure 3.2: Logarithmic Mahalanobis distance plots for $n = 47$ (top) and $n = 48$ (bottom)

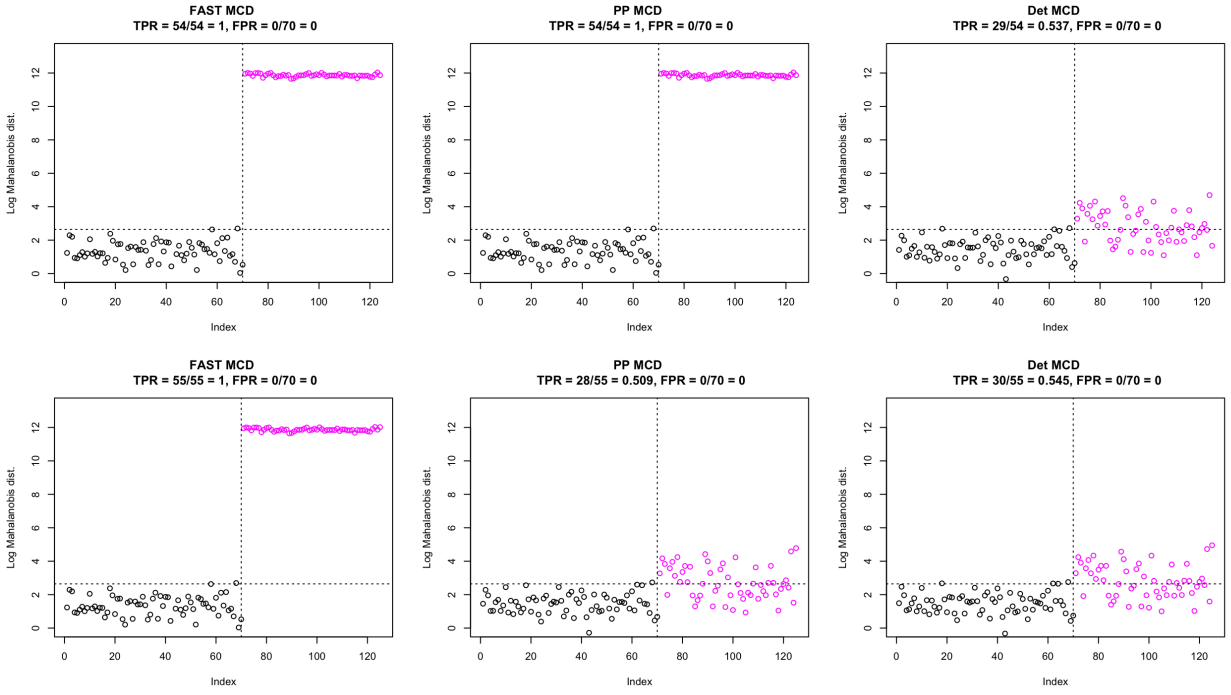

Figure 3.3: Logarithmic Mahalanobis distance plots for $n = 54$ (top) and $n = 55$ (bottom)

Figure 3.1 illustrates the relationship between the purity percentages and the number of outliers m present in the sample. It indicates that the FASTMCD method exhibits comparable or better performance than the other two approaches in terms of purity level for each of the 70 samples. The superiority of the FASTMCD method based on purity percentages becomes apparent when the number of outliers reaches 65 or above, surpassing the efficacy of the two alternative methods. DetMCD displays the poorest performance among the methods, as its bulk purity significantly declines in the presence of over 47 outliers within the sample. And the number of outliers making PP MCD not very robust is at least 55.

Figures 3.2, 3.3 and 3.4 illustrate significant changes of the accuracy of detecting outliers for three methods based on true positive rate (TPR) and false positive rate (FPR) as the number of outliers varies in datasets containing 47, 54, and 65 outliers. Before this, the threshold of Mahalanobis distances d_i^2 for three methods to distinguish outliers and non-

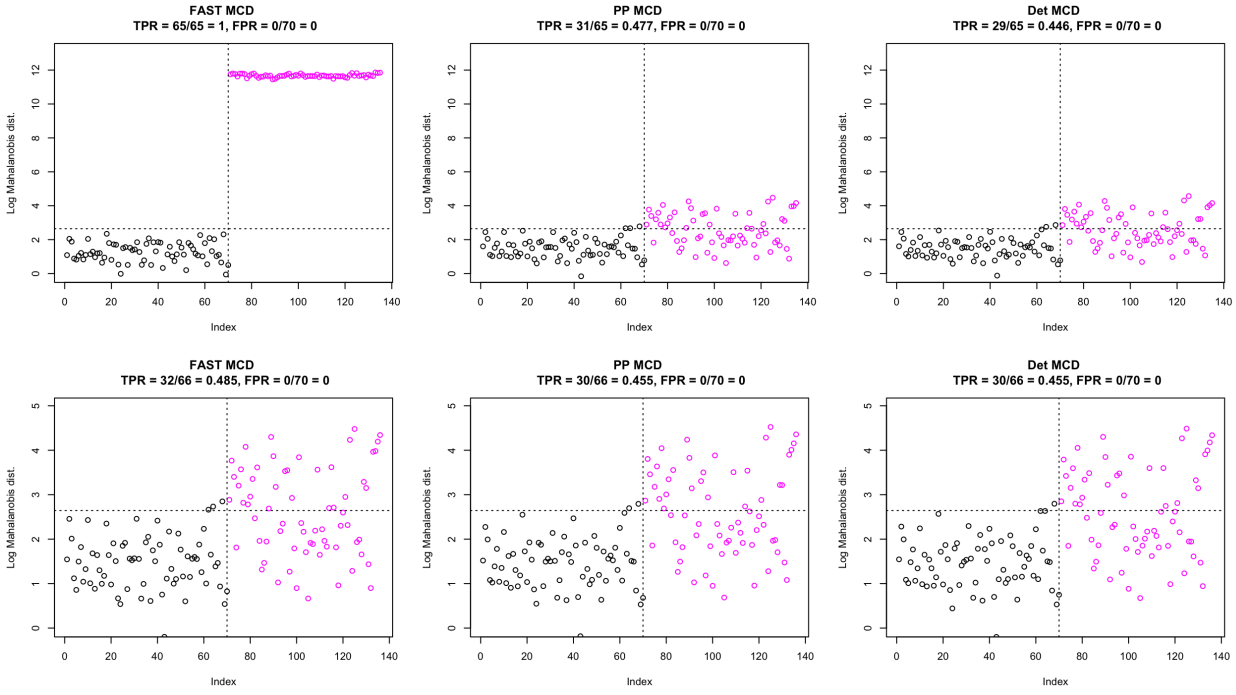

Figure 3.4: Logarithmic Mahalanobis distance plots for $n = 65$ (top) and $n = 66$ (bottom)

outliers should be determined. We performed a simulation to find the cutoff value. In the simulation, significance level 0.05 was chosen and for each method, 5,000 sets of n d_i^2 were generated under the setting of $\mathcal{N}_7(\mathbf{0}, \boldsymbol{I})$ distribution. We proceeded to choose a singular d_i^2 randomly and considered the 95th percentile from these 5,000 sets as our critical value.

Combing Figures 3.1 and 3.2, we find that the TPR values remain constant across all methods when the sample size contains 47 or fewer outliers. However, it is observed that if the count of outliers exceeds 47, the true positive rate (TPR) of DetMCD exhibits a decline. On the other hand, PP MCD and FASTMCD still shows a consistent TPR even under such circumstances. Figures 3.1 and 3.3 demonstrate that TPR of the FASTMCD remains on par with that of the PP MCD, but surpasses that of the DetMCD until the value of m is more than 54. If there are 55 outliers in the sample, FASTMCD will have a significantly higher TPR compared to DetMCD and PP MCD. Figures 3.1 and 3.4 present that TPR of FASTMCD outperforms the other two techniques. Nonetheless, the FASTMCD technique fails to operate as soon as the number of anomalies exceeds 65. To put it briefly, after considering the level of bulk purity and the ability to detect outliers, it can be concluded that the FASTMCD approach is superior to both PP MCD and DetMCD methods when identifying anomalies in a multivariate dataset such as *seeds*.

Chapter 4

Application for a High-Dimensional Dataset

4.1 Data Set *characterA*

In the previous chapter, we examined various techniques for identifying outliers within a multivariate data set. Specifically, the data set we previously utilized is low-dimensional. Here a high-dimensional data set is characterized by having a large number of variables (p), that are equivalent to or exceed the number of observations (n). Although traditional statistical methods exhibit strong performance in data sets with low dimensions, they encounter major obstacles when p exceeds n. The analysis of high-dimensional data poses various challenges for traditional statistical methods. One significant issue is that the scatter matrix is found to be non-invertible. Classical statistical methods do not allow the determination of the inverse of the scatter matrix. Another challenge is computation time. The increase in computation time is more pronounced as p increases compared to n . The magnitude of this increment is significant enough to question the viability of conventional statistical methods when employed on high-dimensional data. Therefore, three robust methods applicable to high dimensions are chosen to evaluate their effectiveness on a high-dimensional data set.

The data set used for this chapter is from R package *mrfDepth*. *characterA* consists of trajectories of the tip of a pen while writing the letter *a*. All samples are from the same writer. This collection of data consists of 100 instances, each with 171 values in two distinct dimensions. These three dimensions refer to time, the count of observations, and the coordinates in the X and Y axes, correspondingly. We divide the data into two sets, namely

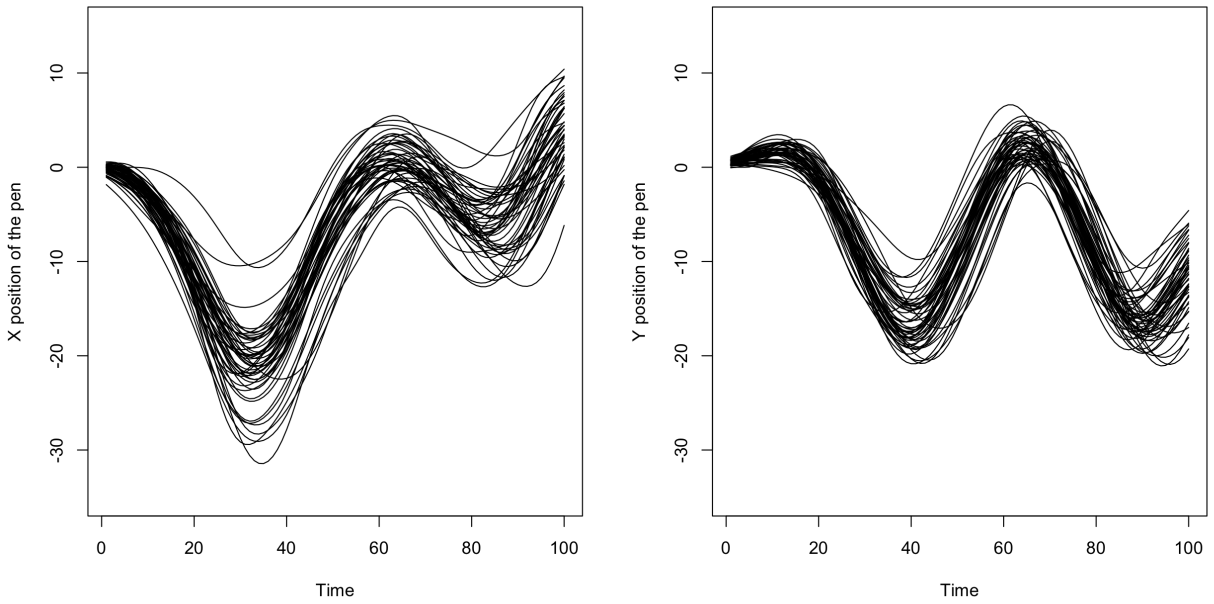

Figure 4.1: X trajectory (left) and Y trajectory (right) of the pen when writing letter a

X and Y. They are made up of X and Y coordinates, separately. For simplicity, We only select the initial 50 entries from both the X and Y datasets. To ensure that each observation is independent, the two data matrices are transposed and then combined. Therefore, we now have a dataset with 100 entries representing different paths of writing the letter a, and 100 separate columns indicating the time of writing process.

Figure 4.1 separately illustrates the shape of X and Y coordinates for the first and last 50 observations of the data matrix. As demonstrated it depicts a noticeable disparity between the X and Y coordinates at a specific time point, indicating a shared characteristic of the data presented in Chapter 3. To carry out the comparisons, the initial 50 observations are regarded as "clean" data whereas the left observations are recognized as anomalies. Hence, when we gradually inject anomalies into "clean" data, the resulting data set always maintains a high-dimensional nature.

4.2 Comparisons and Discussion

For comparisons of high-dimensional data, we choose three methods OGK, PCOut and MRCD. Both PCOut and MRCD techniques were proposed in the last decade and are quite new robust statistical methods for outlier detection. As previously stated, we adopt three measurements of bulk purity, true positive rate and false positive rate to evaluate the efficiencies for three methods in the high-dimensional context. Likewise, we initially create 50 sets of various sizes $n = 51, \ldots, 100$ from the available pool of 100 observations. Every single sample comprises of 50 "clean" data points and extra anomalies of amount $m = n - 50$. For almost all of the samples, the dimensionality $p = 100$ is always larger than the sample size n increasing from 51 to 100. Figure 4.2 displays three plots of bulk purity, TPR and FPR for three methods. It is evident that the OGK technique begins to experience a significant reduction in its purity level in the very early time. Two other techniques are capable of wit-

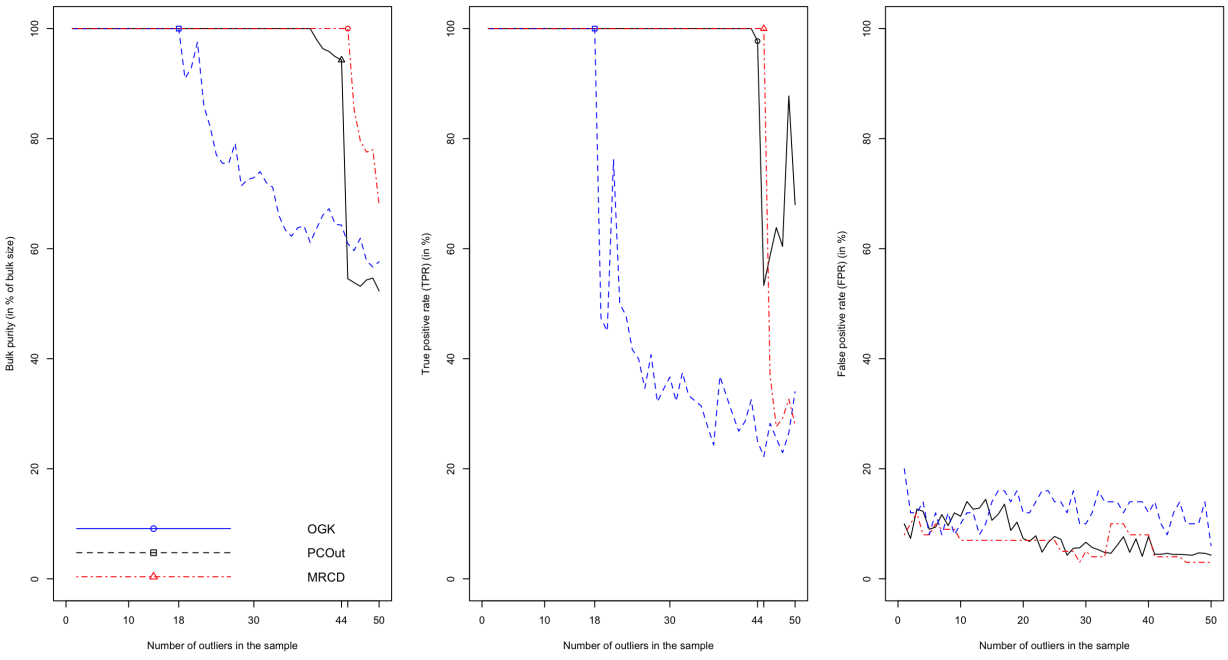

Figure 4.2: Bulk Purity (left), True positive rate (TPR) (middle) and False positive rate (FPR) (right) plots for three methods

hstanding the presence of up to even around 45 outliers in the sample. The performances of PCOut and MRCD methods are strikingly similar. Their "breakthrough" points making methods less effective are close, which were found to be 44 and 45, respectively.

Next, we analyze outlier detection performances of the three methods on important status-changing points. The similar simulation methods in Chapter 3 are utilized to determine the cutoff values. The results are presented in Figures 4.3, 4.4 and 4.5. The threshold is represented by the horizontal line in each of the figures. Any data points exceeding the threshold will be identified as outliers. The vertical line is used to separate the "clean" data and the outliers added from outside. Figures 4.2 and 4.3 depict as long as there are 18 or fewer outliers in the sample, all three techniques exhibit identical true positive rates (TPR). If one more outliers are being included, the OGK technique will no longer be able to identify the actual outliers. Figures 4.2 and 4.4 show that when the number of outliers being added increase to 45, The TPR decreases to .5. Prior to the occurrence of 44 outliers within the sample, it can be observed that the TPR of the PCOut method began to diminish. This suggested the reason that the TPR was at only approximately 88% before the status changed. In the scenario of 46 outliers being added, MRCD method breaks down as is illustrated in Figure 4.5. To sum up, the MRCD method shows the most reliable performance in contrast to the PCOut and OGK methodologies. The OGK technique displays a limited ability to handle even small numbers of outliers present in the data.

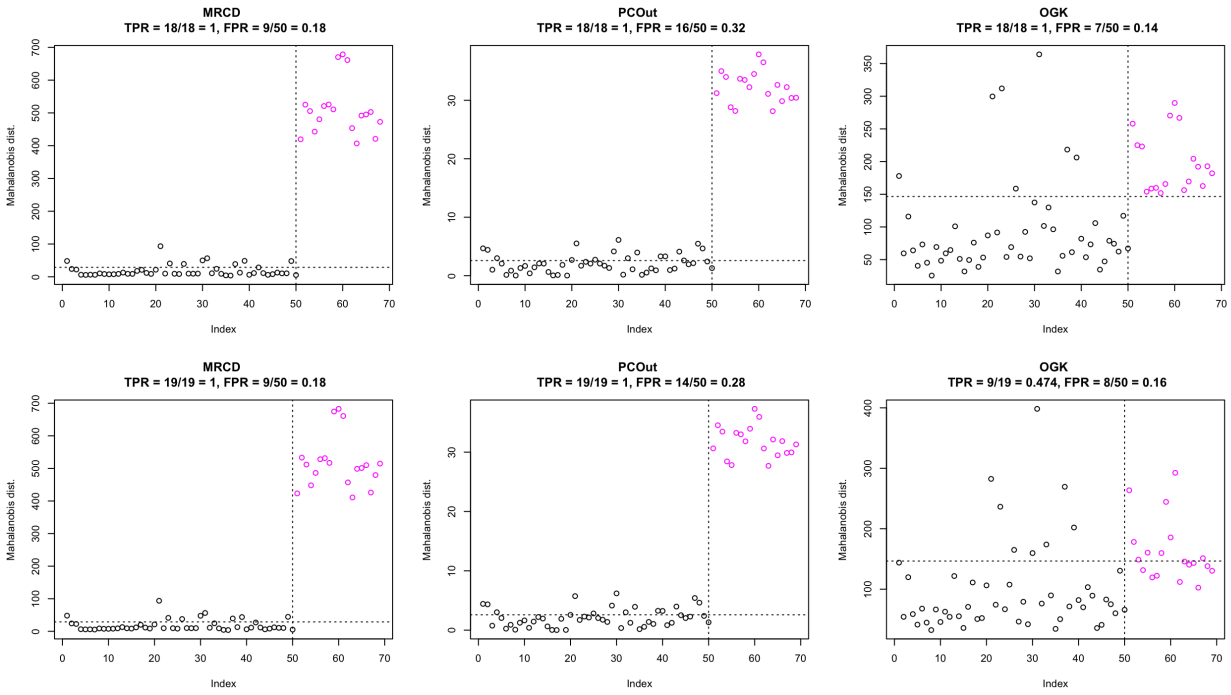

Figure 4.3: Mahalanobis distance plots for $n = 18$ (top) and $n = 19$ (bottom)

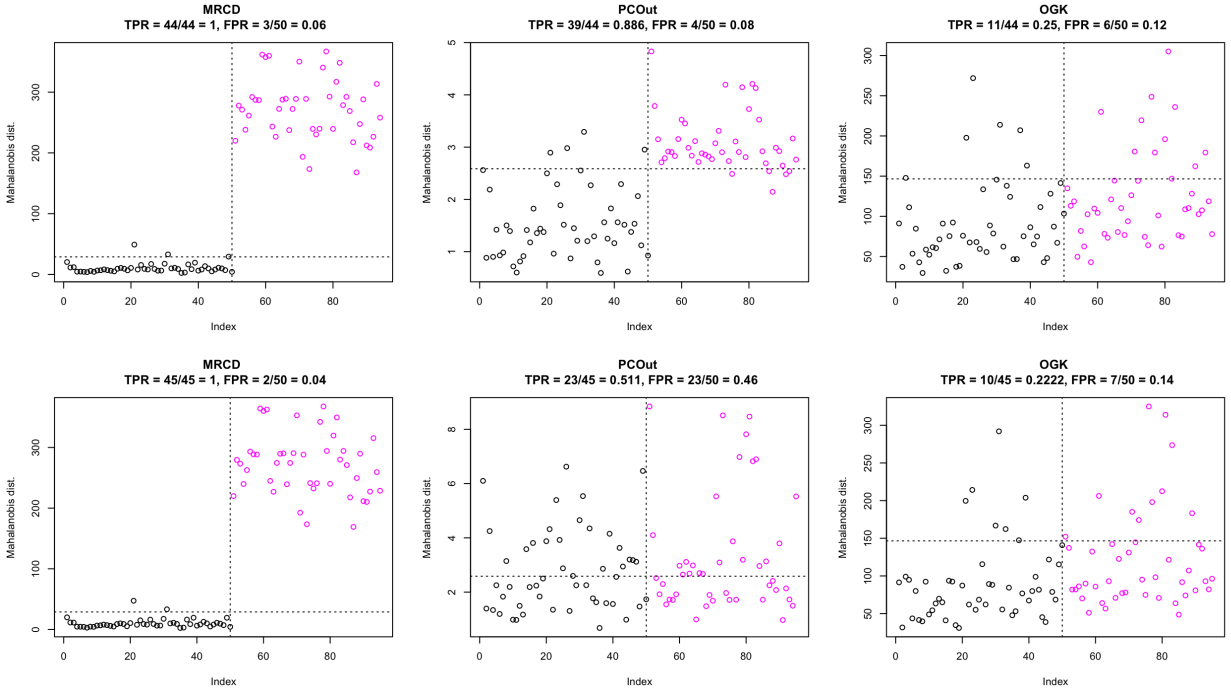

Figure 4.4: Mahalanobis distance plots for $n = 44$ (top) and $n = 45$ (bottom)

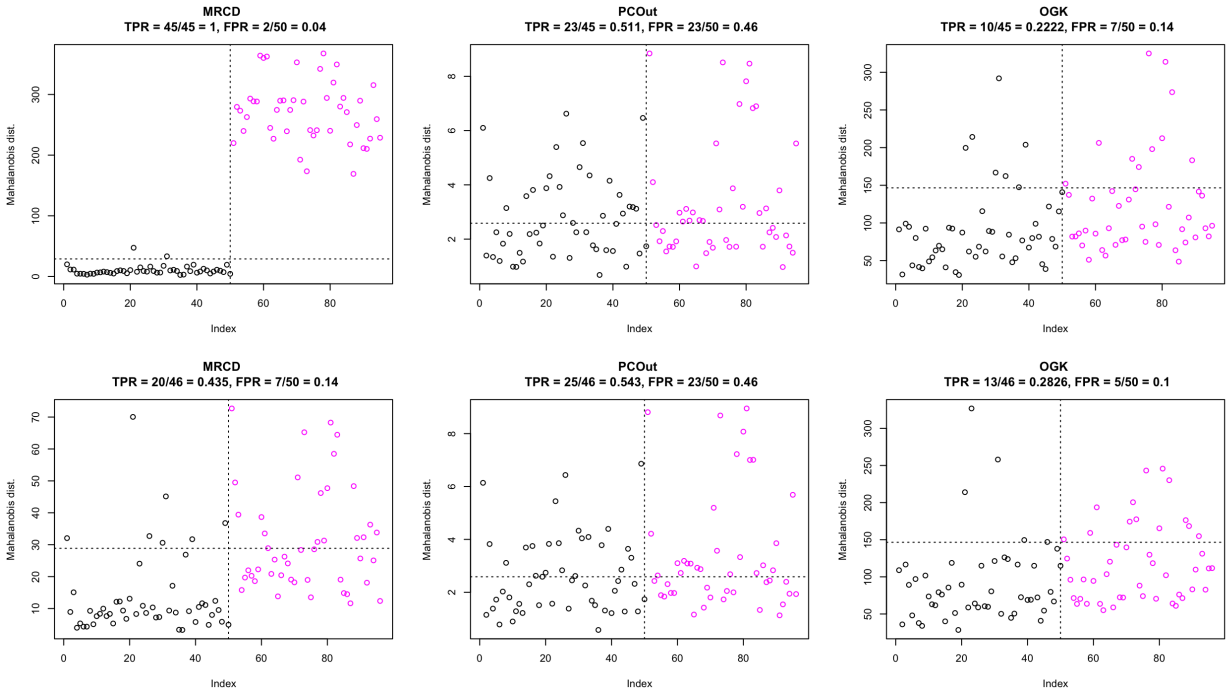

Figure 4.5: Mahalanobis distance plots for $n = 45$ (top) and $n = 46$ (bottom)

Chapter 5

Conclusions

We provided a overview of current robust statistical methods for outlier detection. Two application examples were presented with the purpose of comparing various methods. Assessing the effectiveness of FASTMCD, PP MCD, and DetMCD techniques on a low-dimensional data set, FASTMCD exhibited the best results out of the three methods. The application of DetMCD to the *seeds* data set yielded the least favorable performance. In a high-dimensional scenario, we conducted a comparison of MRCD, OGK, and PCOut methodologies. The results demonstrated that MRCD outperformed both OGK and PCOut. But the difference between MRCD and PCOut approaches was almost negligible. The OGK method was not as robust as we expected. It suffered a breakdown when confronted with a data set containing a mere 19 outliers.

References

Ammann, L. P. (1993). Robust singular value decompositions: A new approach to projection pursuit. *Journal of the American Statistical Association*, 88(422):505–514.

Andrews, D. F. and Hampel, F. R. (2015). *Robust Estimates of Location: Survey and Advances*. Princeton University Press.

Beaton, A. E. and Tukey, J. W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, 16(2):147–185.

Ben-Gal, I. (2005). Outlier detection. In *Data Mining and Knowledge Discovery Handbook*. Springer.

Bendre, S. and Kale, B. (1985). Masking effect on tests for outliers in exponential models. *Journal of the American Statistical Association*, 80(392):1020–1025.

Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604.

Boudt, K., Rousseeuw, P. J., Vanduffel, S., and Verdonck, T. (2020). The minimum regularized covariance determinant estimator. *Statistics and Computing*, 30(1):113–128.

Butler, R., Davies, P., and Jhun, M. (1993). Asymptotics for the minimum covariance determinant estimator. *The Annals of Statistics*, 21(3):1385–1400.

Campbell, N. A. (1980). Robust procedures in multivariate analysis I: Robust covariance estimation. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 29(3):231–237.

Charytanowicz, M., Niewczas, J., Kulczycki, P., Kowalski, P. A., Łukasik, S., and Żak, S.

(2010). Complete gradient clustering algorithm for features analysis of X-ray images. In *Information Technologies in Biomedicine*, pages 15–24. Springer.

Chiang, J.-T. et al. (2007). The masking and swamping effects using the planted mean-shift outliers models. *Int. J. Contemp. Math. Sciences*, 2(7):297–307.

Croux, C. and Haesbroeck, G. (1997). An easy way to increase the finite-sample efficiency of the resampled minimum volume ellipsoid estimator. *Computational Statistics & Data Analysis*, 25(2):125–141.

Croux, C. and Haesbroeck, G. (1999). Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *Journal of Multivariate Analysis*, 71(2):161–190.

Croux, C. and Haesbroeck, G. (2000). Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika*, 87(3):603–618.

Croux, C., Haesbroeck, G., and Rousseeuw, P. J. (2002). Location adjustment for the minimum volume ellipsoid estimator. *Statistics and Computing*, 12(3):191–200.

Davies, L. and Gather, U. (1993). The identification of multiple outliers. *Journal of the American Statistical Association*, 88(423):782–792.

Davies, P. L. (1987). Asymptotic behaviour of s-estimates of multivariate location parameters and dispersion matrices. *The Annals of Statistics*, 15(3):1269–1292.

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

Donoho, D. L. (1982). Breakdown properties of multivariate location estimators. PhD Qualifying paper, Department of Statistics, Harvard University, Boston.

Donoho, D. L. and Huber, P. J. (1983). The notion of breakdown point. In: Bickel P., Doksum K. and Hodges J. L. *A Festschrift for Erich Lehmann.* Belmont, Wadsworth.

Filzmoser, P., Maronna, R., and Werner, M. (2008). Outlier identification in high dimensions. *Computational Statistics & Data Analysis*, 52(3):1694–1711.

Gao, S., Li, G., and Wang, D. (2005). A new approach for detecting multivariate outliers. *Communications in Statistics—Theory and Methods*, 34(8):1857–1865.

Gnanadesikan, R. and Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 28(1):81–124.

Hampel, F. R. (1968). Contributions to the theory of robust estimation. Ph.D. thesis, University of California, Berkeley.

Hampel, F. R. (1971). A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, 42(6):1887–1896.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P., and Stahel, W. A. (1986). *Robust Statistics: the Approach Based on Influence Functions.* Wiley-Interscience, New York.

Hawkins, D. M. (1980). *Identification of Outliers.* Springer.

Hawkins, D. M. and McLachlan, G. J. (1997). High-breakdown linear discriminant analysis. *Journal of the American Statistical Association*, 92(437):136–143.

Hettich, S. and Bay, S. (1999). The UCI KDD Archive. `http://kdd.ics.uci.edu`.

Hodges Jr, J. L. (1967). Efficiency in normal samples and tolerance of extreme values for some estimates of location. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability.* University of California Press.

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

Huber, P. J. (1977). Robust covariances. In *Statistical Decision Theory and Related Topics*, pages 165–191. Elsevier.

Huber, P. J. (1984). Finite sample breakdown of m-and p-estimators. *The Annals of Statistics*, 12(1):119–126.

Huber, P. J. (1992). Robust estimation of a location parameter. In *Breakthroughs in Statistics: Methodology and Distribution*. Springer.

Hubert, M., Rousseeuw, P. J., and Van Aelst, S. (2008). High-breakdown robust multivariate methods. *Statistical Science*, 23(1):92–119.

Hubert, M., Rousseeuw, P. J., and Verdonck, T. (2012). A deterministic algorithm for robust location and scatter. *Journal of Computational and Graphical Statistics*, 21(3):618–637.

Iglewicz, B. and Martinez, J. (1982). Outlier detection using robust measures of scale. *Journal of Statistical Computation and Simulation*, 15(4):285–293.

Juan, J. and Prieto, F. J. (2001). Using angles to identify concentrated multivariate outliers. *Technometrics*, 43(3):311–322.

Kannan, K. S. and Manoj, K. (2015). Outlier detection in multivariate data. *Applied Mathematical Sciences*, 47(9):2317–2324.

Kirschstein, T., Liebscher, S., and Becker, C. (2013). Robust estimation of location and scatter by pruning the minimum spanning tree. *Journal of Multivariate Analysis*, 120:173–184.

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

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

Maronna, R., Bustos, O., and Yohai, V. (2006). Bias-and efficiency-robustness of general m-estimators for regression with random carriers. In *Smoothing Techniques for Curve Estimation: Proceedings of a Workshop held in Heidelberg, April 2–4, 1979*, pages 91–116. Springer.

Maronna, R. A. (1976). Robust M-estimators of multivariate location and scatter. *The Annals of Statistics*, 4(1):51–67.

Maronna, R. A., Martin, R. D., Yohai, V. J., and Salibián-Barrera, M. (2019). *Robust Statistics: Theory and Methods (with R)*. John Wiley & Sons.

Maronna, R. A. and Yohai, V. J. (1995). The behavior of the Stahel-Donoho robust multivariate estimator. *Journal of the American Statistical Association*, 90(429):330–341.

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

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

Penny, K. I. and Jolliffe, I. T. (2001). A comparison of multivariate outlier detection methods for clinical laboratory safety data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 50(3):295–307.

Pokojovy, M. and Jobe, J. M. (2022). A robust deterministic affine-equivariant algorithm for multivariate location and scatter. *Computational Statistics & Data Analysis*, 172:107475.

Pyke, R. (1965). Spacings. *Journal of the Royal Statistical Society: Series B (Methodological)*, 27(3):395–436.

Roelant, E., Van Aelst, S., and Willems, G. (2009). The minimum weighted covariance determinant estimator. *Metrika*, 70(2):177–204.

Ronchetti, E. M. and Huber, P. J. (2009). *Robust Statistics*. John Wiley & Sons Hoboken, NJ, USA.

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications*, 8(283-297):37.

Rousseeuw, P. J. (2005). Discussion of "breakdown and groups" by P. L. Davies and U. Gather. *The Annals of Statistics*, 33(3):1004–1009.

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

Rousseeuw, P. J. and Hubert, M. (2011). Robust statistics for outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):73–79.

Rousseeuw, P. J., Van Aelst, S., Van Driessen, K., and Gulló, J. A. (2004). Robust multivariate regression. *Technometrics*, 46(3):293–305.

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1).

Stahel, W. A. (1981). Robust estimation: Infinitesimal optimality and covariance matrix estimators. Ph.D. thesis, ETH, Zurich.

Taskinen, S., Croux, C., Kankainen, A., Ollila, E., and Oja, H. (2006). Influence functions and efficiencies of the canonical correlation and vector estimates based on scatter and shape matrices. *Journal of Multivariate Analysis*, 97(2):359–384.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. *Contributions to Probability and Statistics*, pages 448–485.

Walczak, B. and Massart, D. (1995). Robust principal components regression as a detection tool for outliers. *Chemometrics and Intelligent Laboratory Systems*, 27(1):41–54.

Zuo, Y. (2000). A note on finite sample breakdown points of projection based multivariate location and scatter statistics. *Metrika*, 51(3):259–265.

Zuo, Y. (2004). Projection-based affine equivariant multivariate location estimators with the best possible finite sample breakdown point. *Statistica Sinica*, 14(4):1199–1208.

Vita

Yuanhong Wu was born in China in 1997. He received his Bachelor of Science degree at Hubei University of Arts and Sciences, Hubei, China. After graduation, he went to USA to pursue a Master's degree in Statistics and Data Science at the University of Texas at El Paso (UTEP) from 2021 to 2023.

During his two years at UTEP, he worked as a tutor at Math Resource Center for Students (MaRCS) in the first year, helping students with Math classes. In the second year, he worked as a Teaching Assistant with several Professors. In March 2023, he was offered an assistantship to pursue a PhD degree at Fordham University. He will continue his studies in Computer Science Program at Fordham University in August 2023.