A New Contamination Model for Robust Estimation with Large High-Dimensional Data Sets

by

Fatemah Ali Alqallaf

B.A. (Mathematics) Kuwait University, 1994

M.Sc. (Statistics) University of British Columbia, 1999

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE STUDIES

Department of Mathematics
(Institute of Applied Mathematics)

we accept this thesis as conforming
to the required standard

The University of British Columbia

April 2003

Abstract

Data sets can be very large, highly multidimensional and of mixed quality. This thesis provides feasible and robust methods for estimating multivariate location and scatter matrix for such data. Our estimates scale well to very large sample sizes and dimensions and are resistant to the presence of multivariate outliers.

Statisticians use *contamination* or *mixture models* to study the performance of robust alternatives to classical statistical procedures. Most multivariate contamination models for numeric data proposed to date (see Hampel et al., 1986) assume that the majority of the observations comes from a nominal distribution such as a multivariate normal distribution, while the remainder comes from another multivariate distribution that generates outliers. We stress that such outliers could be "bad" data due to recording errors of all kinds, or they could be a highly informative subset of the data that leads to the discovery of unexpected knowledge in areas such as business operations, credit card fraud, and even the analysis of performance statistics of professional athletes. Unfortunately, the previously available models do not adequately represent reality for many multivariate data sets that arise in practice. It may often happen that outliers occur in each of the variables independently of the other variables or in special dependency patterns.

We introduce a new contamination model that overcomes the main drawbacks of the current models by taking into account different sources of variability in the data, and allowing greater flexibility. Moreover, our model permits for situations where extreme values of one or more variables (not necessarily outliers) may increase the likelihood of outliers or gross errors in other variables.

There is a large statistical literature on robust covariance and correlation matrix estimates, with an emphasis on affine equivariant estimates that possess high breakdown points and small worst case biases. All such estimates have unacceptable exponential complexity 2^p in the number of variables p. And one of the more attractive of these estimates, the Stahel-Donoho estimate, has an unacceptable quadratic complexity n^2 in the number of observations n. These estimates may be applied in large data applications with large p and n only by the use of adhoc sampling methods that render the robustness properties of the estimates unclear.

In this thesis we focus on *pairwise* robust scatter matrix estimates and *coordinate-wise* location estimates. The pairwise scatter estimates are based on coordinate-wise robust transformations (the quadrant correlation estimate, and the coordinate-wise Huberized estimates). We show that such estimates are computationally simple, and have attractive robustness properties under the existing and the newly proposed contamination models.

Contents

List of Tables

List of Figures

Acknowledgements

Praise God the Almighty for being merciful unto me and blessing me with his grace.

Many people helped me during the years I spent at UBC.

First of all, my gratitude goes to my supervisor, Dr. Ruben Zamar, for directing the course of my studies. The work on this thesis was carried out under his expert direction and supervision. Without his guidance and support, this thesis could not have been completed.

I have also been fortunate to have Dr. Paul Gustafson, Dr. Harry Joe, Dr. Raymond Ng and Dr. Bruno Zumbo on my PhD committee. Their input has added strength to this thesis. Thank you also to my university examiners (Dr. Laks Lakshmanan and Dr. Bertrand Clarke), my external examiner (Dr. Christopher Field) and the chair of my oral defence (Dr. Keith Head) for their comments and analysis. Dr. Edwin Perkins deserves special acknowledgement, since on many occasions he provided me with invaluable suggestions and advice.

I would like to acknowledge Kuwait university for providing me with the financial assistance needed to pursue my postgraduate studies in the form of a postgraduate scholarship.

I feel a deep sense of gratitude for my father Ali Alqallaf and my mother Parween Taqi who formed part of my vision and taught me the good things that really matter in life. I am grateful for all my family back home who always believed in us and unconditionally supported us all these years. Special thanks to my brothers and sisters (AbdulMohsen, Mohammad, Hani, Yousef, Hussein, Liala, Najlaa, Amna and Zainab). Of course, I would like to thank my two brothers in law AbdulReda and Khaled.

Other people who have helped me in different ways are Lindsey Turner for being such a good friend and colleague, Isabella Ghement who was always ready for a chat, Christine Graham and Lee Tran (what would we do without them?) and last but not least the entire Statistics Department for making me feel so at home.

I thank the Institute of Applied Mathematics for providing the necessary resources and to those people within the department for their friendship and interesting discussions. Thank you to the director Dr. Bernie Shizgal and to Dr. Roman Baranowski, the former research/IT manager.

I owe much more than what I can express here to my friends; Rosalía Aguirre-Hernández and Alberto Molina-Escobar.

Finally, no words can express my gratitude to my husband Mohammad Bahzad whose patient love enabled me to complete this work.

FATEMAH ALI ALQALLAF

The University of British Columbia
April 2003

I dedicate this thesis to my parents Ali Alqallaf and Parween Taqi, and to my husband Mohammad Bahzad—an outlier.

Chapter 1

Introduction

It is desirable to develop methods for extracting reliable and useful information from large high-dimensional data sets of mixed quality. Our thesis is that the existing contamination models used to represent high-dimensional data of mixed quality are not completely satisfactory. Therefore we propose a new contamination model and robust estimation procedures which are feasible/scalable to higher dimensions and appear to work relatively well regardless of the size, dimension and quality of the data set.

The focus of our work is on the robust estimation of multivariate location and scatter matrices (i.e. covariance and correlation matrices). These quantities are of great importance as they form the underpinnings of linear estimation theory.

We begin by discussing the role of robust estimations in statistics, and by providing some motivation for our work. Then, we provide a problem statement, a list of thesis contributions, and an outline of this thesis.

1.1 Robust Estimates

Statistics are extracted from data sets to infer properties of their underlying source distribution. Usually, these statistics are estimates of the parameters of the distribution. In the derivation of an estimate, modeling assumptions are made about the source distribution, e.g., i.i.d. (independent and identically distributed) data points or restriction to a particular parametric family of distributions. Those estimates which offer better performance usually make strict assumptions on the data; however, when these assumptions are invalid, the quality of the estimates can be quite poor. One aspect of robust statistics is to address the scenario where most, but not all, of the data points are drawn i.i.d. from a particular distribution. We wish to characterize this distribution. For example, consider the sample mean and median as estimates of the mean of a Gaussian distribution. The sample mean is the minimum variance estimate in this case, but it is not robust as it can be made arbitrarily bad by corrupting a single data point. The median, on the other hand, is very robust as at least 50% of the data points have to be corrupted to make the estimate arbitrarily bad; however, this robustness comes at the price of a significantly higher variance on the estimate. This example illustrates a fundamental trade-off, resistance versus efficiency. This leads to the primary objective in robust statistics the search for estimates which are not only resistant to model deviations but also perform well under the correct model. It should be emphasized that one should not infer that the ultimate goal of robust statistics is to ignore outlying data points. Such a naïve use of robust statistics could waste possible information contained in the outliers themselves. Robust estimates should merely reflect the bulk of the data points. Nonetheless, possessing a robust estimate often makes it easier to detect outliers which tend to be hidden in non-robust statistics. These outliers can then be separately analyzed for their own structure and information.

In the following section, we illustrate the dramatic effects that outliers can have on non-robust estimates.

1.1.1 Applications and Uses of Robust Estimates

Covariance and correlation matrices estimated from data sets are used for a variety of purposes. For example, pairwise sample correlation coefficients are often examined in an exploratory data analysis (EDA) stage to determine which variables are highly correlated with one another. Estimated covariance matrices are used as the basis for computing principal components for both general principal components analysis (PCA),

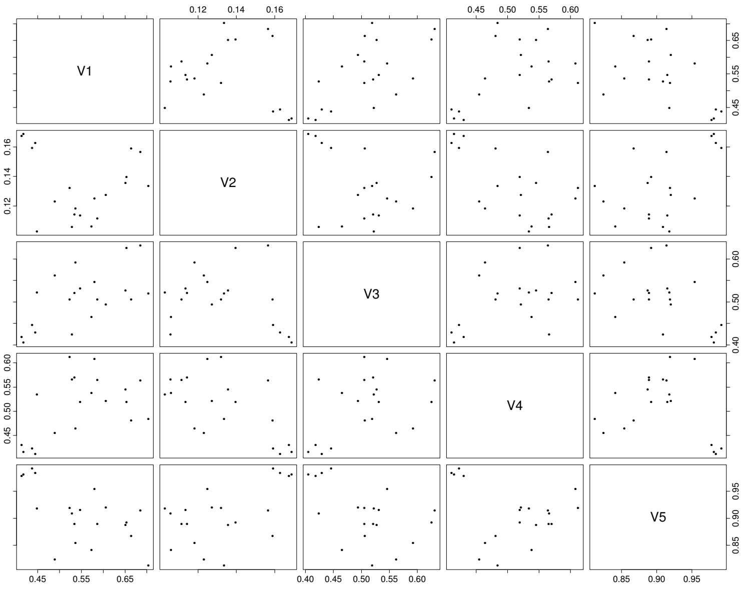
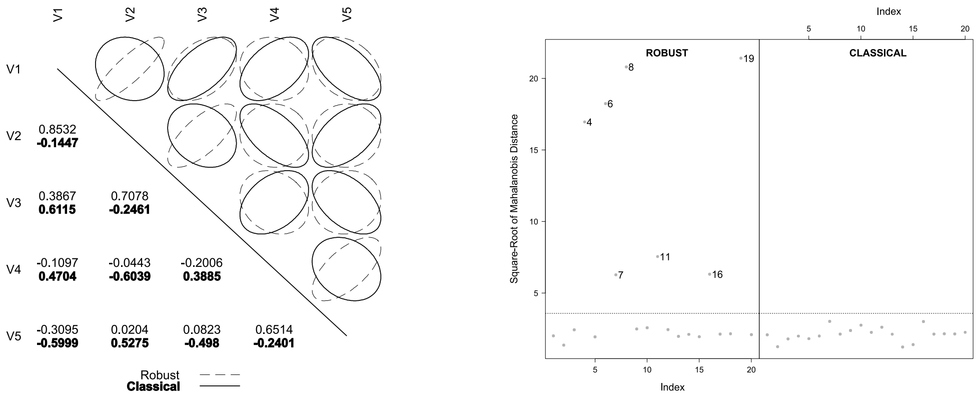


Figure 1.1: Woodmod 5-D Data with Outliers.

and for manual or automatic dimensionality reduction and variable selection. Estimated covariance matrices are also the basis for detecting multidimensional outliers through computation of the so-called Mahalanobis distances of the cases (rows) of a data set.

Unfortunately, the classical sample covariance and correlation matrix estimates, motivated by either Gaussian maximum likelihood or simple method of moments principles, are very sensitive to the presence of multidimensional outliers. Even a small fraction of outliers can distort these classical estimates to the extent that the estimates are very misleading, and virtually useless in any of the above applications. To cope with the problem of outliers, statisticians have invented robust methods that are not much influenced by outliers for a wide range of problems, including estimation of covariance and correlation matrices. We illustrate the extent to which outliers can distort classical correlation matrix estimates and the value of having a robust correlation matrix estimate with the small five-dimensional data set example illustrated in Figures 1.1 – 1.2

Figure 1.1 shows all pairwise scatter plots of the 5-dimensional data set called “Woodmod”. This data set clearly has at least several multidimensional outliers that show up as



(a) Classical and Robust Correlations.

(b) Classical and Robust Mahalanobis Distances with Square-Root 95% Chi-Squared Threshold.

Figure 1.2: Classical and Robust Correlations and Mahalanobis Distances for Woodmod Data.

a cluster in several of the scatterplots. Note that while these outliers are clearly outliers in two-dimensional space, they are not univariate outliers, i.e., they do not show up as well-detached outliers in any of the variables. Figure 1.2(a) shows the result of computing all pairwise classical correlations by both the classical method (sample correlation coefficients) and a particular robust method known as the Fast MCD (FMCD). The lower left triangle of values shows both the classical and robust correlation coefficient estimates, while the upper right triangle of ellipses visually represent the contours of a bivariate Gaussian density with zero means, unity variances, and correlation coefficients given by the classical and robust correlation coefficient estimates. A nearly circular ellipse indicates an estimated correlation coefficient of nearly zero. A narrow ellipse with its major axis oriented along the +45 degree (-45 degree) direction indicates a large positive (negative) estimated correlation coefficient. From the visual representation, we immediately see differences between the classical and robust correlations, sometimes very substantial differences, including changes of sign. For example, the classical correlation between V4 and V5 is -.24 whereas the robust correlation is +.65. The latter is quite consistent with what we might expect if we deleted the small cluster of outliers occurring in the scatterplot of V4 versus V5 in Figure 1.1.

A common way of detecting multidimensional outliers is to use the classical Mahalanobis distance:

$$d(\mathbf{x}_i) = (\mathbf{x}_i - \hat{\boldsymbol{\mu}})'\widehat{\mathbf{C}}^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}). \tag{1.1}$$

In the above expression $\mathbf{x}_i$ is the i-th data vector of dimension p (the transpose of the i-th row of the data set), $\hat{\boldsymbol{\mu}}$ is the vector of sample means of the variables (columns) of the data set, and $\widehat{\mathbf{C}}$ is the usual sample covariance matrix estimate. Under the assumption that the data is multivariate normal and that we use known values $\boldsymbol{\mu}$ and $\mathbf{C}$ in place of the above estimates, the $d(\mathbf{x}_i)$ would have chi-squared distribution with p degrees of freedom. With reasonably large sample sizes, the sample mean vector and sample covariance matrix will be close to their true values, and it is common practice to use the square-root of chi-squared (with p degrees of freedom) percent point such as .95 or .99 as a threshold to compare the square-root of $d(\mathbf{x}_i)$ with, and declare $\mathbf{x}_i$ to be an outlier if it exceeds this threshold. If we follow this classical approach for the Woodmod data of Figure 1.1, we get the results in the right-hand panel of Figure 1.2(b). The horizontal dashed line is the square-root of the 95% point of a chi-squared distribution with 5 degrees of freedom. Clearly, no points are declared outliers by the classical Mahalanobis distance approach. This is because the outliers have distorted the classical $\widehat{\mathbf{C}}$ so much that it does not produce reliable Mahalanobis distances. On the other hand, the left-hand panel of Figure 1.2(b), based on a robust $\widehat{\mathbf{C}}$ (and a robust $\hat{\boldsymbol{\mu}}$) result in detection of not only the cluster of four very large outliers evident on the scatterplots of Figure 1.1, but also three additional moderate-sized outliers.

The above example serves to vividly illustrate the inadequacy of classical correlation and covariance matrices in the presence of outliers and the valuable role of robust alternatives.

1.1.2 Robust Proposals of Scatter Estimates

Statistical literature contains a substantial number of papers proposing and studying the properties of robust scatter matrix estimation. An important early approach was that of M-estimates, first suggested by Hampel (1973), and studied by Maronna (1976) and Huber (1977, 1981). These estimates are positive definite, affine equivariant and relatively easy to compute, but have as a substantial limitation the fact that their breakdown point (BP) – i.e., the maximum proportion of outliers that the estimate can safely tolerate – is at most $1/p$ where p is the dimension of the data. This is not satisfactory, because it means that the breakdown point becomes smaller with increasing dimension, where there are more opportunities for outliers to occur.

Subsequently, there has been considerable emphasis on obtaining positive definite, affine equivariant estimates with a high breakdown point, namely a breakdown point of one-half. The best known is probably the minimum volume ellipsoid (MVE) estimate introduced by Rousseeuw (1984) and discussed by Rousseeuw and Leroy (1987) and Rousseeuw and Van Zomeren (1990). It consists of taking as location estimate the center of the smallest regular ellipsoid containing half the points of the data set. The scatter estimate is then defined by the shape matrix of that ellipsoid. However, Davies (1992) showed that the MVE estimate is not $\sqrt{n}$ consistent, making it less attractive for efficiency reasons. The MVE estimate has also been generalized to multivariate S-estimates (Davies, 1987; Lopuhaä, 1989; Lopuhaä and Rousseeuw, 1991). Rousseeuw (1985) introduced the minimum covariance determinant (MCD) estimate which has the normal rate of convergence. The MCD location and scatter estimates are the average and covariance matrix computed on that half of the data which attain the smallest determinant of their covariance matrix. Croux and Haesbroeck (1999) showed that MCD is more efficient than MVE in high-dimensions, and therefore recommend the use of the MCD.

Another important class of affine equivariant high breakdown point estimates are those based on projections: the Stahel-Donoho (SD) estimate proposed by Stahel (1981) and Donoho (1982) and studied by Maronna and Yohai (1995); P-estimates (Maronna, Stahel and Yohai, 1992); and a recent proposal by Peña and Prieto (2001).

1.2 Problem, Motivation and Approach

Exact computation of robust estimates is feasible only for small data sets. An alternate remedy for large data sets is the approximate computation which is usually based on subsampling. The algorithm of subsampling is composed of taking a number N_s of subsamples, generally of size $p+1$, to obtain an initial set of solutions, which are the starting point for the search for a (hopefully global) extremum. Ruppert (1992) developed a heuristic procedure for S-estimates. Even though the subsampling algorithms tend to lessen the computational burden for robust estimation, the numerical complexity of subsampling algorithms becomes critical for high-dimensional data sets. In order to ensure a given breakdown point, the value of N_s must increase exponentially with p. A high enough value of N_s is also necessary to ensure stability of the result. In general, the subsampling methods are feasible for moderate p, but computing them for large p in a reasonable time requires using values of N_s which imply giving up a high breakdown point. Woodruff and Rocke (1993, 1994) proposed procedures to deal with this problem. Rousseeuw and Van Driesen (1999) proposed the Fast MCD (FMCD), a procedure much more effective than naïve subsampling for minimizing the objective function of the MCD, which seems capable of yielding “good” solutions without requiring huge values of N_s. But FMCD still requires substantial running time for large p. Recently Peña and Prieto (2001) proposed a fast algorithm based on the Kurtosis of projections, which does not require subsampling. However, the main drawback remains the lack of feasible methods to compute the estimates for large high-dimensional data sets.

Much faster estimates with high breakdown points can be computed if one is willing to drop the requirements of positive definiteness and affine equivariance. Early proposals of robust procedures are of this type, see Bickel (1964) and Sen and Puri (1971). A straightforward approach for multivariate location is to simply calculate a robust location estimate for each individual variable. In the case of multivariate scatter, one can similarly apply a robust covariance or correlation coefficient estimate to each pair of variables. Estimates of this type are called *coordinate-wise* and *pairwise.*

There are many proposals for robust univariate location estimates (see for example Hampel et al., 1986). Many researchers obtain several multivariate versions of typically univariate notions such as medians, L-estimates and R-estimates. The multivariate medians are known as the spatial median (also called mediancenter or L_1-median), the Tukey or halfspace median, the Oja median and the Liu or simplicial median; proposed respectively by Haldane (1948), Tukey (1975), Oja (1983) and Liu (1990).

There are also several proposals for the robust estimation of covariance or correlation of a pair of variables. The simplest methods are based on: (i) classical ranks, such as the Spearman's ρ and Kendall's τ (see Abdullah, 1990); (ii) classical correlations applied after coordinate-wise outlier-insensitive transformations, such as the quadrant correlation (QC) and 1-D "Huberized" data (Huber, 1981, page 204); and (iii) bivariate outlier resistant methods such as the method proposed by Gnanadesikan and Kettenring (1972) and studied by Devlin, Gnanadesikan and Kettenring (1981).

Unfortunately, the resulting multivariate location and scatter matrix estimates are not affine equivariant and the scatter matrix is not guaranteed to be positive definite. Rousseeuw and Molenberghs (1993) proposed several methods to deal with the problem of negative eigenvalues. Note that, although the scatter matrices obtained by approaches (i) and (ii) are positive definite, they require a correction to make them consistent for normal data, and the correction destroys their positive definiteness.

In recognition of this opportunity, Maronna and Zamar (2002) recently proposed a new method based on a modification of approach (iii) that preserves positive definiteness and has an "almost affine equivariant" property. However, the particular pairwise estimate they used is not nearly as fast as one might like.

In this thesis, we follow in a similar spirt of Maronna and Zamar (2002). We consider the use of the quadrant correlation and Huberized estimates of approach (ii) above, which are very transparent in the way they work, and enable fast scalable computation for large data applications, with complexity $O(n) \cdot O(p^2)$ for the resulting $p \times p$ covariance or correlation matrix.

1.3 Contributions and Outline of the Thesis

Most multivariate contamination models for numeric data proposed to date (see for example, Hampel et al. 1986) assume that the majority of the observations comes from a nominal distribution such as a multivariate normal distribution, while the remainder comes from another multivariate distribution that generates outliers. The inadequacy of the exiting contamination models for large multivariate data sets has been pointed out by many researchers including Rey (2001) and Zamar and Alqallaf (2001).

The contributions of this thesis are as follows:

- We introduce a flexible multivariate contamination model which allows for different types of contamination in the data. We show several real data examples from the literature suggesting the need for a more general and flexible model. We argue that the proposed model is more appropriate than the existing ones because it is more flexible and better describes different types of possible correlation structures that can occur in practice. We incorporate this feature in our model by allowing its different components to be correlated.

Affine equivariant multivariate location and scatter estimates do not scale well for large sample sizes and highly dimensional data sets. In addition, the new contamination model reveals the possible lack of robustness of these estimates, and suggests that the coordinate-wise and pairwise approaches may be useful to overcome some of the robustness problems.

- We study pairwise robust estimates of scatter matrices based on coordinate-wise robust transformations (the quadrant correlation and the coordinate-wise Huberized estimates). We assess the performance of the proposed pairwise estimates and compare them with the Fast MCD using Monte Carlo simulations and the new contamination model.

- We study the asymptotic properties (consistency and asymptotic normality) of the Huberized correlation coefficient estimates and obtain the mathematical expression for the asymptotic variance of these estimates. Using this expression, we construct an estimate for the variance of the estimated Huberized correlation coefficient. We also verify, using extensive Monte Carlo simulations, that estimated variances approximate well the finite sample variances of the correlation coefficient estimates.

- We distinguish between two kinds of bias in the quadrant and the Huberized correlation coefficient estimates due to the fraction of contamination and because of the structure of the estimates. We then show how to correct for the bias caused by the structure of the estimates. This correction has the drawback that the corrected correlation matrix may no longer be positive definite.

- We show that an improved scalability of the positive definite scatter matrix estimate, proposed by Maronna and Zamar (2002), can be obtained by using the quadrant correlation coefficient estimates instead of the bivariate outlier resistant method (proposed by Gnanadesikan and Kettenring, 1972).

- We extend Huber's (1981) asymptotic maximum bias (maxbias) derivations of the Huberized correlation coefficient estimates to more general cases, where locations and scales are unknown. In particular, we analytically derive the maxbias of the quadrant correlation coefficient and implement numerical computation of the maxbias of the Huberized correlation coefficients.

- We provide the minimaxity properties of the coordinate-wise median in the context of the new contamination model.

- We give numerical evidence suggesting that affine equivariant estimates break down for high-dimensional data under the new contamination model.

The rest of the thesis is organized as follows. Chapter 2 provides the background material on robust estimation in the univariate setting. We present the robust estimation of multivariate location and covariance, in which we generalize familiar concepts in the univariate case and discuss the difficulties that occur in the transition to higher dimensions. We describe three different types of multivariate location and covariance estimates; the M-estimates, the S-estimates and the MCD-estimate.

In chapter 3, we introduce the new class of multivariate contamination model and show that it is more appropriate than the existing contamination models. In particular, we give some real data examples from the literature to illustrate that. We study the different correlation structures that the contamination indicators of the variables may have in the contamination model. We also illustrate the dependency situations among the components of the contamination model.

In chapter 4, we present the pairwise robust estimates of scatter matrices. We report the results of Monte Carlo studies that assess the performance of the pairwise estimates and compare with the Fast MCD. We study the asymptotic properties (consistency and asymptotic normality) of the Huberized correlation coefficient estimates. We present the asymptotic maximum bias of the Huberized correlation coefficient estimates. Finally, we illustrate the implementation of the quadrant and Huberized correlation coefficient estimates on three real data sets. We show that the proposed methods are capable of computing robust location and covariance estimates and detecting multidimensional outliers on arbitrarily large data sets.

Chapter 5 discusses the coordinate-wise robust multivariate location estimates. We study the coordinate-wise median estimate, in which we show its minimaxity properties under the new contamination model.

Chapter 6 contains a brief list of the results obtained in this thesis, the challenges that remain to be solved and the directions we foresee for future work.

To facilitate the reading of the thesis, some of the proofs for the results presented will be relegated to the chapter appendix.

Chapter 2

Background and Related Work

In this chapter, we review robust estimation techniques specifically tailored to estimating multivariate location and scatter. Particular attention will be paid to three methods: M-estimates, S-estimates and the minimum covariance determinant (MCD) estimate. The intuition behind these estimates is to find the location and covariance estimate by trying to simultaneously identify and down-weight outliers in the estimates; although, they do so in different ways. These methods and their attributes will be discussed in this chapter.

To briefly introduce these estimates, let $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \in \mathbb{R}^p$ denote a collection of data points. The majority of them are assumed to be i.i.d. from a distribution whose mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ we wish to estimate, but some of the data points are drawn from another unknown and arbitrary distribution. The estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ will be denoted as $\boldsymbol{t}$ and $\boldsymbol{C}$, respectively. The particular estimate to which it corresponds will be made clear from the context.

An M-estimate $(\boldsymbol{t}, \boldsymbol{C})$ of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is obtained as the solution to the system of equations

$$\frac{1}{n}\sum_{i=1}^{n} v_1(d_i)(\boldsymbol{x}_i - \boldsymbol{t}) = \boldsymbol{0};$$

$$\frac{1}{n}\sum_{i=1}^{n} v_2(d_i^2)(\boldsymbol{x}_i - \boldsymbol{t})(\boldsymbol{x}_i - \boldsymbol{t})^T = \boldsymbol{C},$$

where $d_i^2 = d(\boldsymbol{x}_i, \boldsymbol{t}; \boldsymbol{C})^2 = (\boldsymbol{x}_i - \boldsymbol{t})'\boldsymbol{C}^{-1}(\boldsymbol{x}_i - \boldsymbol{t})$ and $v_1(\cdot)$ and $v_2(\cdot)$ are weighting functions that control the influence of points that are distant (with respect to $\boldsymbol{C}^{-1}$) from $\boldsymbol{t}$. If $v_1(\cdot) = v_2(\cdot) = 1$, then $\boldsymbol{t}$ and $\boldsymbol{C}$ are the sample mean and covariance. By taking $v_1(\cdot)$

and $v_2(\cdot)$ to be decreasing functions, we can reduce the effects of outliers based upon how different (in terms of second order statistics) they are from the rest of the data.

An S-estimate $(\boldsymbol{t}, \boldsymbol{C})$ of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is obtained as the solution to the optimization problem over the set $\Theta = \mathbb{R}^p \times \text{PDS}(p)$, where $\text{PDS}(p)$ is the set of all positive definite symmetric $p \times p$ matrices,

$$\min_{(\boldsymbol{t},\boldsymbol{C}) \in \mathbb{R}^p \times \text{PDS}(p)} \{|\boldsymbol{C}|\};$$

such that

$$\frac{1}{n} \sum_{i=1}^{n} \rho(d_i) = b_0,$$

where $\rho(\cdot)$ is a monotonically increasing function and b_0 is an appropriately defined constant. In the case of $\rho(d) = d^2$, the optimization constraint says that the sum of the squared Mahalanobis distances is constant, i.e. the likelihood is held constant under the assumption of Gaussian data. Not too surprisingly, this is equivalent to least squares estimation and results in the sample mean and covariance as its estimates. Now, in choosing a $\rho(\cdot)$ which rises slower than quadratically, we can relax the weighting associated with points distant from the center of the ellipsoid, lowering the influence of outliers.

The MCD estimate is very intuitive. It does not involve solving a system of nonlinear equations nor a nonlinear optimization problem as above; but instead, it finds the subset of $h (n/2 < h \leq n)$ points that are most tightly clustered and bases its estimate on that subset. In particular, out of all subsets of size h, it finds the one whose sample covariance determinant is minimal, and takes the MCD estimate $(\boldsymbol{t}, \boldsymbol{C})$ of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ as the sample mean and covariance of this subset.

The remainder of this chapter is structured as follows. Section 2.1 provides background material on robust estimation in the univariate setting. Section 2.2 provides the introduction to robust estimation of multivariate location and scatter generalizing familiar concepts in the univariate case and discussing difficulties in the transition to higher dimensions. Section 2.3 discusses the M-estimates in the multivariate context. Section 2.4 describes the S-estimates. Section 2.5 discusses the MCD estimate. The chapter concludes with Section 2.6 which consists of a summary of the material and a closing comment on the direction of robust estimation of multivariate location and covariance.

2.1 Univariate Robust Statistics

In this section, we provide a brief introduction to robust statistics in the univariate setting. Naturally, such a vast field cannot be summarized in a section of a chapter, so we only cover the concepts with which we will be directly interested in the multivariate case. For further details see Hampel et al. (1986) and Huber (1981).

To start, we consider to what kind of model deviations we wish to be robust. For this, we adopt the popular *ϵ-replacement* model, which is also commonly called the Gross Error Model (GEM). Under this model, i.i.d. points $(x_1, x_2, \ldots) \in \mathbb{R}$ are drawn from the distribution

$$G_\theta(x) = (1-\epsilon)F_\theta(x) + \epsilon H(x), \tag{2.1}$$

where $F_\theta(x)$ is a strict parametric model parameterized by θ and $H(x)$ is an arbitrary distribution. The intuition behind this model is that $(1-\epsilon)$ of the sample arise from the parametric model, but ϵ of the sample have been replaced by points from an arbitrary distribution.

The derived statistic should estimate the parameter θ. A statistic based on a sample set of size n will be denoted by $T_n = T_n(x_1, \ldots, x_n)$, $T_n : \mathbb{R}^n \mapsto \mathbb{R}$. We will only consider statistics which generalize to *statistical functionals* which are mappings $T(G)$: domain$(T) \mapsto \mathbb{R}$ with the property that $T(G_n) = T_n \to T(G)$, where G_n is the empirical distribution of the data. A desirable property of statistical functions is Fisher consistency, i.e. $T(F_\theta) = \theta$ so that under the exact parametric distribution, the estimate returns the correct parameter value. Note that this definition of Fisher consistency of functionals encompasses the more common requirement in estimation theory that $T_n \overset{\mathcal{P}}{\to} \theta$ for $x_1, \ldots, x_n \sim$ i.i.d. F_θ.

2.1.1 Influence Function

We follow the statistics literature in using the term *efficiency* to relate to having a small variance. Note however that it does not have to do with achieving the Cramer-Rao lower bound. A fundamental concept in measuring robustness and efficiency of a statistical functional T is its *influence function* (IF). The IF of T at a distribution G is given by

$$\mathrm{IF}(x; T, G) = \lim_{t \to 0} \frac{T((1-t)G + t\Delta_x) - T(G)}{t}, \tag{2.2}$$

for those x where the limit exists, where Δ_x represents point measure of 1 at x. The IF can be interpreted as a directional derivative of T in the space of distributions. So intuitively, an estimate T with a "nice" smooth $\mathrm{IF}(x; T, G)$ should be robust as slight changes to the underlying distribution result in slight changes to the estimate. By applying a von Mises expansion, which resembles a Taylor expansion for functionals, of T at a G_n "near" G, we get

$$\begin{aligned}
T_n = T(G_n) &= T(G) + \int \mathrm{IF}(x; T, G)\ d(G_n - G)(x) + \text{remainder} \\
T_n &= T(G) + \int \mathrm{IF}(x; T, G)\ dG_n(x) + \text{remainder} \\
\sqrt{n}(T_n - T(G)) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathrm{IF}(x_i; T, G) + \text{remainder},
\end{aligned}$$

where the second line follows because $\int \mathrm{IF}(x; T, G)\ d(G) = 0$ as a property of von Mises functionals. Now, if the "remainder" is negligible, which it is for many statistics, then by the Central Limit Theorem, the error asymptotically becomes a zero mean Gaussian with the variance determined by the IF, i.e. $\sqrt{n}(T_n - T(G)) \xrightarrow{\mathcal{D}} \mathrm{N}(0, \mathrm{AV}(T, G))$, where

$$\mathrm{AV}(T, G) = \int \mathrm{IF}(x; T, G)^2\ dG(x) \tag{2.3}$$

is the asymptotic variance. Of course, the above derivation is not precise, but it is meant to help provide insight to the IF and a rigorous derivation can be found in Fernholz (1983).

2.1.2 Robustness Measures

Equation (2.3) illustrates the importance of the IF in characterizing the efficiency of an estimate, however it has equally important role in characterizing robustness. Recall the IF's interpretation as a directional derivative in the space of distributions. Thus it characterizes the sensitivity of the statistic to slight deviations from the model distribution. This motivates the definition the gross error sensitivity (GES) for T at F_θ as

$$\gamma^* = \sup_x |\text{IF}(x; T, F_\theta)|. \tag{2.4}$$

The GES thus measures the worst influence an infinitesimally small fraction of contamination can have on the estimate, i.e. it is an upper bound for the standardized bias induced by the contaminated distribution (what we are calling bias here is *not* a bias in the standard sense of the word, but a bias attributed to a change in the underlying distribution). For this reason, we say that an estimate T at F_θ is *B-robust* if γ^* is finite, where the B stands for bias. An estimate is said to be *most B-robust* if it achieves the minimal GES over all Fisher consistent estimates. As is frequently the case in nature, robustness and efficiency are conflicting goals. Thus, robust statistics will frequently search for *optimally B-robust* estimates which minimize asymptotic variance given a bound on the GES.

From its nature as a derivative, the IF (and thus GES) is only a local characterization of robustness in a small neighborhood of the model distribution F_θ. To provide a global measure of robustness, the concept of the breakdown point was developed. The *finite-sample replacement breakdown point* ϵ_n^* of an estimate T_n at the sample $(x_1, \ldots, x_n)$ is defined as

$$\epsilon_n^* = \min\left\{ \frac{m}{n} \;\middle|\; \max_{i_1,\ldots,i_m} \sup_{y_1,\ldots,y_m} |T_n(z_1, \ldots, z_n) - T_n(x_1, \ldots, x_n)| = \infty \right\}, \tag{2.5}$$

where $(z_1, \ldots, z_n)$ is obtained by replacing the m data points $x_{i1}, \ldots, x_{im}$ with arbitrary values $y_1, \ldots, y_m$. What this definition says is that the breakdown points is the smallest possible fraction of points which must be corrupted to take the estimate across all bounds. Thus, a non-robust estimate, in the sense of breakdown point, has $\epsilon_n^* = 1/n$. Note that there is another prevalent definition of the finite-sample breakdown point in the literature, particularly in the univariate setting for which

$$\tilde{\epsilon}_n^* = \max\left\{ \frac{m}{n} \,\middle|\, \max_{i_1,\ldots,i_m} \sup_{y_1,\ldots,y_m} |T_n(z_1,\ldots,z_n)| < \infty \right\}.$$

The difference between the two is minor in that ϵ_n^* is the smallest fraction of replacements which can cause the estimate to become unbounded; $\tilde{\epsilon}_n^*$ is the largest fraction of replacements that the estimate can tolerate while still guaranteed to be bounded. Thus, their relation is $\epsilon_n^* = \tilde{\epsilon}_n^* + 1/n$. For the remainder of this chapter, we will only consider the initial definition given in equation (2.5); however, in reading other papers on this subject, it is important to differentiate between the two. Even though the definition in equation (2.5) uses $(x_1, \ldots, x_n)$, the breakdown point almost never depends on the points for interesting estimates. There also exists a definition of breakdown point based on distributions instead of points, but it is considerably more involved and conveys the same idea as equation (2.5), see Huber (1981). This distribution breakdown point is denoted as ϵ^* and for reasonable estimates $\epsilon_n^* \to \epsilon^*$.

The concept of gross error sensitivity measures the maximum effect that an infinitesimal amount of point-mass contamination can have on a functional. A stronger robustness concept is to measure the maximum effect or bias that any type of contamination can have on a functional. Define the contamination neighborhood of F

$$\mathcal{F}_\epsilon = \{G : G = (1-\epsilon)F + \epsilon H; H \text{ any distribution}\},$$

for a given fraction of contamination ϵ $(0 \leq \epsilon \leq 1)$. The *maximum contamination bias* function is defined to be

$$B(\epsilon; T, F) = \sup_{G \in \mathcal{F}_\epsilon} |T(G) - T(F)|. \tag{2.6}$$

The maximum bias function is related to the breakdown point, which is a measure of global robustness, as well as to the contamination sensitivity, which is a measure of local robustness. The breakdown point of T at F over contamination neighborhood is defined to be

$$\epsilon^*(T, F) = \inf\{\epsilon > 0 | B(\epsilon, T, F) = \infty\}, \tag{2.7}$$

and the contamination sensitivity is defined to be

$$\gamma(T, F) = \limsup_{\epsilon \to 0} B(\epsilon; T, F)/\epsilon. \tag{2.8}$$

Under certain regularity conditions, the contamination sensitivity and the gross error sensitivity are equal, see Hampel et al. (1986) for further discussion. In general, though, it readily follows that

$$\gamma(T, F) \geq \sup_x \{\limsup_{\epsilon \to 0} |T(G(\epsilon, x)) - T(F)|/\epsilon\} = \gamma^*(T, F). \tag{2.9}$$

There are other common measures of robustness such as qualitative robustness, continuity of a statistical functional, local-shift sensitivity, and rejection point. These will not be discussed further, but the interested reader should see Hampel et al. (1986).

2.1.3 M-Estimates for Location

Having introduced some basic concepts in robustness theory, we now illustrate them with the well established M-estimate for univariate location. In addition to clarifying the concepts introduced above, it may help establish intuition for its extension to the multivariate setting in Section 2.3.

The popular maximum likelihood (ML) estimate chooses the statistic as the parameter value θ which maximizes the likelihood of the sample data, i.e.

$$T_n = \arg\max_{\hat{\theta}} \left\{ \prod_{i=1}^{n} f_{\hat{\theta}}(x_i) \right\} = \arg\max_{\hat{\theta}} \left\{ \sum_{i=1}^{n} \ln\left(f_{\hat{\theta}}(x_i)\right) \right\}$$

where f_θ represents a member of a family of pdf's parameterized by θ. Unfortunately, ML estimation is frequently non-robust. In order to robustify that approach, Huber (1964) considered generalizing the objective function to a function $\rho(x,\theta)$ with derivative $\psi(x,\theta) = \frac{\partial\rho(x,\theta)}{\partial\theta}$ and proposed to calculate

$$T_n = \arg\max_{\hat{\theta}} \left\{ \sum_{i=1}^{n} \rho(x_i, \hat{\theta}) \right\} \tag{2.10}$$

as an estimate. Any solution of this will then solve

$$\sum_{i=1}^{n} \psi(x_i, T_n) = 0. \tag{2.11}$$

Because for reasonable choices of ρ, a solution of equation (2.11) will solve equation (2.10), a solution of either equation (2.10) or equation (2.11) is called an M-estimate, where the M comes from "generalized Maximum likelihood". Because it is frequently simpler to work directly with ψ, little use will be made of ρ and we will associate ψ with the M-estimate it defines.

Extending equation (2.11) to a statistical functional, we get that an M-estimate is a solution of the equation

$$\int \psi(x, T(G))\ dG(x) = 0. \tag{2.12}$$

Applying this to the contaminated distribution $G_{t,x} = (1-t)F + t\Delta_x = F + t(\Delta_x - F)$, differentiating with respect to t, and taking the limit $t \to 0$, we get

$$\begin{aligned}
0 &= \int \psi\left(y, T(F + t[\Delta_x - F])\right)\ d(F + t[\Delta_x - F])(y) \\
0 &= \int \psi\left(y, T(F)\right)\ d(\Delta_x - F)(y) + \frac{\partial}{\partial t}[T(G_t, x)]_{t=0} \int \frac{\partial}{\partial\theta}[\psi(y,\theta)]_{T(F)}\ dF(y) \\
0 &= \int \psi\left(y, T(F)\right)\ d\Delta_x(y) + \mathrm{IF}(x; T, F) \int \frac{\partial}{\partial\theta}[\psi(y,\theta)]_{T(F)}\ dF(y)
\end{aligned}$$

$$\mathrm{IF}(x;\psi,F) = \mathrm{IF}(x;T,F) = \frac{\psi\left(x, T(F)\right)}{-\int \frac{\partial}{\partial\theta}[\psi(y,\theta)]_{T(F)} dF(y)} \tag{2.13}$$

where the denominator is assumed nonzero. Thus, for an M-estimate, one can straightforwardly calculate the IF and its derived quantities such as asymptotic variance and gross-error sensitivity from ψ.

Hampel (1968) derives an optimal M-estimate. He considers that ψ which minimizes the asymptotic variance given a constraint on the GES. His result essentially states that under certain regularity conditions on the set of allowable distributions, the minimum variance M-estimate, subject to a bound on the GES, is that for which the ψ function is taken to be a vertical shift and "clipping" of the maximum likelihood score function $s(x, \theta_*) = \frac{\partial}{\partial\theta}[\ln(f_\theta(x))]_{\theta_*}$. More precisely, for $\theta_* \in \Theta$ a convex set

THEOREM 2.1 *Assume that*

- $s(x, \theta_*)$ *exists for all* x;
- $\int s(x, \theta_*)\ dF_{\theta_*} = 0$ *(a regularity condition);*
- *the Fisher information* $J(F_{\theta_*}) = \int s(x, \theta_*)^2\ dF_{\theta_*}$ *satisfies* $0 < J(F_{\theta_*}) < \infty$.

Then for any $b > 0, \exists\ a \in \mathbb{R}$ *such that*

$$\tilde{\psi}(x) = [s(x, \theta_*) - a]_{-b}^{b} \tag{2.14}$$

satisfies $\int \tilde{\psi}\ dF_{\theta_*} = 0$ *and* $d = \int \tilde{\psi}(y) s(y, \theta_*)\ dF_{\theta_*}(y) > 0$. *This* $\tilde{\psi}$ *minimizes the asymptotic variance*

$$AV(\psi, F_{\theta_*}) = \frac{\int \psi^2(y)\ dF_{\theta_*}(y)}{\left[\int \psi(y) s(y, \theta_*)\ dF_{\theta_*}(y)\right]^2}$$

among all ψ *satisfying*

$$\int \psi(y)\ dF_{\theta_*}(y) = 0$$

$$\int \psi(y) s(y, \theta_*)\ dF_{\theta_*}(y) \neq 0,$$

and

$$\sup \left| \frac{\psi(x)}{\int \psi(y) s(y, \theta_*) F_{\theta_*}(y)} \right| < c = \frac{b}{d},$$

where $[\cdot]_{-b}^{b} = \max\{-b, \min\{\cdot, b\}\}$ denotes clipping the function to the range $[-b, b]$.

Thus, one can control the degrees of B-robustness ($\gamma^* = b/d$) by varying the clipping level b. The shift a is necessary to maintain Fisher consistency, which for M-estimates simplifies to

$$\int \psi(y, \theta_*) \; dF_{\theta_*}(y) = 0, \; \forall \; \theta_*.$$

Note that in the case of placing no bound on the GES, we get the ML estimate as expected.

In the case of location estimation, we model $F_\theta(x) = F(x - \theta)$. Under this model, it is natural to only consider ψ of the form $\psi(x, \theta) = \psi(x - \theta)$ with $\int_{-\infty}^{\infty} \psi(x) \; dF(x) = 0$ for Fisher consistency. From equation (2.13), the IF can be written as

$$\mathrm{IF}(x; \psi, G) = \frac{\psi(x - T(G))}{-\int \psi'(y - T(G)) \; dG(y)}$$

which at the model distribution F becomes

$$\mathrm{IF}(x; \psi, F) = \frac{\psi(x)}{-\int \psi'(y) \; dF(y)}$$

where the denominator is again assumed to be nonzero. Thus, the IF for M-estimates of location are proportional to ψ, so up to a scaling factor, we can specify the IF through the definition of ψ. For a symmetric model distribution F, it is natural to choose a ψ function which is skew-symmetric, i.e. $\psi(-x) = -\psi(x)$. If ψ is also monotonic, then Hampel (1971) obtains the following results:

1. If ψ is bounded then the resulting estimate is B-robust and has a breakdown point of 1/2;

2. If ψ is not bounded then the resulting estimate is not B-robust and has a breakdown point of 0.

The B-robustness can be seen from equation (2.4) and the fact that ψ is proportional to the IF. The breakdown result can be seen by the following reasoning. If ψ is bounded and monotonic then it must eventually approach a constant. Thus when a minority of the data samples grow arbitrarily large, the effect of each corrupted point on the sum in equation (2.11) is limited and is essentially the same as a much lower magnitude value, thus limiting the effect it can have. This point can be more visually appreciated with the example of an optimal M-estimate for location presented next.

For an example, we apply the optimality result in Theorem 2.1 to location estimation. If F is the standard normal distribution, then the score function becomes $s(x-\theta) = x-\theta$ and because F is symmetric $a = 0$. This results in the optimal ψ function

$$\psi_H(x-\theta, b) = [x-\theta]_{-b}^{b} \tag{2.15}$$

known as Huber's ψ function. In the case of $b \to \infty$, we have $\psi_H \to x$ and the estimate becomes

$$\begin{aligned} 0 &= \sum_{i=1}^{n} \psi_H(x_i - T_n, \infty) = \sum_{i=1}^{n} (x_i - T_n) \\ T_n &= \frac{1}{n} \sum_{i=1}^{n} x_i. \end{aligned}$$

However, when we clip ψ to $[-b, b]$, we limit the effect of a sample distant from T_n to be the same that of a sample at a distance b from T_n, i.e. Huber's ψ function effectively draws in points that are further than b from T_n. Thus, arbitrarily large points can have only a limited effect. This "drawing in" of points is why bounded monotonic M-estimates have a breakdown point of 1/2. It should be pointed out that Huber's ψ function does not correspond to an α-Winsorized mean or an α-trimmed mean both of which are not in the class of M-estimates (these fall into a class called L-estimates which will not be discussed here).

Huber (1964) derives an optimal location M-estimate according to a different criterion than Hampel, but arises at the same answer. Huber's analysis was strictly for the case of univariate location estimation. Difficulties are encountered in trying to generalize it. Huber's approach does not utilize the IF as he believes that its interpretation of dealing with infinitesimally small contaminations is not consistent with the spirit of robust statistics where model deviations are more than just infinitesimally small. Instead, he considers a minimax condition on the variance over a neighborhood around the model distribution. Using a symmetrized GEM (Gross Error Model) neighborhood

$$\mathcal{F} = \{G \mid G = (1-\epsilon)F + \epsilon H, H \text{ symmetric}\},$$

he finds the saddle point (ψ_0, G_0) that satisfies

$$V(\psi_0, G) \leq V(\psi_0, G_0) \leq V(\psi, G_0), \ \forall \ \psi \text{ and } G \in \mathcal{F}.$$

Under some regularity condition on the distribution and set of allowable distributions he shows that $\psi_0 = [s(x-\theta)]_{-b}^{b}$, where b is determined from ϵ and G_0. Thus, the solution to Huber's minimax problem has the same form as Hampel's constrained minimization problem. Li and Zamar (1991) extended Huber's (1964) minimax result to the case when the scale parameter is unknown and must be estimated along with the location parameter.

2.2 Robust Estimation in the Multivariate Setting

Having presented an introduction covering location estimation in the univariate case, we now proceed to the primary focus of this thesis which is the investigation of robust estimates of multivariate location and scatter. Most of the discussion in this section follows Lopuhaä and Rousseeuw (1991), with a few exceptions which will be noted.

For a set of points $X_n = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ with $\boldsymbol{x}_i \in \mathbb{R}^p$, we wish to find robust estimates $\boldsymbol{t}_n(X_n) \in \mathbb{R}^p$ of the mean and $\boldsymbol{C}_n(X_n) \in \text{PDS}(p)$ (set of positive definite symmetric $p \times p$

matrices) of the covariance which describe the bulk of the data. The majority of the points are modelled as i.i.d. from an elliptical distribution $F_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$ with density

$$f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x}) = |\boldsymbol{\Sigma}|^{-1/2}\varphi\left((\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{2.16}$$

where $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ is scaled so that a valid density is produced.

We will consider some estimates which satisfy the property of affine equivariance which is defined as follows.

Definition 2.1 – **Affine Equivariance** – *For any invertible $p \times p$ matrix $A, \boldsymbol{v} \in \mathbb{R}^p$, and data set $X_n \in \mathbb{R}^{p\times n}$*

- *A location estimate $\boldsymbol{t}_n$ is affine equivariant if $\boldsymbol{t}_n(AX_n+\boldsymbol{v}) = A\boldsymbol{t}_n(X_n)+\boldsymbol{v}$;*
- *A covariance estimate $\boldsymbol{C}_n$ is affine equivariant if $\boldsymbol{C}_n(AX_n+\boldsymbol{v}) = A\boldsymbol{C}_n(X_n)A^T$,*

where $AX_n+\boldsymbol{v} = (A\boldsymbol{x}_1+\boldsymbol{v},\ldots,A\boldsymbol{x}_n+\boldsymbol{v})$, and A^T is the transpose of A. We will also at times mention translation and coordinate-wise scale equivariant estimates of location, which are estimates that satisfy the weaker conditions,

$$\boldsymbol{t}_n(X_n+\boldsymbol{v}) = \boldsymbol{t}_n(X_n)+\boldsymbol{v};$$

$$\boldsymbol{t}_n(DX_n) = D\boldsymbol{t}_n(X_n),$$

for all diagonal $p \times p$ matrix D.

The definition of breakdown point in equation (2.5) is primarily suited for univariate location estimation. Extending it to our broader context, the location breakdown point is defined as

$$\epsilon_n^*(\boldsymbol{t}_n, X_n) = \min\left\{\frac{m}{n}\ \middle|\ \max_{i_1,\ldots,i_m}\sup_{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_m}\|\boldsymbol{t}_n(X_n)-\boldsymbol{t}_n(Z_n)\| = \infty\right\} \tag{2.17}$$

where as before $Z_n = (\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n)$ is obtained by replacing the m data points $\boldsymbol{x}_{i1},\ldots,\boldsymbol{x}_{im}$ with arbitrary values $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_m$, and the interpretation is the same as for the scalar case. The breakdown point for covariance is quite similar except that we have to account for the undesirable possibility of a singular covariance estimate (it is assumed that the covariance of the uncontaminated source distribution is full rank). Thus, protecting against eigenvalues approaching ∞ and 0, the breakdown point for covariance estimates is defined as

$$\epsilon_n^*(\boldsymbol{C}_n, X_n) = \min\left\{ \frac{m}{n} \;\middle|\; \max_{i_1,\ldots,i_m} \sup_{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_m} D(\boldsymbol{C}_n(X_n), \boldsymbol{C}_n(Z_n)) = \infty \right\}$$

where $D(A,B) = \max\{|\lambda_{max}(A) - \lambda_{max}(B)|, |\lambda_{min}(A)^{-1} - \lambda_{min}(B)^{-1}|\}$ with $\lambda_{max}(A)$ and $\lambda_{min}(A)$ denoting the largest and smallest eigenvalues of A. Thus, covariance estimates are considered to be broken if they produce eigenvalues which are arbitrarily large or close to 0.

In the univariate setting, the median and bounded M-estimates are examples of maximally robust with respect to the breakdown point affine equivariant estimates with $\epsilon_n^* = [(n+1)/2]/n$. A natural question is what happens to the breakdown point when the data dimensionality is increased. Lopuhaä and Rousseeuw (1991) addressed this issue. Let $[\cdot]$ be the greatest integer function, they show that if $\boldsymbol{t}_n$ is translation equivariant, then

$$\epsilon_n^*(\boldsymbol{t}_n, X_n) \leq \frac{[(n+1)/2]}{n} = \begin{cases} \frac{1}{2} & : n \text{ odd}, \\ \frac{1}{2} + \frac{1}{2n} & : n \text{ even}. \end{cases}$$

Because affine equivariant estimates are a subclass of translation equivariant estimates, the above inequality holds for them as well. This result fits with ones intuition because any estimate which fits the majority of the data must fail if more than half of the data points can be arbitrarily corrupted.

Unfortunately, the maximal breakdown point for an affine equivariant covariance estimate is slightly lower than that for a location estimate. Davies (1987) proves that for any affine equivariant covariance estimate $\boldsymbol{C}_n$,

$$\epsilon_n^*(\boldsymbol{C}_n, X_n) \leq \frac{[(n-p+1)/2]}{n}.$$

Thus, covariance estimate have a slightly smaller maximal breakdown point than location estimates. Lopuhaä and Rousseeuw (1991) suggested that a possible reason for this difference is that location estimate can only breakdown if the estimates can be made arbitrarily large, while covariance estimates can breakdown for eigenvalues tending to both 0 and ∞.

In addition to the breakdown point, the other essential quantity that needs to be extended to the multivariate scenario is the IF. First note that the parameter of the distribution in our new setting is now the vector $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \boldsymbol{\Theta} = \mathbb{R}^p \times \mathrm{PDS}(p)$. The extension of the IF is then straightforward and given by Lopuhaä (1989).

Definition 2.2 – **Influence Function** – *consider a statistical functional $\boldsymbol{T}(\cdot)$ mapping a set of distributions into the parameter space $\boldsymbol{\Theta}$ and $G \in dom(\boldsymbol{T})$. The IF of $\boldsymbol{T}$ at G is defined as*

$$IF(\boldsymbol{x}; \boldsymbol{T}, G) = \lim_{t \to 0} \frac{\boldsymbol{T}\left((1-t)G + t\Delta_{\boldsymbol{x}}\right) - \boldsymbol{T}(G)}{t} \tag{2.18}$$

if the limit exists for all $\boldsymbol{x} \in \mathbb{R}^p$.

For affine equivariant estimates and elliptical distributions, it is only necessary to determine the IF under the spherically symmetric distribution $F_{\mathbf{0},\boldsymbol{I}}$ as the IF for any other set of valid parameters can be obtained by

$$\mathrm{IF}(\boldsymbol{x}; \boldsymbol{t}, F_{\boldsymbol{\mu},\boldsymbol{\Sigma}}) = A \ \mathrm{IF}(A^{-1}(\boldsymbol{x} - \boldsymbol{\mu}); \boldsymbol{t}, F_{\mathbf{0},\boldsymbol{I}}) \tag{2.19}$$

$$\mathrm{IF}(\boldsymbol{x}; \boldsymbol{C}, F_{\boldsymbol{\mu},\boldsymbol{\Sigma}}) = A \ \mathrm{IF}\left(A^{-1}(\boldsymbol{x} - \boldsymbol{\mu}); \boldsymbol{C}, F_{\mathbf{0},\boldsymbol{I}}\right) A^T \tag{2.20}$$

where $AA^T = \boldsymbol{\Sigma}$.

The formula for the asymptotic variance under some regularity conditions will generalize to

$$\mathrm{AV}(\boldsymbol{T}; G) = \int \mathrm{IF}(\boldsymbol{x}; \boldsymbol{T}, G)\mathrm{IF}(\boldsymbol{x}; \boldsymbol{T}, G)^T dG(\boldsymbol{x}).$$

We consider the multivariate location model to illustrate the maximum bias of multivariate location. Let $\boldsymbol{X} = (X_1, \ldots, X_p)$ be a random vector with distribution $F_{\boldsymbol{\mu}}(\boldsymbol{x}) = F_{\mathbf{0}}(\boldsymbol{x} - \boldsymbol{\mu})$, where $F_{\mathbf{0}}$ is symmetric around $\mathbf{0}$. To study the robustness property of the multivariate location we will consider a contamination neighborhood of the target distribution. Given a fraction of contamination $\epsilon > 0$, the corresponding contamination neighborhood of $F_{\boldsymbol{\mu}}$ is defined by

$$\mathcal{V}_\epsilon(F_{\boldsymbol{\mu}}) = \{F = (1-\epsilon)F_{\boldsymbol{\mu}} + \epsilon F^* : F^* \text{ any distribution on } \mathbb{R}^p\}.$$

It is natural to require that an estimating functional $\boldsymbol{T}$ have the Fisher consistency property $\boldsymbol{T}(F_{\boldsymbol{\mu}}) = \boldsymbol{\mu}$. In general, given $F \in \mathcal{V}_\epsilon(F_{\boldsymbol{\mu}})$ we will have $\boldsymbol{T}(F) \neq \boldsymbol{\mu}$. Then, we define the asymptotic bias of $\boldsymbol{T}$ in F by

$$b(\boldsymbol{T}, F, \boldsymbol{\mu}) = \left((\boldsymbol{T}(F) - \boldsymbol{\mu})' \boldsymbol{\Sigma}_{F_{\mathbf{0}}}^{-1} (\boldsymbol{T}(F) - \boldsymbol{\mu})\right)^{1/2}, \tag{2.21}$$

where $\boldsymbol{\Sigma}_{F_{\mathbf{0}}}$ is an affine equivariant scatter functional.

The maximum asymptotic bias of an estimating functional $\boldsymbol{T}$ for fraction of contamination ϵ is defined by

$$B(\boldsymbol{T}, \epsilon, F_{\boldsymbol{\mu}}) = \sup_{F \in \mathcal{V}_\epsilon(F_{\boldsymbol{\mu}})} b(\boldsymbol{T}, F, \boldsymbol{\mu}). \tag{2.22}$$

For the univariate case ($p = 1$), the maximum bias of a location estimate T at an arbitrary distribution G_0 reduces to

$$B(T, \epsilon, G_0) = \sup_{H \in \mathcal{V}_\epsilon(G_0)} \left| \frac{T(G) - T(G_0)}{\sigma(G_0)} \right|,$$

where $\sigma(\cdot)$ is a dispersion functional.

If the functional $\boldsymbol{T}$ is equivariant, the maximum bias does not depend on $\boldsymbol{\mu}$, and can therefore be denoted by $B(\boldsymbol{T}, \epsilon, F_{\mathbf{0}})$.

He and Simpson (1992) introduced the contamination sensitivity of an estimate $\boldsymbol{T}$ as

$$\gamma(\boldsymbol{T}, F, \boldsymbol{\mu}) = \left.\frac{\partial B(\boldsymbol{T}, \epsilon, F_{\boldsymbol{\mu}})}{\partial \epsilon}\right|_{\epsilon=0}.$$

Observe that $\gamma(\boldsymbol{T}, F_{\boldsymbol{\mu}}) = \gamma(\boldsymbol{T}, F_{\boldsymbol{0}})$ because of the invariance of the bias.
For small ϵ, the maximum bias can be approximated by

$$B(\boldsymbol{T}, \epsilon, F_{\boldsymbol{\mu}}) \approx \epsilon\gamma(\boldsymbol{T}, F_{\boldsymbol{\mu}}). \tag{2.23}$$

The contamination sensitivity $\gamma(\boldsymbol{T}, F_{\boldsymbol{\mu}})$ is closely related to Hampel's (1971) gross error sensitivity $\gamma^*(\boldsymbol{T}, F_{\boldsymbol{\mu}})$. In fact it is easy to show that always

$$\gamma(\boldsymbol{T}, F_{\boldsymbol{\mu}}) \geq \gamma^*(\boldsymbol{T}, F_{\boldsymbol{\mu}}),$$

where

$$\gamma^*(\boldsymbol{T}, F_{\boldsymbol{\mu}}) = \sup_{\boldsymbol{c}\in\mathbb{R}^p} \left\| \lim_{\epsilon\to 0} \frac{\boldsymbol{T}((1-\epsilon)F_{\boldsymbol{\mu}} + \epsilon\delta_{\boldsymbol{c}}) - \boldsymbol{T}(F_{\boldsymbol{\mu}})}{\epsilon} \right\|,$$

and $\delta_{\boldsymbol{c}}$ stands for a point mass contamination. Under very general regularity conditions $\gamma^*(\boldsymbol{T}, F_{\boldsymbol{\mu}}) = \gamma(\boldsymbol{T}, F_{\boldsymbol{\mu}})$.

Huber (1964) proved that if F_0 is a univariate symmetric distribution with unimodal density f_0 and $F_\mu = F_0(x-\mu)$, then the maximum bias of the median estimating functional T_M is minimax among the affine equivariant estimates, i.e., if T is another affine equivariant estimating functional, then

$$B(T, \epsilon, F_\mu) \geq B(T_M, \epsilon, F_\mu) = F_0^{-1}\left(\frac{1}{2(1-\epsilon)}\right) = d_1(\epsilon, F_0), \tag{2.24}$$

where d_1 stands for the maximum bias of the median i.e., the percentile-0.5 under contamination at infinity.

He and Simpson (1993) obtained a lower bound for the maximum bias of equivariant estimates. Using this result Adrover and Yohai (2002) prove that $d_1(\epsilon, H_0)$ (provided by Theorem 2.1 of He and Simpson, 1993) is a lower bound for any equivariant multivariate location estimate when the central model is elliptical. Croux et al. (1997) derived a similar result when the covariance is known and the central model is multivariate normal.

2.3 M-Estimates

In Section 2.1.3, we discussed the M-estimate for univariate location. Here we describe M-estimates for multivariate location and covariance as presented by Maronna (1976). The definition of the M-estimate $\hat{\boldsymbol{\theta}} = (\boldsymbol{t}, \boldsymbol{C})$ of multivariate location and covariance $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ based on a set of points $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \sim$ i.i.d. $F_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$ is the solution to the equations

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} v_1\left(d(\boldsymbol{x}_i, \boldsymbol{t}; \boldsymbol{C})\right)(\boldsymbol{x}_i - \boldsymbol{t}) = \mathbf{0} \\ \frac{1}{n}\sum_{i=1}^{n} v_2\left(d(\boldsymbol{x}_i, \boldsymbol{t}; \boldsymbol{C})^2\right)(\boldsymbol{x}_i - \boldsymbol{t})(\boldsymbol{x}_i - \boldsymbol{t})^T = \boldsymbol{C}. \end{aligned} \tag{2.25}$$

where v_1 and v_2 are weighting functions to be specified later, and, for ease of notation, we define $d(\boldsymbol{x}, \boldsymbol{t}; \boldsymbol{C}) = [(\boldsymbol{x} - \boldsymbol{t})^T \boldsymbol{C}^{-1}(\boldsymbol{x} - \boldsymbol{t})]^{1/2}$. It directly follows that this is an affine equivariant estimate. To make the relation to the univariate definition in equation (2.11) clear, we can write

$$\Psi\left(\boldsymbol{x}, (\boldsymbol{t}, \boldsymbol{C})\right) = \begin{bmatrix} v_1\left(d(\boldsymbol{x}, \boldsymbol{t}; \boldsymbol{C})\right)(\boldsymbol{x} - \boldsymbol{t}) \\ v_2\left(d(\boldsymbol{x}, \boldsymbol{t}; \boldsymbol{C})\right)(\boldsymbol{x} - \boldsymbol{t})(\boldsymbol{x} - \boldsymbol{t})^T - \boldsymbol{C}. \end{bmatrix} \tag{2.26}$$

Then, equation (2.25) become

$$\frac{1}{n}\sum_{i=1}^{n} \Psi\left(\boldsymbol{x}, (\boldsymbol{t}, \boldsymbol{C})\right) = \mathbf{0}. \tag{2.27}$$

To facilitate understanding and comparisons with the univariate case, we define $\psi_i(s) = sv_i(s)$, for $i = 1, 2$.

We consider an example to help establish intuition as to what the M-estimate is doing. First, if we let $v_1(s) = v_2(s^2) = 1$, then no down-weighting is performed and we obtain the sample mean and covariance as our estimates. Now, if we take $\psi_1(s) = \psi_2(s) = \psi_H(s, K)$, i.e.

$$v_1(s) = \psi_H(s, K)/s$$

and

$$v_2(s^2) = \psi_H(s^2, K^2)/s^2,$$

then this behaves like a multivariate extension of the optimal univariate location estimate in the sense that at the solution $(\boldsymbol{t}_n, \boldsymbol{C}_n)$, points further than K from $\boldsymbol{t}_n$ according to Mahalanobis distance $d\,(\boldsymbol{x}, \boldsymbol{t}_n; \boldsymbol{C}_n)$ are "pulled in" to behave as if they were at distance K.

For all the results shown by Maronna (1976), the following four conditions are assumed:

1. $v_1(s)$ and $v_2(s)$ are nonnegative, nonincreasing, and continuous functions for $s \geq 0$;

2. $\psi_1(s)$ and $\psi_2(s)$ are bounded with $K_i = \sup_{s \geq 0}\{\psi_i(s)\}$;

3. $\psi_2(s^2)$ is nondecreasing and is strictly increasing on the interval where $\psi_2 < K_2$;

4. There exists s_0 such that $\psi_2(s_0^2) > p$ thus $K_2 > p$ and that $v_1(s) > 0$ for $s \leq s_0$.

From this point forward, and generally in the literature, use of the term M-estimate implies adherence to these four hypotheses. As an example, Huber's Multivariate Proposal (1964) satisfies the above conditions. Huber's proposal is to take

$$\psi_1(s) = \psi_H(s, k_1)$$

and

$$\psi_2(s^2) = \psi_H(s^2, k_2^2)/\beta,$$

where $\beta = \mathbb{E}_{F_{0,\boldsymbol{I}}}[\psi_H(\|\boldsymbol{x}\|^2, k_2^2)]$, and thus $K_1 = k_1$ and $K_2 = k_2^2/\beta$.

Properties

With the above conditions on the M-estimate, Maronna proves several important properties about the M-estimate he defines.

- For both a continuous distribution and an empirical distribution based on enough data points, the existence of a solution is guaranteed.
- For a unimodal and symmetric distribution around some point there is a unique solution, and under additional restrictions one can include empirical distributions, but only in the univariate case. Maronna conjectures that it holds for $p > 1$, but is not able to prove it. We are unaware of any result which proves his conjecture.

The problem of not having a uniqueness result for empirical distributions is somewhat mitigated by a convergence result for M-estimates. If the distribution $F_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$ producing the points $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ satisfies a probability measure hypothesis and the M-estimate satisfies the above four conditions, then that distribution has a unique M-estimate $(\boldsymbol{t}_*, \boldsymbol{C}_*)$. Furthermore, an M-estimate $(\boldsymbol{t}_n, \boldsymbol{C}_n)$ based on $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ exists for each n sufficiently large. Maronna shows convergence and asymptotic normality of these M-estimates under general regularity conditions. Thus, even though we may not have unique solutions for finite sample sizes, the estimates do converge to a unique solution.

Maronna also calculates the IF for the M-estimate, but only provides it for the location estimate $\boldsymbol{t}$, omitting the IF for the covariance due to its difficult expression. He shows that the IF for location under a spherically symmetric distribution is

$$\mathrm{IF}(\boldsymbol{x}; \boldsymbol{t}, F_{\boldsymbol{0},\boldsymbol{I}}) = c v_1(\|\boldsymbol{x}\|)\boldsymbol{x} \tag{2.28}$$

for a specified constant c which Maronna (1976) derives. This is reminiscent of the univariate case where the IF is proportional to the ψ function. Looking at equation (2.26), one could make the definition

$$\Psi(\boldsymbol{x}, (\boldsymbol{t}, \boldsymbol{C})) = \begin{bmatrix} \Psi_1(\boldsymbol{x}, (\boldsymbol{t}, \boldsymbol{C})) \\ \Psi_2(\boldsymbol{x}, (\boldsymbol{t}, \boldsymbol{C})) \end{bmatrix} = \begin{bmatrix} v_1\left(d(\boldsymbol{x}, \boldsymbol{t}; \boldsymbol{C})\right)(\boldsymbol{x} - \boldsymbol{t}) \\ v_2\left(d(\boldsymbol{x}, \boldsymbol{t}; \boldsymbol{C})\right)(\boldsymbol{x} - \boldsymbol{t})(\boldsymbol{x} - \boldsymbol{t})^T - \boldsymbol{C} \end{bmatrix}$$

Thus defining vector mappings Ψ_1 and Ψ_2 instead of the scalar mappings ψ_1 and ψ_2 that Maronna uses. From this definition, we see that the IF for location is proportional to

$\Psi_1(\boldsymbol{x}) = v_1\left(d(\boldsymbol{x}, \boldsymbol{t}; \boldsymbol{C})\right)\boldsymbol{x}$. Thus, we immediately can infer that the GES of the location estimate is directly controlled by the bounding level applied to Ψ_1. Unfortunately, nothing is said about the form of the IF of the covariance estimate except that its derivation is similar too, but more laborious, than the one for location.

In addition to presenting the IF for location, Maronna also gives an upper bound on the breakdown point of the M-estimate under the following model $G_{\boldsymbol{\mu},\boldsymbol{\Sigma}} = (1-\epsilon)F_{\boldsymbol{\mu},\boldsymbol{\Sigma}} + \epsilon\Delta_{\boldsymbol{y}}$ with $\|\boldsymbol{y}\| \to \infty$. Note that this is not as general as the GEM in equation (2.1), as Maronna models contaminations as point mass distributions at a point near infinity in comparison to the arbitrary distribution in the GEM. The upper bound Maronna gives for the asymptotic value of the breakdown point (Maronna refers the reader to his thesis (1974) for the derivation) is

$$\epsilon^* = \min\{\epsilon^*(\boldsymbol{t}, F_{\boldsymbol{\mu},\boldsymbol{\Sigma}}), \epsilon^*(\boldsymbol{C}, F_{\boldsymbol{\mu},\boldsymbol{\Sigma}})\} \leq \min\left\{\frac{1}{K_2}, 1 - \frac{p}{K_2}\right\}. \tag{2.29}$$

Thus, we see that not only does the clipping level K_2 on ψ_2 affect the breakdown point (recall that for univariate location estimation we have $\epsilon^* = 1/2$ if ψ is bounded and $\epsilon^* = 0$ otherwise), but having either too large or too small a K_2 will lead to low robustness with respect to breakdown point. Note that if $K_2 > p$, the bound on the breakdown point can be written as

$$\epsilon^* \leq \frac{1}{p+1} \tag{2.30}$$

Thus, M-estimates necessarily have a low breakdown point in high-dimensional spaces.

Maronna presents several important theorems which supports practical use of his multivariate M-estimate; however, nothing is said regarding Fisher consistency. He stated that we will converge to a unique solution $(\boldsymbol{t}_*, \boldsymbol{C}_*)$ so long as the conditions have been met, but this solution is not necessarily the parameter $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ of the underlying distribution. Naturally, not all choices of v_1 and v_2 produce Fisher consistent estimates, but Maronna does not present conditions that guarantee Fisher consistency.

A second criticism is the lack of an optimal M-estimate. He presented the IF for the location estimate and elsewhere derives the expression for the covariance estimate. A natural question is to wonder if Hampel's optimality Theorem 2.1 extends to the multivariate M-estimates that Maronna has presented. This combined with a lack of Fisher consistency prevents us from choosing the weighting functions v_1 and v_2 in a principled manner.

2.4 S-Estimates

The M-estimate studied in the previous section is the natural generalization of the univariate M-estimate to multivariate location and covariance. Several nice properties were shown, but a significant deficiency is its low breakdown point in high dimensions. This issued a search for multivariate estimates which possess a high breakdown that is independent of the dimension. This and the following section discuss two of the more popular estimates with this property that have emerged. This section focuses on the S-estimate as described by Lopuhaä (1989) and Davies (1987).

The S-estimate was originally introduced by Rousseeuw and Yohai (1984) in the context of linear regression. They proposed the S-estimate as the solution of the optimization problem

$$\min_{\sigma>0,\boldsymbol{\alpha}\in\mathbb{R}^p}\{\sigma(\boldsymbol{\alpha})\}$$

such that

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{y_i-\boldsymbol{\alpha}^T\boldsymbol{x}_i}{\sigma(\boldsymbol{\alpha})}\right)=b_0$$

where b_0 satisfies $0 < b_0 < a_0 = \sup\{\rho\}$. In setting $\rho(s) = s^2$, a least squares regression is obtained. By bounding ρ, we limit the maximal effect that any point can have, thus robustifying the least squares technique.

Davies (1987) and Lopuhaä (1989) extend this regression estimate to the S-estimate for multivariate location and covariance, although they do so in slightly different ways.

We use Lopuhaä's definition as it is a little simpler. Lopuhaä defines the S-estimate $\hat{\boldsymbol{\theta}} = (\boldsymbol{t}, \boldsymbol{C})$ as the solution to the optimization problem

$$\min_{(\boldsymbol{t},\boldsymbol{C})\in(\mathbb{R}^p,\ \mathrm{PDS}(p))} \{|\boldsymbol{C}|\} \tag{2.31}$$

such that

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(d(\boldsymbol{x}_i, \boldsymbol{t}; \boldsymbol{C})\right) = b_0 \tag{2.32}$$

where b_0 satisfies $0 < b_0 < a_0 = \sup\{\rho\}$. The constant b_0 will effect the scaling of the covariance estimate and should thus be chosen in accordance with the underlying model distribution. In particular, if $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ are from an elliptical distribution $F_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$, then it is natural to choose $b_0 = \mathbb{E}_{F_{\boldsymbol{\mu},\boldsymbol{\Sigma}}}[\rho\left(d(\boldsymbol{x}_i, \boldsymbol{\mu}; \boldsymbol{\Sigma})\right)]$ which is the limit of the average in equations (2.31)–(2.32) if $\boldsymbol{t}$ and $\boldsymbol{C}$ are Fisher consistent. It is straightforward to show that S-estimates are affine equivariant. Thus b_0 can be selected without knowing $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (which of course we do not know) by choosing $b_0 = \mathbb{E}_{F_{0,\boldsymbol{I}}}[\rho\left(\|\boldsymbol{x}\|\right)]$ utilizing only the normalized parametric distribution.

Just as choosing $\rho(s) = s^2$ yields the least square solution for the regression problem, it also produces the least square solution for the location-covariance problem. Choosing $b_0 = p$ for appropriate scaling of the covariance matrix, the S-estimate produces the sample mean and covariance as the unique solution (Grübel, 1988).

Properties

Lopuhaä and Davies prove many of the same results such as existence, convergence, consistency, and asymptotic normality but under different conditions. Davies applies weaker constraints on the ρ function, but only considers elliptical distributions. Lopuhaä's constraints on ρ are slightly more restrictive, but some of his results apply to more general distributions. Because in most conceivable circumstances, one would apply these techniques to a sample set which they believe to have arisen from an elliptical distribution, we will present the properties from Davies.

- Davies first demonstrates the existence of a unique solution of the S-functional at the true distribution and that this solution is the correct value, i.e. it is Fisher consistent. For an elliptical distribution with some conditions the S-functional has the unique solution $(\boldsymbol{t}_*, \boldsymbol{C}_*) = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Although this should be expected of a good estimate, recall that M-estimates were not shown to be Fisher consistent, but instead, only that a unique solution existed for the M-functional.

- Davies then shows that for large enough sample sizes, equations (2.31)–(2.32) have a solution.

- Furthermore, any sequence of these solutions will converge to the true parameter values of the source distribution.

- Davies derives the asymptotic for the S-estimate and shows that it also has a limiting normal distribution.

Lopuhaä derives the IF for the S-estimate for distributions more general than elliptical. For

$$\Psi(\boldsymbol{x}, (\boldsymbol{t}, \boldsymbol{C})) = \begin{bmatrix} \tilde{v}_1\left(d(\boldsymbol{x}, \boldsymbol{t}; \boldsymbol{C})\right)(\boldsymbol{x} - \boldsymbol{t}) \\ p\tilde{v}_1\left(d(\boldsymbol{x}, \boldsymbol{t}; \boldsymbol{C})\right)(\boldsymbol{x} - \boldsymbol{t})(\boldsymbol{x} - \boldsymbol{t})^T - \tilde{v}_3\left(d(\boldsymbol{x}, \boldsymbol{t}; \boldsymbol{C})\right)\boldsymbol{C} \end{bmatrix}$$

where $\tilde{v}_1(s) = \psi(s)/s$, $\tilde{v}_3(s) = \psi(s)/s - \rho(s) + b_0$, and $\lambda_F(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathbb{E}_F[\Psi(\boldsymbol{x}, (\boldsymbol{\mu}, \boldsymbol{\Sigma}))]$, λ_F has a nonsingular derivative Λ at $(\boldsymbol{t}(F), \boldsymbol{C}(F))$.
Then the IF exists and is

$$\mathrm{IF}(\boldsymbol{x}; S, F) = -\Lambda^{-1}\Psi\left(\boldsymbol{x}, [\boldsymbol{t}(F), \boldsymbol{C}(F)]\right). \tag{2.33}$$

Now if $F = F_{\boldsymbol{0},\boldsymbol{I}}$ is the spherical distribution and ρ satisfies some conditions then, the IF for the associated location S-estimate is

$$\mathrm{IF}(\boldsymbol{x}; \boldsymbol{t}, F_{\boldsymbol{0},\boldsymbol{I}}) = c\psi(\|\boldsymbol{x}\|)\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} = c\tilde{v}_1(\boldsymbol{x}\|)\boldsymbol{x} \tag{2.34}$$

for a constant c given by Lopuhaä. As expected, this expression is parallel that of equation (2.28) for M-estimates.

Thus far, we have one significant advantage of S-estimates over M-estimates, i.e. consistency. However, Davies demonstrates an even more significant attribute of S-estimates, which is the ability to achieve a high breakdown point independent of dimension defined as,

$$\epsilon_n^* = \min\{\epsilon_n^*(\boldsymbol{t}, F_{\boldsymbol{\mu},\boldsymbol{\Sigma}}), \epsilon^*(\boldsymbol{C}, F_{\boldsymbol{\mu},\boldsymbol{\Sigma}})\} = \frac{[nb_0/a_0]+1}{n}. \tag{2.35}$$

Two simple, but important corollaries follow here.

Corollary 2.1

$$\epsilon^* = \lim_{n\to\infty} \epsilon_n^* = b_0/a_0. \tag{2.36}$$

Corollary 2.2 *Setting* $b_0/a_0 = 1/2 - (p+1)/2n$ *yields*

$$\epsilon_n^* = \frac{\frac{[n-p+1]}{2}}{n} \tag{2.37}$$

which is the upper-bound for the breakdown point of affine equivariant estimates, which is proved in Davies (1987).

The S-estimate is a significant improvement on the M-estimate in two ways. First, with an appropriate choice of ρ, it can achieve the maximal breakdown point which is asymptotically 1/2 regardless of dimension. This contrasts with the M-estimates increasing susceptibility to outliers as dimensionality increases. Second, Fisher consistency is proven for S-estimates along with a stronger convergence proof. For elliptical distributions, M-estimates are shown to converge in probability to a unique solution though not necessarily the underlying parameters, whereas S-estimates are shown to converge almost surely to the underlying parameters.

A significant drawback of the S-estimate is that there is no optimality theorem on how to choose the function ρ. Lopuhaä uses as an example Tukey's Biweight function

$$\rho_T(s, c_0) = \begin{cases} \frac{s^2}{2} - \frac{s^4}{2c_0^2} + \frac{s^6}{6c_0^4} & \text{if } |s| \leq c_0 \\ \frac{c_0^2}{6} & \text{if } |s| \geq c_0, \end{cases}$$

but there is no justification for its selection, other than it resembles a smooth clipped parabola, thus approximating robust least-squares estimation. Using Tukey's Biweight function in the S-estimate and Huber's proposed M-estimate, Lopuhaä performs simulations comparison of the two. The general trend is that at the model distribution, the M-estimate achieves a lower error variance, but when the model distribution does not match the actual distribution, the S-estimate performs better. These simulations must be cautiously interpreted however because there may be better choices of weighting functions which could result in a different conclusion.

2.5 MCD Estimate

In this section, we will discuss the third of the three classes of estimates surveyed, the MCD estimate. We mostly follow Butler et al. (1993). The previous two estimates are rather abstract in that the M-estimate is the solution of a system of nonlinear equations, and the S-estimate is the solution of a nonlinear optimization problem. The MCD technique presented here is much more intuitive. Given $0.5 < \alpha \leq 1$, the MCD estimate can be described as follows. Consider all subsets of $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ of size $h = [\alpha n]$, where $[x]$ denotes the greatest integer smaller than or equal to x. For each of these sets, compute their sample mean and covariance. Then, the MCD estimate $(\boldsymbol{t}_n, \boldsymbol{C}_n)$ is the sample mean and covariance from the set whose sample covariance has the minimum determinant.

As intuition would lead to believe, the MCD corresponds to finding the ellipsoid which covers the most dense cluster of h points and then taking a weighted sample mean and covariance where the weighting is the indicator function (divided by h for normalization) of the ellipsoid. Inherent in this is an assumption of unimodality due to the clustering. For the rest of the material presented in this section, we assume that the underlying distribution $F_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$ is unimodal and elliptical. Thus, we can assume it has a density of the form

$$f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x}) = |\boldsymbol{\Sigma}|^{-1/2}\varphi\left((\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{2.38}$$

with $\varphi : \mathbb{R}_+ \mapsto \mathbb{R}_+$ being a decreasing function appropriate scaled so that the density is valid. Furthermore, because the MCD is affine equivariant (which follows from its use of the sample mean and covariance), we can assume the parameters of the underlying distribution are $\boldsymbol{\mu} = \boldsymbol{0}$ and $\boldsymbol{\Sigma} = I$.

Properties

Although the MCD has an intuitively simple interpretation, there are not many results on its properties. The existence of a solution for sample sets is obvious as it takes a minimal covariance from a finite set of solutions. This solution can be seen to be unique (with probability 1) but only if the underlying distribution is continuous. Butler et al. show the existence and uniqueness of a solution under the model distribution. As one would expect, the solution is the sphere centered at the origin in accordance with the underlying distribution $(\boldsymbol{\mu},\boldsymbol{\Sigma}) = (\boldsymbol{0},\boldsymbol{I})$. For further details see Butler et al. (1993).

Butler et al. also studied the convergence of the MCD estimates. He found that the MCD estimate of location converges to the true value, but the covariance estimate does not. Like the M-estimate and S-estimate, the MCD location estimate is asymptotically normal with asymptotic covariance matrix $c\boldsymbol{\Sigma}$ for some constant c. The constant c is given by Butler et al. (1993).

The breakdown point of the MCD estimate has not been mentioned in Butler's article, but due to the straightforwardness of its definition, it is readily apparent that if $n \geq p+1$

then

$$\epsilon_n^*(\boldsymbol{t}) = \frac{n-h+1}{n} \to 1-\alpha \tag{2.39}$$

$$\epsilon_n^*(\boldsymbol{C}) = \min\left\{\frac{n-h+1}{n}, \frac{h-p}{n}\right\} \to \min\{1-\alpha, \alpha\}. \tag{2.40}$$

The breakdown point of the location estimate follows from that if h of the original points remain, then the location estimate will still be finite, but if fewer than h points remain, then the estimate can become unbounded. The same reasoning applies for keeping the maximal covariance eigenvalue bounded. However, we must also consider the other direction of breakdown where the minimum eigenvalue approach zero. Here, we only need to replace $h-p$ points to force an eigenvalue to zero. In particular, we find the p densest sample of points whose convex hull contains no other data points. Then we replace any $h-p$ of the other points with points lying at the center of this convex hull, which is contained within a hyperplane by definition. Now, the most dense cluster of h points is the cluster of points in this convex hull and the covariance estimate has a zero eigenvalue.

Note that, one can choose α so that they can get an asymptotic breakdown point of $\epsilon^* = 1/2$. However, it is clear that by choosing a smaller α yielding a larger ϵ^*, one is utilizing fewer points in the sample statistics and thus will have a higher error variance on the estimates. Choosing h as small as possible, i.e. $h = [(n+1)/2]$, yields the maximal location breakdown point,

$$\epsilon_n^*(\boldsymbol{t}) = \frac{[(n+1)/2]}{n}. \tag{2.41}$$

Similarly, choosing $h = [(n+p+1)/2]$ yields the maximal covariance breakdown point,

$$\epsilon_n^*(\boldsymbol{C}) = \frac{[(n-p+1)/2]}{n}. \tag{2.42}$$

Thus, the MCD can achieve the maximal location and covariance breakdown points, but not at the same time. This is not really a hindrance since the difference is minuscule and both breakdown points asymptotically approach 1/2.

The computational complexity is a major issue regarding the MCD. To find an exact solution requires searching the entire space of all possible groupings of h out of n data samples. Thus, the computational burden grows combinatorially with the sample size. However, there are fast approximations to the MCD, the most prevalent of which is proposed by Rousseeuw and Van Driessen (1999) and briefly described in the next section. The algorithm is an approximation to the MCD and produces only a locally optimal solution. An ellipse $\mathcal{E}$ is considered locally optimal if the determinant of its sample covariance cannot be decreased by switching one point in $\mathcal{E}$ with a point not in $\mathcal{E}$.

The development and availability of fast algorithms Hawkins (1994); Rousseeuw and Van Driessen (1999) for computing the minimum covariance determinant (MCD) has brought renewed interest to this estimate. Asymptotic properties were given in Bulter et al. (1993), but the asymptotic variance of the MCD scatter part remained unknown. In the particular case of one dimension the influence function of the MCD scale was computed by Croux and Rousseeuw (1992a). Croux and Haesbroeck (1999) worked out the influence function of the MCD scatter matrix estimate in arbitrary dimensions, and used it to evaluate the asymptotic efficiency of this estimate. It follows that the MCD scale estimate has a bounded influence function, which is re-descending to zero for the off-diagonal elements, but not for the on-diagonal elements. It is not sufficient to consider only breakdown point and efficiency of robust estimates, but maxbias curves should be computed. This has been done for the MCD estimate in the univariate case by Croux and Haesbroeck (1999), but the multivariate case seems to be rather hard to handle.

The MCD is straightforward to compute and can be approximated with a fairly fast algorithm as described in the next section.

FAST-MCD Algorithm

We now give a brief overview of the FAST-MCD algorithm proposed by Rousseeuw and Van Driessen (1999). As in the previous section, we will let h denote the size of the subsets to examine. The basis for their algorithm is the following theorem.

THEOREM 2.2 *Let H_1 be a subset of $\{1, \ldots, n\}$ of size h with associated sample statistics*

$$\boldsymbol{t}_n^1 = \frac{1}{h} \sum_{i \in H_1} \boldsymbol{x}_i \tag{2.43}$$

$$\boldsymbol{C}_n^1 = \frac{1}{h} \sum_{i \in H_1} (\boldsymbol{x}_i - \boldsymbol{t}_n^1)(\boldsymbol{x}_i - \boldsymbol{t}_n^1)^T. \tag{2.44}$$

If $|\boldsymbol{C}_n^1| > 0$ then define the distances $d_1(i) = d(\boldsymbol{x}_i, \boldsymbol{t}_n^1; \boldsymbol{C}_n^1)$. Now, define the set H_2 to be those points with the h smallest distances $d_1(i)$, i.e. $\{d_1(i) \mid i \in H_2\} = \{(d_1)_1, (d_1)_2, \ldots, (d_1)_h\}$ where $(d_1)_1 \leq (d_1)_2 \leq \ldots \leq (d_1)_n$ are the ordered distances. Now, define $\boldsymbol{t}_n^2$ and $\boldsymbol{C}_n^2$ as in equations (2.43)–(2.44) but with H_2 in place of H_1. Then,

$$|\boldsymbol{C}_n^2| \leq |\boldsymbol{C}_n^1|. \tag{2.45}$$

The construction of H_2 from H_1 is called the C-step where C stands for covariance. Thus, recursively defining sets H_i by repeatedly taking the points which minimize the Mahalanobis distance based on the previous iteration leads to a local minimum (the determinant of $\boldsymbol{C}_n^i$ is a local minimum if it cannot be decreased by switching one point in H_i with a point not in H_i) of the covariance determinant. Using this intuitive principle, the FAST-MCD algorithm can be described as follows:

1. Initialize H_0 by randomly selecting $p+1$ points. Compute the sample mean $\boldsymbol{t}_n^0$ and covariance $\boldsymbol{C}_n^0$ for H_0. Construct the set H_1 as in Theorem 2.2, i.e. choose the h points which minimize the Mahalanobis distance with respect to $\boldsymbol{t}_n^0$ and $\boldsymbol{C}_n^0$.

2. Perform k C-steps (Rousseeuw and Van Driessen recommend $k = 2$ C-step).

3. Perform steps 1 and 2 many times and take the l best solutions H_k^l which have the smallest covariance determinant (Rousseeuw and Van Driessen recommend $l = 10$). For each of these "survivors" repeatedly perform the C-step until convergence (which is guaranteed by Theorem 2.2 and boundedness of the determinant).

4. Take as your solution, the H_∞^l which has the minimum covariance determinant.

For larger sample sets, Rousseeuw and Van Driessen recommend partitioning the data set into several groups and running the FAST-MCD on each of them, then taking the m best estimates from each group and running the FAST-MCD on the entire data set initialized with each of the m solutions from each group and taking the minimum result as the solution.

2.6 Conclusions

The focus of this chapter is primarily on three techniques: the M-estimate, S-estimate and MCD. The first estimate discussed is the M-estimate, which is a generalization of the well known M-estimate for univariate location. The M-estimate is defined by two weighting functions v_1 and v_2 which control the influence of outliers on the location and covariance estimates. The M-estimate is then the solution of a system of equations of the weighted sample moments. Maronna (1976) shows several important properties of the M-estimate under certain conditions on the weighting functions and distribution. In particular, existence, uniqueness and convergence are shown; although, the convergence is not necessarily to the true underlying parameters as conditions for Fisher consistency are not shown. Another notable characteristic of the M-estimate is its low breakdown point which decreases with increasing dimensionality of the data. Furthermore, the conditions on the weighting functions seem to imply an inherent assumption of unimodality on the underlying distribution.

The second estimate surveyed is the S-estimate, which originated in the context of linear regression, but is extended to multivariate location and covariance estimation. This estimate obtains its solution via an optimization problem which minimizes the determinant of the covariance estimate subject to a cost constraint which can be interpreted as the sum of weighted Mahalanobis distances of the samples under the covariance and location estimates. Robustness is endowed to this estimate by limiting the maximal cost that a single data sample can contribute and thus limiting its influence on the estimate. Of the three estimates, this perhaps has the most properties shown about it. These include existence, uniqueness, Fisher consistency and convergence of estimates to the true parameter values. Furthermore, S-estimates are shown to be maximally robust with respect to the breakdown point. S-estimates also satisfy the form of M-estimates, but violate the constraints on the weighting functions and are thus not considered M-estimates.

The MCD estimate is based on the intuition that for unimodal elliptical symmetric distributions, the most reliable points on which to base the estimate are those which are closely clustered. The MCD estimate finds the subset of data points which has the smallest sample covariance and takes as its estimates, the sample mean and covariance from this set. Unlike the M-estimate and S-estimate, there are no weighting/cost functions to choose here, only the cluster size parameter α.

All three of these estimates asymptotically converge to limiting values with a Gaussian distribution. Furthermore, the variance on the "error" goes down as $1/n$. Note that there are other well known estimates which actually have a slower decay on the variance, e.g. the asymptotic variance of the minimum volume ellipsoid (MVE) estimate goes down as $n^{-2/3}$ and does not converge in distribution to a Gaussian.

Chapter 3

Multivariate Contamination Models

3.1 Classical Contamination Model

Statisticians use *contamination or mixture models* to study the performance of robust alternatives to classical statistical procedures when these procedures are applied to messy data sets that contain outliers. Most studies on robustness in statistics are centered around a concept of contamination introduced by Tukey (1960). The best known and most broadly used contamination model is the so called ϵ-*contamination neighborhood* introduced by Tukey (1962) and extended by Huber (1964). These models can be thought of as "testing grounds" where statistical procedures are tested and continuously improved. The contamination model was originally introduced to handle one-dimensional data. Assume that, given a sample $X_1, \dots, X_n$, the majority of the data follows the nominal distribution H_0 while a small fraction ϵ follows an arbitrary distribution $\tilde{H}$. This contamination model, which we will call *classical contamination model*, can be written as

$$H(x) = (1 - \epsilon) H_0(x) + \epsilon \tilde{H}(x), \qquad 0 \leq \epsilon < \frac{1}{2}. \tag{3.1}$$

For example H_0 may be a normal distribution with mean μ and standard deviation σ, i.e., $H_0 = N(\mu, \sigma)$, and $\tilde{H}$ an arbitrary distribution. Robustness in the sense of the Princeton's Group (Tukey, Huber, Hampel et al., etc.) addresses the estimation of the dominant component H_0 and clearly, for this component to be dominant, the amount of contamination ϵ must be less than one half.

The classical contamination model (3.1) has been adopted for multivariate data sets as well (see for example Rocke and Woodruff, 1996), although it is not necessarily appropriate in that context. Given a sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, where $\boldsymbol{X}_i \in \mathbb{R}^p, i = 1, \ldots, n$, the majority of the data follows the nominal distribution H_0 while a small fraction ϵ follows an arbitrary distribution $\tilde{H}$. Then, the classical multivariate contamination model can be written as

$$H(\boldsymbol{x}) = (1-\epsilon)H_0(\boldsymbol{x}) + \epsilon\tilde{H}(\boldsymbol{x}), \qquad 0 \leq \epsilon < \frac{1}{2}. \tag{3.2}$$

For example H_0 may be a multivariate normal distribution with mean $\boldsymbol{\mu}$ and scatter matrix $\boldsymbol{\Sigma}$, i.e., $H_0 = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and $\tilde{H}$ an arbitrary distribution. Under this model a fraction $(1-\epsilon)$ of the cases on average are distributed according to H_0 and are therefore the majority or "core" data, while a fraction ϵ of the cases are from $\tilde{H}$ and generate outliers that deviate from the core behavior of the data. Hence, in this model each data point is either "100 % perfect" coming from H_0 or "100 % spoiled" coming from $\tilde{H}$.

An alternative representation of the classical multivariate contamination model (3.2) can be written as

$$\boldsymbol{X} = (1-b)\mathbf{Y} + b\boldsymbol{Z}, \tag{3.3}$$

where $\boldsymbol{Y}, \boldsymbol{Z}, b$ are independent with

$$\begin{aligned} \boldsymbol{Y} &\sim H_0 \\ \boldsymbol{Z} &\sim \tilde{H} \\ b &\sim \text{BINOMIAL}(1, \epsilon). \end{aligned}$$

This can be shown as follows.

$$\begin{aligned} P(\boldsymbol{X} \leq \boldsymbol{x}) &= P((1-b)\boldsymbol{Y} + b\boldsymbol{Z} \leq \boldsymbol{x}) \\ &= P(\boldsymbol{Z} \leq \boldsymbol{x})P(b=1) + P(\boldsymbol{Y} \leq \boldsymbol{x})P(b=0) \\ &= \tilde{H}(\boldsymbol{x})\epsilon + H_0(\boldsymbol{x})(1-\epsilon) = H(\boldsymbol{x}). \end{aligned}$$

The classical multivariate contamination model (3.2) is a model of "independent mixture of two independent populations", for example a "normal population" and an "abnormal outliers generating population". Unfortunately this model does not adequately represent reality for many multivariate data sets that arise in practice. It may often happen in applications that outliers occur in each of the variables independently of the other variables or in special dependency patterns other than the complete dependency pattern which, we will see, the classical multivariate contamination model enforces. In addition, the classical multivariate contamination model (3.3) does not allow the possibility of dependency among the uncontaminated vector, $\boldsymbol{Y}$, the *contamination indicator*, b, and the contamination vector, $\boldsymbol{Z}$.

Now we present some real data examples from the literature that illustrate the need for a more general and flexible multivariate contamination model.

3.2 Real Data Examples

EXAMPLE 3.1 **Hertzsprung-Russell Diagram of the Star Cluster CYG OB1**
Consider the Hertzsprung-Russell data set (see Rousseeuw and Leroy, 1987). This two-dimensional data set consists of 47 stars in the direction of Cygnus. The first variable is the effective temperature at the surface of the star and the second variable is its light intensity. The scatterplot of the logarithm of the light intensity versus the logarithm of the temperature is shown in Figure 3.1. We can see from the plot that the data have two groups of points, the majority which seems to follow a steep band and the four stars in the upper left corner. These groups are well known in astronomy. The majority of the points are said to lie on the main sequence and astronomers explain the four points with indices 11, 20, 30 and 34 as giant stars.

EXAMPLE 3.2 **Body and Brain Weights**
Consider the brain and body weight of 28 animals as published in Rousseeuw and

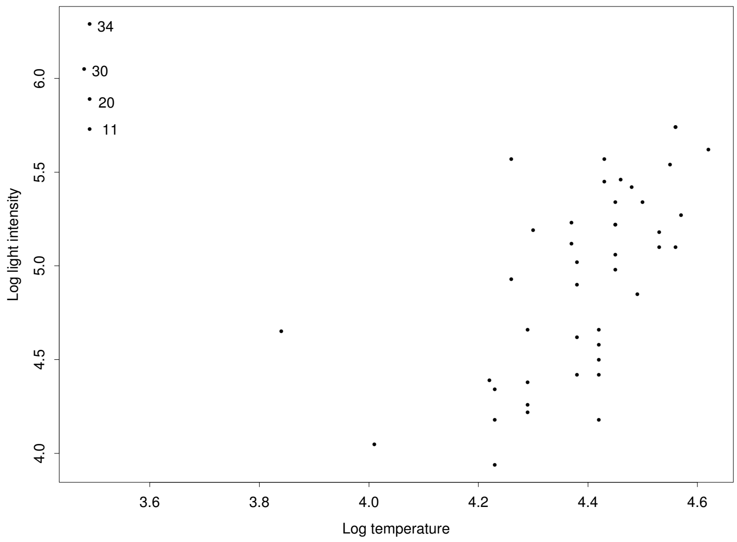


Figure 3.1: Hertzsprung-Russell Diagram of the Star Cluster CYG OB1.

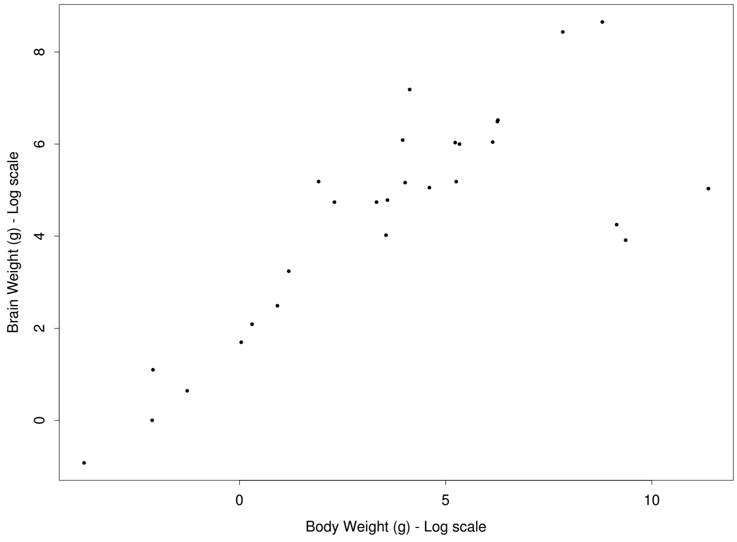


Figure 3.2: Body and Brain Weight Data Set.

Leroy (1987, page 57). This sample was taken from larger data set in Weisberg (1985). Figure 3.2 contains a scatter plot of the logarithm of the brain weight versus the logarithm

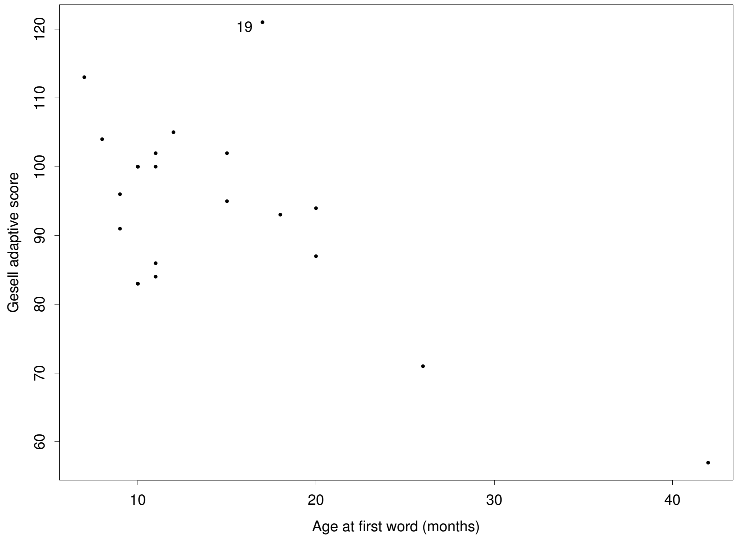


Figure 3.3: Gesell Adaptive Score versus Age at First Word.

of the body weight. The scatter plot shows that the majority of the data follow a clear pattern, except the three observations in the lower right region. These three observations correspond to dinosaurs, each of which possessed a small brain with a heavy body.

Example 3.3 **Gesell Adaptive Score**

This data set was first reported by Mickey et al. (1967) and widely cited in the statistical literature. We obtained the data set from Rousseeuw and Leroy (1987). The study was conducted on 21 children, giving their age (in months) at first spoken word and a score which is a measure of the development of the child. The Gesell adaptive assessment is a standard procedure for direct observation of a child's growth and development. The Gesell assessment is conducted by a trained examiner who makes discriminating observations of a child's behavior and then evaluates these observations by comparison with normal behavior patterns. A normal behavior pattern is a criterion of maturity which has been defined by systematic studies of the average healthy course of child development. The scatterplot of Gesell adaptive score versus age at first word is shown in Figure 3.3.

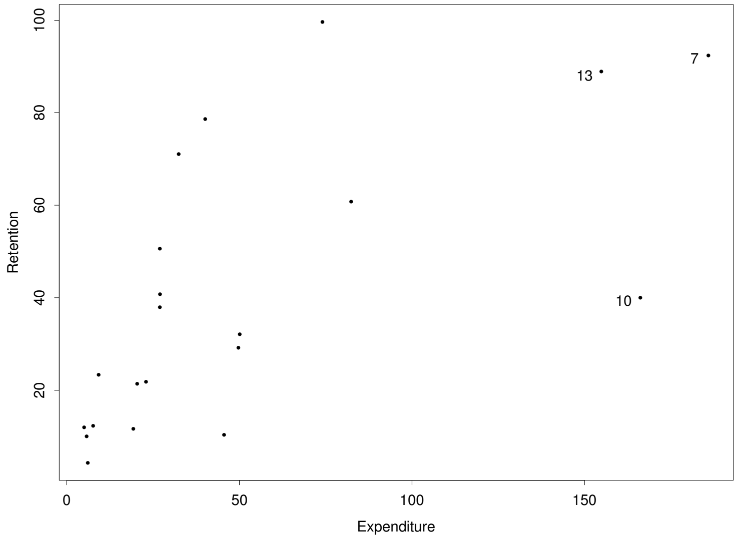


Figure 3.4: Advertising Yield versus Spending.

Case 19 does not follow the general pattern of the remaining data points. Mickey et al. (1967) also decided that this observation is an outlier, by mean of a sequential approach to detect outliers via stepwise regression. Since the value of the score is subjective, this outlier could be explained due to a possible error in the observed Gesell adaptive score given to the child.

Example 3.4 **TV Ad Yields**

Consider the TV Ad Yields data of 21 advertisements published in the Wall Street Journal, March 1, 1984 (available at http://lib.stat.cmu.edu/DASL/Datafiles/tvadsdat.html). The advertisements were selected by an annual survey conducted by Video Board Tests, Inc., a New York ad-testing company, based on interviews with 20,000 adults who were asked to name the most outstanding TV commercial they had seen, noticed and liked. The retained impressions were based on a survey of 4,000 adults in which regular product users were asked to cite a commercial they had seen for that product category in the past week. Figure 3.4 contains a scatterplot of the retention versus the expenditure,

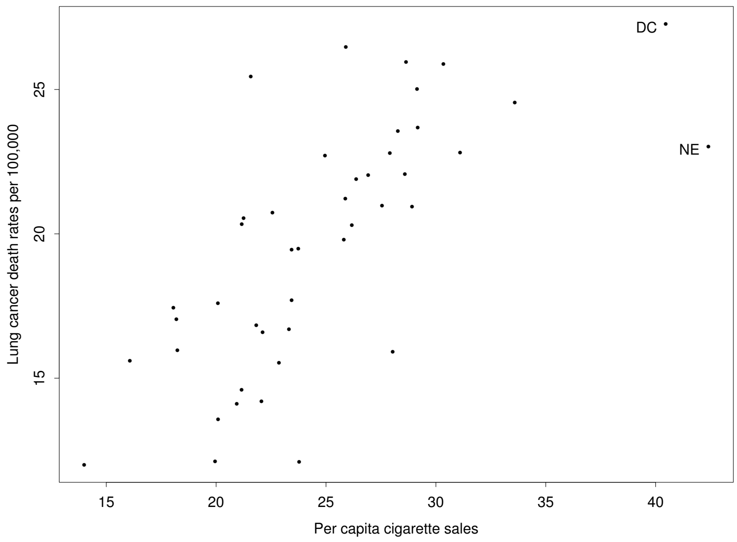


Figure 3.5: Scatterplot of Cigarette Consumption versus Lung Cancer.

which is the TV advertising budget, 1983 ($ Millions). The scatterplot shows that the majority of the data follows an increasing pattern, except the three observations with indices 7, 10 and 13 (McDonald's, Ford and ATT/BELL). These outliers are likely due to an unsuccessful advertising campaign, and therefore the retention figures are lower than expected for the large amount of expenditure. Specifically, the point with index 10 appears to have a low retention for an extremely large value of expenditure.

Example 3.5 **Smoking and Cancer**

Researchers wanted to examine the effect of smoking on cancer development. Data for 43 states and the District of Columbia were collected on per capita numbers of cigarettes smoked (sold) in 1960 together with death rates per thousand population from various forms of cancer, see Fraumeni (1968). Scatterplots of the cigarette consumption versus the lung cancer and the leukemia death rates are shown in Figures 3.5 – 3.6, respectively. From the scatterplots we can see that Nevada (NE) and the District of Columbia are outliers in the distribution of cigarette consumption (sale) per capita. The ready expla-

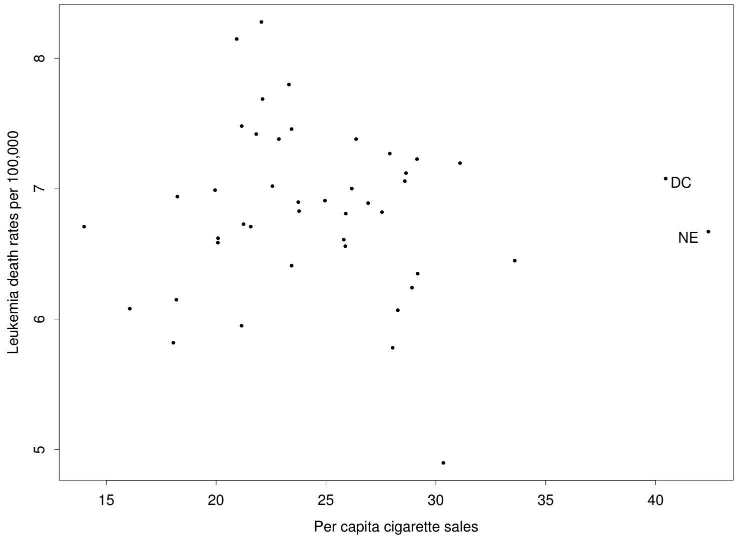


Figure 3.6: Scatterplot of Cigarette Consumption versus Leukemia Death Rates.

nation for the outliers is that cigarette sales are swelled by tourism (Nevada) and tourism and commuting workers (District of Columbia).

Example 3.6 **Air Pollution and Mortality**

Researchers at General Motors collected data on 60 United States Standard Metropolitan Statistical Areas (SMSAs) in a study of whether air pollution contributes to mortality. The variable of main interest is age adjusted mortality and is labelled “Mortality”. The data include variables measuring demographic characteristics of the cities, variables measuring climate characteristics and variables recording the pollution potential of three different air pollutants. These properties were collected from a variety of sources and they are available at http://lib.stat.cmu.edu/DASL/Datafiles/SMSA.html. Figure 3.7 shows all pairwise scatterplots of the variables; mortality, median education, population density, percentage of non-whites, annual rainfall (inches) and logarithm of the Nitrous Oxide (NOx). Clearly these data have several multidimensional outliers that show up as a cluster in several of the scatterplots.

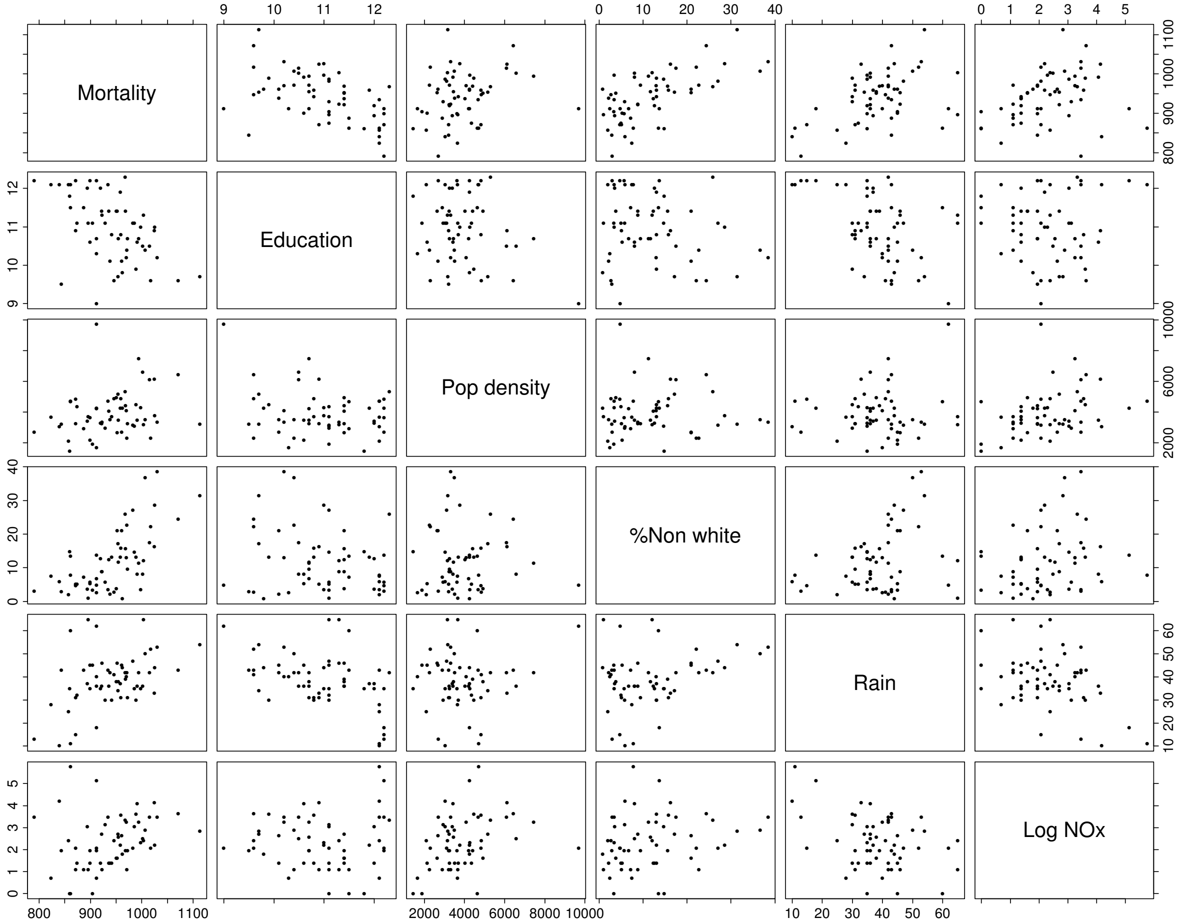


Figure 3.7: Air Pollution and Mortality Data Set.

EXAMPLE 3.7 **Salinity**

Consider the salinity data set (see Ruppert and Carroll, 1980). These data consist of measurements of water salinity (i.e. its salt concentration) and river discharge taken from North Carolina's Pamlico Sound. Figure 3.8 shows all pairwise scatterplots of the variables; water salinity, salinity lagged by two weeks, the trend which is the number of biweekly periods elapsed since the beginning of the spring season and the volume of river discharge into the Sound. Carroll and Ruppert (1985) described the physical background of the data. The scatterplots of the salinity lagged, the trend and water salinity versus

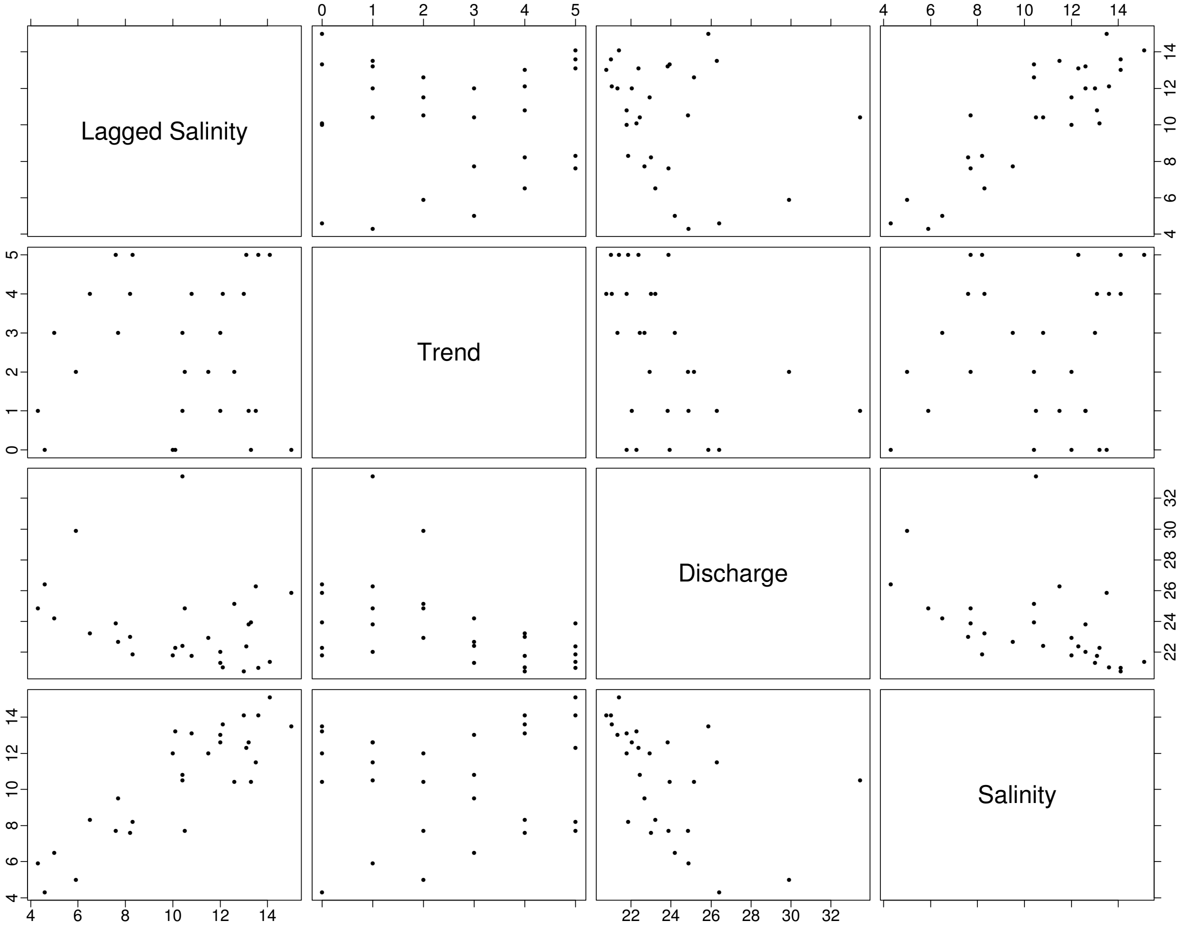


Figure 3.8: Salinity Data Set.

the volume of river discharge each show two points that do not follow the pattern of the remaining data. Carroll and Ruppert (1985) indicate that these outliers are cases 5 and 16 which correspond to periods of very heavy discharge.

Example 3.8 **Wages and Hours**

The data (available at http://lib.stat.cmu.edu/DASL/Datafiles/wagesdat.html) are from a national sample of 6000 households with a male head and earnings of less than $15,000 annually in 1966. Thirty-nine demographic subgroups were formed for analysis of the relation between average hours worked during the year and average hourly wages ($) and

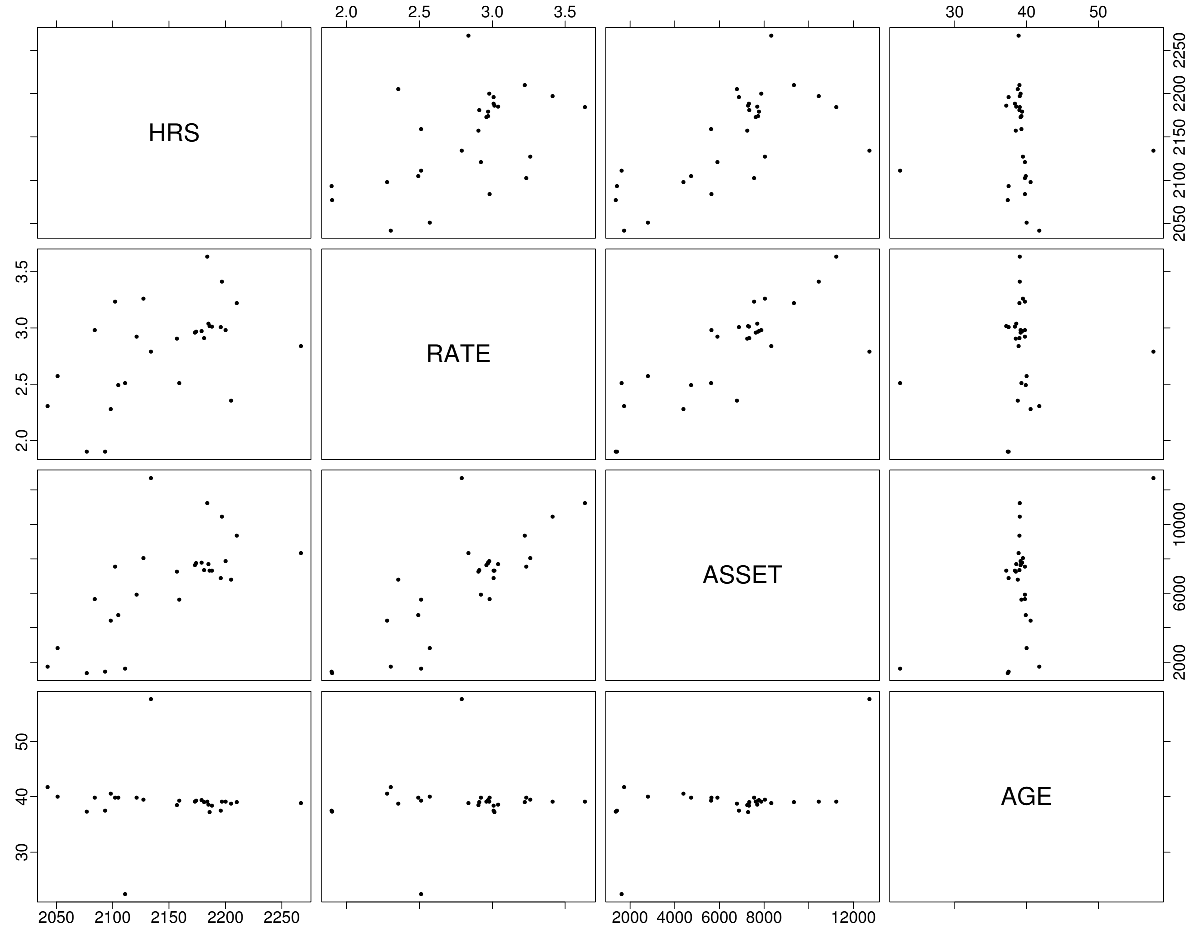


Figure 3.9: Wages and Hours Data Set.

other variables. The study was undertaken in the context of proposals for a guaranteed annual wage (negative income tax). At issue was the response of labor supply (hours worked) to increasing income and effective hourly wages. Figure 3.9 shows all pairwise scatterplots of the average hours, average wages (rate), average family asset holdings and average age of the respondent. Clearly these data have several multidimensional outliers. In particular, the scatterplots of the average hours, the average wages and the average asset versus the age each show outliers in the left and right of the plots. It appears that the age variable has a certain range and any age outside this range shows as an outlier.

Discussion

The Tukey-Huber contamination model (3.2) seems appropriate for the data in Examples 3.1 and 3.2. In the case of Example 3.1 we can assume that there are two subpopulations of stars (the main sequence and the giant stars) and that the given measurements correspond to either one of these two subpopulations with probabilities $(1-\epsilon)$ and ϵ, where ϵ represents the proportion of giant stars. Hence, we can consider these data as coming from an independent mixture of two independent populations, as required by the classical multivariate contamination model (3.2). In the case of Example 3.2 there are also two subpopulations (regular animals and dinosaurs) and the given measurements correspond to each one of these two subpopulations with probabilities ϵ and $(1-\epsilon)$. Since we do not know the criterion used to include animals in the data set, the interpretation of ϵ is less clear, in this case.

It seems difficult, however, to justify the use of the independent mixture model (3.2) for the remaining examples.

In Example 3.3, it would be hard to imagine the existence of a subpopulation of children from which the outlying Case 19 has been drawn. A more likely scenario is that the Gessel adaptive score for this child has been erroneously assigned or recorded. A point to notice here is that one of the two variables (namely, Gessel adaptive score) appears to have a larger probability of gross errors and unusual values than the other variable (namely, age at first spoken word). If the contaminating distribution were to retain the value for the variable age at first spoken word and only contaminate the variable Gessel adaptive score, then the assumption of independence would be violated. A serious limitation of the classical multivariate contamination model (3.2) is the requirement that each data point is either 100% perfect or 100% spoiled.

Model (3.2) is also restrictive in that it fails to allow for cases where the probability of occurrence of discordant values and gross errors in some of the variables depend on the values of other variables. For instance, in Example 3.4, the generally increasing pattern observed for the majority of the data ceases to apply for cases 7, 10 and 13 (McDonald's, Ford and ATT/BELL). Case 10 has a suspiciously low retention value and may be a gross error. The main point in this example is that the outliers were likely to be produced by "problems" affecting the variable retention when the variable expenditure assumes extreme values. Example 3.7 and Example 3.8 are other examples that represent this situation (extreme values of one of the variables may cause the occurrence of outliers or gross errors in other variables). In Example 3.7, it is known that the outliers, cases 5 and 16, correspond to periods of very high discharge. These extreme values of the variable discharge may have affected the values of the other variables. For instance, the variable salinity for cases 5 and 16 has high values compared to the general decreasing pattern as shown in the plot in Figure 3.8. In Example 3.8, each plot in Figure 3.9 involving the variable age has two unusually extreme values. These values may have caused outliers in some of the other variables, due to the fact that the general pattern does not apply when age is too high or low.

The assumption of independent mixture of two independent populations would also be inappropriate in Example 3.5 where the variables were collected from independent sources (agencies) and the outliers are likely to be due to special circumstances regarding cigarette sales in the District of Columbia and Nevada. A similar situation (measurements from independent sources) arises in Example 3.6.

3.3 New Contamination Model

The given examples highlight the need for a more flexible contamination model. Very often, the p-dimensional observation vector collects measurements from different sources, each being inclined to have its own statistical errors. There are also situations where there is a strong dependency between the contaminated and uncontaminated entries and between the uncontaminated entries and their contamination indicator. For example, extreme values of one or more of the variables may increase the likelihood of outliers or gross errors in other variables.

Suppose that $\mathbf{Y} \in \mathbb{R}^p$ has an elliptical distribution H_0 with center $\boldsymbol{\mu}$ and scatter matrix $\boldsymbol{\Sigma}$, for instance $H_0 = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We consider situations when sometimes $\mathbf{Y}$ cannot be perfectly measured and the actual observations can be represented by the contamination model:

$$\mathbf{X} = (I - B)\mathbf{Y} + B\mathbf{Z}, \tag{3.4}$$

where I is a $p \times p$ identity matrix, $\mathbf{Z}$ is an arbitrary random vector and

$$B = \begin{pmatrix} B_1 & 0 & \dots & 0 \\ 0 & B_2 & \dots & 0 \\ & & \ddots & \\ 0 & 0 & \dots & B_p \end{pmatrix} = \text{Diag}(B_1, B_2, \dots, B_p) \tag{3.5}$$

is such that each B_j $(j = 1, \dots, p)$ is a Bernoulli random variable with

$$P(B_j = 1) = 1 - P(B_j = 0) = \epsilon_j, \qquad j = 1, \dots, p. \tag{3.6}$$

That is, ϵ_j represents the probability that the j-th entry of $\boldsymbol{X}$ is contaminated. Notice that in this model the *contamination indicator matrix*, B, is allowed to depend on the uncontaminated vector, $\boldsymbol{Y}$, therefore, $\epsilon_j = \mathbb{E}\{P(B_j = 1|\boldsymbol{Y})\}$. Moreover, the contamination vector, $\boldsymbol{Z}$, can depend on the contamination indicator matrix, B, and the uncontaminated vector, $\boldsymbol{Y}$. Thus, the contamination model (3.4) can be expressed in the following hierarchical way.

$$\boldsymbol{Y}, \quad B|\boldsymbol{Y}, \quad \boldsymbol{Z}|B, \boldsymbol{Y}.$$

Notice that different values of ϵ_j and different dependence structures of the diagonal elements of the contamination indicator matrix, B, generate different contamination patterns. This will be further discussed in Section 3.4 of this chapter. To gain some insight into the contamination model (3.4) we will consider some special situations.

1. **Tukey-Huber contamination model**. If $\boldsymbol{Y}, B$ and $\boldsymbol{Z}$ are independent, and the diagonal matrix B of Bernoulli random variables has the special completely dependent structure

$$P(B_1 = B_2 = \ldots = B_p) = 1, \tag{3.7}$$

then the contamination model generating $\boldsymbol{X}$ reduces to the Tukey-Huber multivariate mixture distribution (3.2). In this case the contamination model represents independent mixture of two independent populations, the normal population and the abnormal population. The Hertzsprung-Russell data set and the body and brain weights data set are examples of this situation.

2. **Independent-contamination model**. In this case $\boldsymbol{Y}, B$ and $\boldsymbol{Z}$ are independent. That is the probability of contamination for the different entries of $\boldsymbol{X}$ does not depend on the uncontaminated vector, $\boldsymbol{Y}$. Also, the contamination vector, $\boldsymbol{Z}$, does not depend on which entries are being contaminated and their values. This situation can be expressed as,

$$\boldsymbol{Y}, \quad B, \quad \boldsymbol{Z}.$$

3. B **and** $\boldsymbol{Z}$ **are independent of** $\boldsymbol{Y}$. In this case the probability of contamination for the different entries of $\boldsymbol{X}$ and the contamination vector, $\boldsymbol{Z}$, are independent of the uncontaminated vector, $\boldsymbol{Y}$. But, the contamination vector, $\boldsymbol{Z}$, depends on which entries are being contaminated. This situation can be expressed as,

$$\boldsymbol{Y}, \quad B, \quad \boldsymbol{Z}|B.$$

4. B **and** $\boldsymbol{Y}$ **are independent**. In this case the probability of contamination for the different entries of $\boldsymbol{X}$ is independent of the uncontaminated vector, $\boldsymbol{Y}$. But, the contamination vector, $\boldsymbol{Z}$, depends on which entries are being contaminated and their values. This situation can be expressed as,

$$\boldsymbol{Y}, \quad B, \quad \boldsymbol{Z}|B, \boldsymbol{Y}.$$

5. $\boldsymbol{Y}$ **and** $\boldsymbol{Z}$ **are independent**. In this case the contamination vector, $\boldsymbol{Z}$, is independent of the uncontaminated vector, $\boldsymbol{Y}$. But, the probability of contamination for the different entries of $\boldsymbol{X}$ depends on the uncontaminated vector, $\boldsymbol{Y}$. Also, the contamination vector, $\boldsymbol{Z}$, depends on which entries are being contaminated. This situation can be expressed as,

$$\boldsymbol{Y}, \quad B|\boldsymbol{Y}, \quad \boldsymbol{Z}|B.$$

Equivariance Considerations

The classical contamination model (3.2) is translation-scale equivariant and affine equivariant. On the other hand, the proposed contamination model (3.4) is translation-scale equivariant but not affine equivariant. To show that the contamination model is not affine equivariant, suppose that the random vector $\boldsymbol{X}$ follows the contamination model

$$\boldsymbol{X} = (I - B)\boldsymbol{Y} + B\boldsymbol{Z},$$

and A is an invertible $(p \times p)$ matrix. Let

$$\begin{aligned} \boldsymbol{V} &= A\boldsymbol{X} = A(I - B)\boldsymbol{Y} + AB\boldsymbol{Z} \\ &\neq (I - B)A\boldsymbol{Y} + BA\boldsymbol{Z}, \end{aligned}$$

unless $AB = BA$ (e.g. A is diagonal).

The lack of affine equivariance of the contamination model (3.4) is a bit surprising given that the uncontaminated vector, $\mathbf{Y}$, exhibits this property. However, this lack of affine equivariance is consistent with the fact that some affine transformations of the observable data, $\mathbf{X}$, may considerably worsen the overall quality of the resulting data. For instance, if $B_1, \ldots, B_p$ are independent with $P(B_j = 1) = \epsilon_j$, then $(1 - \epsilon_j)100\%$ of the measurements $X_{1j}, \ldots, X_{nj}$ $(j = 1, \ldots, p)$ in each coordinate (variable) are expected to be "good", but only $[(1 - \epsilon_1)(1 - \epsilon_2) \ldots (1 - \epsilon_p)]100\%$ of linear combinations $a_1 X_{i1} + a_2 X_{i2} + \ldots + a_p X_{ip}$ $(i = 1, \ldots, n)$, of all the coordinates, can be expected to be "good". We

X_1	X_2	X_3	$X_1 + X_2 + X_3$
0.891	-0.482	10.000	10.409
0.902	-0.769	-1.722	-1.589
1.246	-0.110	1.667	2.803
0.025	-1.198	-0.117	-1.290
-0.861	10.000	0.167	9.306
-0.215	-1.464	-1.265	-2.945
-1.157	-0.294	-0.527	-1.979
-0.149	-1.598	0.910	-0.838
10.000	-1.091	-0.740	8.169
-0.671	-0.007	0.884	0.206

Table 3.1: A Numerical Example of Increased Percentage of Contamination in a Linear Combination of Variables.

illustrate this phenomenon for a small data set with dimension $p = 3$ in Table 3.1. The table exhibits the values of each coordinate and the linear combination of the coordinates. The coordinates are separately and independently 10% contaminated. However, we can see that the linear combination of the coordinates are 30% contaminated. In particular notice that in the contamination model (3.4), where $\epsilon_1 = \epsilon_2 = \ldots = \epsilon_p = \epsilon$, $(1 - \epsilon)$ no longer represents the fraction of "good" data vectors (cases). Most of the observations, $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n, \boldsymbol{X}_i \in \mathbb{R}^p$ $(i = 1, \ldots, n)$ can be contaminated and this is often the case when the dimension, p, is large.

Assuming that $B_1, \ldots, B_p$ from the contamination model (3.4) are independent with constant $\epsilon_1 = \ldots = \epsilon_p = \epsilon$, we generate another simple model of some practical interest. This model represents situations when a certain proportion of outliers occurs independently on each variable. For example, this could be the case if several measurements are performed on the same individuals or items by several laboratories (*calibration model*).

In the next section, we present the different correlation structures that the components of the contamination indicator matrix B (3.5) may have in the contamination model (3.4). Dependence structures that can be covered are general dependence, with both positive and negative dependence.

3.4 Dependence Structures in the Contamination Indicator Matrix

To study the different dependency patterns of the components of B, we consider the bivariate Bernoulli distribution of $B = \text{Diag}(B_1, B_2)$ where $P(B_1 = 1) = P(B_2 = 1) = \epsilon$ as shown in Table 3.2. From the table we have that the expected value of B_j $(j = 1, 2)$ is ϵ with variance $\epsilon(1-\epsilon)$. The expected value of $B_1 B_2$ is δ with covariance $\delta - \epsilon^2$ and correlation $\frac{\delta-\epsilon^2}{\epsilon(1-\epsilon)}$, in which

		B_1		
		0	1	
B_2	0	$1-2\epsilon+\delta$	$\epsilon-\delta$	$1-\epsilon$
	1	$\epsilon-\delta$	δ	ϵ
		$1-\epsilon$	ϵ	1

Table 3.2: Bivariate Distribution of B_1 and B_2.

$$\frac{-\epsilon}{1-\epsilon} \leq \frac{\delta-\epsilon^2}{\epsilon(1-\epsilon)} \leq 1, \qquad \text{as} \quad 0 \leq \delta \leq \epsilon.$$

We consider three special cases of the dependence structures of the diagonal elements of B. Firstly, the *independent* case described in Table 3.3. The joint distribution of B_1 and B_2 in this case is given by $P(B_1 = 1, B_2 = 1) = P(B_1 = 1)P(B_2 = 1) = \epsilon^2$. Secondly, the *perfect correlation* case described in Table 3.4, which is the classical contamination model where $P(B_1 = B_2) = 1$. The joint distribution of B_1 and B_2 in this case is given by $P(B_1 = 1, B_2 = 1) = P(B_1 = 1) = \epsilon$. Lastly, the *perfect rejection* case presented in

		B_1		
		0	1	
B_2	0	$(1-\epsilon)(1-\epsilon)$	$\epsilon(1-\epsilon)$	$1-\epsilon$
	1	$\epsilon(1-\epsilon)$	ϵ^2	ϵ
		$1-\epsilon$	ϵ	1

Table 3.3: Bivariate Distribution of B_1 and B_2, Independent Case.

		B_1		
		0	1	
B_2	0	$1-\epsilon$	0	$1-\epsilon$
	1	0	ϵ	ϵ
		$1-\epsilon$	ϵ	1

Table 3.4: Bivariate Distribution of B_1 and B_2, Perfect Correlation.

		B_1		
		0	1	
B_2	0	$1-2\epsilon$	ϵ	$1-\epsilon$
	1	ϵ	0	ϵ
		$1-\epsilon$	ϵ	1

Table 3.5: Bivariate Distribution of B_1 and B_2, Perfect Rejection.

Table 3.5 is an example of negative dependence. The conditional distribution is given by $P(B_2 = 1|B_1 = 1) = 0$, which implies that the joint distribution of B_1 and B_2 in this case is given by $P(B_1 = 1, B_2 = 1) = 0$ and the conditional distribution $P(B_2 = 1|B_1 = 0) = \frac{\epsilon}{1-\epsilon}$. The three cases presented differ mainly in δ, the expected value of B_1B_2 (see Table 3.2). In the case of independence $\delta = \epsilon^2$, in the perfect correlation case $\delta = \epsilon$ and in the perfect rejection case $\delta = 0$.

Notice that we only consider bivariate cases in Tables 3.3 – 3.5 and situations where $P(B_1 = 1) = P(B_2 = 1) = \epsilon$. On the other hand, the more general multivariate distribution of a Bernoulli vector $(B_1, \ldots, B_p)$ with $P(B_1 = b_1, \ldots, B_p = b_p) = P(b_1, \ldots, b_p)$, $b_j = 1$ or 0 for $j = 1, \ldots, p$ has too many parameters, namely $(2^p - 1)$ parameters. Joe (1997) has an effective proposal for drastically reducing the number of parameters and still attaining a wide range of correlations structures. Joe (1997) proposed the exchangeable mixture model as a reasonable choice for some applications. The exchangeable mixture model is constructed as follows. Conditional on a random parameter θ the variables $B_1, \ldots, B_p$ are i.i.d. Bernoulli(θ). Therefore, the unconditional joint density for $B_1, \ldots, B_p$ is given by

$$P(b_1, \ldots, b_p) = P(B_1 = b_1, \ldots, B_p = b_p) = \int_0^1 \theta^k (1 - \theta)^{p-k} dG(\theta), \tag{3.8}$$

where $k = \sum_{j=1}^{p} b_j$, and $G(\theta)$ is the some specified distribution for θ with support on $[0, 1]$.

Joe (1997) indicates that the exchangeable mixture model only includes non-negative dependence structures. For some cases, the functional form but not the mixture representation of the exchangeable mixture model can be extended to include negative dependence. This feature will help to deal with the negative dependence of the perfect rejection structure in the contamination model (3.4).

Note, for the rest of the thesis we mainly consider situations in which $B_1, \ldots, B_p$ are i.i.d. Binomial$(1, \epsilon)$.

The family of distribution functions H generated by the contamination model (3.4) will be denoted by $\mathcal{H}$. Notice that $\mathcal{H}$ constitutes a contamination neighborhood for the central elliptical distribution H_0. To gain further insight into the family $\mathcal{H}$, in the next section we consider some dependence situations among the contamination vector, $\boldsymbol{Z}$, the contamination indicator matrix, B, and the uncontaminated vector, $\boldsymbol{Y}$.

3.5 Dependence Structures in the Contamination Model

In this section we consider different kinds of dependence structures, other than the dependence of the components of the contamination indicator matrix B (3.5). We illustrate the three dependence situations 2, 3 and 4 discussed in Section 3.3.

3.5.1 Independent-Contamination Model

We consider the independent-contamination model where the probability of contamination for the different entries of $\boldsymbol{X}$ does not depend on the uncontaminated vector, $\boldsymbol{Y}$. Also, the contamination vector, $\boldsymbol{Z}$, does not depend on which entries are being contaminated and their values. Let $\boldsymbol{Y}$, $\boldsymbol{Z}$ and B be independent. Then the model can be written as follows.

$$\boldsymbol{X} = (I - B)\boldsymbol{Y} + B\boldsymbol{Z}, \tag{3.9}$$

where for example $\boldsymbol{Z} = \boldsymbol{\mu} + N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{\mu} \in \mathbb{R}^p$. Suppose that F and G are the distribution functions of $\mathbf{Y}$ and $\mathbf{Z}$, respectively. Then, for the case $p = 2$ the distribution function H of $\mathbf{X}$ can be written as,

$$H(x_1, x_2) = \epsilon_{00}F(x_1, x_2) + \epsilon_{10}G_1(x_1)F_2(x_2) + \epsilon_{01}F_1(x_1)G_2(x_2) + \epsilon_{11}G(x_1, x_2),$$

where $\epsilon_{kj} = P(B_{i1} = k, B_{i2} = j)$ for $k, j = 0, 1$. More generally, we write

$$\begin{aligned}
H(x_1, \ldots, x_p) &= P(B_{i1} = 0, B_{i2} = 0, \ldots, B_{i(p-1)} = 0, B_{ip} = 0)F(x_1, x_2, \ldots, x_{p-1}, x_p) \\
&\quad + P(B_{i1}{=}1, B_{i2}{=}0, \ldots, B_{i(p-1)}{=}0, B_{ip}{=}0)G(x_1)F(x_2, \ldots, x_{p-1}, x_p) \\
&\quad \vdots \\
&\quad + P(B_{i1}{=}1, B_{i2}{=}1, \ldots, B_{i(p-1)}{=}1, B_{ip}{=}0)G(x_1, x_2, \ldots, x_{p-1})F(x_p) \\
&\quad + P(B_{i1} = 1, B_{i2} = 1, \ldots, B_{i(p-1)} = 1, B_{ip} = 1)G(x_1, x_2, \ldots, x_{p-1}, x_p).
\end{aligned}$$

For simplicity we have omitted the subscripts indicating the corresponding marginal distributions of F and G. For instance, $G(x_i)$ stands for $G_i(x_i)$, $G(x_i, x_j)$ stands for

$G_{ij}(x_i, x_j)$, etc. Note that this model will be mainly used throughout the rest of the thesis.

3.5.2 Contamination Vector as a Function of Contamination Indicator Matrix

We consider the dependence situation where the contamination vector, $\boldsymbol{Z}$, depends on which entries are being contaminated. The probability of contamination for the different entries of $\boldsymbol{X}$ and the contamination vector, $\boldsymbol{Z}$, are independent of the uncontaminated vector, $\boldsymbol{Y}$. Let $\boldsymbol{Y}$ and B be independent and given B, let $\boldsymbol{Y}$ and $\boldsymbol{Z}$ be independent. Then the model can be written as follows.

$$\boldsymbol{X} = (I - B)\boldsymbol{Y} + B\boldsymbol{Z}(B),$$

where for example $\boldsymbol{Y} \sim N(\boldsymbol{0}, I)$ and $\boldsymbol{Z} \sim N(\boldsymbol{\mu}(B), \boldsymbol{\Sigma})$. For simplicity, let $p = 2$ and B be a bivariate Bernoulli random variable with $P(B_1 = 1) = P(B_2 = 1) = \epsilon$, as shown in Table 3.2, and $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \tau \\ \tau & 1 \end{pmatrix}$ with $-1 \leq \tau \leq 1$. For different values of B_1 and B_2 we define $\boldsymbol{\mu}(B)$ as follows.

$$\begin{aligned} \boldsymbol{\mu}\left(\begin{pmatrix} 1 \\ 0 \end{pmatrix}\right) &= \boldsymbol{\mu}\left(\begin{pmatrix} 0 \\ 1 \end{pmatrix}\right) = \begin{pmatrix} k_1 \\ k_2 \end{pmatrix}, \\ \boldsymbol{\mu}\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}\right) &= \begin{pmatrix} \tilde{k}_1 \\ \tilde{k}_2 \end{pmatrix}. \end{aligned}$$

where $k_1, k_2, \tilde{k}_1, \tilde{k}_2$ are arbitrary constants.

Therefore, when $B_1 = 1, B_2 = 0$, $\boldsymbol{X}$ can be expressed as

$$\boldsymbol{X} = \begin{pmatrix} Z_1 \\ Y_2 \end{pmatrix} \sim N\left(\begin{pmatrix} k_1 \\ 0 \end{pmatrix}, I\right).$$

When $B_1 = 0, B_2 = 1$, $\boldsymbol{X}$ can be expressed as

$$\boldsymbol{X} = \begin{pmatrix} Y_1 \\ Z_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ k_2 \end{pmatrix}, I\right).$$

When $B_1 = B_2 = 1$, $\boldsymbol{X}$ can be expressed as

$$\boldsymbol{X} = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \tilde{k}_1 \\ \tilde{k}_2 \end{pmatrix}, \boldsymbol{\Sigma}\right).$$

3.5.3 Contamination Vector as a Function of Contamination Indicator matrix and Uncontaminated Vector

We consider the dependence situation where the contamination vector, $\boldsymbol{Z}$, depends on which entries are being contaminated and their values. The probability of contamination for the different entries of $\boldsymbol{X}$ is independent of the uncontaminated vector, $\boldsymbol{Y}$. Let $W \sim N(0,1)$, $\boldsymbol{Y} \sim N(\boldsymbol{0}, I)$ and B be independent. Then the model can be written as follows.

$$\boldsymbol{X} = (I - B)\boldsymbol{Y} + Bg(\boldsymbol{Y}, B, W).$$

For simplicity, let $p = 2$ and B be a bivariate Bernoulli random variable with $P(B_1 = 1) = P(B_2 = 1) = \epsilon$, as shown in Table 3.2 . For different values of B_1 and B_2 we define $\boldsymbol{Z} = g(\boldsymbol{Y}, B, W)$ as follows.

For $B_1 = 1$ and $B_2 = 0$, $\boldsymbol{Z}$ can be written as

$$\boldsymbol{Z} = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} \tau Y_2 + \sqrt{1-\tau^2}W \\ Y_2 \end{pmatrix} + \begin{pmatrix} k_1 \\ k_2 \end{pmatrix}.$$

This implies that the distribution of $\boldsymbol{Z}$ is bivariate normal with mean $\boldsymbol{K} = \begin{pmatrix} k_1 \\ k_2 \end{pmatrix}$ and covariance $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \tau \\ \tau & 1 \end{pmatrix}$, where k_1 and $k_2 \in \mathbb{R}$, and $-1 \leq \tau \leq 1$. Notice that τ is the correlation coefficient between the contaminated coordinates and the uncontaminated coordinates in the effective cases.

For $B_1 = 0$ and $B_2 = 1$, $\boldsymbol{Z}$ can be written as

$$\boldsymbol{Z} = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} Y_1 \\ \tau Y_1 + \sqrt{1-\tau^2}W \end{pmatrix} + \begin{pmatrix} k_1 \\ k_2 \end{pmatrix}.$$

This implies that the distribution of $\boldsymbol{Z}$ is bivariate normal with mean $\boldsymbol{K} = \begin{pmatrix} k_1 \\ k_2 \end{pmatrix}$ and covariance $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \tau \\ \tau & 1 \end{pmatrix}$.

For $B_1 = B_2 = 1$, $\boldsymbol{Z}$ can be written as

$$\boldsymbol{Z} = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} Y_1 \\ \tau Y_1 + \sqrt{1-\tau^2} W \end{pmatrix} + \begin{pmatrix} \tilde{k}_1 \\ \tilde{k}_2 \end{pmatrix}.$$

This implies that the distribution of $\boldsymbol{Z}$ is bivariate normal with mean $\widetilde{\boldsymbol{K}} = \begin{pmatrix} \tilde{k}_1 \\ \tilde{k}_2 \end{pmatrix}$ and covariance $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \tau \\ \tau & 1 \end{pmatrix}$, where $\tilde{k}_1$ and $\tilde{k}_2 \in \mathbb{R}$.

When $B_1 = 1, B_2 = 0$, $\boldsymbol{X}$ can be expressed as

$$\boldsymbol{X} = \begin{pmatrix} Z_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} \tau Y_2 + \sqrt{1-\tau^2} W \\ Y_2 \end{pmatrix} + \begin{pmatrix} k_1 \\ 0 \end{pmatrix},$$

which implies that $\boldsymbol{X}$ has a bivariate normal distribution with mean $\boldsymbol{K} = \begin{pmatrix} k_1 \\ 0 \end{pmatrix}$ and covariance $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \tau \\ \tau & 1 \end{pmatrix}$.

When $B_1 = 0, B_2 = 1$, $\boldsymbol{X}$ can be expressed as

$$\boldsymbol{X} = \begin{pmatrix} Y_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} Y_1 \\ \tau Y_1 + \sqrt{1-\tau^2} W \end{pmatrix} + \begin{pmatrix} 0 \\ k_2 \end{pmatrix},$$

which implies that $\boldsymbol{X}$ has a bivariate normal distribution with mean $\boldsymbol{K} = \begin{pmatrix} 0 \\ k_2 \end{pmatrix}$ and covariance $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \tau \\ \tau & 1 \end{pmatrix}$.

When $B_1 = B_2 = 1$, $\boldsymbol{X}$ can be expressed as

$$\boldsymbol{X} = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} Y_1 \\ \tau Y_1 + \sqrt{1-\tau^2}W \end{pmatrix} + \begin{pmatrix} \tilde{k}_1 \\ \tilde{k}_2 \end{pmatrix},$$

which implies that $\boldsymbol{X}$ has a bivariate normal distribution with mean $\widetilde{\boldsymbol{K}} = \begin{pmatrix} \tilde{k}_1 \\ \tilde{k}_2 \end{pmatrix}$ and covariance $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \tau \\ \tau & 1 \end{pmatrix}$.

A more concise way to express the bivariate distribution of $\boldsymbol{X}$ is in terms of densities. Suppose that $k = k_1 = k_2$ and $\tilde{k} = \tilde{k}_1 = \tilde{k}_2$. Let

$$f_{\boldsymbol{x}}(x_1, x_2) = \sum_{i,j=0,1} h(x_1, x_2 | i, j) p(i, j).$$

For different values of B_1 and B_2, we have the following density functions:

$$\begin{aligned} h(x_1, x_2|1,0) &= \phi\left(\frac{x_1 - k - \tau x_2}{\sqrt{1-\tau^2}}\right)\frac{1}{\sqrt{1-\tau^2}}\phi(x_2) = \phi_\tau(x_1 - k, x_2); \\ h(x_1, x_2|0,1) &= \phi\left(\frac{x_2 - k - \tau x_1}{\sqrt{1-\tau^2}}\right)\frac{1}{\sqrt{1-\tau^2}}\phi(x_1) = \phi_\tau(x_1, x_2 - k); \\ h(x_1, x_2|0,0) &= \phi\left(\frac{x_2 - \tau x_1}{\sqrt{1-\tau^2}}\right)\frac{1}{\sqrt{1-\tau^2}}\phi(x_1) = \phi_\tau(x_1, x_2); \\ h(x_1, x_2|1,1) &= \phi\left(\frac{x_2 - 2\tilde{k} - \tau x_1}{\sqrt{1-\tau^2}}\right)\frac{1}{\sqrt{1-\tau^2}}\phi(x_1 - \tilde{k}) = \phi_\tau(x_1 - \tilde{k}, x_2 - \tilde{k}), \end{aligned}$$

where $\phi(\cdot)$ is the standard normal density function, and $\phi_\tau(\cdot,\cdot)$ denotes the joint density function of the bivariate normal with means equal to zero, variances equal to one and correlation equal to τ.

3.6 Robust Estimation of Multivariate Location and Scatter: Problems and Motivation

Regarding the processing of large data sets, "traditional" affine equivariant high breakdown point robust multivariate location and scatter estimates have two main shortcomings:

- Computational complexity.
- Possible lack of robustness under the contamination model (3.4).

All known affine equivariant high breakdown point estimates are solutions to a highly non-convex optimization problem and as such pose a serious computational challenge. The main challenge is to find good initial estimates from which one searches for a nearest optimum in hopes that it produces a global optimum. The initial estimates are invariably obtained by using some form of repeated random sub-sampling of N_s cases of the original data set, with the number of samples N_s determined in order to achieve a high breakdown point with high probability, e.g., with probability .99 or .999 (see for example Rousseeuw and Leroy, 1987). It happens that achieving this latter condition results in computational algorithms that have exponential complexity of order 2^p in terms of the dimension p of the data set. This rules out the use of such estimates for many data mining applications where one has in excess of 200 – 300 variables. In addition, the robust covariance matrix based on projections has a computational complexity n^2 in the number of observations if implemented in a naïve manner. Empirical evidence indicates that a clever implementation can reduce this to approximately $n * log(n)$. Since many data mining applications involve hundreds of thousands if not millions of rows (cases), the current projection estimates are not feasible for large data sets.

In order to deal with such severe scalability limitations, Rousseeuw and Van Driesen (1999) proposed a "Fast MCD" (FMCD) method that is much more effective than naïve subsampling for minimizing the objective function of the MCD. The FMCD seems capable

Number of Variables	2	5	15	20	25	50	100
% Rows Spoiled (all variables)	10	23	54	64	72	92	99
% Rows Spoiled (pairwise variables)	10	13	16	17	18	19	21

Table 3.6: Percentage of Rows with at Least One Contaminated Entry when each Variable Independently 5% Contaminated ($\epsilon = .05$) in the Independent-Contamination Model.

of yielding "good" solutions without requiring huge values of N_s. But FMCD still requires substantial running times for large p, and it no longer retains a high breakdown point with high probability when n is large.

We consider the possible lack of robustness of the traditional robust estimates under the contamination model (3.4) to be a more serious problem than their computational complexity. Traditional robust estimates have a high breakdown point for all p under the classical contamination model (3.2); in fact, if conveniently tuned, these estimates may attain the maximum breakdown point (BP $= 1/2$) for affine equivariant estimates (Davies, 1987). However, we will show that the traditional affine equivariant robust estimates may not satisfy the robust properties of high breakdown point under the contamination model (3.4).

The possible lack of robustness of affine equivariant estimates can be hinted from the simple probability calculations shown in Table 3.6. For small fraction of contamination in each variable $\epsilon = .05$, consider the independent-contamination model (3.9) where $B_1, B_2, \ldots, B_p$ are i.i.d. Binomial$(1, \epsilon)$.

Given the dimension of the data set p, the second row of Table 3.6 exhibits the percentage of cases (rows) with at least one contaminated entry considering all variables (columns). The third row of this table exhibits the percentage of cases (rows) with at least one contaminated entry considering all pairs of variables one pair at the time. This probability has been numerically calculated with the details given in the chapter appendix.

We can see from Table 3.6 that when we consider all the variables the percentage of contaminated cases increases dramatically for large number of variables. Traditional affine equivariant robust estimates were not designed to cope with these situations, since they work globally with all the variables and require that the majority of the data (more than half) be uncontaminated. On the other hand, if we consider two variables at the time the percentage of contaminated cases is not high even for large number of variables.

In the following two chapters we offer a solution to the problems described above. We focus on considering the fewest possible dimensions of the vector we observe. We propose a coordinate-wise location estimate and a pairwise scatter estimate. Thus the proposed schemes do not involve all the coordinates of the vector, but rather use one dimension at the time for the location estimate and two dimensions at the time for the scatter estimate. In this way, we lessen the computational burden and attain the high breakdown point property under the contamination model (3.4). Using the smallest possible dimensions minimizes the fraction of contaminated cases which are used at each step in the computation of the estimates as suggested by the probability calculations in Table 3.6.

The proposed estimates are not affine equivariant, but this often is an unnecessary property in large data sets such as data mining applications. Also, this is not a disadvantage of the estimates since there are many practical situations where there exists a natural representation for the data (e.g. the form in which they have been measured) except perhaps for a shift and/or some unit changes and in which affine equivariance may not be necessarily a desirable property.

3.7 Chapter Appendix

We wish to calculate the maximum proportion of contaminated rows that may occur under the independent-contamination model (3.9) when we consider all the $\begin{pmatrix} p \\ 2 \end{pmatrix}$ pairs of columns in a p-dimensional data set. Let B_{kj} $(k = 1, \ldots, n, j = 1, \ldots, p)$ be independent Bernoulli random variables with $P\left(B_{kj} = 1\right) = \epsilon$. Let

$$S_{ij} = \sum_{k=1}^{n} \max\{B_{ki}, B_{kj}\}, \qquad i, j = 1, \ldots, p.$$

Then we are interested in studying

$$p_n = \mathbb{E}\left\{\frac{1}{n} \max_{i<j} S_{ij}\right\}. \tag{3.10}$$

To investigate the behavior of p_n, we generated 1000 random matrices of the form

$$A_k = \begin{pmatrix} A_{k11} & A_{k12} & \cdots & A_{k1p} \\ A_{k21} & A_{k22} & \cdots & A_{k2p} \\ \vdots & \vdots & & \vdots \\ A_{kn1} & A_{kn2} & \cdots & A_{knp} \end{pmatrix}, \qquad k = 1, \ldots, 1000,$$

where the A_{klj}'s are independent Bernoulli random variables with $P\left(A_{klj} = 1\right) = \epsilon$, for all $k = 1, \ldots, 1000$, $l = 1, \ldots, n$ and $j = 1, \ldots, p$. The value of p_n (3.10) is then estimated by

$$\frac{1}{1000} \sum_{k=1}^{1000} \left(\max_{i<j} \frac{1}{n} \sum_{l=1}^{n} \max\left\{A_{kli}, A_{klj}\right\} \right).$$

Our results for the case $\epsilon = 0.05$, $n = 1000$ and $p = 2, 5, 15, 20, 25, 50, 100$ are presented in Table 3.6. Similar results were obtained for other values of n.

Chapter 4

Robust Estimation of Multivariate Scatter

4.1 Classical Scatter Estimate

Suppose that $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, where $\boldsymbol{X}_i \in \mathbb{R}^p$, $i = 1, \ldots, n$, are independent and identically distributed according to a multivariate distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The classical and best known estimate of the covariance matrix $\boldsymbol{\Sigma}$ is the method of moment estimate (MME) which is defined as follows.

$$\widehat{\boldsymbol{\Sigma}}_{MME} = \frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}\right) \left(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}\right)', \tag{4.1}$$

where $\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i$. Note that estimation of the correlation matrix R can always be derived from the relation $R = D\boldsymbol{\Sigma}D$, where $D = \text{Diag}\left(1/\sqrt{\sigma}_{11}, \ldots, 1/\sqrt{\sigma}_{pp}\right)$ and $\sigma_{11}, \ldots, \sigma_{pp}$ are the diagonal elements of the covariance matrix $\boldsymbol{\Sigma}$.

The breakdown point is an important feature of the reliability of an estimate, as it indicates, roughly speaking, the smallest proportions of arbitrary values (outliers) that bring the estimate out of the boundaries of the parameter space. The definition of the breakdown point for the covariance matrix estimates is given in Section 2.2 of Chapter 2. Unfortunately, the breakdown point of the method of moment estimate (4.1) is $1/n$, which indicates very poor resistance to outliers.

In the last three decades, many attempts to overcome the poor resistance properties of the classical sample dispersion matrix (i.e. covariance and correlation matrices) have been made. The robust proposals can be classified in two main categories: robust pairwise estimation and robust global estimation of the dispersion matrix. The first one has the advantage of being able to deal with missing values in the data set, but is not affine equivariant and often does not provide a positive definite matrix directly. The second category usually ensures affine equivariance and positive definiteness, but is less appropriate to deal with missing data. In addition, the main drawback remains the computational feasibility of such methods for high dimensional data sets. At this point, it seems appropriate to revisit old proposals for estimating scatter based on using only two variables at a time.

The discussion above motivates the construction of robust dispersion matrices by using pairwise robust correlation (or covariance) coefficients as basic building blocks. Several such methods have been around for many years, but they have been mostly ignored because: (a) of the lack of affine equivariance, and (b) that the resulting dispersion matrix built up from the pairwise estimates lacks positive definiteness. We are motivated to re-examine the pairwise approach because: (1) the lack of affine equivariance is not necessarily important for large data sets such as in data mining applications, and (2) there exist good methods for obtaining positive definiteness, such as Maronna and Zamar (2002), and Rousseeuw and Molenberghs (1993) who proposed three methods; respectively, the shrinking method, the eigenvalue method and the scaling method. When the covariance itself is the quantity of interest, one should transform it to a positive definite matrix using one of these methods; while if some particular entries in the matrix are the values of interest, then the estimated values should provide a good estimation of the real values. The simplest pairwise methods are based on pairwise robust correlation or covariance estimates such as: (i) classical rank based methods, such as the Spearman's ρ and Kendall's τ (see for example Abdullah, 1990); (ii) classical correlations applied after coordinate-wise outlier insensitive transformations such as the quadrant correlation and 1-D "Huberized" data (see Huber, 1981, page 204); and (iii) bivariate outlier resistant methods such the method proposed by Gnanadesikan and Kettenring (1972) and studied by Devlin, Gnanadesikan and Kettenring (1981). The pairwise approach is appealing in that one can achieve high breakdown point on a pairwise basis that results in a high breakdown point for the overall covariance or correlation matrix, and at the same time reduces the computational complexity in the data dimension p from exponential to quadratic (from 2^p to p^2). This greatly increases the range of large data sets problems to which robust covariance and correlation estimates can be applied, e.g., 200 – 300 variables becomes quite feasible.

In this chapter, we will concentrate on estimates in the class (ii) of pairwise robust estimates originally introduced by Huber (1981).

4.2 Simple Class of Pairwise Robust Scatter Estimate

The method we are proposing to estimate the scatter matrix draws on work done by Huber (1981). The work remains largely uninvestigated due to the fact that it is not an affine equivariant estimate. We focus on the estimation of correlation matrices, since estimation of covariance matrices can be derived in the same way. Huber defines robust correlation coefficient estimates as follows.

Suppose $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ is a multivariate sample where $\boldsymbol{X}_i \in \mathbb{R}^p$, $i = 1, \ldots, n$. Let s_j $(j = 1, \ldots, p)$ be some robust scale estimates and let t_j $(j = 1, \ldots, p)$ be location M-estimates defined by the following equation:

$$\sum_{i=1}^{n} \psi\left(\frac{X_{ij} - t_j}{s_j}\right) = 0,$$

where $\psi(x)$ is an appropriate score function. The following two cases are of primary interest:

- **Huber Function**

$$\psi_c(x) = \min\left\{\max\left\{-c, x\right\}, c\right\},$$

where $c \in \mathbb{R}_+$ is a user-chosen constant.

- **Sign Function**

$$\psi(x) = \mathrm{SGN}(x),$$

where SGN(x) has the values +1 for $x > 0$, -1 for $x < 0$, and 0 for $x = 0$.

The robust correlation coefficient estimate $\hat{r}_{jk}$ is now defined as the Pearson correlation coefficient computed on the transformed data $Y_{ij} = \psi\left((X_{ij} - t_j)/s_j\right), j = 1, \ldots, k$; $j = 1, \ldots, p$,

$$\hat{r}_{jk} = \frac{\frac{1}{n}\sum_{i=1}^{n} Y_{ij}Y_{ik}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} Y_{ij}^2 \frac{1}{n}\sum_{i=1}^{n} Y_{ik}^2}}.$$

Notice that $\bar{Y}_j = \bar{Y}_k = 0$ by definition of t_j and t_k.

To save computing time and still gain robustness, we can use another robust location estimate $t_j = \mathrm{median}\{X_{ij}\}$ and therefore the robust correlation coefficient estimate has the following form:

$$\hat{r}_{jk} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(Y_{ij} - \bar{Y}_j\right)\left(Y_{ik} - \bar{Y}_k\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_{ij} - \bar{Y}_j\right)^2 \frac{1}{n}\sum_{i=1}^{n}\left(Y_{ik} - \bar{Y}_k\right)^2}}. \tag{4.2}$$

When ψ is the Huber function, we call this the *Huberized correlation* coefficient, and when ψ is the sign function the estimate is the so-called *quadrant correlation* (QC) coefficient, which is the Huberized correlation coefficient with tuning constant $c = 0$, since $\lim\limits_{c \to 0} \psi_c(x) = \mathrm{SGN}(x)$.

In the case of n observations of a p-dimensional random vector, we use the estimate $\hat{r}_{jk}$ to estimate every correlation between $\boldsymbol{X}_j$ and $\boldsymbol{X}_k$ $(j, k = 1, \ldots, p)$ to get the (j, k) entry of the correlation matrix R. The pairwise Huberized correlation matrix estimate can, therefore, be defined as $\widehat{R} = (\hat{r}_{jk})_{j,k=1,\ldots,p}$.

4.3 Performance of Pairwise Huberized Scatter Estimate

In this section we report the results of a Monte Carlo study on the performance of the pairwise Huberized covariance matrix estimates in the contamination model (3.4). We considered sample sizes $n = 10p$ where p is the number of variables taking values $10, 30$ and 50. The data followed the independent-contamination model:

$$\boldsymbol{X} = (I - B)\boldsymbol{Y} + B\boldsymbol{Z},$$

where $\boldsymbol{Y}$, B and $\boldsymbol{Z}$ are independent and the diagonal elements of B, $B_1, \ldots, B_p$ are i.i.d. $\text{Binomial}(1, \epsilon)$.

In view of the lack of equivariance of the pairwise Huberized covariance matrix estimates their behavior may depend on the covariance structure; hence we generated correlated data as follows: Generate $\boldsymbol{Y}_i$ as p-variate normals $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ for $i = 1, \ldots, n$, and $\boldsymbol{Z}_i$ as $N_p(\boldsymbol{\mu}_0, \delta I)$ for some $\boldsymbol{\mu}_0 = (\mu_0, \ldots, \mu_0)_p$; where $\delta = 0.1$. Generate $B_1, \ldots, B_p$ as i.i.d. $\text{Binomial}(1, \epsilon)$.

The covariance matrix $\boldsymbol{\Sigma}$ generated in our simulation study was obtained as follows:

1. Using the condition number, which is defined as the square root of the largest eigenvalues divided by the smallest eigenvalues $\text{CN} = \sqrt{\frac{\lambda_1}{\lambda_p}}$, we generated multivariate normal data with mean vector $\mathbf{0}$ and covariance matrix, $\boldsymbol{\Sigma}_\lambda = \text{Diag}(\lambda_1, \ldots, \lambda_p)$ as follows:

   - Set the largest eigenvalue $\lambda_1 = 1$ and the smallest eigenvalue $\lambda_p = \frac{1}{CN^2}$ with equally spaced eigenvalues in between.
   - Generate $\boldsymbol{X}_i$ as p-variate normals $N_p(\mathbf{0}, \boldsymbol{\Sigma}_\lambda)$ for $i = 1, \ldots, N = 100,000$.

2. Using random orthogonal matrix, the correlation structure of the data set X, where $X = \boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$, is decided at random by rotating the data set using a method proposed by Fang and Zhang (1990), which we describe briefly as follows:

   - Consider the $n \times p$ matrix X;
   - Generate a random matrix Y as $N_{p\times p}(0, I_{p\times p})$;
   - Let $U = Y(Y'Y)^{-1/2}$. Thus U has a uniform distribution over the Stiefel manifold, the set consisting of all orthogonal random matrices;

- Rotate the matrix X, $X_r = XU$ which has a different correlation matrix.

3. The covariance matrix $\mathbf{\Sigma} = \mathrm{Cov}(X_r)$.

The pairwise Huberized covariance matrix estimate $\widehat{\mathbf{\Sigma}}$ was obtained from the Huberized correlation coefficient estimates using the median as the location estimate and the median absolute deviation (MAD) as the scale estimate. We compared the estimated covariance $\widehat{\mathbf{\Sigma}}$ with the true covariance $\mathbf{\Sigma}$ using the following two metrics:

1. Euclidean distance or the straight line distance between the coordinates of $\mathbf{\Sigma}$ and $\widehat{\mathbf{\Sigma}}$, which is given by

$$d(\mathbf{\Sigma}, \widehat{\mathbf{\Sigma}}) = \sqrt{\frac{1}{p(p+1)} \sum_{i=1}^{p} \sum_{j=i}^{p} |\hat{\sigma}_{ij} - \sigma_{ij}|^2},$$

where $\hat{\sigma}_{ij}, \sigma_{ij}$ $(i, j = 1, \ldots, p)$ are elements of $\widehat{\mathbf{\Sigma}}$ and $\mathbf{\Sigma}$, respectively.

2. Another choice for the metric is to use the determinant of the covariance matrix $\mathbf{\Sigma}$, which is the generalized variance. The generalized variance converts the information on all the variances and covariances into a single number. Generalized variance also has interpretations in the p-space scatterplot representation of the data. The most intuitive interpretation concerns the spread of the scatter about the mean vector. The metric of the determinants of $\widehat{\mathbf{\Sigma}}$ and $\mathbf{\Sigma}$ called *eigenvalue distance* is defined as follows.

$$\begin{aligned} d(\mathbf{\Sigma}, \widehat{\mathbf{\Sigma}}) &= \left| \log\left(\frac{\det \widehat{\mathbf{\Sigma}}}{\det \mathbf{\Sigma}} \right) \right|^{1/p} \\ &= \left| \log\left(\frac{\Pi_{i=1}^{p} \hat{\lambda}_i^{1/p}}{\Pi_{i=1}^{p} \lambda_i^{1/p}} \right) \right| \\ &= \left| \frac{1}{p} \sum_{i=1}^{p} \log \frac{\hat{\lambda}_i}{\lambda_i} \right| \\ &= \left| \frac{1}{p} \sum_{i=1}^{p} \left[\log(\hat{\lambda}_i) - \log(\lambda_i) \right] \right|, \end{aligned}$$

where $\hat{\lambda}_i$ and λ_i $(i = 1, \ldots, p)$ are eigenvalues of $\widehat{\mathbf{\Sigma}}$ and $\mathbf{\Sigma}$, respectively.

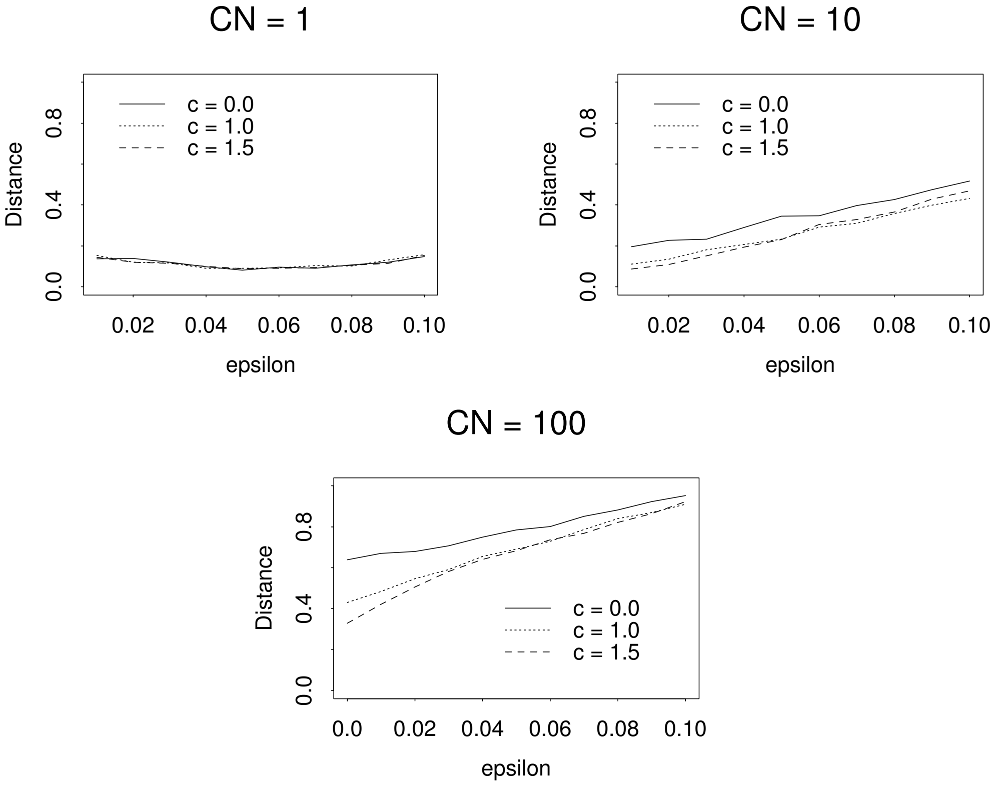


Figure 4.1: Performance of Pairwise Huberized Covariance Estimates using Eigenvalue Metric, for Data Sets with Size of Contamination $\mu_0 = 100$ and $p = 10$.

We ran 1,000 Monte Carlo simulations from the above distributions with CN $= 1, 10, 20, 50$ and 100 and size of contamination $\mu_0 = 5, 10$ and 100. The samples were the same for all estimates and for each combinations n, p, ϵ and c, where c is the tuning constant of the Huber score function. We considered $\epsilon = 0, .01, .02, \ldots, .10$ with $c = 0, 1, 1.25$ and 1.5. For all values of c, the results for the pairwise Huberized covariance estimates indicated (tables not shown here) that the eigenvalue distances increased with increasing values of the condition number; however, the Euclidean distances were not affected. The size of the contamination did not affect the results either. Figures 4.1 – 4.3 illustrate the varying degrees of change in the eigenvalue distances for various values of the condition number, CN = 1, 10 and 100 when the size of the contamination is large, $\mu_0 = 100$.

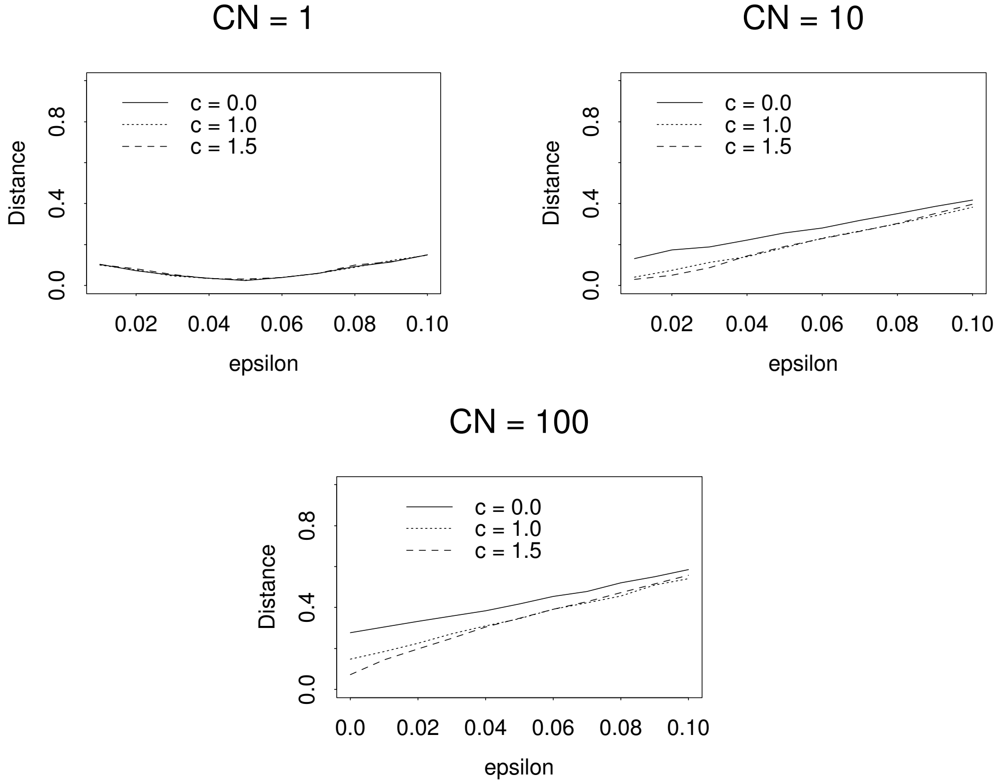


Figure 4.2: Performance of Pairwise Huberized Covariance Estimates using Eigenvalue Metric, for Data Sets with Size of Contamination $\mu_0 = 100$ and $p = 30$.

For $p = 10, 30$ and 50 respectively, each figure plots the eigenvalue distance against the fraction of contamination ϵ, for $c = 0, 1$ and 1.5.

From the plots, we can see that for CN $= 1, 10$ and 100, the eigenvalue distance between the estimated covariance and the true covariance decreases as the dimension of the data increases. In addition, for all data dimensions the effect of the different values of the tuning constant c is minimal as the fraction of contamination ϵ increases. Moreover, when CN $= 1$ the performance of the pairwise Huberized covariance estimates were not affected with the value of the tuning constant c.

The performance of the Fast MCD (FMCD) covariance estimates were also monitored.

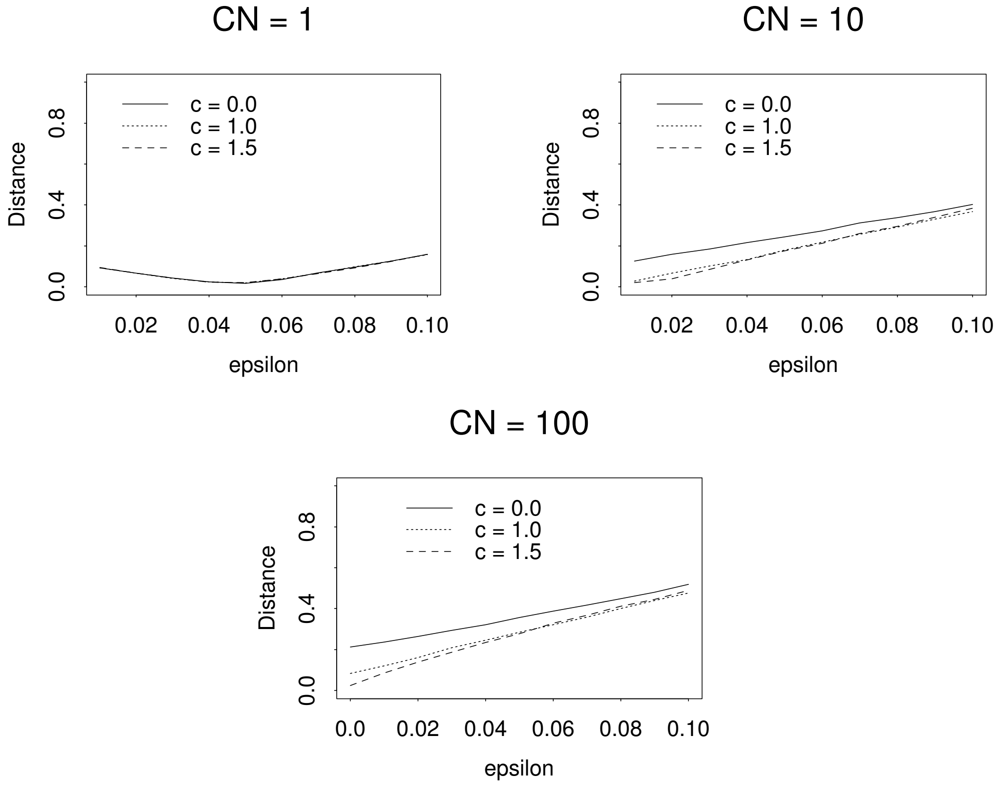


Figure 4.3: Performance of Pairwise Huberized Covariance Estimates using Eigenvalue Metric, for Data Sets with Size of Contamination $\mu_0 = 100$ and $p = 50$.

We used the same sampling situations with CN $= 1$ since the Fast MCD covariance estimates are affine equivariant. The performance of the Fast MCD covariance estimates for the size of contamination, $\mu_0 = 5, 10$ and 100 is shown in Figure 4.4. The figure displays plots of the eigenvalue distance versus the fraction of contamination ϵ, for $p = 10, 20$ and 30. From the plots, we can see that in general, the Fast MCD estimates perform poorly for large contamination sizes and that their performance worsen considerably as the dimension p increases.

We also considered comparing the performance of the pairwise Huberized covariance estimates with the performance of the Fast MCD covariance estimates using eigenvalue

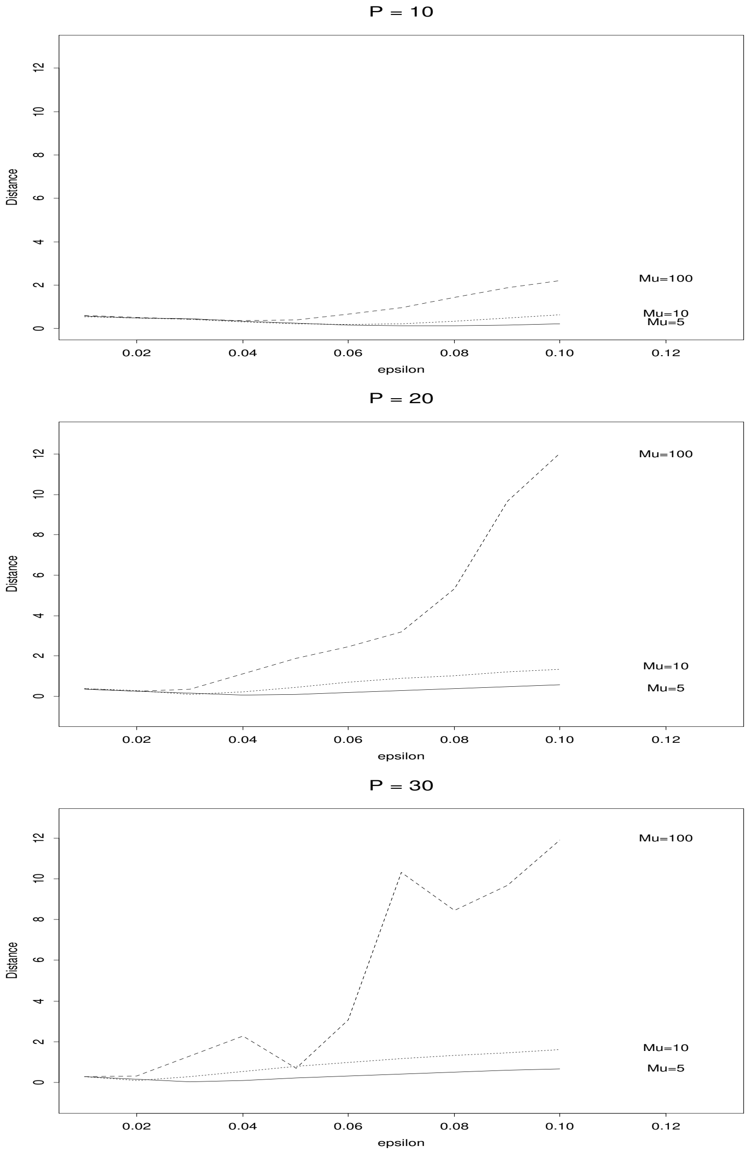


Figure 4.4: Performance of Fast MCD Covariance Estimates using Eigenvalue Metric, for Data Sets with Sizes of Contamination $\mu_0 = 5, 10, 100$ and $p = 10, 20, 30$.

and the Euclidean distances. We generated 1,000 data sets from the above distributions with CN = 1, $\mu_0 = 5, 10$ and 100, and $\epsilon = .05, .10, .15, .20, .25$ and .30. We used $c = 0$ for the Huber score function, which is the quadrant correlation (QC) coefficient.

The performance results of the pairwise Huberized covariance estimates and the Fast MCD covariance estimates are displayed in Tables 4.1, 4.2 and 4.3 for $p = 10$, 20 and 30, respectively.

	Small				Medium				Large			
	Pairwise		FMCD		Pairwise		FMCD		Pairwise		FMCD	
Eps	d1	d2	d1	d2	d1	d2	d1	d2	d1	d2	d1	d2
0.05	0.073	0.022	0.192	0.023	0.072	0.022	0.207	0.030	0.063	0.021	0.276	2.036
0.10	0.200	0.034	0.268	0.103	0.192	0.031	0.641	0.368	0.194	0.035	2.274	35.684
0.15	0.379	0.056	0.639	0.168	0.375	0.054	1.311	0.633	0.366	0.057	3.787	73.045
0.20	0.588	0.092	0.908	0.254	0.599	0.099	1.775	1.147	0.581	0.099	4.887	116.669
0.25	0.856	0.159	1.128	0.337	0.841	0.151	2.118	1.204	0.848	0.156	5.681	137.926
0.30	1.194	0.295	1.287	0.386	1.162	0.252	2.323	1.523	1.166	0.221	6.075	155.076

d1: Eigenvalue Distance ; d2: Euclidean Distance

Small: $\mu_0 = 5$; Medium: $\mu_0 = 10$; Large: $\mu_0 = 100$

Table 4.1: Performance of Pairwise Huberized ($c = 0$) and Fast MCD Covariance Estimates for Data Sets with $p = 10$.

	Small				Medium				Large			
	Pairwise		FMCD		Pairwise		FMCD		Pairwise		FMCD	
Eps	d1	d2	d1	d2	d1	d2	d1	d2	d1	d2	d1	d2
0.05	0.070	0.008	0.132	0.021	0.065	0.009	0.463	0.081	0.057	0.008	1.812	10.475
0.10	0.215	0.016	0.610	0.067	0.211	0.015	1.359	0.286	0.205	0.015	35.573	∞
0.15	0.392	0.026	0.947	0.108	0.389	0.028	1.914	0.487	0.379	0.027	166.804	∞
0.20	0.598	0.044	1.179	0.150	0.606	0.046	2.264	0.590	0.594	0.044	∞	∞
0.25	0.856	0.072	1.355	0.177	0.853	0.073	2.511	0.791	0.857	0.083	∞	∞
0.30	1.170	0.123	1.486	0.219	1.177	0.121	2.695	0.838	1.184	0.093	∞	∞

d1: Eigenvalue Distance ; d2: Euclidean Distance

Small: $\mu_0 = 5$; Medium: $\mu_0 = 10$; Large: $\mu_0 = 100$

Table 4.2: Performance of Pairwise Huberized ($c = 0$) and Fast MCD Covariance Estimates for Data Sets with $p = 20$.

	Small				Medium				Large			
	Pairwise		FMCD		Pairwise		FMCD		Pairwise		FMCD	
Eps	d1	d2	d1	d2	d1	d2	d1	d2	d1	d2	d1	d2
0.05	0.063	0.005	0.267	0.018	0.065	0.006	0.834	0.106	0.068	0.005	10.926	∞
0.10	0.212	0.010	0.730	0.049	0.211	0.010	1.347	0.186	0.211	0.018	88.570	∞
0.15	0.384	0.018	1.042	0.077	0.388	0.020	1.640	0.273	0.390	0.019	251.030	∞
0.20	0.589	0.032	1.263	0.105	0.595	0.031	2.150	0.501	0.600	0.029	∞	∞
0.25	0.849	0.047	1.425	0.127	0.855	0.052	2.935	1.259	0.857	0.049	∞	∞
0.30	1.166	0.082	1.544	0.148	1.178	0.085	4.067	4.330	1.176	0.082	∞	∞

d1: Eigenvalue Distance ; d2: Euclidean Distance

Small: $\mu_0 = 5$; Medium: $\mu_0 = 10$; Large: $\mu_0 = 100$

Table 4.3: Performance of Pairwise Huberized ($c = 0$) and Fast MCD Covariance Estimates for Data Sets with $p = 30$.

We can see that, in general, for both metrics the Fast MCD covariance estimates perform poorly for large contamination sizes and that their performance worsen considerably as the dimension p increases. However, the performance of the pairwise Huberized covariance estimates were not affected as the dimension p and the size of contamination μ_0 increase. The performance of the two estimates becomes dramatically different for large p as the fraction of contamination ϵ increases.

4.4 Asymptotic Properties of Huberized Correlation Coefficients

The objective of this section is to show that under certain regularity conditions the Huberized correlation coefficient estimates are consistent and asymptotically normal.

4.4.1 Consistency of Huberized Correlation Coefficients

The next theorem shows that under certain regularity conditions, if the location and the scale are consistent estimates, then the Huberized correlation coefficient estimates are also consistent.

THEOREM 4.1 – **Consistency of Huberized Correlation Coefficients** –
Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ *be a random sample from a bivariate distribution. Let* $\hat{\mu}_X$ *and* $\hat{\mu}_Y$ *be location estimates, and* $\hat{\sigma}_X$ *and* $\hat{\sigma}_Y$ *be scale estimates. Let* $\psi : \mathbb{R} \to \mathbb{R}$ *satisfy the following:*

P.1 $\psi(-u) = -\psi(u)$, $u \geq 0$;

P.2 $\psi(u)$ *is non-decreasing and* $\lim_{u \to \infty} \psi(u) > 0$;

P.3 ψ *is continuously differentiable;*

P.4 ψ, ψ' *and* $\psi'(u)u$ *are bounded.*

Then if,

$$\begin{aligned} \hat{\mu}_X &\longrightarrow \mu_X && \textit{(a.s.)} \\ \hat{\mu}_Y &\longrightarrow \mu_Y && \textit{(a.s.)} \\ \hat{\sigma}_X &\longrightarrow \sigma_X && \textit{(a.s.)} \\ \hat{\sigma}_Y &\longrightarrow \sigma_Y && \textit{(a.s.)} \end{aligned}$$

as $n \to \infty$, *then* $\hat{r} \to r$ *almost surely as* $n \to \infty$ *where*

$$\hat{r} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(\psi\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right) - \frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\right)\left(\psi\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right) - \frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\psi\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right) - \frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\right)^2 \frac{1}{n}\sum_{i=1}^{n}\left(\psi\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right) - \frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\right)^2}},$$

and

$$\begin{aligned} r &= \frac{\mathbb{E}\left\{\left(\psi\left(\frac{X-\mu_X}{\sigma_X}\right) - \mathbb{E}\psi\left(\frac{X-\mu_X}{\sigma_X}\right)\right)\left(\psi\left(\frac{Y-\mu_Y}{\sigma_Y}\right) - \mathbb{E}\psi\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right)\right\}}{\sqrt{\mathbb{E}\left\{\psi\left(\frac{X-\mu_X}{\sigma_X}\right) - \mathbb{E}\psi\left(\frac{X-\mu_X}{\sigma_X}\right)\right\}^2 \mathbb{E}\left\{\psi\left(\frac{Y-\mu_Y}{\sigma_Y}\right) - \mathbb{E}\psi\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right\}^2}} \\ &\equiv \frac{Cov\left(\psi\left(\frac{X-\mu_X}{\sigma_X}\right), \psi\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right)}{\sqrt{Var\left(\psi\left(\frac{X-\mu_X}{\sigma_X}\right)\right) Var\left(\psi\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right)}}. \end{aligned}$$

The relation between r and $\rho = \mathrm{Corr}(X, Y)$ will be studied in Section 4.5.

4.4.2 Asymptotic Normality of Huberized Correlation Coefficients

Having shown the consistency of the Huberized correlation coefficient estimates, we turn our attention to their asymptotic distribution. We will focus on the MM-location estimates, an important special case of robust location estimates, which is defined below.

We need some definitions and assumptions that will be used in the statement of Theorem 4.2 and its proof.

Assume that $\psi : \mathbb{R} \to \mathbb{R}$ satisfies P.1 – P.4 from Theorem 4.1. Moreover, we will assume that the real function $\chi : \mathbb{R} \to \mathbb{R}_+$ satisfies the following:

A.1 $\chi(0) = 0$, $\chi(-u) = \chi(u)$, $u \geq 0$ and $\sup_{u\in\mathbb{R}} \chi(u) = 1$;

A.2 $\chi(u)$ is non-decreasing in $u \geq 0$;

A.3 χ is continuously differentiable;

A.4 χ, χ', $\chi'(u)u$ are bounded.

We now define the S-scale family of estimates (Rousseeuw and Yohai, 1984).

DEFINITION 4.1 – **S-scale estimates** – *Let $X_1, \dots, X_n$ be a random sample and $0 < b \leq 1/2$. The S-scale $\hat{\sigma}_n$ is defined as*

$$\hat{\sigma}_n = \inf_{t \in \mathbb{R}} s_n(t),$$

where $s_n(t)$ is given by

$$\frac{1}{n} \sum_{i=1}^{n} \chi\left(\frac{X_i - t}{s_n(t)}\right) \quad = \quad b. \tag{4.3}$$

Naturally associated with this family are the S-location estimates.

DEFINITION 4.2 – **S-location estimates** – *Let $X_1, \dots, X_n$ be a random sample, and for each $t \in \mathbb{R}$ let $s_n(t)$ be as in (4.3). The S-location estimate $\tilde{\mu}_n$ is*

$$\tilde{\mu}_n = \arg\inf_{t \in \mathbb{R}} s_n(t).$$

In analogy with Yohai (1987) we will refer to the M-location estimates calculated with an S-scale as MM-estimates.

DEFINITION 4.3 – **MM-location estimates** – *Let $X_1, \dots, X_n$ be a random sample and $\hat{\sigma}_n$ be an S-scale estimate. The solution $\hat{\mu}_n$ of*

$$\frac{1}{n} \sum_{i=1}^{n} \psi\left(\frac{X_i - \hat{\mu}_n}{\hat{\sigma}_n}\right) = 0.$$

is called the MM-location estimate of $X_1, \dots, X_n$.

The following theorem states the asymptotic normality of the Huberized correlation coefficient estimates under certain regularity conditions.

THEOREM 4.2 - **Asymptotic Normality of Huberized Correlation Coefficients** - *Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be a random sample of independent and identically distributed random vectors with elliptically symmetric distribution. We consider the Huberized correlation coefficient estimate defined as follows.*

$$\hat{r} \quad = \quad \frac{\frac{1}{n}\sum_{i=1}^{n} \psi\left(\frac{X_i - \hat{\mu}_X}{\hat{\sigma}_X}\right) \psi\left(\frac{Y_i - \hat{\mu}_Y}{\hat{\sigma}_Y}\right)}{\sqrt{\left[\frac{1}{n}\sum_{i=1}^{n} \psi^2\left(\frac{X_i - \hat{\mu}_X}{\hat{\sigma}_X}\right)\right]\left[\frac{1}{n}\sum_{i=1}^{n} \psi^2\left(\frac{Y_i - \hat{\mu}_Y}{\hat{\sigma}_Y}\right)\right]}}, \tag{4.4}$$

where $\hat{\mu}_X$ and $\hat{\mu}_Y$ are MM-location estimates, $\hat{\sigma}_X$ and $\hat{\sigma}_Y$ are S-scale estimates. Then,

$$\sqrt{n}\,(\hat{r} - r) \longrightarrow_d N(0, AV),$$

as $n \to \infty$, where

$$r = \frac{I\!\!E\left\{\psi\left(\frac{X-\mu_X}{\sigma_X}\right)\psi\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right\}}{\sqrt{I\!\!E\left\{\psi^2\left(\frac{X-\mu_X}{\sigma_X}\right)\right\}I\!\!E\left\{\psi^2\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right\}}}.$$

The variance of the limiting distribution of $\sqrt{n}(\hat{r} - r)$ can be expressed as

$$AV = \frac{\sigma_{11}}{vw} + (1/4)\frac{\sigma_{22}}{v^2}\frac{u^2}{vw} + (1/4)\frac{\sigma_{33}}{w^2}\frac{u^2}{vw} - \sigma_{12}(\frac{u}{v^2w}) - \sigma_{13}(\frac{u}{w^2v}) + (1/2)\sigma_{23}\frac{u^2}{v^2w^2},$$

where

$$\begin{aligned}
\sigma_{11} &= Var\left\{\psi\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\psi\left(\frac{X-\mu_X}{\sigma_X}\right)\right\};\\
\sigma_{12} &= Cov\left\{\psi\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\psi\left(\frac{X-\mu_X}{\sigma_X}\right), \psi^2\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right\};\\
\sigma_{13} &= Cov\left\{\psi\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\psi\left(\frac{X-\mu_X}{\sigma_X}\right), \psi^2\left(\frac{X-\mu_X}{\sigma_X}\right)\right\};\\
\sigma_{22} &= Var\left\{\psi^2\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right\};\\
\sigma_{23} &= Cov\left\{\psi^2\left(\frac{Y-\mu_Y}{\sigma_Y}\right), \psi^2\left(\frac{X-\mu_X}{\sigma_X}\right)\right\};\\
\sigma_{33} &= Var\left\{\psi^2\left(\frac{X-\mu_X}{\sigma_X}\right)\right\},
\end{aligned}$$

and

$$\begin{aligned}
u &= I\!\!E\left\{\psi\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\psi\left(\frac{X-\mu_X}{\sigma_X}\right)\right\};\\
v &= I\!\!E\left\{\psi^2\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right\};\\
w &= I\!\!E\left\{\psi^2\left(\frac{X-\mu_X}{\sigma_X}\right)\right\}.
\end{aligned}$$

To simplify the notation, define

$$c_{ij} = \mathbb{E}\left\{\psi^i\left(\frac{X-\mu_X}{\sigma_X}\right)\psi^j\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right\},$$

where μ_X, μ_Y are locations and σ_X, σ_Y are scales of X and Y, respectively. Then the asymptotic variance can be written as follows.

$$\begin{aligned} AV &= \frac{c_{22}-c_{11}^2}{c_{02}c_{20}} + (1/4)\frac{c_{04}-c_{02}^2}{c_{02}^2}\frac{c_{11}^2}{c_{02}c_{20}} \\ &+(1/4)\frac{c_{40}-c_{20}^2}{c_{20}^2}\frac{c_{11}^2}{c_{02}c_{20}} - (c_{13}-c_{11}c_{02})\frac{c_{11}}{c_{02}^2c_{20}} \\ &-(c_{31}-c_{11}c_{20})\frac{c_{11}}{c_{20}^2c_{02}} + (1/2)(c_{22}-c_{02}c_{20})\frac{c_{11}^2}{c_{02}^2c_{20}^2}. \end{aligned} \tag{4.5}$$

The proofs of Theorems 4.1 and 4.2 are relatively straightforward and given in Sections 4.9.1 and 4.9.2 of the chapter appendix.

Estimating the Variance of the Huberized Correlation Coefficients

To estimate the asymptotic variance (4.5) of the Huberized correlation coefficient estimate, replace c_{ij} by $\hat{c}_{ij}$ which is defined as follows.

$$\hat{c}_{ij} = \frac{1}{n}\sum_{k=1}^{n}\psi^i\left(\frac{X_k-\hat{\mu}_X}{\hat{\sigma}_X}\right)\psi^j\left(\frac{Y_k-\hat{\mu}_Y}{\hat{\sigma}_Y}\right),$$

where $\hat{\mu}_X$ and $\hat{\mu}_Y$ are MM-location estimates and $\hat{\sigma}_X$ and $\hat{\sigma}_Y$ are S-scale estimates of X and Y, respectively.

We provide a C source called within Splus. The program computes the Huberized correlation coefficient estimate $\hat{r}$ (4.4) and its standard error SE($\hat{r}$). The latter is calculated from the asymptotic variance of the estimate (4.5). Below we give the skeleton of the program for computing the robust estimate. The input is a data set containing n 2-D points of the form (X_i, Y_i) and the tuning constant of the Huber score function c.

1. For each variable X and Y, do:

   (a) Compute the MM-location estimates and S-scale estimates to get $\hat{\mu}_X$, $\hat{\mu}_Y$, $\hat{\sigma}_X$ and $\hat{\sigma}_Y$.

   (b) For $i = 1$ to n, do: $\psi_c\left(\frac{X_i - \hat{\mu}_X}{\hat{\sigma}_X}\right)$ and $\psi_c\left(\frac{Y_i - \hat{\mu}_Y}{\hat{\sigma}_Y}\right)$.

2. Compute the Huberized correlation coefficient estimate

$$\hat{r} = \frac{\hat{c}_{11} - \hat{c}_{10}\hat{c}_{01}}{\sqrt{(\hat{c}_{20} - \hat{c}_{10}^2)(\hat{c}_{02} - \hat{c}_{01}^2)}}.$$

3. Compute the standard error $\mathrm{SE}(\hat{r}) = \sqrt{\frac{\mathrm{AV}}{n}}$.

4. Return the Huberized correlation coefficient estimate and its standard error: $(\hat{r}, \mathrm{SE}(\hat{r}))$.

The score function used to define the Huberized correlation coefficient estimate is not continuously differentiable. Since Theorem 4.2 requires this property, it is uncertain that the asymptotic variance formula (4.5) can be used to estimate the standard error of the Huberized correlation coefficient estimate. However we will show here, by means of the following numerical experiment, that the formula (4.5) can still be used.

To evaluate the accuracy of the estimated standard error of the Huberized correlation coefficient estimates, $\mathrm{SE}(\hat{r})$, we conducted a Monte Carlo experiment which consists of the following: Generate 5000 Monte Carlo samples of size n from a bivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. Compute $\hat{r}_i$ and $\mathrm{SE}(\hat{r}_i)$ for each sample ($i = 1, \ldots, 5000$), using the above computer program. Compute the empirical approximation to the standard deviation; $\mathrm{SD}(\hat{r}_1, \ldots, \hat{r}_{5000}) = \mathrm{SD}(\hat{\boldsymbol{r}})$, and the mean of the standard errors; $\mathrm{mean}(\mathrm{SE}(\hat{r}_1), \ldots, \mathrm{SE}(\hat{r}_{5000})) = \overline{\mathrm{SE}}$. To give some measure of variability of the standard errors, compute the standard deviation of the standard errors; $\mathrm{SD}(\mathrm{SE}(\hat{r}_1), \ldots, \mathrm{SE}(\hat{r}_{5000}))$.

The experiment was carried out for different sample sizes, $n = 20, 30, 50$ and 100 and with different correlation coefficients, $\rho = 0, .1, .25, .50, .75, .90$ and .99. The results are

	$n = 20$		$n = 30$		$n = 50$		$n = 100$	
ρ	SD($\hat{\boldsymbol{r}}$)	$\overline{\text{SE}}$	SD($\hat{\boldsymbol{r}}$)	$\overline{\text{SE}}$	SD($\hat{\boldsymbol{r}}$)	$\overline{\text{SE}}$	SD($\hat{\boldsymbol{r}}$)	$\overline{\text{SE}}$
0.00	0.228	0.211 (0.021)	0.187	0.176 (0.013)	0.141	0.139 (0.007)	0.099	0.099 (0.003)
0.10	0.224	0.210 (0.023)	0.185	0.175 (0.014)	0.143	0.138 (0.008)	0.100	0.098 (0.004)
0.25	0.220	0.202 (0.028)	0.177	0.168 (0.018)	0.139	0.132 (0.011)	0.097	0.095 (0.005)
0.50	0.194	0.174 (0.040)	0.155	0.143 (0.027)	0.121	0.113 (0.016)	0.083	0.080 (0.008)
0.75	0.140	0.115 (0.044)	0.106	0.094 (0.029)	0.083	0.073 (0.018)	0.057	0.052 (0.009)
0.90	0.076	0.056 (0.031)	0.057	0.046 (0.020)	0.042	0.035 (0.012)	0.029	0.025 (0.006)
0.99	0.010	0.007 (0.005)	0.007	0.005 (0.003)	0.005	0.004 (0.002)	0.003	0.003 (0.001)

Table 4.4: Evaluation of the Asymptotic Standard Errors of the Huberized Correlation Coefficient Estimates with $c = 1.00$.

displayed in Tables 4.4 – 4.6 for the tuning constant of the Huber score function $c = 1, 1.25$ and 1.50, respectively. For each sample size, n, the first column contains the Monte Carlo standard deviation of the Huberized correlation coefficient estimates. The second column contains the mean of the standard errors of the Huberized correlation coefficient estimates and the corresponding Monte Carlo standard deviation within parentheses. From the tables, we can see that the mean standard errors, $\overline{\text{SE}}$, approximate the empirical standard error of the estimates, SD($\hat{\boldsymbol{r}}$), closely. In particular, for large sample sizes and high correlations the difference between them is small and they have small standard deviation. Tables for different values of c are fairly similar, indicating that there is not much loss of efficiency by using relatively small value of c such as $c = 1$.

	$n = 20$		$n = 30$		$n = 50$		$n = 100$	
ρ	SD($\hat{r}$)	$\overline{\text{SE}}$	SD($\hat{r}$)	$\overline{\text{SE}}$	SD($\hat{r}$)	$\overline{\text{SE}}$	SD($\hat{r}$)	$\overline{\text{SE}}$
0.00	0.228	0.209 (0.025)	0.184	0.175 (0.016)	0.144	0.138 (0.009)	0.101	0.099 (0.004)
0.10	0.230	0.208 (0.027)	0.183	0.174 (0.016)	0.143	0.137 (0.010)	0.097	0.098 (0.005)
0.25	0.219	0.199 (0.033)	0.178	0.166 (0.021)	0.136	0.131 (0.012)	0.095	0.094 (0.006)
0.50	0.191	0.167 (0.042)	0.153	0.139 (0.029)	0.117	0.109 (0.017)	0.080	0.078 (0.009)
0.75	0.126	0.106 (0.042)	0.102	0.087 (0.027)	0.074	0.068 (0.017)	0.053	0.048 (0.009)
0.90	0.064	0.050 (0.026)	0.049	0.041 (0.017)	0.036	0.031 (0.010)	0.025	0.022 (0.005)
0.99	0.009	0.006 (0.004)	0.006	0.005 (0.003)	0.004	0.003 (0.001)	0.003	0.002 (0.001)

Table 4.5: Evaluation of the Asymptotic Standard Errors of the Huberized Correlation Coefficient Estimates with $c = 1.25$.

	$n = 20$		$n = 30$		$n = 50$		$n = 100$	
ρ	SD($\hat{r}$)	$\overline{\text{SE}}$	SD($\hat{r}$)	$\overline{\text{SE}}$	SD($\hat{r}$)	$\overline{\text{SE}}$	SD($\hat{r}$)	$\overline{\text{SE}}$
0.00	0.229	0.207 (0.029)	0.186	0.174 (0.019)	0.143	0.137 (0.011)	0.102	0.099 (0.005)
0.10	0.228	0.206 (0.030)	0.182	0.173 (0.019)	0.141	0.136 (0.011)	0.101	0.098 (0.006)
0.25	0.218	0.196 (0.035)	0.178	0.164 (0.023)	0.134	0.130 (0.014)	0.097	0.093 (0.007)
0.50	0.185	0.162 (0.043)	0.146	0.135 (0.029)	0.113	0.106 (0.018)	0.079	0.076 (0.009)
0.75	0.123	0.101 (0.040)	0.096	0.084 (0.028)	0.071	0.065 (0.016)	0.049	0.046 (0.008)
0.90	0.059	0.046 (0.024)	0.045	0.038 (0.016)	0.034	0.029 (0.009)	0.023	0.021 (0.005)
0.99	0.007	0.005 (0.003)	0.006	0.004 (0.002)	0.004	0.003 (0.001)	0.003	0.002 (0.001)

Table 4.6: Evaluation of the Asymptotic Standard Errors of the Huberized Correlation Coefficient Estimates with $c = 1.50$.

4.5 Bias in Quadrant and Huberized Correlation Coefficients

In this section, we discuss the bias that the quadrant and the Huberized correlation coefficient estimates may have due to the fraction of contamination in the data and because of the structure of the estimates. Therefore, we need to distinguish between two kinds of bias in the quadrant and the Huberized correlation coefficient estimates.

For simplicity and without loss of generality, we will restrict our attention to the case $p = 2$. Since in this case there are only two variables involved there is not much loss in using the classical contamination neighborhood,

$$\mathcal{H}_\epsilon(\rho) = \left\{ H : H = (1-\epsilon)\, H_0 + \epsilon \tilde{H} \right\}, \tag{4.6}$$

where ρ is the true correlation coefficient of the two variables under the nominal distribution H_0, $\rho(H_0)$. For example $H_0 = N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ and $\tilde{H}$ is an arbitrary and unspecified distribution.

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ and $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ be i.i.d. H and H_0, respectively, where $\boldsymbol{X}_i$ and $\boldsymbol{Y}_i \in \mathbb{R}^2$, $i = 1, \ldots, n$. Using the consistency result in Theorem 4.1, then

$$\begin{aligned} \hat{r}(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n) &\rightarrow r(H) \qquad \text{a.s.} \\ \hat{r}(\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_n) &\rightarrow r(H_0) \qquad \text{a.s.} \end{aligned}$$

as $n \rightarrow \infty$.

Because of the fraction of contamination included in H, $r(H)$ will typically be asymptotically biased. The asymptotic bias of $r(H)$ with $H \in \mathcal{H}_\epsilon(\rho)$, where $\mathcal{H}_\epsilon(\rho)$ is the family distribution of H generated by (4.6), can be written as

$$b(r, H) = |r(H) - r(H_0)|,$$

and the corresponding maximum asymptotic bias over $\mathcal{H}_\epsilon(\rho)$ can be expressed as

$$B_r(\epsilon) = \sup_{H \in \mathcal{H}_\epsilon(\rho)} |r(H) - r(H_0)|.$$

The other bias is the intrinsic bias that occurs at the nominal model H_0 because of the structure of the estimate, which requires transforming the data. This transformed data has a slightly different correlation than the correlation of the original data, so $r(H_0) \neq \rho(H_0)$. Thus, we define the maximum "*overall bias*" (OB) of the correlation coefficient $r(H)$ as follows:

$$\text{OB} = \sup_{H \in \mathcal{H}_\epsilon(\rho)} |r(H) - \rho(H_0)|.$$

To make this idea clear, we re-write the maximum overall bias as follows:

$$\text{OB} = \sup_{H \in \mathcal{H}_\epsilon(\rho)} |r(H) - r(H_0) + r(H_0) - \rho(H_0)|. \tag{4.7}$$

Hence, we can see that the overall bias (4.7) is composed of two different biases:

- Asymptotic bias, $|r(H) - r(H_0)|$, occurs due to the fraction of contamination in the data set.

- Intrinsic bias, $|r(H_0) - \rho(H_0)|$, happens because of the data transformation, "Huberizing" the data.

It is well known that because of the data transformation, the nature of the data will change. Specifically, when X and Y are jointly normal with correlation ρ, the limiting value r of $\hat{r}$ satisfies $|r| \leq |\rho|$, with strict inequality, except in the trivial cases $|\rho| = 0, 1$. The next theorem states this result.

THEOREM 4.3 *Let X, Y be jointly normal, with marginals $N(0,1)$, with $\mathbb{E}\{XY\} = \rho$. Suppose that f and g are measurable functions such that $\mathbb{E}\{f(X)^2\}$ and $\mathbb{E}\{g(Y)^2\}$ are finite. Then, the correlation of $f(X)$ and $g(Y)$ is less than or equal to ρ in absolute value.*

The folklore traces this result back to Kolmogorov, although we could not find published proof. Therefore, we give the proof of this theorem in Section 4.9.3 of the chapter appendix.

To verify Theorem 4.3 via empirical evidence, we conducted a Monte Carlo simulations using Huber score function with different values of the tuning constant c. A random sample of size $n = 100,000$ was generated from a bivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. Using the Huber score function for a specific value of the tuning constant c, the data were transformed (Huberized data). Then, the correlation coefficient of the transformed data is calculated from the following formula:

$$r = \frac{\mathbb{E}\left\{\psi(X)\psi(Y)\right\}}{\sqrt{\mathbb{E}\left\{\psi^2(X)\right\}\mathbb{E}\left\{\psi^2(Y)\right\}}}. \tag{4.8}$$

We compared the correlation coefficient of the transformed data r with the correlation coefficient ρ of the generated data for different values of the tuning constant c. We show the differences between the two correlations in Figures 4.5 and 4.6, which display plots of the correlation coefficient of the transformed data r versus the correlation coefficient ρ for different values of the tuning constant c. We see that the intrinsic bias decreases for larger values of the tuning constant c and becomes considerably smaller for $c \geq 1$. We also see that for moderate positive correlation coefficients $(.5 < \rho < .7)$ the magnitude of under estimation is larger. On the other hand, for moderate negative correlation coefficients $(-.7 < \rho < -.5)$ the magnitude of over estimation is larger.

To correct the intrinsic bias under the assumption of a Gaussian model we can use an appropriate non-decreasing transformation function,

$$\tilde{r} = g_c(r). \tag{4.9}$$

Specifically, for the quadrant correlation (QC) coefficient Huber (1981) suggested the following transformation:

$$g_{QC}(r) = \sin\left((\pi/2)\, r\right). \tag{4.10}$$

For general cases such as Huber score function with tuning constant $c > 0$, we suggest that $g_c(r)$ can be obtained by numerical means. Using the Monte Carlo simulations, we

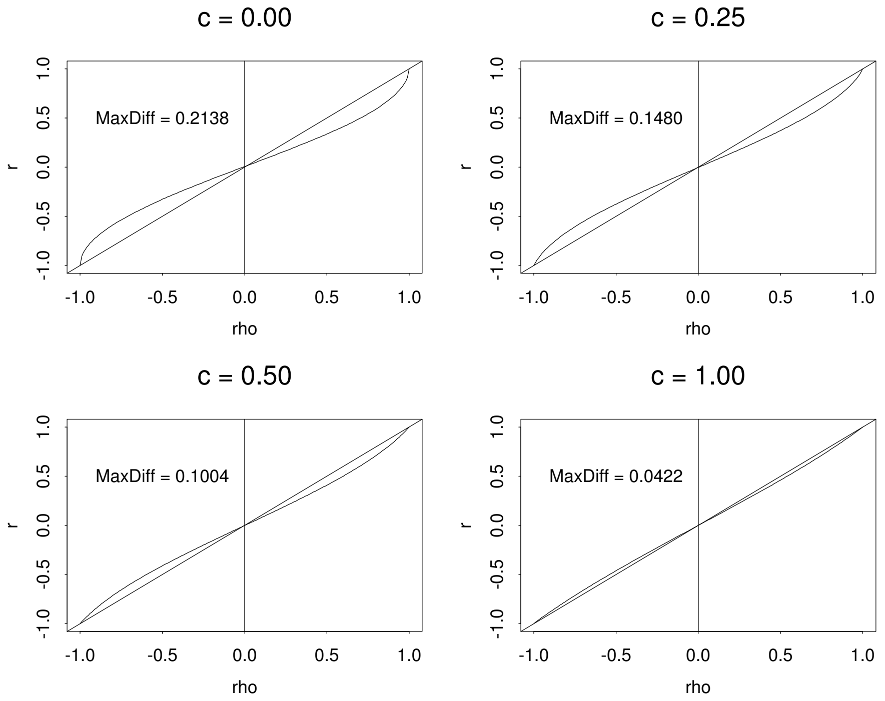


Figure 4.5: Intrinsic Bias of Huberized Correlation Coefficient Estimates.

calculated r using formula (4.8) and denoted by $r = r(\rho, c)$. Therefore, numerical tables can be obtained for different values of c and correlation coefficients, ρ. Each table allows us to read the value of ρ, denoted by $\rho = g_c(r)$, given r and c. However, if the value of r is not in the numerical tables then we can use interpolation to get $g_c(r)$.

The pairwise Huberized correlation matrix estimate $\hat{R} = (\hat{r}_{jk})_{j,k=1,\ldots,p}$ is a positive definite matrix. This is because it is constructed from the Pearson correlation coefficient estimates $\hat{r}_{jk}$ (4.2) of the outlier-free transformed data. However, when the correlation coefficient estimates are corrected for the intrinsic bias, the corrected correlation matrix estimate $\tilde{R} = (\tilde{r}_{jk})_{j,k=1,\ldots,p}$ is not positive definite. Fortunately, we have seen that for large values of the tuning constant c, e.g., $c = 1.00$ or so, the intrinsic bias is very small (less than .05) and, therefore, $\tilde{r}_{jk} \approx \hat{r}_{jk}$. In such cases we recommend using $\hat{r}_{jk}$ to preserve positive definiteness. On the other hand, when the bias correction is needed,

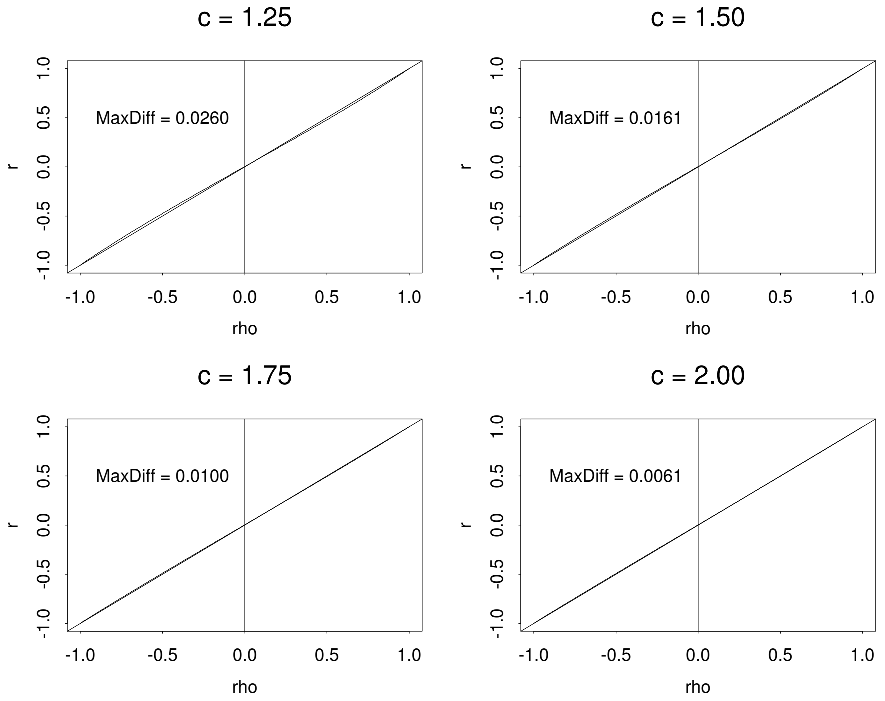


Figure 4.6: Intrinsic Bias of Huberized Correlation Coefficient Estimates.

there are intuitively appealing methods for adjusting the positive definiteness of the resulting scatter matrix. One such method is introduced by Maronna and Zamar (2002) which we will discuss in the next section.

4.6 Positive Definite Pairwise Robust Scatter Estimates

In this section, we describe a general method to obtain positive definite robust scatter estimates. The method was introduced by Maronna and Zamar (2002) for any pairwise robust scatter estimate. They applied their method to the bivariate outlier resistant estimate obtained by Gnanadesikan and Kettenring (1972) and Devlin, Gnanadesikan and Kettenring (1981). However, we will show that when applying this method to the quadrant correlation coefficient estimate we will obtain a pairwise robust scatter estimate which is computationally more feasible.

We now briefly describe Maronna and Zamar's (2002) method for correction of positive definiteness. Recall that if $\mathbf{\Sigma}$ is the covariance matrix of the p-dimensional random vector $\boldsymbol{X}$, then

$$\sigma^2(\mathbf{a}'\boldsymbol{X}) = \mathbf{a}'\mathbf{\Sigma}\mathbf{a}, \tag{4.11}$$

for all $\mathbf{a} \in \mathbb{R}^p$ and σ denotes the standard deviation. Let

$$\widehat{\mathbf{\Sigma}} = \sum_{j=1}^{p} \hat{\lambda}_j \hat{\mathbf{a}}_j \hat{\mathbf{a}}_j',$$

where $\hat{\lambda}_1 < \hat{\lambda}_2 < \ldots < \hat{\lambda}_p$ are eigenvalues of $\widehat{\mathbf{\Sigma}}$ and $\hat{\mathbf{a}}_j$ $(j = 1, \ldots, p)$ are the corresponding eigenvectors. One notices that, when $\widehat{\mathbf{\Sigma}}$ is the sample covariance, the $\hat{\lambda}_j$'s are the variances of the projected data on the direction of the corresponding eigenvectors. To solve the *negative eigenvalues* problem they proposed to replace the eigenvalues by the square of a robust scale estimate of the corresponding principle components in formula (4.11).

$$\hat{\lambda}_j = \hat{\sigma}^2(\mathbf{a}_j'\boldsymbol{X}_1, \mathbf{a}_j'\boldsymbol{X}_2, \ldots, \mathbf{a}_j'\boldsymbol{X}_n), \qquad j = 1, 2, \ldots, p.$$

Furthermore, Maronna and Zamar (2002) show that the estimate can still be improved upon by means of a re-weighting step. Hence, the final output is weighted mean and covariance matrix with weights based on the Mahalanobis distances:

$$d_i = d(\boldsymbol{X}_i) = (\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})'\widehat{\mathbf{\Sigma}}^{-1}(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}).$$

Let W be a weight function and define $\hat{\boldsymbol{\mu}}_w, \widehat{\mathbf{\Sigma}}_w$ as the weighted mean and covariance matrix, where each $\boldsymbol{X}_i$, $i = 1, \ldots, n$ has weight $W_i = W(d_i)$, that is,

$$\hat{\boldsymbol{\mu}}_w = \frac{\sum_{i=1}^n W_i \boldsymbol{X}_i}{\sum_{i=1}^n W_i}, \ \widehat{\mathbf{\Sigma}}_w = \frac{\sum_{i=1}^n W_i (\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_w)(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_w)'}{\sum_{i=1}^n W_i}.$$

They used the simplest W which is the "hard rejection", with $W(d) = I(d \leq d_0)$ and

$$d_0 = \frac{\chi_p^2(\beta)\text{med}(d_1, \ldots, d_n)}{\chi_p^2(.5)},$$

where $\chi_p^2(\beta)$ is the β-quantile of the chi-square distribution with p degrees of freedom.

4.6.1 Preliminary Estimation of Scatter Matrix

We have already seen that the general method can be applied to any robust estimate of the scatter matrix. Since we are interested in comparing the computational performance of various estimates of the scatter matrix, we will discuss two such estimates the Gnanadesikan and Kettenring (GK) and the quadrant correlation (QC).

The Gnanadesikan and Kettenring estimate is based on the following identity:

$$\mathrm{Cov}(X,Y) = \frac{1}{4}\left[\sigma^2(X+Y) - \sigma^2(X-Y)\right],$$

where X and Y are random variables and σ is the standard deviation. Gnanadesikan and Kettenring (1972) proposed to define a robust covariance matrix by using a robust scale as σ; they used a trimmed standard deviation. The resulting matrix is symmetric, but not necessarily positive definite and is not affine equivariant. Genton and Ma (1999) calculated its influence function and asymptotic efficiency.

Maronna and Zamar (2002) suggested the τ-scale introduced by Yohai and Zamar (1988), which is a truncated standard deviation, and a weighted mean to be the robust scale and location respectively. Define the functions:

$$W_c(x) = \left(1 - \left(\frac{x}{c}\right)^2\right)^2 I(|x| \le c); \;\; \rho_c(x) = \min(x^2, c^2).$$

Let $X = X_1, \ldots, X_n$ be a univariate sample; and put

$$\sigma_0 = \mathrm{MAD}(X) = \mathrm{med}\left(|X - \mathrm{med}(X)|\right); \;\; W_i = W_{c_1}\left(\frac{X_i - \mathrm{med}(X)}{\sigma_0}\right),$$

where $I(\cdot)$ is the indicator function and "med" denotes the median. Now, the weighted mean and the τ-scale estimates are defined as follows:

$$\hat{\mu}(X) = \frac{\sum_{i=1}^n X_i W_i}{\sum_{i=1}^n W_i}; \;\; \hat{\sigma}(X)^2 = \frac{\sigma_0^2}{n}\sum_{i=1}^n \rho_{c_2}\left(\frac{X_i - \mu(X)}{\sigma_0}\right).$$

To combine robustness and efficiency, Maronna and Zamar (2002) set $c_1 = 4.5$ and $c_2 = 3$. Ma and Genton (2001) advocated the use of the scale estimate Q_n proposed by Croux and Rousseeuw (1992b) and Rousseeuw and Croux (1993), but Maronna and Zamar prefer τ-scale estimate for reasons of speed.

It is well known that robust estimates suffer from lack of computational efficiency. Thus, we propose that better computational speed can be attained by using the quadrant correlation estimate in place of GK in Maronna and Zamar (2002) method for estimating scatter matrix.

The quadrant correlation is the sample correlation coefficient of the signs of the differences between the data and their respective medians. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be a multivariate sample where $\boldsymbol{X}_i \in \mathbb{R}^p, i = 1, \ldots, n$ with medians M_j $(j = 1, \ldots, p)$. The quadrant correlation coefficient estimate is defined as follows:

$$\hat{\rho}_{lk} = \frac{\sum_{i=1}^{n} \mathrm{SGN}(X_{il} - M_l)\mathrm{SGN}(X_{ik} - M_k)}{n},$$

where $l, k = 1, \ldots, p$.
The quadrant correlation coefficient estimate can be corrected for the intrinsic bias as follows:

$$\tilde{\rho}_{lk} = \sin\left(\frac{\pi}{2}\hat{\rho}_{lk}\right).$$

Let $\mathrm{MAD}(X_{il})$ and $\mathrm{MAD}(X_{ik})$ be the scale estimates of X_{il} and X_{ik}, respectively. Then the covariance estimate can be defined as:

$$\hat{\sigma}_{lk} = \mathrm{MAD}(X_{il})\mathrm{MAD}(X_{ik})\tilde{\rho}_{lk}.$$

From the $\hat{\sigma}_{lk}$ the initial robust covariance matrix estimate is formed:

$$\widehat{\boldsymbol{\Sigma}}_0 = \{\hat{\sigma}_{lk}\}_{l,k=1,\ldots,p}.$$

The final positive definite robust covariance matrix is formed by applying Maronna and Zamar's (2002) method for correction of positive definiteness followed by the final reweighting step.

4.6.2 Computational Complexity and Computing Times

The resulting robust dispersion matrix has different computation complexity for different robust methods. The pairwise approach is appealing in that it reduces the computational complexity in the data dimension p from exponential to quadratic (from 2^p to p^2). The computational complexity of some robust methods are presented below.

- Stahel-Donoho (SD) estimate with naïve implementation has computation complexity $O(2^p) \cdot O(n^2)$; whereas, with better implementation the Stahel-Donoho estimate has computation complexity $O(2^p) \cdot O(n \cdot \log(n))$.

- MCD estimate has computation complexity $O(2^p) \cdot O(n)$; whereas, the FAST MCD estimate has the same computation complexity with a much better constant for $O(n)$.

- The robust pairwise scatter estimate has computation complexity $O(p^2) \cdot O(n)$.

We compared the computing times of the pairwise covariance matrix estimates obtained using the quadrant correlation (QC) estimate and the GK estimate (with τ-scale robust estimate). The Maronna and Zamar (2002) method is applied to correct for positive definiteness of the resulting covariance matrix estimates with the final re-weighting step. Both programs for the two estimates were written in C and called within Splus for Unix. Moreover, we made the programs available as a built-in Splus command in the recently released Splus, namely Splus 6 for Windows and Splus 6.1 for Unix. The command lines for the estimates (using robust library) are the following:

- The Gnanadesikan-Kettenring estimate (with τ-scale robust estimate); covRob(stack.dat, estim = “pairwisegk”);

- The quadrant correlation (QC) estimate; covRob(stack.dat, estim = “pairwiseqc”).

We carried out some timing experiments for a range of sample size n and dimension p. We ran the experiments on Splus (version 6.0, Sun0S 5.6). To make the calculation

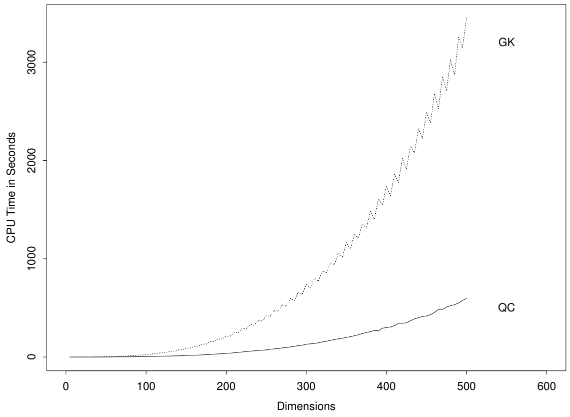


Figure 4.7: Scalability of Dimensions for the Covariance Estimates obtained using QC and GK for $n = 5p$.

of the median run faster, we did not use the built-in Splus command "median", which employs sorting; but a selection algorithm (the procedure "select" in Section 8.5 of Press et al., 1992), which is linear in n.

We generated standard normal random samples with different values of n and p. The computing times of the covariance estimates obtained using QC and GK for different data dimensions with sample sizes $n = 5p$ is shown in Figure 4.7. The figure displays plots of the computing time (time in seconds) versus the data dimension, p. We can see that QC requires less computing time than GK. Also, the computing times of GK and QC tend to increase quadratically as p increases. However, QC has a smaller constant for the quadratic polynomial in p. The computing times of the covariance estimates obtained using QC and GK for different sample sizes with data dimension $p = 50$ is shown in Figure 4.8. The figure displays plots of the computing time (time in seconds) versus the sample size, n. We can see that QC is much faster than GK. Also, the computing times of GK and QC tend to increase linearly as n increases. However, QC has a smaller slope. Hence, from Figures 4.7 and 4.8 we can conclude that the computing time growth of the pairwise scatter estimates is linear in n and quadratic in p, which confirms the above complexity claim.

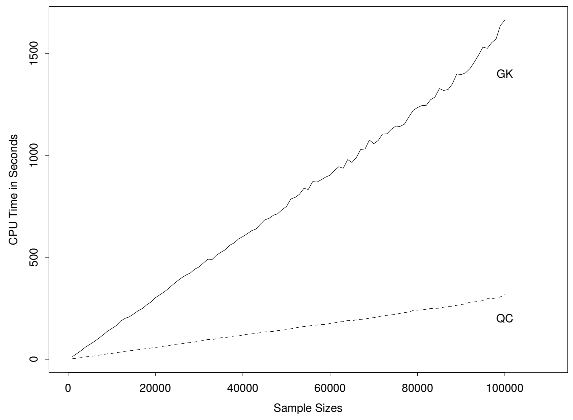


Figure 4.8: Scalability of Sample Sizes for the Covariance Estimates obtained using QC and GK for $p = 50$.

We compared the computing times of the pairwise covariance estimates obtained using QC and GK (with τ-scale robust estimate) with the Fast MCD (FMCD) covariance estimates. The computing times of the covariance estimates obtained using QC, GK and FMCD for different sample sizes with dimensions $p = 10, 30$ and 50 are shown in Figure 4.9, which displays plots of the computing time (time in seconds) versus the sample size, n. We can see that for higher dimensions, FMCD requires a much larger amount of computing times than QC and GK. The computing times of the covariance estimates obtained using QC, GK and FMCD for larger sample sizes are shown in Figure 4.10, which displays plots of the computing time (time in seconds) versus the sample size, n. We can see that for larger sample sizes, QC still requires less computing times than GK and FMCD. On the other hand, for higher dimensions, GK requires a much larger amount of computing times than QC and FMCD. We also notice that, for larger sample sizes, FMCD requires less amount of computing times. The reason that FMCD requires less computing times for a larger sample size is when n is larger than a certain n_0 (the default is 6000) the FMCD algorithm applies an ingenious splitting procedure to reduce the number of evaluations.

We also ran the timing experiments for contaminated data. The contaminated samples were p-variate normal ϵ-contaminated distribution, with p taking the values 10, 30 and 50. Generate $\boldsymbol{X}_i$ as p-variate normals $\mathrm{N}_p(\mathbf{0}, I)$ for $i = 1, \ldots, n{-}m$, where $m = [n\epsilon]$; $[\cdot]$ denotes the integer part, and as $\mathrm{N}_p(\boldsymbol{\mu}_0, \delta^2 I)$ for $i > n{-}m$. We chose $\boldsymbol{\mu}_0 = (100, \ldots, 100)_p$, $\delta = 0.1$ and $\epsilon = 0.20$. The computing times of the covariance estimates obtained using QC, GK and FMCD for different sample sizes with dimensions $p = 10, 30$ and 50 are shown in Figure 4.11. The computing times of the covariance estimates obtained using QC, GK and FMCD for larger sample sizes are shown in Figure 4.12. The figures display plots of the computing time (time in seconds) versus the sample size, n. From the plots, we can see that the three estimates have similar timing behavior compared to the clean data timing situations.

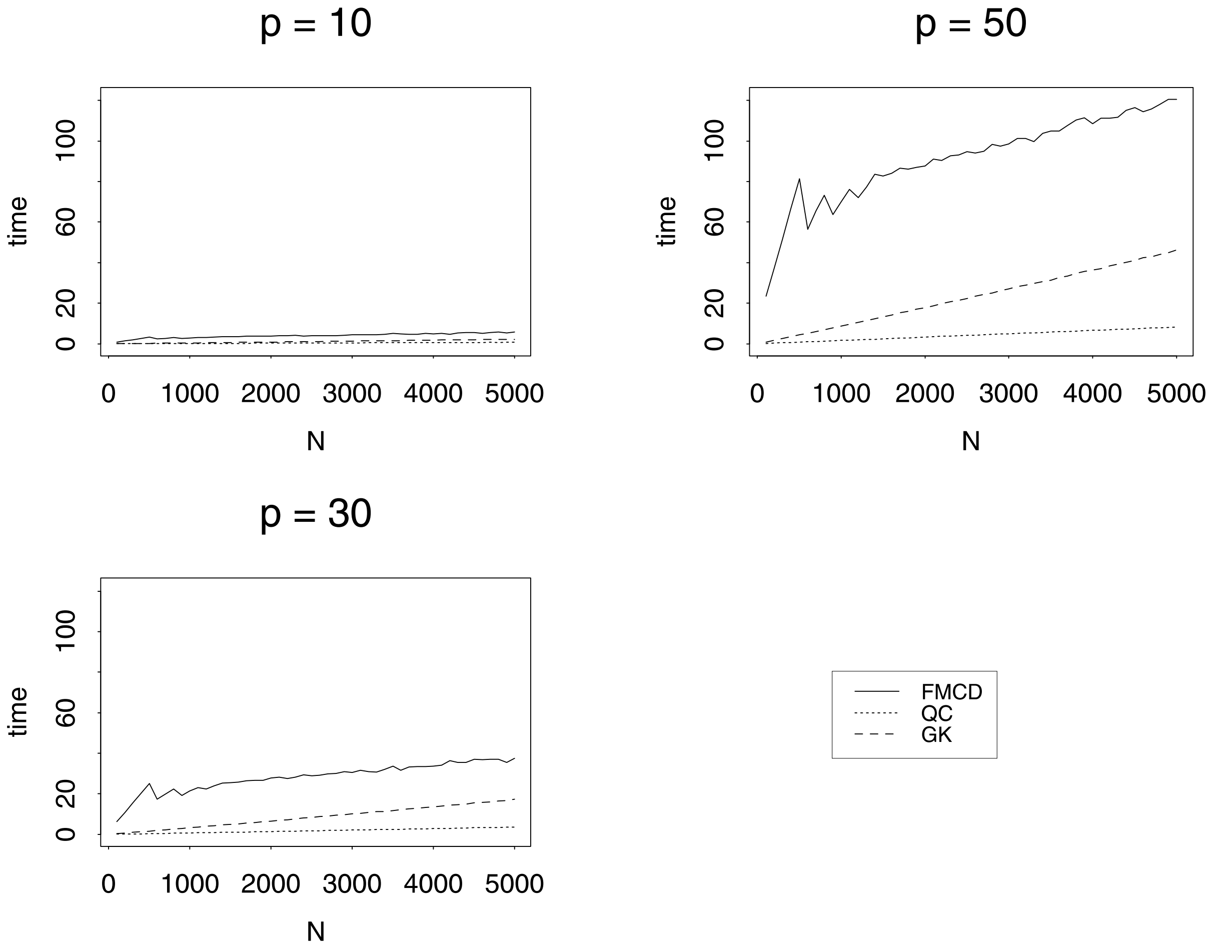


Figure 4.9: CPU Time of the Covariance Estimates obtained using FMCD, QC and GK for Clean Data.

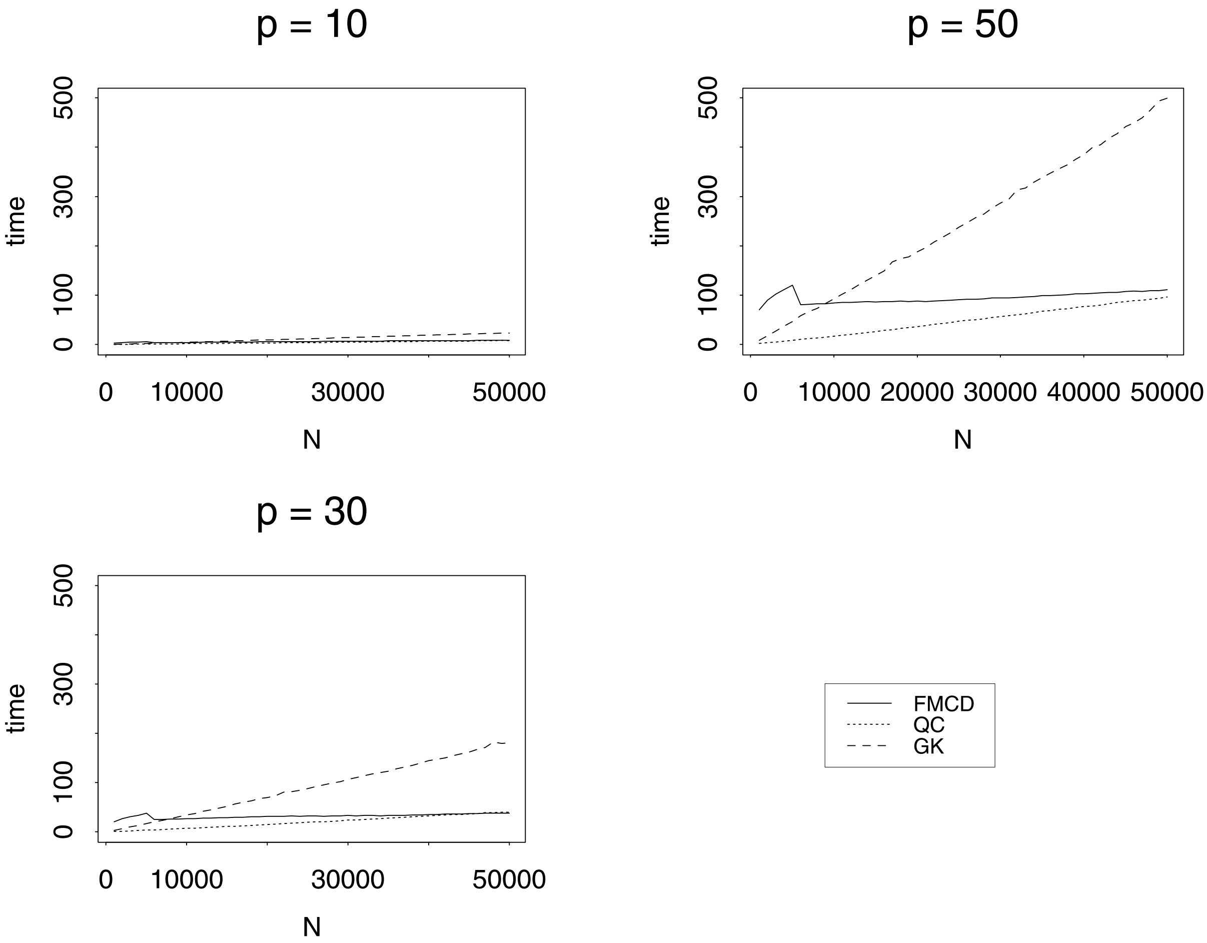


Figure 4.10: CPU Time of the Covariance Estimates obtained using FMCD, QC and GK for Clean Data with Larger Sample Sizes.

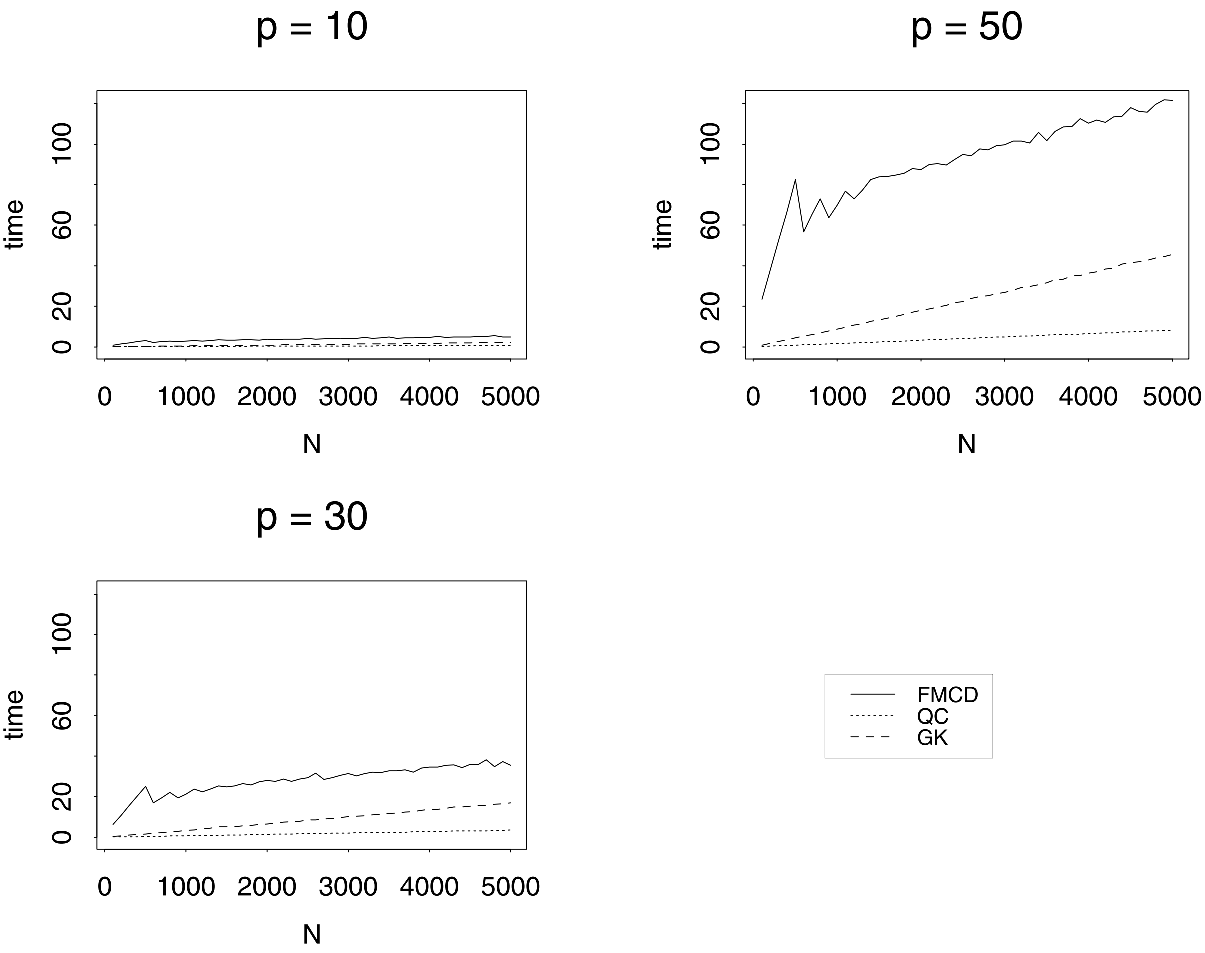


Figure 4.11: CPU Time of the Covariance Estimates obtained using FMCD, QC and GK for 20% Contaminated Data.

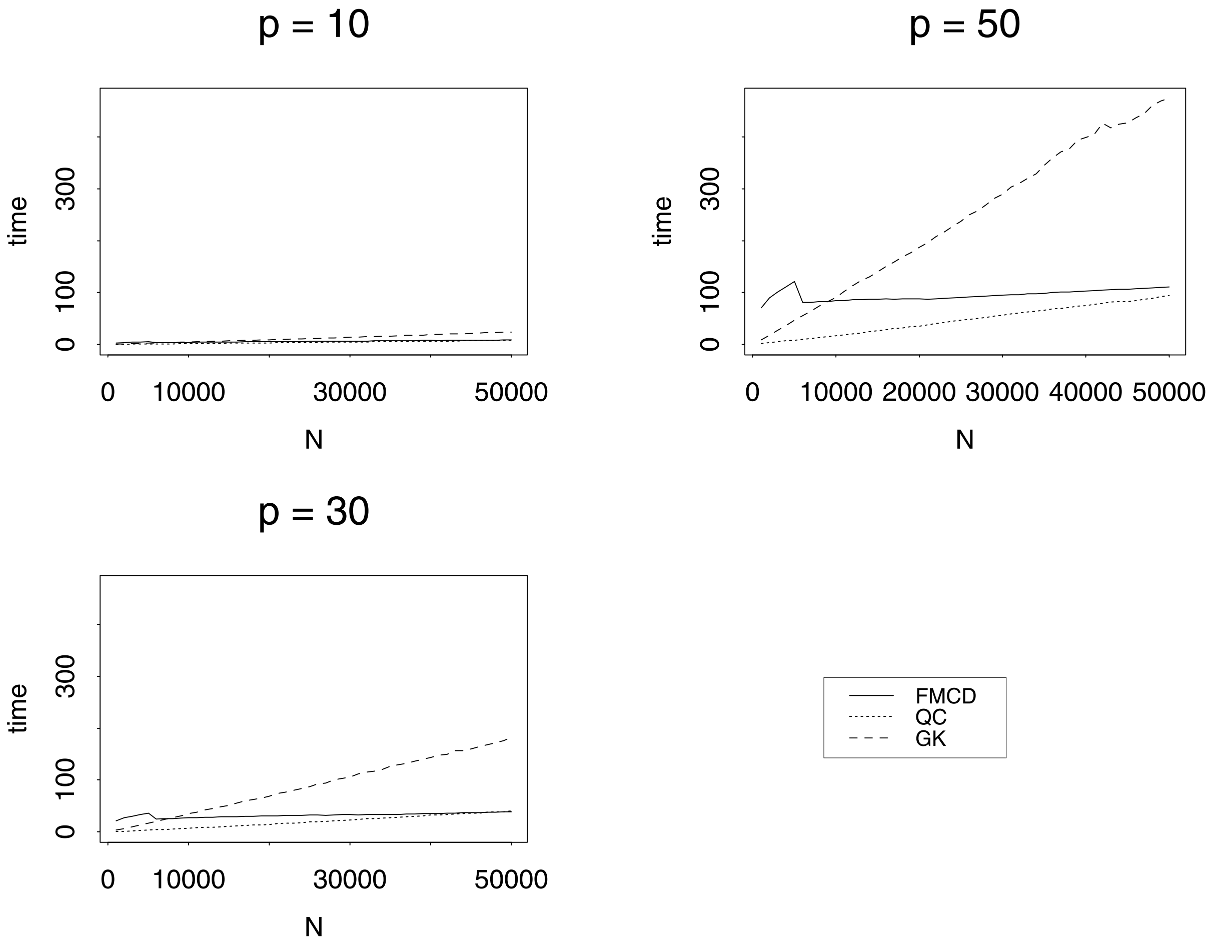


Figure 4.12: CPU Time of the Covariance Estimates obtained using FMCD, QC and GK for 20% Contaminated Data with Larger Sample Sizes.

4.7 Maximum Bias of Quadrant and Huberized Correlation Coefficients

In this section, we study the maximum asymptotic bias (maxbias) of the quadrant and Huberized correlation coefficient estimates in the contamination neighborhoods. We are also interested in comparing their maxbias with the maxbias of the affine equivariant estimates such as the Fast MCD and the Stahel-Donoho.

For simplicity and without loss of generality, we assume that the number of variables $p = 2$. Hence, it is appropriate to use the classical contamination neighborhood (4.6) as a good approximation.

The rest of this section is organized as follows. Section 4.7.1 studies the maxbias of the quadrant and Huberized correlation coefficient estimates when the location and scale parameters are known. Section 4.7.2 considers the maxbias of the quadrant correlation coefficient estimate with unknown location and scale parameters. Section 4.7.3 deals with numerical computation of the maxbias of the Huberized correlation coefficient estimates when the location and scale parameters are unknown. Finally, Section 4.7.4 compares the maxbias of the Huberized correlation coefficient estimates with the maxbias of the Fast MCD and the Stahel-Donoho correlation coefficient estimates.

4.7.1 Maxbias of Quadrant and Huberized Correlation Coefficients with Known Locations and Scales

The following theorem shows the maxbias of the quadrant and Huberized correlation coefficient estimates when the location and scale parameters are known.

THEOREM 4.4 – **Worst Case Bias** – *The maximum asymptotic bias $B(\epsilon)$ under the classical contamination neighborhood (4.6) of size ϵ, $\sup_{H \in \mathcal{H}_\epsilon(\rho)} |r(H) - \rho|$ is given by:*

$$\max\left\{\left|g_c\left(\frac{r(H_0) - \beta}{1+\beta}\right) - \rho\right|, \left|g_c\left(\frac{r(H_0) + \beta}{1+\beta}\right) - \rho\right|\right\}, \qquad (4.12)$$

where $\beta = [\epsilon/(1-\epsilon)]\left[\psi^2(\infty)/I\!\!E\{\psi^2(Z)\}\right]$ *with* $Z \sim N(0,1)$*, and let* $g_c(\cdot)$ *be defined as in (4.9).*

Sketch of the Proof:

Assume without loss of generality that $\psi(\infty) = 1$. Let $A = \mathbb{E}_{H_0}\{\psi(X)\psi(Y)\}$, $B = \mathbb{E}_{H_0}\{\psi^2(X)\}$, $a = \mathbb{E}_{\tilde{H}}\{\psi^2(X)\}$, $b = \mathbb{E}_{\tilde{H}}\{\psi^2(Y)\}$, and

$$r(H) = \frac{\mathbb{E}_H\{\psi(X)\psi(Y)\}}{\sqrt{\mathbb{E}_H\{\psi^2(X)\}\,\mathbb{E}_H\{\psi^2(Y)\}}}.$$

By the Cauchy-Schwarz inequality

$$r(H) \leq \frac{(1-\epsilon)A + \epsilon\sqrt{ab}}{\sqrt{(1-\epsilon)B+\epsilon a}\sqrt{(1-\epsilon)B+\epsilon b}}.$$

Differentiating the right hand side with respect to a and using the Cauchy-Schwarz inequality again we can verify that this derivative is non-negative for all $a \leq b$. Therefore, letting $(\epsilon/(1-\epsilon))/b = \beta$ and noticing that $r(H_0) = A/B$ we can write

$$r(H) \leq \frac{(1-\epsilon)A + \epsilon b}{(1-\epsilon)B + \epsilon b} \leq \frac{(1-\epsilon)A + \epsilon}{(1-\epsilon)B + \epsilon} = \frac{r(H_0)+\beta}{1+\beta}. \tag{4.13}$$

The second inequality follows because

$$[(1-\epsilon)A + \epsilon b]\,/\,[(1-\epsilon)B + \epsilon b]$$

is increasing in b. An analogous reasoning gives

$$r(H) \geq \frac{r(H_0)-\beta}{1+\beta}. \tag{4.14}$$

Huber (1981) states inequalities (4.13) and (4.14) without providing a proof. The result follows now because the function $\rho = g_c(r)$ is non-decreasing. We give a detailed proof of Theorem 4.4 in Section 4.9.4 of the chapter appendix.

4.7.2 Maxbias of Quadrant Correlation Coefficient with Unknown Locations and Scales

Before we derive the maxbias of the quadrant correlation coefficient estimates, when the location and scale parameters are unknown, we need to state the following lemma.

Lemma 4.1 *Suppose that the random vector (X, Y) has a bivariate normal distribution H_0 with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ and let $x_0 = \Phi^{-1}\left(\frac{1}{2(1-\epsilon)}\right)$. For any a and b in the interval $[-x_0, x_0]$ define:*

$$g(a,b) = \mathbb{E}_{H_0}\left\{SGN(X-a)SGN(Y-b)\right\}.$$

Then the following are satisfied:

(a) g(a,b) = g(b,a);

(b) $\max\limits_{|a|,|b|\leq x_0} g(a,b) = g(-x_0,-x_0)$*;*

(c) $\min\limits_{|a|,|b|\leq x_0} g(a,b) = g(-x_0,+x_0)$*.*

Proof:

Assume without loss of generality that $\rho \geq 0$. Part (a) follows because $(X, Y) \sim H_0$ are exchangeable:

$$\begin{aligned} g(a,b) &= \mathbb{E}_{H_0}\left\{\text{SGN}(X-a)\text{SGN}(Y-b)\right\} \\ &= \mathbb{E}_{H_0}\left\{\text{SGN}(Y-a)\text{SGN}(X-b)\right\} = g(b,a). \end{aligned}$$

Because of (a), to show (b) and (c) we only need to consider the upper triangle $\{(a,b) : x_0 \geq b \geq a \geq -x_0\}$. Let

$$\begin{aligned} g(a,b) &= \mathbb{E}_{H_0}\left\{\text{SGN}(X-a)\text{SGN}(Y-b)\right\} \\ &= \mathbb{E}_{H_0}\left\{\mathbb{E}_{H_0}\left\{\text{SGN}(Y-b)|X\right\}\text{SGN}(X-a)\right\}, \end{aligned}$$

where the conditional distribution of Y given $X = x$ is normal with mean ρx and variance $1 - \rho^2$. Now

$$\begin{aligned} \mathbb{E}_{H_0}\left\{\text{SGN}(Y-b)|X=x\right\} &= P\left(Y > b|X=x\right) - P\left(Y < b|X=x\right) \\ &= 1 - \Phi\left(\frac{b-\rho x}{\sqrt{1-\rho^2}}\right) - \Phi\left(\frac{b-\rho x}{\sqrt{1-\rho^2}}\right) \\ &= 1 - 2\Phi\left(\frac{b-\rho x}{\sqrt{1-\rho^2}}\right). \end{aligned}$$

Then

$$g(a,b)=\int_{a}^{\infty}\left[1-2\Phi\left(\frac{b-\rho x}{\sqrt{1-\rho^{2}}}\right)\right]\phi(x)dx-\int_{-\infty}^{a}\left[1-2\Phi\left(\frac{b-\rho x}{\sqrt{1-\rho^{2}}}\right)\right]\phi(x)dx.$$

Fixing b and differentiating with respect to a we get

$$\frac{\partial}{\partial a}g(a,b) \;=\; 2\left[2\Phi\left(\frac{b-\rho a}{\sqrt{1-\rho^{2}}}\right)-1\right].$$

The sign of this expression depends on the sign of $b-\rho a$. To investigate the sign of $b-\rho a$, we break the domain of b into three intervals as follows:

Case I: $-x_0 \leq b \leq -\rho x_0$. For any $a \in [-x_0, x_0]$ we have $b \leq -\rho x_0 \leq \rho a \leq \rho x_0$ and so $b-\rho a \leq 0$. So $g(a,b)$ decreases with a in the given interval. Therefore,

$$\max_{-x_0\leq a\leq x_0} g(a,b)=g(-x_0,b).$$

Case II: $-\rho x_0 < b \leq \rho x_0$. For any $a \in [-x_0, x_0]$, if $-x_0 \leq a \leq b/\rho$, then $-\rho x_0 \leq \rho a \leq b$ and so $b-\rho a \geq 0$. So $g(a,b)$ increases with $a \in [-x_0, b/\rho]$. And if $b/\rho < a \leq x_0$, then $b \leq \rho a \leq \rho x_0$ and so $b-\rho a \leq 0$. So $g(a,b)$ decreases with $a \in (b/\rho, x_0]$. Therefore,

$$\max_{-x_0\leq a\leq x_0} g(a,b)=g(b/\rho,b).$$

Case III: $\rho x_0 < b \leq x_0$. For any $a \in [-x_0, x_0]$, if $0 \leq a \leq x_0$, then $\rho a \leq a$ and so $b-\rho a \geq b-a \geq 0$. And if $-x_0 \leq a < 0$, then $-\rho a > 0$ and so $b-\rho a > 0$. So $g(a,b)$ increases with a in the given interval. Therefore,

$$\max_{-x_0\leq a\leq x_0} g(a,b)=g(x_0,b).$$

For all $-x_0 \leq b \leq -\rho x_0$ we have $g(-x_0,b)=g(b,-x_0)$, then by Case I

$$g(b,-x_0)\leq g(-x_0,-x_0).$$

For all $\rho x_0 < b \leq x_0$ we have $g(x_0,b)=g(b,x_0)$, then by Case III

$$g(b,x_0)\leq g(x_0,x_0)=g(-x_0,-x_0).$$

For all $-\rho x_0 < b \leq \rho x_0$ then $-x_0 \leq b/\rho \leq x_0$. We consider three cases:

Case 1: $b/\rho \leq -\rho x_0$. In this case

$$\begin{aligned} g(b/\rho, b) &= g(b, b/\rho) \\ &\leq g(-x_0, b/\rho) = (b/\rho, -x_0) \leq g(-x_0, -x_0). \end{aligned}$$

Case 2: $-\rho x_0 < b/\rho \leq \rho x_0$. In this case

$$\begin{aligned} g(b/\rho, b) = g(b, b/\rho) &\leq g(b/\rho, b/\rho) \\ &\leq g(b/\rho^2, b/\rho) \\ &= g(b/\rho, b/\rho^2) \\ &\leq g(b/\rho^2, b/\rho^2). \end{aligned}$$

Let k be such that $-\rho x_0 \leq b/\rho^{k-1} \leq \rho x_0$, but not for b/ρ^k. Eventually, $b/\rho^k < -\rho x_0$ or $b/\rho^k > \rho x_0$. Then $g(b, b/\rho) \leq g(-x_0, -x_0)$ or $g(b, b/\rho) \leq g(x_0, x_0)$.

Case 3: $b/\rho \geq \rho x_0$. In this case

$$\begin{aligned} g(b/\rho, b) = g(b, b/\rho) &\leq g(x_0, b/\rho) \\ &= g(b/\rho, x_0) \leq g(x_0, x_0). \end{aligned}$$ ■

Now in the following theorem we derive the maxbias of the quadrant correlation coefficient estimate.

THEOREM 4.5 – **Maxbias of the Quadrant Correlation Coefficient** – *If (X, Y) is distributed according to the classical contamination model H (4.6) where $H_0 = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then the quadrant correlation coefficient $r_{QC}(H)$ has the following properties:*

(a) $\sup_{H \in \mathcal{H}_\epsilon(\rho)} r_{QC}(H) = (1-\epsilon) g(-x_0, -x_0) + \epsilon$;

(b) $\inf_{H \in \mathcal{H}_\epsilon(\rho)} r_{QC}(H) = (1-\epsilon) g(-x_0, +x_0) - \epsilon$.

Proof:

We assume without loss of generality that $\mu_X = \mu_Y = 0$, $\sigma_X = \sigma_Y = 1$ and $\sigma_{XY} = \rho$. Therefore the bivariate random variable (X, Y) is distributed according to model (4.6) with

$$H_0 \;=\; \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right), \qquad -1 \leq \rho \leq 1.$$

Let M_X and M_Y be the medians of X and Y under H, and let

$$r_{QC}(H) \;=\; \mathbb{E}_H \left\{ \mathrm{SGN}(X - M_X)\mathrm{SGN}(Y - M_Y) \right\}$$

be the quadrant correlation coefficient between X and Y under the contaminated distribution H. Since $|M_X| \leq x_0$ and $|M_Y| \leq x_0$, where x_0 is as in Lemma 4.1. Then

$$r_{QC}(H) \;\leq\; (1-\epsilon) \max_{-x_0 \leq M_X, M_Y \leq x_0} g(M_X, M_Y) + \epsilon.$$

By Lemma 4.1 (b) $\max_{-x_0 \leq M_X, M_Y \leq x_0} g(M_X, M_Y) = g(-x_0, -x_0)$, and therefore the right hand side of (a) is upper bound for $r_{QC}(H)$.

Now let $\overline{H} = (1-\epsilon)H_0 + \epsilon\delta_{(-\infty,-\infty)}$, and notice that $M_X(\overline{H}) = M_Y(\overline{H}) = -x_0$ and

$$r_{QC}(\overline{H}) = (1-\epsilon) g_{H_0}(-x_0, -x_0) + \epsilon,$$

proving (a).

The proof of part (b) follows along the same lines with $\overline{H}$ replaced by $\underline{H} = (1-\epsilon)H_0 + \epsilon\delta_{(+\infty,+\infty)}$, and noticing that $M_Y(\underline{H}) = +x_0$ and $M_X(\underline{H}) = -x_0$, and

$$r_{QC}(\underline{H}) = (1-\epsilon) g_{H_0}(-x_0, +x_0) - \epsilon. \qquad \blacksquare$$

4.7.3 Maxbias of Huberized Correlation Coefficients with Unknown Locations and Scales

The derivation of the maxbias of the Huberized correlation coefficient estimates when the location and scale parameters are unknown is not tractable. Therefore, we decided to derive the maxbias using numerical computations which we describe below.

Let X and Y be two random variables jointly distributed under the point mass contamination model:

$$H_{(x_0,y_0)} = (1-\epsilon)H_0 + \epsilon\delta_{(x_0,y_0)}, \tag{4.15}$$

where $\delta_{(x_0,y_0)}$ is a point mass distribution at (x_0, y_0) and

$$H_0 = \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right).$$

Let the median, M, be the robust location estimate and the median absolute deviation (MAD), S, be the robust scale estimate. To obtain the median and the MAD of X, consider the univariate point mass contamination model which can be expressed as follows:

$$F_{x_0}(x) = (1-\epsilon)\Phi(x) + \epsilon\delta_{x_0}(x), \tag{4.16}$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function and

$$\delta_{x_0}(x) = \begin{cases} 0 & x < x_0 \\ 1 & x \geq x_0. \end{cases}$$

Then the distribution function of X can be written as

$$F_{x_0}(x) = \begin{cases} (1-\epsilon)\Phi(x) & x < x_0 \\ (1-\epsilon)\Phi(x) + \epsilon & x \geq x_0. \end{cases}$$

Let $c = \Phi^{-1}\left(\frac{1}{2(1-\epsilon)}\right)$, then the median of X, M_X, can be expressed as

$$M_X = \begin{cases} x_0 & |x_0| < c \\ c \ \mathrm{SGN}(x_0) & |x_0| \geq c. \end{cases}$$

To see this, first notice that to cover the case of discontinuous distributions like (4.16), the definition of median is extended as follows:

$$M_X = \sup\left\{t : F_{x_0}(t) \leq \frac{1}{2}\right\}.$$

We will restrict attention to the case $x_0 \geq 0$. The case of negative x_0 can be dealt with similarly. Let $0 \leq x_0 \leq c$ for all $x \leq x_0$, $F_{x_0}(x) = (1-\epsilon)\Phi(x) \leq (1-\epsilon)\Phi(x_0) \leq (1-\epsilon)\Phi(c) = \frac{1}{2}$. On the other hand, for all $x > x_0$, $F_{x_0}(x) > F_{x_0}(x_0) = (1-\epsilon)\Phi(x_0) + \epsilon \geq (1-\epsilon)\Phi(0) + \epsilon = \frac{1}{2} + \frac{\epsilon}{2} > \frac{1}{2}$. Therefore, $M_X = x_0$. Finally, for $x_0 > c$ since $(1-\epsilon)\Phi(c) = \frac{1}{2}$ and $F'_{x_0}(c) = (1-\epsilon)\phi(c) > 0$, the median is equal to c.

Now to compute MAD of X, $S_X = \text{med}|X - M_X|/\Phi^{-1}(3/4)$. We have to solve for u the following equation:

$$P_{x_0}\left(|X - M_X| \leq u\right) = 1/2,$$

which can be expressed as

$$F_{x_0}(M_X + u) - F_{x_0}(M_X - u) = 1/2. \tag{4.17}$$

For the case $|x_0| < c$, where $M_X = x_0$. Substitute M_X in equation (4.17), we get

$$F_{x_0}(x_0 + u) - F_{x_0}(x_0 - u) = 1/2,$$

where

$$\begin{aligned} F_{x_0}(x_0 + u) &= (1-\epsilon)\Phi(x_0 + u) + \epsilon, \qquad & x_0 + u \geq x_0 \\ F_{x_0}(x_0 - u) &= (1-\epsilon)\Phi(x_0 - u), & x_0 - u < x_0 \end{aligned}$$

which implies that

$$\Phi(x_0 + u) - \Phi(x_0 - u) = \frac{0.5 - \epsilon}{1 - \epsilon}. \tag{4.18}$$

Using Newton-Raphson method to solve for u in equation (4.18), we get $S_X = u/\Phi^{-1}(3/4)$. For the case $|x_0| \geq c$, where $M_X = c \ \text{SGN}(x_0)$. Let $b = c \ \text{SGN}(x_0)$ and substitute M_X in equation (4.17), we get

$$F_{x_0}(b + u) - F_{x_0}(b - u) = 1/2, \tag{4.19}$$

where

$$F_{x_0}(b+u) = \begin{cases} (1-\epsilon)\Phi(b+u) & b+u < x_0 \\ (1-\epsilon)\Phi(b+u)+\epsilon & b+u \geq x_0. \end{cases}$$

$$F_{x_0}(b-u) = \begin{cases} (1-\epsilon)\Phi(b-u) & b-u < x_0 \\ (1-\epsilon)\Phi(b-u)+\epsilon & b-u \geq x_0. \end{cases}$$

Using Newton-Raphson method to solve for u in equation (4.19), we get $S_X = u/\Phi^{-1}(3/4)$. Similarly, we obtain median of Y, M_Y, and MAD of Y, S_Y.

The Huberized correlation coefficient under the point mass contamination model (4.15) is defined as follows:

$$r(H_{(x_0,y_0)}) = \frac{\mathbb{E}\left\{\psi\left(\frac{X-M_X}{S_X}\right)\psi\left(\frac{Y-M_Y}{S_Y}\right)\right\} - \mathbb{E}\left\{\psi\left(\frac{X-M_X}{S_X}\right)\right\}\mathbb{E}\left\{\psi\left(\frac{Y-M_Y}{S_Y}\right)\right\}}{\sqrt{\left[\mathbb{E}\psi^2\left(\frac{X-M_X}{S_X}\right) - \left(\mathbb{E}\psi^2\left(\frac{X-M_X}{S_X}\right)\right)^2\right]\left[\mathbb{E}\psi^2\left(\frac{Y-M_Y}{S_Y}\right) - \left(\mathbb{E}\psi^2\left(\frac{Y-M_Y}{S_Y}\right)\right)^2\right]}},$$

where the numerator can be written as

$$\begin{aligned} & (1-\epsilon)\mathbb{E}_{H_0}\left\{\psi\left(\frac{X-M_X}{S_X}\right)\psi\left(\frac{Y-M_Y}{S_Y}\right)\right\} + \epsilon\psi\left(\frac{x_0-M_X}{S_X}\right)\psi\left(\frac{y_0-M_Y}{S_Y}\right) \\ - & \left[(1-\epsilon)\mathbb{E}_{H_0}\psi\left(\frac{X-M_X}{S_X}\right) + \epsilon\psi\left(\frac{x_0-M_X}{S_X}\right)\right]\left[(1-\epsilon)\mathbb{E}_{H_0}\psi\left(\frac{Y-M_Y}{S_Y}\right)\right. \\ + & \left.\epsilon\psi\left(\frac{y_0-M_Y}{S_Y}\right)\right]. \end{aligned}$$

The first factor of the denominator can be written as

$$\begin{aligned} & (1-\epsilon)\mathbb{E}_{H_0}\psi^2\left(\frac{X-M_X}{S_X}\right) + \epsilon\psi^2\left(\frac{x_0-M_X}{S_X}\right) \\ - & \left[(1-\epsilon)\mathbb{E}_{H_0}\psi\left(\frac{X-M_X}{S_X}\right) + \epsilon\psi\left(\frac{x_0-M_X}{S_X}\right)\right]^2, \end{aligned}$$

and the second factor of the denominator can be written as

$$(1-\epsilon)\mathbb{E}_{H_0}\psi^2\left(\frac{Y-M_Y}{S_Y}\right)+\epsilon\psi^2\left(\frac{y_0-M_Y}{S_Y}\right)$$

$$-\quad\left[(1-\epsilon)\mathbb{E}_{H_0}\psi\left(\frac{Y-M_Y}{S_Y}\right)+\epsilon\psi\left(\frac{y_0-M_Y}{S_Y}\right)\right]^2.$$

To obtain the maxbias of the Huberized correlation coefficient estimates, we constructed a grid of point mass distributions located at (x_0, y_0) with x_0 and y_0 between -10 and 10 with increments of 0.01. We used numerical integration, namely the trapezoidal method to, compute the Huberized correlation coefficient $r(H_{(x_0,y_0)})$, at each point of the grid.

For different correlation coefficients $\rho = 0.1, 0.5$ and 0.9, the maximum and the minimum values of the Huberized correlation coefficients in the grid for each tuning constant $c = 0, .25, .50, 1.00, 1.25, 1.50, 2.0$ and fraction of contamination $\epsilon = 0.01, 0.05, 0.10, 0.15$, 0.20 are shown in Tables 4.7, 4.10 and 4.13. Labels C and U in the tables stand for "Corrected" values, which are corrected for the intrinsic bias using formula (4.10) for the quadrant correlation coefficients ($c = 0$) and the numerical tables for $c > 0$, and "Uncorrected" values which are not corrected for the intrinsic bias. Then, the maxbias is defined as follows.

$$\max\left\{|\max_{(x_0,y_0)}(r(H_{(x_0,y_0)}))-\rho|,|\min_{(x_0,y_0)}(r(H_{(x_0,y_0)}))-\rho|\right\}.$$

For each ρ, Tables 4.8, 4.11 and 4.14 exhibit the maxbiases of the corrected quadrant and Huberized correlation coefficients, given ϵ and c. Tables 4.9, 4.12 and 4.15 display the maxbiases of the uncorrected quadrant and Huberized correlation coefficients for each ϵ and c. Figures 4.13, 4.14 and 4.15 display part of the results in graphical form.

These pictures show at a glance that some of the maxbiases of the corrected quadrant and Huberized correlation coefficients are larger than the maxbiases of the uncorrected quadrant and Huberized correlation coefficients. An explanation for this is that the worst

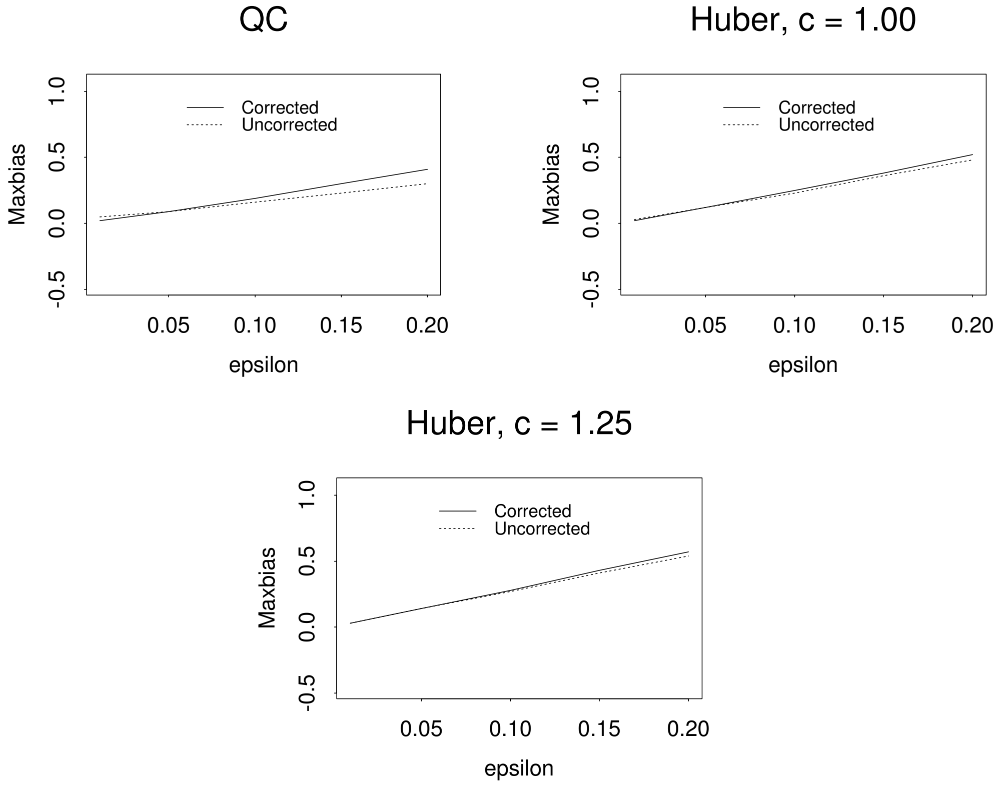


Figure 4.13: Maxbias Comparison of Corrected and Uncorrected Quadrant and Huberized Correlation Coefficient Estimates, $\rho = 0.10$.

contamination bias causes the estimate to become negative and correction for the intrinsic bias makes the estimate even more negative. Therefore, when $\rho > 0$ but the estimate of ρ is negative the maxbias of the corrected quadrant and Huberized correlation coefficients is larger than the maxbias of the uncorrected quadrant and Huberized correlation coefficients. We notice that this phenomenon is more obvious for $\rho = 0.1$ than for $\rho = 0.5$ and 0.9, since for low positive correlations it is more likely that the estimated correlation coefficients will be negative.

The results show that for a fixed value of ϵ the maxbiases of the corrected Huberized correlation coefficients increase as the values of c increase, and therefore the corrected

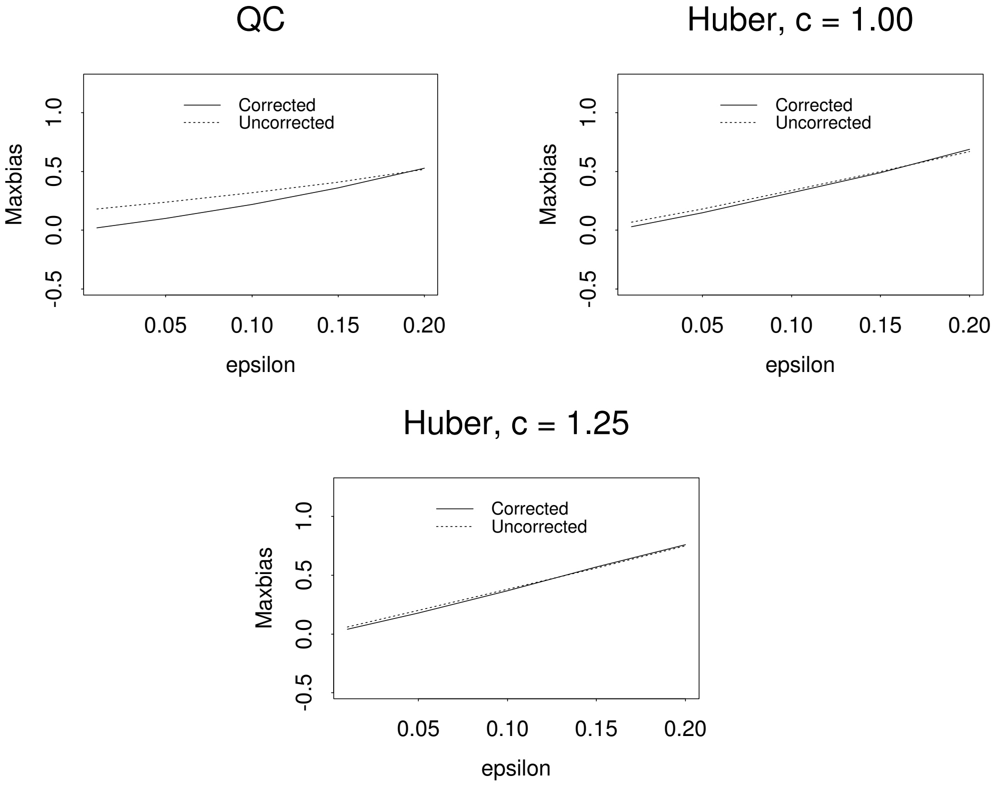


Figure 4.14: Maxbias Comparison of Corrected and Uncorrected Quadrant and Huberized Correlation Coefficient Estimates, $\rho = 0.50$.

quadrant correlation coefficient has the least maxbias. This implies that the contamination bias is an increasing function of c, since the maxbias of the corrected Huberized correlation coefficient will contain only the contamination bias. For different correlation coefficients $\rho = .1, .5$ and $.9$ and $\epsilon = 0.20$, Figure 4.16 displays plots of the maxbiases of the corrected Huberized correlation coefficients versus the tuning constant c.

The results also show that when $\epsilon = 0.01$ and 0.05 the maxbiases of the uncorrected quadrant correlation coefficients are larger than the maxbiases of the uncorrected Huberized correlation coefficients with $c = 1$. An explanation for this occurrence is that for small fractions of contamination the overriding bias in the Huberized correlation coeffi-

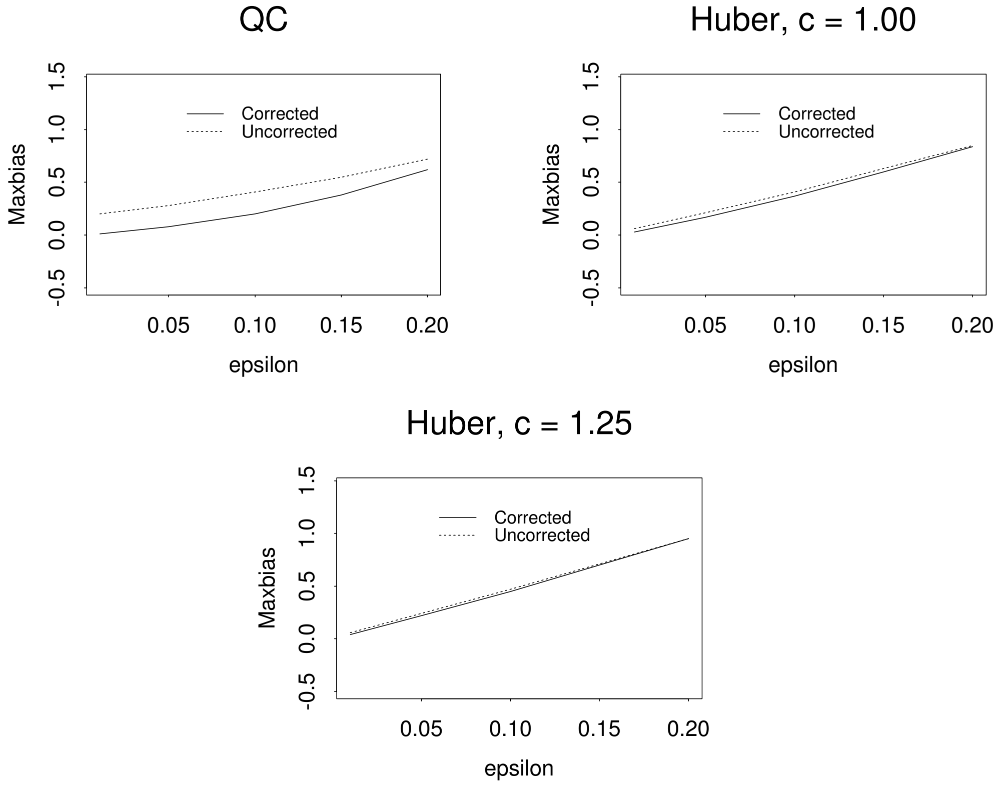


Figure 4.15: Maxbias Comparison of Corrected and Uncorrected Quadrant and Huberized Correlation Coefficient Estimates, $\rho = 0.90$.

cients is the intrinsic bias which is larger for the quadrant correlation coefficients than for the Huberized correlation coefficients with $c = 1$. Figures 4.5 and 4.6 show that the intrinsic bias is a decreasing function of c.

Now we turn our attention to the choice of the tuning constant c in the Huberized correlation coefficients. One important consideration to guide the choice of the tuning constant c is the maxbias over the contamination neighborhoods, which we would like to make as small as possible.

It is important to notice that the intrinsic bias decreases as c increases, whereas the contamination bias increases as c increases. Therefore, in practice it is essential to

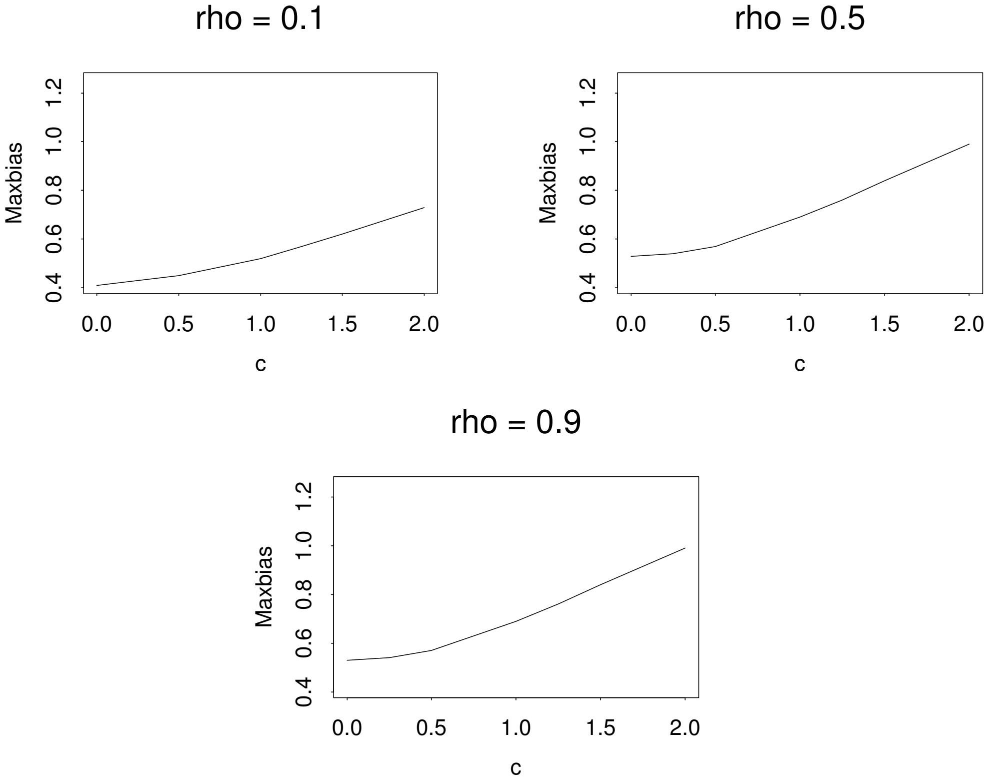


Figure 4.16: Maxbias of Corrected Quadrant and Huberized Correlation Coefficient Estimates, $\epsilon = 0.20$.

choose the tuning constant c to achieve a trade-off between the intrinsic bias and the contamination bias.

The results above suggest that the corrected quadrant correlation coefficient ($c = 0$) has the least maxbias. This implies that we should choose $c = 0$ with correction for the intrinsic bias and thus, as mentioned in Section 4.5, that we should correct the resulting pairwise Huberized scatter matrix for positive definiteness. On the other hand, the results show that the uncorrected Huberized correlation coefficient with $c = 1$ has less maxbias than the uncorrected quadrant correlation coefficient when $\epsilon \leq .05$. We consider the fraction of contamination $\epsilon = .05$ in each variable. Although this value of ϵ might seem small, since each variable is contaminated at this rate, in fact, a large number of the cases will be contaminated in the same way. Therefore, the value of $c = 1$ is a good choice since in this case the correction for the intrinsic bias is not needed and thus the positive definiteness of the resulting pairwise Huberized scatter matrix will be preserved. Figures 4.13, 4.14 and 4.15 show that the maxbiases of the corrected and the uncorrected Huberized correlation coefficients with $c = 1$ are fairly close.

When faced with the question of whether $c = 0$ or $c = 1$ is to be used, the user may have to balance the following issues. Namely, the corrected quadrant correlation coefficient will correct for the intrinsic bias but will yield a scatter matrix which is not necessarily positive definite, while the uncorrected Huberized correlation coefficient with $c = 1$ will provide no correction for the intrinsic bias but will lead to a positive definite scatter matrix. From a computational point of view, the uncorrected Huberized correlation coefficient with $c = 1$ is also to be preferred, because it will be less computationally intense. Also, as seen in Section 4.4.2, that for the choice of $c = 1$ we do not lose much efficiency compared to larger values of c.

Since computational feasibility of the estimate is our important concern, we will focus only on the uncorrected Huberized correlation coefficient (with $c = 1$) in the following section.

4.7.4 Maxbias Comparison of the Correlation Coefficients for Stahel-Donoho and FMCD to Huber

In this section, we compare the maxbias of the uncorrected Huberized correlation coefficient estimates (with $c = 1$) with the maxbias of the Fast MCD (FMCD) and the Stahel-Donoho (SD) correlation coefficient estimates.

We now briefly describe the Stahel-Donoho estimate (see Maronna and Yohai, 1995). Essentially, it is an "outlyingness–weighted" mean and variance, which downweights any point that is many robust standard deviations away from the sample in some univariate projection. The outlyingness measure, r, is based on the idea that if a point is a multivariate outlier then there must be some one-dimensional projection of the data for which the point is a univariate outlier.

Suppose $X = \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ is a multivariate sample where $\boldsymbol{X}_i \in \mathbb{R}^p$, $i = 1, \ldots, n$. The outlyingness r of each point $\boldsymbol{X}_i$ is computed by finding the direction $\boldsymbol{a}_i \in \boldsymbol{A}$, where $\boldsymbol{A} = \{\boldsymbol{a} \in \mathbb{R}^p | \; \|\boldsymbol{a}\| = 1\}$ such that;

$$r(\boldsymbol{X}_i) = \sup_{\boldsymbol{a} \in \boldsymbol{A}} \frac{|\boldsymbol{X}_i'\boldsymbol{a} - \text{med}\{\boldsymbol{X}_j'\boldsymbol{a}\}_{j=1}^n|}{\text{MAD}\{\boldsymbol{X}_j'\boldsymbol{a}\}_{j=1}^n}.$$

The Stahel-Donoho estimate of location and scatter is defined as

$$\hat{\boldsymbol{\mu}} = \frac{\sum_{i=1}^n W_i \boldsymbol{X}_i}{\sum_{i=1}^n W_i},$$

and

$$\widehat{\boldsymbol{\Sigma}} = \frac{\sum_{i=1}^n W_i (\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})'}{\sum_{i=1}^n W_i},$$

with $W_i = W(r(\boldsymbol{X}_i))$.

We used the Splus built-in command in the robust library covRob(stack.dat, estim = “donostah”), in which the weight W_i is computed using the following function of outlyingness;

$$W(r;c) = \begin{cases} 0 & \frac{|r|}{|c|} > 1 \\ a_1 + a_2 \left(\frac{r}{c}\right)^2 + a_3 \left(\frac{r}{c}\right)^4 + a_4 \left(\frac{r}{c}\right)^6 & 0.8 < \frac{|r|}{|c|} \le 1 \\ 1 & \frac{|r|}{|c|} \le 0.8 \end{cases}$$

where $a_1 = -19.71879$, $a_2 = 82.30453$, $a_3 = -105.45267$ and $a_4 = 42.86694$. The tuning constant c is set to be the square-root of the 0.95 quantile of a chi-squared distribution with p degrees of freedom.

To compute the Fast MCD estimate we used the Splus built-in command in the robust library covRob(stack.dat, estim = "MCD"). This implementation uses the Fast MCD algorithm of Rousseeuw and Van Driessen (1999) to approximate the minimum covariance determinant estimate. This algorithm relies on a method called the "C-step" with which, given any approximation to the MCD, it is possible to compute another approximation with a smaller determinant. The Fast MCD algorithm is discussed in Chapter 2.

To obtain the maxbias of the SD and the FMCD correlation coefficient estimates, we constructed a grid of point mass distributions located at (x_0, y_0) with x_0 and y_0 between -2 and 2 with increments of 0.01. At each point of the grid, we generated a bivariate data set from the point mass contamination model (4.15) of sample size $n = 50,000$ for the FMCD estimates and of sample size $n = 5000$ for the SD estimates. The sample size is smaller for the SD estimates due to their computational burden.

For different correlation coefficients $\rho = .1, .5$ and .9, the maximum and the minimum values of the FMCD and the SD correlation coefficient estimates, $\hat{r}$, in the grid for $\epsilon = 0.01, 0.05$, $0.10, 0.15$ and 0.20 are shown in Tables 4.16 and 4.18. Then, the maxbias is defined as.

$$\max\left\{ |\max_{(x_0,y_0)} \hat{r} - \rho|, |\min_{(x_0,y_0)} \hat{r} - \rho| \right\}.$$

Tables 4.17 and 4.19 exhibit the maxbiases of the FMCD and the SD correlation coefficient estimates given ϵ and ρ.

We can now compare the FMCD and the SD maxbiases with the maxbias results of the uncorrected Huberized correlation coefficient estimates (with $c = 1$) from Section 4.7.3. Figure 4.17 shows the maxbiases for each of the estimates plotted versus the fraction of contamination ϵ for three different correlation coefficients, $\rho = .10, .50$ and .90.

From the plots in Figure 4.17, it is evident that for $\rho = .10$ the Huberized approach has the smallest maxbiases for all values of ϵ. The SD approach is better than the FMCD. For $\rho = .50$ the maxbiases of the Huberized approach are almost equal to those of the

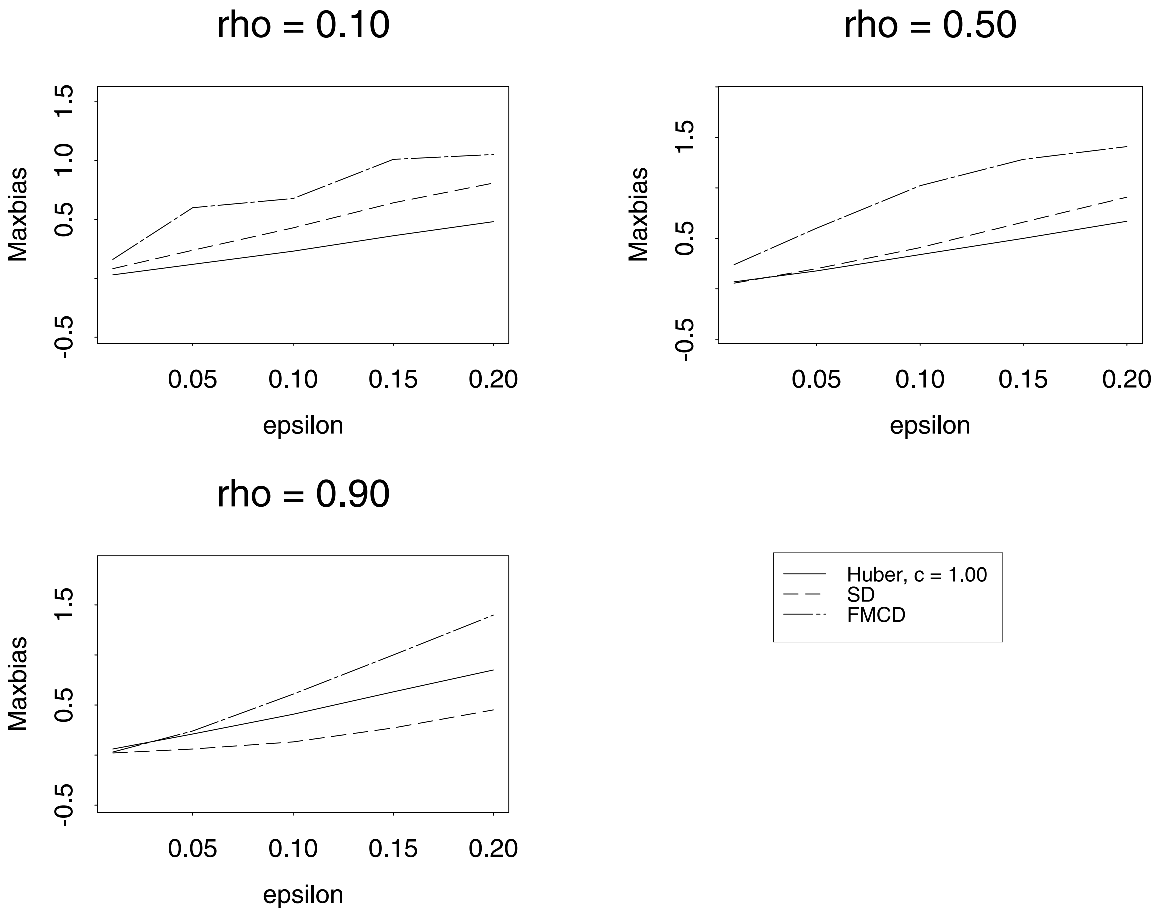


Figure 4.17: Maxbias Comparison of Uncorrected Huberized Correlation Coefficient Estimates ($c = 1$) with SD, FMCD Correlation Coefficient Estimates.

SD approach up to $\epsilon = .05$. For larger values of ϵ, the Huberized approach is better. On the other hand, for $\rho = .90$ the SD approach performs the best, indicating that for structured data the SD approach identifies the outliers very easily. The Huberized approach, however, performs better than the FMCD.

Though the SD approach has the smallest maxbiases for $\rho = .90$, the Huberized approach is a very close competitor. Moreover, the Huberized approach can be very easily coded for computation and is much faster to implement.

ϵ c		0.01		0.05		0.10		0.15		0.20	
		U	C	U	C	U	C	U	C	U	C
0.00	max	0.0734	0.11	0.1127	0.18	0.1662	0.26	0.2288	0.35	0.2974	0.45
	min	0.0529	0.08	0.0067	0.01	-0.0550	-0.09	-0.1260	-0.20	-0.2037	-0.31
0.25	max	0.0821	0.11	0.1282	0.18	0.1914	0.26	0.2611	0.36	0.3395	0.46
	min	0.0588	0.08	0.0067	0.01	-0.0653	-0.09	-0.1463	-0.20	-0.2375	-0.33
0.50	max	0.0915	0.12	0.1453	0.18	0.2182	0.27	0.2981	0.37	0.3871	0.47
	min	0.0642	0.08	0.0032	0.01	-0.0807	-0.10	-0.1740	-0.22	-0.2778	-0.35
1.00	max	0.1090	0.12	0.1834	0.20	0.2814	0.31	0.3837	0.42	0.4904	0.53
	min	0.0700	0.08	-0.0170	-0.02	-0.1325	-0.15	-0.2551	-0.28	-0.3828	-0.42
1.25	max	0.1170	0.13	0.2052	0.22	0.3180	0.34	0.4313	0.45	0.5443	0.57
	min	0.0696	0.08	-0.0352	-0.04	-0.1699	-0.18	-0.3076	-0.33	-0.4444	-0.47
1.50	max	0.1245	0.13	0.2291	0.24	0.3578	0.37	0.4811	0.49	0.5975	0.61
	min	0.0667	0.07	-0.0592	-0.06	-0.2146	-0.22	-0.3656	-0.38	-0.5076	-0.52
2.00	max	0.1395	0.14	0.2843	0.29	0.4438	0.45	0.5793	0.58	0.6931	0.70
	min	0.0536	0.05	-0.1231	-0.12	-0.3174	-0.32	-0.4844	-0.49	-0.6239	-0.63

Table 4.7: Maximum and Minimum Values of Corrected and Uncorrected Quadrant and Huberized Correlation Coefficient Estimates, $\rho = 0.1$.

ϵ	0.01	0.05	0.10	0.15	0.20
$c = 0.00$	0.02	0.09	0.19	0.30	0.41
$c = 0.25$	0.02	0.09	0.19	0.30	0.43
$c = 0.50$	0.02	0.09	0.20	0.32	0.45
$c = 1.00$	0.02	0.12	0.25	0.38	0.52
$c = 1.25$	0.03	0.14	0.28	0.43	0.57
$c = 1.50$	0.03	0.16	0.32	0.48	0.62
$c = 2.00$	0.04	0.22	0.42	0.59	0.73

Table 4.8: Maximum Bias of Corrected Quadrant and Huberized Correlation Coefficient Estimates, $\rho = 0.1$.

ϵ	0.01	0.05	0.10	0.15	0.20
$c = 0.00$	0.05	0.09	0.16	0.23	0.30
$c = 0.25$	0.04	0.09	0.17	0.25	0.34
$c = 0.50$	0.04	0.10	0.18	0.27	0.38
$c = 1.00$	0.03	0.12	0.23	0.36	0.48
$c = 1.25$	0.03	0.14	0.27	0.41	0.54
$c = 1.50$	0.03	0.16	0.31	0.47	0.61
$c = 2.00$	0.05	0.22	0.42	0.58	0.72

Table 4.9: Maximum Bias of Uncorrected Quadrant and Huberized Correlation Coefficient Estimates, $\rho = 0.1$.

c \ ϵ		0.01		0.05		0.10		0.15		0.20	
		U	C	U	C	U	C	U	C	U	C
0.00	max	0.3413	0.51	0.3697	0.55	0.4070	0.60	0.4505	0.65	0.4974	0.70
	min	0.3213	0.48	0.2633	0.40	0.1827	0.28	0.0898	0.14	-0.0171	-0.03
0.25	max	0.3813	0.51	0.4138	0.55	0.4570	0.60	0.5067	0.65	0.5622	0.71
	min	0.3580	0.48	0.2917	0.40	0.1977	0.27	0.0917	0.13	-0.0290	-0.04
0.50	max	0.4188	0.51	0.4555	0.55	0.5045	0.60	0.5591	0.66	0.6190	0.72
	min	0.3915	0.48	0.3129	0.39	0.2029	0.25	0.0804	0.10	-0.0571	-0.07
1.00	max	0.4719	0.51	0.5183	0.56	0.5786	0.62	0.6419	0.68	0.7070	0.74
	min	0.4329	0.47	0.3177	0.35	0.1635	0.18	0.0005	0.01	-0.1704	-0.19
1.25	max	0.4886	0.51	0.5415	0.57	0.6084	0.63	0.6759	0.70	0.7426	0.76
	min	0.4412	0.46	0.3010	0.32	0.1195	0.13	-0.0646	-0.07	-0.2485	-0.26
1.50	max	0.5007	0.51	0.5616	0.58	0.6360	0.65	0.7075	0.72	0.7748	0.79
	min	0.4428	0.46	0.2731	0.28	0.0628	0.07	-0.1398	-0.14	-0.3317	-0.34
2.00	max	0.5169	0.52	0.5988	0.60	0.6884	0.69	0.7650	0.77	0.8295	0.83
	min	0.4310	0.44	0.1913	0.19	-0.0737	-0.07	-0.2989	-0.30	-0.4886	-0.49

Table 4.10: Maximum and Minimum Values of Corrected and Uncorrected Quadrant and Huberized Correlation Coefficient Estimates, $\rho = 0.5$.

ϵ	0.01	0.05	0.10	0.15	0.20
$c = 0.00$	0.02	0.10	0.22	0.36	0.53
$c = 0.25$	0.02	0.10	0.23	0.37	0.54
$c = 0.50$	0.02	0.11	0.25	0.40	0.57
$c = 1.00$	0.03	0.15	0.32	0.49	0.69
$c = 1.25$	0.04	0.18	0.37	0.57	0.76
$c = 1.50$	0.04	0.22	0.43	0.64	0.84
$c = 2.00$	0.06	0.31	0.57	0.80	0.99

Table 4.11: Maximum Bias of Corrected Quadrant and Huberized Correlation Coefficient Estimates, $\rho = 0.5$.

ϵ	0.01	0.05	0.10	0.15	0.20
$c = 0.00$	0.18	0.24	0.32	0.41	0.52
$c = 0.25$	0.12	0.21	0.30	0.41	0.53
$c = 0.50$	0.11	0.19	0.30	0.42	0.56
$c = 1.00$	0.07	0.18	0.34	0.50	0.67
$c = 1.25$	0.06	0.20	0.38	0.56	0.75
$c = 1.50$	0.06	0.23	0.44	0.64	0.83
$c = 2.00$	0.07	0.31	0.57	0.80	0.99

Table 4.12: Maximum Bias of Uncorrected Quadrant and Huberized Correlation Coefficient Estimates, $\rho = 0.5$.

c \ ϵ		0.01		0.05		0.10		0.15		0.20	
		U	C	U	C	U	C	U	C	U	C
0.00	max	0.7161	0.90	0.7290	0.91	0.7444	0.92	0.7626	0.93	0.7823	0.94
	min	0.6957	0.89	0.6166	0.82	0.4950	0.70	0.3484	0.52	0.1793	0.28
0.25	max	0.7933	0.90	0.8067	0.91	0.8243	0.92	0.8437	0.93	0.8642	0.94
	min	0.7698	0.89	0.6790	0.82	0.5424	0.69	0.3817	0.51	0.1966	0.27
0.50	max	0.8406	0.90	0.8532	0.91	0.8693	0.92	0.8866	0.93	0.9043	0.94
	min	0.8131	0.88	0.7068	0.80	0.5530	0.65	0.3790	0.46	0.1843	0.23
1.00	max	0.8811	0.90	0.8928	0.91	0.9077	0.92	0.9228	0.94	0.9374	0.95
	min	0.8420	0.87	0.6909	0.73	0.4881	0.53	0.2744	0.30	0.0516	0.06
1.25	max	0.8900	0.90	0.9024	0.91	0.9176	0.93	0.9325	0.94	0.9467	0.95
	min	0.8427	0.86	0.6612	0.68	0.4267	0.45	0.1893	0.20	-0.0471	-.05
1.50	max	0.8958	0.90	0.9092	0.91	0.9252	0.93	0.9404	0.94	0.9541	0.96
	min	0.8380	0.85	0.6206	0.63	0.3516	0.36	0.0922	0.10	-0.1524	-0.16
2.00	max	0.9021	0.90	0.9190	0.92	0.9373	0.94	0.9529	0.95	0.9656	0.97
	min	0.8164	0.82	0.5121	0.52	0.1766	0.18	-0.1103	-0.11	-0.3508	-0.35

Table 4.13: Maximum and Minimum Values of Corrected and Uncorrected Quadrant and Huberized Correlation Coefficient Estimates, $\rho = 0.9$.

ϵ	0.01	0.05	0.10	0.15	0.20
$c = 0.00$	0.01	0.08	0.20	0.38	0.62
$c = 0.25$	0.01	0.08	0.21	0.39	0.63
$c = 0.50$	0.02	0.10	0.25	0.44	0.67
$c = 1.00$	0.03	0.17	0.37	0.60	0.84
$c = 1.25$	0.04	0.22	0.45	0.70	0.95
$c = 1.50$	0.05	0.27	0.54	0.80	1.06
$c = 2.00$	0.08	0.38	0.72	1.01	1.25

Table 4.14: Maximum Bias of Corrected Quadrant and Huberized Correlation Coefficient Estimates, $\rho = 0.9$.

ϵ	0.01	0.05	0.10	0.15	0.20
$c = 0.00$	0.20	0.28	0.41	0.55	0.72
$c = 0.25$	0.13	0.22	0.36	0.52	0.70
$c = 0.50$	0.09	0.19	0.35	0.52	0.72
$c = 1.00$	0.06	0.21	0.41	0.63	0.85
$c = 1.25$	0.06	0.24	0.47	0.71	0.95
$c = 1.50$	0.06	0.28	0.55	0.81	1.05
$c = 2.00$	0.08	0.39	0.72	1.01	1.25

Table 4.15: Maximum Bias of Uncorrected Quadrant and Huberized Correlation Coefficient Estimates, $\rho = 0.9$.

ρ \ ϵ		0.01	0.05	0.10	0.15	0.20
0.10	max	0.2592	0.6343	0.8475	0.9345	0.9624
	min	-0.0543	-0.5019	-0.7758	-0.9112	-0.9538
0.50	max	0.6169	0.8345	0.9351	0.9688	0.9792
	min	0.3626	-0.0980	-0.5181	-0.7786	-0.9104
0.90	max	0.9278	0.9733	0.9889	0.9936	0.9958
	min	0.8653	0.6554	0.2923	-0.1038	-0.4955

Table 4.16: Maximum and Minimum Values of Fast MCD Correlation Coefficient Estimates.

ϵ	0.01	0.05	0.10	0.15	0.20
$\rho = 0.10$	0.16	0.60	0.68	1.01	1.05
$\rho = 0.50$	0.24	0.60	1.02	1.28	1.41
$\rho = 0.90$	0.03	0.24	0.61	1.00	1.40

Table 4.17: Maximum Bias of Fast MCD Correlation Coefficient Estimates for Different Correlation Coefficients, ρ.

ρ \ ϵ		0.01	0.05	0.10	0.15	0.20
0.10	max	0.1778	0.3374	0.5066	0.6663	0.7588
	min	0.0182	-0.1239	-0.3304	-0.5417	-0.7132
0.50	max	0.5551	0.6512	0.7428	0.8185	0.8679
	min	0.4408	0.3018	0.0862	-0.1633	-0.4134
0.90	max	0.9176	0.9343	0.9527	0.9642	0.9750
	min	0.8827	0.8403	0.7711	0.6372	0.4509

Table 4.18: Maximum and Minimum Values of Stahel-Donoho Correlation Coefficient Estimates.

ϵ	0.01	0.05	0.10	0.15	0.20
$\rho = 0.10$	0.08	0.24	0.43	0.64	0.81
$\rho = 0.50$	0.06	0.20	0.41	0.66	0.91
$\rho = 0.90$	0.02	0.06	0.13	0.27	0.45

Table 4.19: Maximum Bias of Stahel-Donoho Correlation Coefficient Estimates for Different Correlation Coefficients, ρ.

4.8 Application Examples to Real Data

The goal of this section is to illustrate the implementation of the quadrant and Huberized correlation coefficient estimates on three real data sets. For the first two data sets, we implemented the quadrant correlation (QC) coefficient estimates version of the Maronna and Zamar (2002), discussed in Section 4.6. These estimates are used to obtain the robust covariance matrix estimate and outlier detection via robust Mahalanobis distances. The robust Mahalanobis distance,

$$d(\mathbf{x}_i) = (\mathbf{x}_i - \hat{\boldsymbol{\mu}})'\widehat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})$$

is computed using the robust covariance matrix estimate, $\widehat{\boldsymbol{\Sigma}}$, along with the coordinate-wise median, $\hat{\boldsymbol{\mu}}$, as the robust location estimate. In the above expression $\mathbf{x}_i$ is the i-th data vector of dimension p (the transpose of the i-th row of the data). To decide whether or not a matrix row is an outlier, we used the 99-th percentile of the distribution of the maximum of n independent chi-squared random variables with p degrees of freedom. That is, we compare each $d_i(\mathbf{x}_i)$ with the value $d = \chi^2_{.99}(\max)$ given by the equation:

$$P\left(\max_{1\le i\le n} X_i \le d\right) = .99,$$

where $X_1, \ldots, X_n$ are i.i.d. $\chi^2(p)$.

4.8.1 Glass Data

The data set *glass* is a small 214×10 matrix consisting of 9 numeric variables and one categorical variable. The 9 numeric variables are the percentages of various chemical constituents of the glass. We obtained these data from the new Insightful Miner (I-Miner) data mining product. We computed the QC based robust covariance matrix and robust Mahalanobis distances for the sub-matrix consisting of the first five columns of the above matrix.

Upon running the outlier detection computations with threshold point $\chi^2_{.99}(\max) = 27.43$, we found that approximately 34% of the data points are outliers while the remain-

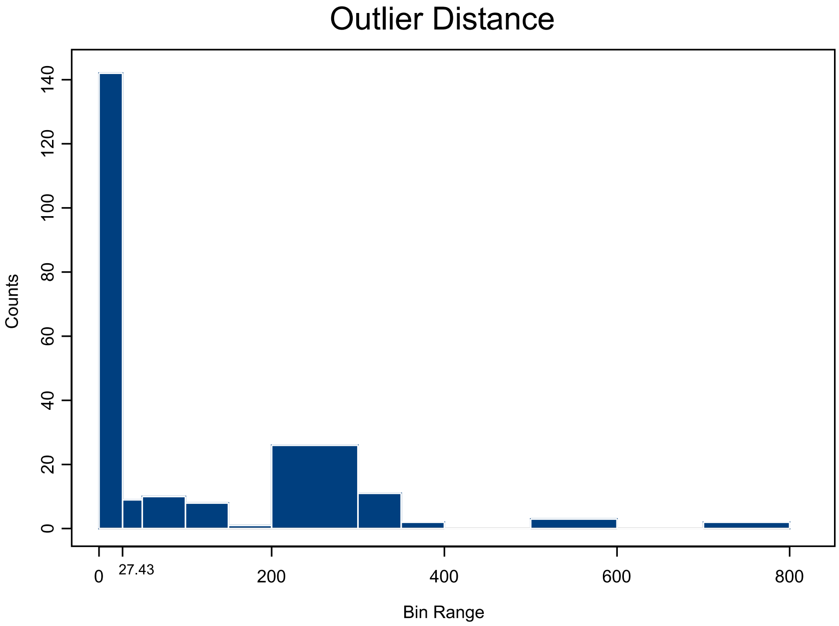


Figure 4.18: Histogram of the Robust Mahalanobis Distances for the Glass Data.

ing 66% of the data represents a central core. The histogram of the robust distances in Figure 4.18 clearly shows a cluster of large distances around 200 to 350 which are much larger than the $\chi^2_{.99}$(max) with 5 degrees of freedom threshold of 27.43 used above.

A visualization of the data by means of all pairwise scatter plots in Figure 4.19 reveals an interesting aspect of the multivariate structure that is reasonably consistent with these observations. One sees that the data appears to have a central core that is roughly elliptical in the pairwise views, along with broadly scattered outliers and the distinctive rod-like structure. The latter is due to the fact that 41 of the observations of the *Mg* variable have value zero. This was evidently because the data values were not recorded or were misplaced, and zero values were substituted for the missing values.

Inspection of the scatterplots in Figure 4.19 reveals that there are roughly an additional 31 diffuse outliers well separated from the elliptical core. So what the outlier detection algorithm with a $\chi^2_{.99}$(max) threshold of 27.43 does is identify the diffuse outliers as well as the extreme outlying rod caused by the zero *Mg*'s. In other words, the

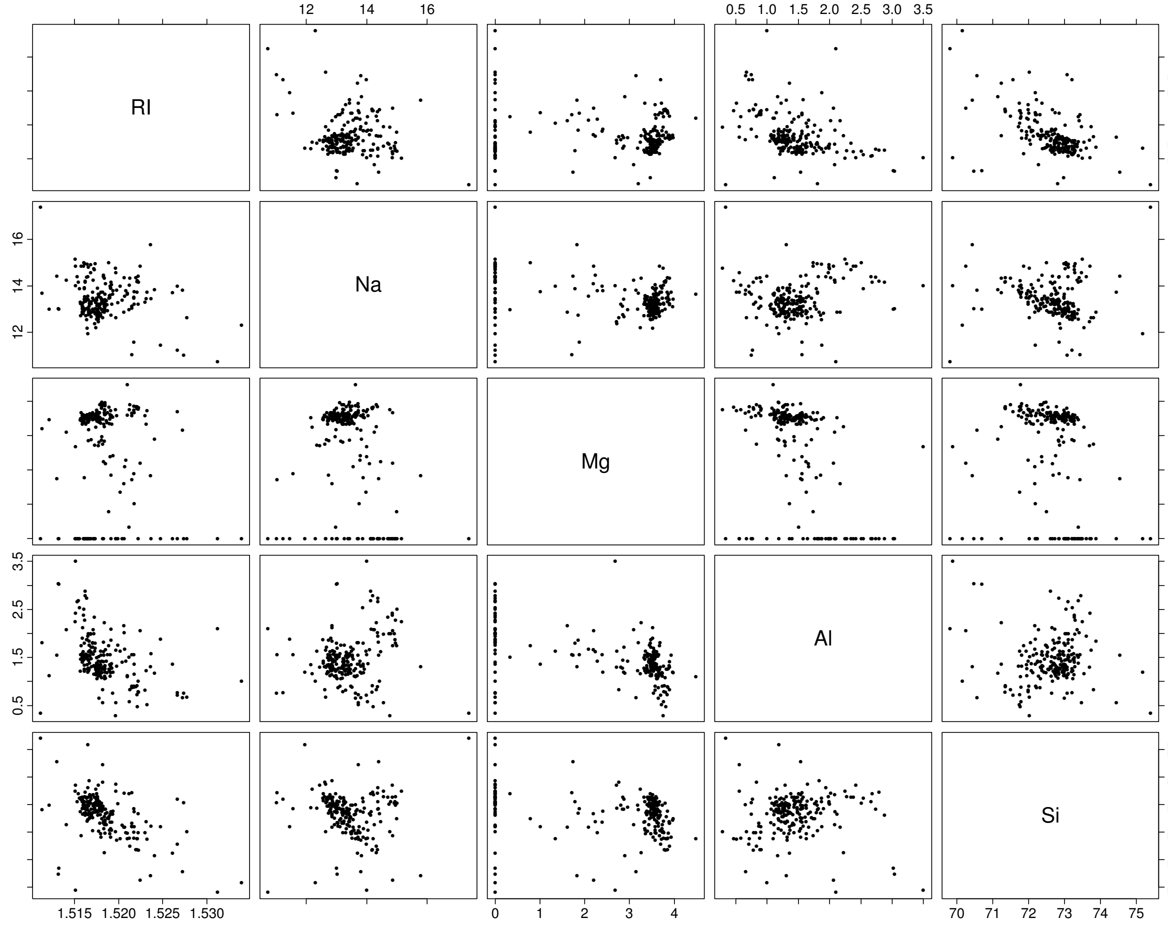


Figure 4.19: Pairwise Plots for the Data Set Glass.

outlier detection algorithm is behaving quite as anticipated.

What the histogram is identifying with its bimodal character is the separation of the pure rod outlier as the most extreme set of distances, distances that are well beyond those of the diffuse outliers closer to central bulk of the data. If we use a threshold of 200 to set aside outliers we will set aside the pure rod as shown in Figure 4.20, and this is not an unreasonable first step. In a second step we will find the remainder of the diffuse outliers.

These observations suggest that one might well use robust covariance matrix based

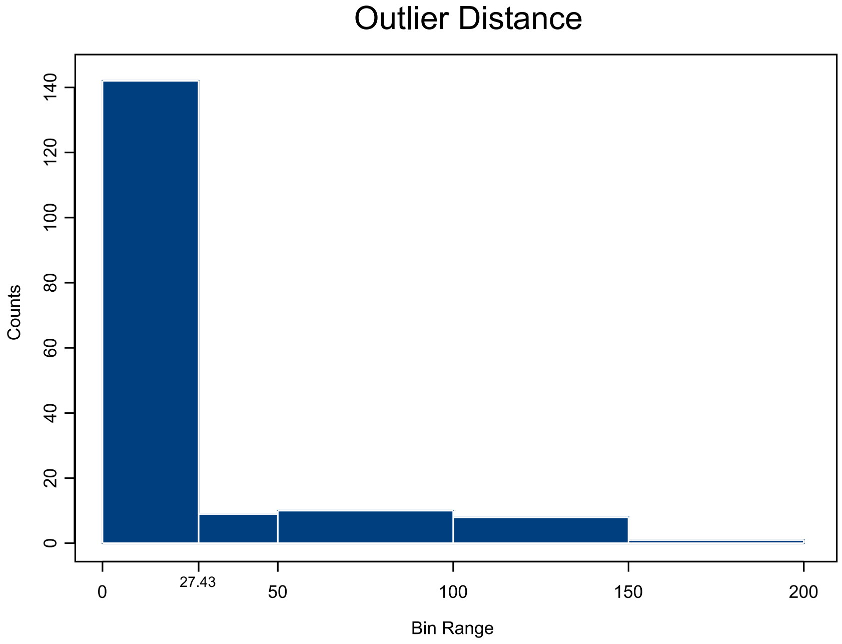


Figure 4.20: Histogram of the Robust Mahalanobis Distances using a Threshold of 200 for the Glass Data.

robust distances to iteratively cluster multivariate data by iterative removal of outlier groups, monitored by histograms or density estimates of the robust distances, and subsequent iteration on the sub-clusters. This possibility bears further investigation.

4.8.2 KDD-CUP-98 PVA Donations Data

This data set was used for the second International Knowledge Discovery and Data Mining tools competition, which was held in conjunction with KDD-98 the fourth International Conference on Knowledge Discovery and Data Mining. The competition task was a regression problem where the goal is to estimate the return from a direct mailing in order to maximize donation profits.

This data set, which we will refer to as the “PVA” data, represents a much more substantial data mining challenge. The original KDD-CUP-98 PVA data set consists of $95,412$ records (rows) and 481 variables (columns). For purposes of this example, we

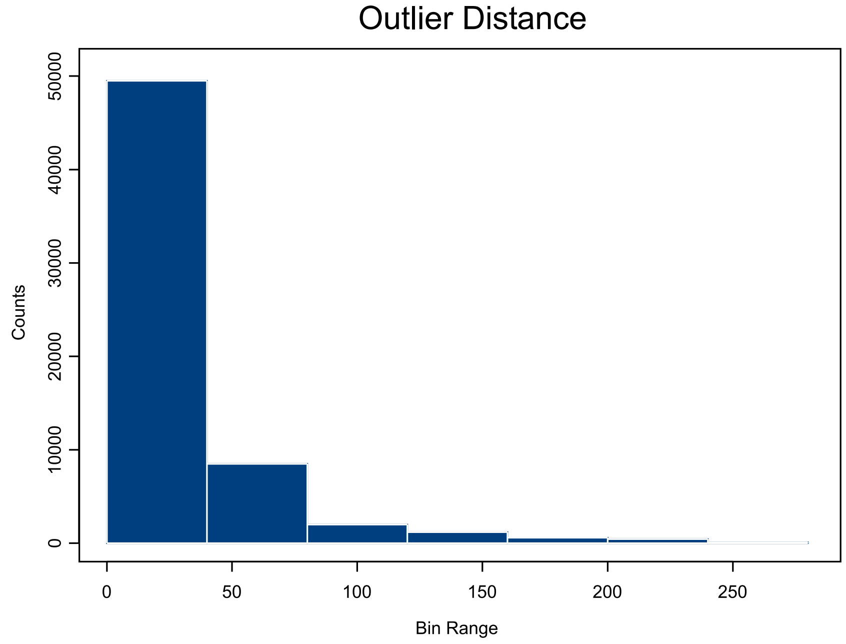


Figure 4.21: Histogram of the Robust Mahalanobis Distances for the PVA Data.

have used 16 of the numeric variables. We computed the QC based robust covariance matrix and robust Mahalanobis distances for the sub-matrix. Upon running the outlier detection computations with threshold point $\chi^2_{.99}(\max) = 64.1$, we found that 17,903 outliers rows, 44,284 non-outlier rows and 33,225 NA rows (missing data).

The histogram of the robust distances is shown in Figure 4.21 (in which we have filtered out a few very extreme outlier distances for purposes of a more detailed display). In Figure 4.22 we show the plot of ordered absolute differences between the classical and robust correlation coefficients obtained from the robust correlation matrix. Figure 4.22 shows that while the vast majority of the absolute differences between the classical and robust correlation coefficients are less than .05, a few differences are fairly large (three are larger than .2 and ten are larger than .1).

In order to more fully test the capabilities of the robust outlier detection method, we modified a subset of the PVA data as follows. We took a subset of 10,000 records from the PVA data set. Then we added 1,000 rows – each identical to the second row except

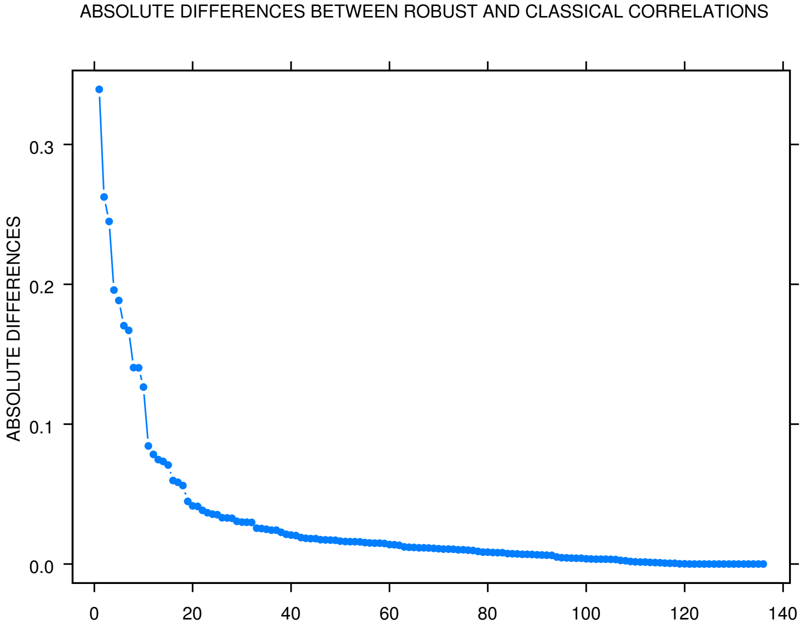


Figure 4.22: Differences between Classical and Robust Correlation Coefficients for the PVA Data.

that the value for the variable “minramnt” was changed to 1 and the value of the variable “avggift” was changed to 50. While this does not result in very extreme outliers, it does result in outliers that are well detached from the bulk of the data. The results for this modified data set, upon running the outlier detection computations, are 3618 outlier rows and 7382 non-outlier rows. The histogram of the robust distances is shown in Figure 4.23 (in which we have filtered out a few very extreme outlier distances for purposes of a more detailed display). Figure 4.24 shows the plot of ordered absolute differences between the classical and robust correlation coefficients obtained from the robust correlation matrix.

In this case the outliers show up as a clear bump in the histogram bar located near 225. This suggests further investigation of the data by deleting all outliers with robust distances greater than 175–200. The overall shape in Figure 4.24 is similar to that of Figure 4.22, except now the largest difference is .5 rather than .35, and several more in the .2–.3 range, reflecting the impact of the added outliers.

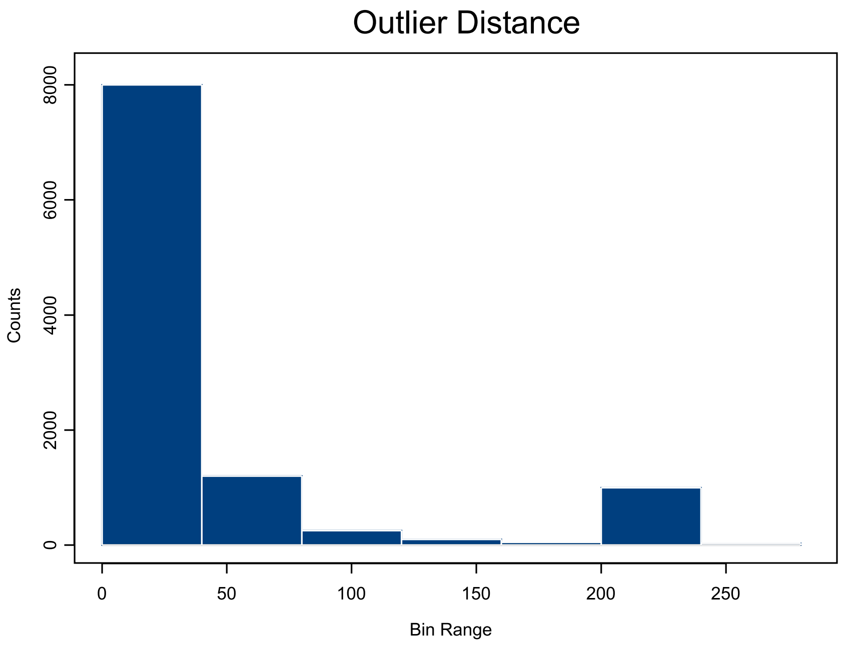


Figure 4.23: Histogram of the Robust Mahalanobis Distances for the Modified PVA Data with Outliers.

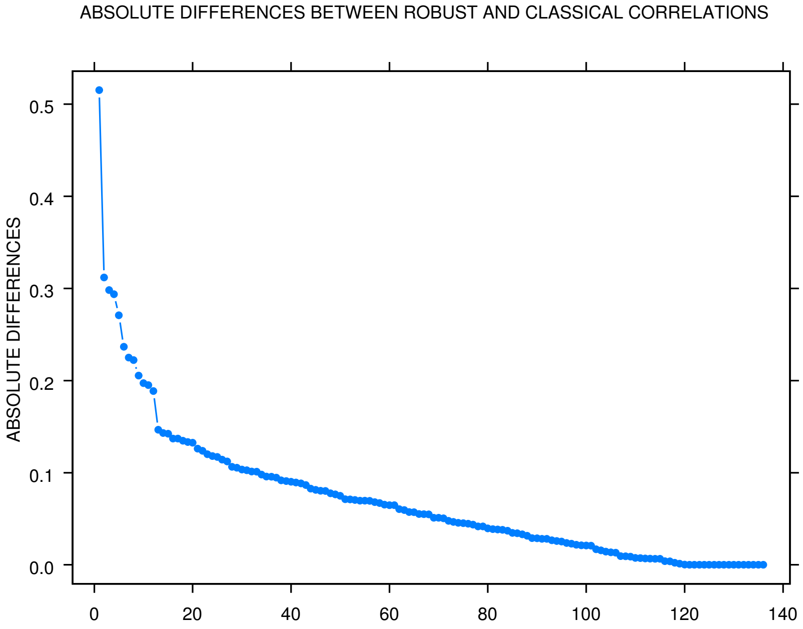


Figure 4.24: Differences between Classical and Robust Correlation Coefficients for the Modified PVA Data.

4.8.3 Daily Pressure in Northern Hemisphere Data

In this example, two sets of data were considered for the analysis. We obtained these data from Dr. Pandolfo (Department of Oceanography, University of British Columbia). The first data set has the daily data of sea-level pressure (SLP) in boreal winter (December, January and February) for 41 years (from December 1958 to February 1999) in Northern Hemisphere from 20^oN to 85^oN partitioned with $2.5^o \times 2.5^o$ grid lines. We have 27 $(= \frac{85^o - 20^o}{2.5^o} + 1)$ latitudinal grid lines and 145 $(= \frac{360^o}{2.5^o} + 1)$ longitudinal grid lines. At each of the 3915 $(= 145 \times 27)$ grid points, the SLP values are available for each of the 3690 $(= 90 \times 41)$ days. February 29 is not considered for the leap year. Thus, the number of observations is $n = 3690$ and the number of variables is $p = 3915$ for the first data set.

The second set of the data has the daily pressure values at 500 hPa geopotential heights in boreal winter for 40 years (from December 1958 to February 1998) for the same grid points as above. The number of days in this case is 3600 $(= 90 \times 40)$. Thus, the number of observation is $n = 3600$ and the number of variables is $p = 3915$ for the second data set. Though 0^o longitude and 360^o longitude are the same line, they are considered different to present the world on a flat page. This is why the number of longitudinal grid lines is 145 (instead of 144).

The results we obtained are almost similar for the two different data sets (because the data sets are similar in nature). Therefore, we discuss only the results for the sea-level pressure data set.

To compare the classical covariance estimate with the robust pairwise Huberized covariance estimate, we implemented principal component analysis using the classical covariance estimate and the pairwise Huberized covariance estimate with $c = 1$. We calculated the proportion of total variance due to each principal component using both estimates. The first 20 of these proportions are presented in the first and second columns of Table 4.20. The values in the two columns are close which indicates that the data do not contain outliers. This also indicates that the pairwise Huberized covariance estimate

Clean Data		1% 3SD		1% 6SD		2% 3SD		2% 6SD	
Classical	Robust	Classical	Robust	Classical	Robust	Classical	Robust	Classical	Robust
10.440	10.045	10.045	10.011	9.613	10.011	9.691	9.978	8.925	9.978
8.227	8.369	7.920	8.340	7.584	8.340	7.641	8.308	7.047	8.308
7.329	7.289	7.047	7.263	6.746	7.263	6.790	7.232	6.253	7.232
6.957	7.060	6.704	7.036	6.422	7.036	6.454	7.010	5.946	7.010
5.654	5.720	5.446	5.697	5.216	5.697	5.256	5.677	4.844	5.677
5.312	5.319	5.109	5.299	4.892	5.299	4.931	5.280	4.547	5.280
4.506	4.358	4.341	4.342	4.160	4.342	4.179	4.324	3.853	4.324
3.591	3.555	3.463	3.544	3.318	3.544	3.339	3.531	3.080	3.531
3.484	3.363	3.353	3.351	3.211	3.351	3.234	3.340	2.982	3.340
2.6487	2.605	2.551	2.596	2.444	2.596	2.462	2.587	2.270	2.587
2.449	2.377	2.357	2.367	2.259	2.367	2.274	2.359	2.098	2.359
2.390	2.329	2.301	2.321	2.204	2.321	2.224	2.313	2.053	2.313
2.018	1.984	1.943	1.978	1.862	1.978	1.874	1.827	1.728	1.827
1.488	1.840	1.782	1.834	1.708	1.834	1.722	1.827	1.590	1.827
1.737	1.663	1.674	1.657	1.604	1.657	1.614	1.651	1.489	1.651
1.637	1.590	1.577	1.585	1.511	1.585	1.524	1.580	1.407	1.580
1.532	1.424	1.478	1.419	1.418	1.419	1.427	1.414	1.318	1.414
1.430	1.394	1.379	1.391	1.321	1.391	1.335	1.386	1.234	1.386
1.403	1.356	1.355	1.352	1.300	1.352	1.305	1.348	1.205	1.348
1.231	1.173	1.186	1.170	1.136	1.170	1.147	1.166	1.060	1.166

Table 4.20: First 20 Proportions of Variation for the Classical and the Pairwise Huberized (with $c = 1$) Covariance Estimates.

works as well as the classical estimate for clean data (data without outliers).

To investigate how the classical and the pairwise Huberized covariance estimates behave in a large data set with outliers, we decided to contaminate 10% of the variables in the data set. We used four different levels of contamination. For each of these 10% variables, first 1% and then 2% of the observations are selected randomly for contamination, and they are replaced by randomly generated values from each of the following two distributions:

- $N(\max + 3\text{SD}, 0.5\text{SD})$
- $N(\max + 6\text{SD}, 0.5\text{SD})$

where max is the maximum observation and SD is the standard deviation of the variables being contaminated. Thus, the four levels of contamination can be described as follows:

- 1%, 3SD
- 1%, 6SD
- 2%, 3SD
- 2%, 6SD

Both the classical and the pairwise Huberized covariance estimates are used for each of these contaminations and the results are presented in Table 4.20. From the table we see that the proportions based on the classical covariance estimate change with increased contamination. However, the proportions based on the pairwise Huberized covariance estimate are hardly affected by contamination. To make the comparison more visible the first six proportions for both clean and contaminated data are plotted in Figure 4.25 for the classical covariance estimates and in Figure 4.26 for the pairwise Huberized covariance estimates. The plots clearly indicate that the classical covariance estimate is very much affected by outliers while the pairwise Huberized covariance estimate is more resistant to the outliers. In the classical approach, if we increase the percentage of contamination we get less favorable results. Also, if we use 6SD contamination instead of 3SD the results further deteriorate. For the pairwise Huberized covariance estimates, if we increase the percentage of contamination, the results get a little bit worse. However, if we use 6SD contamination instead of 3SD we get the same results because of the definition of the Huber function.

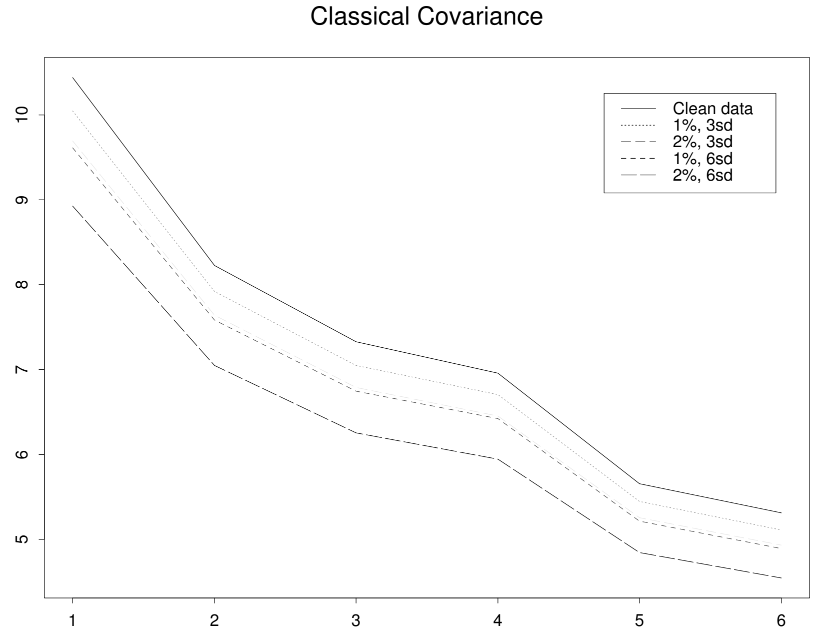


Figure 4.25: Proportions of Variation for the Classical Covariance Estimates.

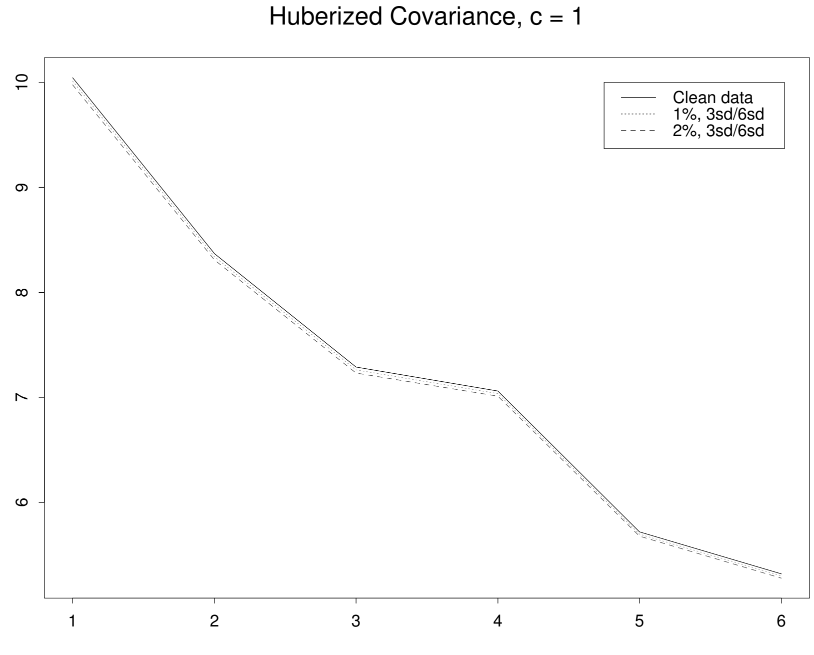


Figure 4.26: Proportions of Variation for the Pairwise Huberized (with $c = 1$) Covariance Estimates.

For large data sets it is difficult or sometimes impossible to identify any outliers. In such cases we can apply both the classical and the Huberized approach. If the results are the same, we will understand that there are no outliers, and if the results are different we should rely on the results of the Huberized approach because of the observations presented above.

4.9 Chapter Appendix

4.9.1 Proof of Theorem 4.1

In this section, we prove Theorem 4.1 that shows under certain regularity conditions the Huberized correlation coefficient estimates are consistent.

STEP I By applying the Taylor expansion for the function $f(x,y) = \frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{X_i - x}{y}\right)$ about the point (μ_X, σ_X) we get

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right) = \frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{X_i-\mu_X}{\sigma_X}\right) \\ - \frac{(\hat{\mu}_X-\mu_X)}{n\tilde{\sigma}_X}\sum_{i=1}^{n}\psi'\left(\frac{X_i-\tilde{\mu}_X}{\tilde{\sigma}_X}\right) \\ - \frac{(\hat{\sigma}_X-\sigma_X)}{n\tilde{\sigma}_X}\sum_{i=1}^{n}\psi'\left(\frac{X_i-\tilde{\mu}_X}{\tilde{\sigma}_X}\right)\left(\frac{X_i-\tilde{\mu}_X}{\tilde{\sigma}_X}\right) \end{aligned} \tag{4.20}$$

for some $(\tilde{\mu}_X, \tilde{\sigma}_X)$ between (μ_X, σ_X) and $(\hat{\mu}_X, \hat{\sigma}_X)$.

STEP II We show that the second and the third terms on the right hand side of equation (4.20) tend to zero as $n \to \infty$. Since ψ' is bounded $|\psi'| \leq M$, for some $M > 0$. So

$$\left|\frac{(\hat{\mu}_X-\mu_X)}{n\tilde{\sigma}_X}\sum_{i=1}^{n}\psi'\left(\frac{X_i-\tilde{\mu}_X}{\tilde{\sigma}_X}\right)\right| \leq M\left|\frac{(\hat{\mu}_X-\mu_X)}{\tilde{\sigma}_X}\right|$$

and the right hand side tends to zero, since $(\hat{\mu}_X - \mu_X) \to 0$ as $n \to \infty$. And for the third term, since $\psi'(X)X$ is bounded $|\psi'(X)X| \leq M$, for some $M > 0$. So

$$\left|\frac{(\hat{\sigma}_X-\sigma_X)}{n\tilde{\sigma}_X}\sum_{i=1}^{n}\psi'\left(\frac{X_i-\tilde{\mu}_X}{\tilde{\sigma}_X}\right)\left(\frac{X_i-\tilde{\mu}_X}{\tilde{\sigma}_X}\right)\right| \leq M\left|\frac{(\hat{\sigma}_X-\sigma_X)}{\tilde{\sigma}_X}\right|$$

and the right hand side tends to zero, since $(\hat{\sigma}_X - \sigma_X) \to 0$ as $n \to \infty$.

STEP III Since the variables $\psi\left(\frac{X_i - \mu_X}{\sigma_X}\right)$, $i = 1, \ldots, n$, are i.i.d., then by the strong law of large numbers

$$\frac{1}{n}\sum_{i=1}^{n} \psi\left(\frac{X_i - \mu_X}{\sigma_X}\right) \longrightarrow \mathbb{E}\left\{\psi\left(\frac{X - \mu_X}{\sigma_X}\right)\right\} \qquad \text{a.s.}$$

as $n \to \infty$. From the preceding steps

$$\frac{1}{n}\sum_{i=1}^{n} \psi\left(\frac{X_i - \hat{\mu}_X}{\hat{\sigma}_X}\right) \longrightarrow \mathbb{E}\left\{\psi\left(\frac{X - \mu_X}{\sigma_X}\right)\right\} \qquad \text{a.s.}$$

as $n \to \infty$.

STEP IV Let us assume that $\epsilon_i = \pm 1$. Then by writing the Taylor expansion for $f(x, y)$ $= \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \psi\left(\frac{X_i - x}{y}\right)$ about the point (μ_X, σ_X) we get

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n} \epsilon_i \left(\psi\left(\frac{X_i - \hat{\mu}_X}{\hat{\sigma}_X}\right) - \psi\left(\frac{X_i - \mu_X}{\sigma_X}\right)\right) \\ = \quad & -\frac{(\hat{\mu}_X - \mu_X)}{n\tilde{\sigma}_X}\sum_{i=1}^{n} \epsilon_i \psi'\left(\frac{X_i - \tilde{\mu}_X}{\tilde{\sigma}_X}\right) \\ & -\frac{(\hat{\sigma}_X - \sigma_X)}{n\tilde{\sigma}_X}\sum_{i=1}^{n} \epsilon_i \psi'\left(\frac{X_i - \tilde{\mu}_X}{\tilde{\sigma}_X}\right)\left(\frac{X_i - \tilde{\mu}_X}{\tilde{\sigma}_X}\right) \end{aligned}$$

for some $(\tilde{\mu}_X, \tilde{\sigma}_X)$ between (μ_X, σ_X) and $(\hat{\mu}_X, \hat{\sigma}_X)$. Now apply the same method as in Step II to show that the two terms on the right hand side of the above equality tend to zero. So

$$\frac{1}{n}\sum_{i=1}^{n} \epsilon_i \left(\psi\left(\frac{X_i - \hat{\mu}_X}{\hat{\sigma}_X}\right) - \psi\left(\frac{X_i - \mu_X}{\sigma_X}\right)\right) \longrightarrow 0 \qquad \text{a.s.}$$

as $n \to \infty$. From here it is clear that

$$\frac{1}{n}\sum_{i=1}^{n} \left|\psi\left(\frac{X_i - \hat{\mu}_X}{\hat{\sigma}_X}\right) - \psi\left(\frac{X_i - \mu_X}{\sigma_X}\right)\right| \longrightarrow 0 \qquad \text{a.s.}$$

as $n \to \infty$.

Step V Set

$$Z = \frac{1}{n}\sum_{i=1}^{n} \psi\left(\frac{X_i - \hat{\mu}_X}{\hat{\sigma}_X}\right).$$

Then

$$\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n}\left(\psi\left(\frac{X_i - \hat{\mu}_X}{\hat{\sigma}_X}\right) - Z\right)^2 \\
= & \frac{1}{n}\sum_{i=1}^{n}\left(\psi\left(\frac{X_i - \mu_X}{\sigma_X}\right) - Z\right)^2 \\
& + \frac{1}{n}\sum_{i=1}^{n}\left(\psi\left(\frac{X_i - \hat{\mu}_X}{\hat{\sigma}_X}\right) - \psi\left(\frac{X_i - \mu_X}{\sigma_X}\right)\right)^2 \\
& + \frac{2}{n}\sum_{i=1}^{n}\left(\psi\left(\frac{X_i - \mu_X}{\sigma_X}\right) - Z\right)\left(\psi\left(\frac{X_i - \hat{\mu}_X}{\hat{\sigma}_X}\right) - \psi\left(\frac{X_i - \mu_X}{\sigma_X}\right)\right).
\end{aligned}$$

Now we show that the second term on the right hand side in the above equality tends to zero. In fact ψ is bounded, so $|\psi| \leq M$ then

$$\begin{aligned}
& \left|\frac{1}{n}\sum_{i=1}^{n}\left(\psi\left(\frac{X_i - \hat{\mu}_X}{\hat{\sigma}_X}\right) - \psi\left(\frac{X_i - \mu_X}{\sigma_X}\right)\right)^2\right| \\
\leq & \frac{2M}{n}\sum_{i=1}^{n}\left|\psi\left(\frac{X_i - \hat{\mu}_X}{\hat{\sigma}_X}\right) - \psi\left(\frac{X_i - \mu_X}{\sigma_X}\right)\right|
\end{aligned}$$

and the right hand side tends to zero by Step IV. So $\frac{1}{n}\sum_{i=1}^{n}\left(\psi\left(\frac{X_i - \hat{\mu}_X}{\hat{\sigma}_X}\right) - \psi\left(\frac{X_i - \mu_X}{\sigma_X}\right)\right)^2$ tends to zero. On the other hand, each of the terms $\psi\left(\frac{X_i - \hat{\mu}_X}{\hat{\sigma}_X}\right)$ is bounded by M so their average is bounded by M that is $|Z| \leq M$. So

$$\begin{aligned}
& \left|\frac{2}{n}\sum_{i=1}^{n}\left(\psi\left(\frac{X_i - \mu_X}{\sigma_X}\right) - Z\right)\left(\psi\left(\frac{X_i - \hat{\mu}_X}{\hat{\sigma}_X}\right) - \psi\left(\frac{X_i - \mu_X}{\sigma_X}\right)\right)\right| \\
\leq & \frac{4M}{n}\sum_{i=1}^{n}\left|\psi\left(\frac{X_i - \hat{\mu}_X}{\hat{\sigma}_X}\right) - \psi\left(\frac{X_i - \mu_X}{\sigma_X}\right)\right|
\end{aligned}$$

and by Step IV the right hand side tends to zero. So

$$\frac{2}{n}\sum_{i=1}^{n}\left(\psi\left(\frac{X_i-\mu_X}{\sigma_X}\right)-Z\right)\left(\psi\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)-\psi\left(\frac{X_i-\mu_X}{\sigma_X}\right)\right)$$

tends to zero. But we already know that

$$\frac{1}{n}\sum_{i=1}^{n}\left(\psi\left(\frac{X_i-\mu_X}{\sigma_X}\right)-Z\right)^2 \longrightarrow \operatorname{Var}\left(\psi\left(\frac{X-\mu_X}{\sigma_X}\right)\right) \qquad \text{a.s.}$$

as $n \to \infty$. Then,

$$\frac{1}{n}\sum_{i=1}^{n}\left(\psi\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)-Z\right)^2 \longrightarrow \operatorname{Var}\left(\psi\left(\frac{X-\mu_X}{\sigma_X}\right)\right) \qquad \text{a.s.}$$

as $n \to \infty$.

Step VI Similarly, we show that,

$$\frac{1}{n}\sum_{i=1}^{n}\left(\psi\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)-W\right)^2 \longrightarrow \operatorname{Var}\left(\psi\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right) \qquad \text{a.s.}$$

as $n \to \infty$, where

$$W=\frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right).$$

Step VII We have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left(\psi\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)+\psi\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)-(Z+W)\right)^2 \\
= \;&\frac{1}{n}\sum_{i=1}^{n}\left(\psi\left(\frac{X_i-\mu_X}{\sigma_X}\right)+\psi\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)-(Z+W)\right)^2 \\
+ \;&\frac{1}{n}\sum_{i=1}^{n}\left(\psi\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)+\psi\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)-\psi\left(\frac{X_i-\mu_X}{\sigma_X}\right)-\psi\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)\right)^2 \\
+ \;&\frac{2}{n}\sum_{i=1}^{n}\left(\psi\left(\frac{X_i-\mu_X}{\sigma_X}\right)+\psi\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)-(Z+W)\right) \\
&\left(\psi\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)+\psi\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)-\psi\left(\frac{X_i-\mu_X}{\sigma_X}\right)-\psi\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)\right)
\end{aligned}$$

Now we show that the second term on the right hand side of the above equality tends to zero

$$\left|\frac{1}{n}\sum_{i=1}^{n}\left(\psi\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)+\psi\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\right.\right.$$

$$\left.\left.-\psi\left(\frac{X_i-\mu_X}{\sigma_X}\right)-\psi\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)\right)^2\right|$$

$$\leq \frac{4M}{n}\left(\sum_{i=1}^{n}\left|\psi\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)-\psi\left(\frac{X_i-\mu_X}{\sigma_X}\right)\right|\right.$$

$$\left.+\sum_{i=1}^{n}\left|\psi\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)-\psi\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)\right|\right)$$

and the right hand side tends to zero. Now we show that the third term on the right hand side of the above equality tends to zero.

$$\left|\frac{2}{n}\sum_{i=1}^{n}\left(\psi\left(\frac{X_i-\mu_X}{\sigma_X}\right)+\psi\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)-(Z+W)\right)\right.$$

$$\left.\left(\psi\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)+\psi\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)-\psi\left(\frac{X_i-\mu_X}{\sigma_X}\right)-\psi\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)\right)\right|$$

$$\leq \frac{8M}{n}\left(\sum_{i=1}^{n}\left|\psi\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)-\psi\left(\frac{X_i-\mu_X}{\sigma_X}\right)\right|\right.$$

$$\left.+\sum_{i=1}^{n}\left|\psi\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)-\psi\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)\right|\right)$$

and the right hand side tends to zero. Now we already know that

$$\frac{1}{n}\sum_{i=1}^{n}\left(\psi\left(\frac{X_i-\mu_X}{\sigma_X}\right)+\psi\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)-(Z+W)\right)^2$$

$$\longrightarrow \operatorname{Var}\left(\psi\left(\frac{X-\mu_X}{\sigma_X}\right)+\psi\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right) \qquad \text{a.s.}$$

as $n \to \infty$. Then,

$$\frac{1}{n}\sum_{i=1}^{n}\left(\psi\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)+\psi\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)-(Z+W)\right)^2$$

$$\longrightarrow \operatorname{Var}\left(\psi\left(\frac{X-\mu_X}{\sigma_X}\right)+\psi\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right) \qquad \text{a.s.}$$

as $n \to \infty$.

Step VIII Now we can write

$$\frac{1}{n}\sum_{i=1}^{n}\left(\psi\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)-Z\right)\left(\psi\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)-W\right)$$

$$= \frac{1}{2n}\left[\sum_{i=1}^{n}\left(\psi\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)+\psi\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)-(Z+W)\right)^2\right.$$

$$\left.-\sum_{i=1}^{n}\left(\psi\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)-Z\right)^2-\sum_{i=1}^{n}\left(\psi\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)-W\right)^2\right]$$

and from the preceding steps, the right hand side of the above equality tends almost surely to

$$\frac{1}{2}\left(\operatorname{Var}\left(\psi\left(\frac{X-\mu_X}{\sigma_X}\right)+\psi\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right)-\operatorname{Var}\left(\psi\left(\frac{X-\mu_X}{\sigma_X}\right)\right)-\operatorname{Var}\left(\psi\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right)\right)$$

as $n \to \infty$.

Step IX Now the Huberized correlation coefficient estimate

$$\hat{r} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(\psi\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)-Z\right)\left(\psi\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)-W\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\psi\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)-Z\right)^2\frac{1}{n}\sum_{i=1}^{n}\left(\psi\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)-W\right)^2}}$$

tends almost surely to

$$r = \frac{\operatorname{Cov}\left(\psi\left(\frac{X-\mu_X}{\sigma_X}\right),\psi\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right)}{\sqrt{\operatorname{Var}\left(\psi\left(\frac{X-\mu_X}{\sigma_X}\right)\right)\operatorname{Var}\left(\psi\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right)}}.$$

as $n \to \infty$. ■

4.9.2 Proof of Theorem 4.2

Here we provide the proof of Theorem 4.2 that gives the asymptotic normality of the Huberized correlation coefficient estimates.

It can be shown that

$$\begin{array}{rclc}\hat{\mu}_X & \longrightarrow & \mu_X & \text{a.s.}\\ \hat{\mu}_Y & \longrightarrow & \mu_Y & \text{a.s.}\\ \hat{\sigma}_X & \longrightarrow & \sigma_X & \text{a.s.}\\ \hat{\sigma}_Y & \longrightarrow & \sigma_Y & \text{a.s.}\end{array}$$

as $n \to \infty$.

For $i = 1, \ldots, n$, note that the estimates $\hat{\mu}_X$, $\hat{\mu}_Y$, $\tilde{\mu}_X$, $\tilde{\mu}_Y$, $\hat{\sigma}_X$ and $\hat{\sigma}_Y$ satisfy:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right) &= 0;\\ \frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right) &= 0;\\ \frac{1}{n}\sum_{i=1}^{n}\chi\left(\frac{X_i-\tilde{\mu}_X}{\hat{\sigma}_X}\right) &= b;\\ \frac{1}{n}\sum_{i=1}^{n}\chi\left(\frac{Y_i-\tilde{\mu}_Y}{\hat{\sigma}_Y}\right) &= b,\end{aligned}$$

where $\tilde{\mu}_X$ and $\tilde{\mu}_Y$ are initial S-location estimates. Using Taylor expansion about the points (μ_X, σ_X) and (μ_Y, σ_Y), we define the following:

$$\begin{aligned}\psi\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right) &= \psi\left(\frac{X_i-\mu_X}{\sigma_X}\right) - \frac{1}{\breve{\sigma}_X}\psi'\left(\frac{X_i-\breve{\mu}_X}{\breve{\sigma}_X}\right)(\hat{\mu}_X-\mu_X)\\ &\quad -\frac{1}{\breve{\sigma}_X}\psi'\left(\frac{X_i-\breve{\mu}_X}{\breve{\sigma}_X}\right)\left(\frac{X_i-\breve{\mu}_X}{\breve{\sigma}_X}\right)(\hat{\sigma}_X-\sigma_X),\end{aligned}$$

for some $(\breve{\mu}_X, \breve{\sigma}_X)$ between (μ_X, σ_X) and $(\hat{\mu}_X, \hat{\sigma}_X)$, and

$$\begin{aligned}\psi\left(\frac{Y_i - \hat{\mu}_Y}{\hat{\sigma}_Y}\right) &= \psi\left(\frac{Y_i - \mu_Y}{\sigma_Y}\right) - \frac{1}{\breve{\sigma}_Y}\psi'\left(\frac{Y_i - \breve{\mu}_Y}{\breve{\sigma}_Y}\right)(\hat{\mu}_Y - \mu_Y) \\ &\quad -\frac{1}{\breve{\sigma}_Y}\psi'\left(\frac{Y_i - \breve{\mu}_Y}{\breve{\sigma}_Y}\right)\left(\frac{Y_i - \breve{\mu}_Y}{\breve{\sigma}_Y}\right)(\hat{\sigma}_Y - \sigma_Y),\end{aligned}$$

for some $(\breve{\mu}_Y, \breve{\sigma}_Y)$ between (μ_Y, σ_Y) and $(\hat{\mu}_Y, \hat{\sigma}_Y)$.

Therefore, $\frac{1}{n}\sum_{i=1}^n \psi\left(\frac{X_i - \hat{\mu}_X}{\hat{\sigma}_X}\right)\psi\left(\frac{Y_i - \hat{\mu}_Y}{\hat{\sigma}_Y}\right)$ can be written as

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n \left[\psi\left(\frac{X_i - \mu_X}{\sigma_X}\right) - \frac{1}{\breve{\sigma}_X}\psi'\left(\frac{X_i - \breve{\mu}_X}{\breve{\sigma}_X}\right)(\hat{\mu}_X - \mu_X)\right. \\ &- \left.\frac{1}{\breve{\sigma}_X}\psi'\left(\frac{X_i - \breve{\mu}_X}{\breve{\sigma}_X}\right)\left(\frac{X_i - \breve{\mu}_X}{\breve{\sigma}_X}\right)(\hat{\sigma}_X - \sigma_X)\right] \\ &\times \left[\psi\left(\frac{Y_i - \mu_Y}{\sigma_Y}\right) - \frac{1}{\breve{\sigma}_Y}\psi'\left(\frac{Y_i - \breve{\mu}_Y}{\breve{\sigma}_Y}\right)(\hat{\mu}_Y - \mu_Y)\right. \\ &- \left.\frac{1}{\breve{\sigma}_Y}\psi'\left(\frac{Y_i - \breve{\mu}_Y}{\breve{\sigma}_Y}\right)\left(\frac{Y_i - \breve{\mu}_Y}{\breve{\sigma}_Y}\right)(\hat{\sigma}_Y - \sigma_Y)\right].\end{aligned}$$

Using Serfling's lemma (Serfling, 1980, page 253), it is easy to show that

$$\begin{aligned}A_n &= \frac{1}{n\breve{\sigma}_Y}\sum_{i=1}^n \psi\left(\frac{X_i - \mu_X}{\sigma_X}\right)\psi'\left(\frac{Y_i - \breve{\mu}_Y}{\breve{\sigma}_Y}\right)\left(\frac{Y_i - \breve{\mu}_Y}{\breve{\sigma}_Y}\right) \to A \qquad \text{a.s.} \\ B_n &= \frac{1}{n\breve{\sigma}_X}\sum_{i=1}^n \psi\left(\frac{Y_i - \mu_Y}{\sigma_Y}\right)\psi'\left(\frac{X_i - \breve{\mu}_X}{\breve{\sigma}_X}\right)\left(\frac{X_i - \breve{\mu}_X}{\breve{\sigma}_X}\right) \to B \qquad \text{a.s.} \\ C_n &= \frac{1}{n\breve{\sigma}_Y}\sum_{i=1}^n \psi\left(\frac{X_i - \mu_X}{\sigma_X}\right)\psi'\left(\frac{Y_i - \breve{\mu}_Y}{\breve{\sigma}_Y}\right) \to 0 \qquad \text{a.s.} \\ D_n &= \frac{1}{n\breve{\sigma}_X}\sum_{i=1}^n \psi\left(\frac{Y_i - \mu_Y}{\sigma_Y}\right)\psi'\left(\frac{X_i - \breve{\mu}_X}{\breve{\sigma}_X}\right) \to 0 \qquad \text{a.s.,}\end{aligned}$$

as $n \to \infty$, where

$$A = \mathbb{E}\left\{\psi\left(\frac{X - \mu_X}{\sigma_X}\right)\psi'\left(\frac{Y - \mu_Y}{\sigma_Y}\right)\left(\frac{Y - \mu_Y}{\sigma_Y}\right)\right\},$$

and

$$B = \mathbb{E}\left\{\psi\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\psi'\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{X-\mu_X}{\sigma_X}\right)\right\}.$$

Moreover, all the other cross products are $o\left(1/\sqrt{n}\right)$. For example

$$\begin{aligned}
&\frac{1}{\breve{\sigma}_X\breve{\sigma}_Y\sqrt{n}}\sum_{i=1}^{n}\psi'\left(\frac{X_i-\breve{\mu}_X}{\breve{\sigma}_X}\right)\left(\hat{\mu}_X-\mu_X\right)\psi'\left(\frac{Y_i-\breve{\mu}_Y}{\breve{\sigma}_Y}\right)\left(\hat{\mu}_Y-\mu_Y\right)\\
= \quad & \left[\frac{1}{n\breve{\sigma}_X\breve{\sigma}_Y}\sum_{i=1}^{n}\psi'\left(\frac{X_i-\breve{\mu}_X}{\breve{\sigma}_X}\right)\psi'\left(\frac{Y_i-\breve{\mu}_Y}{\breve{\sigma}_Y}\right)\right]\\
&\times\left[\sqrt{n}\left(\hat{\mu}_Y-\mu_Y\right)\right]\left(\hat{\mu}_X-\mu_X\right)\to 0
\end{aligned}$$

as $n \to \infty$. Therefore,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\psi\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\\
\doteq \quad & \frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{X_i-\mu_X}{\sigma_X}\right)\psi\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)\\
&-\frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{X_i-\mu_X}{\sigma_X}\right)\psi'\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)\left(\hat{\sigma}_Y-\sigma_Y\right)\\
&-\frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)\psi'\left(\frac{X_i-\mu_X}{\sigma_X}\right)\left(\frac{X_i-\mu_X}{\sigma_X}\right)\left(\hat{\sigma}_X-\sigma_X\right),
\end{aligned}$$

where "$\doteq$" means "asymptotically equivalent". Hence,

$$\frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\psi\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)$$

is asymptotically equivalent to

$$\frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{X_i-\mu_X}{\sigma_X}\right)\psi\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)-A\left(\hat{\sigma}_Y-\sigma_Y\right)-B\left(\hat{\sigma}_X-\sigma_X\right). \qquad (4.21)$$

In addition,

$$0 = \frac{1}{n}\sum_{i=1}^{n}\chi\left(\frac{Y_i - \tilde{\mu}_Y}{\hat{\sigma}_Y}\right) - b$$

$$= \frac{1}{n}\sum_{i=1}^{n}\chi\left(\frac{Y_i - \mu_Y}{\sigma_Y}\right) - b - \left[\frac{1}{n}\sum_{i=1}^{n}\chi'\left(\frac{Y_i - \mu_Y}{\sigma_Y}\right)\right](\hat{\mu}_Y - \mu_Y)$$

$$- \left[\frac{1}{n}\sum_{i=1}^{n}\chi'\left(\frac{Y_i - \mu_Y}{\sigma_Y}\right)\left(\frac{Y_i - \mu_Y}{\sigma_Y}\right)\right](\hat{\sigma}_Y - \sigma_Y) + o\left(\frac{1}{\sqrt{n}}\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\chi\left(\frac{Y_i - \mu_Y}{\sigma_Y}\right) - b - \left[\frac{1}{n}\sum_{i=1}^{n}\chi'\left(\frac{Y_i - \mu_Y}{\sigma_Y}\right)\left(\frac{Y_i - \mu_Y}{\sigma_Y}\right)\right](\hat{\sigma}_Y - \sigma_Y) + o\left(\frac{1}{\sqrt{n}}\right),$$

because

$$1/n\sum\chi'\left(\frac{Y_i - \mu_Y}{\sigma_Y}\right) \to \mathbb{E}\left\{\chi'\left(\frac{Y - \mu_Y}{\sigma_Y}\right)\right\} = 0.$$

Analogously,

$$0 = \frac{1}{n}\sum_{i=1}^{n}\chi\left(\frac{X_i - \tilde{\mu}_X}{\hat{\sigma}_X}\right) - b$$

$$= \frac{1}{n}\sum_{i=1}^{n}\chi\left(\frac{X_i - \mu_X}{\sigma_X}\right) - b - \left[\frac{1}{n}\sum_{i=1}^{n}\chi'\left(\frac{X_i - \mu_X}{\sigma_X}\right)\left(\frac{X_i - \mu_X}{\sigma_X}\right)\right](\hat{\sigma}_X - \sigma_X) + o\left(\frac{1}{\sqrt{n}}\right).$$

Therefore,

$$\hat{\sigma}_Y - \sigma_Y \doteq \frac{\sum_{i=1}^{n}\chi\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right) - nb}{\sum_{i=1}^{n}\chi'\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)}, \tag{4.22}$$

and

$$\hat{\sigma}_X - \sigma_X \doteq \frac{\sum_{i=1}^{n}\chi\left(\frac{X_i-\mu_X}{\sigma_X}\right) - nb}{\sum_{i=1}^{n}\chi'\left(\frac{X_i-\mu_X}{\sigma_X}\right)\left(\frac{X_i-\mu_X}{\sigma_X}\right)}. \tag{4.23}$$

Replacing (4.22) and (4.23) in (4.21) we get

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\psi\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\\
\doteq\ &\frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{X_i-\mu_X}{\sigma_X}\right)\psi\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)\\
&-\frac{1}{n}\left[\sum_{i=1}^{n}\psi\left(\frac{X_i-\mu_X}{\sigma_X}\right)\psi'\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)\right]\left[\sum_{i=1}^{n}\chi\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)-nb\right]\\
&/\left[\sum_{i=1}^{n}\chi'\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)\right]\\
&-\frac{1}{n}\left[\sum_{i=1}^{n}\psi\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)\psi'\left(\frac{X_i-\mu_X}{\sigma_X}\right)\left(\frac{X_i-\mu_X}{\sigma_X}\right)\right]\left[\sum_{i=1}^{n}\chi\left(\frac{X_i-\mu_X}{\sigma_X}\right)-nb\right]\\
&/\left[\sum_{i=1}^{n}\chi'\left(\frac{X_i-\mu_X}{\sigma_X}\right)\left(\frac{X_i-\mu_X}{\sigma_X}\right)\right]\\
\doteq\ &\frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{X_i-\mu_X}{\sigma_X}\right)\psi\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)\\
&-\alpha\frac{1}{n}\sum_{i=1}^{n}\left(\chi\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)-b\right)-\beta\frac{1}{n}\sum_{i=1}^{n}\left(\chi\left(\frac{X_i-\mu_X}{\sigma_X}\right)-b\right), \qquad (4.24)
\end{aligned}
$$

where

$$\alpha=\frac{\mathbb{E}\left\{\psi\left(\frac{X-\mu_X}{\sigma_X}\right)\psi'\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right\}}{\mathbb{E}\left\{\chi'\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right\}},$$

and

$$\beta=\frac{\mathbb{E}\left\{\psi\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\psi'\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{X-\mu_X}{\sigma_X}\right)\right\}}{\mathbb{E}\left\{\chi'\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{X-\mu_X}{\sigma_X}\right)\right\}}.$$

Similar reasoning gives

$$\frac{1}{n}\sum_{i=1}^{n}\psi^2\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\doteq\frac{1}{n}\sum_{i=1}^{n}\psi^2\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)-2\delta\frac{1}{n}\sum_{i=1}^{n}\left(\chi\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)-b\right), \quad (4.25)$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\psi^2\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right) \doteq \frac{1}{n}\sum_{i=1}^{n}\psi^2\left(\frac{X_i-\mu_X}{\sigma_X}\right) - 2\delta\frac{1}{n}\sum_{i=1}^{n}\left(\chi\left(\frac{X_i-\mu_X}{\sigma_X}\right)-b\right), \quad (4.26)$$

where

$$\begin{aligned}\delta &= \frac{\mathbb{E}\left\{\psi\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\psi'\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right\}}{\mathbb{E}\left\{\chi'\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right\}} \\ &= \frac{\mathbb{E}\left\{\psi\left(\frac{X-\mu_X}{\sigma_X}\right)\psi'\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{X-\mu_X}{\sigma_X}\right)\right\}}{\mathbb{E}\left\{\chi'\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{X-\mu_X}{\sigma_X}\right)\right\}}.\end{aligned}$$

From (4.24), (4.25) and (4.26) it follows that

$$\sqrt{n}\left[\begin{pmatrix}\frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\psi\left(\frac{X_i-\hat{\mu}_Y}{\hat{\sigma}_X}\right)\\ \frac{1}{n}\sum_{i=1}^{n}\psi^2\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\\ \frac{1}{n}\sum_{i=1}^{n}\psi^2\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\end{pmatrix} - \begin{pmatrix}\mathbb{E}\left\{\psi\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\psi\left(\frac{X-\mu_X}{\sigma_X}\right)\right\}\\ \mathbb{E}\left\{\psi^2\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right\}\\ \mathbb{E}\left\{\psi^2\left(\frac{X-\mu_X}{\sigma_X}\right)\right\}\end{pmatrix}\right]$$

$$\longrightarrow_d \; N\left\{\begin{pmatrix}0\\0\\0\end{pmatrix}, \begin{pmatrix}\sigma_{11} & \sigma_{12} & \sigma_{13}\\ \sigma_{21} & \sigma_{22} & \sigma_{23}\\ \sigma_{31} & \sigma_{32} & \sigma_{33}\end{pmatrix}\right\}, \quad (4.27)$$

as $n \to \infty$, where

$$\begin{aligned}\sigma_{11} &= \mathrm{Var}\left\{\psi\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\psi\left(\frac{X-\mu_X}{\sigma_X}\right)\right\};\\ \sigma_{12} &= \mathrm{Cov}\left\{\psi\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\psi\left(\frac{X-\mu_X}{\sigma_X}\right), \psi^2\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right\};\\ \sigma_{13} &= \mathrm{Cov}\left\{\psi\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\psi\left(\frac{X-\mu_X}{\sigma_X}\right), \psi^2\left(\frac{X-\mu_X}{\sigma_X}\right)\right\};\\ \sigma_{22} &= \mathrm{Var}\left\{\psi^2\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right\};\\ \sigma_{23} &= \mathrm{Cov}\left\{\psi^2\left(\frac{Y-\mu_Y}{\sigma_Y}\right), \psi^2\left(\frac{X-\mu_X}{\sigma_X}\right)\right\};\end{aligned}$$

and

$$\sigma_{33} \;=\; \mathrm{Var}\left\{\psi^2\left(\frac{X-\mu_X}{\sigma_X}\right)\right\}.$$

To simplify the notation set

$$\begin{aligned}
U_n &= \frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\psi\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right), \quad u=\mathbb{E}\left\{\psi\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\psi\left(\frac{X-\mu_X}{\sigma_X}\right)\right\};\\
V_n &= \frac{1}{n}\sum_{i=1}^{n}\psi^2\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right), \quad v=\mathbb{E}\left\{\psi^2\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right\};\\
W_n &= \frac{1}{n}\sum_{i=1}^{n}\psi^2\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right), \quad w=\mathbb{E}\left\{\psi^2\left(\frac{X-\mu_X}{\sigma_X}\right)\right\}.
\end{aligned}$$

From (4.27) and using the δ-method (see Billingsley, 1986) we obtain

$$\begin{aligned}
\sqrt{n}\left(\hat{r}-r\right) &= \sqrt{n}\left(\frac{\frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\psi\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)}{\sqrt{\left[\frac{1}{n}\sum_{i=1}^{n}\psi^2\left(\frac{Y_i-\hat{\mu}_Y}{\hat{\sigma}_Y}\right)\right]\left[\frac{1}{n}\sum_{i=1}^{n}\psi^2\left(\frac{X_i-\hat{\mu}_X}{\hat{\sigma}_X}\right)\right]}}\right.\\
&\qquad\left. - \frac{\mathbb{E}\left\{\psi\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\psi\left(\frac{X-\mu_X}{\sigma_X}\right)\right\}}{\sqrt{\mathbb{E}\left\{\psi^2\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right\}\mathbb{E}\left\{\psi^2\left(\frac{X-\mu_X}{\sigma_X}\right)\right\}}}\right)\\
&= \sqrt{n}\left(\frac{U_n}{\sqrt{V_nW_n}}-\frac{u}{\sqrt{vw}}\right)\\
&= \sqrt{n}\left(g\left(U_n,V_n,W_n\right)-g\left(u,v,w\right)\right)\to_d N\left\{\mathbf{0},\nabla_g'\boldsymbol{\Sigma}\nabla_g\right\},
\end{aligned}$$

as $n\to\infty$, where

$$\boldsymbol{\Sigma}=\begin{pmatrix}\sigma_{11} & \sigma_{12} & \sigma_{13}\\ \sigma_{21} & \sigma_{22} & \sigma_{23}\\ \sigma_{31} & \sigma_{32} & \sigma_{33}\end{pmatrix},$$

$$g\left(u,v,w\right)=\left(\frac{u}{\sqrt{vw}}\right),$$

and

$$\nabla_g = \nabla_g(u,v,w) = \begin{pmatrix} (\partial/\partial u)\, g(u,v,w) \\ \\ (\partial/\partial v)\, g(u,v,w) \\ \\ (\partial/\partial w)\, g(u,v,w) \end{pmatrix} = \begin{pmatrix} 1/\sqrt{vw} \\ \\ -0.5\,(1/v)\,(u/\sqrt{vw}) \\ \\ -0.5\,(1/w)\,(u/\sqrt{vw}) \end{pmatrix}. \quad \blacksquare$$

4.9.3 Proof of Theorem 4.3

A famous result of Itô (1951), see Øksendal (1998) page 38, gives the following formula for n times iterated Itô integrals:

$$n! \int_{0\le u_1 \le \ldots \le u_n \le t} \ldots (\int (\int d\bar{B}_{u_1}) d\bar{B}_{u_2}) \ldots d\bar{B}_{u_n} = t^{\frac{n}{2}} h_n\left(\frac{\bar{B}(t)}{\sqrt{t}}\right), \tag{4.28}$$

where $\bar{B}$ is a given Brownian motion and h_n is the Hermite polynomial of degree n, defined by

$$h_n(x) = (-1)^n \exp\left(\frac{x^2}{2}\right) \frac{d^n}{dx^n} \exp\left(-\frac{x^2}{2}\right); \qquad n = 0, 1, 2, \ldots.$$

Thus the first Hermite polynomials are

$$\begin{aligned} h_0(x) &= 1, h_1(x) = x, h_2(x) = x^2 - 1, h_3(x) = x^3 - 3x, \\ h_4(x) &= x^4 - 6x^2 + 3, h_5(x) = x^5 - 10x^3 + 15x, \ldots. \end{aligned}$$

Then the normalized Hermite polynomials will be $H_n(x) = \frac{1}{n!} h_n(x)$.

Let $X(t)$ and $Y(t)$ be two Brownian motions such that $<X, Y>_t = \rho t$, where $<X, Y>_t$ is the quadratic variation of X and Y. In principle, we can construct $X(t)$ and $Y(t)$ from a couple of i.i.d. standard Brownian motions. Let B_1, B_2 and B be i.i.d. standard Brownian motions and define

$$\begin{aligned} X(t) &= \sqrt{1-\rho} B_1(t) + \sqrt{\rho} B(t), \\ Y(t) &= \sqrt{1-\rho} B_2(t) + \sqrt{\rho} B(t). \end{aligned}$$

Then

$$\mathbb{E}\left\{X(t)Y(t)\right\} = \rho t. \tag{4.29}$$

Note that $\int_0^1 d <X, Y>_s = \int_0^1 \rho ds = \rho = \mathbb{E}\{X(1)Y(1)\}$ by (4.29). Taking successive integrals, we construct the sequence $\{X_n(t)\}$ by

$$\begin{aligned} X_1(t) &= X(t) \\ X_2(t) &= \int_0^t X_1(s) dX(s) \\ &\vdots \\ X_{n+1}(t) &= \int_0^t X_n(s) dX(s). \end{aligned}$$

Similarly we construct the sequence $\{Y_m(t)\}$ by

$$\begin{aligned} Y_1(t) &= Y(t) \\ Y_2(t) &= \int_0^t Y_1(s) dY(s) \\ &\vdots \\ Y_{m+1}(t) &= \int_0^t Y_m(s) dY(s). \end{aligned}$$

Now applying (4.28) with $t = 1$ for $X(t)$, we obtain

$$\begin{aligned}
X_1(1) &= \int_0^1 dX(s) = X(1) - X(0) = X(1) = h_1(X(1)) = H_1(X(1)) \\
X_2(1) &= \int_0^1 X_1(u_1)dX(u_1) = \int_0^1 \left(\int_0^{u_1} dX(s)\right) dX(u_1) \\
&= \frac{1}{2!}h_2(X(1)) = H_2(X(1)) \\
X_3(1) &= \int_0^1 X_2(u_1)dX(u_1) \\
&= \int_0^1 \left(\int_0^{u_1} X_1(u_2)dX(u_2)\right) dX(u_1) \\
&= \int_0^1 \left(\int_0^{u_1} \left(\int_0^{u_2} dX(u_3)\right) dX(u_2)\right) dX(u_1) = \frac{1}{3!}h_3(X(1)) = H_3(X(1)) \\
&\vdots \\
X_n(1) &= \int_0^1 X_{n-1}(u_1)dX(u_1) = \frac{1}{n!}h_n(X(1)) = H_n(X(1)) = H_n(X).
\end{aligned}$$

Similarly, $Y_m(1) = H_m(Y(1)) = H_m(Y)$. Hence the sequences of $X_n(1)$ and $Y_m(1)$ are Hermite polynomials.

Now we show that for every n, we have

$$\mathbb{E}\left\{X_n(t)Y_n(t)\right\} = \frac{\rho^n t^n}{n!}.$$

Using the definition of stochastic integral, we obtain for every n and m,

$$\begin{aligned}
\mathbb{E}\left\{X_n(t)Y_m(t)\right\} &= \mathbb{E}\left\{\int_0^t X_{n-1}(s)dX(s)\int_0^t Y_{m-1}(s)dY(s)\right\} \\
&= \mathbb{E} < \int_0^t X_{n-1}(s)dX(s), \int_0^t Y_{m-1}(s)dY(s) > \\
&= \mathbb{E}\int_0^t X_{n-1}(s)Y_{m-1}(s)d < X, Y >_s \\
&= \int_0^t \mathbb{E}\left\{X_{n-1}(s)Y_{m-1}(s)\right\} \rho ds \\
&= \rho\int_0^t \mathbb{E}\left\{X_{n-1}(s)Y_{m-1}(s)\right\} ds.
\end{aligned}$$

By induction, we see that if $m = n$, then we obtain

$$\mathbb{E}\left\{X_n(t)Y_n(t)\right\} = \frac{\rho^n t^n}{n!},$$

while if $m \neq n$, then

$$\begin{aligned} \int_0^t \mathbb{E}\left\{Y_j(s)\right\} ds &= 0, \\ \int_0^t \mathbb{E}\left\{X_j(s)\right\} ds &= 0. \end{aligned}$$

Now $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ are orthogonal in L^2 space. Thus

$$\begin{aligned} \mathbb{E}\{X_i(t)X_j(t)\} &= 0, \\ \mathbb{E}\{X_i^2(t)\} &= t. \end{aligned}$$

Let $X = X(1)$ and $Y = Y(1)$, then we showed that the families $H_n(X)$ and $H_m(Y)$ are orthonormal with

$$\mathbb{E}\left\{H_n(X)H_m(Y)\right\} = \delta_{mn}\rho^n,$$

where

$$\delta_{mn} = \begin{cases} 1 & n = m \\ 0 & n \neq m. \end{cases}$$

Since the Hermite polynomials are complete, i.e. the functions $f(X)$ and $g(Y)$ with $\mathbb{E}\{f^2(X)\} < \infty$ and $\mathbb{E}\{g^2(Y)\} < \infty$, can be approximated by Hermite polynomials, thus

$$f(X) = \sum_{n=0}^{\infty} a_n H_n(X),$$

and

$$f(Y) = \sum_{m=0}^{\infty} a_m H_m(Y).$$

Without loss of generality, we assume that $f(X)$ and $g(Y)$ have mean zero and variance one. Consequently, $a_0 = 0$, $b_0 = 0$ and then,

$$\begin{aligned}
\mathbb{E}\left\{f(X)g(Y)\right\} &= \mathbb{E}\left\{\sum_{n=1}^{\infty} a_n H_n(X) \sum_{m=1}^{\infty} b_m H_m(Y)\right\} \\
&= \sum_{n=1}^{\infty}\sum_{m=1}^{\infty} a_n b_m \mathbb{E}\left\{H_n(X)H_m(Y)\right\} \\
&= \sum_{n=1}^{\infty} a_n b_n \mathbb{E}\left\{H_n(X)H_n(Y)\right\} \qquad \text{(by orthogonality)} \\
&= \sum_{n=1}^{\infty} a_n b_n \rho^n \\
&= \rho \sum_{n=1}^{\infty} a_n b_n \rho^{n-1} \\
&\leq |\rho| \sum_{n=1}^{\infty} |a_n||b_n| \qquad (|\rho| \leq 1) \\
&\leq |\rho| \sqrt{\sum_{n=1}^{\infty} a_n^2} \sqrt{\sum_{m=1}^{\infty} b_m^2} \qquad \text{(by Cauchy-Schwarz)} \\
&= |\rho| \mathbb{E}\{f^2(X)\}^{1/2} \mathbb{E}\{g^2(Y)\}^{1/2} \\
&= |\rho|.
\end{aligned}$$

Since $\mathbb{E}\left\{f^2(X)\right\}^{1/2} = 1$ and $\mathbb{E}\left\{g^2(Y)\right\}^{1/2} = 1$. ■

4.9.4 Proof of Theorem 4.4

Let H_0 and $\tilde{H}$ be elliptically symmetric distributions in $\mathbb{R}^2$ and assume that (X, Y) is distributed according to the following model:

$$H = (1 - \epsilon)H_0 + \epsilon\tilde{H}.$$

Assume without loss of generality that the location and scale parameters of X and Y are known, such that $\mu_X = \mu_Y = 0$ and $\sigma_X = \sigma_Y = 1$. We will show that the correlation coefficient $r(H)$ of $\psi(X)$ and $\psi(Y)$ satisfies:

$$(1 - \eta)r(H_0) - \eta \leq r(H) \leq (1 - \eta)r(H_0) + \eta,$$

where $\frac{\eta}{1-\eta} = \frac{\epsilon}{1-\epsilon} \cdot \frac{\psi^2(\infty)}{\mathbb{E}_{H_0}\{\psi^2(X)\}}$.

The Huberized correlation coefficient of X and Y is defined as follows:

$$\begin{aligned} r(H) &= \frac{\mathbb{E}_H \psi(X)\psi(Y)}{\sqrt{\mathbb{E}_H \psi^2(X) \mathbb{E}_H \psi^2(Y)}} \\ &= \frac{(1-\epsilon)\mathbb{E}_{H_0}\psi(X)\psi(Y) + \epsilon \mathbb{E}_{\tilde{H}}\psi(X)\psi(Y)}{\sqrt{(1-\epsilon)\mathbb{E}_{H_0}\psi^2(X) + \epsilon \mathbb{E}_{\tilde{H}}\psi^2(X)}\sqrt{(1-\epsilon)\mathbb{E}_{H_0}\psi^2(Y) + \epsilon \mathbb{E}_{\tilde{H}}\psi^2(Y)}}. \end{aligned}$$

By the Cauchy-Schwarz inequality

$$\leq \frac{(1-\epsilon)\mathbb{E}_{H_0}\psi(X)\psi(Y) + \epsilon \sqrt{\mathbb{E}_{\tilde{H}}\psi^2(X)\mathbb{E}_{\tilde{H}}\psi^2(Y)}}{\sqrt{(1-\epsilon)\mathbb{E}_{H_0}\psi^2(X) + \epsilon \mathbb{E}_{\tilde{H}}\psi^2(X)}\sqrt{(1-\epsilon)\mathbb{E}_{H_0}\psi^2(Y) + \epsilon \mathbb{E}_{\tilde{H}}\psi^2(Y)}}.$$

By symmetry, the worst $\tilde{H}$ must have the same marginals

$$\leq \frac{(1-\epsilon)\mathbb{E}_{H_0}\psi(X)\psi(Y) + \epsilon \mathbb{E}_{H_\infty}\psi^2(X)}{(1-\epsilon)\mathbb{E}_{H_0}\psi^2(X) + \epsilon \mathbb{E}_{H_\infty}\psi^2(X)}.$$

Because

$$\begin{aligned} |\mathbb{E}_{H_0}\psi(X)\psi(Y)| &\leq \mathbb{E}_{H_0}\psi^2(X) \\ &\leq \frac{(1-\epsilon)\mathbb{E}_{H_0}\psi(X)\psi(Y) + \epsilon\psi^2(\infty)}{(1-\epsilon)\mathbb{E}_{H_0}\psi^2(X) + \epsilon\psi^2(\infty)} \\ &= \frac{(1-\epsilon)r(H_0) + \epsilon\frac{\psi^2(\infty)}{\mathbb{E}_{H_0}\psi^2(X)}}{(1-\epsilon) + \epsilon\frac{\psi^2(\infty)}{\mathbb{E}_{H_0}\psi^2(X)}} \\ &= \frac{r(H_0) + \frac{\epsilon}{1-\epsilon}\frac{\psi^2(\infty)}{\mathbb{E}_{H_0}\psi^2(X)}}{1 + \frac{\epsilon}{1-\epsilon}\frac{\psi^2(\infty)}{\mathbb{E}_{H_0}\psi^2(X)}}. \end{aligned}$$

Set

$$\begin{aligned} \frac{\eta}{1-\eta} &= \frac{\epsilon}{1-\epsilon}\frac{\psi^2(\infty)}{\mathbb{E}_{H_0}\psi^2(X)} = A \\ \eta &= (1-\eta)A = A - \eta A \\ \eta &= \frac{A}{1+A}. \end{aligned}$$

Then

$$\begin{aligned} r(H) &\leq \frac{r(H_0)+A}{1+A} \\ &= (1-\eta)\left[r(H_0)+\frac{\eta}{1-\eta}\right] \\ &= (1-\eta)r(H_0)+\eta. \end{aligned}$$

Analogously

$$r(H) \geq \frac{r(H_0)-A}{1+A}.$$

Therefore

$$\frac{r(H_0)-A}{1+A} \leq r(H) \leq \frac{r(H_0)+A}{1+A}.$$

Now the contamination bias:

$$\begin{aligned} \frac{r(H_0)-A}{1+A} - r(H_0) &= \frac{r(H_0)-A-r(H_0)-Ar(H_0)}{1+A} \\ &= \frac{-A}{1+A}(1+r(H_0)), \end{aligned}$$

$$\frac{r(H_0)+A}{1+A} - r(H_0) = \frac{A}{1+A}(1-r(H_0)).$$

Therefore

$$\frac{-A}{1+A}(1+r(H_0)) \leq r(H)-r(H_0) \leq \frac{A}{1+A}(1-r(H_0)).$$

So

$$|r(H)-r(H_0)| \leq \frac{A}{1+A}(1+r(H_0)) \qquad \forall\ r(H_0) > 0,$$

and

$$|r(H)-r(H_0)| \leq \frac{A}{1+A}(1-r(H_0)) \qquad \forall\ r(H_0) > 0.$$

In summary

$$|r(H)-r(H_0)| \leq \frac{A}{1+A}(1+|r(H_0)|). \qquad \blacksquare$$

Note that

1. The worst case bias corresponds to $|r(H_0)| = 1$:

$$|r(H) - r(H_0)| \leq 2\frac{A}{1+A}.$$

2. The smallest value of A is $A = 1$ if $\psi(X) = \mathrm{SGN}(X)$, in other words the quadrant correlation is asymptotically minimax with respect to bias:

$$|r(H) - r(H_0)| \leq (1 + |r(H_0)|)\frac{\frac{\epsilon}{1-\epsilon}}{1 + \frac{\epsilon}{1-\epsilon}} = (1 + |r(H_0)|)\epsilon.$$

Chapter 5

Robust Estimation of Multivariate Location

The preceding chapter dealt with the estimation of the scatter matrix. However, the same idea of using fewer coordinates can also be applied to the estimation of the multivariate location. In this chapter, we discuss coordinate-wise estimate of a multivariate location which involves one dimensional data at a time. Particularly, we study the coordinate-wise median. Some notions of robustness are considered, such as, the minimaxity properties of the coordinate-wise median under the new contamination model (3.4).

5.1 Classical Multivariate Location Estimate

In this section, we consider the estimation of the "center" or location of a distribution in $\mathbb{R}^p$. Suppose that we have a multivariate random sample

$$\begin{aligned} X &= \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \\ &= (X_{11}, X_{12}, \ldots, X_{1p}), \ldots, (X_{n1}, X_{n2}, \ldots, X_{np}), \end{aligned}$$

so that the sample consists of n data points (rows) each of p dimensions (columns). A multivariate location estimate can be described as a $\mathbb{R}^p$-valued function, $\boldsymbol{T}_n$, defined for each sample size n, mapping the set of data points into some point $\boldsymbol{T}_n(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = \boldsymbol{T}_n(X)$, which is an approximation of the location of the distribution. A location estimate,

$\boldsymbol{T}_n$, is said to be translation and coordinate-wise scale equivariant if

$$\mathbf{T}_n(X+\mathbf{b}) = \mathbf{T}_n(X)+\mathbf{b},$$

for all constant vectors $\mathbf{b} \in \mathbb{R}^p$, where $X+\mathbf{b} = \{\boldsymbol{X}_1+\mathbf{b}, \ldots, \boldsymbol{X}_n+\mathbf{b}\}$, and

$$\mathbf{T}_n(XA) = \mathbf{T}_n(X)A,$$

for all diagonal $p \times p$ matrices $A = \mathrm{Diag}(a_1, \ldots, a_p)$, where $XA = \{\boldsymbol{X}_1 A, \ldots, \boldsymbol{X}_n A\}$.

The most well-known estimate of multivariate location is the arithmetic mean

$$\boldsymbol{T}_n(X) = \bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i,$$

which is basically the least squares estimate because it minimizes $\sum_{i=1}^{n} \|\boldsymbol{X}_i - \boldsymbol{T}\|^2$, where $\|\cdot\|$ is the Euclidean norm. However, it is not necessary that such an estimate is useful in all situations. It is well known that the arithmetic mean is not robust, because a single "bad" outlier in the sample can take $\bar{\boldsymbol{X}}$ arbitrarily far away. Using the definition of the multivariate location breakdown point (2.17) in Section 2.2 of Chapter 2, we can show that the multivariate arithmetic mean possesses a breakdown point of $1/n$. We often consider the limiting breakdown point as $n \to \infty$. Therefore we can say that the multivariate mean has 0 breakdown point, see Rousseeuw and Leroy (1987).

It is obvious that no translation location equivariant estimate can have a breakdown point larger than .50, because one could build a configuration of outliers which is just a translation image of the "good" data points, making it impossible for the estimate to choose. In univariate situations, this upper bound of .50 breakdown point can be attained, for example, by the sample median. Therefore, several multivariate generalizations of the median have been constructed, as well as some other proposals to achieve a certain amount of robustness.

5.2 Robust Multivariate Location Estimates

Robust alternatives to the arithmetic mean for estimating locations have a history going back at least to Laplace (see Stigler 1986, page 54). Fisher (1922) drew attention to the inefficiency of the arithmetic mean as an estimate of location for some distributions belonging to the family of Pearson curves near the normal. Using his normal contamination models, Tukey (1960) dramatically demonstrated how inefficient the mean can become when contamination increases. The same paper also shows how alternative location estimates such as the median or trimmed means can achieve higher asymptotic efficiency than the mean.

We distinguish between two classes of robust location estimates: those that are affine equivariant and those that are not. We say that, $\boldsymbol{T}_n$ is affine equivariant if and only if

$$\boldsymbol{T}_n(XA + \mathbf{b}) = \boldsymbol{T}_n(X)A + \mathbf{b},$$

for all nonsingular $p \times p$ matrices A and $\mathbf{b} \in \mathbb{R}^p$, where

$$XA + \mathbf{b} = \{\boldsymbol{X}_1 A + \mathbf{b}, \ldots, \boldsymbol{X}_n A + \mathbf{b}\}.$$

Sometimes we do not consider equivariance with respect to all affine transformations, but only for those that preserve Euclidean distances. An estimate is said to be orthogonal equivariant if

$$\boldsymbol{T}_n(XA + \mathbf{b}) = \boldsymbol{T}_n(X)A + \mathbf{b},$$

for all orthogonal $p \times p$ matrices A (i.e. $A' = A^{-1}$) and $\mathbf{b} \in \mathbb{R}^p$, where

$$XA + \mathbf{b} = \{\boldsymbol{X}_1 A + \mathbf{b}, \ldots, \boldsymbol{X}_n A + \mathbf{b}\}.$$

For instance, the L_1-location estimate defined as

$$\boldsymbol{T}_n = \arg\min_{\boldsymbol{T}} \sum_{i=1}^{n} \|\boldsymbol{X}_i - \boldsymbol{T}\|,$$

is orthogonal equivariant because it only depends on Euclidean distances. The L_1-estimate, also known as the spatial median or the mediancenter, is a generalization of the univariate median and its breakdown point is .50. Even though it is not affine equivariant, the L_1-estimate is clearly translation equivariant. Also, the foregoing minimum is known to be unique, except for degenerate cases. For some background and properties, refer to Small (1990); more recent results and references can also be found in Chaudhuri (1996) or Chakraborty and Chaudhuri (1999).

Although the affine equivariance property seems a natural requirement for an estimate, there are many practical situations in which it is not necessarily desirable from our point of view. For instance, requiring affine equivariance may be too restrictive when the data set is large and high-dimensional (e.g. data mining applications) or when there is a natural representation of the data up to a shift and/or change of units (arising from the form of measurement). We have seen that traditional affine equivariant estimates are prohibitively expensive for large multivariate data sets. We also have shown numerically that under the assumption of the new contamination model, affine equivariant estimates break down.

5.3 Coordinate-wise Location Estimates

The simplest and the straightforward approach is to consider each variable separately and simply calculate the robust location estimate for each of the individual variable. Indeed, for each variable the points $X_{1j}, X_{2j}, \ldots, X_{nj}$ can be considered as a one dimensional data set with n points (for $j = 1, \ldots, p$). Estimates of this type are called *coordinate-wise*, in which one applies the one-dimensional robust location estimate to each coordinate and combine the results into a p-dimensional estimate. This procedure inherits the breakdown point of the original estimate; however, although it is not affine equivariant it is translation-scale equivariant.

Note that, the multivariate arithmetic mean is an affine equivariant and can be computed coordinate-wise. However, this is an exception. Indeed, Donoho (1982, Proposition 4.6) shows that the only measurable location estimate that is both affine equivariant and computable as a vector of one-dimensional location estimates is the arithmetic mean.

A simple way to obtain coordinate-wise location estimates that are translation-scale equivariant with high breakdown point is to use a one-dimensional M-estimate to construct its multivariate analogue coordinate-wise. Given the p-dimensional distribution H, let H_i be the corresponding i-th marginal distribution. The coordinate-wise location M-estimate is defined as

$$\mathbf{T}(H) = \begin{pmatrix} T(H_1) \\ T(H_2) \\ \vdots \\ T(H_p) \end{pmatrix}, \tag{5.1}$$

where $T(H_i)$ is the corresponding M-estimate for the i-th marginal distribution H_i. To save computing time and still attain a high breakdown point, we consider the coordinate-wise median which is defined as

$$\mathbf{MED}(H) = \begin{pmatrix} med(H_1) \\ med(H_2) \\ \vdots \\ med(H_p) \end{pmatrix}, \tag{5.2}$$

with $med(H_i) = \text{median}(H_i)$.

5.4 Bias-Robustness Properties of Coordinate-wise Median

In this section, we focus on the bias-robustness of the coordinate-wise location estimates in the context of the contamination model. We will consider contamination model of the form:

$$\boldsymbol{X} = (I - B)\boldsymbol{Y} + B\boldsymbol{Z}, \tag{5.3}$$

where $\boldsymbol{Y}$, B and $\boldsymbol{Z}$ are independent, $\boldsymbol{Y}$ is multivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, $\boldsymbol{Z}$ is an arbitrary random vector and the diagonal elements of B, $B_1, \dots, B_p$ are i.i.d. Binomial(1,ϵ).

We will use the letters H, H_0 and $\tilde{H}$ to denote the joint distribution functions of $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{Z}$. Also, we will denote by $\mathcal{H}_\epsilon$ the set of all distribution functions for vectors (5.3) which is called *independent-contamination* neighborhood of size ϵ.

Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_n$ ($\boldsymbol{X}_i \in \mathbb{R}^p$) be i.i.d. H. We will consider estimates that satisfy the following property:

$$\mathbf{T}_n(\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_n) \;\rightarrow\; \mathbf{T}(H) \qquad \text{a.s.} \qquad \text{as} \quad n \rightarrow \infty.$$

In particular, let $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \dots, \boldsymbol{Y}_n$ ($\boldsymbol{Y}_i \in \mathbb{R}^p$) be i.i.d. H_0. Then,

$$\mathbf{T}_n(\boldsymbol{Y}_1, \boldsymbol{Y}_2, \dots, \boldsymbol{Y}_n) \;\rightarrow\; \mathbf{T}(H_0) \qquad \text{a.s.} \qquad \text{as} \quad n \rightarrow \infty.$$

Because of the contamination included in H, $\mathbf{T}(H)$ will typically be asymptotically biased. The asymptotic bias of $\mathbf{T}(H)$ with $H \in \mathcal{H}_\epsilon$ is defined as follows.

$$b(\mathbf{T},H) \;=\; \left\| \mathrm{Diag}(\sigma_{11}^{-1/2}, \sigma_{22}^{-1/2}, \dots, \sigma_{pp}^{-1/2})(\mathbf{T}(H) - \mathbf{T}(H_0)) \right\|,$$

where $\|\cdot\|$ is an arbitrary norm in $\mathbb{R}^p$ and if $\|\cdot\|$ is the Euclidean norm then,

$$b(\mathbf{T},H) \;=\; \sqrt{\sum_{i=1}^{p} \left[T_i(H) - T_i(H_0)\right]^2 / \sigma_{ii}}, \tag{5.4}$$

where σ_{ii}^2 is the diagonal i-th element of the covariance matrix $\boldsymbol{\Sigma}$.

Notice that the asymptotic bias (5.4) is translation and coordinate-wise scale invariant. The corresponding maximum asymptotic bias (maxbias) over the contamination neighborhood $\mathcal{H}_\epsilon$ is defined as follows.

$$B_{\mathbf{T}}(\epsilon) = \sup_{H \in \mathcal{H}_\epsilon} b(\mathbf{T},H). \tag{5.5}$$

Since we will only consider translation-scale equivariant estimates, we can assume without loss of generality that $\mu_1 = \mu_2 = \ldots = \mu_p = 0$ and $\sigma_{11} = \sigma_{22} = \cdots = \sigma_{pp} = 1$. Therefore, the asymptotic bias of $\mathbf{T}(H)$ satisfies:

$$b(\mathbf{T},H) = \|\mathbf{T}(H)\| = \sqrt{\sum_{i=1}^{p} T_i(H)^2},$$

and the maxbias can be written as

$$B_{\mathbf{T}}(\epsilon) = \sup_{H \in \mathcal{H}_\epsilon} \|\mathbf{T}(H)\|.$$

The following results highlight the good bias-robustness properties of the coordinate-wise multivariate median estimate. The next theorem show that when $\boldsymbol{Y}$ in (5.3) has multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma} = \mathrm{Diag}(\sigma_{11}^2, \sigma_{22}^2, \ldots, \sigma_{pp}^2)$, then the coordinate-wise median is the minimax-bias among translation-scale equivariant multivariate location estimates.

THEOREM 5.1 – ***Independent Normal Distributions*** – *The coordinate-wise median* $\mathbf{MED}(H)$ *(see (5.2)) minimizes the maximum asymptotic bias (5.5) at the independent-contamination neighborhood* $\mathcal{H}_\epsilon$ $(0 < \epsilon < 1/2)$ *centered at the multivariate normal distribution with mean* $\boldsymbol{\mu}$ *and covariance matrix* $\boldsymbol{\Sigma} = Diag(\sigma_{11}^2, \sigma_{22}^2, \ldots, \sigma_{pp}^2)$ *(in (5.3)) among translation-scale equivariant multivariate location estimates.*

The proof of Theorem 5.1 is given in Section 5.5.1 of the chapter appendix.

Notice that Theorem 5.1 holds regardless of the $\mathbb{R}^p$-norm used to define the asymptotic bias. This generality is of practical importance because the choice of an appropriate norm could depend on the given problem.

We also notice that the maximum bias depends on the correlation structure of the "core" data. Therefore we should consider different correlation structures, in addition to the independent case addressed by Theorem 5.1. Unfortunately we could not extend Theorem 5.1 for correlated normal distributions in its complete generality. However, we obtained the following result for the class of "*marginally-consistent*" estimates.

Suppose that H is a p-dimensional distribution with the following property: for each $i = 1, 2, \ldots, p$ the i-th one-dimensional marginal distribution of the H is symmetric about μ_i. Then

$$\mathbf{T}(H) = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}.$$

THEOREM 5.2–***General Normal Distributions***–*The coordinate-wise median* $\mathbf{MED}(H)$ *(see (5.2)) minimizes the maximum asymptotic bias (5.5) at the independent-contamination neighborhood* $\mathcal{H}_\epsilon$ $(0 < \epsilon < 1/2)$ *centered at the multivariate normal distribution with mean* $\boldsymbol{\mu}$ *and covariance matrix* $\boldsymbol{\Sigma}$ *(in (5.3)) among translation-scale equivariant, marginally-consistent multivariate location estimates.*

The proof of Theorem 5.2 is given in Section 5.5.2 of the chapter appendix.

5.5 Chapter Appendix

5.5.1 Proof of Theorem 5.1

We will use an approach similar to Huber's (1964) proof of the minimaxity of the univariate median. For simplicity of notation, we consider the case $p = 2$. A similar approach can be followed for the cases $p > 2$.

Because of the translation-scale invariance of the maximum asymptotic bias (5.5) and the location-scale equivariance of the coordinate-wise estimate, we can assume without loss of generality that $\mu_1 = \mu_2 = 0$ and $\sigma_{11} = \sigma_{22} = 1$.

Consider the independent-contamination neighborhood $\mathcal{H}_\epsilon$ (5.3) for H_0, with $P(B_i = 1) = \epsilon$, $i = 1, 2$. Let f be the joint density function of a distribution $F \in \mathcal{H}_\epsilon$. Then f

must be of the form:

$$\begin{aligned} f(x_1,x_2) &= f(x_1,x_2|B_1=0,B_2=0)P(B_1=0,B_2=0) \\ &\quad +f(x_1,x_2|B_1=0,B_2=1)P(B_1=0,B_2=1) \\ &\quad +f(x_1,x_2|B_1=1,B_2=0)P(B_1=1,B_2=0) \\ &\quad +f(x_1,x_2|B_1=1,B_2=1)P(B_1=1,B_2=1) \\ &= (1-\epsilon)^2\,\varphi(x_1)\varphi(x_2)+\epsilon\,(1-\epsilon)\,\varphi(x_1)h_2(x_2) \\ &\quad +\epsilon\,(1-\epsilon)\,h_1(x_1)\varphi(x_2)+\epsilon^2 h_3(x_1,x_2), \qquad (5.6) \end{aligned}$$

where φ is the standard normal density and $h_1(x_1)$, $h_2(x_2)$ and $h_3(x_1,x_2)$ are arbitrary densities.

The quantity $x_0 = \Phi^{-1}\left(\frac{1}{2(1-\epsilon)}\right)$ corresponds to the maxbias of the univariate median (see Huber, 1964) and will play a central role in our derivations below. Huber considered the case of a single ϵ-contaminated normal density

$$f(x) = (1-\epsilon)\varphi(x)+\epsilon h(x),$$

and the corresponding contamination neighborhood is denoted by $\mathcal{F}_\epsilon$. Huber constructed two ϵ-contaminated normal distributions F_+ and F_-, which are symmetric about x_0 and $-x_0$, respectively and which are translations of each other; that is

$$F_-(x) = F_+(x+2x_0).$$

The densities for F_+ and F_- are defined respectively as follows.

$$f_+(x) = \begin{cases} (1-\epsilon)\varphi(x) & x \le x_0 \\ (1-\epsilon)\varphi(x-2x_0) & x > x_0, \end{cases} \qquad (5.7)$$

$$f_-(x) = \begin{cases} (1-\epsilon)\varphi(x+2x_0) & x \le -x_0 \\ (1-\epsilon)\varphi(x) & x > -x_0. \end{cases} \qquad (5.8)$$

Therefore, for any translation equivariant location estimate, T, we have

$$T(F_+) - T(F_-) = 2x_0,$$

which implies that there is not any translation equivariant functional that can have an absolute bias smaller than x_0 at F_+ and F_- simultaneously. In fact, for any translation invariant functional

$$\begin{aligned} |T(F_+) - T(F_-)| &\leq |T(F_+)| + |T(F_-)| \\ 2x_0 &\leq |T(F_+)| + |T(F_-)|, \end{aligned}$$

and so either $|T(F_+)| > x_0$ or $|T(F_-)| > x_0$. Then T has larger maxbias than that of the median. Therefore,

$$\sup_{F \in \mathcal{F}_\epsilon} |T(F)| \geq x_0.$$

Huber's result then follows because f_+ and f_- belong to the ϵ-contamination neighborhood of the standard normal density. To see that let

$$f_+(x) = (1-\epsilon)\varphi(x) + \epsilon h_+(x), \qquad f_-(x) = (1-\epsilon)\varphi(x) + \epsilon h_-(x).$$

With

$$h_+(x) = \frac{f_+(x) - (1-\epsilon)\varphi(x)}{\epsilon}, \quad h_-(x) = \frac{f_-(x) - (1-\epsilon)\varphi(x)}{\epsilon}.$$

The claim follows then provided that:

1. $h_+(x) \geq 0$, $h_-(x) \geq 0$;

2. $\int_{-\infty}^{\infty} h_+(x)dx = \int_{-\infty}^{\infty} h_-(x)dx = 1$.

To see Part 1, notice that

$$f_+(x) - (1-\epsilon)\varphi(x) = \begin{cases} 0 & x < x_0 \\ \epsilon(1-\epsilon)[\varphi(x-2x_0) - \varphi(x)] & x \geq x_0, \end{cases}$$

$$f_-(x) - (1-\epsilon)\varphi(x) = \begin{cases} \epsilon(1-\epsilon)[\varphi(x+2x_0) - \varphi(x)] & x < -x_0 \\ 0 & x \geq -x_0. \end{cases}$$

This follows because of the inequalities:

$$\begin{aligned} \varphi(x - 2x_0) - \varphi(x) &\geq 0 \\ (x-2x_0)^2 &\leq x^2 \\ x_0(x_0 - x) &\leq 0 \\ x_0 &\leq x, \end{aligned}$$

and

$$\begin{aligned} \varphi(x + 2x_0) - \varphi(x) &\geq 0 \\ (x+2x_0)^2 &\leq x^2 \\ x_0(x_0 + x) &\leq 0 \\ x_0 &\leq -x. \end{aligned}$$

To see Part 2 holds, notice that

$$\begin{aligned} (1/\epsilon)\int_{-\infty}^{\infty} [f_+(x) - (1-\epsilon)\varphi(x)]\,dx &= \frac{1-\epsilon}{\epsilon}\int_{x_0}^{\infty} [\varphi(x-2x_0) - \varphi(x)]\,dx \\ &= \frac{1-\epsilon}{\epsilon}[1 - \Phi(-x_0) - 1 + \Phi(x_0)] \\ &= \frac{1-\epsilon}{\epsilon}[2\Phi(x_0) - 1] \\ &= \frac{1-\epsilon}{\epsilon}\left[2\Phi\left(\Phi^{-1}\left(\frac{1}{2(1-\epsilon)}\right)\right) - 1\right] \\ &= \frac{1-\epsilon}{\epsilon}\left[\frac{1}{(1-\epsilon)} - 1\right] \\ &= 1. \end{aligned}$$

A similar calculation shows that

$$(1/\epsilon)\int_{-\infty}^{\infty} [f_-(x) - (1-\epsilon)\varphi(x)]\,dx = 1.$$

Because F_+ and F_- are translations of each other and T is translation equivariant, then we have

$$\begin{aligned} F_-(x) &= F_+(x + 2x_0) \\ T(F_-(x)) &= T(F_+(x + 2x_0)) \\ T(F_-(x)) &= T(F_+(x)) - 2x_0 \\ T(F_+) - T(F_-) &= 2x_0. \end{aligned}$$

Now we turn our attention to the case $p = 2$. We define

$$g_+(x_1, x_2) = f_+(x_1)f_+(x_2),$$

and

$$g_-(x_1, x_2) = f_-(x_1)f_-(x_2).$$

With f_+ and f_- given by (5.7) and (5.8). We now proceed as follows:

1. Use the fact that $f_+(x) = (1 - \epsilon)\varphi(x) + \epsilon h_+(x)$ and $f_-(x) = (1 - \epsilon)\varphi(x) + \epsilon h_-(x)$ to show that $g_+(x_1, x_2)$ and $g_-(x_1, x_2)$ are of the form (5.6) and therefore the corresponding distribution functions G_+ and G_- belong to the independent-contamination neighborhood $\mathcal{H}_\epsilon$.

2. Show that $g_+(x_1, x_2)$ and $g_-(x_1, x_2)$ are translations of each other.

For Part 1, we write

$$\begin{aligned} g_+(x_1, x_2) &= f_+(x_1)f_+(x_2) \\ &= \left[(1 - \epsilon)\varphi(x_1) + \epsilon h_+(x_1)\right]\left[(1 - \epsilon)\varphi(x_2) + \epsilon h_+(x_2)\right] \\ &= (1 - \epsilon)^2\varphi(x_1)\varphi(x_2) + (1 - \epsilon)\epsilon\varphi(x_1)h_+(x_2) \\ &\quad + \epsilon(1 - \epsilon)h_+(x_1)\varphi(x_2) + \epsilon^2 h_+(x_1)h_+(x_2), \end{aligned}$$

and

$$\begin{aligned}
g_{-}(x_1,x_2) &= f_{-}(x_1)f_{-}(x_2) \\
&= [(1-\epsilon)\varphi(x_1)+\epsilon h_{-}(x_1)]\,[(1-\epsilon)\varphi(x_2)+\epsilon h_{-}(x_2)] \\
&= (1-\epsilon)^2\varphi(x_1)\varphi(x_2)+(1-\epsilon)\epsilon\varphi(x_1)h_{-}(x_2) \\
&\quad +\epsilon(1-\epsilon)h_{-}(x_1)\varphi(x_2)+\epsilon^2 h_{-}(x_1)h_{-}(x_2).
\end{aligned}$$

Hence $g_{+}(x_1,x_2)$ and $g_{-}(x_1,x_2)$ belong to the independent-contamination model (5.6). For Part 2, notice that since

$$\begin{aligned}
f_{-}(x_1) &= f_{+}(x_1+2x_0), \\
f_{-}(x_2) &= f_{+}(x_2+2x_0).
\end{aligned}$$

Then

$$g_{-}(x_1,x_2)=g_{+}(x_1+2x_0,x_2+2x_0).$$

Now, similar to Huber (1964) for all translation equivariant estimate $\mathbf{T}$ we have

$$\mathbf{T}(G_{+})-\mathbf{T}(G_{-})=2\begin{pmatrix} x_0 \\ x_0 \end{pmatrix},$$

and so

$$\|\mathbf{T}(G_{+})-\mathbf{T}(G_{-})\|=\sqrt{(2x_0)^2+(2x_0)^2}=\sqrt{2}(2x_0).$$

Moreover,

$$\begin{aligned}
\sqrt{2}(2x_0) &= \|\mathbf{T}(G_{+})-\mathbf{T}(G_{-})\| \\
&\le \|\mathbf{T}(G_{+})\|+\|\mathbf{T}(G_{-})\|,
\end{aligned}$$

and so

$$\sqrt{2}(2x_0)\le 2\sup_{H\in\mathcal{H}_\epsilon}\|\mathbf{T}(H)\|=2B_{\mathbf{T}}(\epsilon).$$

This yields,

$$B_{\mathbf{T}}(\epsilon)\ge\sqrt{2}x_0. \tag{5.9}$$

On the other hand,

$$\begin{aligned}\sup_{H\in\mathcal{H}_\epsilon} \|\mathbf{MED}(H)\| &= \sup_{H\in\mathcal{H}_\epsilon} \sqrt{\text{med}_1^2(H)+\text{med}_2^2(H)}\\ &\le \sqrt{\sup_{H\in\mathcal{H}_\epsilon} \text{med}^2(H_1)+\sup_{H\in\mathcal{H}_\epsilon} \text{med}^2(H_2)},\end{aligned}$$

where

$$\mathbf{MED}(H)=\begin{pmatrix}\text{med}(H_1)\\ \text{med}(H_2)\end{pmatrix}$$

is the coordinate-wise median. Since H_1 and H_2 are the distribution functions for $(1-B_1)X_1+B_1\tilde{X}_1$ and $(1-B_2)X_2+B_2\tilde{X}_2$ with $X_1\sim N(0,1)$ and $X_2\sim N(0,1)$, by Huber (1964)

$$\sup_{H\in\mathcal{H}_\epsilon} \|\mathbf{MED}(H)\| \le \sqrt{x_0^2+x_0^2}=\sqrt{2}x_0. \tag{5.10}$$

The result now follows from (5.9) and (5.10). ■

5.5.2 Proof of Theorem 5.2

For simplicity of notation, we consider the case $p=2$. A similar approach can be followed for the case $p>2$. We can assume without loss of generality that $\boldsymbol{\mu}=\mathbf{0}$ and

$$\boldsymbol{\Sigma}=\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}. \tag{5.11}$$

Let

$$\mathbf{X} = \begin{pmatrix}X_1\\ X_2\end{pmatrix}=\begin{pmatrix}(1-B_1)\,Y_1\\ (1-B_2)\,Y_2\end{pmatrix}+\begin{pmatrix}B_1Z_1\\ B_2Z_2\end{pmatrix},$$

where Y_1 and Y_2 are jointly normal with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$ (5.11). In addition, $B=\text{Diag}\,(B_1,B_2)$, $\mathbf{Y}=(Y_1,Y_2)'$ and $\mathbf{Z}=(Z_1,Z_2)'$ are independent and the variables Z_1 and Z_2 are independent with common density

$$h_+(x) = \begin{cases}0 & x<x_0\\ \frac{1-\epsilon}{\epsilon}[\varphi(x-2x_0)-\varphi(x)] & x\ge x_0,\end{cases}$$

where as before, $x_0 = \Phi^{-1}\left(\frac{1}{2(1-\epsilon)}\right)$. We notice that for $i = 1, 2$

$$X_i \;=\; (1 - B_i)\, Y_i + B_i Z_i,$$

has density

$$g_+(x) \;=\; \begin{cases} 0 & x < x_0 \\ (1-\epsilon)\, \varphi(x - 2x_0) & x \geq x_0, \end{cases}$$

which is symmetric at x_0. Therefore

$$\mathbf{T}\,(\mathbf{X}) \;=\; \begin{pmatrix} x_0 \\ x_0 \end{pmatrix}.$$

On the other hand, by similar arguments

$$\begin{aligned} \tilde{\mathbf{X}} \;&=\; -\mathbf{X}{=}\begin{pmatrix} -X_1 \\ -X_2 \end{pmatrix} = \begin{pmatrix} (1-B_1)\,(-Y_1) \\ (1-B_2)\,(-Y_2) \end{pmatrix} + \begin{pmatrix} B_1\,(-Z_1) \\ B_2\,(-Z_2) \end{pmatrix} \\ &=\; \begin{pmatrix} (1-B_1)\,\tilde{Y}_1 \\ (1-B_2)\,\tilde{Y}_2 \end{pmatrix} + \begin{pmatrix} B_1 \tilde{Z}_1 \\ B_2 \tilde{Z}_2 \end{pmatrix}, \end{aligned}$$

where $\tilde{Y}_1$ and $\tilde{Y}_2$ are jointly normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}$ (5.11). Since $B = \mathrm{Diag}\,(B_1, B_2)\,,\ \tilde{\mathbf{Y}} = \left(\tilde{Y}_1, \tilde{Y}_2\right)'$ and $\tilde{\mathbf{Z}} = \left(\tilde{Z}_1, \tilde{Z}_2\right)'$ are independent, the joint distribution of $\tilde{\mathbf{X}}$ belongs to $\mathcal{H}_\epsilon$. Moreover, the variables $\tilde{Z}_1$ and $\tilde{Z}_2$ are independent with common density

$$h_-(x) \;=\; \begin{cases} \frac{1-\epsilon}{\epsilon}[\varphi(x + 2x_0) - \varphi(x)] & x < -x_0 \\ 0 & x \geq -x_0, \end{cases}$$

and so, for $i = 1, 2$

$$\tilde{X}_i \;=\; (1 - B_i)\, \tilde{Y}_i + B_i \tilde{Z}_i,$$

has density

$$g_-(x) \quad = \quad \begin{cases} (1-\epsilon)\,\varphi(x-2x_0) & x < -x_0 \\ 0 & x \geq -x_0, \end{cases}$$

which is symmetric at $-x_0$. Therefore

$$\mathbf{T}\left(\tilde{\mathbf{X}}\right) \quad = \quad \begin{pmatrix} -x_0 \\ -x_0 \end{pmatrix}.$$

Hence, for all marginally consistent and translation-scale equivariant estimate $\mathbf{T}$ we have

$$\mathbf{T}(G_+) - \mathbf{T}(G_-) = 2\begin{pmatrix} x_0 \\ x_0 \end{pmatrix},$$

where G_+ and G_- are the distribution functions for G_+ and G_-, respectively. So as in the proof of Theorem 5.1, we obtain

$$B_{\mathbf{T}}(\epsilon) \quad \geq \quad \sqrt{2}x_0. \tag{5.12}$$

The theorem now follows from (5.12) and (5.10). ■

Chapter 6

Conclusion

Our study may be divided into three parts. In Part I we introduced a new contamination model that is more suitable than the existing contamination models for the multivariate setting. Part II dealt with the estimation of multivariate scatter where we revisited Huber's (1981) proposal and extended some of his results. Finally, in Part III, we studied the estimation of the multivariate location, in which we considered the coordinate-wise estimation.

The following is an outline of the main results obtained in the thesis, the problems that we encountered, and the directions we foresee for future work.

- **A New Contamination Model**:

  - ▼ We introduced a new multivariate contamination model that adequately represents reality for many multivariate data sets that arise in practice. This model resolves the deficiency of the current contamination models by allowing more flexibility and certain forms of dependency that the existing contamination models do not address.

  - ▼ We gave some arguments and numerical evidence which indicate that the breakdown point of affine equivariant estimates tends to zero when the dimension p tends to infinity under the new contamination model.

  - ▼ Our study concentrated on the multivariate location and scatter matrix estimates. It is desirable to investigate the performance of robust regression analysis and some other related models (e.g., error in variables) using the new contamination model. It is also of interest to revisit the problem of outliers detection in time series in the context of the new contamination model. This will be based on an embedding of the time series which allows us to regard the time series as a multivariate sample with identically distributed but non independent observations. Thus, multivariate outlier identifiers can be transferred into the context of time series. This gives interesting insights in some features of outliers in time dependent data, which are not recognizable by other methods, see Gather et al. (2003).

- **Simple Robust Pairwise Scatter Estimates**:

  ▼ The criterion for the selection of a good robust estimate includes small maxbias, high breakdown point and computational feasibility. Based on this criterion, we singled out a particular robust pairwise scatter estimate, namely, the Huberized correlation coefficient (with $c = 1$) based scatter matrix estimates. We found that this estimate is computationally simpler than the Fast MCD and other fast scatter estimates recently proposed (see Maronna and Zamar, 2002). We also showed that the Huberized estimate is more stable than the Fast MCD estimate when the data are contaminated according to the independent-contamination model.

  ▼ We studied the consistency and asymptotic normality of the Huberized correlation coefficient estimates. It remains to establish the asymptotic distribution of the Huberized correlation coefficient estimates using other than MM-location and scale estimates.

  ▼ We added scalability to Maronna and Zamar (2002) pairwise robust scatter estimate by replacing the robustified Gnanadesikan and Kettenring (1972) robust scale-based scatter estimate by the quadrant correlation estimate. We showed its scalability to high dimensions and large sample sizes.

- ▼ We studied the maxbias and the intrinsic bias of the Huberized correlation coefficient estimates. We then extended Huber's (1981) maxbias formulas and derived the analytical form of the maxbias for the quadrant correlation (QC) coefficient when the locations and scales are unknown. It is desirable to show that the QC is also minimax in this more general context.

- **Coordinate-wise Robust Location Estimates**:

  - ▼ In this part we used the same criterion as in the case of the multivariate scatter matrix to select a good robust estimate. The coordinate-wise medians appeared to fulfil our requirements.

  - ▼ We studied the minimaxity properties of the coordinate-wise median for two special cases. The independent situation and the correlated situation; for the latest case we restricted attention to the class of marginally-consistent estimates. We have not been able to show that the coordinate-wise median is minimax in general. However, we conjecture that this is true and, therefore, deserves further study.

  - ▼ Our proposed Huberized covariance matrix estimates and the coordinate-wise medians can be used to define a robustified Mahalanobis distance. It is of interest to study its approximate distribution properties.

Bibliography

Abdullah, M. B. (1990). On a robust correlation coefficient. *The Statistician*, 39:455–460.

Adrover, J. and Yohai, V. J. (2002). Projection estimates of multivariate location. Tentatively accepted in the Annals of Statistics.

Bickel, P. J. (1964). On some alternative estimates for shift in the p-variate one sample problem. *Ann. Math. Statist.*, 35:1079–1090.

Billingsley, P. (1986). *Probability and Measure.* John Wiley and Sons, 2 edition.

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

Carroll, R. J. and Ruppert, D. (1985). Transformation in regression: A robust analysis. *Technometrics*, 27:1–12.

Chakraborty, B. and Chaudhuri, P. (1999). A note on the robustness of multivariate medians. *Statist. Probab. Letters*, 45:269–276.

Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association*, 91:862–872.

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

Croux, C., Haesbroeck, G., and Rousseeuw, P. J. (1997). Location adjustment for the minimum volume ellipsoid estimator. Technical report, University of Brussels (ULB). Available at http://homepages.ulb.ac.be/˜ccroux/public.htm.

Croux, C. and Rousseeuw, P. J. (1992a). A class of high-breakdown scale estimators based on subranges. *Communications in Statistics, Theory Methods*, 21:1935–1951.

Croux, C. and Rousseeuw, P. J. (1992b). Time-efficient algorithms for two highly robust estimators of scale. *Computational Statistics*, 2:411–428.

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

Davies, P. L. (1992). The asymptotic of Rousseeuw's minimum volume ellipsoid. *Annals of Statistics*, 20:1828–1843.

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

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

Fang, K. T. and Zhang, Y. (1990). *Generalized Multivariate Analysis.* Springer and Science Press, Berlin and Beijing.

Fernholz, L. (1983). *Von Mises Calculus for Statistical Functionals.* Springers, New York.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Astr. Soc London Ser. A*, 222:309–368.

Fraumeni, J. F. (1968). Cigarette smoking and cancers of the united track: Geographic variations in the united states. *Journal of the National Cancer Institute*, 41:1205–1211.

Gather, U., Bauer, M., and Fried, R. (2003). The identification of multiple outliers in online monitoring data. In process.

Genton, M. G. and Ma, Y. (1999). Robustness properties of dispersion estimators. *Statistics and Probability Letters*, 44:343–350.

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

Grübel, R. (1988). A minimal characterization of the covariance matrix. *Metrika*, 35:49–52.

Haldane, J. B. S. (1948). Note on the median of multivariate distribution. *Biometrika*, 35:414–415.

Hampel, F. (1968). *Contributions to the Theory of Robust Estimation.* PhD thesis, Univ. Calif., Berkeley.

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

Hampel, F. (1973). Robust estimation: A condensed partial survey. *Zeitschrift für Wahrscheinlichkeitstheorie and Verwandte Gebiete*, 27:87–104.

Hampel, F., Ronchetti, E., Rousseeuw, P. J., and Stahel, W. (1986). *Robust Statistics: The Approach Based on Influence Functions.* John Wiley and Sons, Inc. New York.

Hawkins, D. (1994). The feasible solution algorithm for the minimum covariance determinant estimator in multivariate data. *Comput. Statist. Data. Anal.*, 17:197–210.

He, X. and Simpson, D. (1992). Robust direction estimation. *Annals of Statistics*, 20:351–369.

He, X. and Simpson, D. (1993). Lower bounds for contamination bias: Globally minimax versus locally linear estimation. *Annals of Statistics*, 21:314–337.

Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.*, 35:73–101.

Huber, P. J. (1977). Robust covariances. In Gupta, S. S. and Moore, D. S., editors, *Statistical Decision Theory and Related Topics 2*, pages 165–191. Academic Press.

Huber, P. J. (1981). *Robust Statistics.* John Wiley and Sons.

Itô, K. (1951). Multiple Wiener integral. *Journal of Mathematical Society*, 3:157–169. Japan.

Joe, H. (1997). *Multivariate Models and Dependence Concepts.* Chapman and Hall, 1 edition.

Li, B. and Zamar, R. H. (1991). Min-max asymptotic variance of M-estimates of location when scale is unknown. *Statistics and Probability Letters*, 11:139–145.

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

Lopuhaä, H. (1989). On the relation between S-estimators and M-estimators of multivariate location and covariance. *Annals of Statistics*, 17:1662–1683.

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

Ma, Y. and Genton, M. G. (2001). Highly robust estimation of dispersion matrices. *Journal of Multivariate Analysis*, 78(1):11–36.

Maronna, R. A. (1974). *Estimacion Robusta De Locacion Y Dispersion Multivariadas.* PhD thesis, Universidad de Buenos Aires.

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

Maronna, R. A., Stahel, W. A., and Yohai, V. J. (1992). Bias-robust estimation of multivariate scatter based on projections. *Journal of Multivariate Analysis*, 42:141–161.

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

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

Mickey, M. R., Dunn, O. J., and Clark, V. (1967). Note on the use of stepwise regression in detecting outliers. *Comput. Biomed. Res.*, 1:105–111.

Oja, H. (1983). Descriptive statistics for multivariate distributions. *Statist. Probab. Letters*, 1:327–332.

Øksendal, B. K. (1998). *Stochastic Differential Equations: An Introduction With Applications.* Springer-Verlag, Berlin Heidelberg, 5 edition.

Peña, D. and Prieto, F. J. (2001). Multivariate outlier detection and robust covariance matrix estimation. *Technometrics*, 43:286–301.

Press, W., Teukolsky, S. A., Vetterlling, W. T., and Flannery, B. P. (1992). *Numerical Recipes in Fortran.* Cambridge University Press.

Rey, W. J. J. (2001). On 100% multivariate breakdown point. In *International Conference on Robust Statistics*, Vorau, Austria.

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

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

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. In Grossmann, W., Pflig, G., Vincze, I., and Wertz, W., editors, *Mathematical Statistics and Applications*, pages 283–297. Reidel Publishing, Dodrecht.

Rousseeuw, P. J. and Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88:1273–1283.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection.* John Wiley and Sons.

Rousseeuw, P. J. and Molenberghs, G. (1993). Transformation of non positive semidefinite correlation matrices. *Communications in Statistics: Theory and Methods*, 22:965–984.

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

Rousseeuw, P. J. and Van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85:633–339.

Rousseeuw, P. J. and Yohai, V. J. (1984). Robust regression by means of S-estimates. In *Robust and Nonlinear Time Series Analysis*, volume 26, pages 256–272. Springer. Lecture notes in statistics.

Ruppert, D. (1992). Computing S-estimates for regression and multivariate location/dispersion. *Journal of Computational and Graphical Statistics*, 1:253–270.

Ruppert, D. and Carroll, R. J. (1980). Trimmed least squares estimation in the linear model. *Journal of the American Statistical Association*, 75:828–838.

Sen, P. K. and Puri, M. L. (1971). *Nonparametric Methods in Multivariate Analysis.* John Wiley and Sons, New York.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* John Wiley and Sons, New York.

Small, C. G. (1990). A survey of multidimensional medians. *Int. Statist. Rev.*, 58:263–277.

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

Stigler, S. (1986). *The History of Statistics.* The Belknap Press of Harvard University Press, Cambridge, Ma.

Tukey, J. W. (1960). A survey of sampling from contaminated distribution. In Olkin, I., editor, *Contributions to Probability and Statistics*, pages 448–485. Stanford University Press, Stanford, Calif.

Tukey, J. W. (1962). The future of data analysis. *Ann. Math. Statist.*, 33:1–67.

Tukey, J. W. (1975). Mathematics and picturing data. In *International Congress of Mathematics*, volume 2, pages 523–531.

Weisberg, S. (1985). *Applied Linear Regression.* John Wiley and Sons, New York.

Woodruff, D. L. and Rocke, D. M. (1993). Heuristic search algorithms for the minimum volume ellipsoid. *Journal of Computational and Graphical Statistics*, 2:69–95.

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

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

Yohai, V. J. and Zamar, R. H. (1988). High breakdown point estimates of regression by means of the minimization of an efficient scale. *Journal of the American Statistical Association*, 86:403–413.

Zamar, R. H. and Alqallaf, F. (2001). New contamination model for high dimensional data sets. In *Joint Statistical Meeting Conference*, Atlanta.