University of London

Robust Estimation of Multivariate Location and Scatter with Application to Financial Portfolio Selection

Simona Costanzo

London School of Economics and Political Science

PhD thesis

Department of Statistics, Houghton Street, London WC2A 2AE

December 2003

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

The thesis studies robust methods for estimating location and scatter of multivariate distributions and contributes to the development of some aspects regarding the detection of multiple outliers.

A variety of methods have been designed for detecting single point outliers which, when applied to groups of contaminated data, lead to problems of "masking", that is when an outlier appears as a "good" data. Robust high-breakdown estimators overcome the masking effect, also allowing for a high tolerance of "bad" data. The Minimum Volume Ellipsoid (MVE) and the Minimum Covariance Determinant estimator (MCD) are the most widely used high-breakdown estimators.

The central problem when identifying an anomaly is setting a decision rule. The exact distribution of the MCD and MVE is not known, implying that the diagnostics constructed as a function of these robust estimates have also an unknown distribution. Single point oultiers can be recognized using Mahalanobis distances; multivariate outliers are detected by robust (via MCD and MVE) distances of Mahalanobis type. The thesis obtains the small sample distribution of the first ones in an alternative simpler way than the proof existing in the literature. Furthermore, some empirical evidences show the need of a correction factor to improve the approximation to the expected distribution. Some graphical devices are suggested to enhance the results.

One of the limiting aspects of the literature on robustness is the lack of real data applications beside the literature examples. The personal interest in financial subjects has driven the thesis to consider applications in this area. Particular attention is paid to methods for optimal selection of financial portfolios. Mean-Variance portfolio theory selects the assets which maximize the return and minimize the risk of the investment using Maximum Likelihood Estimates (MLE). However, MLE are known to be sensitive to relatively small fractions of outliers. Furthermore, a wide financial literature provides evidence of the non-gaussian distribution of the stock

returns. All these reasons motivate the construction of a robust portfolio selection model proposed in the thesis.

To my friend Paolo

"Il mare immenso, l'oceano mare, che infinito corre
oltre ogni sguardo, l'immane mare onnipotente -
c'è un luogo dove finisce, e un istante - l'immenso mare,
un luogo piccolissimo e un istante da nulla."

Alessandro Baricco

Acknowledgements

During these years of research, when the motivation for completing the PhD has not always been "strong", I have found support and inspiration from many people.

I thank Sankarshan and George who made my start at the LSE smooth and enjoyable.

My thanks go to the whole LSE Statistics Department, with special mention to Dr Martin Knott and Dr Irini Moustaki, who have been an irreplaceable support. "Obrigada" to my friend and office mate Teresa, with whom I shared many difficulties encountered during research.

The project was sponsored by BSI (Banca della Svizzera Italiana) to whom I am immensely indebted. My thanks go particularly to Alberto Di Stefano, Andrea Laurent, Fabiano Cavadini, Fransiska Bignasca. Dr Fabio Trojani, from University of Lugano, Switzerland, gave an important contribution to the original idea for the project.

I am grateful to my family who has pushed me to follow my aspirations without any limitations.

I thank my encouraging friends from Italy: Anna, Bettina, Claudia, Giada. Thanks to Yasmine for taking care of social life in London when I was still a "fresher": the thesis was written despite her distractions.

Last but not least, my thanks are dedicated to my supervisor Prof. Anthony Atkinson who made my PhD thesis possible and represented a unique guidance from whom I have gained a valuable knowledge.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The word *robustness* is widely used in many fields of scientific research to signify the most diverse meanings. Scientific experiment are constructed according to a framework which determines the validity of the results. Sometimes these initial assumptions are too restrictive and do not match what happens in reality. In general, the results of an experiment are termed *robust* if they are not affected by changes in the initial framework. This section clarifies the meaning of *statistical robustness* and how it is related to outlier diagnostics.

Huber (1981) defines the word *robust* as

> *insensitive to small deviations from statistical assumptions.*

Hampel, Ronchetti, Rousseeuw, and Stahel (1986) restrict the concept of robustness in the following way:

> *robust statistics, as a collection of related theories, is the statistics of approximate parametric methods.*

According to Hampel's definition, robust statistics does not include semiparametric and non parametric models. These two models are generalizations of the parametric methods obtained by relaxing some distributional assumptions. Non parametric methods allow for a wider range of probability distributions than semiparametric ones. In other words, robust statistics is considered as a broader class of parametric statistics, including also "approximate" models, that are neighborhoods of parametric models where some initial assumptions do not hold. Although there is a fine line

between semiparametric and robust models, we believe that, in general terms, the former models allow for a wider range of distributions than the latter models. Furthermore, while semiparametric methods model all the data, the robust statistics reject the observations believed to be inconsistent with the reference model or, at least, reduce their importance through down-weighting. There are various implications on the properties of the robust and semiparametric estimators which can be studied referring to the specific literature: Bickel, Klaassen, Ritov, and Wellner (1993), Horowitz (1998), Powell (1994). For the reasons explained above, Hampel's definition seems more appropriate than the Huber's to describe the robust tools studied in this thesis.

The approximations of parametric models are determined by gross errors that are measurement or transcription errors, distributional mis-specifications, rounding or grouping, or by the presence of some correlation structure in the data. These errors generate *outliers* or "strange" observations. The outlier diagnostic literature is vast and includes many different approaches mainly depending on the model considered. Most of the studies concern regression models, although there is an increasing amount of work on time series and categorical data.

1.1 The Outlier Problem

The most common definition of an *outlier* is

> *an observation lying far from the rest of the data,*

although this is not sufficient to identify an anomaly. A remote observation is an outlier only if it is judged inconsistent with the remainder of the data. The purpose of statistical methods is to introduce some objectivity in the identification and treatment of the "strange" points.

Regarding the treatment of outliers, rejection of strange points is not always the optimal solution. The most common criticism from the "antagonists" of robust methods is that blind deletion could result in a loss of some relevant information. Simply, if these points are generated from errors in reading, recording or grouping,

they should be eliminated from the data. In other cases, when the anomalies derive from a different probability distribution or a different deterministic model, there are various approaches for treating the outliers other than plain deletion. A first approach consists in applying a transformation, when possible and appropriate, in order to adapt the model to the furthest points. A second approach is to reduce the importance of outliers through down-weighting.

Huber's data offers a simple and clear example of how even only one distant observation can influence the model's fit. The observations are only six for each of the two variables (Table 1.1). A linear model, $\mathbf{y} = \mathbf{X}^T\boldsymbol{\beta}$, where $\mathbf{y}$ is the (6×1) response vector and $\mathbf{X}$ the (6×2) matrix of carriers, including the dependent variable and a constant term, is fitted. Panel (a) suggests that there might be one outlier, observation 6, in the direction of the x axis. These types of outliers are called *leverage points.* It is also an influential observation since it changes the direction of the fitted line. Panels (b) and (c) are the Q-Q Normal plots of the residuals from a linear and quadratic fit: $y_i = \sum_{k=0}^{2} \beta^k x_i^k$, $i = 1, 2, ..., 6$. The second fit appears to be a better solution than the first one: the residuals lie closer to the straight line. A good result comes also from fitting a straight line after the deletion of the "bad" observation shown in panel (d). Further investigations are needed to decide between the two approaches. In addition, the data are not enough to show if the *extreme* point is an outlier or simply a sample variation of the data.

The meanings of some terms that will be recurrent in the thesis have been implicitly defined. An extreme observation is a point far from the rest of the data. This is an outlier if the distance is considered "unusual". There are different types of anomalies according to how these points are related to the model considered: leverage points and influential observations (sometimes called good and bad leverage points) are outliers for a regression model.

1.2 Contribution of the Thesis

The thesis studies robust modelling methods for estimating location and scatter of multivariate distributions and contributes to the development of some aspects

regarding the detection of multiple outliers. The computational work is substantial. Large use is also made of graphical tools, which are the most direct and simple approach to detect anomalies.

The identification of multivariate outliers is a particularly difficult topic to cope with. A variety of methods have been designed for detecting single point outliers which, when applied to groups of contaminated data, lead to problems of "masking" (meaning when an outlier appears as a "good" datum). Robust high-breakdown estimators overcome the masking effect, also allowing for a high tolerance of "bad" data. On the contrary, most of the robust statistics have breakdown at a fraction $1/(p+1)$ of contaminated data, where p is the dimension. Therefore, high-breakdown estimators are particularly useful in high dimensional sets. Many different methods have been offered by the literature as well as feasible algorithms for their computation. The Minimum Volume Ellipsoid (MVE) and the Minimum Covariance Determinant estimator (MCD) are the most popular ones (Chapter 3). The second one has better statistical properties than the first one, but its use has been limited by the lack of a fast and efficient algorithm. There are three main algorithms developed for the computation of MCD estimates: the FSA (Feasible Solution Algorithm) by Hawkins (1994) is computationally heavy and relatively slow; the Fast Algorithm (Rousseeuw and Van Driessen 1999) solves problems of speed; and the Forward Search for the MCD (Atkinson and Cheng 2000) applies a simple and efficient criterion. In addition to increasing the velocity of the algorithm on which various authors seem mostly focused, there are other computational aspects to be discussed. The first one is the choice of the size of the starting subset which is the outlier-free set. Including too few data in the initial set can compromise the efficiency of the estimates. However, if we start with too many data, including outliers, the result is a loss of robustness. We discuss this problem and propose some practical solutions.

Robust methods allow us both to find estimates for the location and the variability of a multivariate cloud according to robustness criteria and to detect groups of outliers at the same time. The central problem when identifying an anomaly is setting a decision rule. In this context, the distributional aspects of the robust diagnostics become very important. The asymptotic properties have been studied

in the literature, although the exact distribution of the MCD and MVE is not known. This implies that the outlier diagnostics constructed as a function of the robust estimates also have an unknown distribution. Single outlying points can be recognised using Mahalanobis distances as a diagnostic tool; multivariate outliers are detected by the robust (via MCD and MVE) distances of Mahalanobis type (Chapter 4). The thesis obtains the small sample distribution of the Mahalanobis distances in an alternative simpler way than the proof existing in the literature. Furthermore, some empirical experiments show the need of a correction factor for the approximation of the robust distances to their asymptotic distribution. Simulation envelopes, introduced for the first time by Atkinson (1985) in regression models, are found to be a valuable tool to detect outliers, overcoming the problems deriving from the unsatisfactory approximation of the robust distances by the theoretical distribution.

It has been noted that one of the limiting aspects of the literature on robustness is the lack of real data applications beside the canonical examples that are usually referred to by experts in the topic. My personal interest in financial subjects has driven the thesis to consider applications in this area. In particular, the attention is on methods for optimal selection of financial portfolios. The objective of these methods is to select the quota of the budget to invest in different financial assets (stocks, bonds etc.). Markowitz (1952) develops what is known as the Mean-Variance theory whose general and simple idea is to select the portfolio which maximizes the return and minimizes the risk of the investment, requiring estimates for the mean and variance-covariance matrix of asset returns. Markowitz (1952) considers Maximum Likelihood Estimates (MLE), known to be sensitive to relatively small fractions of outliers. Furthermore, a wide financial literature provides evidence of the non-gaussian distribution of the stock returns. Finally, there are motivations lying in the difference between *tactical* and *strategic* portfolios. Strategic portfolios are long-term portfolios, determining the general investment policy of the financial institution. Tactical portfolios are those trying to anticipate the market movements in the short term. From a strategic point of view, it is desirable that the composition of the portfolio does not vary much over time, mainly because of

the high transaction costs. All these points motivate the construction of a robust portfolio selection model as proposed in the thesis.

It has been mentioned that stock returns follow a non-gaussian distributional model. Although various authors disagree on the specific distribution, the common idea is that returns are longer-tailed than the Normal. The Student-t is suggested as a possible model. The advantages of using such a distribution compared to other long-tailed forms derive mainly from its simplicity and closeness to the Normal. The thesis explores the possibility of robust modelling using the multivariate Student-t and compares it with the high-breakdown optimizer. Some distributional aspects of the estimates are also discussed.

1.3 The Outline

Chapter 2 explores a financial data set which will be often used in the thesis.

The literature regarding robust estimators of multivariate location and scatter is reviewed in Chapter 3, with particular attention to the MVE and the MCD estimators. This Chapter also examines some computational aspects of the MCD method regarding the choice of the size of the initial set.

The detection of groups of outliers in multivariate data is studied in Chapter 4. The diagnostics used in the literature are reviewed and the distribution of Mahalanobis distances is analytically derived. The critical regions commonly used to detect outliers in a multivariate data set are quantiles of the Chisquare. It is shown that this approximation leads to the rejection of too many points, with a consequent loss in efficiency of the estimator. Analytical and empirical evidence are provided.

Chapter 5 studies an alternative way of modelling robustly via the Student-t distribution. Obtaining the MLE for a multivariate-t requires the use of the EM algorithm traditionally used when there are some missing observations in the data assumed Normal and non-contaminated. The EM can also be adapted to the framework of the multivariate t distribution. Finally, detection of multiple outliers is explored and the goodness of this model is compared in some applications with the high-breakdown estimator methods.

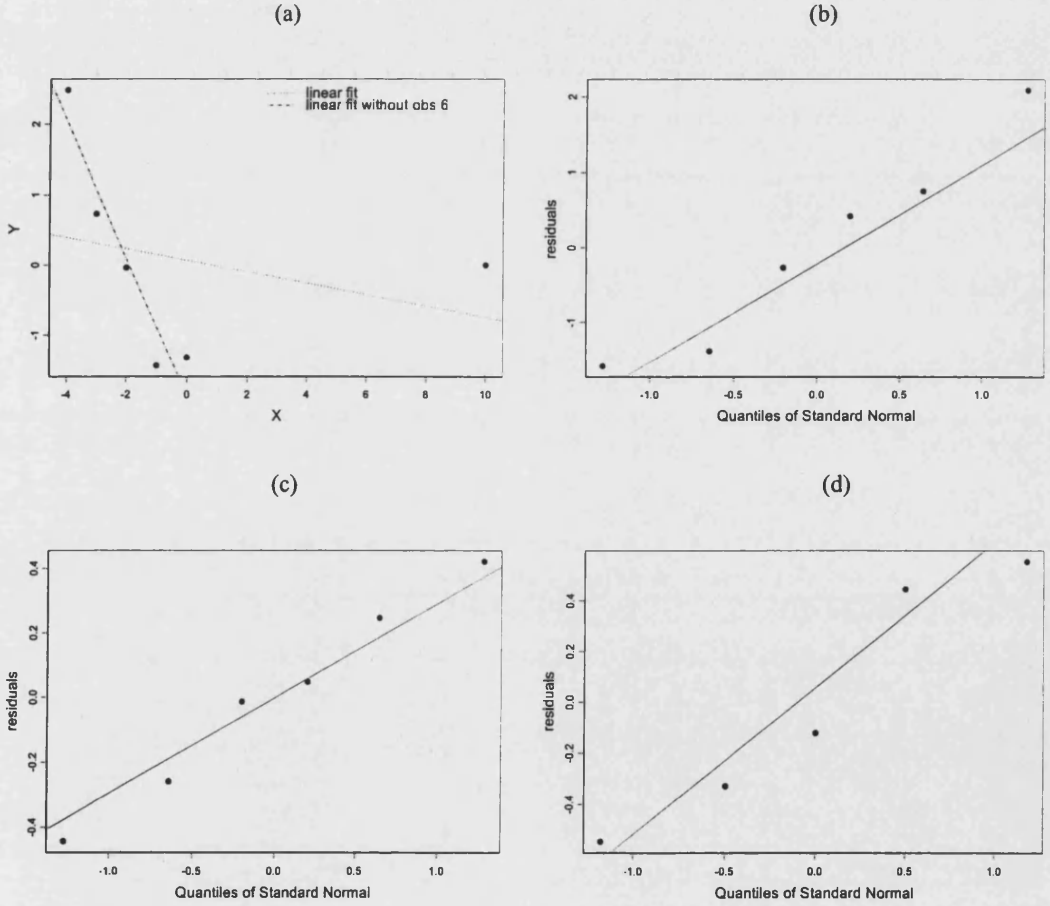
The robust construction of a model for optimisation of portfolios is proposed in Chapter 6. The Chapter provides analytical and empirical motivations for the use of a robust model. The performances of the model are analysed through a simulation study and an example using real data.

In Chapter 7, we summarize the results found in the thesis and propose ideas for further extension of the work.

Table 1.1: Huber's data (a) and residuals from the linear (lm), quadratic (qm) and linear without outlier fits (b)-(d).

observation	x	y	residual (lm)	residual (qm)
1	-4	2.48	2.09	0.25
2	-3	0.73	0.41	-0.26
3	-2	-0.04	-0.27	0.05
4	-1	-1.44	-1.59	-0.44
5	0	-1.32	-1.39	0.42
6	10	0.00	0.75	-0.01

Figure 1.1: Huber's data example: scatter-plot, (a), and Q-Q plots, (b)-(d).



Chapter 2

A Financial Data Set

2.1 Introduction

This Chapter introduces a data set coming from a real investment decision problem. The data is a recurrent example in the thesis. It consists of 10 monthly price indexes of the bond and stock markets. For each type of asset, there are indexes on 5 countries: United Kingdom, Japan, USA, Germany and Switzerland. The bond market indexes are provided by Salomon and the stock data by Morgan Stanley, with the exception of the stocks for Europe, produced by BSI (Banca della Svizzera Italiana). The observations are 175: from January 1985 to July 1999 and are expressed in local currencies.

The interest is on returns rather than asset prices. Therefore, each variable has been transformed applying differences of logarithms of the prices in two subsequent times:

$$y_t = \ln(p_{t-1}/p_t),$$

where y is the rate of return and p_t is the price at time t. The resulting observations are then expressed as annualized percentages, that is $y_t \times 1200$, where y_t is the return of one asset at time t. The currency of reference is Swiss Francs. The final data set is displayed in Table 2.1, Table 2.2 and Table 2.3, where the variable names are Datastream codes.

2.2 Data Description

There is a considerable amount of financial literature describing financial return data, particularly stocks. These are shown to be longer tailed than the Normal distribution and weakly autocorrelated.

Our data are indexes and, therefore, better behaved than series of single stocks. There are no significant autocorrelations of order one, although it is still possible to find some correlations of higher order or of some function of the initial variables (Figure 2.1). This means that the observations are not independent, although, as a first approximation, they will be treated as such.

Since we assume the assets are time independent, the scatter-matrix plot is a very useful graph to display the data structure and the relationship between pairs of variables. On the diagonal panels, there are the box-plots for each asset-market. The bond distribution is symmetrical and approximates quite well the Normal. The stock indexes are roughly symmetrical, but more scattered than the Normal distribution. These results are confirmed by looking at the Q-Q plots of Figure 2.2 and at Table 2.4. The skewness and kurtosis in Table 2.4 are computed using moment sample estimates:

$$\begin{aligned} \hat{\gamma}_1 &= \frac{1}{T}\sum_{t=1}^{T}\left(\frac{y_t-\bar{y}}{\hat{\sigma}}\right)^3, \\ \hat{\gamma}_2 &= \left[\frac{1}{T}\sum_{t=1}^{T}\left(\frac{y_t-\bar{y}}{\hat{\sigma}}\right)^4\right]-3, \end{aligned}$$

where $\bar{y}$ is the average return and $\hat{\sigma}$ is the sample standard deviation. Values around 0 indicate that the distribution is symmetric and mesokurtic. The kurtosis coefficient is positive for all the stocks, which means that the distribution is leptokurtic. The negative $\hat{\gamma}_1$ shows that the distribution of the stocks is also skewed to the left, which is confirmed by the Q-Q plots.

The panels in the off-diagonals are scatter-plots of each pair of assets. These show high correlation among the stock markets (the last five rows and column panels). The bonds appear less highly associated, although the pairwise correlation

coefficients are still significant (Table 2.5). The only exception is the Swiss bond index, which appears to be significantly related to the German bond one.

Both the scatter and the Q-Q plots evidence a couple of observations lying far from the bulk of the data. This confirms what we expected: because the data are monthly indexes (weighted averages of single stocks resulting from aggregations of daily data) there are only a few outliers. October 1987 represents the well known crash of the New York Stock Exchange, the largest drop of the returns after 1929, which affected most of European and Asian markets. August 1998 represents the Russian crisis, following the crash of the Asian markets, which extended to Europe and the US.

2.3 Comments

The description of the data confirms what it is well known in the literature on finance: stock returns are non Gaussian. Their distribution is rather longer tailed than the Normal. Because the data are indexes, there are only a few outliers. We leave further exploration of the data set to the following chapters.

Table 2.1: Monthly stock and bond indexes (y_t). Source: Data Stream, BSI.

LEGEND		
	Bonds	Stocks
Switzerland	SBSZ37L	MSSWITL(RI)
United States	SBUS37L	MISSAL(RI)
Germany, Europe	SBDM37L	STOXXEU50BSI
United Kingdom	SBUK37L	MUTUAL(RI)
Japan	SBJY37L	MESCAL(RI)

Date	SBSZ37L	SBUS37L	SBDM37L	SBUK37L	SBJY37L	MSSWITL(RI)	MISSAL(RI)	STOXXEU50BSI	MUTUAL(RI)	MESCAL(RI)
1/1/85	2.5174	56.2395	26.7255	-1.1508	25.4840	76.7025	126.1974	126.0549	62.4253	30.9413
2/1/85	-7.3270	59.6121	-0.5116	27.9873	44.0468	1.6221	93.2483	46.8113	6.8622	113.6219
3/1/85	7.8059	-91.7769	1.9017	68.4115	-54.6567	0.7427	-110.4207	0.8232	74.1913	-40.0159
4/1/85	9.1817	28.3743	8.9213	25.5755	3.6613	44.0614	-1.7541	32.5133	22.4297	-45.7103
5/1/85	7.3424	44.7257	22.6118	44.6741	6.9486	45.0671	61.6917	51.7512	62.6166	35.1587
6/1/85	1.8875	-3.1556	10.3021	15.7650	7.0215	71.0386	0.4176	5.7002	-68.5997	31.5318
7/1/85	5.5275	-122.3319	-5.7396	2.5853	-52.5109	22.5693	-120.5570	-47.6467	7.9975	-112.7739
8/1/85	10.6304	20.4906	23.1196	5.7987	4.8271	63.2446	-8.6431	49.4068	74.3063	24.6043
9/1/85	3.9478	-47.7692	8.5456	-33.7403	74.1518	-54.5447	-98.7550	-6.2000	-95.8873	69.1907
10/1/85	7.8568	-4.3161	-15.1854	10.9701	-30.7666	131.6516	28.3999	74.3282	93.0693	8.9719
11/1/85	6.8900	-14.2081	26.3975	3.2563	40.0093	92.0339	41.8992	71.0019	56.0589	-18.0677
12/1/85	9.2392	20.2493	32.9603	-52.4226	22.1545	110.9900	48.1399	97.0060	-62.0387	59.4697
1/1/86	6.9112	-10.0951	18.5053	-52.3144	36.1432	-62.0750	-11.3118	30.1307	-20.4120	12.6846
2/1/86	9.6766	-52.2324	18.0537	-6.5831	31.7353	-26.5818	2.0883	55.7405	38.3409	46.3278
3/1/86	3.6925	75.3216	6.2259	130.4371	71.7382	93.2789	102.4053	157.9020	163.8412	262.8500
4/1/86	-3.1323	-78.4317	15.3041	2.1916	2.4878	5.4800	-103.4301	55.0955	-28.2184	-24.8557
5/1/86	-1.1207	51.8362	-23.3005	15.8785	5.7216	-8.4883	134.9323	-43.7711	-26.5906	61.5331
6/1/86	5.4812	-53.7683	-13.3767	-45.2823	3.3250	-21.1030	-72.2133	-57.8128	3.4458	32.6218
7/1/86	7.3436	-64.4622	-8.5306	-120.3665	1.3274	-97.5118	-148.3373	-1.6194	-187.5721	64.7905
8/1/86	8.8414	7.1840	21.6926	-12.7498	-12.8619	121.6507	61.1531	127.0497	47.3946	78.5381
9/1/86	2.9693	-13.7094	8.8722	-94.4361	1.3402	-47.7390	-103.7615	-66.7450	-103.5142	33.3898
10/1/86	7.6643	66.0348	23.7020	23.4353	-4.5053	62.6992	114.3181	39.1111	73.3340	-120.5698
11/1/86	4.5752	-36.5426	10.0290	-23.1327	-31.9214	40.9853	-22.8931	26.4305	-10.4718	22.8446
12/1/86	6.8302	-20.9030	15.2072	50.4737	15.9695	4.1925	-53.2581	9.1941	48.0359	75.2455
1/1/87	8.8323	-40.1591	28.3492	2.4842	3.4985	-40.2852	97.9208	2.9757	76.0137	118.0991
2/1/87	1.0727	1.5449	2.0078	53.4525	19.2528	-48.4575	36.1847	-11.5287	126.4723	13.1643
3/1/87	1.1788	-32.0789	9.0302	57.1025	55.0598	19.8949	10.7237	46.4946	37.1977	104.7951
4/1/87	5.4504	-47.4943	-9.9093	37.5072	29.0170	-34.5531	-36.8510	-8.2672	51.4274	148.5117
5/1/87	4.4701	23.5136	22.3437	19.6423	7.4556	-15.0464	35.0990	-6.6545	95.4947	24.1267
6/1/87	4.2418	15.7630	-7.4444	-20.0085	-40.8588	66.1222	62.9618	25.9884	41.0719	-87.9216
7/1/87	4.3323	19.1541	-7.6244	-16.4290	-23.6679	97.7571	74.2822	58.3728	42.9237	-24.8097
8/1/87	0.4218	-40.3753	-6.1927	-28.0438	26.1372	-1.0693	9.3985	-3.9029	-61.3229	103.5110
9/1/87	-3.3788	13.3653	7.7136	47.7019	-42.3054	49.5485	7.9956	-14.9470	93.9164	-9.5807
10/1/87	15.9652	-44.7401	2.4287	35.5661	23.2435	-317.1861	-371.2046	-300.0344	-375.8926	-175.8801
11/1/87	10.0784	-67.3962	12.4906	19.8633	-2.2298	-140.9541	-179.9611	-117.3088	-131.5528	-30.6677
12/1/87	1.8609	-50.2876	-9.9306	-36.7722	54.2717	-36.8571	17.9660	-26.6282	81.5703	-46.1179
1/1/88	7.8257	120.0100	17.9760	26.7023	33.0493	8.5917	134.1123	-16.0263	73.9947	145.3859
2/1/88	6.2444	34.7220	30.7187	30.6374	22.8676	93.8231	70.1879	141.0486	13.1307	108.9576
3/1/88	-2.8623	-33.0469	3.0150	73.3404	30.0759	-40.0794	-64.7032	-10.3339	33.4560	64.7892
4/1/88	9.5821	21.6196	6.8236	22.1397	28.3713	16.5656	37.1650	56.7170	57.3257	37.4760
5/1/88	-4.9854	36.2990	3.2120	23.3551	28.2685	0.5100	52.6499	23.8271	5.9611	-13.0141
6/1/88	4.4776	72.5164	-9.5634	-56.0021	-34.0193	71.2829	106.2367	72.1785	14.2257	6.6693
7/1/88	-6.0079	36.7502	-7.2468	47.5168	54.0550	13.7037	37.9767	36.0878	52.1242	91.3999
8/1/88	4.0764	18.0340	20.1176	-13.9960	-19.4834	-4.4252	-22.1766	-14.1442	-61.3708	-70.9815
9/1/88	7.4040	21.5787	20.8966	28.5866	34.3703	44.5201	48.6225	66.6054	55.4695	53.8204
10/1/88	8.3631	-44.9204	17.9699	6.4333	31.9212	45.7415	-27.6607	41.0965	18.9351	23.2352
11/1/88	-0.5022	-53.1015	-11.4831	-10.9285	8.2572	-8.3811	-57.3025	-25.5191	-17.4812	69.6061
12/1/88	2.1076	37.6336	10.1822	29.5229	11.2157	35.9581	58.6403	64.4285	15.9630	45.3490
1/1/89	-15.6443	88.3585	-6.7835	65.0622	20.8954	35.0491	163.5538	61.1935	200.6722	71.2669
2/1/89	-5.0904	-41.5877	-11.1619	-43.9328	-18.1078	-5.8385	-70.4697	-42.2647	-62.7810	-5.7522
3/1/89	1.5294	85.2429	44.5729	48.6040	25.7342	70.0307	106.4610	93.6321	87.4240	39.6630
4/1/89	-6.2317	31.7881	26.7572	13.1862	7.9345	46.4882	71.3385	58.5770	27.9446	6.7431
5/1/89	-14.2184	51.8155	-41.9614	-67.7263	-63.2668	-44.4697	69.1123	-13.7817	-61.6701	-40.5063
6/1/89	22.8930	16.8962	10.5582	-26.1377	-38.6946	129.5777	-23.9030	19.3891	-17.3384	-85.2741
7/1/89	5.4785	-22.9003	18.4656	73.4627	26.0104	98.9113	52.8016	59.7982	121.1402	98.8254
8/1/89	-4.4621	35.5697	-2.0667	-7.2640	-4.4866	56.8080	77.9081	53.4257	30.1803	-25.9091
9/1/89	-8.1557	-42.9723	-0.1981	-34.6695	-7.2596	-45.4700	-50.8143	5.5785	-57.9136	26.0552
10/1/89	3.7789	22.6687	12.0546	-16.9582	-45.9371	-69.5103	-33.3292	-60.4224	-123.9919	-37.4865
11/1/89	4.2752	-5.5823	16.8399	-30.6052	-22.5087	50.5231	6.4574	61.9883	47.7854	40.7299
12/1/89	-3.7654	-34.5076	35.1148	19.1094	-47.7597	-3.2487	-9.6803	90.7779	76.8806	-35.3005
1/1/90	-19.5243	-42.7391	-40.7690	0.9838	-66.2231	-38.5172	-111.6611	-72.0638	-23.1126	-107.4553
2/1/90	-13.0204	-7.9146	-44.8386	-13.9292	-48.9009	-1.3200	8.2082	-45.6866	-39.7900	-139.3451
3/1/90	2.8244	3.6693	25.5509	-45.7239	-79.4369	-31.5565	31.0163	68.5869	-30.1044	-251.7675
4/1/90	2.9220	-41.3910	-31.5096	-52.3471	-32.2059	-55.5983	-59.6642	-35.8597	-117.4435	-23.4944
5/1/90	30.9701	12.2017	-19.1356	71.6775	61.5367	158.2145	90.2731	16.9241	146.9803	141.7156
6/1/90	9.5102	3.5268	23.9073	63.7768	-19.9337	13.1828	-18.4474	3.3448	55.2124	-73.4509
7/1/90	11.0341	-39.0535	16.6371	27.6824	-22.2763	-33.8072	-60.6308	-12.2919	-5.2352	-69.1638
8/1/90	-15.2740	-45.6508	-36.0385	-3.9804	-43.9572	-178.3026	-151.4519	-185.9663	-102.2297	-163.3113
9/1/90	10.4716	4.0599	0.5507	-16.2272	31.6484	-156.1686	-65.5450	-142.7399	-120.1802	-222.5033
10/1/90	4.7026	7.5406	45.0298	79.2005	112.4993	62.5396	-15.0996	92.9781	81.2722	249.5936
11/1/90	3.2908	12.0027	12.2366	17.3181	-19.9244	-30.5475	68.3879	-15.1847	53.5471	-158.0806
12/1/90	4.2745	16.0714	6.2457	-8.1745	-4.3264	1.4295	30.3640	-20.9269	-8.8624	53.8644

Table 2.2: Monthly stock and bond indexes (y_t). Source: Data Stream, BSI.

Date	SBSZ37L	SBUS37L	SBDM37L	SBUK37L	SBJY37L	MSSWITL(RI)	MISSAL(RI)	STOXXEU50BSI	MUTUAL(RI)	MESCAL(RI)
1/1/91	21.9269	-3.9340	13.7939	41.0428	39.4815	52.9025	39.3265	5.6836	28.3670	17.6252
2/1/91	12.6001	68.0483	48.6171	53.1921	63.4787	128.5694	144.0443	159.1326	151.7209	199.3105
3/1/91	10.1769	117.5825	-14.4970	9.7043	32.4183	52.6834	140.6609	22.5320	43.5458	33.3756
4/1/91	8.3836	9.3669	-2.4456	-10.1275	40.2245	18.8347	-0.3605	5.6784	-3.4825	34.3092
5/1/91	16.2190	33.3980	25.1549	16.2848	19.8578	55.9981	76.0381	54.1290	17.2482	14.5341
6/1/91	-1.4995	59.5751	7.3367	5.5342	61.2557	-52.1123	4.0409	-35.9040	-39.7596	-28.0129
7/1/91	4.2127	-9.4138	19.2985	45.4066	-0.5326	45.3765	29.3709	29.2199	110.4885	14.8732
8/1/91	1.1209	26.8407	13.3842	20.2864	27.1653	2.2096	30.5373	21.7813	30.6123	-67.5539
9/1/91	3.4495	-38.0199	6.5401	12.4674	-1.6380	-47.5795	-83.7940	-13.7424	-16.6955	37.6209
10/1/91	-20.2790	27.6196	18.0893	11.9623	42.8852	16.5434	33.6756	7.9016	-15.4272	69.4521
11/1/91	15.1478	-6.7882	15.6063	-9.3093	-8.0789	-40.4062	-73.0163	-31.4313	-83.0361	-117.6181
12/1/91	21.0377	-36.2929	32.6541	9.7774	1.6635	29.2244	59.8158	19.9956	29.4906	-32.8112
1/1/92	12.9751	47.1835	15.4033	38.0221	71.3415	70.6555	43.5998	64.7149	51.7595	4.5251
2/1/92	4.8976	48.8063	32.0834	40.7302	15.7643	52.1650	59.2678	66.3370	28.7556	-55.9177
3/1/92	-17.1379	5.1953	6.5941	-25.0192	-11.8886	-6.8599	-8.7066	-10.6213	-56.0910	-123.5319
4/1/92	6.0444	26.2500	12.9234	84.8490	8.3883	32.4244	47.6062	50.5300	151.8830	-63.8457
5/1/92	-7.7902	-27.7733	-5.7995	10.0079	12.2734	38.5728	-43.6499	4.4492	16.7102	44.9406
6/1/92	0.5516	-48.7045	-3.3493	-21.7630	-31.9153	-31.1015	-88.2796	-71.8240	-114.5559	-184.5274
7/1/92	9.7937	-23.0146	-12.3229	-44.1083	-34.2672	-31.8791	-0.5631	-92.6885	-98.0967	-65.3531
8/1/92	12.0631	-39.1341	17.5207	-25.0558	-3.6014	-32.9608	-85.1375	-33.7325	-60.7184	139.4384
9/1/92	35.9165	-3.2071	8.3360	-83.5481	21.5868	71.7125	-9.9935	-20.1949	-39.8383	-51.4374
10/1/92	30.7820	112.6141	55.7153	35.4416	108.7644	18.3612	139.9546	49.4111	28.3583	78.0555
11/1/92	2.2175	41.2529	20.2530	2.0637	43.3013	9.2962	88.6217	40.2151	69.6022	81.5370
12/1/92	25.0439	39.9399	19.1821	44.3983	25.2501	98.2701	36.0329	35.0072	67.7700	1.8945
1/1/93	20.3549	46.9948	38.3291	21.8148	36.3539	5.6619	28.0287	38.9504	-11.8021	13.8944
2/1/93	21.0639	52.9686	37.1928	-2.1188	119.0506	18.0247	47.9751	98.2292	8.4466	83.2823
3/1/93	16.4139	-21.0600	3.3796	44.6517	-9.2477	52.1360	-1.0829	21.0752	57.3984	133.0507
4/1/93	2.3039	-37.3904	-29.3933	-14.3655	-9.8542	-21.5688	-76.3183	-48.0170	-15.9964	143.8181
5/1/93	8.3839	-18.8368	-11.5344	-9.9605	28.6967	82.6888	16.3750	-4.5024	-4.6391	21.5920
6/1/93	8.5605	96.6417	6.3034	46.1016	100.2898	60.3306	79.6315	40.1611	51.3940	46.4118
7/1/93	3.5164	11.0311	-2.6354	24.4798	47.1443	19.0277	8.2510	46.3079	22.5018	93.0576
8/1/93	7.7776	-10.6073	29.8291	-10.1543	-13.1885	44.7237	9.6898	96.2049	41.7024	-3.6639
9/1/93	9.7285	-37.3349	-2.5037	-32.1277	-37.7084	-2.7720	-54.2091	-52.7550	-55.7860	-104.9878
10/1/93	13.1542	47.6051	34.7154	54.5854	38.2219	110.2156	66.8542	88.1237	92.9895	42.9878
11/1/93	4.9337	4.4010	-1.0732	24.9922	30.4619	16.5059	4.3971	-38.2517	16.3656	-204.2103
12/1/93	8.1530	-7.3468	-13.4754	10.1561	-28.9226	89.6746	2.0674	66.6025	77.1097	26.1969
1/1/94	3.6809	-0.5367	-18.7357	3.2728	-22.1113	91.1323	27.9900	32.6261	37.6657	164.7496
2/1/94	-9.6392	-52.5054	-25.0808	-63.7198	10.5419	-104.0258	-65.9328	-51.4971	-92.6517	22.0041
3/1/94	0.0756	-39.9929	12.0669	-49.4968	11.3829	-30.1108	-69.3077	-30.8732	-99.4447	-68.6742
4/1/94	-6.4437	-11.3629	3.3451	16.4360	11.3683	-17.7651	15.8843	64.0365	41.6422	49.0893
5/1/94	1.4434	-2.9270	-1.8714	-31.6590	-25.9560	-2.1580	15.7495	-54.6294	-68.6531	21.8959
6/1/94	-10.1402	-57.2593	-22.7361	-26.6974	-2.9768	-44.3287	-90.6291	-82.8653	-52.3929	1.3320
7/1/94	7.7843	23.8498	27.1842	11.6834	-12.8898	-20.8596	46.8435	87.1337	69.6596	-36.9499
8/1/94	-10.6206	-10.7067	-15.8629	-6.1810	-31.8856	35.7999	32.5321	9.0635	52.8644	-5.6673
9/1/94	0.7672	-52.4703	-26.9118	-12.8165	-8.4270	-46.1424	-64.1408	-90.9353	-95.4126	-68.5410
10/1/94	3.6758	-28.9137	12.9379	23.5042	-6.8441	-16.2289	-0.5184	25.3561	39.8302	4.5711
11/1/94	12.3991	56.9781	33.6911	37.3289	50.3564	40.1641	23.8946	20.0719	12.1501	3.5235
12/1/94	10.3982	-13.1320	-10.9907	-22.5869	-18.1163	11.9894	-0.2769	-10.5158	-16.6664	-4.1299
1/1/95	6.6586	-7.7371	15.7927	2.1810	-15.3986	-39.6648	2.4577	-21.9960	-48.6707	-102.3713
2/1/95	8.4013	-6.2844	20.5305	-32.1282	12.4414	37.0873	13.5930	5.8870	-26.2372	-92.3188
3/1/95	14.9464	-110.6032	-8.0610	-64.5568	60.7120	-41.6955	-84.8765	-50.1610	-29.5574	-11.1194
4/1/95	18.2267	32.5030	19.9577	14.7441	64.5406	40.2132	54.3022	62.6653	33.5900	74.3169
5/1/95	11.3330	67.8693	25.7853	46.7274	58.2921	82.0985	72.7430	31.7992	54.1096	-50.1508
6/1/95	6.6210	-7.5231	-1.0586	-28.4046	-17.5388	19.3661	13.5020	-12.3947	-14.6371	-74.5386
7/1/95	9.6170	-7.2888	14.6506	25.4001	-46.1877	6.8579	34.5021	53.1959	59.4283	86.2246
8/1/95	11.4257	71.0716	0.3353	38.9238	-78.7383	31.4837	61.0716	-15.0867	32.2574	11.6753
9/1/95	21.4341	-49.9817	-4.8565	-23.4955	-36.8644	42.1448	-5.1820	-33.3795	-22.3523	-47.8457
10/1/95	12.6112	2.2681	15.3322	0.3072	-54.3693	36.8083	-13.4278	-21.1067	-14.5326	-82.0277
11/1/95	15.3624	57.0753	32.9349	30.2079	62.9480	54.2326	90.5360	46.7269	45.9934	107.8038
12/1/95	0.8700	-11.3028	-1.4225	6.2972	-56.6384	22.8338	-5.4922	27.7351	7.6341	36.2700
1/1/96	2.1389	72.0617	26.0333	34.7536	11.1246	-13.4712	102.7184	87.2333	58.6238	44.6270
2/1/96	6.2611	-26.9539	-14.5237	-9.1568	-6.8029	41.1505	1.1264	8.9128	-1.7967	-32.9714
3/1/96	5.7659	-18.6518	-5.2909	-15.2476	-6.7407	93.0705	4.2119	9.8096	-14.7856	33.6232
4/1/96	8.9582	46.2847	20.3994	46.6077	64.1370	6.1404	71.3014	63.2542	76.6320	119.9641
5/1/96	-21.0520	5.1960	13.2138	49.1298	-16.5639	-33.1714	39.1767	31.4949	21.2408	-56.0033
6/1/96	5.3975	13.3221	2.3172	23.3763	-14.1927	58.1227	6.7339	13.5932	-7.6114	5.9333
7/1/96	8.4141	-50.6485	-1.3843	-42.7201	-22.8833	-77.5501	-107.2534	-76.5435	-45.1514	-108.3965
8/1/96	11.3676	0.6848	7.3778	17.1901	4.7412	52.0852	27.4343	8.5135	61.1670	-54.1154
9/1/96	19.5253	72.8455	36.2525	68.5260	32.1893	26.0392	120.3963	89.1661	82.2734	95.7540
10/1/96	9.2937	28.6858	23.6296	55.0110	-1.9412	1.8845	35.5038	24.2743	60.4621	-78.2250
11/1/96	-0.5748	57.2784	44.9292	93.7174	47.1452	51.4946	128.5312	112.3802	107.4351	63.6869
12/1/96	12.0125	26.3424	32.8688	55.3996	7.6341	13.1309	13.9569	61.5064	77.3750	-49.1412

Table 2.3: Monthly stock and bond indexes (y_t). Source: Data Stream, BSI.

Date	SBSZ37L	SBUS37L	SBDM37L	SBUK37L	SBJY37L	MSSWITL(RI)	MISSAL(RI)	STOXXEU50BSI	MUTUAL(RI)	MESCAL(RI)
1/1/97	9.5744	74.7019	11.7894	7.4767	29.8790	105.6006	149.6007	79.7345	34.2259	-68.2440
2/1/97	21.0821	42.7409	13.3528	75.6228	51.6899	58.2165	51.5837	62.4512	79.8997	70.7713
3/1/97	-9.8420	-37.8815	-21.9129	-31.6718	-56.2246	39.8694	-83.0292	11.7438	-15.7496	-68.3505
4/1/97	8.7318	41.9079	-5.3191	29.4856	-7.6113	56.1374	103.7971	8.0903	48.8434	69.9209
5/1/97	11.1777	-39.1830	-29.1036	-27.8197	43.1087	37.9584	17.5542	-2.1565	4.9549	77.0842
6/1/97	2.7477	47.6083	26.1180	62.7452	70.6149	130.5464	89.7072	121.3827	52.4102	122.6449
7/1/97	0.8536	71.6684	-15.8377	33.4068	18.4470	62.0040	135.4158	107.3446	98.8279	6.9011
8/1/97	-3.1124	-30.5702	-1.7055	-28.2977	-33.0915	-139.1065	-95.6406	-124.4578	-42.1713	-131.1415
9/1/97	0.8552	-10.5596	3.8755	-1.8417	-23.6806	91.3612	35.3007	87.7348	73.8932	-44.1136
10/1/97	-5.0176	-30.6692	-18.0236	-1.5019	-31.6283	-54.0375	-80.0428	-119.5148	-93.5355	-164.2579
11/1/97	6.6652	26.7846	1.3417	31.4593	-47.4335	62.3691	81.8804	40.5550	30.8210	-50.9106
12/1/97	16.8335	38.8069	16.5029	16.4291	5.6733	92.9430	45.2028	67.8332	71.6922	-43.1172
1/1/98	21.1583	33.5407	14.0026	25.7112	38.6261	59.0965	29.4819	51.6146	70.9260	116.5983
2/1/98	10.1187	-10.5272	9.0928	2.4684	6.8313	102.6929	75.0455	98.7466	74.0028	-0.1115
3/1/98	-9.8824	48.9587	26.4701	80.6916	-14.4133	71.6792	106.5087	136.1582	107.2659	-38.6966
4/1/98	-6.8714	-13.8997	16.1756	-10.3343	-0.4894	-25.3503	-5.3448	15.2088	-16.8920	-24.2427
5/1/98	7.8751	-8.1867	2.7037	-37.3874	-61.2522	33.5840	-41.2969	43.6164	-62.2084	-84.8164
6/1/98	-5.2641	39.3403	23.9509	46.5516	22.0854	31.1916	81.2923	31.0084	48.9689	47.5215
7/1/98	5.5001	-16.7537	4.7654	-33.7321	-62.7575	64.0437	-32.7405	6.1286	-45.1999	-36.6773
8/1/98	16.4671	-8.0656	-0.3846	21.7019	0.4843	-240.0942	-215.7832	-212.9149	-126.5056	-181.2601
9/1/98	10.2577	-17.9891	24.7959	-6.1049	3.5296	-160.9774	23.9937	-112.9134	-77.1837	-85.5727
10/1/98	8.7967	-24.4138	-14.3311	-34.1506	161.9383	136.6581	63.1604	64.2981	44.5205	159.8180
11/1/98	2.6322	26.0439	19.4500	28.7505	-36.6215	90.9050	111.7035	104.0860	85.1664	86.4998
12/1/98	2.2842	-11.6698	11.4512	15.8598	47.0799	23.0659	53.3349	54.1701	21.7299	30.1731
1/1/99	7.3937	46.1246	13.8651	36.7584	10.4095	11.3370	90.1631	5.6122	26.8214	48.1879
2/1/99	2.8317	0.1889	-28.0273	-20.8058	17.2463	-14.3851	-9.0182	-17.7416	51.2867	-1.4985
3/1/99	0.7351	34.0346	17.1439	42.0591	43.1069	12.4190	74.3244	32.0321	62.1392	181.4388
4/1/99	6.1462	37.0513	21.8782	24.5044	38.0270	31.1474	76.5675	80.6489	79.5188	83.1669
5/1/99	-5.8636	-9.8213	-19.1991	-7.6294	-16.0338	-59.2658	-25.7049	-49.7742	-65.7690	-66.9299
6/1/99	-19.8889	17.9177	-12.9776	-10.3837	9.5999	3.1424	85.4184	40.8547	25.6343	131.1979
7/1/99	-2.4159	-41.1099	-10.1395	-21.8949	14.3143	-1.6709	-88.0262	-41.2932	-30.4864	65.8171

Table 2.4: Annualized return data summary.

	SBSZ37L	SBUS37L	SBDM37L	SBUK37L	SBJY37L	MSSWITL.RI.	MSUSAML.RI.	STOXXEU50BSI	MSUTDKL.RI.	MSJPANL.RI.
Min:	-21.0520	-122.3319	-44.8386	-120.3665	-79.4369	-317.1861	-371.2046	-300.0344	-375.8926	-251.7675
1st Qu.:	0.8626	-28.3435	-5.5293	-16.3281	-18.1120	-20.9813	-23.3980	-15.1357	-26.4139	-49.6460
Mean:	5.0355	4.8028	7.2121	8.4417	7.6638	15.5536	13.5304	15.5727	13.1778	6.2621
Median:	5.4812	1.5449	8.9213	10.0079	6.8313	19.0277	17.9660	21.0752	26.8214	12.6846
3rd Qu.:	9.6468	35.9343	20.0376	34.0802	31.6919	56.4727	63.0611	58.4749	59.0260	64.7898
Max:	35.9165	120.0100	55.7153	130.4371	161.9383	158.2145	163.5538	159.1326	200.6722	262.8500
Std Dev.:	9.4766	43.1883	18.7356	38.1279	39.1750	66.7969	75.6998	65.3687	71.7114	87.4453
Skewness:	-0.1146	0.0932	-0.2734	-0.1257	0.4947	-1.2766	-1.0432	-1.0673	-1.0820	-0.1287
Kurtosis:	1.1806	0.0131	0.0079	0.5199	1.0360	4.1852	3.2067	3.3651	4.2681	0.4331

Table 2.5: Annualized return data correlations.

	SBSZ37L	SBUS37L	SBDM37L	SBUK37L	SBJY37L	MSSWITL.RI.	MSUSAML.RI.	STOXXEU50BSI	MSUTDKL.RI.
SBUS37L	0.1245								
	(0.1006)*								
SBDM37L	0.3754	0.3956							
	(0.0000)	(0.0000)							
SBUK37L	0.0906	0.4539	0.4138						
	(0.2330)	(0.0000)	(0.0000)						
SBJY37L	0.2705	0.3269	0.3390	0.3128					
	(0.0003)	(0.0000)	(0.0000)	(0.0000)					
MSSWITL.RI.	0.2221	0.3989	0.2330	0.2677	0.1723				
	(0.0031)	(0.0000)	(0.0019)	(0.0003)	(0.0226)				
MSUSAML.RI.	0.0687	0.7661	0.3433	0.3894	0.2922	0.6648			
	(0.3666)	(0.0000)	(0.0000)	(0.0000)	(0.0001)	(0.0000)			
STOXXEU50BSI	0.0845	0.4930	0.4975	0.3717	0.2870	0.7849	0.7353		
	(0.2664)	(0.0000)	(0.0000)	(0.0000)	0.0001	(0.0000)	(0.0000)		
MSUTDKL.RI.	0.0808	0.4562	0.3235	0.6257	0.2547	0.6688	0.7260	0.7363	
	(0.2876)	(0.0000)	(0.0000)	(0.0000)	(0.0007)	(0.0000)	(0.0000)	(0.0000)	
MSJPANL.RI.	0.0791	0.2425	0.1712	0.2999	0.5240	0.4043	0.3942	0.5031	0.4453
	(0.2979)	(0.0012)	(0.0235)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)

* p-values for the test on the correlation using Pearson coefficient.

Figure 2.1: Autocorrelation function for annual return data. Squared total return for Swiss bonds (a), squared total return for Swiss stocks (b), absolute total returns for Japanese bonds (c) and cube total returns for UK bonds (d).

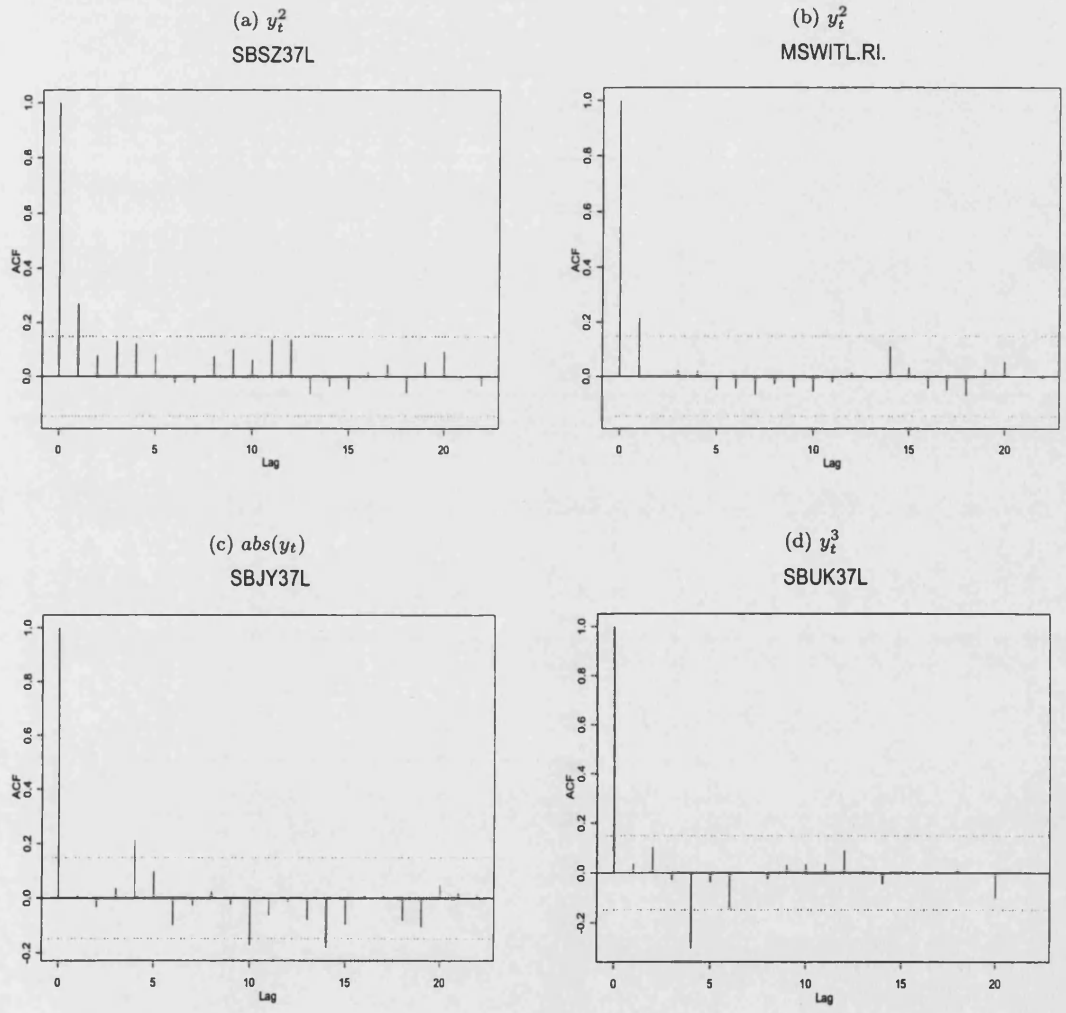


Figure 2.2: Q-Q plots of the annual returns. The panels on the left are the bonds indexes; those on the right are the stock data.

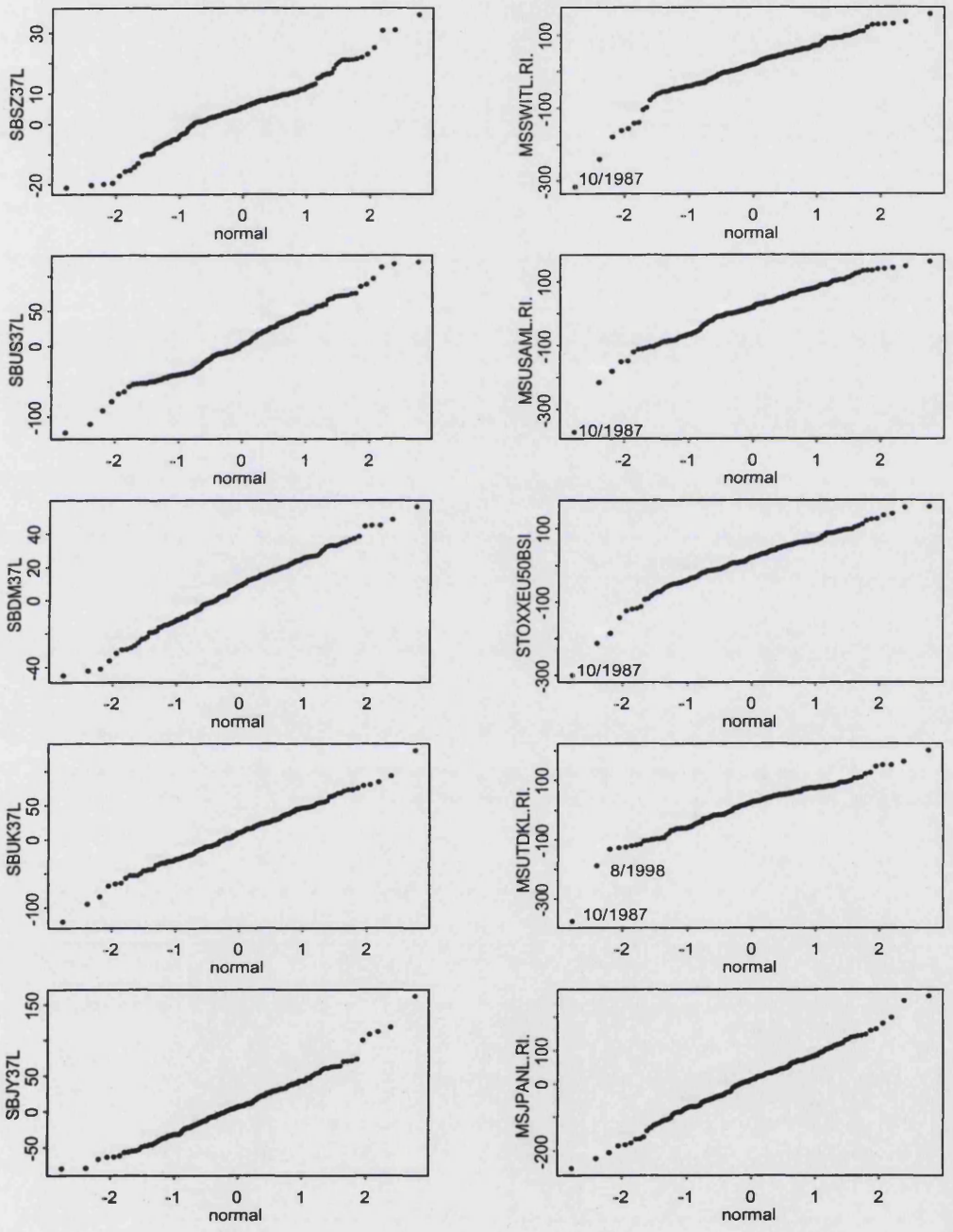
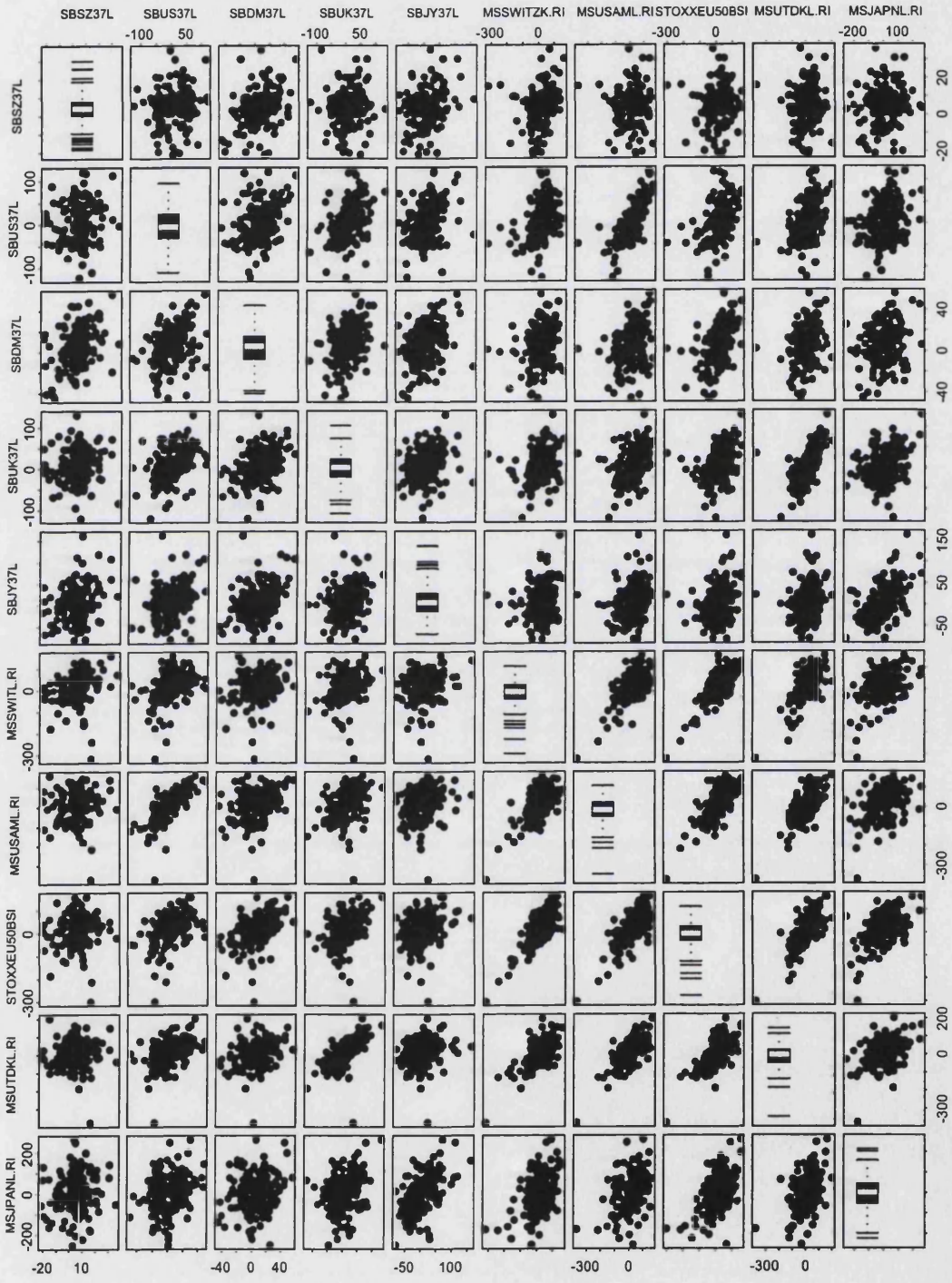


Figure 2.3: Scatter and boxplots of the annual returns.



Chapter 3

Background for Robust Multivariate Estimation and Computational Issues

3.1 Basic Concepts

The literature on robust statistics has produced a considerable amount of work on multivariate estimation of location and dispersion matrices, mainly because they allow for a wide range of applications. Let us consider a sample of n points, $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n\}$, independent and identically distributed, observed on a p-dimensional real space. The sample mean vector and the variance-covariance matrix of $\mathbf{Y}$ are:

$$\mathbf{y}(n) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i \tag{3.1}$$

$$\mathbf{S}(n) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \bar{\mathbf{y}}(n))(\mathbf{y}_i - \bar{\mathbf{y}}(n))^T, \tag{3.2}$$

where the notation $\bar{\mathbf{y}}(n)$ and $\mathbf{S}(n)$ emphasises that the estimates are calculated on a sample of size n. Multivariate data analysis techniques are based on sample estimates that are known to be "highly" sensitive to "small" fractions of outliers (Huber 1981). This motivates the search for robust alternatives.

The underlying class of distributions we are interested in includes only elliptical families (multinormal, cauchy, gamma, multivariate-t, etc.), although some of the

robust statistics that are here illustrated can be generalized to a broader class. Our objective is estimating robustly the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ characterizing the elliptical distribution $F_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$ with density of the form:

$$f(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma},g) = (\det(\boldsymbol{\Sigma}))^{-1/2} g\{(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\}, \tag{3.3}$$

where g is a non-negative function; $\boldsymbol{\mu} \in \Re^p$ and $\boldsymbol{\Sigma} \in PDS(p)$, the set of positive definite symmetric matrices of dimensions $p \times p$. These models are obtained starting from a spherically symmetric density:

$$f(\mathbf{x},g) = (\det(\boldsymbol{\Sigma}))^{-1/2} g\{\mathbf{x}^T\mathbf{x}\}, \tag{3.4}$$

where $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ and replacing $\mathbf{x}$ with $\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})$, that is applying some type of *affine* transformation. An affine transformation is a linear transformation in a multivariate space, equivalent to stretching, rotating or translating the axes. In the Normal model $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$, $\boldsymbol{\Sigma}$ is the variance-covariance matrix. For the whole class of elliptical distributions $\boldsymbol{\Sigma}$ is called *the pseudo-covariance matrix* or, even more generally, *the scatter matrix*, while $\boldsymbol{\mu}$ is the location.

Robust inference includes methods of estimation unaffected by outliers. There are some desirable properties that a robust estimator should satisfy: a condition for most robust estimators is *equivariance* under some transformations. If $T(\mathbf{Y}) \in \Re^p$ is a location estimate for $\mathbf{Y}$, T is *translation equivariant* if $T(\mathbf{Y}+\mathbf{b}) = T(\mathbf{Y}) + \mathbf{b}$ for all $\mathbf{b} \in \Re^p$, where $\mathbf{Y}+\mathbf{b} = \{\mathbf{y}_1+\mathbf{b}, \mathbf{y}_2+\mathbf{b}, ..., \mathbf{y}_n+\mathbf{b}\}$. For all non-singular $p \times p$ matrices $\mathbf{A}$ and for $\mathbf{b} \in \Re^p$, if

$$T(\mathbf{AY}+\mathbf{b}) = \mathbf{A}T(\mathbf{Y}) + \mathbf{b}$$

holds, then T is an *affine equivariant* estimator, where

$$\mathbf{AY}+\mathbf{b} = \{\mathbf{Ay}_1+\mathbf{b}, \mathbf{Ay}_2+\mathbf{b}, ..., \mathbf{Ay}_n+\mathbf{b}\}.$$

A *PDS* covariance matrix estimate $\mathbf{C}(\mathbf{Y})$ is affine equivariant if

$$\mathbf{C}(\mathbf{AY} + \mathbf{b}) = \mathbf{AC}(\mathbf{Y})\mathbf{A}^T$$

for all $\mathbf{b} \in \Re^p$. Although some robust estimators are not affine equivariant, this property is useful in many applications, for example, in robust principal component analysis, when one may wish to commute the estimator with the rotation and scaling of the axis.

Any estimator, robust and non-robust, should be consistent in the sense that

$$T(F) = \lim_{n \to \infty} T(F_n).$$

F is the underlying theoretical distribution and F_n is the empirical one (edf) defined by:

$$F_n(\mathbf{Y}) = n^{-1} \sum_{i=1}^{n} \Delta_{\mathbf{y}_i}, \tag{3.5}$$

where $\Delta_{\mathbf{y}}$ is the point mass 1 at $\mathbf{y}$, that is the probability distribution concentrated in $\mathbf{y} \in \Re^p$. Estimators are often represented as functions of the edf. Therefore, for $T_n = T_n(\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n)$ we can write:

$$T_n = T(F_n),$$

where T is computed assuming the model described by F_n. In robustness, consistency is usually intended according to Fisher's definition. Let T be an estimator and $\mathbf{Y}$ a sample whose underlying distribution is elliptical. T is said to be *Fisher consistent* if

$$T(F_{\boldsymbol{\mu},\boldsymbol{\Sigma}}) = (\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

This means that under the cumulative distribution function (cdf) $F_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$, the estimator asymptotically tends to the true population values.

Because robust estimators deal with mixed distribution models, it is difficult to know their behaviour in small samples. Therefore, the focus is on the asymptotic distribution. In many cases *asymptotic normality* is assumed:

$$\sqrt{n}(T(F_n) - T(F)) \xrightarrow{d} N(\mathbf{0}, V(T,F)),$$

where $\xrightarrow{d}$ indicates convergence in distribution and $\mathbf{0}$ is the p-dimensional null vector, $T(F)$ is Fisher consistent, and $V(T,F)$ is the asymptotic variance of $T(F_n)$. Asymptotic variances are generally computed using the expression:

$$V(T,F) = \int IF(\mathbf{y},T,F)IF(\mathbf{y},T,F)^T dF(\mathbf{y}).$$

IF is the influence function of the multivariate estimator T at the distribution F and is defined as follows:

$$IF(\mathbf{y};T,F) = \lim_{\epsilon \to 0} \frac{T((1-\epsilon)F + \epsilon\Delta_{\mathbf{y}}) - T(F)}{\epsilon}, \tag{3.6}$$

where $\epsilon \in [0,1]$.

We can rewrite (3.6) as:

$$IF(\mathbf{y};T,F) = \lim_{\epsilon \to 0} \frac{T(F + \epsilon(\Delta_{\mathbf{y}} - F)) - T(F)}{\epsilon},$$

which, for a finite sample of n observations, becomes:

$$\lim_{n \to \infty} n \left\{ T(F_{n-1} + \frac{1}{n}(\Delta_{\mathbf{y}} - F_{n-1})) - T(F_{n-1}) \right\},$$

since ϵ tends to 0 when the sample becomes infinitely large. Furthermore:

$$\begin{aligned} & n\{T(F_{n-1} + 1/n(\Delta_{\mathbf{y}} - F_{n-1})) - T(F_{n-1})\} \\ = \; & n\{T(1/n\Delta_{\mathbf{y}} + (n-1)/n\, F_{n-1})) - T(F_{n-1})\} \\ = \; & n\{T(F_n) - T(F_{n-1})\}. \end{aligned}$$

Provided that the limit exists (Hampel, Ronchetti, Rousseeuw, and Stahel 1986), if $\epsilon = 1/n$ and F is replaced by the empirical distribution F_{n-1}, we find the expression for the sensitivity curve that measures how much T changes when we add one observation $\mathbf{y}$ to the sample of size $n-1$. In other words, the influence function describes the stability of the estimates towards an infinitesimal ϵ at the point mass $\mathbf{y}$, where the effect of the contamination size is eliminated through standardization.

The efficiency of an estimator is computed by means of the asymptotic variance and, therefore, of the influence function. The efficiency of the diagonal element of a generic variance covariance matrix estimator $\mathbf{C}$ is defined by Hampel, Ronchetti, Rousseeuw, and Stahel (1986) as:

$$EFF(\mathbf{C}_{ii}, F) = \frac{1}{V_{ii}(\mathbf{C}, F) J_{ii}(\Sigma, F)},$$

where $i = 1, 2, ..., n$ and $J(T, F)$ is the Fisher information matrix. For a generic vector of parameters $\boldsymbol{\theta} = \{\theta_1, \theta_2, ..., \theta_p\}$:

$$J(T, F_{\boldsymbol{\theta}}) = E\left(\frac{\partial L}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial L}{\partial \boldsymbol{\theta}}\right)^T,$$

where $L = lnf(\mathbf{Y}|\boldsymbol{\theta})$ and $f(\mathbf{Y}|\boldsymbol{\theta})$ is the density of $\mathbf{Y}$ with respect to $\boldsymbol{\theta}$. However, the thesis will refer to a relative, rather than absolute efficiency. The asymptotic relative efficiency (AREFF) is defined as:

$$AREFF(\boldsymbol{\Sigma}_{ii}, F) = \frac{V_{ii}(\mathbf{C}, F)}{V_{ii}(\mathbf{S}, F)},$$

where $\mathbf{S}$ is the MLE under normality.

When constructing a robust estimator, the first condition to satisfy is that its influence function is bounded, so that the estimates do not explode ($\rightarrow \infty$) or implode ($\rightarrow 0$) at a contaminated distribution. The upper bound of the asymptotic bias is defined as *gross error sensitivity*:

$$\gamma^* = \sup_{\mathbf{y}}\{\|IF(\mathbf{y}, T, F)\|\},$$

where $\|.\|$ is the Euclidean norm. The maximum fraction of outliers that can be tolerated by T before the asymptotic bias becomes unbounded is *the breakdown-point* of the estimator:

$$\epsilon_n^*(T, \mathbf{Y}) = \min\left\{\frac{m}{n};\ bias(m; T, \mathbf{Y}) \text{ is infinite}\right\},$$

where m is the number of outliers. In other words, the breakdown is the smallest fraction of contamination that can cause T to take values arbitrarily far from $T(\mathbf{Y})$.

Under the point of view of global robustness, that is the general stability of the model, given any neighbourhood of a generic model P under a metric d, the breakdown point $\epsilon_n^*(T, \mathbf{Y})$ is defined as:

$$\epsilon_n^*(T, P, d) = \inf\left\{\epsilon > 0 : \sup_{P' \in B_\epsilon(P,d)} \|T(P')\| = \infty\right\},$$

where the expression $B_\epsilon(P, d)$ defines the neighbourhood of P under the metric d as the following:

$$B_\epsilon(P, d) = \{P' : d(P, P') < \epsilon\}.$$

3.2 M and S-Estimators

The idea behind constructing a robust estimator is weighting the remote observations to obtain a bounded asymptotic bias. Historically, the first class of multivariate robust estimators is a generalization of the maximum likelihood estimators for location and scatter $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ of elliptical distributions, (3.3), introduced by Maronna (1976). It is presented as the solutions $\mathbf{t} \in \Re^p$ and $\mathbf{V} \in PDS(p)$ to the simultaneous set of equations:

$$\mathbf{t} = \frac{\sum_{i=1}^n v_1(MD_i)\mathbf{y}_i}{\sum_{i=1}^n v_1(MD_i)}, \tag{3.7}$$

$$\mathbf{V} = \frac{\sum_{i=1}^n v_2(MD_i)(\mathbf{y}_i - \mathbf{t})(\mathbf{y}_i - \mathbf{t})^T}{\sum_{i=1}^n v_3(MD_i)}, \tag{3.8}$$

where v_1, v_2 and v_3 are real valued non-negative functions and

$$MD_i = \{(\mathbf{y_i} - \mathbf{t})^T \mathbf{V}^{-1} (\mathbf{y}_i - \mathbf{t})\}^{1/2}$$

is the Mahalanobis Distance of the i^{th} observation, $i = 1, 2, ..., n$.

The class of S-estimators for multivariate location and scatter was introduced for the first time by Davies (1987), who extended the idea of the S-estimators for regression (Rousseeuw and Yohai 1984). They are defined as the solution $(\mathbf{t}, \mathbf{V}) \in \Re^p \times PDS(p)$ to:

$$\min \det(\mathbf{V}), \tag{3.9}$$

subject to

$$\frac{1}{n} \sum_i \rho(MD_i) = b_0, \tag{3.10}$$

where $b_0 > 0$ and ρ is a symmetric, continuously differentiable and non decreasing function of MD_i (Lopuhäa 1989).

The properties of M and S-estimators have been widely discussed in the literature. Hampel, Ronchetti, Rousseeuw, and Stahel (1986), Lopuhäa (1989) and Tyler (1991) are some of the main references. The uniqueness of a solution for the system of equations (3.7)-(3.8) has been shown in Maronna (1976) and Huber (1981) only for $v_3 = 1$. This is consistent at a $n^{-1/2}$ rate and asymptotically Normal. M-estimators are locally robust since their influence function is bounded, but have a breakdown of $\epsilon = 1/(p+1)$, which means they cannot tolerate many outliers in high dimensions.

S-estimators are also asymptotically Normal and $n^{1/2}$ consistent, but their breakdown is higher than the M-estimators, reaching asymptotically 50% of the data.

It can be shown (Lopuhäa 1989) that the solution to the minimization problem for the S-estimators satisfies also the system of equations (3.7) and (3.8). In other words, S-estimators satisfy the first order conditions for the M-estimators as defined

in Huber (1981). This result might lead one to believe that any solution of the minimization problem (3.9)-(3.10) is an M-estimate. However, M-estimators have generally low breakdown points, which is not the case for the S-estimators. The difference lies in the specification of the weights $v(.)$. Huber (1981) chooses $v_2 \geq$ to be monotone and v_3 to be equal to a constant to prove the uniqueness of the solution for the M-estimator equation set. However, (3.9) and (3.10) do not satisfy those restrictions, as shown in Lopuhäa (1989). In other words, S-estimates are a solution (not the only one) with high breakdown points, that satisfy simultaneously the system of equations (3.7)-(3.8) and the minimization problem (3.9)-(3.10).

3.3 The MVE and MCD Estimators: the General Idea

If ρ is taken as an indicator function assuming values $\{0, 1\}$, (3.9) and (3.10) define what in the literature is known as the Minimum Volume Ellipsoid (MVE) estimator. Introduced for the first time by Rousseeuw (1983) and Rousseeuw (1985), the MVE is described as:

$$T(\mathbf{Y}) = \textit{centre or scatter of the ellipsoid with minimum volume covering at least } h \textit{ points of the sample } \mathbf{Y}.$$

The volume of a dataset is the square rooted determinant of its scatter matrix (Woodruff and Rocke 1994). If M is a subset of size m, where $m = p + 1, ..., n$, generating the sample estimates $\bar{\mathbf{y}}(m)$ and $\mathbf{S}(m)$, the volume of the ellipsoid covering h points is:

$$V(h) = \{\det(\mathbf{S}(m)) \times m_M^p\}^{1/2}, \tag{3.11}$$

where m_M is a correction ensuring that h points are included in the set with covariance $\mathbf{S}(m)$ and mean $\bar{\mathbf{y}}(m)$: m_M is the h-th order statistic of the squared Mahalanobis distance $MD^2_{[h]}$, the smallest h-th distance for all the ellipsoids of size m. In

general, the squared Mahalanobis distance of the k-th observation from the subset of m points is given by:

$$MD_k^2(m) = (\mathbf{y}_k - \bar{\mathbf{y}}(m))^T \mathbf{S}^{-1}(m)(\mathbf{y}_k - \bar{\mathbf{y}}(m)), \tag{3.12}$$

where $k = 1, 2, ..., n$.

Rousseeuw (1983) defines h, the proportion of the data for which the MVE has the maximum breakdown, as $h = [n/2] + 1$, where [.] is the largest integer part of n/2. Later, Rousseeuw and Leroy (1987) show that the MVE has maximum breakdown when $h = [(n+p+1)/2]$, which asymptotically reaches half of the data.

If H is the subset of $\mathbf{Y}$ with minimum volume, then the MVE estimates are the sample estimates for H:

$$\bar{\mathbf{y}}_{mve}(h) = \frac{1}{h}\sum_{i \in H} \mathbf{y}_i \tag{3.13}$$

$$\mathbf{S}_{mve}(h) = c \times \{\frac{1}{h}\sum_{i \in H}(\mathbf{y}_i - \bar{\mathbf{y}}_{mve}(h))^T(\mathbf{y}_i - \bar{\mathbf{y}}_{mve}(h))\}, \tag{3.14}$$

where the covariance is scaled with the constant c to attain consistency to the multivariate Normal distribution. The problem of scaling the estimates will be broadly examined in Chapter 4.

Similarly to the MVE, the Minimum Covariance Determinant (MCD) estimator (Rousseeuw 1983 and Rousseeuw 1985) is the centre and scatter of the data of size h whose covariance matrix has minimum determinant.

The objective function to be minimized is the same as for the MVE. The difference is that the resulting covariance estimate for the MCD is constrained to cover 50% of the data, rather than the covariance defining an ellipsoid covering the half. This implies that, when the fraction of outliers $\epsilon \to 0$, the MVE tends to the centre (or covariance) of the smallest ellipsoid covering all the data, whereas the MCD converges to the sample estimates.

As far as the breakdown point of the MCD is concerned, it is the same as for the MVE, that is $\epsilon = ([n/2] - p + 1)/n$.

3.4 The MVE and MCD Estimators: Some Properties

Davies (1992) and Nolan (1991) study the asymptotic properties of the MVE. The first work raises an interesting point of discussion regarding the existence and uniqueness of the MVE solution, although the main result concerns its asymptotic behaviour. The MVE solutions for location and scatter are shown to be consistent for the population values, although they converge weakly, at a rate of $n^{-1/3}$, to a non-gaussian random process. The consequence of this result affects the efficiency of the estimates: small perturbations in the data determine large perturbations in the estimate Maronna and Yohai (1998). This characteristic makes the MVE less attractive than the MCD.

For the theoretical properties of the MCD we will mainly refer to three works: Croux and Rouseeuw (1992), Butler, Davies, and Juhn (1993) (BDJ) and Croux and Haesbroeck (1999).

BDJ shows that for elliptical symmetric distributions, defined as in (3.3), where the derivative g' is assumed to be strictly negative in order to have a unimodal distribution, the MCD problem has a unique solution, given by the ellipsoid:

$$E = \{\mathbf{y} : (\mathbf{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \leq q_\epsilon\},$$

where $\epsilon \in [0, 1]$ is the fraction of the data not determining the MCD: q_ϵ is chosen so that the probability of being included in the ball of centre 0 and radius $\sqrt{q_\epsilon}$ is $1 - \epsilon$, where $h = n(1 - \epsilon)$.

In addition, BDJ also shows that the MCD estimates satisfy:

$$\begin{aligned} T(F_{\boldsymbol{\mu},\boldsymbol{\Sigma}}) &= \boldsymbol{\mu}, \\ \Sigma(F_{\boldsymbol{\mu},\boldsymbol{\Sigma}}) &= \Big(\frac{c_\epsilon}{1-\epsilon} \int_{z^T z \leq q_\epsilon} z^2 dF_{0,I}(z)\Big)\boldsymbol{\Sigma}, \end{aligned}$$

where for a bounded measurable set A:

$$T_A(F) = \frac{\int_A y dF(y)}{1-\epsilon},$$
$$\Sigma_A(F) = \frac{\int_A (y - T_A(F))(y - T_A(F))^T dF(y))}{1-\epsilon}.$$

To obtain Fisher consistency at the theoretical distribution F, c_ϵ is chosen so that $\Sigma(F_{\mu,\Sigma}) = \Sigma$, that is:

$$c_\epsilon = \frac{1-\epsilon}{\int_{z^T z \leq q_\epsilon} z^2 dF_{0,I}(z)}.$$

In addition to the uniqueness and consistency of the MCD solution, BDJ show that both the location and the scatter estimators converge weakly to a Gaussian random variable and obtain a form for the asymptotic variance of the first one.

Croux and Haesbroeck (1999) obtain a form for the influence function of both location and scatter estimators. The IF for location in the univariate case was already derived by Croux and Rouseeuw (1992). Furthermore, they complete the results of BDJ by obtaining an expression for the asymptotic variance of the MCD scatter matrix. This last is used to compute the asymptotic efficiency of the estimator.

It is noted that the efficiency varies depending on both the dimension p and the trimmed proportion of the data ϵ. In the paper it is shown that increasing the fraction trimmed the efficiency decreases when the data do not contain outliers. The loss is particularly high for low dimensions and when $\epsilon \approx 0.5$, the maximum breakdown. This motivates the authors suggestion of considering $\epsilon = 0.25$ as "a compromise between robustness and efficiency", according to a rule of thumb.

The trade-off between global reliability and accuracy of robust estimates has been the object of several studies. Extensions of S-estimators, such as τ-estimators (Lopuhäa 1991), M (CM)-estimators (Kent and Tyler 1996) and compound estimators (Woodruff and Rocke 1994) manage to combine high-breakdown with efficiency. Another way of keeping a high breakdown while improving the efficiency is to reweight the estimates. Rousseeuw and Van Driessen (1999) use a weighting function w_i defined as follows:

$$w_i = \begin{cases} 0 & \text{when} \quad RD_i^2(h) > \chi^2_{p,\,0.975} \\ 1 & \text{otherwise,} \end{cases} \tag{3.15}$$

where:

$$RD_i^2(h) = (\mathbf{y}_i - \bar{\mathbf{y}}_{mcd}(h))^T \mathbf{S}_{mcd}^{-1}(h)(\mathbf{y}_i - \bar{\mathbf{y}}_{mcd}(h)) \tag{3.16}$$

is the Mahalanobis distances computed on the MCD estimates; $\bar{\mathbf{y}}_{mcd}(h)$ and $\mathbf{S}_{mcd}(h)$ were defined in the previous paragraph, where $h = n(1-\epsilon)$. The reweighted estimates are given by:

$$\begin{aligned} \bar{\mathbf{y}}_{mcd}(h) &= \frac{\sum_{i=1}^n w_i \mathbf{y}_i}{\sum_{i=1}^n w_i} \\ \mathbf{S}_{mcd}(h) &= c \times \left\{ \frac{\sum_{i=1}^n w_i (\mathbf{y}_i - \bar{\mathbf{y}}_{mcd}(h))(\mathbf{y}_i - \bar{\mathbf{y}}_{mcd}(h))^T}{\sum_{i=1}^n w_i - 1} \right\}. \end{aligned}$$

Lopuhäa (1999) derives the asymptotic properties of the MCD reweighted estimates and in Croux and Haesbroeck (1997) the asymptotic efficiencies of the reweighted estimates, sample and S-estimates are compared in a simulation study.

3.5 MCD Estimator: Computation

The literature on robust estimators has produced a large amount of work on the search for fast and efficient algorithms. Woodruff and Rocke (1994) underline the importance of the computational aspect by writing that "an algorithm is an estimator itself", implying that using different algorithms is like using different estimators. Woodruff and Rocke (1994) refer to probabilistic algorithms, for which there is not a unique solution, but a set of feasible ones. For example, the "ideal" algorithm for the MCD would select all the possible subsets of size h and retain the solution with the lowest determinant. For computational cost reasons, an algorithm of this type cannot be used in practice unless very small sample sizes and dimensions are

considered. As a consequence, the global optimum is replaced by an approximate solution, chosen, within a feasible set, according to different possible criteria.

The MCD has better theoretical properties than the MVE and only recently the problems concerning its computation have found a solution. This section describes the main algorithms proposed in the literature.

The framework is a mixed distribution model with a sample $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n\}$ of p-dimensional observations where there are h "good cases" following a multinormal distribution, $N(\mu, \Sigma)$, and a remaining $n - h$ (the outliers) belonging to some other model.

The Feasible Solution Algorithms (FSA) for the MVE and MCD (Hawkins 1993 and Hawkins 1994) provide a computationally efficient solution obtained from a combinatorial algorithm. The FSA for the MCD starts from a trial subset of size h and retains its mean and variance-covariance matrix. In a following step, the algorithm selects all the possible pairwise swaps with observations from subsets of size $n - h$ and chooses the set whose scatter matrix has minimum determinant. The chances of getting the global optimum solution are increased by considering different initial choices. The algorithm has been recently refined by considering sorted Mahalanobis distances in order to identify the minimum determinant subset (Hawkins and Olive 1999).

Atkinson (1994) and Atkinson and Cheng (2000) propose a Forward Search (FS) algorithm for the MVE and MCD. The search starts from a random subset M of size $m = \alpha \times n$, where $0 < \alpha < 1$. Usually $m = p + 1$, the smallest size allowing a non-zero covariance matrix determinant. At each step, M is incremented by s observations, usually $s = 1$. The added observations are those having the smallest Mahalanobis distances, computed with respect to the location and scale of the M set. Grübel (1988) shows that the smallest Mahalanobis distances characterize the subset with minimum covariance matrix determinant. The FS was born as a algorithm to detect groups of outliers, opposing the backward selection methods, which can generate problems of "masking". At each step, the outliers are detected by the largest Mahalanobis distances.

Rousseeuw and Van Driessen (1999) propose a Fast algorithm for the MCD. The

kernel of the algorithm is the C-step:

- *Step 0*: start from a subset H_0 of size h, with $h \leq n$, whose location and scatter are given by $(\bar{\mathbf{y}}_0(h), \mathbf{S}_0(h))$. The H_0-set is constructed from a subset M of size $m = p + 1$, where $\det(\mathbf{S}(m)) > 0$. The Mahalanobis distances are computed as in (3.12), where $k = 1, 2, ..., n$, and sorted. The H-set will include the observations with minimum distances:

$$\{MD^2_{[1]}(m), MD^2_{[2]}(m), ..., MD^2_{[h]}(m)\},$$

  where $MD^2_{[i]}(m)$ is the i-th smallest distance calculated on the basis of sample estimates on a set of size m.

- *Step 1*: update the H_0 set by calculating:

$$MD^2_k(h) = (\mathbf{y}_k - \bar{\mathbf{y}}_0(h))^T \mathbf{S}_0^{-1}(h)(\mathbf{y}_k - \bar{\mathbf{y}}_0(m)),$$

  with $k = 1, 2, ..., n$. The smallest h-th distances characterize the updated set H_1.

- *Step 2*: repeat *Step 1* many times until there is no longer a significant reduction in the determinant of $\mathbf{S}$.

Steps 0-2 are repeated for different random choices of M. The algorithm also treats cases of "exact fit", that is when at least h observations lie on a hyperplane, which implies that the MCD covariance matrix is singular. If the sample size is very large (usually $n > 600$) the velocity of the algorithm is increased by using a nested structure similar to the "Branch and Bound" algorithm by Agullò (1996). The "Branch and Bound" structure starts from a set of N initial solutions. The algorithm is run simultaneously on the N sets, which at the next step are reduced to $N - K$; the algorithm repeats until reaching a unique solution.

Some computational issues regarding the MCD are worth a deeper discussion.

3.5.1 Forward Search for the Choice of the Initial Set

The computation of the MCD or MVE requires the choice of h, the number of observations to be included in the "good set". Ideally, we should know in advance the number of outliers in the data, which is very unlikely to happen in practice.

Usually, there are two possible ways of setting h. The first one considers the "worst case scenario" , that is when the breakdown is maximum. Therefore, the subset size is $h = [(n + p + 1)/2]$. The alternative consists in using some prior information. For example, previous experiences on similar sets of data could suggest that no more than a fixed proportion of outliers can appear. The first method has a drawback: if too many data are trimmed, leaving out some "good" observations as well as outliers, the resulting estimates would lose efficiency while still being consistent. On the other hand, if h is too big, the estimates would, as a result, be biased by the inclusion of some outliers in the "good set". The trade-off between robustness and efficiency has been long discussed with no real solution.

We suggest a third method, consisting in comparing the stability of the estimates for different choices of h, according to a Forward Search approach. Figure 3.1 (a) and (b) display how changes in the size of the h affect the robust estimate of the scatter matrix, under two possible contamination sizes. The simulation study generates 100 replicates of size 100 from a mixture of two bivariate normals with shifted mean and different covariance structure:

$$\mathbf{Y}_{good} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix} \begin{bmatrix} .5 & 0 \\ 0 & .5 \end{bmatrix}\right)$$

and

$$\mathbf{Y}_{outlier} \sim N\left(\begin{bmatrix} 0 \\ 5 \end{bmatrix} \begin{bmatrix} .5 & .2 \\ .2 & .5 \end{bmatrix}\right).$$

At each iteration the size of the good set is incremented by one observation. The graph shows that the variability of the robust scatter matrix does not inflate when the proportion of fitted data does not exceed the "good set" of observations. In addition, the determinants remain stable even when few outliers are included in the fit.

Having chosen h, the next step would be deciding how the H-set is selected. The FSA simply performs a random selection, while both in the Fast-MCD and FS, the algorithm starts from a random M-set of small size, usually $m = p + 1$. Following the last method, the M-set is more likely to be outlier-free. The size of M is then increased by choosing the h observations with the smallest distances from the M-set. The second approach is more sensible, although it poses the problem of choosing m. We suggest the use of the FS algorithm to show the differences in the estimates for increasing dimensions of M. The procedure works as follows:

- *Step 0* Choose an initial subset of size $m + 1$
- *Step 1* Estimate the sample mean and covariance, $\bar{\mathbf{y}}(m)$ and $\mathbf{S}(m)$
- *Step 2* Compute Mahalanobis distances as functions of $\bar{\mathbf{y}}(m)$ and $\mathbf{S}(m)$
- *Step 3* Choose H the set of the h-th smallest sorted distances
- *Step 4* Take as MCD estimates the centre and covariance of the H-subset
- *Step 5* Repeat steps 1-4 for $m = m + 1$ until $m = n$

Figure Figure 3.2, panels (a) and (b), plots the determinants of the H-sets resulting from the above procedure. The samples are generated in the same way as in the previous example. The plot shows that the determinants converge to a unique solution when m increases approaching n. Therefore, the smaller is m the more different are the initial H-sets. When the initial subset size approaches that of whole sample, the random choices of M become similar in variability, leading to the same choice of the H-set. The reason for the increase in the determinants after the convergence of the algorithm is that the outliers, being the observations with largest distances, are included in the H-set only at the end, producing the change in the variability visible at the right end side of the plot. The convergence to the unique solution is faster for a smaller number of outliers.

3.6 Conclusions

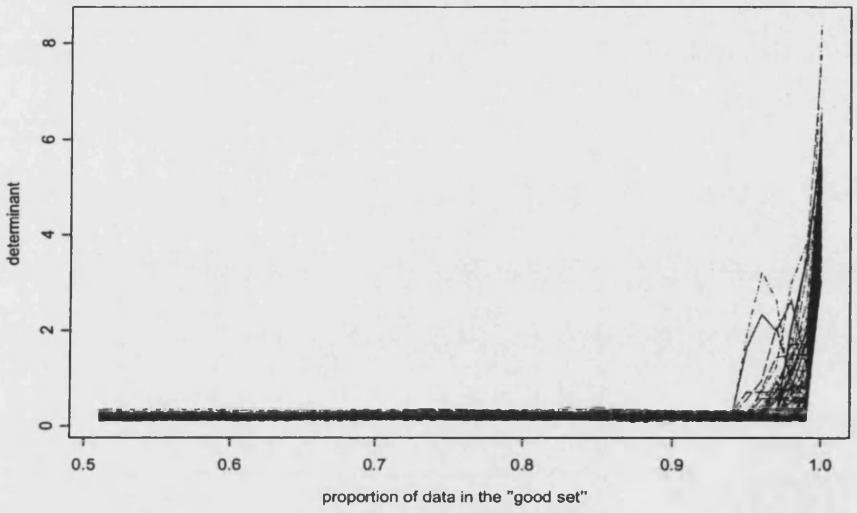
The development of robust statistics has been particularly intensive in the past twenty years. In addition to the high-breakdown and affine equivariance, most of the endeavours are focused on the problems of high dimensionality, efficiency and algorithms for optimization of the objective functions. It is understood that an affine equivariant estimator with high-breakdown point is an essential property.

There is no perfect estimator fulfilling all the desired properties of robustness (Davies 1993). For this reasons, some authors have directed their efforts towards compound estimators, trying to incorporate both high-breakdown and efficiency. Nevertheless, a "very robust" estimator may sometimes fail to identify observations lying too far from the bulk of the data because of some defects in the algorithm. The computational aspect is, therefore, as important as the estimator itself.

One of the problems arising in the computation of a robust estimator is selecting starting set. The consequences of including some outliers, although a small error is admissible, is a dramatic increase of the determinant of the robust scatter, indicating a large bias of the covariance matrix estimate. It is common practice to take advantage of some apriori knowledge of the data and on the possible number of outliers. Alternatively, we propose to use a Forward Search approach that offers a profile of solutions according to different choices of h, allowing us to choose the maximum size of the H-set before the bias explodes. The FS is again used to compare the differences in the estimates for different initial choices of M. The results confirm that the best approach is choosing m as small as possible to guarantee that the different choices of the H-sets are heterogeneous.

Figure 3.1: The change in the MCD covariance determinant for increasing sizes of the "good set". The simulated data are sampled from a mixture of two Normal distributions. The size of the contamination is 15% in(a); 25% in (b).

(a) $\epsilon = 0.15$



(b) $\epsilon = 0.25$

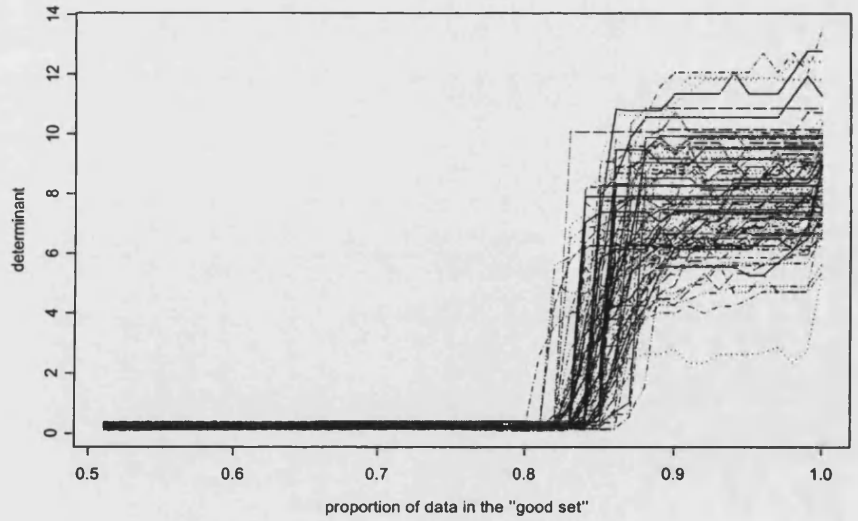
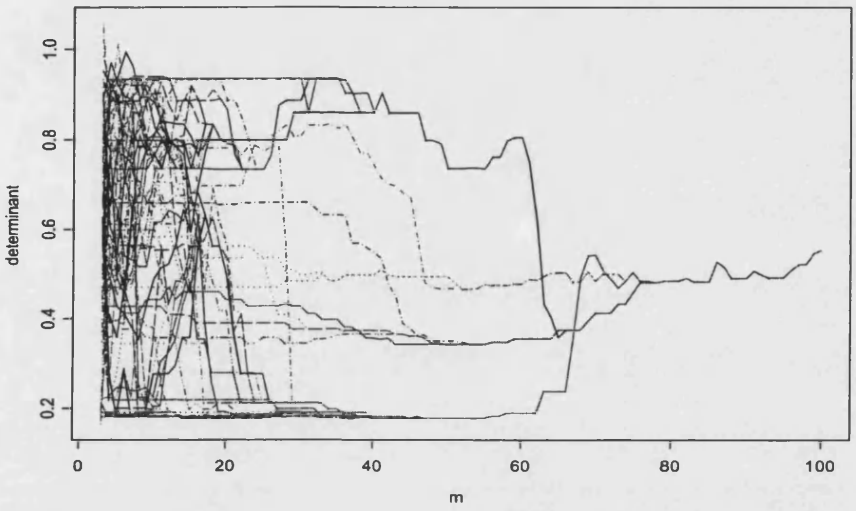
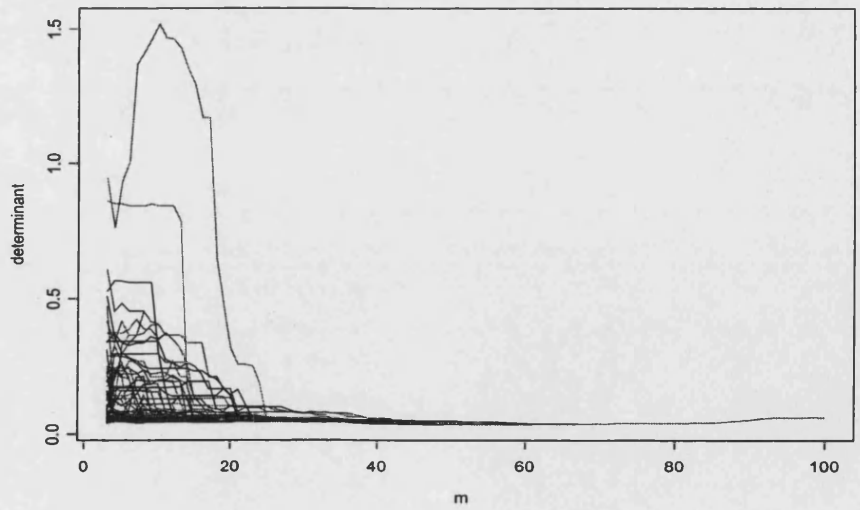


Figure 3.2: The change in the H-set covariance determinant for increasing sizes of the initial M-set. The data have 45% in (a) and 15% in (b).

(a) $\epsilon = 0.45$



(b) $\epsilon = 0.15$



Chapter 4

Detection of Multivariate Outliers

4.1 Introduction

The identification of multiple outliers is a widely discussed topic mainly because of the difficulties involved with the "traditional" methods of detection. These problems are known as "masking" and "swamping" effects. For a long time squared Mahalanobis Distances, (3.12), from now on referred as MD, have been the most common diagnostic: a large distance for a point may indicate that this observation is an outlier. However, these distances are functions of the sample estimates for the mean and variance-covariance matrix, that are highly sensitive to outliers. If a group of observations is lying "far" from the bulk of the data, the mean will shift in the direction of the small cluster and the covariance will inflate. As a consequence, some distances may have small values for observations belonging to the small cluster ("masking") and some others large values for cases belonging to the main bulk of the data ("swamping"). Furthermore, the difficulty in detecting multivariate outliers also depends on the way these points are arranged in the data in addition to the size of the sample and the dimensions. Woodruff and Rocke (1996) give good insights about this matter.

A first improvement of the literature is due to Campbell (1980) who introduces a diagnostic obtained from MD as functions of the robust M-estimates. Another step forward was using high-breakdown point estimators of location and scatter Rousseeuw and Van Zomeren (1991). The robust approach solves the masking problem meanwhile raising new issues. One concern is how to determine when a

point is too extreme to be consistent with the majority of the data. The literature follows a parametric approach: the distances are approximated by a Chisquare distribution and the cut-off is at the 0.975 tolerance ellipsoid.

This Chapter examines the distribution of the main diagnostics already existing in the literature. For small samples, we obtain the distribution of MD in an alternative simpler way than the proof of Wilks (1962). An extensive simulation work studies the convergence of some diagnostics to the theoretical distribution. Furthermore, we propose the use of simulation envelopes (Atkinson 1985) as a diagnostic plot for multivariate outliers. The Chapter is concluded with some examples on simulated and real data sets already popular in the literature.

4.2 Standardized Residuals

The diagnostics for multivariate outliers can be derived from the classical linear regression model:

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

where there are n independent cases on the continuous response $\mathbf{y} = (y_1, y_2, ..., y_n)$, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ is the matrix of regressors of dimensions $n \times p$ and $\boldsymbol{\epsilon}$ is the vector of errors.

The studentized residuals are defined as:

$$r_i = \frac{y_i - \hat{y}_i}{s\sqrt{(1-h_i)}},$$

where $s = \sqrt{\sum (y_i - \hat{y}_i)^2/(n-p)}$ is the residual mean square and $h_i = x_i^T(\mathbf{X}^T\mathbf{X})^{-1}x_i$ is the diagonal element of the hat matrix $\mathbf{H}$, often used to detect leverage points. A point of high leverage has extreme values for one or more explanatory variables and has the effect of driving the model close to the response observed for that point. The variance of the residuals is $\sigma^2(1-h_i)$ and $\sigma^2 h_i$ is the variance of the fitted value

$\hat{y}_i$, implying that points with high leverage have a smaller residual variance while the variance of the fitted value is high.

Cook and Weiseberg (1986) define the standardized residuals in a slight different way:

$$r'_i = \frac{y_i - \hat{y}_i}{s\sqrt{(1 - v_i)}},$$

where $v_i = x_i^T(\mathbf{X}^T\mathbf{X})^{-1}x_i$. The r'_i are referred as internally studentized residuals, although their distribution is not a Student-t. In §2.2 they show that the distribution of ${r'_i}^2/(n-p)$, where ${r'_i}^2 \in [0, (n-p)]$, is a beta with parameters $1/2$ and $(n-p-1)/2$. Since the maximum value is bounded, although homoskedastic, the residuals are not independent.

4.2.1 Deletion Residuals

An alternative way of standardizing the regression residuals is dividing by a variance estimator that is independent of e_i. This is obtained by computing the residual mean square after deleting the i-th observation. The residual sum of squares can be written as:

$$(n-p)s^2 = \mathbf{y}^T\mathbf{y} - \hat{\boldsymbol{\beta}}^T\mathbf{X}^T y. \tag{4.1}$$

After deletion of the i-th row, (4.1) becomes:

$$(n-p-1)s^2_{(i)} = \mathbf{y}^T\mathbf{y} - y_i^2 - \hat{\boldsymbol{\beta}}^T_{(i)}(\mathbf{X}^T\mathbf{y} - \mathbf{x}_i^T y_i), \tag{4.2}$$

where:

$$\hat{\boldsymbol{\beta}}_{(i)} = ({\mathbf{X}_{(i)}}^T\mathbf{X}_{(i)})^{-1}(\mathbf{X}^T\mathbf{y} - \mathbf{x}_i^T y_i) = \hat{\boldsymbol{\beta}} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i e_i/(1-h_i), \tag{4.3}$$

since $e_i = (1-h_i)y_i$ and ${\mathbf{X}_{(i)}}^T\mathbf{X}_{(i)} = X^TX - \mathbf{x}_i\mathbf{x}_i$. Combining (4.2) and (4.3)

$$(n-p-1)s^2_{(i)} = (n-p)s^2 - e_i^2/(1-h_i) \tag{4.4}$$

and therefore:

$$s^2_{(i)} = s^2\left(\frac{n-p-r_i^2}{n-p-1}\right). \tag{4.5}$$

For a complete analytical derivation the reader can refer to Atkinson (1985) §2.2 or (Atkinson and Riani 2000) .

Finally, the expression for the deletion studentized residuals is:

$$r_i^* = \frac{y_i - \hat{y}_i}{s_{(i)}\sqrt{(1-h_i)}}, \tag{4.6}$$

where the supremum of r_i^* is unbounded. Since e_i follows a $N(0,(1-h_i)\sigma^2)$ and $s_{(i)}$ is independent of e_i, we can rearrange the above expression in order to obtain:

$$r_i^* = \frac{e_i/(\sigma\sqrt{(1-h_i)})}{s_{(i)}/\sigma},$$

where the numerator is a $N(0,1)$ and the denominator is $\sqrt{\chi^2_{(n-p-1)}/(n-p-1)}$. Therefore, the distribution of r_i^* is a Student-t with $(n-p-1)$ degrees of freedom.

4.3 Mahalanobis Distances

This section studies the distribution of the MD_i^2 (MD), (3.12), from which most of the robust diagnostics derive. Asymptotically, the distances follow a Chisquare. However, our interest is rather focused on the small sample distribution, known in the literature by two results: Wilks (1962) and Penny (1996). The proof of Wilks (1962) is rather cumbersome; we propose an alternative simpler method. Our derivation makes use of two diagnostics referred to as *out-of-sample* and *deletion* MD introduced in the next two subsections.

4.3.1 Out-of-sample MD

If the linear regression model fits only the constant, the squared standardized residuals become:

$$r_i^2 = \frac{(y_i - \bar{y})^2}{s^2}, \tag{4.7}$$

where $\bar{y} = \sum_i y_i/n$. (4.7) is the expression for the squared "classical" MD of a sample of n points independently observed on the $\mathbf{y}$ variable. Generalizing for a sample of size n drawn from a p-variate Normal population, $\mathbf{Y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the MD of an extra observation $\mathbf{y}_{out}$, of dimension $p \times 1$, independent of $\mathbf{Y}$, is:

$$MD^2_{\mathbf{y}_{out}} = (\mathbf{y}_{out} - \bar{\mathbf{y}})^T \mathbf{S}^{-1} (\mathbf{y}_{out} - \bar{\mathbf{y}}), \tag{4.8}$$

where $\bar{\mathbf{y}} = n^{-1}(\mathbf{Y}^T\mathbf{1})$ and $\mathbf{S} = (n-1)^{-1}[\mathbf{Y}^T\mathbf{Y} - n\bar{\mathbf{y}}\bar{\mathbf{y}}^T]$ are the sample unbiased estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. This is a simpler notation from that in (3.1) and (3.2), where in the new notation it is implicit that the estimates are computed on a sample of n observations. If the sample estimates are replaced by the population moments, it is straightforward that, being a sum of squares of standardized normals,

$$(\mathbf{y}_{out} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y}_{out} - \boldsymbol{\mu}) \sim \chi^2_p.$$

Definition 4.3.1

The Hotelling T^2 distribution with parameters p and m is defined as the following:

$$m\mathbf{d}^T\mathbf{M}^{-1}\mathbf{d} \sim T^2(p, m) \tag{4.9}$$

where $\mathbf{d}$ and $\mathbf{M}$ are independently distributed as $N_p(\mathbf{0}, \mathbf{I})$ and a standard Wishart $W_p(\mathbf{I}, m)$.

Definition 4.3.2

A matrix $\mathbf{M}$ is said to have a Wishart distribution with scale matrix $\boldsymbol{\Sigma}$ and degrees of

freedom m if it can be written as $\mathbf{M} = \mathbf{X}^T\mathbf{X}$, where $\mathbf{X}$ of size $m \times p$ has distribution $N_p(\mathbf{0}, \boldsymbol{\Sigma})$. If $\boldsymbol{\Sigma} = \mathbf{I}$, then the Wishart distribution is said to be of the standard form (Mardia, Kent, and Bibby 1982, page 66).

Since $(\mathbf{y}_{out} - \bar{\mathbf{y}}) \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}(n+1)/n)$ and $(n-1)\mathbf{S} \sim W_p(\boldsymbol{\Sigma}, n-1)$ (Mardia, Kent, and Bibby 1982, page 68-69), by setting:

$$\mathbf{d}^* = \left(\frac{n+1}{n}\right)^{-1/2} \boldsymbol{\Sigma}^{-1/2}(\mathbf{y}_{out} - \bar{\mathbf{y}})$$

and

$$\mathbf{M}^* = (n-1)\boldsymbol{\Sigma}^{-1/2}\mathbf{S}\boldsymbol{\Sigma}^{-1/2},$$

we satisfy the requirements for (4.9). Therefore, we can write:

$$(n-1)\mathbf{d}^{*T}\mathbf{M}^{*-1}\mathbf{d}^* \sim T^2(p, n-1),$$

which leads to:

$$\left(\frac{n}{n+1}\right)(\mathbf{y}_{out} - \bar{\mathbf{y}})^T\mathbf{S}^{-1}(\mathbf{y}_{out} - \bar{\mathbf{y}}) \sim T^2(p, n-1).$$

Furthermore, by Theorem 3.5.2 (Mardia, Kent, and Bibby 1982, page 74):

$$T^2(p, n-1) = \frac{p\,(n-1)}{n-p} F_{p,\,n-p},$$

where $F_{p,n-p}$ is a Fisher distribution with p and $n-p$ degrees of freedom respectively. Therefore, trivially:

$$MD^2_{\mathbf{y}_{out}} \sim \frac{(n^2-1)p}{n(n-p)} F_{p,\,n-p}. \tag{4.10}$$

When the sample size grows, p being constant,

$$p\, F_{p,\, n-p} \rightarrow \chi^2_p.$$

The result follows by considering that $F_{p,n-p} = (\chi^2_p/p)/(\chi^2_{n-p}/(n-p))$ and that

$$\frac{\chi^2_n}{n} = \sum_{i=1}^{n} z_i^2/n \overset{n\uparrow\infty}{\longrightarrow} 1,$$

where $z_i \overset{i.i.d}{\sim} N(0,1)$. Usually, good approximations are already attained when $n/p > 5$, for $n \geq 5$.

4.3.2 Deletion MD

Similarly to the deletion studentized residuals (4.6), deletion MD are defined as the following:

$$MD^2_{(i)} = (\mathbf{y}_i - \bar{\mathbf{y}}_{(i)})^T \mathbf{S}^{-1}_{(i)} (\mathbf{y}_i - \bar{\mathbf{y}}_{(i)}),$$

where:

$$\mathbf{y}_i - \bar{\mathbf{y}}_{(i)} = \frac{n}{n-1} (\mathbf{y}_i - \bar{\mathbf{y}}) \tag{4.11}$$

and in a similar way as the derivation of (4.4), it is possible to get the diagonal and off-diagonal elements of $\mathbf{S}$:

$$\begin{aligned}
(n-2)\mathbf{S}_{(i)jj} &= (n-1)\mathbf{S}_{jj} - \left(\frac{n}{n-1}\right)(y_{jj} - \bar{y}_{.j})^2, \\
(n-2)\mathbf{S}_{(i)jk} &= (n-1)\mathbf{S}_{jk} - \left(\frac{n}{n-1}\right)(y_{ij} - \bar{y}_{.j})(y_{ik} - \bar{y}_{.k}).
\end{aligned}$$

Generalizing:

$$(n-2)\mathbf{S}_{(i)} = (n-1)\mathbf{S} - \left(\frac{n}{n-1}\right)\mathbf{e}_i\mathbf{e}_i^T, \tag{4.12}$$

where the $\mathbf{e}_i = \mathbf{y}_i - \bar{\mathbf{y}}$. The inverse of $\mathbf{S}_{(i)}$ can be found by applying the formula for updating the inverse of a matrix of the type $\mathbf{A} + \mathbf{a}^T\mathbf{b}$:

$$(\mathbf{A} + \mathbf{a}^T\mathbf{b})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{a}^T(\mathbf{I} + \mathbf{b}\mathbf{A}^{-1}\mathbf{a}^T)^{-1}\mathbf{b}\mathbf{A}^{-1} \tag{4.13}$$

where $\mathbf{A}$ is a full rank $(m \times m)$ matrix and $\mathbf{a}$, $\mathbf{b}$ are $(q \times m)$ of rank q. The formula (4.13) can be applied to (4.12) by letting $\mathbf{A} = (n-1)\mathbf{S}$, $\mathbf{a} = -\{n/(n-1)\}^{1/2}(\mathbf{y}-\bar{\mathbf{y}})^T$ and $\mathbf{b} = \{n/(n-1)\}^{1/2}(\mathbf{y}-\bar{\mathbf{y}})$. The resulting expression for $\mathbf{S}_{(i)}^{-1}$, together with (4.11), substituted in (4.8), allows deriving the squared deletion distances as a monotone function of the classical MD (Atkinson and Mulira 1993):

$$MD_{(i)}^2 = \frac{(n-2)n^2}{(n-1)^3}\left\{\frac{MD_i^2}{1 - nMD_i^2/(n-1)^2}\right\}. \tag{4.14}$$

Deletion distances do not add much more informative value than classical MD as a diagnostic tool. However, our interest is not in their interpretation but rather in their properties that will be useful for the derivation of the MD critical points discussed in the next subsection.

4.3.3 In-sample MD

In Section 4.3.1 we showed that, in small samples, MD follow an F rather than a χ^2 distribution. However, in

$$(\mathbf{y}_{out} - \bar{\mathbf{y}})^T\mathbf{S}^{-1}(\mathbf{y}_{out} - \bar{\mathbf{y}}) \sim \{(n^2-1)p/[n(n-p)]\}F_{p,n-p}$$

$\mathbf{y}_{out}$ does not belong the data. This condition is obviously not true when testing to detect outliers. Penny (1996) discusses this problem and obtains new critical values for the MD when $\mathbf{y}$ is not independent from the rest of the sample. The paper refers to Wilks (1962), who gives the distribution of the scatter ratio for multivariate Normal samples. The ratio is defined as:

$$R_i = \frac{|\mathbf{A}_i|}{|\mathbf{A}|},$$

where:

$$\mathbf{A} = \mathbf{Y}^T\mathbf{Y} - n\bar{\mathbf{y}}\bar{\mathbf{y}}^T = (n-1)\mathbf{S}$$

and $\mathbf{A}_i = \frac{n}{n-1}(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^T$ is $\mathbf{A}$ computed with the i-th observation of $\mathbf{Y}$ deleted. The ratio is computed for each observation of the sample; the furthest point from the bulk of the data is the observation with the greatest reduction in the determinant $|\mathbf{A}|$ and, therefore, will have the smallest scatter ratio R_i.

By using the moment generating function of the Wishart distribution, Wilks (1962) shows that:

$$R_i \sim B_e\left(\frac{n-p-1}{2}, \frac{p}{2}\right). \tag{4.15}$$

Penny (1996) refers to (4.15) to show that:

$$MD_i^2 \sim \frac{p\,(n-1)^2 F_{p,n-p-1}}{n\,(n-p-1+p\,F_{p,n-p-1})}.$$

Furthermore, empirical results evidence that there is a significant difference in the outliers detected with this critical region rather than $p\,(n-1)/(n-p)F_{p,n-p-1}$.

Alternatively, we provide a simpler way of obtaining the same critical points. From (4.14):

$$\frac{(n-1)^3}{(n-2)n^2}\,MD_{(i)}^2 = MD_i^2\left\{1 - \frac{n}{(n-1)^2}\,MD_i^2\right\}^{-1}$$

and therefore:

$$\frac{1}{MD_i^2} = \frac{n}{(n-1)^2} + \frac{(n-2)\,n^2}{(n-1)^3 MD_{(i)}^2}. \tag{4.16}$$

Since $\mathbf{y}_i$ is independent of $\bar{\mathbf{y}}_{(i)}$ and $\mathbf{S}_{(i)}$ we can use the result of Mardia, Kent, and Bibby (1982), (4.10), to derive the distribution of the deletion distances:

$$MD_{(i)}^2 \sim \frac{np\,(n-2)}{(n-1)(n-p-1)}\,F_{p,n-p-1}. \tag{4.17}$$

The substitution of (4.17) in (4.16) gives:

$$\frac{1}{MD_i^2} \sim \frac{n}{(n-1)^2} + \frac{(n-p-1)\,n}{(n-1)^2 p\,F_{p,n-p-1}}, \tag{4.18}$$

leading to the same result as Penny (1996):

$$MD_i^2 \sim \frac{p\,(n-1)^2 F_{p,n-p-1}}{n\,(n-p-1+p\,F_{p,n-p-1})}. \tag{4.19}$$

The F distribution is a scaled ratio of Chisquares, so:

$$\left(\frac{p}{n-p-1}\right) F_{p,n-p-1} = \frac{\chi_p^2}{\chi_{n-p-1}^2} = \frac{\chi_p^2 + \chi_{n-p-1}^2}{\chi_{n-p-1}^2} - 1 \tag{4.20}$$

The Beta distribution can also be written as a ratio of Chisquares. If $X_1 \sim \Gamma(\alpha, \delta)$ and $X_2 \sim \Gamma(\beta, \delta)$ are independent Gammas with the same scale parameter,

$$Y = \frac{X_1}{X_1 + X_2} \sim Be(\alpha, \beta)$$

Furthermore, $\chi_\nu^2 = \Gamma\left(\frac{\nu}{2}, 2\right)$. Therefore:

$$\left(\frac{\chi_{n-p-1}^2 + \chi_p^2}{\chi_{n-p-1}^2}\right)^{-1} = Be\left(\frac{n-p-1}{2}, \frac{p}{2}\right),$$

which, together with (4.20), leads to:

$$F_{p,n-p-1} = \frac{p}{n-p-1}\left[Be\left(\frac{n-p-1}{2}, \frac{p}{2}\right)^{-1} - 1\right]. \tag{4.21}$$

Substituting (4.21) in (4.19):

$$MD_i^2 \sim \frac{(n-1)^2}{n}\left\{Be\left(\frac{n-p-1}{2},\frac{p}{2}\right)\right\}. \tag{4.22}$$

Knowing the distribution of the MD, we can easily derive that of the scatter ratio. This last can be rewritten according to (4.12) in the following way:

$$R_i = \frac{|\mathbf{A}_i|}{|\mathbf{A}|} = \frac{|\mathbf{A} - \left(\frac{n}{n-1}\right)(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^T|}{|\mathbf{A}|}. \tag{4.23}$$

If a matrix $\mathbf{Z}$ can be partitioned in the following way:

$$\mathbf{Z} = \begin{pmatrix} \mathbf{C} & \mathbf{a} \\ \mathbf{a}^T & -1 \end{pmatrix}$$

where $\mathbf{C}$ is a $p \times p$ matrix and $\mathbf{a}$ is a $p \times 1$ vector, then Cook and Weiseberg (1986) show that:

$$|\mathbf{Z}| = -|-\mathbf{C} - \mathbf{a}\mathbf{a}^T|.$$

Setting $\mathbf{C} = -\mathbf{A}$, $\mathbf{a} = (\mathbf{y}_i - \bar{\mathbf{y}})(\frac{n}{n-1})^{1/2}$, and $\mathbf{Z} = \mathbf{A}_i$, we find (4.23).

Rao (1973) for the same partition shows:

$$|\mathbf{Z}| = -|\mathbf{C}|\{1 + \mathbf{a}^T\mathbf{C}^{-1}\mathbf{a}\}.$$

(4.24) is referred to as *expansion of a bordered determinant*, and therefore:

$$|\mathbf{A}_i| = |\mathbf{A}|\{1 - \left(\frac{n}{n-1}\right)(\mathbf{y}_i - \bar{\mathbf{y}})^T.\mathbf{A}^{-1}(\mathbf{y}_i - \bar{\mathbf{y}})\}$$

Since $\mathbf{A} = (n-1)\mathbf{S}$, we can write:

$$R_i = 1 - \frac{n}{(n-1)^2} MD_i^2.$$

Finally:

$$R_i \sim \frac{n-p-1}{n-p-1+p\,F_p, n-p-1},$$

from which, using the relation between F and *Beta* distributions, we obtain:

$$\left(1+\frac{p}{n-p-1}F_{p,n-p-1}\right)^{-1} = B_e\left(\frac{n-p-1}{2}, \frac{p}{2}\right),$$

confirming the same result as Wilks (1962).

4.4 Robust MD

This section is dedicated to the distribution of the RD. These are MD written as functions of the MCD estimator for multivariate location and scatter ($\bar{\mathbf{y}}_{MCD}$ and $\mathbf{S}_{MCD}$). The expression for the $RD_i^2(h)$ (RD), already seen in (3.16), is:

$$RD_i^2(h) = (\mathbf{y}_i - \bar{\mathbf{y}}_{MCD}(h))^T \mathbf{S}_{MCD}^{-1}(h)(\mathbf{y}_i - \bar{\mathbf{y}}_{MCD}(h)),$$

where $i = 1, 2, ..., n$ and h is the number of observations fitted with the MCD method. The small sample distribution of the estimates is not known; however, we will refer to the asymptotic results provided by the literature. Butler, Davies, and Juhn (1993) show Fisher consistency for the location estimator, that is:

$$\bar{\mathbf{y}}_{MCD}(F_{\boldsymbol{\mu},\boldsymbol{\Sigma}}) = \boldsymbol{\mu} \tag{4.24}$$

and:

$$\mathbf{S}_{MCD}(F_{\boldsymbol{\mu},\boldsymbol{\Sigma}}) = c\boldsymbol{\Sigma}, \tag{4.25}$$

where the robust estimators are written as functionals of the elliptical distribution $F_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$. The scatter matrix needs to be scaled for consistency. This result has been

generalized by Grübel and Rocke (1990) to any affine equivariant estimator. (4.24) and (4.25) can be rewritten as:

$$(\bar{\mathbf{y}}_{MCD}, \mathbf{S}_{MCD}) \xrightarrow{p} (\boldsymbol{\mu}, c\boldsymbol{\Sigma}), \tag{4.26}$$

where $\xrightarrow{p}$ denotes convergence in probability. According to Slutzky theorem, for any random variables X_n, X and continuous function g,

$$\text{if } X_n \xrightarrow{p} X, \text{ then } g(X_n) \xrightarrow{p} g(X).$$

The theorem can be applied to (4.26) to obtain:

$$\{(\mathbf{y}_i - \bar{\mathbf{y}}_{MCD}(h))^T \mathbf{S}_{MCD}(h)^{-1} (\mathbf{y}_i - \bar{\mathbf{y}}_{MCD}(h))\} \xrightarrow{p} c \times \{(\mathbf{y}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \boldsymbol{\mu})\} \tag{4.27}$$

In other words, we can either scale the distances or the covariance estimates to the shape of the Normal distribution. Rousseeuw and Van Driessen (1999) standardize by the median robust distance. Similar forms of scaling have been used by Woodruff and Rocke (1996) who standardize by the h-th order statistics divided by the h/n quantile of a $\sqrt{\chi^2_p}$, where $h = [(n + p + 1)/2]$. Atkinson (1994) uses the sum of the squared distances, $p(n-1)$, divided by the average of the total distances over k simulations.

4.5 Simulation Envelopes for Robust and Mahalanobis Distances

According to (4.27), the robust distance asymptotic distribution is some scaled form of a Chisquare. MD also converge to a Chisquare. These approximations set the critical regions for multiple outlier detection: the common approach of the literature chooses a cut-off at the 97.5% percentile. Furthermore, in the Fast-algorithm for the computation of the MCD, the 97.5% tolerance ellipse is the decision rule for reweighting the observations (3.15).

In Section 4.6, Section 4.7 and in this section, we show empirically that this approximation is not satisfactory.

Figure 4.1 includes some descriptive graphs of Mahalanobis and RD. The data are a sample of 100 observations generated from a standard bivariate Normal distribution. In the Q-Q plots, the RD deviate more than MD from the expected Chisquare. The differences appear to be more evident on the upper tail, where the observations lie far above the 45 degrees line. The box-plots, panels (g) and (h), confirm that the RD have longer tails than the classical MD.

In some cases the sampling distortions make it difficult, just by looking at the Q-Q plots, to assess whether the points are distributed as expected. To cope with this problem, Atkinson (1985) introduces the simulation envelopes, a graphical tool applied to the residual diagnostic in linear regression. The idea is to generate an empirical region where the points should be included, if they are well-behaved.

We suggest a different use of simulation envelopes, that is to compare the empirical distribution of the 0.975 quantile of Mahalanobis and RD with the theoretical one.

The simulation envelope for the distances is constructed as follows: a sample of n distance quantiles is replicated k times from a p-variate Normal. As a result, we obtain an array of dimension $(k \times n)$. Next, the array is ordered within the columns, obtaining k ordered vectors of n distances ordered from the smallest to the largest. The envelope is built so that the probabilities of one observation falling outside the lower and the upper bounds are respectively 2.5% and 97.5%. In other words, if 119 quantiles are generated, the lower and upper bounds will be respectively the 3-rd smallest and the 117-th largest vector of quantiles, since $3/120 = 0.025$ and $117/120 = 0.975$. So, if $d^2_{[k]}[i]$ is the i-th ordered distance generated in the k-th simulation, the bounds of the envelope in 119 simulations are:

$$\begin{aligned} d^2_l &= d^2_{[3]}[1], d^2_{[3]}[2], ..., d^2_{[3]}[n] \\ d^2_u &= d^2_{[117]}[1], d^2_{[117]}[2], ..., d^2_{[117]}[n]. \end{aligned}$$

If the distribution of the distances is well approximated by the Chisquare, we

would expect that the lower and upper bound of the envelopes were sufficiently close to the correspondent order statistics of the theoretical distribution.

Figure 4.2 displays the envelope: the upper and lower bounds are plotted against the quantiles of a Chisquare distribution. The solid line is the vector of average distances over the simulations.The plot shows that while MD are well behaved, the robust distance (computed using the Fast-algorithm) cut-off at 97.5% lies well above the Chisquare distribution, confirming that they are longer tailed than the null distribution. The Chisquare approximation is still poor for a relatively large spread between n and p (Figure 4.1, panel (d)). The consequence in outlier detection is that the use of the Chisquare tolerance ellipse leads to an "overestimation" of the number of outliers. In Section 4.6 a simulation experiment gives further evidence of this result.

4.6 Further Empirical Evidence

As a confirmation of what has been evidenced in the previous section, this simulation study reports the average number of observations lying outside the $\chi^2_{p,0.975}$ tolerance ellipse.

The experiment replicates 500 samples of n RD from a p-variate standard Normal. For each sample, the observations labelled as outliers are those for which:

$$RD_i^2(h) > \sqrt{\chi^2_{p,0.975}}$$

where $i = 1, 2, ..., n$.

If the distances follow the theoretical distribution and in presence of no outliers, about 2.5% of the data will fall outside the chosen tolerance region. Table 4.1 shows that, for relatively small samples and dimensions, the proportion of observations that is not covered by the tolerance region is higher than expected. However, the results sensibly improve when the sample size grows compared to the dimensions.

Figure 4.3 plots the distribution of the mean and the median robust and MD against χ^2_3. The RD show a change in the pattern around the h-th point: the reason is that the MCD ellipsoids are constrained to cover at least h observations. The

smallest h distances will be, therefore, closer to each other. This irregularity in the plot becomes less evident when the spread between n and p increases.

Table 4.1: Average and median proportion of observations lying outside the tolerance ellipse of $\chi^2_{p,0.975}$.

n	p	mean	median
20	2	0.23	0.25
50	2	0.11	0.10
100	2	0.06	0.07
150	2	0.04	0.05
20	3	0.30	0.32
50	3	0.15	0.14
100	3	0.07	0.07
150	3	0.06	0.05
20	4	0.37	0.4
50	4	0.21	0.20
100	4	0.09	0.08
150	4	0.06	0.06

4.7 A Monte Carlo Test

In order to support the graphics with some data, we carry out a goodness-of-fit test for the distribution of the RD. There are three parameters affecting the convergence to the asymptotic distribution: the sample size, the dimensions and the proportion of observations fitted robustly. In simple terms, the question is for which combination of n, p and h the distances approximate to a χ^2_p. The convergence of the RD is compared with that of the MD.

The test performed is the Anderson-Darling based on the comparison between the theoretical and the empirical cumulative distributions. The following expression describes the whole class of Cramér-von Mises measures of discrepancy:

$$Q = n \int_{-\infty}^{\infty} \{F_n(y) - F(y)\}^2 \psi(y) dF(y), \tag{4.28}$$

where $F(y)$ is the cdf and $F_n(y)$ is the edf defined as in (3.5). The Anderson-Darling statistic A^2 is obtained by replacing $\psi(y) = [\{F(y)\}\{1 - F(y)\}]^{-1}$ in (4.28). The weighting function ψ has the effect of giving a greater importance to the tails of the distribution, which attributes good power properties to the test. The statistic is computed by:

$$A^2 = -n - (1/n)\sum_{i=1}^{n}\{2i - 1\}\{\ln z_{(i)} + \ln(1 - z_{(n+1-i)})\} \tag{4.29}$$

where $z_{(i)}$ are the ordered values of $z = F(y)$. Further computational aspects are discussed in D'Agostino and Stephens (1986).

The simulation experiment starts by generating a sample of K=1000 vectors of distances from a multivariate standard Normal distribution. The MCD is computed using the Fast-algorithm. We have tested each vector of distances and counted how many times the null distribution, χ^2_p, has been rejected over the K replications. Having set the level of significance at $\alpha = 0.05$, if the distances are distributed as a Chisquare, we expect the simulated proportion of rejection to be very close to α.

The distribution of the A^2 is known, also for finite samples, in the case where the parameters of the null distribution are fully specified. The percentage points are found assuming that the $z_{(i)}$ have a uniform distribution under the null hypothesis. However, if there are one or more unknown parameters, the $z_{(i)}$ are no longer uniformly distributed, but rather have a higher density on low values. By consequence, also the percentage points of A^2 change. These depend on the distribution tested, the sample size and the method of estimation. As a result, if (4.29) is calculated on the sample moments or other type of estimates of the population parameters, even for large samples, the test produces exceptionally small p-values. This odd situation does not occur if the estimates are replaced by the population moments.

The appropriate correction factors, in addition to the tables of percentage points for the Anderson-Darling and other quadratic edf statistics, are available for a limited number of distribution functions and in the case when one or all the parameters are estimated by maximum likelihood (D'Agostino and Stephens 1986). For the distribution of MD, we refer to the tables of the Gamma when the scale parameter

α is known and the shape β is unknown. This is equivalent to testing if y_i scaled for an unknown constant belongs to a χ^2_p with known degrees of freedom and where $\alpha = p/2$ and $\beta = 2$. For the robust estimates there are no available tables. The half-sample method (Stephens 1978) offers a possibility of dealing with unknown population parameters. It consists in using only a random half of the sample to estimate the parameters. The A^2 statistic is then computed using the whole data set. Stephens (1978) shows that, under these circumstances, the A^2 converges quite fast ($n \geq 20$) to the distribution where all the parameters are known.

The simulation experiment tests a sample of distances where the observations come from a common parent population with unknown parameters, although they are not independent. However, the comparisons of the results between simulated independent and correlated distances showed insignificant differences. D'Agostino and Stephens (1986) also refers to a similar case where a sample of linear regression residuals are tested for normality and conclude that when $n \geq 20$ the correlation among the observations does not change much the asymptotic points of the A^2 statistic.

4.7.1 The Test Results

The first two rows of Table 4.2 are the sample size and the dimensions. δ is the proportion of observations fitted through the MCD, that should be at least around 50% of the data. When $\delta = 1$ the unweighted MCD are equal to the sample estimates, implying that when the size of the fitted set increases the RD approach to the classical MD. The percentage of rejections in the Anderson-Darling test are computed for MD, MCD estimates and the one step re-weighted MCD and compared with the results for the theoretical distribution.

The table confirms that for a relatively small sample size of 50 and 2 up to 4 dimensions, MD are well approximated by the Chisquare. When the number of variables increase from 4 to 6, the sample size must grow of four times to achieve the convergence. The numbers in brackets refer to the half-sample method. These do not deviate much from the results from the available tables for small dimensions ($p = 2$). However, when the dimensions grow the convergence is much slower.

The RD also cover only half of the sample, although the test results for the two fits are completely different. While the half-sample method looks for a random subset, the MCD is the solution to an optimization procedure leading to the subset with minimum volume. The test is rejected more than 50% of the times, even for relatively large spreads between n and p. When δ is increased, the test improves sensibly. Croux and Haesbroeck (1997) provide an empirical study where it is shown that the efficiency of the MCD improves with δ approaching 1 and agree that $\delta = 0.75$ is a reasonable compromise between robustness and efficiency. When $\delta = 1$ the unweighted MCD are the sample estimates: the table shows that in this case the percentage of rejections are perfectly in line with those for the MD. However, excluding the case of $\delta = 1$, Table 4.2 shows that, even when the efficiency is improved by including more observations in the fit, the probability of rejection is still extremely high: approximately 86% when $n = 100$, $p = 2$ and $\delta = 0.9$, implying a very poor approximation.

An alternative way of improving the fit while δ is unchanged is to re-weight the estimates. The approximation to the theoretical distribution of the re-weighted MCD is better than the un-weighted estimates, although still not satisfactory. The test is never rejected less than 45% of the times.

The empirical results suggests that a correction factor is needed to improve the approximation of the Chisquare distribution to the RD. This factor can be computed empirically by following a similar approach used by Rousseeuw and Van Zomeren (1991) for the MVE.

4.7.2 Accuracy of the Results

The relative precision of the rejection proportions depends on the assumption that the distances are Chisquare random variables. Under the null hypothesis, we expect the percentage to be as close as possible to the theoretical $\alpha = 5\%$, although we need to consider the sampling error. At our simulation model the number of rejections is a binomial $B(K, \alpha)$, where K is the number of drawn samples. Therefore, the number of rejections would fall in the 95% confidence interval:

$$\alpha \pm 1.96\sqrt{\frac{\alpha(1-\alpha)}{K}},$$

using the central limit theorem. In other words, 95% of the probability of rejections at the null distribution are within 0.036 and 0.063. This is a reasonable error margin, although it could still be improved by increasing the number of simulations.

4.8 Robust Envelopes for Outlier Detection

The empirical results show that the Chisquare does not provide a good asymptotic approximation of the robust distance distribution. The evidences provided in Section 4.6 and Section 4.5 show that the distances are longer tailed than expected, leading to exclude "good" data as well as outliers. We suggest the use of empirical tolerance regions to cope with this problem. In Section 4.5 simulation envelopes were used as a diagnostic tool to compare the quantiles of the sample data with the theoretical distribution. In this section we propose the use of simulation envelopes as a diagnostic tool for multivariate outliers.

4.8.1 The Mean Shift Outlier Model

When simulating a contaminated data set, the outliers can be arranged in different ways. According to the mean shift outlier model, in a p-variate sample $\mathbf{Y}$ of n cases, there are h "good" observations following a Normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $n-h$ misplaced points from a $N(\boldsymbol{\mu}+\mu_0, \boldsymbol{\Sigma})$. In other words:

$$\mathbf{Y} \sim (1-\epsilon)N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \epsilon N(\boldsymbol{\mu}+\mu_0, \boldsymbol{\Sigma}),$$

where $\epsilon = (n-h)/n$ is the mixture proportion.

This situation is the worst case scenario. According to Woodruff and Rocke (1996), the type of outliers hardest to detect is the one with the same shape as the "good" data. In this case, MD of the "bad" data have a smaller expected value than in the case where also the covariance matrix is inflated. The result is that the outliers are "masked" among the "good" observations.

For outliers that are easier to detect, less sophisticated methods, other than the robust ones, can be used. For most of these cases, it is sufficient to look at the classical MD.

As a first example Figure 4.4, we consider a bivariate Normal distribution. The data are 50 observations of which the last five 5 are outliers. These are already evident in the scatter plot on the top right corner of panel (a). The simulation envelopes, panel (b), separate quite clearly the group of "good" data from the outliers. However, there are still a few observations lying above the 97.5% cut-off.

4.8.2 Example 1: Modified Wood Gravity Data

The wood gravity data often recurs in the literature on outlier diagnostic. The original set from Draper and Smith (1966) consists of five explanatory variables and one response framed in a linear regression model. Rousseeuw and Leroy (1987) contaminated the data by replacing some of the cases with outliers. The "bad" observations are units 4, 6, 8 and 19. Since these are leverage points, the response variable is redundant. Therefore, our data set includes only the five carriers.

The computed MCD covariances are singular. This usually occurs when one or more eigenvalues is zero or approaches the machine precision and, therefore, rounded off to zero. As a result, it is impossible to compute the inverse of the covariance matrix:

$$\mathbf{S}^{-1} = \mathbf{U}[\text{diag}(1/\lambda_j)]\mathbf{U}^T$$

where $\mathbf{U}$ is the column-orthogonal matrix of eigenvectors and $[\text{diag}(1/\lambda_j)]$ is the diagonal matrix whose elements are the reciprocals of the eigenvalues. To remove this inconvenience, the diagonal elements of the covariance have been multiplied by the machine floating-point precision 10^{-6}. This is probably not the most correct way of solving the problem. A better solution would have been multiplying the whole matrix by a scalar.

Figure 4.5, panel (b), is a Distance-Distance plot (Rousseeuw and Van Driessen 1999) where the MD are plotted against the robust diagnostics. The two lines are the

cut-off values at the $\chi^2_{5,0.975}$. If the data did not contain outliers, all the observations would fall within the lower left rectangle. Furthermore, since the classical and RD have different distributions and because of the relatively small size of the sample, in an outlier-free set we would not expect the points to lie on the 45 degrees line. MD don't detect any anomalies, although observations 8 and 19 are very close to the cut-off line. However, there are 7 observations lying beyond the boundary of the RD, with observations 4,6,8 and 19 being the most extreme points.

Figure 4.6, panels (a) and (b), are the simulation envelopes for Mahalanobis and RD plotted on a logarithmic scale. Also in this case, MD don't reveal any extreme points. Differently from the MD, the RD have a twisted pattern which deviates from a straight line, because of the presence of the outliers. In panel (a), observations 4,6,8 and 19 are outside the envelope with case 19 very close to the boundary. However, after deleting the 4 outliers panel (b), no observations fall outside the robust envelope. The pattern of the distances is still rather far from the expected distribution due to the relatively small size of the sample.

The shape of the simulated envelope looks very different from the envelope of the expected χ^2_5. This last is marked by the dotted line and lie far below the envelope, revealing 3 extra outliers, which confirms what has previously appeared from the D-D plot.

4.8.3 Example 2: Hawkins, Bradu and Kass Data

The data (Bradu, Hawkins, and Kass 1984) is artificial: there are 75 observations on 3 explanatory variables and one response. For our purposes, we have considered only the matrix of regressors. The first 14 observations in the sample are leverage points. These are already evident in the scatter matrix plot, Figure 4.8, panel (a). The data allows showing how effective the RD are, even when the size of the contamination is quite large.

The classical MD detect only two extreme points: observations 12 and 14. However, RD evidence clearly the group of outlying points. In this case, the outliers lie very distant from the rest of the observations and can be identified by the Chisquare critical points. No additional information is provided by the robust envelopes of the

RD.

4.8.4 Example 3: Stack Loss Data

The stackloss data (Brownlee 1965) is another classical example coming from outlier diagnostic in linear regression. There are 20 observations on 4 variables, of which 3 regressors and 1 dependent variable. In this case , we consider the whole data set without excluding the response Figure 4.9, panel (a). The four outliers known from the previous literature (1,3,4 and 21) are not leverage points, implying that they can't be detected if analyzing the matrix of the carriers only.

The Distance-Distance plot looks very similar to that for the woodgravity data. There are 8 points lying beyond the cut-off for the RD, split in two small clusters. The furthest four observations are the known outliers; cases 20, 14 and 13 are closer to the boundary and observation 2 is in between the two groups. MD don't detect any outliers.

The 97.5% envelope for the distances excludes 3 observations (3,21,1) and observations 4, 20, 2 are lying on the border.

4.9 Conclusions

The Chapter has discussed the distributional properties of the main multivariate outlier diagnostics. We have provided a proof for the small sample distribution of in-sample MD. The result agree with the cumbersome proof existing in the literature.

Empirical evidences have shown that the approximation of the RD to the Chisquare is poor. Increasing the efficiency of the estimates by including more observations in the fit via MCD or by re-weighting the estimates does not lead to significant improvements. We conclude that a correction factor, which can be found via simulation, is needed.

Because RD are longer tailed, the use of Chisquare tolerance regions leads to reject too many points. We have proposed the use of robust simulation envelopes as a possible graphical tool to detect groups of outliers avoiding problems of "over-identifications".

Table 4.2: Proportion of rejections for the AD test on Mahalanobis Distances (MD) and Robust Distances (RD). The unweighted RD are obtained from the MCD estimates not reweighted for efficiency. The size of the test is α for 1000 replications. δ is the proportion of observations fitted robustly.

δ	p	2			4			6				
	n	50	100	200	50	100	200	60	100	200	500	800
	expected	0.041	0.050	0.060	0.054	0.051	0.046	0.052	0.049	0.054	0.060	0.040
	MD	0.040	0.056	0.036	0.047	0.048	0.061	0.076	0.066	0.050	0.052	0.050
	MD (H-S)	(0.086)	(0.084)	(0.062)	(0.210)	(0.105)	(0.083)	(0.374)	(0.262)	(0.138)	(0.104)	(0.062)
0.5	RD (H-S)	(0.895)	(0.786)	(0.706)	(0.999)	(0.982)	(0.915)	(1)	(0.999)	(0.992)	(0.924)	(0.946)
	unwght RD (H-S)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)
0.7	RD (H-S)	(0.775)	(0.654)	(0.581)	(0.995)	(0.920)	(0.805)	(1)	(0.995)	(0.950)	(0.846)	(0.898)
	unwght RD (H-S)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)
0.75	RD (H-S)	(0.722)	(0.616)	(0.54)	(0.947)	(0.883)	(0.759)	(0.998)	(0.979)	(0.922)	(0.820)	(0.880)
	unwght RD (H-S)	(1)	(0.998)	(1)	(0.999)	(1)	(1)	(1)	(1)	(1)	(1)	(1)
0.85	RD (H-S)	(0.595)	(0.536)	(0.489)	(0.919)	(0.774)	(0.669)	(0.992)	(0.954)	(0.848)	(0.774)	(0.848)
	unwght RD (H-S)	(0.987)	(0.994)	(1)	(0.983)	(0.994)	(1)	(1)	(1)	(1)	(1)	(1)
0.90	RD (H-S)	(0.547)	(0.478)	(0.460)	(0.862)	(0.715)	(0.615)	(0.948)	(0.910)	(0.804)	(0.744)	(0.840)
	unwght RD (H-S)	(0.811)	(0.868)	(0.990)	(0.932)	(0.930)	(0.994)	(0.970)	(0.982)	(0.998)	(1)	(1)
1	RD (H-S)	(0.055)	(0.230)	(0.295)	(0.303)	(0.293)	(0.332)	(0.476)	(0.426)	(0.436)	(0.544)	(0.704)
	unwght RD (H-S)	(0.086)	(0.084)	(0.062)	(0.210)	(0.105)	(0.083)	(0.374)	(0.262)	(0.138)	(0.104)	(0.062)

Figure 4.1: Q-Q and Box Plots for Mahalanobis and robust distances in three independent samples from a simulated Normal distribution.

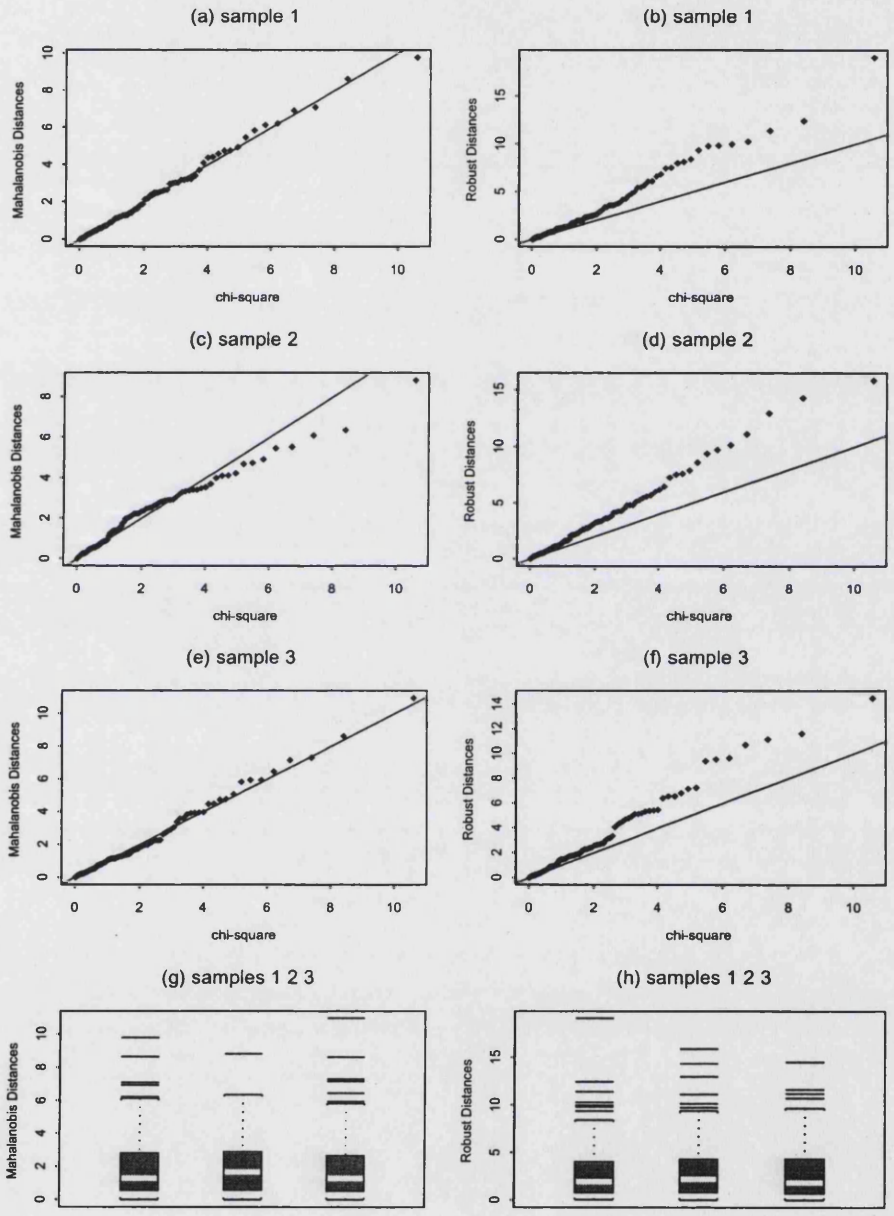


Figure 4.2: 97.5% Simulation envelopes for Mahalanobis and robust distances generated from simulated Normal data and theoretical order statistics of the χ^2_p.

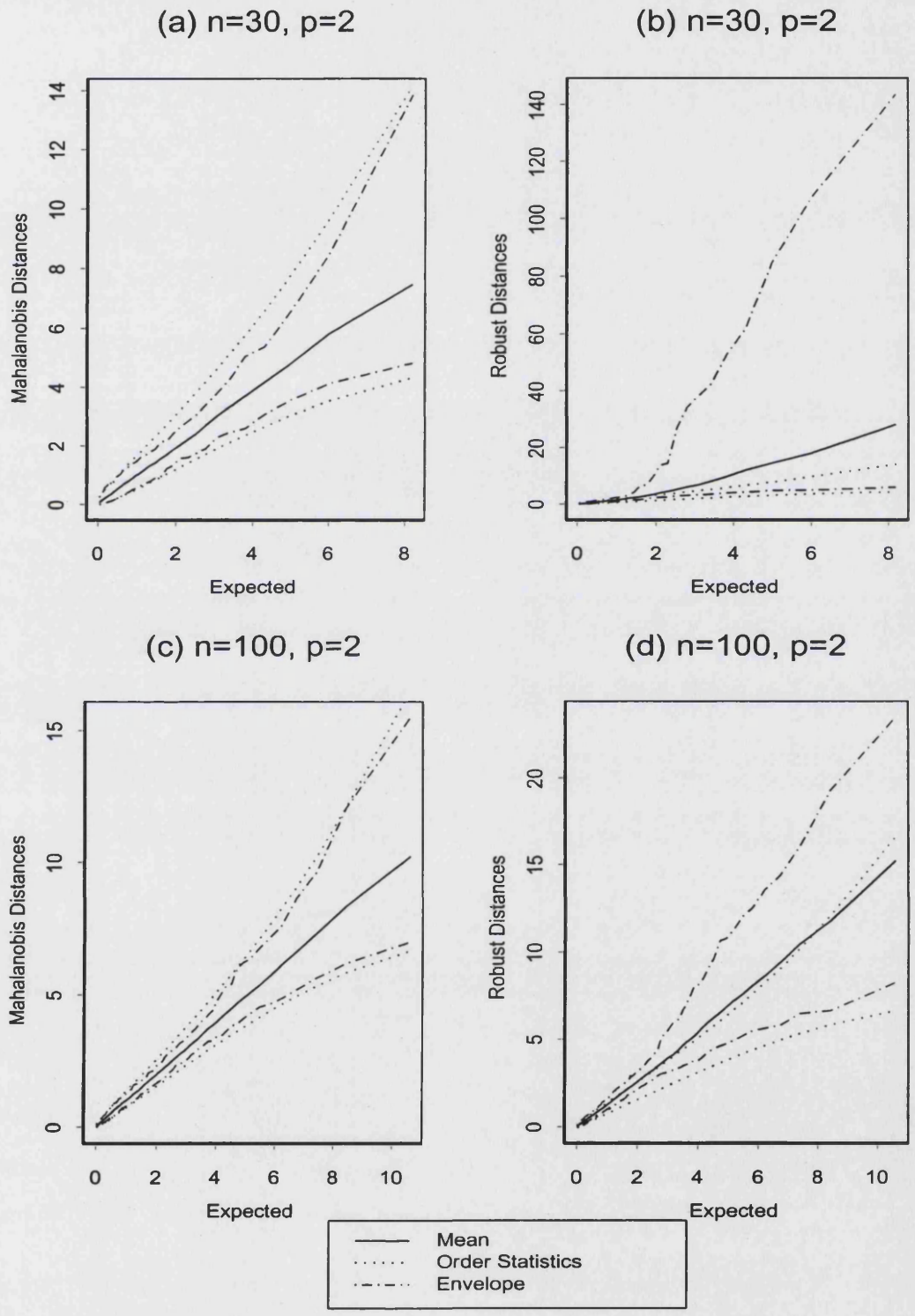


Figure 4.3: Quantile plots for mean and median of robust and Mahalanobis distances.

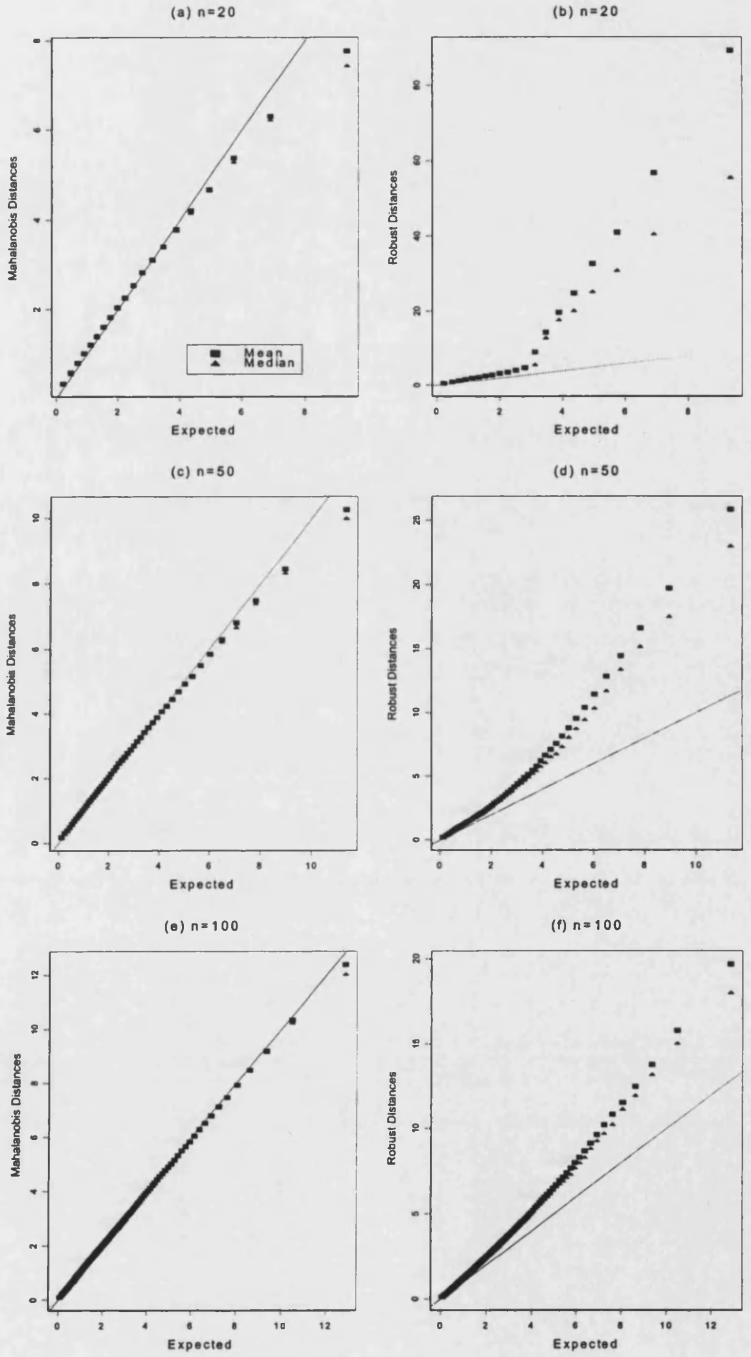
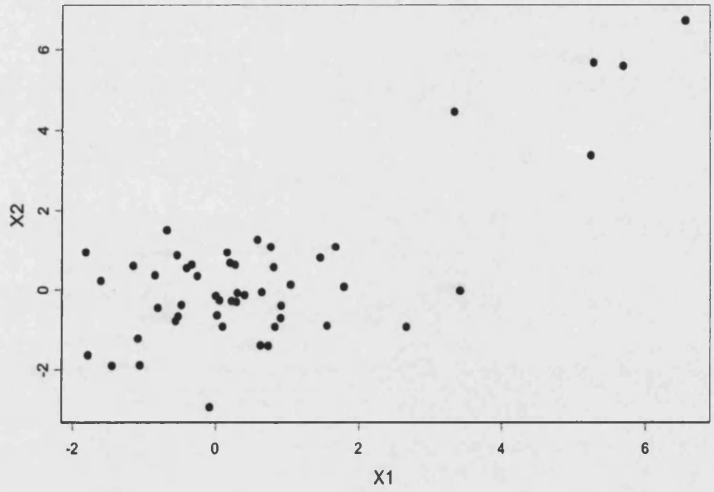


Figure 4.4: Mean-shift outlier model.

(a) Scatter plot



(b) Simulation Envelope

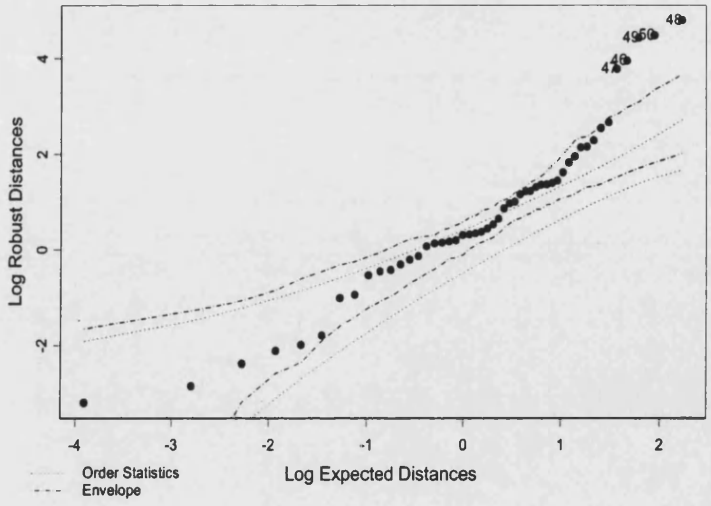
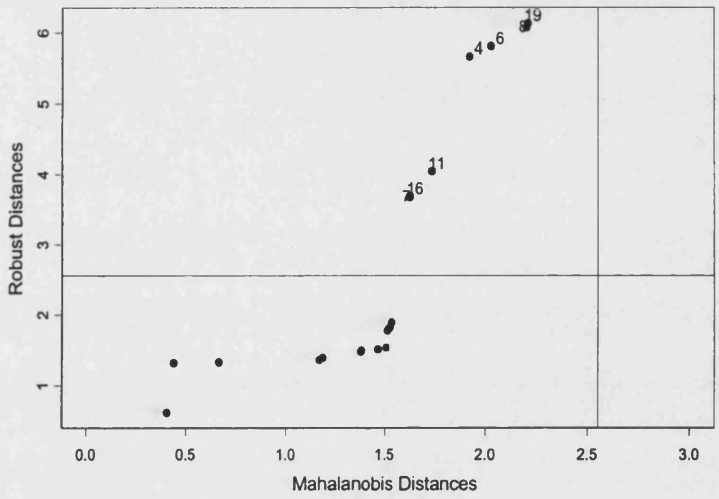


Figure 4.5: Example 1, Woodgravity data.

(a) Distance-Distance Plot



(b) 97.5% Simulation Envelopes for Mahalanobis Distances

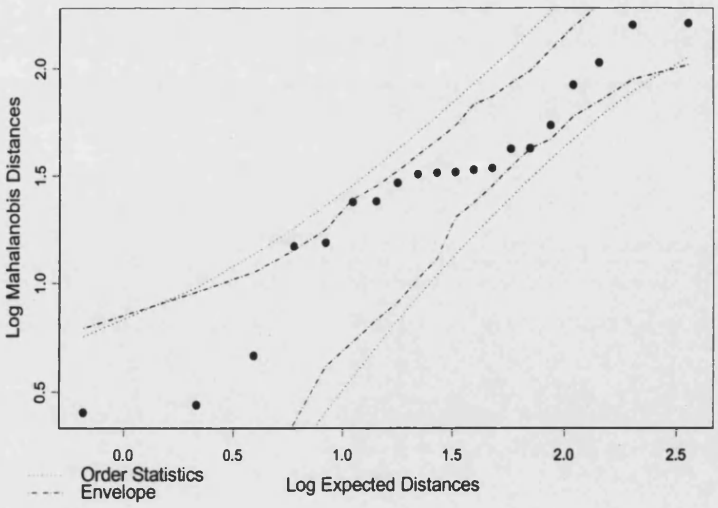
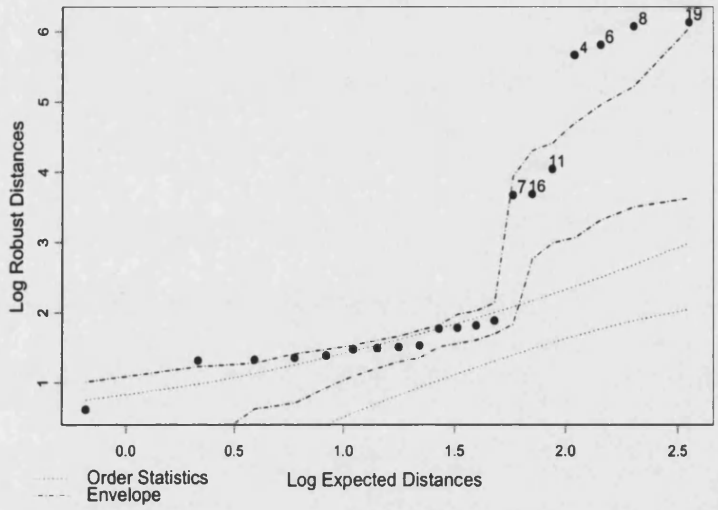


Figure 4.6: Example 1, Woodgravity data, simulation envelopes.

(a) 97.5% Simulation Envelopes for Robust Distances



(b) 97.5 % Simulation Envelope for Robust Distances after deleting the outliers

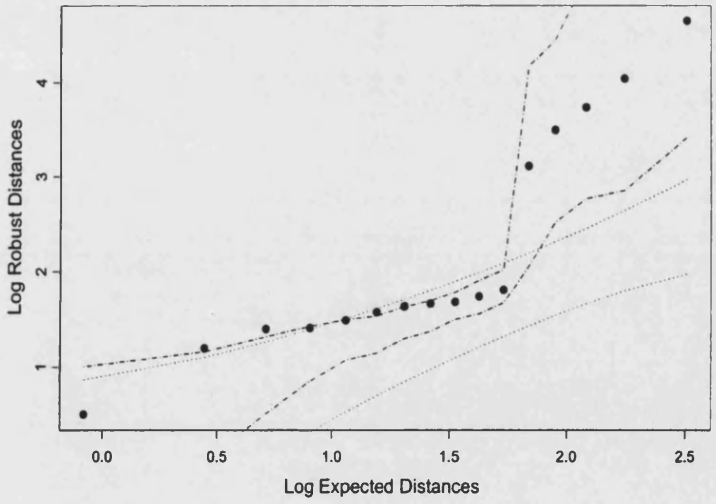
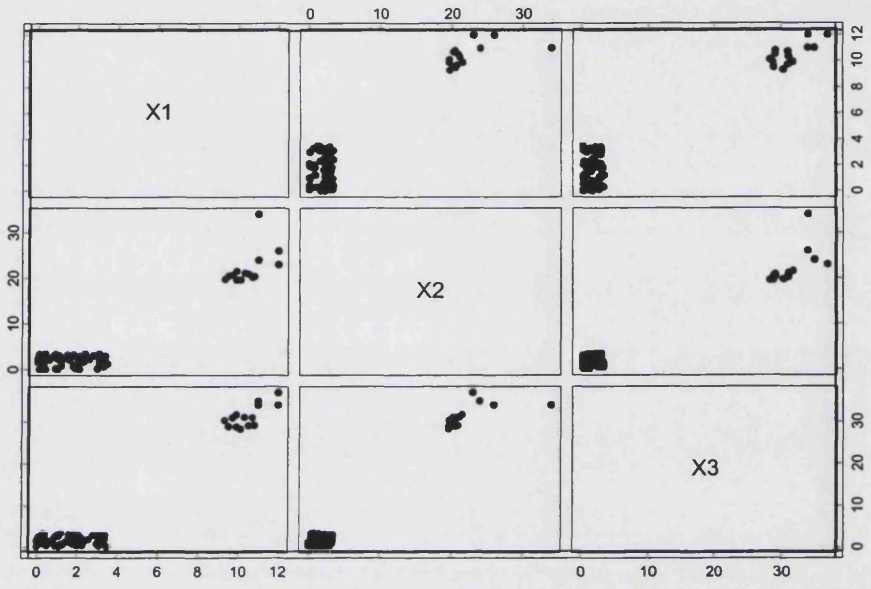


Figure 4.7: Example 2, Hawkins, Bradu and Kass data.

(a) Scatter Plot Matrix



(b) Distance-Distance Plot

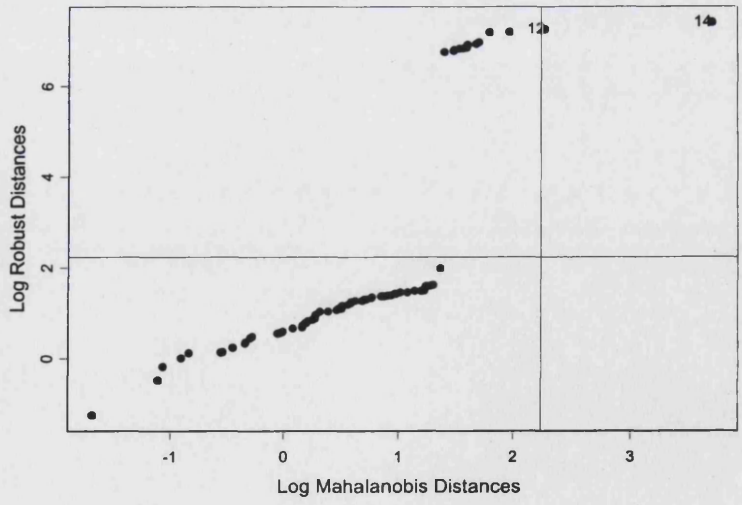
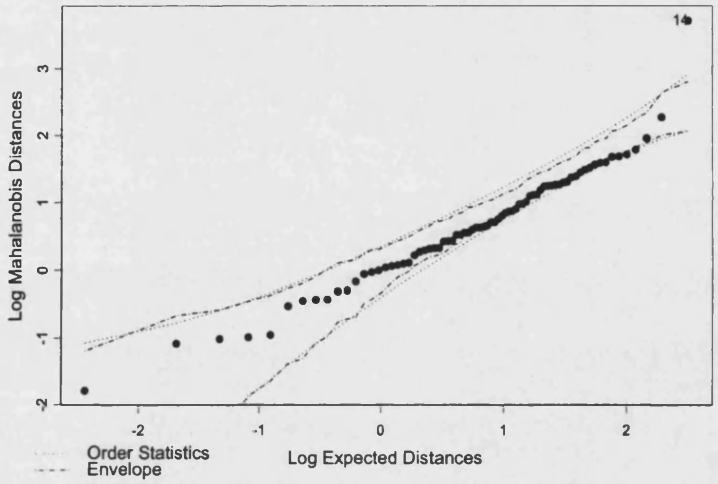


Figure 4.8: Example 2, Hawkins, Bradu and Kass data, simulation envelopes.

(a) 97.5% Simulation Envelopes for Mahalanobis Distances



(b) 97.5% Simulation Envelopes for Robust Distances

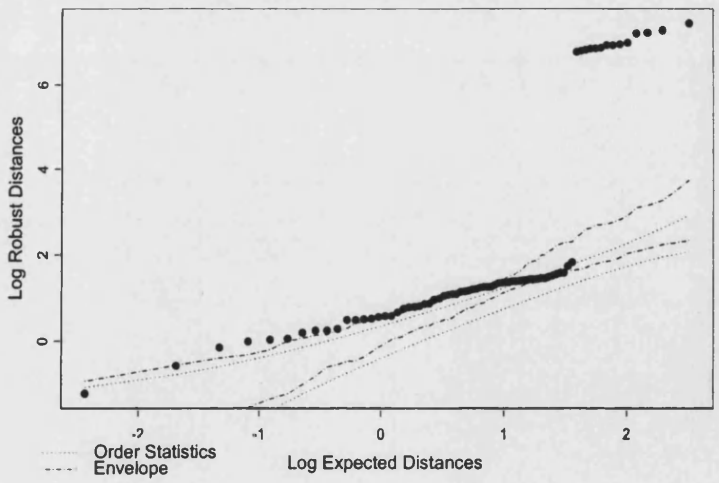
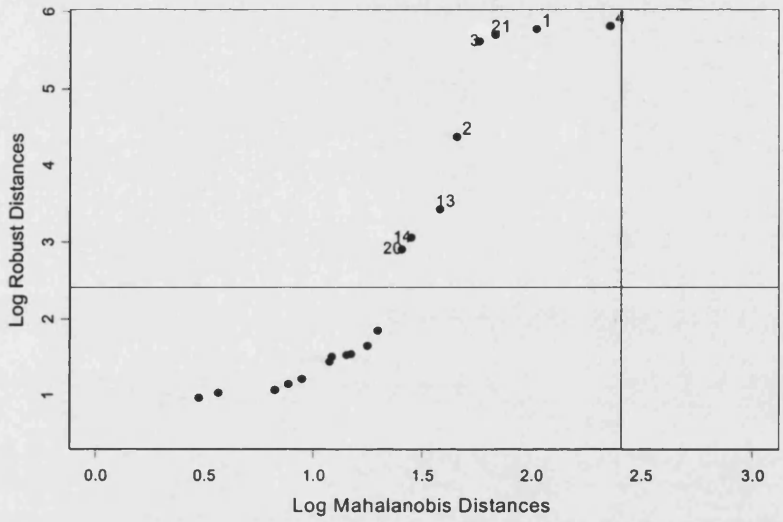
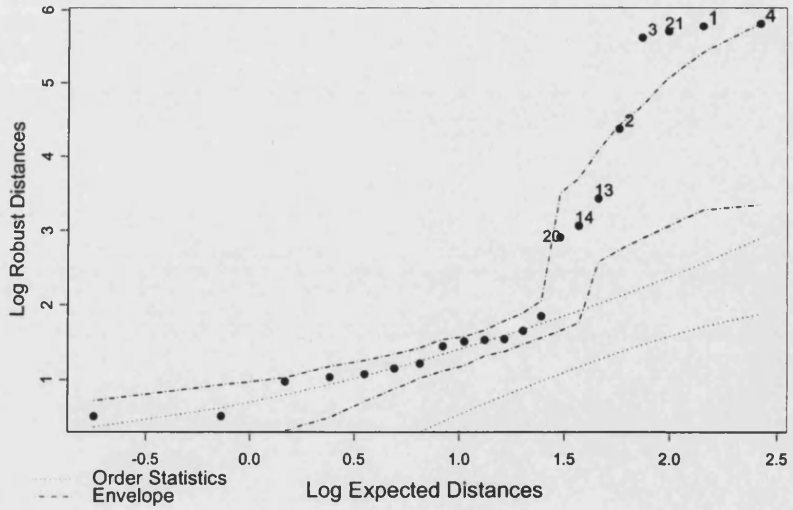


Figure 4.9: Example 3, Stackloss data.

(a) Distance-Distance Plot



(b) 97.5% Simulation Envelopes for Robust Distances



Chapter 5

Robust Detection of Outliers using the Student-t distribution

5.1 Introduction

Chapter 4 has underlined the sensitivity of the Normal distribution to extreme points and how this can affect the detection of multiple outliers. We have coped with this problem by estimating robustly the location and scatter of the distribution, which also allows us to detect groups of outliers avoiding masking and swamping effects. The underlying assumption is that the majority of the data is normally distributed.

An alternative method to deal with extreme points is to remove the assumption of normality and replace it with a more "robust" one. The approach considered consists in assuming a multivariate Student-t distribution, longer tailed than the Normal, in statistical models. The Chapter gives the literature background: the first three sections introduce the multivariate Student-t distribution and examine the properties of the MLE for its parameters. Dempster, Laird, and Rubin (1977) show how to use the EM algorithm to find MLE for the Student-t distribution; an illustration of the method follows in Section 5.4. In addition to the literature, some empirical results on the efficiency of the estimates are shown in Section 5.3.1 and Section 5.5. Finally, we compare this method with high-breakdown point estimators in outlier detection.

5.2 The Multivariate Student-t Distribution

The univariate central t distribution is defined as:

$$t(\nu) = \frac{z}{\sqrt{\chi^2_\nu/\nu}}\,, \tag{5.1}$$

where $z \sim N(0,1)$ and is independent from χ^2_ν.

A generalization of the Student-t distribution comes from considering a vector of independent univariate t's, $y_j = x_j(\sqrt{\chi^2_\nu/\nu})^{-1}$, where $j = 1, 2, ..., p$, where the joint density of the p-variates is the product of the individual densities. In order to introduce a covariance structure, it is assumed that the vector $\mathbf{x} = (x_1, x_2, ..., x_p)$ has a joint multivariate Normal distribution with some nonzero covariances. Therefore, let $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n\}$ be a sample where $\mathbf{y}_i \in \Re^p$, $i = 1, 2, ..., n$. The multivariate Student-t distribution is generated from a multinormal sample of observations in the following way:

$$\mathbf{y}_i = \frac{\mathbf{x}_i}{\sqrt{\tau_i}} \qquad \text{for} \quad i = 1, 2, ..., n\,, \tag{5.2}$$

where $\mathbf{x}_i \sim N_p(\boldsymbol{\mu}\sqrt{\tau_i}, \boldsymbol{\Psi})$ and $\nu\tau_i \sim \chi^2_\nu$. Furthermore, τ_i and $\mathbf{x}_i$ are independent.

If the vector $\boldsymbol{\tau} = (\tau_1, \tau_2, ..., \tau_n)$ is considered as observed, $\mathbf{Y}$ has a weighted Normal distribution, that is:

$$\mathbf{y}_i \overset{ind}{\sim} N_p(\boldsymbol{\mu}, \boldsymbol{\Psi}/\tau_i) \qquad \text{for} \quad i = 1, 2, ..., n\,. \tag{5.3}$$

It should be noted that if $Z \sim \frac{1}{2}\chi^2_{2\alpha}$, then $(1/\beta)Z \sim Gamma(\alpha, \beta)$, where the density of $X \sim Gamma(\alpha, \beta)$ is defined as:

$$f(x; \alpha, \beta) = \frac{1}{\Gamma(\alpha)}\beta^\alpha x^{\alpha-1}\exp\{-\beta x\}.$$

where, for $k > 0$:

$$\Gamma(k) = \int_0^\infty t^{k-1} exp(-t) dt.$$

Therefore, by setting $Z = (\nu/2)\tau$, $\alpha = \nu/2$ and $\beta = \nu/2$, it follows that:

$$\tau_i \overset{iid}{\sim} Gamma(\nu/2, \nu/2). \tag{5.4}$$

The density function of a multivariate Student-t with parameters $\boldsymbol{\mu}$, $\boldsymbol{\Psi}$ and ν, $\mathbf{y} \sim t_p(\boldsymbol{\mu}, \boldsymbol{\Psi}, \nu)$ is:

$$f(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Psi}, \nu) = \frac{\Gamma(\frac{\nu+p}{2})}{\Gamma(\frac{\nu}{2})(\pi\nu)^{p/2}} \times \frac{1}{|\boldsymbol{\Psi}|^{1/2}} \times \{1 + \frac{1}{\nu}\, \delta_{\mathbf{y}}^2\}^{-(\frac{\nu+p}{2})}, \tag{5.5}$$

where $\delta_{\mathbf{y}}^2 = (\mathbf{y} - \boldsymbol{\mu})^T \boldsymbol{\Psi}^{-1} (\mathbf{y} - \boldsymbol{\mu})$.

If $\nu = 1$, (5.5) defines the Cauchy distribution with non-existing moments. If $\nu > 1$, $\boldsymbol{\mu}$ is the mean vector, $\boldsymbol{\Psi}$ is the inner-product matrix and ν the degrees of freedom. If $\nu > 2$, the variance covariance matrix, similarly to the univariate case, is given by $\nu/(\nu - 2)\boldsymbol{\Psi}$.

5.3 Maximum Likelihood Estimation

The log-likelihood function of the model in (5.5) is:

$$\begin{aligned} l_T(\boldsymbol{\mu}, \boldsymbol{\Psi}, \nu; \mathbf{Y}) &= constant + n \ln\left(\Gamma\left(\frac{\nu+p}{2}\right)\right) - n \ln\left(\Gamma\left(\frac{\nu}{2}\right)\right) - \frac{n}{2} |\boldsymbol{\Psi}| \\ &\quad - \frac{np}{2} \ln(\nu) - \left(\frac{\nu+p}{2}\right) \sum_{i=1}^{n} \ln\left(1 + \frac{1}{\nu}\, \delta_i^2\right), \end{aligned} \tag{5.6}$$

where $\delta_i^2 = (\mathbf{y}_i - \boldsymbol{\mu}) \boldsymbol{\Psi}^{-1} (\mathbf{y}_i - \boldsymbol{\mu})^T$.

If the vector of weights is considered as a vector of observations as well as $\mathbf{Y}$, the likelihood can be expressed in an alternative form. In this case, the whole data is $\{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n; \tau_1, \tau_2, ..., \tau_n\}$. Therefore, the density of the data is given by:

$$f(\mathbf{Y}, \tau; \boldsymbol{\mu}, \boldsymbol{\Psi}, \nu) = f_N(\mathbf{Y}|\tau) \times f_G(\tau).$$

As a result, the likelihood of the model can be factorized as follows:

$$l_T(\boldsymbol{\mu}, \boldsymbol{\Psi}, \nu; \mathbf{Y}) = l_N(\boldsymbol{\mu}, \boldsymbol{\Psi}; \mathbf{Y}, \tau) + l_G(\nu; \tau),$$

where, from (5.3) and (5.4):

$$\begin{aligned} l_N(\boldsymbol{\mu}, \boldsymbol{\Psi}; \mathbf{Y}, \tau) &= constant - \frac{n}{2}\ln|\boldsymbol{\Psi}| - \frac{1}{2}\sum_{i=1}^{n}\tau_i(\mathbf{y}_i - \boldsymbol{\mu})^T\boldsymbol{\Psi}^{-1}(\mathbf{y}_i - \boldsymbol{\mu}) \\ &= -\frac{n}{2}\ln|\boldsymbol{\Psi}| - \frac{1}{2}\,\mathrm{tr}\left\{\boldsymbol{\Psi}^{-1}\sum_{i=1}^{n}\tau_i\mathbf{y}_i\mathbf{y}_i^T\right\} + \\ &\quad + \frac{1}{2}\,\boldsymbol{\mu}^T\boldsymbol{\Psi}^{-1}\sum_{i=1}^{n}\tau_i\mathbf{y}_i - \frac{1}{2}\,\boldsymbol{\mu}^T\boldsymbol{\Psi}^{-1}\boldsymbol{\mu}\sum_{i=1}^{n}\tau_i \end{aligned} \tag{5.7}$$

and

$$l_G(\nu; \tau) = constant - n\ln\left(\frac{\nu}{2}\right) + \frac{n\nu}{2}\ln\left(\frac{\nu}{2}\right) + \frac{\nu}{2}\sum_{i=1}^{n}\{\ln(\tau_i) - \tau_i\}. \tag{5.8}$$

The maximization of (5.7) leads to the following estimates of the location and inner-product matrix:

$$\begin{aligned} \hat{\boldsymbol{\mu}}_{MLE} &= \frac{\sum_{i=1}^{n}\tau_i\mathbf{y}_i}{\sum_{i=1}^{n}\tau_i} \\ \hat{\boldsymbol{\Psi}}_{MLE} &= \frac{1}{n}\sum_{i=1}^{n}\tau_i(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_{MLE})(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_{MLE})^T. \end{aligned}$$

For the multivariate Student-t, $\hat{\boldsymbol{\mu}}_{MLE}$ $\hat{\boldsymbol{\Psi}}_{MLE}$ are the weighted MLE (sample mean and variance-covariance matrix) of the Normal model. Conditions for the uniqueness of the solution to the location-scatter estimation problem in the multivariate Student-t case are discussed in Kent, Tyler, and Vardi (1994).

Since there is no analytical solution, the degrees of freedom ν are estimated by solving, via some numerical methods (Newton-Raphson, bisection, etc.), the following equation:

$$-\frac{d\ln(\Gamma(\nu/2))}{d(\nu/2)} + \ln\left(\frac{\nu}{2}\right) + \frac{1}{n}\sum_{i=1}^{n}\{\ln(\tau_i) - \sum_{i=1}^{n}\tau_i\} = 0. \tag{5.9}$$

5.3.1 Distribution of $\hat{\mu}_{MLE}$ and $\hat{\Psi}_{MLE}$

Assuming that the vector of weights $\boldsymbol{\tau} = \tau_1, \tau_2, ..., \tau_n$ is known, the distribution of $\hat{\boldsymbol{\mu}}_{MLE}$ is:

$$\hat{\boldsymbol{\mu}}_{MLE} | \boldsymbol{\tau} \sim N_p(\boldsymbol{\mu}, (\boldsymbol{\tau}^T\mathbf{1})^{-1}\boldsymbol{\Psi}),$$

where $\mathbf{1}$ is a $n \times 1$ vector of ones.

We show that, as in the multinormal case, the conditional distribution of the covariance matrix is a Wishart (Definition 4.3.1).

$$\hat{\boldsymbol{\Psi}}_{MLE} = \frac{1}{n}\sum_{i=1}^{n}\tau_i(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_{MLE})(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_{MLE})^T = \frac{1}{n}\sum_{i=1}^{n}\tau_i\{\mathbf{y}_i\mathbf{y}_i^T - \hat{\boldsymbol{\mu}}_{MLE}\hat{\boldsymbol{\mu}}_{MLE}^T\}.$$

Let $\mathbf{D}$ be a diagonal matrix

$$\mathbf{D} = \begin{bmatrix} \tau_1 & & 0 \\ & \ddots & \\ 0 & & \tau_n \end{bmatrix}.$$

Then:

$$\hat{\boldsymbol{\Psi}}_{MLE} = n^{-1}\{\mathbf{Y}^T\mathbf{D}\mathbf{Y} - \mathbf{Y}^T\mathbf{D}\mathbf{1}\mathbf{1}^T\mathbf{D}\mathbf{Y}(\boldsymbol{\tau}^T\mathbf{1})^{-1}\},$$

where:

$$\hat{\boldsymbol{\mu}}_{MLE} = \frac{\sum_{i=1}^{n}\tau_i\mathbf{y}_i}{\sum_{i=1}^{n}\tau_i} = \mathbf{Y}^T\mathbf{D}\mathbf{1}(\boldsymbol{\tau}^T\mathbf{1})^{-1}.$$

Letting $\mathbf{H} = \mathbf{D} - \mathbf{D}\mathbf{1}\mathbf{1}^T\mathbf{D}(\boldsymbol{\tau}^T\mathbf{1})^{-1}$, the covariance matrix is rewritten as

$$n\hat{\boldsymbol{\Psi}}_{MLE} = \mathbf{Y}^T\mathbf{H}\mathbf{Y}. \tag{5.10}$$

$\mathbf{D}$ is a diagonal matrix, symmetric but not idempotent and $\mathbf{H}$ has elements:

$$\begin{aligned} h_{ii} &= \tau_i - \tau_i^2 / \sum_i \tau_i \\ h_{ij} &= -\tau_i\tau_j / \sum_i \tau_i, \qquad \text{for} \quad i \neq j\,. \end{aligned}$$

Since $\mathbf{H}$ is symmetric but not idempotent the distribution of (5.10) is a weighted sum of independent Wishart distributions where the weights are the eigenvalues of $\mathbf{H}$ (Mardia, Kent, and Bibby 1982, Theorem 3.4.4(a)). Some simulation studies are carried out to describe the efficiency and bias of $\hat{\boldsymbol{\mu}}_{MLE}$ and $\hat{\boldsymbol{\Psi}}_{MLE}$ in Section 5.5.

5.4 The EM algorithm: general idea

The EM is an iterative procedure designed to provide MLE when the data have some missing observations. The framework comprises two sample spaces X and Y which generate two sets of realizations: a complete set, $\mathbf{x}$ and a latent, or incomplete one, $\mathbf{y}$. The algorithm works on a data augmentation principle: at each iteration the observed (incomplete) data are augmented by missing data and/or parameter estimates allowing the updating of the maximum-likelihood estimates. In other words, the complete set can't be observed directly, but only through $\mathbf{y}$.

Let $\theta = \{\theta_1, \theta_2, ..., \theta_n\}$ be a set of parameters and $g(\mathbf{y}; \theta)$ the density of the vector of observations $\mathbf{y}$, that is the likelihood for θ, given the set of observed data, $L(\theta|\ \mathbf{y})$. At each step, the incomplete set is "estimated" through the relationship:

$$g(\mathbf{y}; \theta) = \int_Y f(\mathbf{x}; \theta) d\mathbf{x}, \tag{5.11}$$

where $f(\mathbf{x}; \theta)$ is the density of the complete data vector. Therefore, given a set $\mathbf{y}$ of observations, the purpose of the algorithm is to find the parameter estimate that maximizes the likelihood of $\mathbf{y}$, found through (5.11).

The algorithm consists of two stages. The expectation step involves the computation of the log-likelihood for the parameter θ from the complete data, $\log f(\mathbf{x})$. Since the latter are not observed, the likelihood is replaced by its conditional expectation.

Let $Q(\theta^{(t)}|\ \theta^{(t-1)})$ be the expectation of the log-likelihood function at iteration t, given the current estimate of the parameter. The E-step computes:

$$Q(\theta^{(t)}|\ \theta^{(t-1)}) = E[\log f(\mathbf{x} \mid \mathbf{y}, \theta^{(t)})|\ \mathbf{y}, \theta^{(t-1)}],$$

where $\theta^{(t)}$ is the parameter estimates at the step t.

In the maximization step the aim is finding θ that maximizes $Q(\theta|\ \theta^{(t)})$, where

$$Q(\theta|\ \theta^{(t)}) = f(\mathbf{x}; \theta).$$

The parameter estimates allow updating the conditional expectations in the E-step. The E and M-steps are iterated until the difference

$$L(\theta^{(t+1)}) - L(\theta^{(t)})$$

is small, the accuracy being chosen arbitrarily.

The paper by Dempster, Laird, and Rubin (1977) discusses some basic properties of the EM. It is shown that in two subsequent steps of the algorithm the likelihood is not decreased, that is:

$$L(\theta^{(t+1)}) \geq L(\theta^{(t)}),$$

where the equality holds if the conditional expected likelihood satisfies:

$$Q(\theta^{(t+1)}|\ \theta^{(t)}) \geq Q(\theta|\ \theta^{(t)}).$$

The paper also gives evidence of convergence of the estimates to the maximum of the likelihood function, where the rate depends on the information about θ contained in the observed data. This implies that if there are many data missing the convergence to the optimum will be slower.

5.4.1 ML estimation with known degrees of freedom

In Section 5.2, when defining the multivariate Student-t t distribution, the vector $\boldsymbol{\tau}$ was assumed to be known. Nevertheless, in practice, $\boldsymbol{\tau}$ is a latent, rather than an observed variable. In the simplest case, $\boldsymbol{\tau}$ is missing and ν is known so that "only" the MLE for $\{\boldsymbol{\mu}, \boldsymbol{\Psi}\}$ need to be found.

The E-Step imputes the missing observations $\{\tau_1, \tau_2, ..., \tau_n\}$ with their conditional expectation given the observed data $\mathbf{Y}$ and the $\{\hat{\boldsymbol{\mu}}_{MLE}, \hat{\boldsymbol{\Psi}}_{MLE}\}$. From (5.6) it follows that, at step $(t+1)$:

$$\hat{\tau}_i^{(t+1)} = E(\tau_i \mid \hat{\boldsymbol{\mu}}^{(t)}, \hat{\boldsymbol{\Psi}}^{(t)}, \nu) = \frac{\nu + p}{\nu + \delta_i^{2\,(t)}},$$

where $\delta_i^{2\,(t)}$ is a function of the current estimates for $\boldsymbol{\mu}$ and $\boldsymbol{\Psi}$. The procedure can be easily extended to the case where there are some missing observations in $\mathbf{Y}$. Therefore, the E-step would also compute the conditional expectations for the missing data.

The M-step finds the MLE of the parameters by calculating:

$$\begin{aligned}
\hat{\boldsymbol{\mu}}^{(t+1)} &= \frac{\sum_{i=1}^n \hat{\tau}_i^{(t+1)} \mathbf{y}_i}{\sum_{i=1}^n \hat{\tau}_i^{(t+1)}} \\
\hat{\boldsymbol{\Psi}}^{(t+1)} &= \frac{1}{n} \sum_{i=1}^n \hat{\tau}_i^{(t+1)} (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_{MLE}^{(t+1)})(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_{MLE}^{(t+1)})^T,
\end{aligned}$$

that are the weighted least squares estimators for $\{\boldsymbol{\mu}, \boldsymbol{\Psi}\}$.

The algorithm iterates until convergence of the likelihood (5.6).

In the case of ν fixed, the EM algorithm involves the calculation of the weights $\hat{\tau}_i$, given the observations $\mathbf{Y}$ and the current estimate $\{\hat{\boldsymbol{\mu}}^{(t)}, \hat{\boldsymbol{\Psi}}^{(t)}\}$. When ν is known, the likelihood has the form of an exponential family (5.7) where the sufficient statistics for the parameters are $S_{\tau \mathbf{y}} = \sum_i \tau_i \mathbf{y}_i$, $S_{\tau \mathbf{y}\mathbf{y}} = \sum_i \tau_i \mathbf{y}_i \mathbf{y}_i^T$ and $S_\tau = \sum_i \tau_i$. Therefore, the weighted least squares estimates at the M-step $(t+1)$ are:

$$\hat{\boldsymbol{\mu}}^{(t+1)} = \frac{S_{\tau \mathbf{y}}^{(t+1)}}{S_\tau^{(t+1)}}$$

$$\hat{\Psi}^{(t+1)} = \frac{1}{n}\left(S_{\tau yy}^{(t+1)} - \frac{1}{S_{\tau}^{(t+1)}} S_{\tau y}^{(t+1)} (S_{\tau y}^{(t+1)})^T\right).$$

Arslan, Constable, and Kent (1995) propose two alternative ways of accelerating the EM algorithm for estimation of location and scatter, given the degrees of freedom.

5.4.2 ML estimation with unknown degrees of freedom

When ν is not observed, the EM finds simultaneously the MLE for $\{\mu, \Psi, \nu\}$. The algorithm is modified as follows: in the E-step, ν is replaced by the current estimate $\hat{\nu}^{(t)}$. Therefore,

$$\hat{\tau}_i^{(t+1)} = E(\tau_i \mid \hat{\mu}^{(t)}, \hat{\Psi}^{(t)}, \hat{\nu}^{(t)}) = \frac{\hat{\nu}^{(t)} + p}{\hat{\nu}^{(t)} + \delta_i^{2\,(t)}}.$$

From (5.8) the sufficient statistics for ν is $S_{\tau\tau} = \sum_{i=1}^{n}\{\ln(\tau_i) - \tau_i\}$.

The estimation of the degrees of freedom requires the computation of the conditional expectation of $S_{\tau\tau}$, given the observed data and the current estimates for the parameters $\{\mu, \Psi, \nu\}$:

$$\begin{aligned}\hat{S}_{\tau\tau} = E(S_{\tau\tau} \mid \hat{\mu}^{(t)}, \hat{\Psi}^{(t)}, \hat{\nu}^{(t)}) &= \sum_{i=1}^{n}\left\{\phi\left(\frac{p+\hat{\nu}^{(t)}}{2}\right) - \ln\left(\frac{p+\hat{\nu}^{(t)}}{2}\right)\right\} \\ &\quad + \sum_{i=1}^{n}\{\ln(\hat{\tau}_i^{(t+1)}) - \hat{\tau}_i^{(t+1)}\},\end{aligned}$$

where $\phi(x) = d\ln(\Gamma(x))/d(x)$.

The M-step consists in the separate maximization of the L_N (5.7) with respect to (μ, Ψ) and of L_G (5.8) over ν. Therefore, the M-step for (μ, Ψ) in case of unknown degrees of freedom is the same as for ν fixed. The estimation of ν requires a more difficult and computationally expensive procedure, that is finding the root of the equation:

$$-\phi\left(\frac{\nu}{2}\right) + \ln\left(\frac{\nu}{2}\right) + \frac{1}{n}\hat{S}_{\tau\tau} + 1 = 0. \tag{5.12}$$

(5.12) differs from the corresponding (5.9) by the replacement of the sufficient statistics for τ_i with their conditional expectations and by an additional term that can be interpreted as a correction for the mean value imputation for the missing observations $\tau_1, \tau_2, ..., \tau_n$. For further reference regarding this problem, see Liu and Rubin (1995).

Therefore, the E-step is the same as in Section 5.4.1 with the exception of the additional calculation of the conditional expectations for the sufficient statistics of ν. The M-step differs from the previous case because it requires the maximization of the Gamma likelihood to find the current estimate of ν.

The convergence of the EM when the degrees of freedom are unknown is very slow, since the convergence of two likelihood functions is required. There is a variety of studies on possible extensions of the EM that can be more efficient. The ECM algorithm by Meng and Rubin (1993), the ECME by Meng and Rubin (1993) and a further expansion of this last by Liu and Rubin (1995) are some of the main suggestions offered by the literature.

5.5 Empirical Results on $\hat{\mu}_{MLE}$ and $\hat{\Psi}_{MLE}$

An empirical study on the bias and efficiency (relative to the Normal distribution) has been conducted. The results are summarized in Table 5.1 to Table 5.8.

In Table 5.1 and Table 5.2 the data are simulated from a multivariate Student-t with 3 degrees of freedom from which we have drawn 500 independent samples. The MLE for the multivariate Student-t, t3, are assumed to be known. These are compared with the MLE from a Normal model and with the robust estimates via MCD. In other words, the aim is to measure the drop in the bias and the change in the variability of the estimates when we fit a Normal or a robust MCD model. We also observe the differences in the variability and bias for increasing sample sizes and dimensions.

Table 5.1 shows the bias range across the different models. As expected, the bias is minimum for the Student-t, where the difference between the model performances is greater for $\hat{\Psi}$ than for $\hat{\mu}$. The MCD fit has a better bias and variability than

the Normal on a t. With the sample size growing, the performances of the models become more similar.

Table 5.2 reflects the same behaviour of the estimates as Table 5.1.

The same study is repeated in Table 5.3 to Table 5.4, where bias of the first two elements in the mean vector and that of the first diagonal and off-diagonal elements in the scatter matrix Ψ are displayed. In order to reduce the sample variability new observations are added to the data, rather than sampling independent sets of observations. The results confirm those in Table 5.1 and Table 5.2: the Normal model has a larger bias compared to the other two models (t3 and MCD). The difference in the performance improves when sample size increases with respect to the dimension. The bias for the Normal fit is the same for increasing dimensions since we are considering only the first two variables. When n grows from 20 to 50 the bias for both the mean vector and the scatter matrix increase before smoothing down for n=100. The jump in the variability occurs when the furthest observations are included in the sample. This is confirmed by Figure 5.1 that plots the change in the determinant of the sample variance-covariance matrix for increasing sample sizes. The graph shows the behaviour of one sample, although different samples would have a similar pattern, that is one or a few peak occurring when the most remote observations are included in the sample.

Table 5.7 compares the range in bias and efficiency (with respect to the Normal model) similarly to Table 5.1. In this case sample size and dimensions are fixed: 500 samples of 100 observations on 3 variables are replicated from a multivariate Student-t with 4 degrees of freedom. The results of fitting Student-t by fixing different values of degrees of freedom are compared with the adaptive robust procedure, where ν is estimated, and with the Normal model, for which the average estimate of ν and, in brackets, its standard deviation are calculated. The table show that a "wrong guess" of the degrees of freedom does not affect significantly the bias and the efficiency of the MLE for $\hat{\mu}$. On the contrary, bias and efficiency of $\hat{\Psi}_{MLE}$ increase sensibly to changes in ν. The estimate for the degrees of freedom have a reasonably small bias, .65, although they vary quite significantly over the samples.

The same experiment is repeated in Table 5.8, where the bias and efficiency of

some single elements of $\hat{\mu}$ and $\hat{\Psi}$ are compared. The comments are similar to those for Table 5.7.

The results found from this simulation study appear in favour of the model with unknown degrees of freedom, if we look at the efficiency for the MLE of μ and Ψ. Nevertheless, the degrees of freedom are estimated with low precision.

Furthermore, the bias and the standard deviation of the estimates are much smaller compared to the Normal model, particularly for the inner product matrix $\hat{\Psi}$. This confirms that the Normal model is not well suited for data with extreme points.

5.6 Outlier Diagnostics: Weighted Mahalanobis Distances

Assuming τ known, we define the Weighted Mahalanobis Distances (WMD) as:

$$WMD_i = \tau_i \delta_i^2 = \tau_i (\mathbf{y}_i - \boldsymbol{\mu})^T \Psi^{-1} (\mathbf{y}_i - \boldsymbol{\mu}),$$

where τ_i is the scaling parameter that, mixed with a Normal distribution, gives the Student-t, defined as in (5.2). If these weights are observed, then:

$$\tau_i \delta_i^2 \sim \chi_p^2. \tag{5.13}$$

Alternatively, $\nu\tau_i$ is a random variable with distribution χ_ν^2. From (5.13), δ_i^2 is a χ_p^2, assuming τ_i are 1 for all i's. Therefore:

$$\delta_i^2 / \tau_i p \sim F_{p,\nu}. \tag{5.14}$$

When $\nu \longrightarrow \infty$, τ_i converges to 1 and $\mathbf{y}_i$ becomes a multinormal with mean μ and variance covariance matrix Ψ. The WMD are equivalent to the MD_i for a Normal distribution, described in (3.12). As a result, τ_i can be interpreted as a robustifying parameter, assigning low weights to the observations with large distances. This last result can be observed from the expression for the expected value of $\boldsymbol{\tau}$:

$$E(\tau_i \mid \mu, \Psi, \nu) = \frac{\nu + p}{\nu + \delta_i^2} = 1 + \frac{p - \delta_i^2}{\nu + \delta_i^2}.$$

For large distances the difference $p - \delta_i^2$ is negative with the result of down-weighting the i-th observation.

In addition to τ, also ν affects the robustness of the model. From (5.14), for lower ν the degree with which the observations are down-weighted increases.

The behaviour of the Mahalanobis-like distances is explained by Figure 5.2. The plot shows the maximum, 99, 98.5, 98 and 97.5% envelopes of the distances and their average for four different combinations of sample sizes and dimensions. The quantiles are plotted against the Chisquare, that is the asymptotic distribution. The Weighted Mahalanobis Distances are longer tailed than the equivalent distances for the Normal case, with the extreme points lying far from the remaining observations. Increasing the degrees of freedom, the envelope becomes more narrow approaching to the Normal model type, Figure 4.2. On the contrary, when the dimension grows, the distribution becomes longer tailed. In conclusion, when tuning the parameter τ for robustness, dimension and size of the sample in addition to the degrees of freedom should be taken into account.

5.7 Example 1: Univariate Linear Regression on Stackloss Data

This example refers to Lange, Little, and Taylor (1989). It is linear regression model where the dependent variable has a longer tailed distribution than the Normal:

$$y_i \overset{ind}{\sim} t(\mu(\boldsymbol{\beta}, \mathbf{x}_i), \Psi^2, \nu) \qquad \text{for} \quad i = 1, 2, ..., n\,. \tag{5.15}$$

$\mathbf{y} = \{y_1, y_2, ..., y_n\}$ depends linearly on a matrix $n \times p$ of regressors, $\mathbf{X}$, and $\boldsymbol{\beta}$ is the vector of parameters, $k \times 1$, therefore, $\mu(\boldsymbol{\beta}, \mathbf{x}_i) = \mathbf{x}_i^T \boldsymbol{\beta}$. Similarly to (5.3):

$$y_i \mid \tau_i \overset{ind}{\sim} N(\mu(\boldsymbol{\beta}, \mathbf{x}_i), \Psi^2/\tau_i) \qquad \text{for} \quad i = 1, 2, ..., n\,.$$

The maximum likelihood estimates are obtained via the EM given that ν is known. The general idea is to find the parameter estimates by fitting re-weighted least squares iteratively, until convergence of the likelihood. At iteration (t), the E-step computes the weights, that is the conditional expectation of τ_i:

$$\hat{\tau}_i^{(t+1)} = E(\tau_i \mid y_i, \beta^{(t)}, \Psi^{2\,(t)}, \nu) = \frac{\nu + 1}{\nu + \delta_i^{2\,(t)}},$$

where $\delta_i^{2\,(t)} = \{y_i - \mathbf{x}_i^T \beta^{(t)}\}/\Psi^{2\,(t)}$. Given the estimates of the weights, the M-Step finds the parameter estimates by minimizing a weighted sum of squares:

$$\beta^{(t+1)} = \min_{\beta} \{\sum_i \hat{\tau}_i^{(t)} (y_i - \mathbf{x}_i^T \beta)^2\}.$$

The parameter Ψ^2 is updated as it follows:

$$\Psi^{2\,(t+1)} = \frac{1}{n} \sum_i \{y_i - \mathbf{x}_i^T \beta^{(t)}\}^2.$$

The log-likelihood has a very similar form as the Normal model. Considering ν observed and ignoring constants:

$$l(\beta, \Psi) \propto -\frac{n}{2} \ln(\Psi^2) - \frac{1}{2\Psi^2} \sum_{i=1}^{n} \hat{\tau}_i (y_i - \mathbf{x}_i^T \beta)^2.$$

The Stackloss data set has been presented in Section 4.8.4. There are 3 regressors and one dependent variable observed on 21 units. The outliers identified by the literature are observations 1, 3, 4 and 21. Table 5.9 illustrates the results from fitting Student-t models with different degrees of freedom and the Normal model. The fit of the Student-t model is good: the log-likelihoods appear to be between the Normal and the Normal fitted without the 4 outliers. Furthermore, the likelihoods of the t-models are closer to the Normal without outliers than to the Normal fit.

Figure 5.3 is a plot of the sorted weights. The smallest 4 weights are the outliers, that correspond to those identified by Brownlee and previously in Chapter 4. For these observations, clearly separated from the rest of the data, the weights are 0.44, 0.36, 0.19 and 0.13 (Table 5.12).

5.7.1 Example 2: Stackloss Data

The stackloss data example is here proposed assuming there is no linear relationship between the stackloss and the carriers. The fitted models give robust estimates for the mean and the scatter matrix of the data, including both the regressors and the dependent variable. Table 5.10 displays the result from the fit. For $\nu = 10$ the t-model provides estimates for the mean that are very close to the Normal model. The convergence of Ψ appears, instead, slower. The table displays also the results from the robust MCD estimates. The coefficients for these lasts are very close to the outlier-free ones, showing a better fit than the Student-t. Compared to the regression model, there are 2 more outliers detected from the t3 model: observations 2 and 17. Observation number 2 is also labelled as an outlier in the MCD model of Figure 4.9.

Figure 5.4 is the profile likelihood as a function of the degrees of freedom. The plot shows the t-model convergence to the Normal (top line) for increasing degrees of freedom.

5.7.2 Example 3: Hawkins, Bradu and Kass

This artificial data set evidences clearly the outliers (the first fourteen observations) just by a simple scatter plot of the data, Section 4.8.3. This justifies its popularity in the robust statistic literature.

The results from the fit of the Normal and robust models are shown in Table 5.11, commented in a similar way as those for the Stackloss data example. The best fit in this case comes from the MCD providing the closest estimates to the fit of the Normal outlier-free model.

The last two graphs (Figure 5.6, panels (a) and (b)) are Distance-Distance plots for the Normal, Student-t and MCD fits. The solid line is the Chisquare 0.975 quantile. The MCD model seems the "most robust", clearly separating the data into two groups of observations. The Weighted Mahalanobis Distances are able to detect the outliers, but they are closer to the Normal model. The classical MD are not able to separate the outliers from the main data.

5.8 Conclusions

The Chapter explores outlier detection and robust modelling using the multivariate Student-t. The distributions of the MLE for the mean and inner product matrix are discussed. Since it is not possible to find a theoretical form for the distribution of Ψ, a simulation study has been carried out. If we assume ν fixed (not estimated), it is found that the efficiency of the estimates for the mean does not vary significantly for changes in the degrees of freedom. On the contrary, bias and efficiency of $\hat{\Psi}$ are highly affected by a "wrong guess" of ν. In addition the precision of the estimate of ν is very low.

Furthermore, some real data examples of maximum likelihood estimation of the Student-t parameters are compared to the estimates from the Normal and "very" robust model (MCD). The results confirm the Student-t ability of dealing with outliers in a wide range of settings. The quality of the fits appears good and close to the Normal model without outliers.

The purpose of using a Student-t distribution is to build a model capable to capture the observations lying far from the center of the distribution. Allowing different degrees of freedom, it is possible to vary the "longtailness" of the distribution.

Given sufficient data, ν can be estimated, as well as μ and Ψ, by likelihood methods, Section 5.4.2. However, the approach used in this Chapter chooses the degrees of freedom a priori. The computational effort and the low accuracy of the estimates for ν make this method less attractive for practical purposes. Since usually ν varies in $\Re^+$, the estimates will correspond to local rather than global maxima. Furthermore, the EM algorithm can sometimes converge to points where the likelihood is not bounded, as shown in a study of radioimmunossay data (Lange, Little, and Taylor 1989, Fernandez and Steel 1999; Lange and Sinsheimer 1993).

It can be argued that there are distributions, other than the t, that could be used for the same purpose. Some authors consider the slash or other Normal/independent type distributions. Lange and Sinsheimer (1993) describe the properties of the Normal/independent type and compare their performances in some robust regression examples on real data sets. A common feature of these distributions, in addition

to the long tails, is the possibility of applying the EM algorithm to get maximum likelihood estimates of the parameters. We have chosen the Student-t because of its simplicity: the likelihood has a simple form, similar to the Normal distribution (if ν is fixed, that is not estimated) and there is only one parameter, ν, to robustify the distribution.

Table 5.1: Minimum and maximum (.) bias for $\hat{\mu}_{MLE}$ and $\hat{\Psi}_{MLE}$ on a multivariate Student-t with $\nu = 3$. t3 is estimated by fixing the degrees of freedom.

$\hat{\mu}_{MLE}$									
n	20	50	100	20	50	100	20	50	100
p	2			4			8		
t3	-0.2439	-0.2655	-0.2448	-0.1716	-0.1922	-0.1788	-0.0171	-0.013	-0.0073
	(0.1718)	(0.1624)	(0.1563)	(0.1858)	(0.1692)	(0.1774)	(0.0208)	(0.0112)	(0.0067)
Normal	-0.2537	-0.2619	-0.2396	-0.1855	-0.1834	-0.1718	-0.0179	-0.0108	-0.0159
	(-0.1907)	(0.1665)	(0.1588)	(0.1922)	(0.1774)	(0.1738)	(0.0181)	(0.0096)	(0.0116)
MCD	-0.239	-0.2607	-0.2462	-0.1705	-0.1868	-0.1803	-0.0314	-0.0086	-0.009
	(0.164)	(0.158)	(0.1588)	(0.165)	(0.17)	(0.178)	(0.0237)	(0.0157)	(0.013)

$\hat{\Psi}_{MLE}$									
n	20	50	100	20	50	100	20	50	100
p	2			4			8		
t3	0.2396	0.1416	0.1003	0.2386	0.1526	0.1039	0.2328	0.1385	0.0921
	(0.2761)	(0.1668)	(0.1117)	(0.2952)	(0.1713)	(0.1247)	(0.2952)	(0.1668)	(0.122)
Normal	0.3597	0.2257	0.1478	0.3409	0.2148	0.1534	0.3473	0.2018	0.1508
	(0.4111)	(0.252)	(0.1575)	(0.4424)	(0.2761)	(0.1916)	(0.4458)	(0.2679)	(0.1899)
MCD	0.2852	0.1603	0.1111	0.2875	0.1775	0.119	0.2776	0.1663	0.1059
	(0.3399)	(0.1835)	(0.1241)	(0.3631)	(0.2041)	(0.142)	(0.3499)	(0.2027)	(0.1424)

Table 5.2: Minimum and maximum (.) standard deviation for $\hat{\mu}_{MLE}$ and $\hat{\Psi}_{MLE}$.

$\hat{\mu}$									
n	20	50	100	20	50	100	20	50	100
p	2			4			8		
t3	-0.4545	-0.4397	-0.4501	-1.383	-1.3871	-1.3808	-0.0774	-0.0789	-0.0759
	(1.143)	(1.1045)	(1.0989)	(1.3092)	(1.3087)	(1.2801)	(0.1701)	(0.1127)	(0.1151)
Normal	-0.4659	-0.076	-0.0447	-1.7176	-1.76	-1.7371	-0.6705	-0.5031	-0.4221
	(3.4077)	(3.0229)	(2.9054)	(3.4285)	(3.6703)	(3.9302)	(2.4579)	(2.2589)	(2.4757)
MCD	-0.6045	-0.4856	-0.4436	-1.3348	-1.3568	-1.3712	-0.094	-0.2652	-0.2094
	(0.9548)	(1.0425)	(1.1125)	(1.0724)	(1.1584)	(1.2372)	(0.293)	(0.0649)	(0.0664)

$\hat{\Psi}$									
n	20	50	100	20	50	100	20	50	100
p	2			4			8		
t3	0.2551	0.157	0.0995	0.2476	0.1465	0.104	0.2339	0.1304	0.096
	(0.4833)	(0.2684)	(0.1941)	(0.5129)	(0.32)	(0.2291)	(0.5749)	(0.3364)	(0.2256)
Normal	6.7486	1.5283	0.9494	1.4031	1.8588	1.7061	1.8176	1.1128	0.8577
	(8.3417)	(4.0192)	(1.7286)	(4.1135)	(4.7136)	(7.4994)	(7.3008)	(10.1487)	(4.2035)
MCD	0.3576	0.224	0.1635	0.3324	0.1912	0.1386	0.3512	0.1593	0.1161
	(0.5837)	(0.3486)	(0.2688)	(0.6256)	(0.3819)	(0.2747)	(1.0513)	(0.3622)	(0.2601)

Table 5.3: Bias of the first two elements of the estimated location vector of a simulated multivariate Student-t

n	20	50	100	20	50	100	20	50	100
p	2			4			8		
$\hat{\mu}_1$(t3)	0.2021	0.1305	0.0929	0.1930	0.1244	0.0881	0.1952	0.1201	0.0872
$\hat{\mu}_2$(t3)	0.2302	0.1388	0.1025	0.2130	0.1356	0.0970	0.2157	0.1292	0.0914
$\hat{\mu}_1$(Normal)	0.2762	0.1929	0.1345	0.2762	0.1929	0.1345	0.2762	0.1929	0.1345
$\hat{\mu}_2$(Normal)	0.3049	0.1987	0.1450	0.3049	0.1987	0.1450	0.3049	0.1987	0.1450
$\hat{\mu}_1$(MCD)	0.2478	0.1478	0.0997	0.2323	0.1481	0.0993	0.2394	0.1489	0.1008
$\hat{\mu}_2$(MCD)	0.2740	0.1633	0.1182	0.2758	0.1649	0.1101	0.2602	0.1615	0.1067

Table 5.4: Bias of the first diagonal and off-diagonal elements of the estimated scatter matrix of a simulated multivariate Student-t

n	20	50	100	20	50	100	20	50	100
p	2			4			8		
$\hat{\Psi}_{1,1}$(t3)	0.3307	0.2135	0.1516	0.3181	0.1993	0.1435	0.3190	0.1936	0.1381
$\hat{\Psi}_{1,2}$(t3)	0.2332	0.1440	0.1100	0.2242	0.1404	0.1051	0.2177	0.1395	0.1002
$\hat{\Psi}_{1,1}$(Normal)	1.6823	2.0068	1.9147	1.6823	2.0068	1.9147	1.6823	2.0068	1.9147
$\hat{\Psi}_{1,1}$(Normal)	0.8417	0.8599	0.6454	0.8417	0.8599	0.6454	0.8417	0.8599	0.6454
$\hat{\Psi}_{1,1}$(MCD)	0.4417	0.2893	0.2079	0.4099	0.2600	0.1814	0.4529	0.2545	0.1766
$\hat{\Psi}_{1,1}$(MCD)	0.3391	0.2343	0.1671	0.3090	0.1927	0.1398	0.3313	0.1703	0.1204

Table 5.5: Efficiency of the first two elements of the estimated location vector of a simulated multivariate Student-t

n	20	50	100	20	50	100	20	50	100
p	2			4			8		
$\hat{\mu}_1$(t3)	0.5331	0.4612	0.4537	0.4754	0.4171	0.4037	0.4798	0.3949	0.3935
$\hat{\mu}_2$(t3)	0.5437	0.4718	0.5004	0.4885	0.4386	0.4432	0.4948	0.4059	0.3919
$\hat{\mu}_1$(MCD)	0.8139	0.5834	0.5369	0.7022	0.5804	0.5107	0.7144	0.5751	0.5375
$\hat{\mu}_2$(MCD)	0.7535	0.6469	0.6365	0.7975	0.6618	0.5673	0.6943	0.6217	0.5385

Table 5.6: Efficiency of the first diagonal and off-diagonal elements of the estimated scatter matrix of a simulated multivariate Student-t

n	20	50	100	20	50	100	20	50	100
p	2			4			8		
$\hat{\Psi}_{1,1}$(t3)	0.0181	0.0055	0.0082	0.0169	0.0050	0.0073	0.0178	0.0047	0.0068
$\hat{\Psi}_{1,2}$(t3)	0.0313	0.0061	0.0099	0.0284	0.0057	0.0091	0.0280	0.0057	0.0082
$\hat{\Psi}_{1,1}$(MCD)	0.0305	0.0103	0.0148	0.0248	0.0073	0.0120	0.0406	0.0061	0.0097
$\hat{\Psi}_{1,2}$(MCD)	0.0702	0.0174	0.0241	0.0574	0.0104	0.0168	0.0695	0.0089	0.0127

Table 5.7: Minimum and Maximum bias and efficiency of the MLE for μ and Ψ on a simulated multivariate Student-t with $\nu = 3$.

Model	ν	$\hat{\mu}$			
		Min Bias	Max Bias	Min Eff.	Max Eff.
t3	3	0.0486	0.2084	0.0018	0.0195
t4	4	0.1389	0.3409	0.0148	0.0457
t5	5	0.2094	0.4484	0.0336	0.0748
t6	6	0.2682	0.5393	0.0551	0.1045
t7	7	0.3177	0.6171	0.0773	0.1334
t8	8	0.3604	0.6851	0.0995	0.1613
t$\hat{\nu}$	4.65 (1.67)*	0.1565	0.3849	0.0188	0.0552
Normal	-	1.1426	1.8903	-	-
Model	ν	$\hat{\Psi}$			
		Min Bias	Max Bias	Min Eff.	Max Eff.
t3	3	4.6003	142.2621	0.0061	0.0022
t4	4	2.9822	8.7962	0.0000	0.0009
t5	5	9.1313	131.2022	0.0052	0.0086
t6	6	14.2782	233.3098	0.0164	0.0211
t7	7	18.6811	320.3971	0.0309	0.0362
t8	8	22.5087	395.9063	0.0472	0.0525
t$\hat{\nu}$	4.65 (1.67)*	4.6739	46.3114	0.0006	0.0023
Normal	-	98.2385	1822.9992	-	-

Table 5.8: Bias and efficiency of the MLE for the first two elements of μ and a diagonal and off-diagonal element of Ψ obtained from a simulated multivariate Student-t with $\nu = 3$.

		Bias		Efficiency	
Model	ν	$\hat{\mu}(1)$	$\hat{\mu}(2)$	$\hat{\mu}(1)$	$\hat{\mu}(2)$
t3	3	0.0881	0.0970	0.4037	0.4432
t4	4	0.0886	0.0971	0.4071	0.4453
t5	5	0.0893	0.0975	0.4131	0.4505
t6	6	0.0901	0.0981	0.4202	0.4571
t7	7	0.0908	0.0988	0.4276	0.4644
t8	8	0.0916	0.0996	0.4351	0.4719
t$\hat{\nu}$	3.23 (0.78)*	0.0881	0.0973	0.4043	0.4481
Normal	-	1.4146	1.9103	-	-
		Bias		Efficiency	
Model	ν	$\hat{\Psi}_{1,1}$	$\hat{\Psi}_{1,2}$	$\hat{\Psi}_{1,1}$	$\hat{\Psi}_{1,2}$
t3	3	0.1435	0.1051	0.0073	0.0091
t4	4	0.1819	0.1150	0.0087	0.0108
t5	5	0.2340	0.1240	0.0101	0.0125
t6	6	0.2867	0.1322	0.0115	0.0141
t7	7	0.3365	0.1397	0.0129	0.0157
t8	8	0.3826	0.1465	0.0143	0.0172
t$\hat{\nu}$	3.24 (0.78)*	0.1531	0.1068	0.0082	0.0095
Normal	-	0.6350	1.5936	-	-

* Average and (standard deviation) for the MLE of ν; ν is estimated simultaneously as μ and Ψ.

Table 5.9: Estimates from fitting a regression on Stackloss Data.

Model	Intercept	Airflow	Temperature	Acid	log-likelihood
t3	-39.10	0.90	0.70	-0.10	-31.80
t4	-40.10	0.90	0.70	-0.10	-32.10
t5	-40.50	0.80	0.80	-0.10	-32.30
t6	-40.70	0.80	0.90	-0.10	-32.50
t7	-40.70	0.80	0.90	-0.10	-32.60
t8	-40.70	0.80	1.00	-0.10	-32.70
Normal	-39.90	0.70	1.30	-0.20	-33.00
Normal minus outliers	-37.70	0.80	0.60	-0.10	-10.10

Table 5.10: Results from the fit of the multivariate Student-t model on Stackloss Data.

Model	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\mu}_4$	$\hat{\Psi}_{1,1}$	$\hat{\Psi}_{1,2}$	$\hat{\Psi}_{1,3}$	$\hat{\Psi}_{1,4}$	Log-L
t3	58.44	20.69	85.97	15.48	51.37	14.61	17.03	51.93	-236.79
t4	58.72	20.74	86.01	15.81	56.43	15.96	17.87	57.72	-235.91
t5	58.95	20.79	86.05	16.07	60.18	16.95	18.52	61.95	-235.37
t6	59.13	20.83	86.08	16.27	62.99	17.68	19.04	65.06	-235
t7	59.28	20.86	86.11	16.43	65.16	18.24	19.45	67.42	-234.74
t8	59.39	20.89	86.13	16.56	66.82	18.65	19.79	69.19	-234.54
t9	59.49	20.91	86.14	16.66	68.16	18.99	20.08	70.6	-234.39
t10	59.57	20.92	86.16	16.74	69.27	19.26	20.32	71.74	-234.26
MCD	56.15	20.23	85.38	13.15	28.31	8.79	20.1	24.64	N/A
Normal	60.43	21.1	86.29	17.52	84.06	22.66	24.57	85.76	-233.2
Normal minus outliers	53.71	19.14	82.71	10	56.57	3.71	26.9	21.67	-36.87

Table 5.11: Results from the fit of the multivariate Student-t model on Hawkins, Bradu and Kass data.

Model	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\Psi}_{1,1}$	$\hat{\Psi}_{1,2}$	$\hat{\Psi}_{1,3}$	Log-L
t3	1.81	2.4	2.53	3.24	5.15	7.83	-526.99
t4	2.07	2.97	3.41	5.36	9.82	14.85	-527.87
t5	2.25	3.39	4.05	6.82	13.03	19.66	-528.08
t6	2.39	3.69	4.5	7.8	15.2	22.91	-528.21
t7	2.48	3.9	4.83	8.5	16.75	25.23	-528.37
t8	2.56	4.07	5.08	9.03	17.92	26.97	-528.57
t9	2.62	4.2	5.28	9.44	18.83	28.31	-528.8
t10	2.66	4.31	5.44	9.76	19.56	29.39	-529.05
MCD	1.54	1.8	1.66	1.11	0	0.16	N/A
Normal	3.19	5.6	7.23	13.38	28.5	41.32	-541.38
Normal minus outliers	1.52	1.78	1.69	1.12	0.02	0.12	-98.97

Figure 5.1: Determinant for the sample variance-covariance matrix of a simulated multivariate Student-t data. The dimensions of the sample varies from 20 to 200 observations on 4 variables.

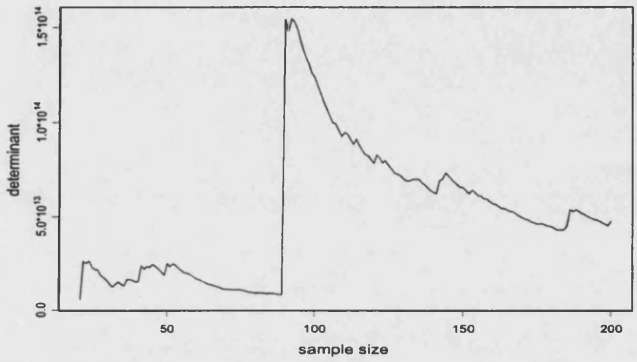


Figure 5.2: Maximum, 99, 98.5, 98 and 97.5% simulation envelopes for Mahalanobis distances on multivariate Student-t data.

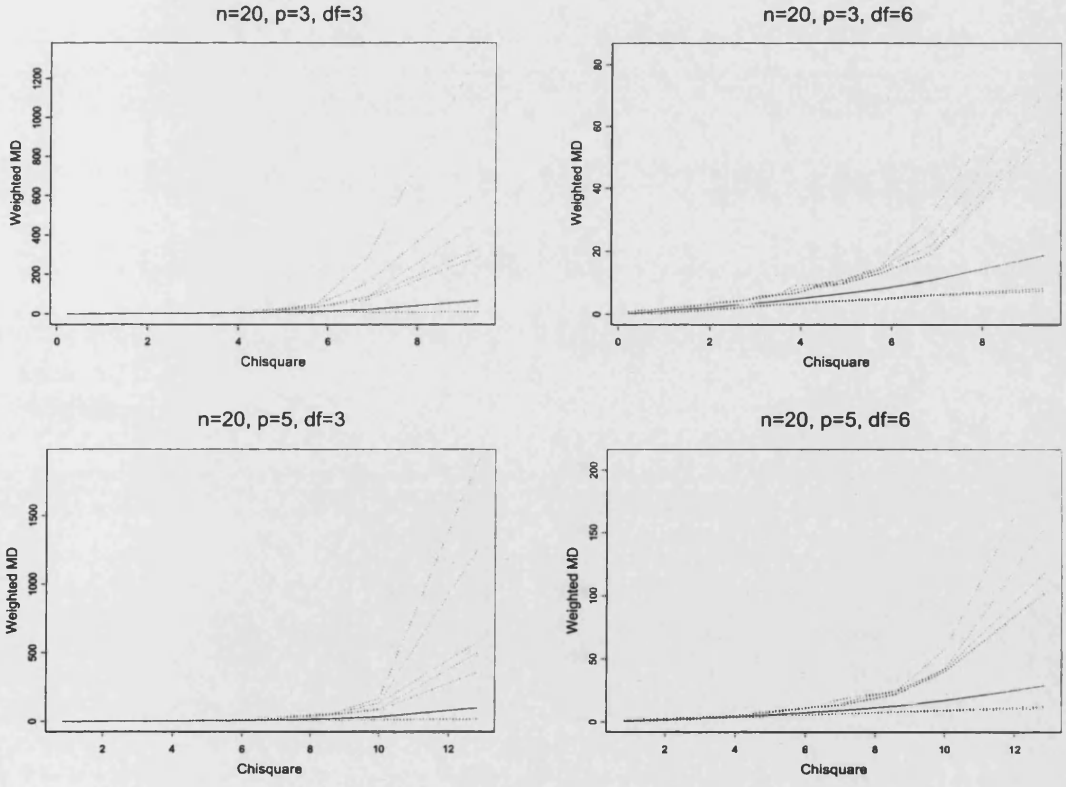


Figure 5.3: Plot of the sorted weights ($\hat{\tau}$) from the t3 model.

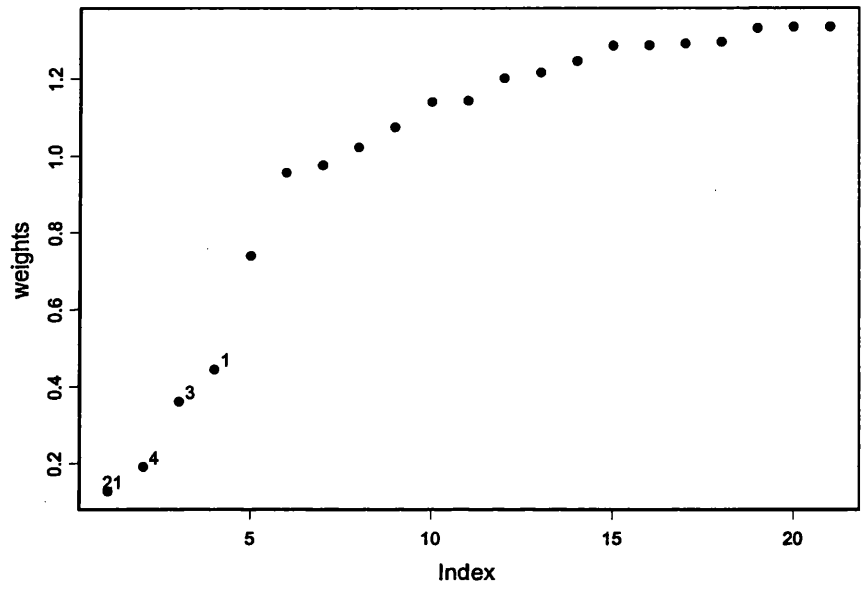


Table 5.12: Example 1: Stackloss data. Weights of the t3 model. The odd columns are the observations; the even columns the weight values.

obs	weight	obs	weight	obs	weight
1	0.445	8	1.333	15	0.977
2	1.246	9	1.076	16	1.295
3	0.361	10	1.333	17	1.286
4	0.191	11	1.202	18	1.330
5	1.141	12	1.290	19	1.285
6	0.958	13	0.740	20	1.024
7	1.217	14	1.144	21	0.127

Figure 5.4: Profile Likelihood for the multivariate Student-t degrees of freedom.

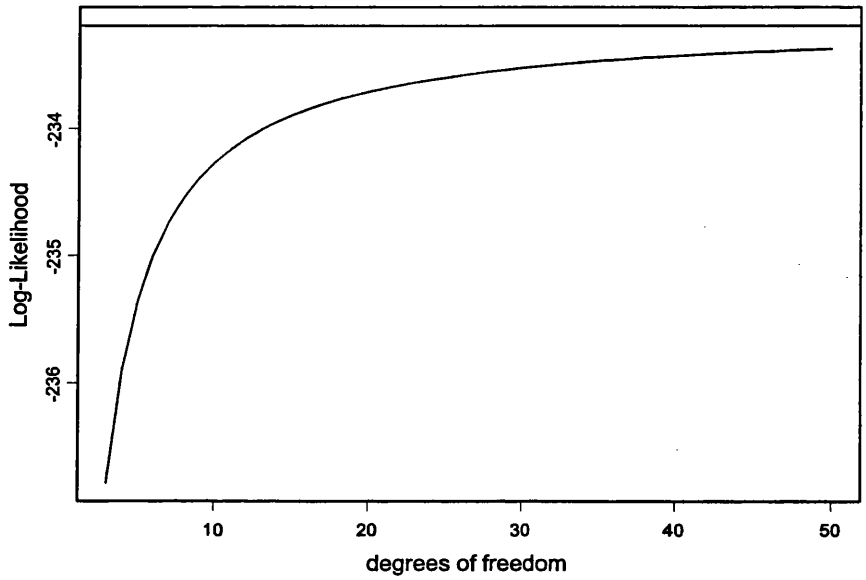


Figure 5.5: Plot of the sorted weights for the multivariate Student-t with $\nu = 3$ fitted on Stackloss Data.

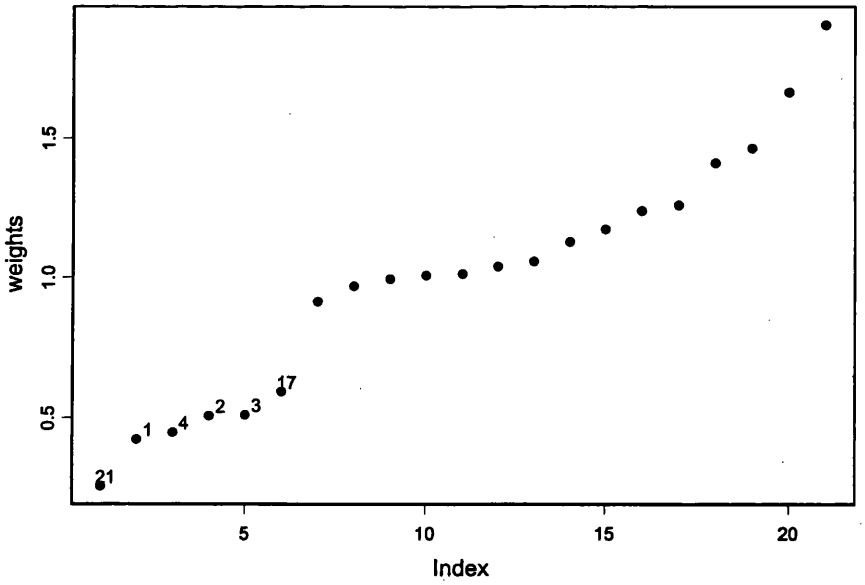
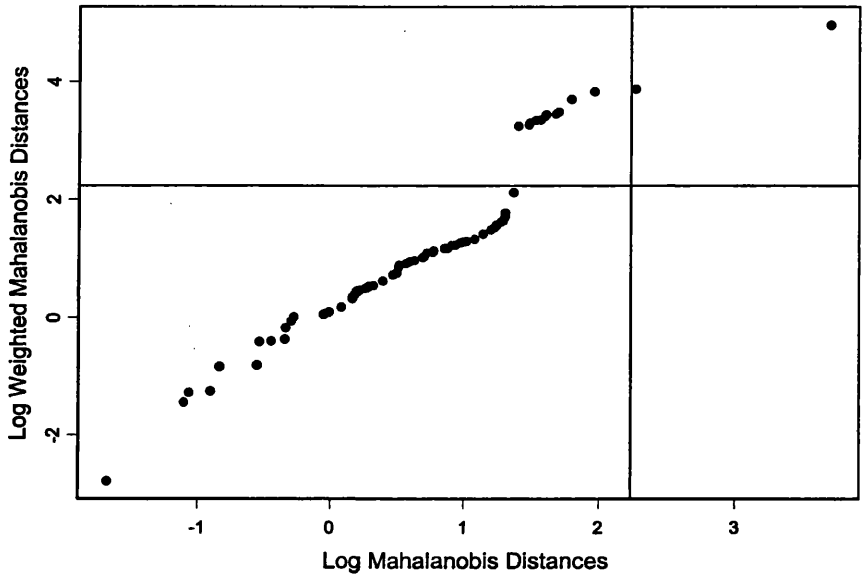
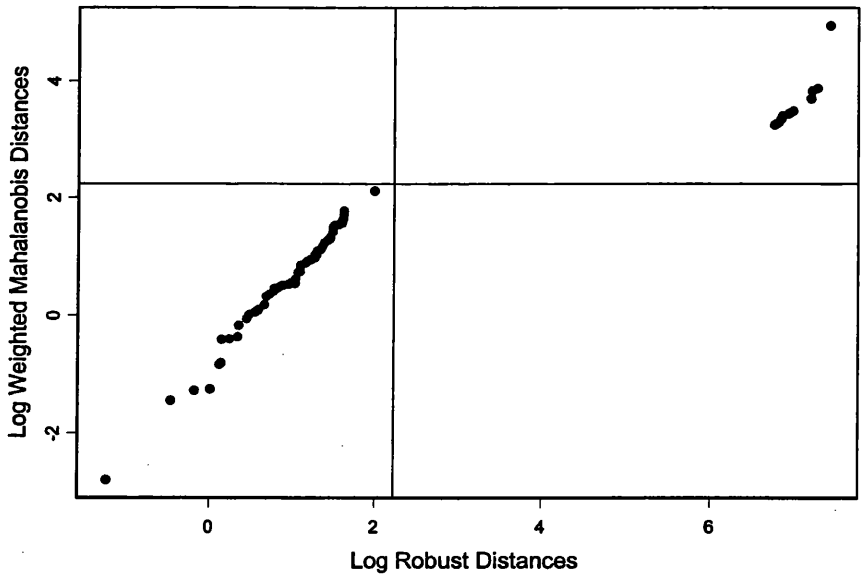


Figure 5.6: D-D Plot of Mahalanobis-type Distances fitted on Bradu Hawkins and Kass data. The solid lines is the 97.5% Chisquare quantile.

(a)



(b)



Chapter 6

Robust Modelling for Financial Portfolio Selection

6.1 Introduction

Despite of criticism of the well established Mean-Variance models for the selection of financial portfolios, they are still widely used in practice. These models require the estimation of the covariance matrix and the mean vector of asset returns.

Generally, the literature tends to consider the sample estimates that are maximum likelihood when the returns have a multivariate Normal distribution. It is well known that equity returns have heavier tails than the Normal distribution. Kon (1984), Roll (1988), Fama (1965) or, more recently, Linden (2001), Vilasuso and Katz (2000) are some examples of works reporting evidence on the distribution of stock returns. It is also known that the sample estimates are very sensitive to the presence of outliers (Huber 1981).

The first section of the Chapter gives an overview of the classical portfolio selection problem and motivations for a robust solution. The second section explains the robust models. Finally, we give some performances on both simulated and real data.

6.1.1 Basic Notions on Financial Portfolios

There are many different ways of investing in financial markets. For simplicity, we will consider the direct form of investments, for which the investor purchases

a security directly from a government or a private institution. The direct investments include both money market instruments, that are high liquidity securities and capital market securities, which have maturity of more than one year (corporate, government bonds) or no maturity at all, like the stocks. These are the instruments considered in our portfolio analysis. Other types of capital market securities, which we exclude from our portfolios, are the derivatives. These are contracts (options or futures) whose value depends on an underlying security.

Any security is described by a return and a risk over a time period. The return, generally expressed as a percentage, is meant as the change in the value of a security plus any income received during that time divided by the initial value of the security. The risk is the uncertainty in the outcome of an operation and is measured by the variability, usually the standard deviation, of the returns. Some assets are more risky than others depending generally on the length of the maturity and on the issuer. Assets with a longer maturity are more risky than short term debt instruments. Stocks have more risk than bonds because the issuer has a higher risk. As a consequence of this, the risk-less assets are those able to guarantee a constant return over time.

The portfolio selection problem is the decision of the quota of the available budget to invest within a set of assets. There are two parameters to consider in this framework: the time of duration of the investment and the number of assets. Both of them are assumed to be finite. Time is indicated by t ($t = 1, 2, ..., T$) and the assets by i ($i = 1, 2, ..., p$). Let W be the wealth of the investor; at time t the investor can spend:

$$\sum_{i}^{p} v_i p_i = W,$$

where v_i is the quantity of asset i and p_i is its price. It is often convenient to normalize by the total amount of wealth, that is:

$$\mathbf{1}^T \mathbf{w} = 1,$$

where $\mathbf{w} = (w_1, w_2, ..., w_p)$ and $w_i = v_i p_i / W$ is the portfolio weight assigned to the asset i. The normalization constraint is also called budget constraint. It is possible, although unlikely in practice, that the one or more weights are negative. In that case the assets are sold "short". The general idea of a short-sale is that the investor can sell assets that he does not own by borrowing them from another investor with the promise to make payments compensating those due on the assets (for example the dividends in the case of a stock) and to re-purchase and replace the assets when the transaction is closed. Short positions are often motivated by speculative reasons: if a short seller expects the price of the security to go down and he wants to profit from the decrease in the price. Sometimes short sales are accompanied by long positions in order to diversify the market risk exposure. Models that exclude the possibility of speculative operations impose the restriction of no short sales by setting $w_i \geq 0$, $i = 1, 2, ..., p$.

6.1.2 Notation

Mean-variance theory has historically been one of the earliest formalizations of the portfolio selection problem (Markowitz 1952).

The general idea is to determine the quota to be invested in each asset according to a "mean-variance efficiency" criterion. In other words, the aim is to find the combination of assets giving the optimal equilibrium between the risk and the expected return of the portfolio.

The starting point is that investors make their choices under uncertainty, that is: every market operation is characterized by an expected return and a risk.

Markowitz defines the return and the risk of a portfolio in the following way: the portfolio expected return is the weighted average of the returns on the individual assets, where the weights are the quota invested in each asset.

$$\mu_P = E_t(\sum_{i=1}^{p} w_i y_{it})$$

where $i = 1, 2, ..., p$ and $t = 1, 2, ..., T$. y_{it} is the return on asset i at time t; w_i is the weight assigned to the asset i.

The portfolio risk depends on the risk of the individual assets and on their joint risk. The risk of the individual assets is defined as their variance and the joint risk of two assets i and k as their covariance:

$$\begin{aligned} \sigma_i^2 &= E_t(y_{it} - \mu_i)^2, \\ \sigma_{ik} &= E_t[(y_{it} - \mu_i)(y_{kt} - \mu_k)]; \end{aligned}$$

where $i = 1, 2, ..., p$, $k = 1, 2, ..., p$, with $k \neq i$, and $\mu_i = E_t(y_{it})$. According to Markowitz, when investing in a "basket" of assets, the risk of having bad outcomes is reduced if the assets are "diversified". This principle, that may appear common sense, is fundamental in mean-variance theory and motivates the definition of pair-wise covariances as a measure of the joint risk. High covariances between the assets indicate a less diversified portfolio, that translates into a high risk exposure. Therefore, the portfolio risk is:

$$\sigma_P^2 = \sum_{i=1}^{p} w_i^2 \sigma_i^2 + \sum_{i=1}^{p} \sum_{k \neq i}^{p} w_i w_k \sigma_{ik}.$$

The above notation holds since Markowitz theory assumes normality of the asset returns. Since the portfolio is a weighted linear combination of the returns and, since the Normal distribution is completely characterized by its mean and variance, all portfolios coming from those assets are defined by their means and variances.

6.2 Standard Mean-Variance Portfolio Problem

Risk and return can be rewritten in matrix notation:

$$\sigma_P^2 = \mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w}, \tag{6.1}$$

$$\mu_P = \mathbf{w}^T \boldsymbol{\mu}, \tag{6.2}$$

where $\mathbf{w}$ is the vector of asset weights, that is the quota of the budget invested in each asset. Furthermore, $\mathbf{y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\mathbf{y} = (y_1, y_2, ..., y_p)$.

As a basic example, we consider the simplest model. The assumptions are:

1. *Absence of riskless assets*

2. *Possibility of short-selling*

3. *The expected return level is fixed*

4. *There are no transaction costs*

Although it is realized that most of these assumptions are not realistic, this model is only considered as a first approximation. However, extensions to more complicated situations are still possible (the absence of risk-less assets and fixed level of expected returns are removed in Section 6.4).
The selected portfolio is the one that maximizes the utility function of the investor. Since, according to the mean-variance theory, there is a trade-off between risk and return, the optimal portfolio maximizes the asset returns for a fixed level of risk or, equivalently, minimizes the risk for given expected return objective.

The formulation due to Merton (1972) is:

$$\min_{\mathbf{w}} \mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w} \tag{6.3}$$

subject to:

$$\mathbf{w}^T \mathbf{1} = 1, \tag{6.4}$$

$$\mathbf{w}^T \boldsymbol{\mu} = q, \tag{6.5}$$

where $\mathbf{1}$ is the unit vector and q is the level of expected return required to the portfolio. The solution of the problem for this case is simple, since it is an optimization with linear constraints. We proceed by forming the Lagrangian:

$$L \equiv \mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w} + \lambda_1 (1 - \mathbf{w}^T \mathbf{1}) + \lambda_2 (q - \mathbf{w}^T \boldsymbol{\mu})$$

from which the first-order conditions are:

$$\frac{\partial L}{\partial w_i} = \frac{\partial L}{\partial \lambda_j} = 0 \tag{6.6}$$

$i = 1, 2, ..., p; \quad j = 1, 2.$

The solution is:

$$\mathbf{w} = \frac{C - qB}{AC - B^2}\Sigma^{-1}\mathbf{1} + \frac{qA - B}{AC - B^2}\Sigma^{-1}\boldsymbol{\mu}, \tag{6.7}$$

where:

$$A = \mathbf{1}^T\Sigma^{-1}\mathbf{1}, \quad B = \mathbf{1}^T\Sigma^{-1}\boldsymbol{\mu} \text{ and } C = \boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}. \tag{6.8}$$

The above system allows solutions provided that:

1. Σ is non-singular and, therefore, invertible.

2. $\boldsymbol{\mu} \neq k\mathbf{1}$ i.e. the assets do not have the same mean, therefore, $AC - B^2 \neq 0$.

Substituting the above results in the expression for the portfolio variance (6.1), we obtain:

$$\sigma_P^2 = \frac{Aq^2 - 2Bq + C}{AC - B^2} \tag{6.9}$$

The plot of σ_P^2 for each value of q is the parabola of all *minimum-variance* portfolios, the *Mean-Variance Frontier.* Usually the frontier is plotted in a mean-standard deviation, rather than in a mean-variance space. The result is an hyperbola instead of a parabola, Figure 6.1, panel (a). The *efficient* solutions are the portfolios on the positive slope frontier, for which the return increases with a higher risk exposure. In addition to the efficient solutions, it desirable to choose portfolios that are closer to the minimum rather than to the top end of the frontier, where for moderate increase in the risk the returns grow at a lower rate.

Empirical and theoretical results have shown that the efficient frontier has a convex shape and it is absolutely differentiable (Ingersoll 1987). In other words,

it allows a global optimum, which is the portfolio with *global minimum risk*. The minimum weights are function of the variance-covariance matrix only:

$$w_{\min} = \frac{\Sigma^{-1}\mathbf{1}}{\mathbf{1}^T\Sigma^{-1}\mathbf{1}}. \tag{6.10}$$

6.2.1 Motivations for a Robust Model

The Markowitz optimizer takes as inputs the expected returns and their variance-covariance matrix and gives an output vector of weights. This method poses the problem of estimating the first two moments of the return distribution.

The sample estimates, commonly used, are maximum-likelihood when the underlying distribution is multivariate Normal. The financial literature has produced large evidences that the equity returns are not Normal, but rather tend to a heavy tailed and skewed distribution. Many studies support the assumption of a mixture of Normals: Fama (1965), Kon (1984), Roll (1988) and Richardson and Smith (1994).

Secondly, sample estimates are very sensitive to outliers, Huber (1981). The consequence is that an "extremely" good or bad outcome for even a single time-observation would bias the estimates of the weights.

Finally, there is a "practical" motivation. Every time the portfolio is re-balanced, that is when new weights are assigned to the assets, there are transaction costs involved. These costs are reduced by controlling the variability of the portfolio weights. These are function of (μ, Σ). In other words, transaction costs are decreased by obtaining relatively stable estimates for the return location and scatter.

The example in Figure 6.2 shows the sensitivity of the mean-variance frontier to deviations from the Normal distribution. In panel (a) the data are generated from a shifted-mean model (Section 4.8.1): it is a sample of 200 observations on 4 variables from a Normal distribution where there are 2 outliers. The data are generated with the parameters being equal to the sample mean and variance-covariance matrix of the data illustrated in Chapter 2, reduced to the first 4 bond markets. The plots compare the mean-variance frontiers. The outliers, extremely high returns, are shifted upwards; this causes the efficient frontier to shift upwards as well when the contamination is included in the data. Figure 6.2 (b) shows the opposite situation

when the outliers are extremely bad outcomes. In both cases, the contaminated frontiers are more to the right than the true and outlier-free ones, which means that for fixed expected returns (that vary within moderate values) the risk exposure is higher than for the outlier-free frontier. Furthermore, the latter is closer to the true one, which gives motivation for robust modelling. For higher return objectives, the contaminated frontier lies above the other two frontiers and the outlier-free frontier lies further from the true portfolios than the contaminated frontier. However, portfolios with very high returns are not interesting since they are not realistically achievable in terms of risk tolerance.

Panel (c) shows the sensitivity of the optimal portfolio when the underlying distribution has heavy tails. The data are generated from a multivariate Student-t distribution with 3 degrees of freedom (Chapter 5); mean and inner-product matrix are the same as the two moments of the Normal distribution in panel (a) and (b). The frontier fitted with maximum-likelihood estimates on a Normal model is shifted far to the right from the true frontier and to the fit on the multivariate Student-t (assuming the degrees of freedom are known). This last is closer to the true portfolios towards the global minimum solution. The results are in line with those found in the examples of panel (a) and (b).

6.2.2 The Robust Models

The Chapter proposes two alternative ways of selecting a robust portfolio. The first one assumes that the portfolio returns have a multivariate Student-t distribution. Therefore:

$$\mathbf{y}_t \sim t_p(\boldsymbol{\mu}, \boldsymbol{\Psi}, \nu) = N_p(\boldsymbol{\mu}, \boldsymbol{\Psi}/\tau_t)$$

defined as in (5.3), where τ_t are the "latent" weights for the time-observations. (6.11) implies:

$$P = \mathbf{w}^T\mathbf{y} \sim N_p(\mathbf{w}^T\boldsymbol{\mu}, \tau_y^{-1}\mathbf{w}^T\boldsymbol{\Psi}\mathbf{w}),$$

that is:

$$P \sim t_p(\mathbf{w}^T\boldsymbol{\mu}, \mathbf{w}^T\Psi\mathbf{w}, \nu). \tag{6.11}$$

Estimates of $\boldsymbol{\mu}$ and Ψ replace those for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in the Normal model. MLE for the multivariate Student-t are obtained as it was shown in Section 5.3.

The second model assumes that the "majority" of the data is normally distributed and estimates $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ robustly via the MCD method (Section 3.3).

6.3 The Influence of Outliers on Markowitz Portfolio

The influence function described Chapter 3 measures the sensitiveness of an estimator to an infinitely small contamination. Martin (1999) computes the IF for the mean-variance frontier based on classical Normal maximum-likelihood estimates; following his result we recalculate the IF for the multivariate Student-t portfolio.

Let us consider a mixed distribution:

$$F_\epsilon = (1-\epsilon)F + \epsilon\Delta_y. \tag{6.12}$$

The majority of the data $1-\epsilon$ has distribution $F = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and a smaller contaminated fraction of observations ϵ, density mass Δ_y, (3.5). Under normality, the portfolio has the form:

$$\mathbf{w}^T\mathbf{y} \sim N(\mu_P, \sigma_P^2),$$

where μ_P and σ_P^2 are given by (6.2) and (6.1). The minimum variance portfolio is given by (6.10), that substituted in (6.1) and (6.2), leads to the following expressions:

$$\begin{aligned}
\sigma_P^2(F_\epsilon) &= (\mathbf{1}^T\boldsymbol{\Sigma}(F_\epsilon)^{-1}\mathbf{1})^{-1} \\
\mu_P(F_\epsilon) &= (\mathbf{1}^T\boldsymbol{\Sigma}(F_\epsilon)^{-1}\mathbf{1})^{-1}\mathbf{1}^T\boldsymbol{\Sigma}(F_\epsilon)^{-1}\boldsymbol{\mu}(F_\epsilon),
\end{aligned}$$

where the parameter estimates are written as functionals of the contaminated model:

$$\begin{aligned}\mu(F_\epsilon) &= \int \dots \int \mathbf{y} dF_\epsilon \\ \Sigma(F_\epsilon) &= \int \dots \int (\mathbf{y} - \mu(F_\epsilon))(\mathbf{y} - \mu(F_\epsilon))^T dF_\epsilon.\end{aligned}$$

The influence function defined in (3.6) is the differential of the contaminated model at $\epsilon = 0$. Therefore:

$$IF_{\sigma_P^2} = \left. \frac{d\sigma_P^2(F_\epsilon)}{d\epsilon} \right|_{\epsilon=0}.$$

Analogously for μ_P.The result is:

$$IF_{\sigma_P^2} = (\mathbf{1}^T \Sigma(F_\epsilon)^{-1} \mathbf{1})^{-2} \mathbf{1}^T \Sigma(F_\epsilon)^{-1} \left. \frac{d\Sigma(F_\epsilon)}{d\epsilon} \right|_{\epsilon=0} \Sigma(F_\epsilon)^{-1} \mathbf{1} \tag{6.13}$$

$$\begin{aligned}IF_{\mu_P} = {} & IF_{\sigma_P^2} \mathbf{1}^T \Sigma(F_\epsilon)^{-1} \mu(F_\epsilon) + \sigma_P^2(F_\epsilon) \times \\ & \left[\mathbf{1}^T \Sigma(F_\epsilon)^{-1} \left. \frac{d\mu(F_\epsilon)}{d\epsilon} \right|_{\epsilon=0} - \mathbf{1}^T \Sigma(F_\epsilon)^{-1} \left. \frac{d\Sigma(F_\epsilon)}{d\epsilon} \right|_{\epsilon=0} \Sigma(F_\epsilon)^{-1} \mu(F_\epsilon) \right],\end{aligned} \tag{6.14}$$

where $d\Sigma^{-1}(F_\epsilon)/d\epsilon = -\Sigma(F_\epsilon)^{-1} d\Sigma(F_\epsilon)/d\epsilon \Sigma(F_\epsilon)^{-1}$. From (6.13) and (6.14), it is concluded that the asymptotic bias of the portfolio minimum variance depends on the influence function of $\mu(F_\epsilon)$ and $\Sigma(F_\epsilon)$. Therefore, the solution will be more or less robust depending on the robustness of the estimator for location and scatter.

In our model, it is assumed normality for the majority of the observations. The sensitivities for μ and Σ at the mixed model (6.12) are just a generalization of the sample mean and variance sensitivities in the univariate case:

$$\left. \frac{d\Sigma(F_\epsilon)}{d\epsilon} \right|_{\epsilon=0} = -\Sigma(F) + (\mathbf{y} - \mu(F))(\mathbf{y} - \mu(F))^T \tag{6.15}$$

$$\left. \frac{d\mu(F_\epsilon)}{d\epsilon} \right|_{\epsilon=0} = \mathbf{y} - \mu(F), \tag{6.16}$$

where F is the Normal distribution and:

$$\begin{aligned}\boldsymbol{\mu}(F) &= \int ... \int \mathbf{y} dF \\ \boldsymbol{\Sigma}(F) &= \int ... \int (\mathbf{y} - \boldsymbol{\mu}(F))(\mathbf{y} - \boldsymbol{\mu}(F))^T dF..\end{aligned}$$

Substituting (6.15) and (6.16) in (6.13) and (6.14), we finally find the expression for the IF of the minimum variance portfolio. We conclude that the asymptotic bias is unbounded for infinitely large values of $\mathbf{y}$.

We recalculate the influence function for the parameter assuming that F, the distribution for the majority of the data is a multivariate Student-t, $t_p(\boldsymbol{\mu}, \boldsymbol{\Psi}, \nu)$. The estimates for the variance-covariance matrix and mean are replaced by those for $\boldsymbol{\mu}$ and $\boldsymbol{\Psi}$ defined as:

$$\begin{aligned}\frac{d\boldsymbol{\mu}(F_\epsilon)}{d\epsilon} &= \tau_y \mathbf{y} - \boldsymbol{\mu}(F) \\ \frac{\boldsymbol{\Psi}(F_\epsilon)}{d\epsilon} &= -\boldsymbol{\Psi}(F) + \tau_y (\mathbf{y} - \boldsymbol{\mu}(F))(\mathbf{y} - \boldsymbol{\mu}(F))^T,\end{aligned}$$

where:

$$\begin{aligned}\boldsymbol{\mu}(F) &= \int ... \int \tau_y \mathbf{y} dF \\ \boldsymbol{\Psi}(F) &= \int ... \int \tau_y (\mathbf{y} - \boldsymbol{\mu}(F))(\mathbf{y} - \boldsymbol{\mu}(F))^T dF.\end{aligned}$$

The degrees of freedom are assumed known. The portfolio distribution changes as described in (6.11).

As expected, the sensitivity of the robust portfolio depends on the weights τ_y: when the latter approach 1, the IF of the portfolio approaches the one of the classical Markowitz model. On the contrary, the smaller are the weights, the more bounded becomes the influence function. The robustifying parameter is the degrees of freedom, as shown in Figure 6.4 and Figure 6.3. These are plots of the influence functions for the moments of the minimum variance portfolio of a bivariate Normal and Student-t distributions with mean and variance respectively:

$$\mu = (.1, .1), \quad \Sigma = \begin{pmatrix} .0025 & .00125 \\ .00125 & .0025 \end{pmatrix}.$$

The IF of the maximum-likelihood portfolio on the Normal distribution is unbounded. For the multivariate Student-t model the function is smoother on the edges, showing good robustness properties. When the degrees of freedom increase from 3 to 10 the shape approaches that for the Normal model. Since the weights τ_y are unknown, they have been estimated through EM, with an accuracy of (0.9533, 1.0467), computed as the following:

$$\hat{\tau}_y \pm 2 \times \frac{\hat{\sigma}_{\tau_y}}{\sqrt{n}}.$$

6.4 The Weights

Figure 6.5 shows the boxplots of weights obtained from 500 optimizations on multivariate Normal samples of 200 observations. The data are generated from the first two sample moments of the return data described in Chapter 2, of which we consider only the first four assets. Panel (a) is the optimal solution for a 1% monthly expected return constraint. Panel (b) is the global minimum variance solution. The solid dots are the true parameters. Panel (a) shows a considerably high variability of the estimates, which causes a slow convergence to the true values. The global minimum solution does not evidence the same problem. Although Merton's formulation, (6.3)-(6.5), looks very appealing thanks to its simplicity, it imposes little constraint, therefore allowing for the portfolio composition to vary appreciably. The model is therefore modified by setting additional constraints.

It is assumed the possibility of holding a cash position at a fixed interest rate R. Since the cash could also finance the positions in financial assets, the budget constraint is removed. The new model is:

$$\min_{\mathbf{w}} \sigma_P^2 = \mathbf{w}^T \Sigma \mathbf{w} \tag{6.17}$$

with the constraint:

$$(\boldsymbol{\mu} - R\mathbf{1})^T \mathbf{w} = \mu_P - R, \tag{6.18}$$

which leads to the optimal solution:

$$\mathbf{w} = \gamma \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - R\mathbf{1}). \tag{6.19}$$

Substituting the latter in (6.18) and (6.17) gives:

$$\mu_P - R = \gamma(\boldsymbol{\mu} - R\mathbf{1})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - R\mathbf{1}) \tag{6.20}$$

$$\sigma_P^2 = \gamma^2(\boldsymbol{\mu} - R\mathbf{1})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - R\mathbf{1}). \tag{6.21}$$

From (6.20), letting $Sh = (\boldsymbol{\mu} - R\mathbf{1})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - R\mathbf{1})$, it is found that $\gamma = (\mu_p - R)/Sh$ which, replaced in (6.21), gives the solution in the mean-standard deviation space:

$$\sigma_P = \frac{|\mu_P - R|}{\sqrt{Sh}}.$$

This is geometrically represented as two rays with common origin R and slopes $\pm\sqrt{Sh}$ (Figure 6.1). The tangency point with the mean-variance frontier of the first model (6.9) represents the investment made of risky assets only. This solution is the portfolio that, for a fixed R, maximizes the Sharpe-Ratio Sh, that is the standardized excess return over the risk-free investment. The ratio was introduced for the first time by Sharpe (1975) as a measure of fund performance adjusted for the risk. Replacing (6.8) in the expression for the Sharpe ratio leads $Sh = C - 2RB + R^2A$.

Since the tangency solution contains only risky assets, it must satisfy the budget constraint $\mathbf{w}^T\mathbf{1} = 1$ which, replaced in (6.19), gives $\gamma = (B - AR)^{-1}$ and, therefore, the risky portfolio weights:

$$\mathbf{w}_{tan} = \frac{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - R\mathbf{1})}{B - AR}.$$

The interest is obviously in the efficient set of solutions, that is the tangents to the positive-slope ray. Huang and Litzenberger (1988) show that positive tangents are only for risk-free rates R less than B/A, the global minimum variance portfolio return.

Figure 6.6 and Figure 6.7 compare the distribution of the weights from the classical model with MLE on a Normal distribution and the robust optimizations (MCD and M-t). The simulated samples have the same dimension as in the previous example: 200×4 replicated 500 times. The solutions considered are both the global minimum variance and the tangency portfolio with a fixed annual $R = .3\%$.

In absence of outliers (panels (a) and (c)), the distribution of the ML and robust weights do not differ significantly. After introducing a contamination (mean-shift outliers), the robust portfolios out-perform the classical optimizer since they are consistent to the true parameters. Furthermore, the robust MCD performs better than the M-t for higher fractions of outliers. However, this last model improves when ϵ is small. The plots also show that the robust weights have longer tailed distribution than the ML ones. Within the robust portfolios, the multivariate-t (M-t) weights appear to be longer tailed than the MCD ones. After introducing contamination (mean-shift outliers),

6.5 Performances: Turnover, Risk and Return

The performances are evaluated by three variables: the turnover, the portfolio risk and return. The turnover is the average absolute difference of the weights in two subsequent periods. Risk and return are evaluated as in (6.1) and (6.2), where Σ and μ are replaced by the sample estimates.

6.5.1 Performances for the Normal and Outlier-shift models

The average performances on a sample of 200 independent portfolios are displayed in Table 6.1, Table 6.2 and Table 6.5.

In Table 6.1, the data are generated from a Normal distribution whose centre and location are the sample estimates of a real set of assets (Chapter 2). The dimension

of the samples is 120×10. The performances of the "classical" tangency portfolio are compared with the robust MCD and multivariate Student-t ones. The turnover increases for all models with the number of assets increasing. When the portfolio includes only few assets, the model performances do not differ much one from the other. For higher dimensions, the robust (MCD) portfolios have a higher turnover and a higher risk than the other models. This is motivated by the fact that the efficiency of the robust estimates becomes worse when p grows large (Croux and Haesbroeck 1999).

Table 6.2 considers a mixed distribution. The simulated data include 10% of mean-shift-outliers. The MCD portfolio out-performs the other two models in both turnover and risk. The performances for the global minimum portfolio are closer to one another in the case of smaller dimensions, 2 and 4 assets, while they differ more sensibly when 6 or 10 assets are included in the set of choices. The Student-t estimates are less variable than those on the Normal distribution. Also in this case, the turnover improves with higher dimensions because the decrease in the efficiency.

Table 6.5 reports the sensitivity of the MCD performances to the size of the fitted "good" set of observations. According to the results, we are allowed for a 10 or 20% error in "guessing" the number of outliers in the data without affecting the estimates sensibly. The performances are trivially optimal for 90% of fitted observation, since we included 10% of outliers in the simulated set.

Finally, Table 6.6 shows the convergence of the performances of the robust Student-t model to the"classical" ones for increasing degrees of freedom.

6.5.2 Performances for the Multivariate-t Model

The same experiment has been repeated when the underlying distribution is a multivariate Student-t with 3 degrees of freedom. Table 6.3 displays the results. The Student-t model has better performances than the MCD and the Normal model, although there is still a "large" difference from the true portfolio parameters. The bias increases with the growing of the dimensions.

Table 5.1 to Table 5.8 are a simulation study on the efficiency and bias of the multivariate Student-t. The bias of the inner product matrix $\hat{\Psi}$ is reported to be

high compared to $\hat{\mu}$. Figure 6.8 and Figure 6.9 show the squared bias and the standard deviation of the estimates when the sample size increase from 80 to 800 and for 3 dimensions. The draws for different dimensions are not independent: some extra-observations are added to the initial set to increase the sample size. This method has been chosen in order to produce plots with a reduced sampling variability.

The bias for the inner-product matrix estimate is higher compared to the mean. A considerably large sample, at least 400 observations, must be chosen to obtain a good convergence to the true parameter values. The variability of the estimates (Figure 6.9) is also considerably higher for $\hat{\Psi}$ than for $\hat{\mu}$, although it sensibly improves when the sample size grows.

6.6 Out-of-Sample performances

This section discusses a more "realistic" simulation framework. A *buy-and-hold* strategy is followed, consisting of selecting a portfolio at time t on the basis of the previous $t - 120$ observations. The portfolio is then held constant until time $t + 12$ (*out-of-sample period*), when it is liquidated and the process starts all over again.

The result of the experiment is a sample of weights vectors whose performances are tested. Our interest is particularly focused on the "out-of-sample" period, that is the 12 months when the portfolio does not change. The attention is on the portfolio standard deviation and the weights distribution.

Table 6.4 reports the result on a simulated set and Table 6.7 shows a real data example.

The size of the simulated sample is around 2400 on 4 variables, which leads to a sample of 194 portfolios. The performances confirm what we have found previously when simulating a distribution of independent portfolios. The models behave as expected: the classical Markowitz portfolio performs better on Normal data. The robust MCD model has better performances on a distribution contaminated with mean-shift outliers and the M-t model out-performs the others on a simulated multivariate-t distribution.

The real data are monthly return indexes on 5 markets from April 1970 to February 2000. From the scatter matrix Figure 6.10, the data seem to have few outliers and appear generally more scattered than the Normal distribution. Therefore, it is expected that a long tailed distribution model would perform well. On the contrary, the MCD would give less good performances being more robust but less efficient than the ML estimators.

The turnover is lowest for the multivariate Student-t portfolio, although the difference is not small if compared to the classical Markowitz model. On the contrary, MCD definitely performs worse than the other two optimizers, confirming what expected.

The risk is lower in average and variability for the multivariate Student-t case, which also has the highest expected return with the lowest variability in both the tangency and global minimum variance portfolio.

Figure 6.11 is the plot of Mahalanobis distances on the fitted Student-t parameters. The outliers detected have a economic meaning: 1975 and 1987 were notoriously two critical dates for the international financial markets.

6.7 Conclusion

Financial data are notoriously not normally distributed. Robust alternatives to the classical maximum likelihood estimates perform better in the modelling of financial portfolios. According to our simulations, the robust optimization produce portfolios with a lower risk and turnover. The performances are still in favour of the robust approach when the number of assets is increased.

The choice of the robust model depends on the assumptions and on the type of data. The MCD works better on large data with many outliers, where the high robustness properties can compensate the lack of efficiency. When the data are longer tailed than the Normal distribution and have few outliers, the multivariate-t model out-performs the robust MCD one. Nevertheless, MLE for the inner-product matrix Ψ has a large sample variability and not very good consistency properties. When selecting more than 4 assets the sample size has to be above 600 observations

in order to stabilize the sample variance.

However, there is a main drawback: the Markowitz model assumes that the observations are independent and identically distributed. Although financial returns show absence of autocorrelation, since their distribution is not Gaussian, this is not enough to ensure independence. Further possible extension of the work could be directed towards a robust model for the volatility.

Table 6.1: Turnover, return and risk of the tangency and global minimum variance portfolios from a simulated Normal distribution without outliers. Risk free rate= .2

		tangency solution			minimum variance solution		
p	method	turnover	risk	return	turnover	risk	return
2	Markowitz	0.21	0.83	0.45	0.05	0.78	0.42
	MCD	0.31	0.8	0.46	0.08	0.7	0.43
	M-t	0.25	0.71	0.46	0.05	0.64	0.42
	true	-	0.79	0.42	-	0.79	0.42
4	Markowitz	0.78	0.98	0.58	0.18	0.76	0.44
	MCD	1.57	1.12	0.63	0.28	0.69	0.44
	M-t	0.85	0.87	0.59	0.19	0.66	0.44
	true	-	0.81	0.47	-	0.78	0.44
6	Markowitz	1.17	1.23	0.73	0.24	0.74	0.42
	MCD	2.78	1.56	0.76	0.36	0.67	0.42
	M-t	1.3	1.13	0.76	0.25	0.66	0.42
	true	-	0.83	0.49	-	0.83	0.49
10	Markowitz	1.98	1.35	0.88	0.42	0.72	0.42
	MCD	5.45	2.1	1.19	0.72	0.63	0.42
	M-t	2.1	1.27	0.92	0.45	0.66	0.42
	true	-	0.85	0.53	-	0.76	0.42

Table 6.2: Turnover, return and risk of the tangency and global minimum variance portfolios from a simulated Normal distribution with 10% outliers. Risk free rate= .03 (monthly)

		tangency solution			minimum variance solution		
p	method	turnover	risk	return	turnover	risk	return
2	Markowitz	0.4	3.16	1.47	0.2	3.1	1.42
	MCD	0.15	0.75	0.44	0.07	0.72	0.42
	M-t	0.22	1	0.51	0.05	0.96	0.48
	true	-	0.79	0.42	-	0.79	0.42
4	Markowitz	1.64	3.41	1.77	0.73	3.07	1.44
	MCD	0.51	0.79	0.53	0.26	0.71	0.44
	M-t	1.01	1.41	0.76	0.22	1.19	0.55
	true	-	0.81	0.47	-	0.78	0.44
6	Markowitz	2.21	3.73	2.1	0.95	3.04	1.4
	MCD	0.73	0.86	0.6	0.32	0.69	0.41
	M-t	1.77	2.02	1.13	0.35	1.45	0.6
	true	-	0.83	0.49	-	0.83	0.49
10	Markowitz	4.17	4.01	2.51	1.74	2.98	1.4
	MCD	1.51	0.94	0.77	0.66	0.65	0.41
	M-t	3.91	2.98	1.85	0.93	1.93	0.8
	true	-	0.85	0.53	-	0.76	0.42

Table 6.3: Turnover, return and risk of the tangency and global minimum variance portfolios from a simulated multivariate Student-t distribution with 3 degrees of freedom. Risk free rate= .003

		tangency solution			minimum variance solution		
p	method	turnover	risk	return	turnover	risk	return
2	Markowitz	0.95	4.78	0.5	0.13	3.04	0.45
	MCD	1.28	3.43	0.8	0.1	1.85	0.43
	M-t	1.46	3.8	0.48	0.06	1.83	0.43
	true	-	0.79	0.42	-	0.79	0.42
4	Markowitz	4.34	7.32	1.26	0.5	2.67	0.5
	MCD	1.71	2.4	0.86	0.42	1.65	0.48
	M-t	1.35	2.35	0.76	0.28	1.68	0.49
	true	-	0.81	0.47	-	0.78	0.44
6	Markowitz	7.95	10.2	0.69	0.63	2.37	0.37
	MCD	8.07	6.18	0.18	0.5	1.44	0.36
	M-t	6.93	5.78	2.1	0.35	1.53	0.36
	true	-	0.83	0.49	-	0.83	0.49
10	Markowitz	1.29	1.57	1.19	0.78	1.24	0.77
	MCD	1.22	0.93	1.17	0.79	0.74	0.77
	M-t	0.83	1.02	1.12	0.51	0.85	0.78
	true	-	0.85	0.53	-	0.76	0.42

Table 6.4: Out-of-sample performances on four assets portfolios. Risk free rate= .03

		tangency solution			minimum variance solution		
model	method	turnover	risk	return	turnover	risk	return
Normal	Markowitz	0.13	0.81	0.46	0.06	0.74	0.43
	MCD	0.27	0.76	0.46	0.13	0.65	0.43
	M-t	0.14	0.7	0.47	0.06	0.63	0.43
mixed	Markowitz	0.55	1.6	1.51	0.17	0.92	1.38
	MCD	0.24	0.76	1.42	0.11	0.68	1.38
	M-t	0.33	0.96	1.45	0.07	0.64	1.38
Student-t	Markowitz	3.03	7.64	0.1	0.11	2.58	0.36
	MCD	2.18	3.53	0.38	0.18	1.8	0.43
	M-t	0.8	2.8	0.44	0.08	1.76	0.4
	true	-	0.81	0.47	-	0.78	0.44

Table 6.5: Sensitivity of the robust MCD model performances to h, the size of the fitted set. The proportion of the "good" observations in the simulated data is 0.9. Risk free rate= .03 (monthly), p=4.

	tangency solution			minimum variance solution		
h/n	turnover	risk	return	turnover	risk	return
0.5	0.51	0.79	0.53	0.26	0.71	0.44
0.7	0.45	0.81	0.52	0.21	0.73	0.44
0.8	0.45	0.81	0.52	0.2	0.74	0.44
0.9	0.43	0.83	0.52	0.19	0.75	0.44
0.95	0.44	0.84	0.52	0.19	0.76	0.44
1	1.53	2.68	1.3	0.56	2.34	0.99

Table 6.6: Sensitivity of the robust M-t model performances to the fitted degrees of freedom. Risk free rate= .03 (monthly), p=4.

	tangency solution			minimum variance solution		
df	turnover	risk	return	turnover	risk	return
4	0.83	0.88	0.59	0.19	0.67	0.44
8	0.79	0.91	0.58	0.18	0.7	0.44
15	0.78	0.93	0.58	0.18	0.72	0.44
30	0.78	0.95	0.58	0.18	0.74	0.44
Markowitz	0.78	0.98	0.58	0.18	0.76	0.44

Table 6.7: Out-of-sample performances of the tangency portfolio on a real data set. Risk free rate= .03

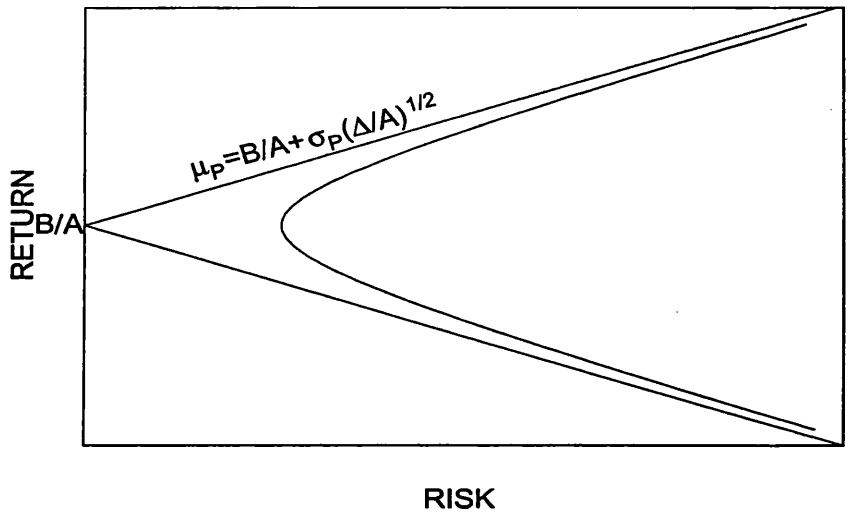
method	turnover	risk	st. dev. risk	return	st. dev. return
Markowitz	0.62	2.03	0.78	0.55	0.55
MCD	0.69	1.53	0.53	0.58	0.58
M-t	0.58	1.56	0.53	0.63	0.63

Table 6.8: Out-of-sample performances of the global minimum variance portfolio.

method	turnover	risk	st. dev. risk	return	st. dev. return
Markowitz	0.15	1.68	0.67	0.56	0.43
MCD	0.27	1.40	0.52	0.64	0.65
M-t	0.14	1.35	0.42	0.67	0.46

Figure 6.1: Portfolio frontiers.

(a) Mean-Standard Deviation Frontier



(b) Tangency Portfolio

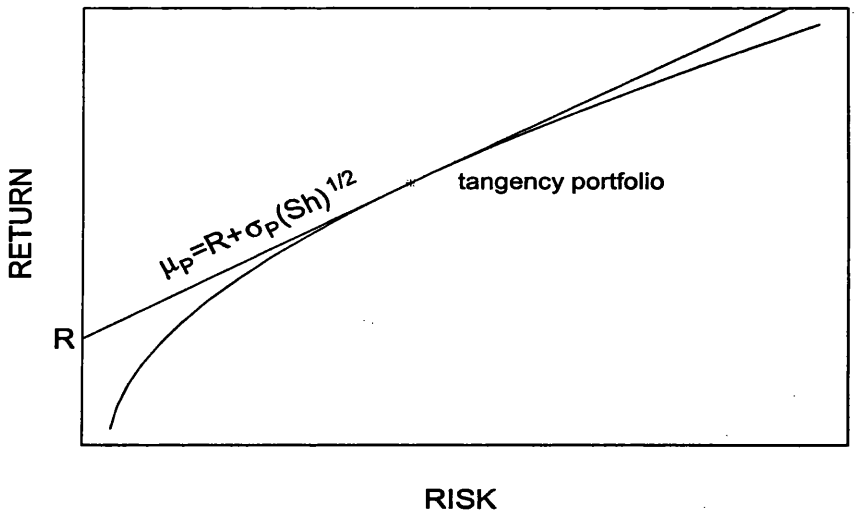
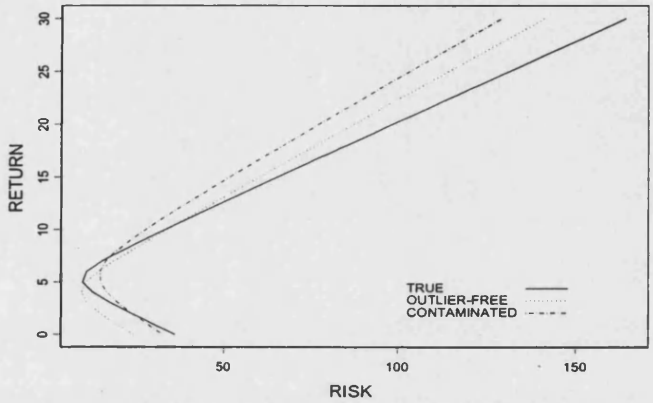
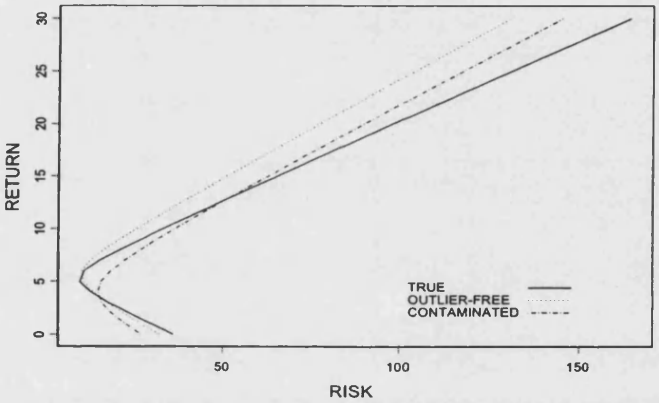


Figure 6.2: Contaminated and outlier-free portfolio mean-variance frontiers.

(a) Positive mean-shift outliers



(b) Negative mean-shift outliers



(c) Outliers generated from a multivariate Student-t

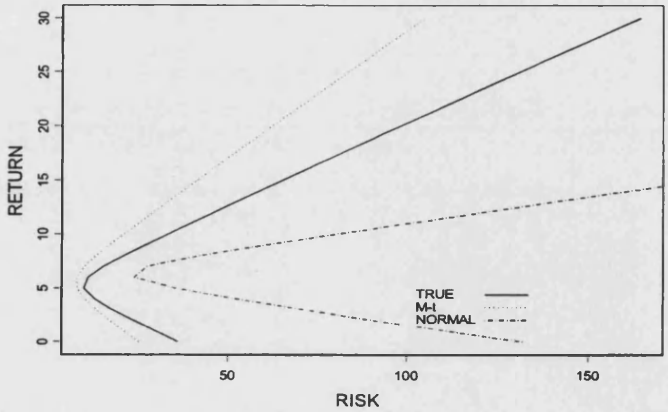
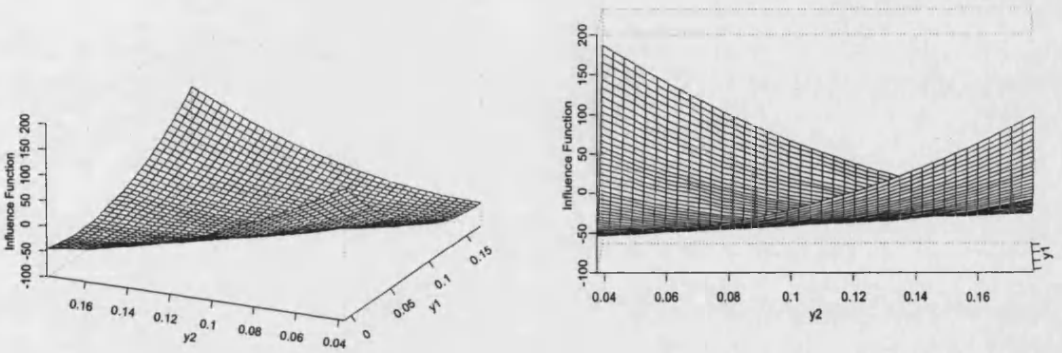
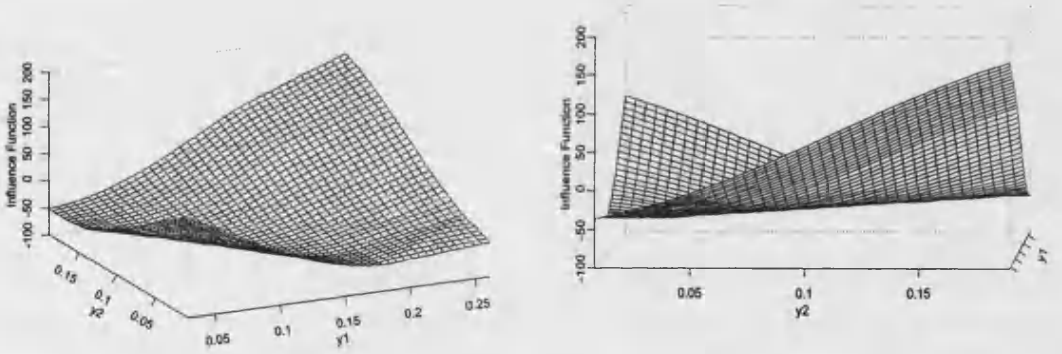


Figure 6.3: Influence function for the ML mean of the global minimum-variance portfolio from a bivariate Normal distribution and a Student-t.

(a) Multivariate Normal distribution



(b) Multivariate Student-t with 10 degrees of freedom



(c) Multivariate Student-t with 3 degrees of freedom

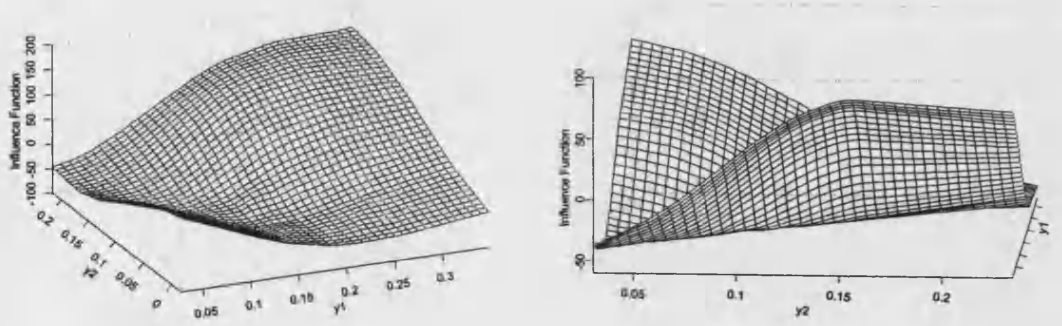
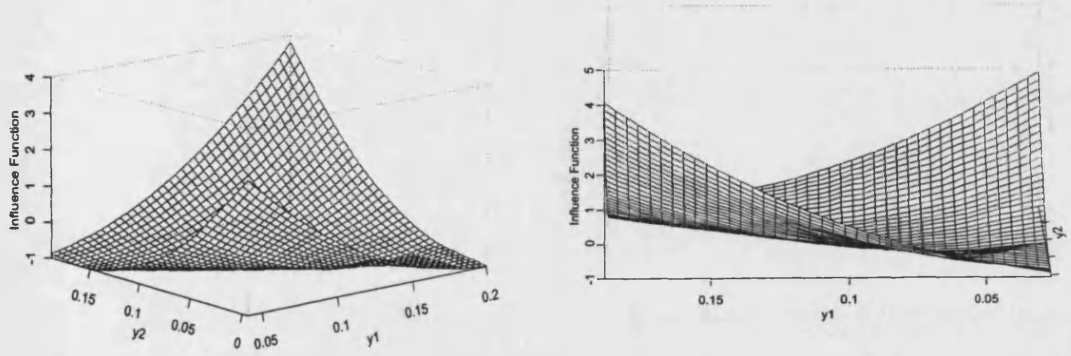
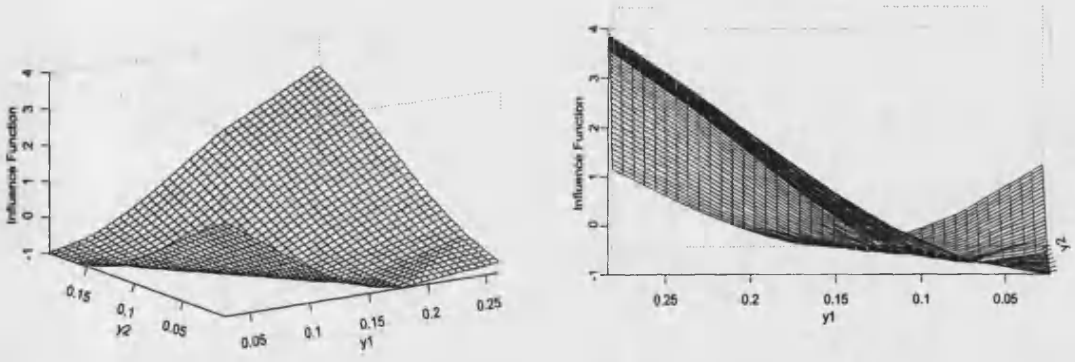


Figure 6.4: Influence function for the ML variance of the global minimum-variance portfolio from a bivariate Normal distribution and Student-t.

(a) Multivariate Normal distribution



(b) Multivariate Student-t with 10 degrees of freedom



(c) Multivariate Student-t with 3 degrees of freedom

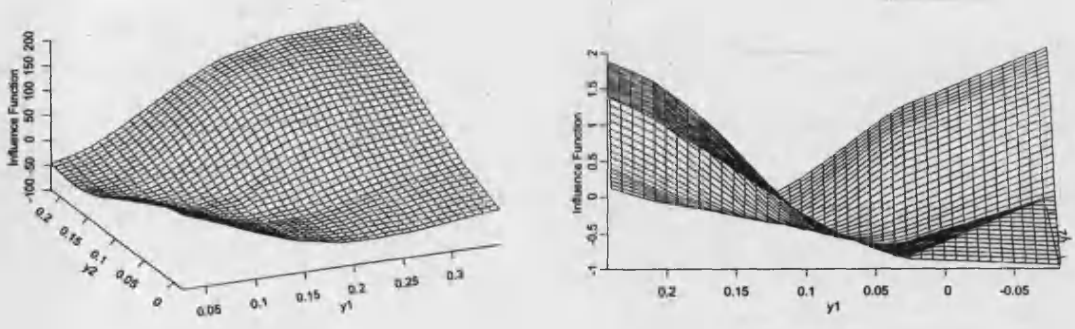
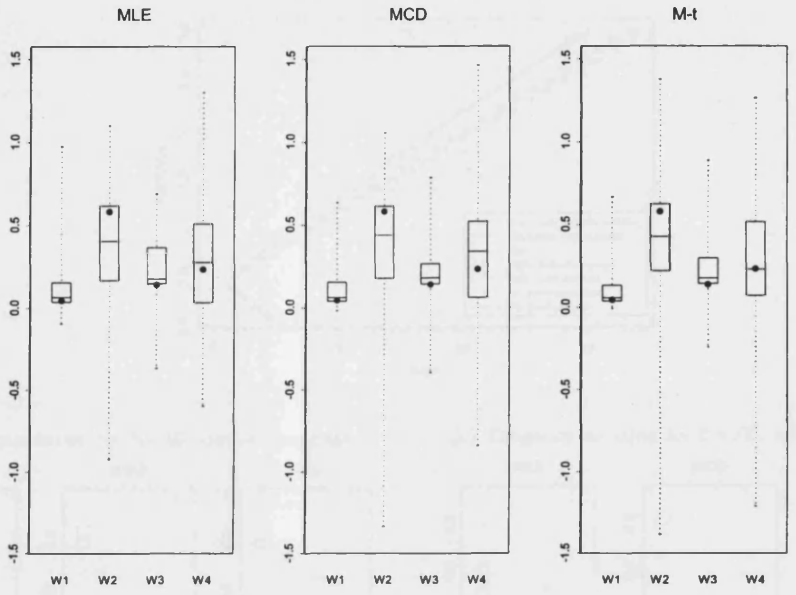


Figure 6.5: Distribution of the weights for Merton's model (6.3)-(6.5) on Normal data. n=200, p=4, q=12% annual.

(a) Optimal solution for q=1



(b) Minimum-variance solution.

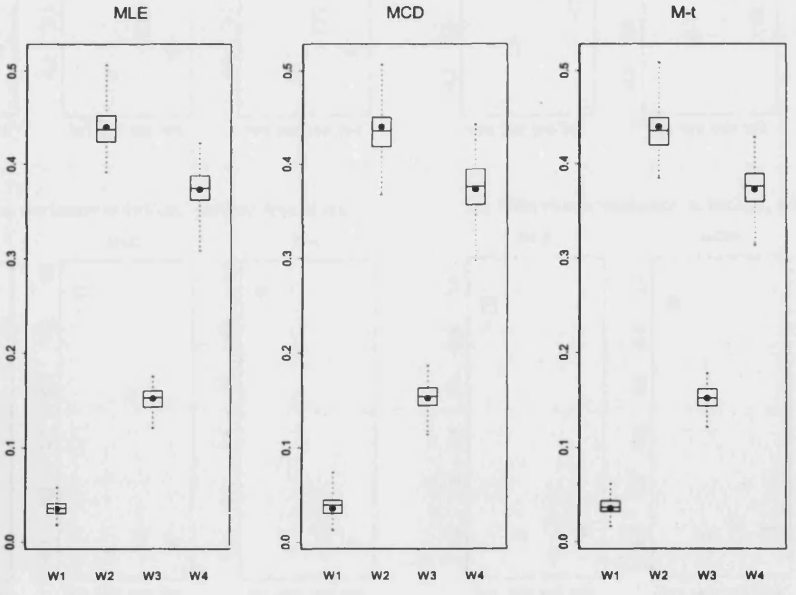
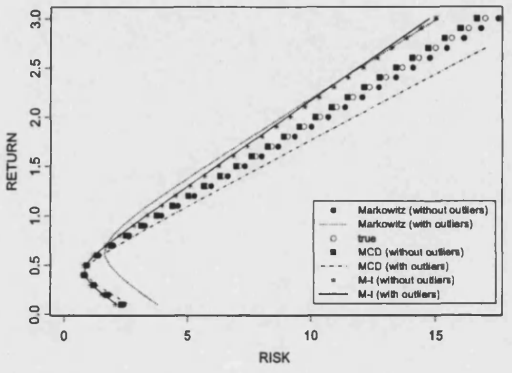
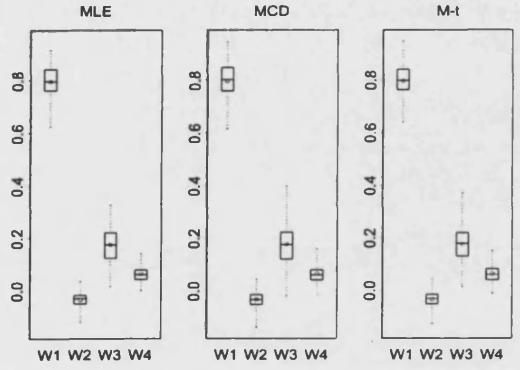


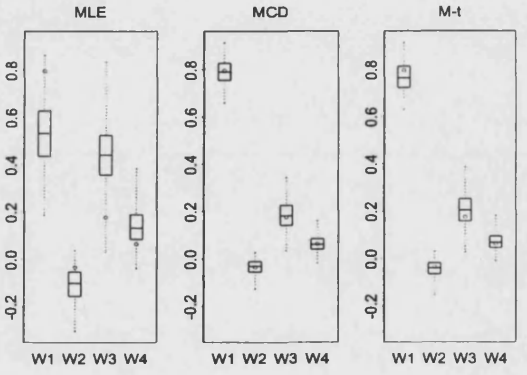
Figure 6.6: (a) Average mean-variance frontiers



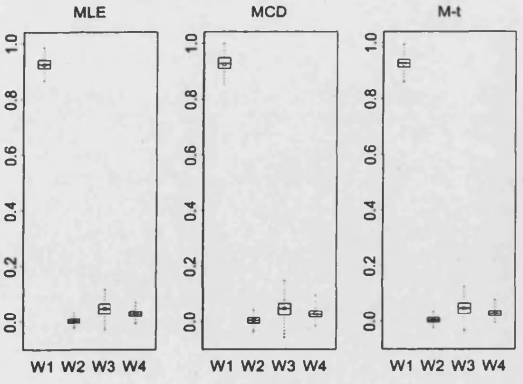
(b) Tangency solution for R=.03, outlier-free data



(c) Tangency solution for R=.03, mixed data, ϵ=.02



(d) Minimum variance solution, outlier-free data



(e) Minimum variance solution, mixed data, ϵ=.02

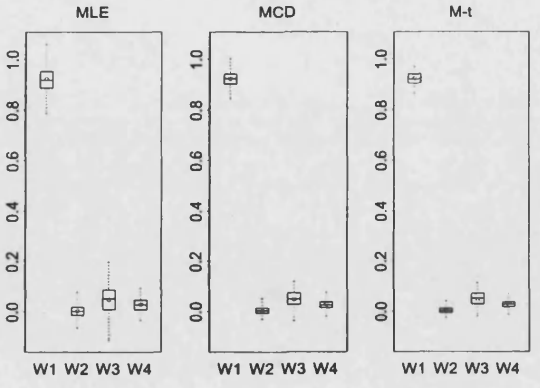
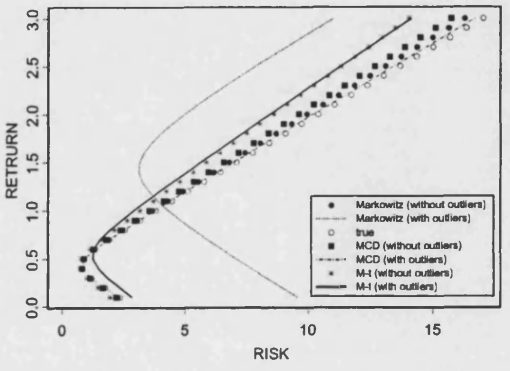
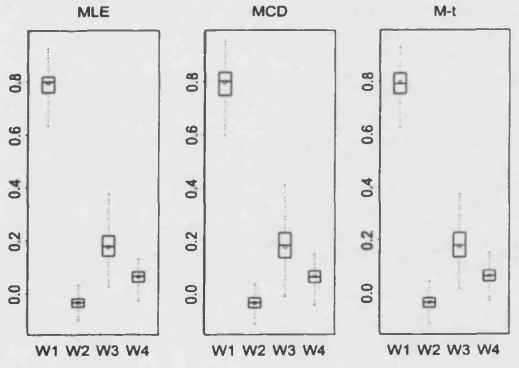


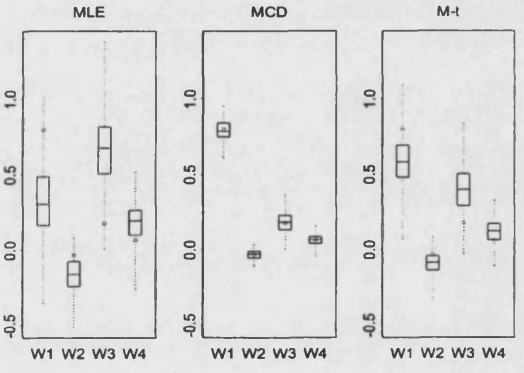
Figure 6.7: (a) Average mean-variance frontiers



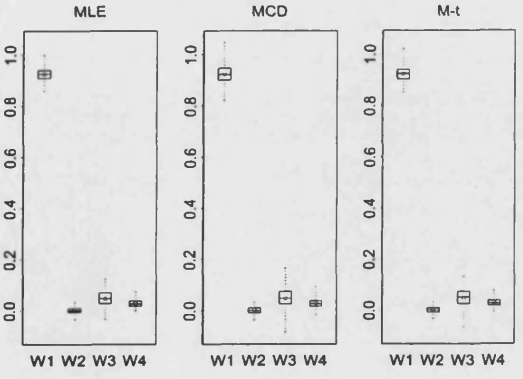
(b) Tangency solution for R=.03, outlier-free data



(c) Tangency solution for R=.03, mixed data, ϵ=.1



(d) Minimum variance solution, outlier-free data



(e) Minimum variance solution, mixed data, ϵ=.1

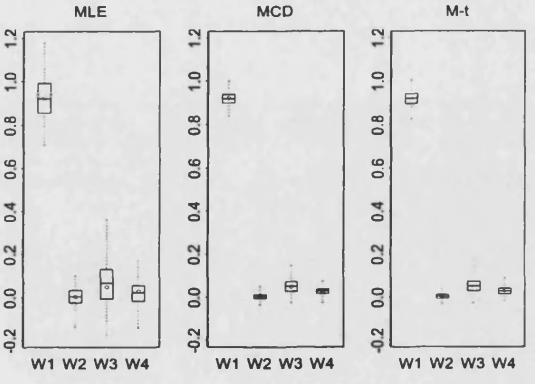


Figure 6.8: Bias for the average μ and for the determinant of Ψ for increasing sample sizes.

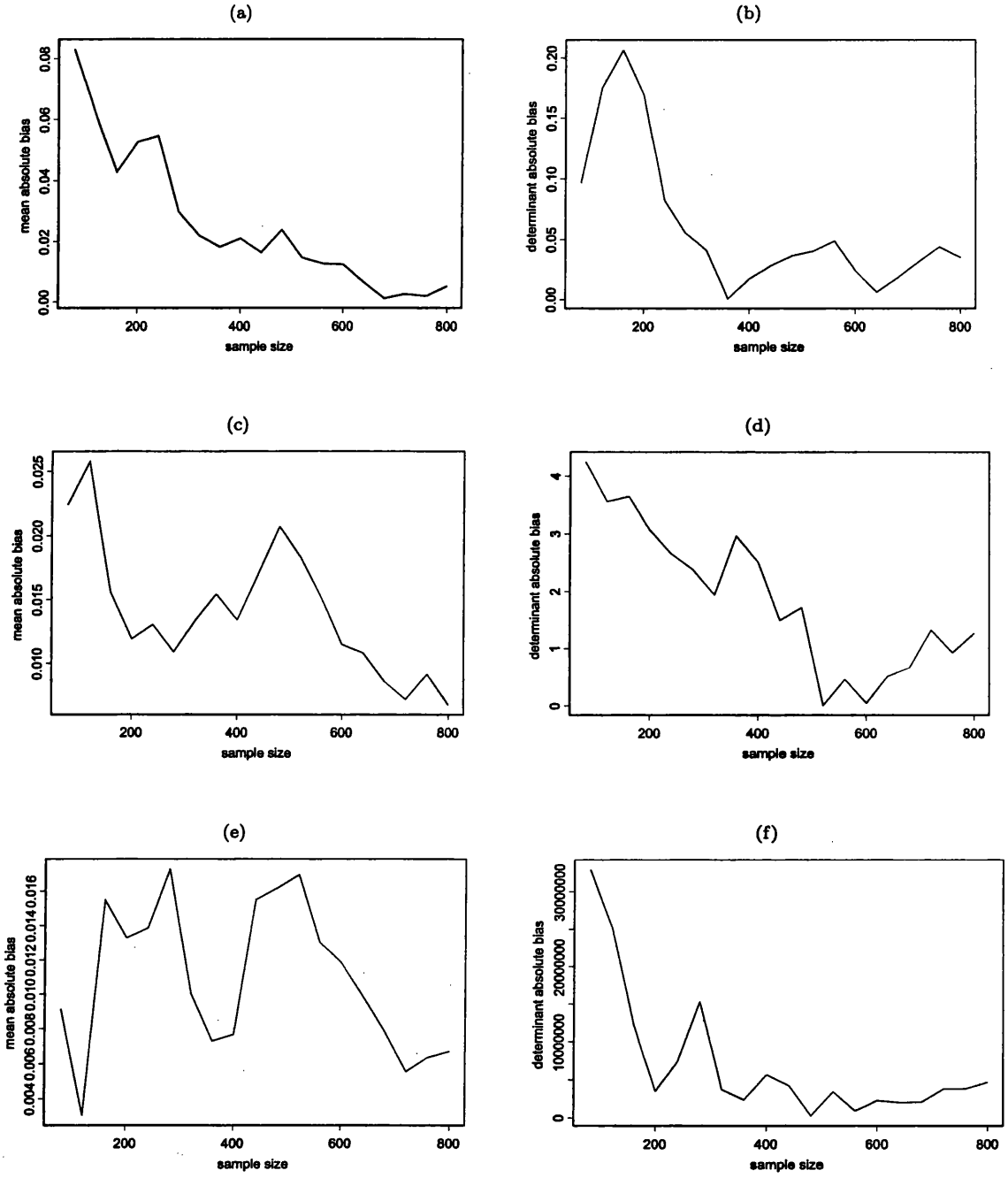


Figure 6.9: Average standard deviation for the MLE on a multivariate Student-t. The sample size varies from 80 to 800 observations, the number of replications are 120.

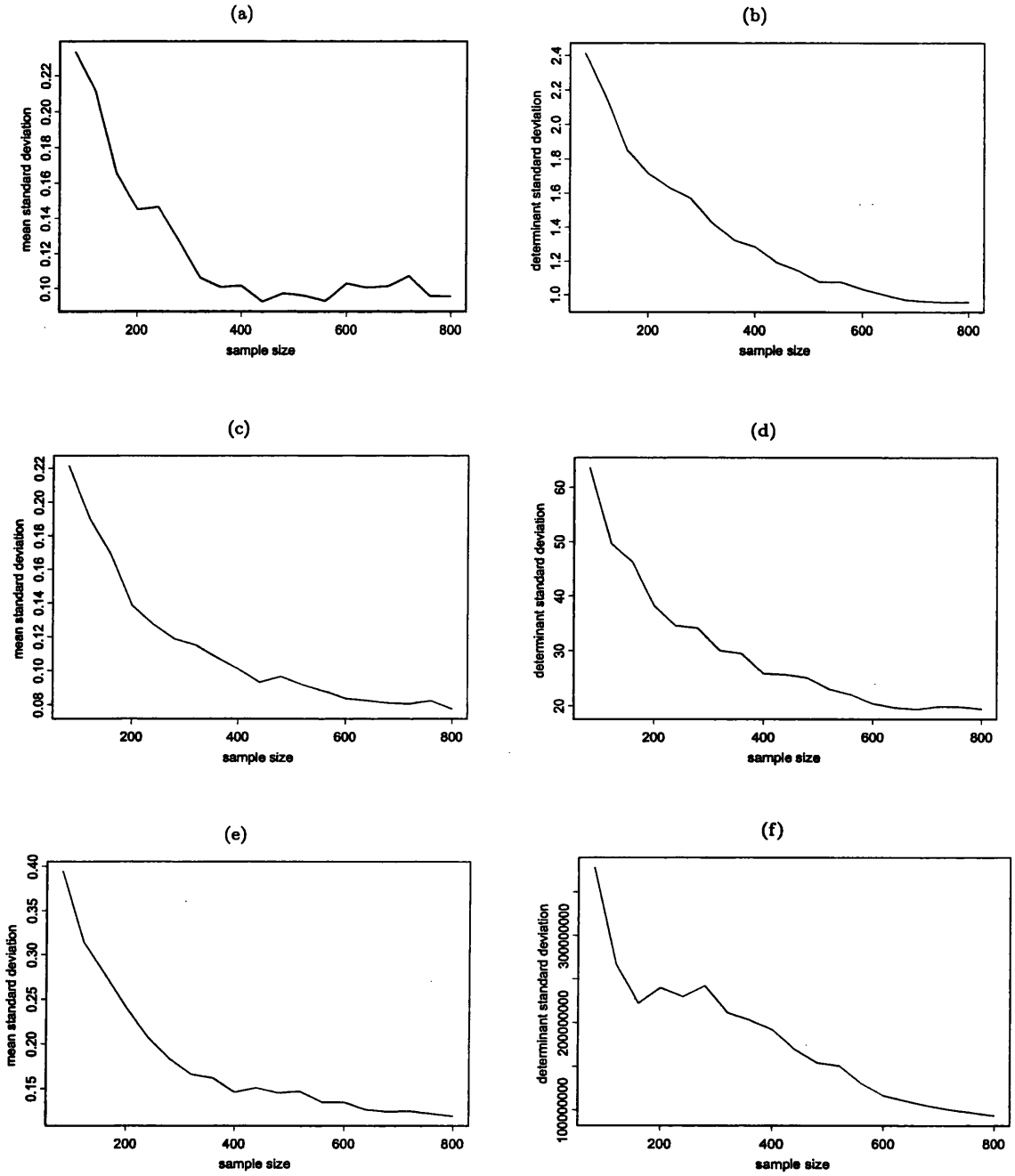


Figure 6.10: Morgan Stanley monthly stock return indexes of 5 countries.

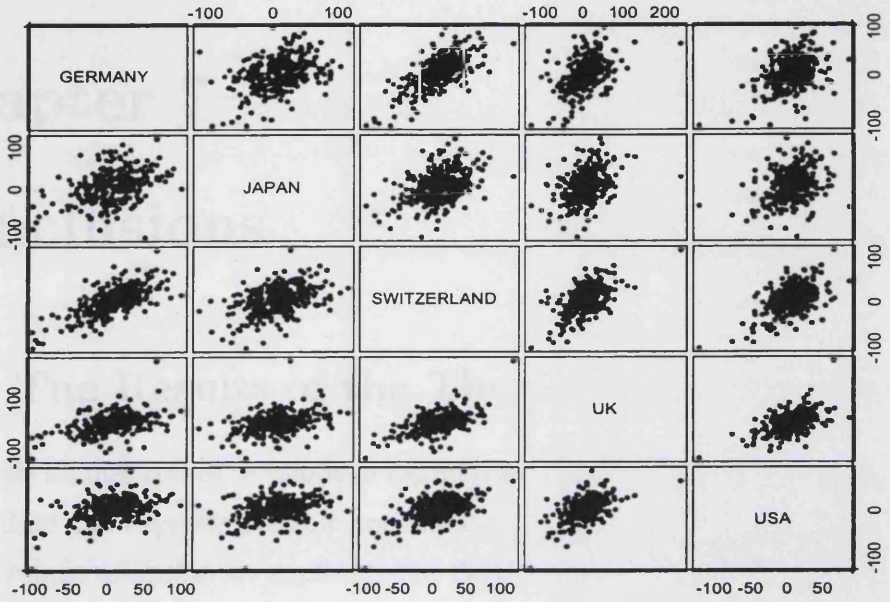
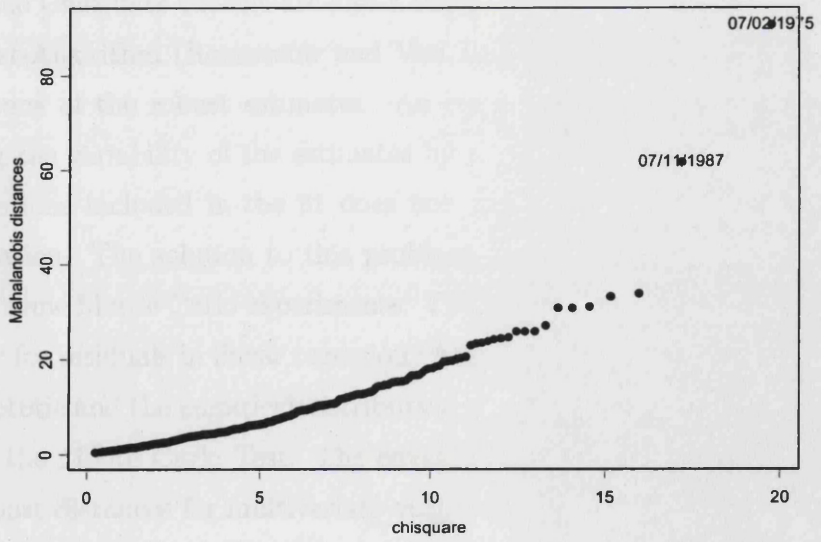


Figure 6.11: Mahalanobis distances on the Student-t fit for the stock data.



Chapter 7

Conclusions

7.1 The Results of the Thesis

The thesis studies robust methods of estimating the location and scatter for multivariate data and suggests possible applications.

Very robust estimates are applied to the detection of multivariate outliers because they avoid swamping and masking problems. The thesis gives evidence that the Chisquare confidence regions for outliers do not well approximate the distribution of the robust distances based on the high-breakdown center and scatter-matrix. These tolerance regions lead to the rejection of outliers as well as "good" observations. Because the Chisquare cut-offs are also a decision rule to reweight the observations in the Fast-Algorithm (Rousseeuw and Van Driessen 1999), the result is a loss in the efficiency of the robust estimates. An extensive simulation study shows that improving the variability of the estimates by reweighting or increasing the number of observations included in the fit does not produce "significant" changes in the approximation. The solution to this problem is a correction factor, which can be found via some Monte Carlo experiments. The simulation envelope, introduced as a diagnostic for residuals in linear regession (Atkinson 1985), is suggested to compare the asymptotic and the empirical distribution of the robust distances, enhancing the results of the Monte Carlo Test. The envelopes are also applied for the first time to the robust distances for multivariate outliers detection avoiding the problems of overidentification. In small samples, the distribution of the Mahalanobis distances is known: Penny (1996) provides the critical points on the basis of the result of Wilks

(1962) for the scatter ratio, a diagnostic used to detect a single outlying point, although the proof is rather difficult. The thesis offers a simpler way to find the small sample behaviour of the distances, agreeing with the results of Penny (1996) and Wilks (1962).

The computational aspect of the robust estimates is as important as their theoretical properties. A clue issue which has not yet found a clear solution is the choice of the subset of data from which the search for the minimum covariance-determinant solution starts. The consequence of a wrong choice is a loss in robustness or efficiency of the estimates, depending on whether we include too many or too few observations in the "good" set. In the context of the robust portfolio model, the performances do not seem to be very sensitive to relatively small errors in estimating the number of outliers. However, in Chapter 3, we suggest an application of the Forward Search algorithm (Atkinson 1994) to identify the initial set of maximum size which does not bias the estimates.

In Chapter 6, high-breakdown point estimates are used to construct a model for selecting optimal financial portfolios. From an exploratory analysis of a real data set, we have shown that stock returns have longer tails than the Normal distribution. This is a "stylized" fact in the financial literature. In presence of few, either positive or negative, outliers the classical Markowitz portfolio produces biased estimates of the true portfolio weights. The same result is shown if the underlying distribution is a Student-t, longer tailed than the Normal. The Influence Function derived on the Markowitz "classical" optimizer shows high sensitivity to outliers. This result is expected since the function depends on the Normal maximum likelihood estimates for location and scatter, known to be non-robust (Huber 1981).

We explore the performances of two robust models: one is based on the MCD estimates of location and scatter and the other assumes a multivariate Student-t distribution for the stock returns. These performances are compared with the classical Markowitz optimizer (Markowitz 1952) via a simulation study, focusing on the distribution of the weights, which determine the budget quota to invest in each asset. When there are no outliers, the results of the robust and the "classical" portfolio models are very similar, with a negligible loss in efficiency for the robust

method. In presence of contaminated data, the robust weights are more stable overtime and the risk is minimum. Under these conditions, the robust portfolios are also shown to be consistent for the true parameters, differently from the weights produced by the "classical" optimizer that are biased by the presence of outliers. From comparing the performances of the two robust methods, the MCD estimates produce less risky and more stable portfolios when it is possible to identify two groups of observations in the data: a main bulk of points and a smaller group generated from a different model. The multivariate Student-t portfolios are however preferable when there are very few or no outliers and the data indicate a longer tailed distribution than the Normal. The results of the simulation studies are confirmed by an example on a real data set introduced in Chapter 2.

The Student-t estimates are produced assuming the degrees of freedom are observed. However, with sufficient data, these can be estimated from the data simultaneously to the mean and inner-product matrix (adaptive procedure). It is shown that the adaptive procedure, in addition to being a computationally heavy method, also provides estimates with a high variability. Even when the degrees of freedom are known, the estimates are less efficient than the Normal distribution. The high sampling variabilty, which is particularly evident for the MLE for Ψ, also determines a slow convergence to the true parameter values. Generally, a large sample size is needed to be to produce "good" estimates. However, in the portfolio selection context, this is not feasible since we are treating monthly observations.

7.2 Suggestions for further Studies

The thesis leaves many open questions. Further work is required for the robust optimizer. The high-breakdown point estimators work under the assumption that the majority of the data are normally distributed and the observations are independent. This restrictive assumption also underlies the Markowitz Mean-Variance theory. Even though stock returns have very low autocorrelations, this does not imply independence. We have shown, although it is well known in the literature, the existence of correlations of higher order or of some function of the initial variables

(Figure 2.1). In other words, we can model the volatility of the returns, which suggests a possible robust dynamic optimizer, definitely worth exploring. The Forward Search, because of its dynamic structure, could be a possible application of robust computation to time series.

Furthermore, the applications of high-breakdown estimators are numerous and still not thoroughly studied. Little has been done on the applications to categorical data models.

Bibliography

Agullò, J. (1996). Exact iterative computation of the multivariate Minimum Volume Ellipsoid estimator with a Branch and Bound algorithm. In A. Prat (Ed.), *Proceedings in Computational Statistics*, Heidelberg, pp. 175–180. Physica-Verlag.

Arslan, O., D. L. Constable, and J. T. Kent (1995). Convergence behaviour of the EM algorithm for the multivariate t-distribution. *Communications in Statistics: Theory and Methods 24*, 2981–3000.

Atkinson, A. (1994). Fast very robust methods for the detection of multiple outliers. *Journal of the American Statistical Association No. 89*, 1329–1339.

Atkinson, A. and T.-C. Cheng (2000). On robust linear regression with incomplete data. *Computational Statistics and Data Analysis 33*, 361–380.

Atkinson, A. and H. M. Mulira (1993). The stalactite plot for the detection of multivariate outliers. *Statistics and Computing 3*, 27–35.

Atkinson, A. C. (1985). *Plots Transformation and Regression.* Oxford Statistical Science Series. Oxford: Oxford University Press.

Atkinson, A. C. and M. Riani (2000). *Robust Diagnostic Regression Analysis.* Springer Series in Statistics. New York: Springer.

Bickel, P. J., C. A. Klaassen, Y. Ritov, and J. A. Wellner (1993). *Efficient and Adaptive Estimation of Semiparametric Models.* Baltimore and London: John Hopkins University Press.

Bradu, D., D. M. Hawkins, and G. Kass (1984). Location of several outliers in multiple regression data using elemental sets. *Technometrics 26*, 197208.

Brownlee, K. A. (1965). *Statistical Theory and Methodology in Science and Engineering.* New York: John Wiley and Sons.

Butler, R. W., P. L. Davies, and M. Juhn (1993). Asymptotics of the Minimum Covariance Determinant estimator. *The Annals of Statistics 21*, 1385–1400.

Campbell, N. A. (1980). Robust procedures in multivariate analysis i: Robust covariance estimation. *Applied Statistics 29*, 231–237.

Cook, R. R. and S. Weiseberg (1986). *Residuals and Influence in Regression.* Monographs on Statistics and Applied Probability. New York: Chapman and Hall.

Croux, C. and G. Haesbroeck (1997). An easy way to increase the finite-sample efficiency of the resampled Minimum Volume Ellipsoid estimator. *Computational Statistics and Data Analysis 25*, 125–141.

Croux, C. and G. Haesbroeck (1999). Influence function and efficiency of the Minimum Covariance Determinant scatter matrix estimator. *Journal of Multivariate Analysis 71*, 161–190.

Croux, C. and P. J. Rouseeuw (1992). A class of high-breakdown scale estimators based on subranges. *Communications in Statistics: Theory and Methods 21*, 1935–1951.

D'Agostino, R. B. and M. A. Stephens (1986). *Goodness-of-Fit Techniques.* New York: Marcel Dekker Inc.

Davies, L. (1987). Asymptotics of s-estimators of multivariate location parameters and dispersion matrices. *The Annals of Statistics 15*, 1269–1292.

Davies, L. (1992). The asymptotics of rousseeuw's Minimum Volume Ellipsoid estimator. *The Annals of Statistics 20*, 1828–1843.

Davies, L. (1993). Aspects of robust linear regression. *The Annals of Statistics 21*, 1843–1899.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). ML for incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B 39*, 1–38.

Draper, N. R. and H. Smith (1966). *Applied Regression Analysis.* New York: John Wiley and Sons.

Fama, E. (1965). The behaviour of stock prices. *Journal of Business 38*, 34–105.

Fernandez, C. and M. Steel (1999). Multivariate student-t regression models. *Biometrika 87*, 153–167.

Grübel, R. (1988). The length of the shorth. *The Annals of Statistics 16*, 619–628.

Grübel, R. and D. M. Rocke (1990). On the cumulants of affine equivariant estimators in elliptical families. *Journal Multivariate Analysis 35*, 1–20.

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

Hawkins, D. M. (1993). A feasible solution algorithm for the Minimum Volume Ellipsoid estimator in multivariate data. *Computational Statistics 8*, 95–107.

Hawkins, D. M. (1994). A feasible solution algorithm for the Minimum Covariance Determinant estimator. *Computational Statistics 17*, 197–210.

Hawkins, D. M. and D. J. Olive (1999). Improved feasible solution algorithms for high breakdown estimation. *Computational Statistics and Data Analysis 30*, 1–11.

Horowitz, J. L. (1998). *Semiparametric methods in econometrics.* Berlin: Springer Verlag.

Huang, C. and R. H. Litzenberger (1988). *Foundations for financial economics.* New York: Prentice Hall.

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

Ingersoll, J. R. (1987). *Theory of Financial Decision Making.* New York: Rowman and Littlefield.

Kent, J. T. and D. E. Tyler (1996). Constrained M-estimation for multivariate location and scatter. *The Annals of Statistics 24*, 1346–1370.

Kent, J. T., D. E. Tyler, and Y. Vardi (1994). A curious likelihood indentity for the multivariate t-distribution. *Communications in Statistics: Simulation and Computation 23*, 441–453.

Kon, S. (1984). Models of stock returns - a comparison. *Journal of Finance 39*, 147–165.

Lange, K. and J. S. Sinsheimer (1993). Normal/independent distributions and their application in robust regression. *The Journal of Computational and Graphical Statistics 2*, 175–198.

Lange, K. L., R. J. A. Little, and J. M. G. Taylor (1989). Robust statistical modelling using the t distribution. *Journal of the American Statistical Association 84*, 881–895.

Linden, M. (2001). A model for stock return distribution. *International Journal of Finance and Economics 6*, 159–169.

Liu, C. and D. B. Rubin (1995). ML estimation of the t dstribution using EM and its extensions ECCM and ECME. *Statistica Sinica 5*, 19–39.

Lopuhäa, H. P. (1989). On the relation between S-estimators and M-estimators of multivariate location and covariance. *Annals of Statistics 17*, 1662–1683.

Lopuhäa, H. P. (1991). Multivariate t-estimators for location and scatter. *Canad. J. Statist. 19*, 307–321.

Lopuhäa, H. P. (1999). Asymptotics of reweighted estimators of multivariate location and scatter. *Annals of Statistics 27*, 1638–1665.

Mardia, K. V., J. T. Kent, and J. M. Bibby (1982). *Multivariate Analysis.* London: Academic Press.

Markowitz, H. (1952). Portfolio selection. *Journal of Finance 7*, 77–91.

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

Maronna, R. A. and V. J. Yohai (1998). Robust estimation of multivariate location and scatter. In S. Kotz, C. Read, and D. Banks (Eds.), *Encyclopedia of Statistical Sciences Update.*

Martin, R. D. (1999). Outliers, influence functions and robust portfolio optimization. S-plus Workshop. City University, London.

Meng, X. L. and D. B. Rubin (1993). Maximum Likelihood Estimation via ECM algorithm: a general framework. *Biometrika 80*, 267–278.

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

Nolan, D. (1991). The excess mass ellipsoid. *Journal of Multivariate Analysis 39*, 348–371.

Penny, K. I. (1996). Appropriate critical values when testing for a single multivariate outlier by using the Mahalanobis distance. *Applied Statistics 45*, 73–81.

Powell, J. L. (1994). Estimation in semiparametric models. In Springer (Ed.), *Handbook of Econometrics*, Volume IV. North Holland: Engle, R. F. and McFadden, D. F.

Rao, C. R. (1973). *Linear Statistical Inference and its applications.* Series in Probability and Mathematical Statistics. New York: John Wiley and Sons.

Richardson, M. and T. Smith (1994). A direct test of the mixture of distributions hypothesis: Measuring the daily flow of information. *Journal of Financial and Quantitative Analysis 29*, 101–116.

Roll, R. (1988). R^2. *Journal of Finance 43*, 541–566.

Rousseeuw, P. (1983). Location M-estimators are characterized by the infinitesimal behavior of their asymptotic variance. *Bulletin de la Socit Mathmatique de Belgique 35 (B)*, 167–176.

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

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown points. In W. Grossmann and et al. (Eds.), *Mathematical Statistics and its Applications Vol B*, pp. 283–297. Dordrecht: Reidel Publishing Company.

Rousseeuw, P. J. and M. A. Leroy (1987). *Robust Regression and Outlier Detection.* New York: John Wiley and Sons.

Rousseeuw, P. J. and K. Van Driessen (1999). A fast algorithm for the Minimum Covariance Determinant estimator. *Technometrics 41*, 212–223.

Rousseeuw, P. J. and B. Van Zomeren (1991). Robust distances simulations and cutoff values. In S. W. A. and S. Weiseberg (Eds.), *Directions in Robust Statistics Volume II*, The IMA Volumes of Mathematics and its Applications, New York. Springer Verlag.

Sharpe, W. (1975). Adjusting for risk in portfolio performance measurement. *Journal of Portfolio Management*, 29–34.

Stephens, M. A. (1978). On the half-sample method for goodness-of-fit. *Journal of the Royal Statistical Society B*, 64–70.

Vilasuso, J. and D. Katz (2000). Estimates of the likelihood of extreme returns in international stock markets. *Journal of Applied Statistics 27*, 119–130.

Wilks, S. S. (1962). *Mathematical Statistics.* New York: John Wiley and Sons.

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

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