Modern Portfolio Optimization using Robust Estimation Techniques

by

Conrad van Straaten (VSTCON001)

Dissertation presented in fulfilment of the requirements of the degree M.Sc in Financial Mathematics

at the

University of Cape Town

Supervisor: Prof C.G. Troskie

Submission Date: February 2005

Acknowledgements

Firstly, I would like to thank my parents. For without their support, I would not have been able to complete this course. Secondly, I would like to thank Peter Ouwehand, who gave me the opportunity to study Financial Mathematics. Finally, I would like to thank Professor Cas G. Troskie. Without your invaluable insight into the problem I would not have been able to complete this dissertation. Thank you for your enthusiasm and willingness to answer all my questions. I hope you enjoyed working with me as much as I enjoyed working with you.

Index

page

Introduction

Introduction 1

Problem Statement 1

Portfolio Layout 1

1 Ordinary Least Squares

1.1 Introduction 3

1.2 Background and Notation 3

1.3 Ordinary Least Squares 5

1.4 Outliers 6

2 Low Breakdown Regression Procedures

2.1 Introduction 9

2.2 *M*-Estimators 11

2.3 Bounded Influence Estimators 13

2.4 The Ψ-function 15

3 Multivariate Location and Scale Estimation

3.1 Introduction 18

3.2 Classical Estimation technique 19

3.3 Outlier Resistant Methods 20

3.3.1 Non Affine Equavariant Estimators

3.3.1.1 Coordinate Median 20

3.3.1.2 Hadi's Forward Search 21

3.3.2 Affine Equivariant Estimators

3.3.2.1 Convex Peeling 23

3.3.2.2 Transformed One-step Weighted dispersion Estimator 23

3.3.3 Affine Equiv4ariant Estimators with High Breakdown Pionts

3.3.3.1 Stahel-Donoho Estimator 24

3.3.3.2 Minimum Volume Ellipsoid Estimator 26

3.3.3.3 Minimum Covariance Determinant Estimator 31

4 High Breakdown Regression Estimators

4.1 Introduction 34

4.2 Least Median of Square 34

4.3 Least Trimmed Squares 37

4.4 One-step Generalized *M*-Estimators 38

4.4.1 Mallows 1-Step Estimator 40

4.4.2 Schweppe 1-Step Estimator 42

4.4.3 Case Study: Stackloss Data 44

4.5 Computational Issues for High breakdown Regression 46

5 The ARCH Model

5.1 Introduction 48

5.2 The ARCH(1) Model 49

5.3 Testing for ARCH Errors 51

5.4 Parameter Estimation 53

6 The GARCH Model

6.1 Introduction 55

6.2 The GARCH(1,1) Model 55

6.3 The GARCH(p,q) Model **56**
6.4 GARCH Model Limitations **57**
6.5 The Exponential GARCH **58**

7 Markowitz Portfolio Selection

7.1 Markowitz Portfolio Theory **60**
7.2 The Efficient Frontier **65**

8 Sharpe Portfolio Selection

8.1 Sharpe Single Index Model Theory **67**
8.2 The Sharpe Multiple Index Model **72**

9 Innovations in the Classical Models

9.1 Generalisation of the Markowitz Formulation **78**
9.2 Innovations to the Sharpe Single Index Model
9.2.1 Bayesian Estimates in the Sharpe Single Index Model **80**
9.2.2 Non-zero Covariance between Residuals of Shares **81**

10 Data Used in Analysis **85**

11 Results

11.1 Introduction **86**
11.2 Case Study 1 **87**
11.3 Case Study 2 **93**
11.4 Case Study 3 **99**
11.5 Case Study 4 **100**

12 Conclusions 103

References 104

Introduction

Introduction

Harry Markowitz developed modern portfolio theory in the 1950's. His development is still used in portfolio construction today. Sharpe, who introduced index models into the portfolio theory, simplified the Markowitz portfolio theory. This simplification was as a result of fewer parameters that have to be estimated in the Sharpe portfolio theory. In both of these models it is assumed that the underlying distribution of the data is Normal. This assumption allows for the use of Ordinary Least Squares (OLS) regression to determine the values of the parameters in the models. This assumption however, does not hold for stock markets and therefore it is a reasonable assumption that OLS regression does not provide the best parameter estimation. This can in turn lead to the misrepresentation of the optimal portfolio.

Problem Statement

Rather than following a normal distribution, share returns and market proxies have been shown to follow skewed distributions, with long tails in some cases. In this dissertation various robust estimation techniques are investigated in an attempt to minimise the influence that outliers may have on the estimation and to better estimate the input parameters for the Markowitz and Sharpe portfolio models. The main goal is to ascertain whether or not the input parameters determined, using the robust procedures, yield better results than the OLS procedure.

Dissertation Layout

This dissertation can be split into four sections. In the first section a brief discussion is given on OLS and why it is not the best regression procedure to pursue when dealing with data that contains influential observations. Robust estimators are then discussed,

with comments being made about difficulties with estimation, shortcomings in the estimators, improvements on some of the estimators etc. The robust procedures discussed in this section is used to determine the covariance matrix used as an input parameter into the Markowitz portfolio problem.

In the second section of this dissertation, ARCH and GARCH procedures are discussed as another possible estimation technique. These procedures estimate the residual errors in the, in an attempt to obtain a better model estimation. These two estimation procedures are used to determine the beta and variance parameters that are used as inputs in the Sharpe portfolio problem.

In the third section the Markowitz and Sharpe portfolio selection procedures are discussed along with innovations that have been made to the models.

In the fourth section the data that is used in the analysis of the data is discussed. Several case studies are performed on the data set using some of the robust procedures discussed. The results are stated, in the form of graphs and tables, and commented on. In the final chapter conclusions that are reached is discussed.

Chapter 1

Ordinary Least Squares

1.1 Introduction

"The world is essentially linear and normal", a joke that frequently circulates in the statistical community. This is because, when dealing with analytic procedures in statistics, the assumption is often made that the data is linear and normal. Under this assumption there exists an overabundance of statistical theory which allows the analysis of data over a broad field of application. It is this assumption that often leads to misguiding results and conclusions.

Hampell, Ronchetti, Rousseeuw and Stahel (1986) states the following, "…routine data are thought to contain 1% to 10% gross errors." In general this proves to be true as real data is generally considered not to be well behaved in terms of linearity and normality. In the field of regression, which studies the relationship between two sets of variables, the data may exhibit a linear trend between the two sets of variables, even though the data may contain points that do not fit the general trend of the data. Examples of these wild observations, referred to as outliers, will be shown in a later section in the chapter. Outliers in the data can occur as a result of human error, in either the execution of the experiment or in the capturing of the data. These outliers may indicate that the proposed model does not fit the data properly. The rest of this chapter deals with classical regression.

1.2 Background and Notation

Regression can be thought of as the relationship between two sets of variables. The two sets are called the regressor variables and the response variables. The regressor variables are often referred to as the independent variables. This set of k regressor

variables, $x_1, x_2, \ldots, x_k$, are assumed to be fixed and without error. The numerical values of the regressor variables can be predetermined or observed as data points. The second set of variables contains the dependant variables and in the case of univariate regression the second set of variables contains only one response variable, y.

As a statistical model, regression models explain the response variable as a function of the regressor variables with the addition of a random error term to account for individual differences. The regression model can be defined as

$$y_i = f(x_{1i}, x_{2i}, \ldots, x_{ki}) + \varepsilon_i, \quad i = 1,2,\ldots,n,$$

where n is the number of observations. The functions for the response variables will be restricted to the linear functions of the type

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i .$$

The linearity of the function is with respect to the p unknown parameters, the vector β.

The standard linear regression notation, the data is organised into a $n \times 1$ response vector, y, and a $n \times p$ regressor matrix, $\mathbf{X}$. If we define the β as a $p \times 1$ parameter vector and the as a $n \times 1$ random error vector, the linear model can now be defined as $\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$. The model can now be written as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

It becomes necessary to eliminate the columns of ones in $\mathbf{X}$ when working with multivariate location and scale estimation. $\mathbf{Z}$ is defined as the $n \times k$ matrix, containing only the k regressor variables and can be written as

$$\mathbf{Z} = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{k1} \\ x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \vdots \\ x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix},$$

with $\mathbf{Z}_y$ representing the augmented matrix that is formed by augmenting the vector $\mathbf{y}$ to $\mathbf{Z}$. The $\mathbf{Z}_y$ can be written as

$$\mathbf{Z}_y = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{k1} & y_1 \\ x_{12} & x_{22} & \cdots & x_{k2} & y_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1n} & x_{2n} & \cdots & x_{kn} & y_n \end{bmatrix}.$$

When referring to estimates the standard "hat" notation is employed. One example is $\hat{\beta}$, this presents the $p \times 1$ vector of parameter estimates. Classical regression also makes assumptions about the random error term in the formulation. It is assumed that the error terms are independent, identically distributed (iid) from a normal distribution with mean 0 and variance σ^2. The variance is assumed to be constant.

1.3 Ordinary Least Squares

The solution to the following minimization problem,

$$\min_{\forall b} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \ldots - b_k x_{ik})^2,$$

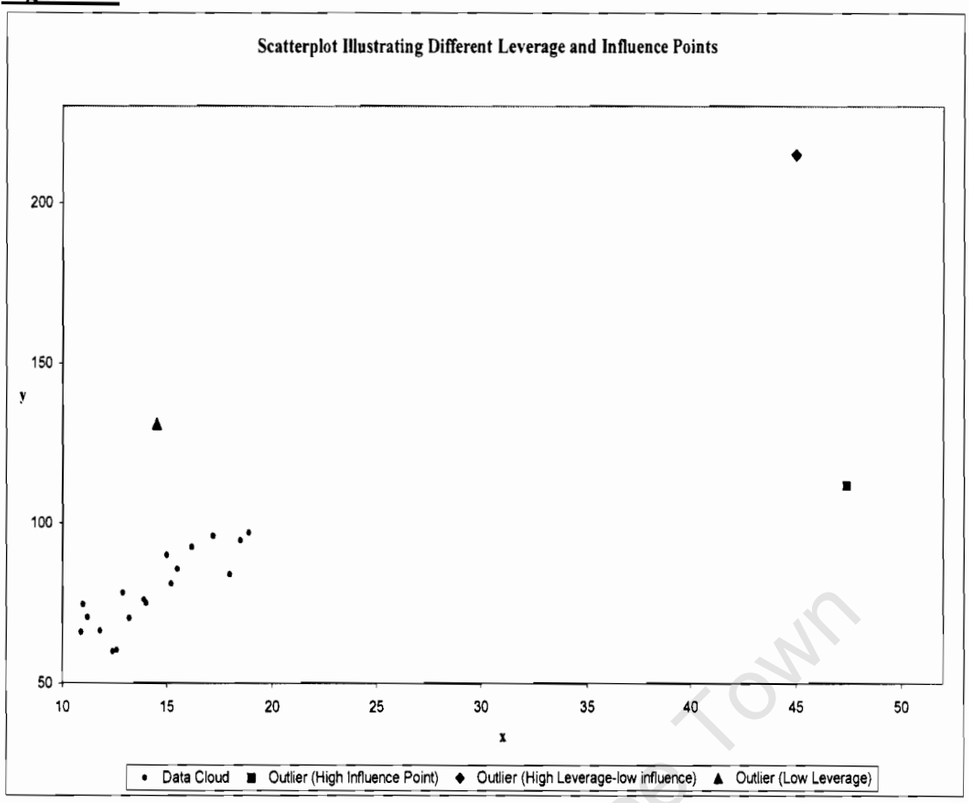
estimates the value of the $\hat{\beta}$ parameter. This minimisation is referred to as *Ordinary Least Squares* (OLS). Essentially this problem minimises the sum of the squared distances between the true value of the response variable, y_i, and their corresponding estimated values, denoted as $\hat{y}_i$, which are based on the regressor variables $\mathbf{x}_i'$. Simply stated, OLS minimizes the sum of squared residuals.

In the event that the error terms are truly iid normal, the OLS estimator has the smallest variance among the possible linear unbiased estimators. In the instance when the OLS has the smallest variance it can be referred to as the "best" linear unbiased estimator (referred to as BLUE). Another result due to the minimum variance property is that the maximum likelihood estimator (MLE) equals the OLS estimator. Outliers can cause the OLS not to attain the correct minimum variance in the model. In the following section we will be discussing outliers in general.

1.4 Outliers

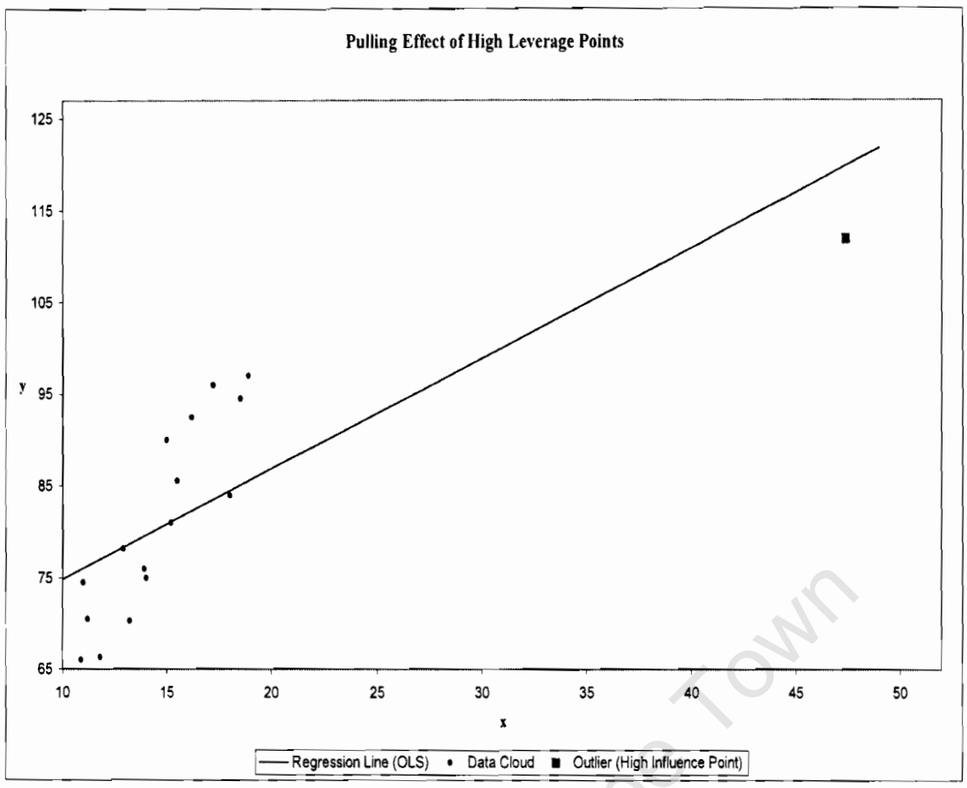
The term *outliers* refer to data points that are considered to be extreme in the response variable, relative to the trend of the general data. The term leverage describes the position of an observation in the regressor space. There are two types of leverage points that can be distinguished between. The first is known as a low leverage point, this point is positioned near the general data cloud. The second is known as a high leverage point and is positioned in an extreme location. Rousseeuw and van Zomeren (1990) distinguished even further between different high leverage points. A 'good leverage point" is an observation that possesses a large leverage in the regressor space, while with respect to the response variable it still fits the general data trend. A "bad leverage point" is a data point that possesses a large leverage in the regressor space and the response variable does not fit the general trend of the data. The differences between these points are illustrated in figure 1.1.

Figure 1.1



The OLS procedure has the drawback that outliers may have such a large residual that upon squaring, the objective function of the OLS is overwhelmed. This overwhelming of the objective function could have the effect that the regression may be pulled towards this one point. This pulling effect of the regression towards this one high leverage point is as a result of the reduction in this particular squared residual. This reduction in the squared residual more than offsets the increase in the residual values of the data points in the general data cloud. In figure 1.2 the pulling effect that such a high influence point can have on a regression line is illustrated. The same data is used as in figure 1.1 with the exceptions that the low leverage and high leverage-low influence points are ommited. Low leverage outliers on the other hand, tend to affect the intercepts of a regression line and as a result produce a fit that is not aligned with the general data trend.

Figure 1.2



It is clear that the OLS procedure lacks the ability to handle as little as one leverage point. This would imply that if an explanatory data analysis is not performed on the data, the OLS procedure can produce misleading coefficients, standard errors, predictions, hypothesis tests, and other numerical measures.

Chapter 2

Low Breakdown Regression Procedures

2.1 Introduction

Robust procedures were developed to minimize the impact that the presence or absence of an outlier would have on an estimation. In effect these procedures make the estimation resistant to the impact that outliers could have.

To achieve this, robust procedures bound the influence of outliers. The method in which the procedures bound the influence depends on the type of problem that they are addressing. Commonly there are three classes of outliers referred to in the regression framework:

1. The first case is where the outliers lie primarily in the direction of the y-axis, as can be seen in figure 1.1. M-estimators are primarily used in robust methods designed to deal with this class of problem.
2. The second case is where problems are encountered with a moderate percentage of multivariate outliers in the covariates space (i.e., outliers in the x-space or leverage points), as can be seen in figure 1.2. Bounded influence estimators are used for this class of problem.
3. The third case is where the frequency, with which outliers occur, with respect to the y- and x-component, tends to be very high. The frequency can be as high as 50%. For this class of problem high breakdown point estimators are used.

The first two cases are referred to as low breakdown point regression procedures. The third case makes use of high breakdown point estimators. The first two cases will be discussed in this chapter and the third in a separate chapter.

It is however firstly important to understand what is meant by a breakdown point. As is stated above, robust procedures were developed to minimize the impact that presence or absence of an outlier/outliers would have on an estimation technique. A low breakdown point simply means that the estimator is only able to handle a few outliers in the data before the estimator takes on arbitrarily large values.

Technically the definition of a breakdown point was first defined by Hodges in 1967 and was restricted to one-dimensional estimation of location. Hampel (1971) gave a much more general formulation; this formulation was unfortunately asymptotic and very mathematical in its nature which may have restricted its dissemination (Rousseeuw and Leroy 1987).

Rousseeuw and Leroy (1987), who cite Donaho and Huber (1983), introduce a finite-sample version of the breakdown point. Take any sample of n data points,

$$Z = \{(x_{11},\ldots,x_{1p},y_1),\ldots,(x_{n1},\ldots,x_{np},y_n)\}, \tag{2.1}$$

and let T be a regression estimator. Applying T to the sample Z would yield a vector regression coefficient:

$$T(Z) = \hat{\theta}. \tag{2.2}$$

If we consider all possible corrupt samples Z' that can be obtained by replacing m of the original data points by arbitrary values (allows for the insertion outliers). The maximum bias (denoted bias$(m;T,Z)$) that can be caused by such a contamination is defined as:

$$\text{bias}(m;T,Z) = \sup_{Z'} \|T(Z') - T(Z)\|, \tag{2.3}$$

where the supremum is over all possible Z'. If $\text{bias}(m;T,Z)$ is infinite this has the meaning that m outliers can have an arbitrarily large effect in T, which in essence means that the estimator "breaks down". Therefore the finite sample breakdown point $\text{BP}(T,Z)$ of an estimator T, given a sample set Z is defined as

$$\text{BP}(T, Z) = \min\left\{\frac{m}{n} = \text{bias}(m;T,Z) = \infty\right\}. \tag{2.4}$$

Simply stated it is the smallest fraction of contamination that can occur in the sample set before the estimator T takes on values arbitrarily far from $T(Z)$.

Further more it is important to note that the breakdown point of any estimator can never be higher than 50% simply because it would become impossible to distinguish between the "good" and "bad" observations in the sample set.

2.2 *M*-Estimators

As was shown in the previous chapter a single erroneous observation in the data set can lead to the complete breakdown of the OLS regression. Taking the original OLS objective function and rewriting it as

$$\min_{\forall b} \sum_{i=1}^{n} \rho(r_i), \tag{2.5}$$

where $\rho(t) = t^2$ represents the OLS argument. This choice of ρ leads to parameter estimates that are the best linear unbiased estimator (BLUE). This makes this choice optimal under classical regression assumptions of iid normal errors. Another consequence of this choice of ρ is that the parameter estimates are also the Maximum Likelihood Estimates (MLE's). These results are academic, as violations to these assumptions are usually found in real data sets (outliers often occur in real data).

Changing the choice of ρ to a symmetric function [i.e., $\rho(t) = \rho(-t)$ for all t] with a unique minimum at zero yields the expression

$$\min_{\forall b} \sum_{i=1}^{n} \rho\left(\frac{y_i - \mathbf{x}_i'\mathbf{b}}{\hat{\sigma}}\right), \tag{2.6}$$

where $\hat{\sigma}$ is an appropriate estimation of σ.

The function of the ρ-function is to downweight the effect that an outlier would have on an estimator. Differentiating this expression with respect to **b** yields the expression

$$\sum_{i=1}^{n} \psi\left(\frac{y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}}{\hat{\sigma}}\right)\mathbf{x}_i = \mathbf{0}, \tag{2.7}$$

where $\psi = \rho^{(1)}$ and $\hat{\boldsymbol{\beta}}$ is the solution for **b** and $\mathbf{x}_i$ is the row vector of explanatory variables of the *i*th case:

$$\begin{aligned} \mathbf{x}_i &= (x_{i1}, \ldots, x_{ip}) \\ \mathbf{0} &= (0, \ldots, 0). \end{aligned} \tag{2.8}$$

In effect a system of p nonlinear equations are formed by (2.8), which are not always easy to solve. In the first case the ρ-function is arbitrary and a choice must be made to the type of function that will be assigned to it. Huber proposed, on theoretical grounds,

$$\begin{aligned} \rho(r) &= c|r| - c^2/2 \qquad \text{for } |r| \geq c \\ \rho(r) &= r^2/2 \qquad\quad\ \text{for } |r| \leq c, \end{aligned}$$

which corresponds to

$$\psi(r) = \max\left[-c, \min(c, r)\right]. \tag{2.9}$$

This function satisfies the minimax asymptotic variance arguments.

The solution to this set of p nonlinear equations is not equivariant with respect to the y axis. The effect that this has is that the residuals have to be standardized by some estimation technique. Generally the estimation choices are limited to robust measures of scale. A frequently used estimation technique is the median absolute deviation (MAD), defined in the following way

$$\hat{\sigma} = c \cdot \underset{\forall i}{med}\left|r_i - \underset{\forall i}{med}\, r_i\right|. \tag{2.10}$$

Under normal errors, assigning the constant c, referred to as the tuning constant, the value of 1.4826 would make the MAD estimate a consistent estimator of σ. Several methods are available to solve the system of p nonlinear equations, the Newton-Raphson and Iterated Reweighted Least Squares method are two examples.

M-estimators with (2.9) have the advantage of not only being more efficient than L_1 regression (at a model with Gaussian errors) but also being robust with respect to outlying y_i. The breakdown point of this method is again however $1/n$. This is due to the effect that an outlying $\mathbf{x}_i$ has on the estimator. This problem lead to the introduction of generalized *M*-estimators (*GM*-estimators).

2.3 Bounded Influence Estimates

As stated in the previous section the fact that *m*-estimators have a breakdown point of $1/n$ lead to the introduction of generalized *M*-estimators. *GM*-estimators have the main purpose of bounding the influence that an outlying $\mathbf{x}_i$ will have on an estimator by means

of some weighting function $w_i = w(x_i)$. $w_i = w(x_i)$ Is a decreasing function which depends on the magnitude of the $\mathbf{x}_i$ (e.g., the distance from the specific data point to centre of the data cloud). In essence the weighting function assigns a weight to all of the data points $\mathbf{x}_i$. The assigned weight will determine how much emphasis a regression estimator puts on an observation, i.e. a good observation should carry weighting value of close to 1. An outlier or high influence point, on the other hand, should carry a weighting value of close to zero or even zero.

Mallows (1975) proposed the function

$$\sum_{i=1}^{n} w(\mathbf{x}_i)\psi\left(\frac{y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}}{\hat{\sigma}}\right)\mathbf{x}_i = \mathbf{0}, \tag{2.11}$$

to replace (2.7) whereas Schweppe proposed

$$\sum_{i=1}^{n} w(\mathbf{x}_i)\psi\left(\frac{y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}}{w(\mathbf{x}_i)\hat{\sigma}}\right)\mathbf{x}_i = \mathbf{0}, \tag{2.12}$$

as an alternative to (2.7). The main difference between Schweppe and Mallows estimators is that the Mallows estimator downweights an outlier regardless of that points' residual value, this can lead to a loss of efficiency in the estimator. Whereas the Schweppe estimator downweights leverage points only if its' corresponding residuals are large.

As stated previously, these estimators were constructed to bound the influence of a single outlier. The effect of such an outlier can be measured with the use of an influence function (Hampel 1974). Using these criteria there have been many suggestions through the years as to how to optimize the w and ψ functions (Hampel 1978, Krasker 1980, Krasker and Welsch 1982, Ronchetti and Rousseeuw 1985, Samarov 1985). This lead to the *GM*-estimators being referred to as bounded influence estimators (*BI* estimators). A

problem with *GM*-estimators is that its breakdown point has an upper bound that is not adequately high enough. Furthermore the value of the upper bound decreases as the number of regression coefficients p increase. This implies that the breakdown point is a decreasing function of p. This leads to the unsatisfactory realisation that as the dimension of the regression coefficients increases, the chances that outlier observations can occur is much greater and at the same time the estimators ability to handle these outliers decrease (as the breakdown point decreases accordingly).

The affect that choosing a w_i (referred to as the *BI* weight) has on the *BI*-estimator is that it determines which data structures the *BI*-estimator can handle. One example is when all the w_i's are chosen to be equal to 1, this reduces the estimator to an *M*-estimator. This causes the breakdown of the estimator when trying to handle outliers at high leverage locations. The "pull" effects that these high leverage location outliers have generally mean that they have small residual values. And as stated earlier the smaller the residual value of an outlier the less downweighting impact the estimator has on that point.

2.4 The Ψ-function

In both the *M* and *BI* regression procedures the sets of nonlinear equations that have to be solved rely on some Ψ-function. This Ψ-function dictates some of the robust properties of the estimator. There are many Ψ-functions that have been developed over the years. In Andrews et. al. (1972) Ψ-functions used in location estimations are studied.

The following two Ψ-functions (Huber and bisquare) are both highly rated in terms of their overall performance. They offer the choice of either a nondescending function or redescending function. A redescending function is one for which $\psi(x) = 0$ whenever $|x| \geq c$, this implies that large outliers will be excluded from the computation. A nondescending function is for which $\psi(x) = -c$ when $x < -c$ and $\psi(x) = c$ when $x > c$. The constant c is referred to as the tuning parameter, assigning different values to the

constant c leads to obtaining different levels of efficiency. The Huber Ψ-function is defined as

$$\psi(t) = \begin{cases} -c_H, & if \quad t < -c_H, \\ t, & if \quad |t| < c_H, \\ c_H, & if \quad t > c_H. \end{cases} \tag{2.13}$$

To attain 95% efficiency under normal errors for the location model, the tuning parameter $c_H = 1.345$. This Ψ-function is referred to as a nondescending function.

Another Ψ-function available is the bisquare weight function defined as

$$\psi(t) = \begin{cases} 0, & if \quad t < -c_B, \\ t\left(1-(t/c_B)^2\right)^2, & if \quad |t| < c_B, \\ 0, & if \quad t > c_B. \end{cases} \tag{2.14}$$

To obtain a 95% efficiency under normal errors for the location model the tuning parameter, c_B, is given the value of 4.685.

The values used above are all derived for the location model and as such are suitable for usage in M regression. For BI regression however the general preference is to incorporate the correction factor

$$c = \sqrt{2pn}/(n-2p).$$

This correction factor is introduced to account for the w_i that is present in the Ψ-function that is used to produce the BI estimator (Walker (1984), Birch and Agard (1993)). The Huber Ψ-functions is defined as

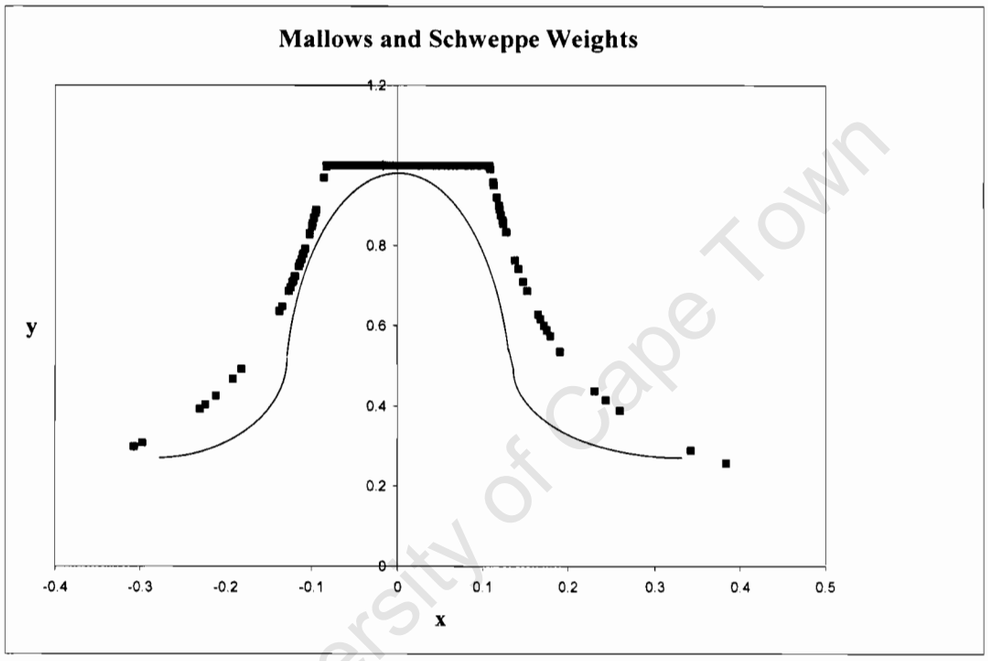
$$c_H = 1.345\sqrt{2pn}/(n-2p),$$

and the bisquare Ψ-function is defined as

$$c_B = 4.685\sqrt{2pn}/(n-2p).$$

In figure 2.1 below a graphical illustration of the difference between the Mallows and Schweppe weighting functions is given.

Figure 2.1 Mallows weights- dotted line, Schweppe weights- line



Chapter 3

Multivariate Location and Scale Estimation

3.1 Introduction

Before high breakdown point regression procedures can be discussed one must first be familiar with multivariate location and scale procedures. The scale estimator in the multivariate setting is called the dispersion matrix, although it is commonly referred to as the covariance matrix.

The dimension of the estimation can be split into two cases.

- The first case is when the n observations in the data are covering the k regressor variables and thus making the dimension of the estimation equal to k.
- The second case is when the n observations in the data are covering the k regressor variables plus the response variable and thus making the dimension of the estimation p.

In the rest of the chapter the assumption is made that the dimension of the estimation is p. Furthermore no intercept variable (a column of ones) is incorporated. Simply stated the data is contained in a $n \times p$ $\mathbf{Z}_y$ matrix. If so desired, the response variable can be easily removed from the analysis by simply replacing the p dimensions of the estimation by k dimensions through the use of only using the k regressor variables. Replace the p with k in the notation and substitute the $\mathbf{Z}_y$ matrix with the $\mathbf{Z}$ matrix.

3.2 Classical Estimation technique

When data is found in the multivariate data cloud, the job of identifying outliers becomes much more complicated than in the univariate case. This leads to the problem that the construction of robust techniques becomes more complicated.

Using the p dimensional data set with n data points

$$\begin{aligned} Z &= \{\mathbf{z}_1, \ldots, \mathbf{z}_n\} \\ &= \{(z_{11}, z_{12}, \ldots, z_{1p}), \ldots, (z_{n1}, z_{n2}, \ldots, z_{np})\}, \end{aligned}$$

we want to estimate the "centre" of this data "cloud". To achieve this we make use of a multivariate location estimator. The most widely known estimator of multivariate location is the $p \times 1$ arithmetic mean

$$T(Z) = \mathbf{m} = \frac{\sum_{i=1}^{n} \mathbf{z}_{y,i}}{n}, \tag{3.1}$$

which is the least squares estimator in this framework because of the fact that it minimizes $\sum_{i=1}^{n} \|\mathbf{z}_i - T\|^2$, where $\|\ldots\|$ is the ordinary Euclidean norm. Whereas the most widely known optimal estimator for dispersion is the $p \times p$ sample covariance matrix,

$$C = \frac{\sum_{i=1}^{n} (\mathbf{z}_{y,i} - \mathbf{m})(\mathbf{z}_{y,i} - \mathbf{m})'}{n-1}.$$

Because both of these estimators are mean based they are very sensitive to the effects that even a single bad outlier can have. The arithmetic mean possesses a breakdown point $1/n$. The breakdown point is often evaluated when $n > \infty$ therefore the multivariate

mean is considered to have a 0% breakdown. The 0% breakdown of the arithmetic mean led to the development of estimators which are outlier resistant.

3.3 Outlier Resistant Methods

In the following section some of the major outlier resistant methods available are listed and briefly discussed. There have been numerous alternative estimators developed over the years to replace the classical method (Lopuhaa 1992). Some of these estimators are computationally more extensive than others they may also differ with respect to some of their theoretical properties.

Furthermore they differ with respect to the affine equivariance property. The definition of affine equivariance is as follows.

An estimator T is said to be affine equivariant if it satisfies the following property

$$T\left(\left\{\left(\mathbf{x}_i\mathbf{A}, y_i\right); i=1,\ldots,n\right\}\right) = \mathbf{A}^{-1}T\left(\left\{\left(\mathbf{x}_i, y_i\right); i=1,\ldots,n\right\}\right), \tag{3.2}$$

for any non-singular square matrix **A.** This means that a linear transformation of the $\mathbf{x}_i's$ should transform the estimator T accordingly. This is because $\hat{y}_i = \mathbf{x}_i T = \left(\mathbf{x}_i\mathbf{A}\right)\left(\mathbf{A}^{-1}T\right)$. This property allows us to utilise a different coordinate system for the explanatory variables, without adversely affecting the estimated $\hat{y}_i$

3.3.1 Non Affine Equivariant Estimators

3.3.1.1 Coordinate Median

Simplicity led to the development of this method as it is based on the idea of evaluating each variable individually. For each of the j variables the numbers $x_{1j}, x_{2j}, \ldots, x_{nj}$ can be considered to be a one dimensional data set with n points. Using a univariate robust

estimator one can estimate a more resistant median for each of the p variables and combine this into a p-dimensional estimate. The dispersion matrix estimator becomes a covariance calculation that is centred by this coordinatewise median.

The definition of the coordinatewise medium is

$$\left(\underset{i}{med}\, x_{i1}, \underset{i}{med}\, x_{i2}, \ldots, \underset{i}{med}\, x_{ip}\right). \tag{3.3}$$

The coordinatewise medium poses a substantial problem in the multivariate setting, in that the position estimator does not necessarily lie in the general data cloud. This is stated in Rousseeuw and Leroy (1987, pg.250) as "This estimator is easily computed, but fails to satisfy some "natural" properties. For instance, it does not have to lie in the convex hull of the sample when $p \geq 3$." Along with the above problem this estimator is also not affine equivariant.

3.3.1.2 Hadi's Forward Search

As the name states Hadi (1992) developed this procedure. This procedure starts by defining the initial location estimator, $\mathbf{m}_0$, as the coordinatewise medium vector. The initial location estimator is then utilized in calculating the initial covariance estimator, $\mathbf{C}_0$, which is defined as

$$\mathbf{C}_0 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{z}_{y,i} - \mathbf{m}_0\right)\left(\mathbf{z}_{y,i} - \mathbf{m}_0\right)'. \tag{3.4}$$

Next a set of n robust Mahalanobis distances are created using these initial estimators, these Mahalanobis distances are defined as

$$RD_i = \sqrt{\left(\mathbf{z}_{y,i} - \mathbf{m}_0\right)'\mathbf{C}_0^{-1}\left(\mathbf{z}_{y,i} - \mathbf{m}_0\right)}. \tag{3.5}$$

Trimmed location and covariance estimators are now determined using the smallest h robust distances, where $h = \lfloor (n + p + 1)/2 \rfloor$. These estimators are in turn used to create a new set of n robust Mahalanobis distances. The observations are now split into two subsets, namely the basic and non-basic. The basic subset contains the $r = p + 1$ smallest robust distances observations. The remaining observations are then placed in the non-basic subset.

The basic subset is now used to determine the usual mean vector and covariance matrix, which in turn is used to determine a new set of robust distances. If the basic subset is of full rank then the robust distances are defined as

$$RD_i = \sqrt{(\mathbf{z}_{y,i} - \mathbf{m}_b)' \mathbf{V}_b \mathbf{W}_b \mathbf{V}_b' (\mathbf{z}_{y,i} - \mathbf{m}_b)},$$

where $\mathbf{V}_b$ is the matrix of normalized eigenvectors of $\mathbf{C}_b$ and $\mathbf{W}_b$ is a diagonal matrix whose j^{th} diagonal element is defined as

$$w_j = \frac{1}{\max\{\lambda_j, \lambda_s\}}, \quad j = 1,2,\ldots,n,$$

with λ_s being the smallest non-zero eigenvalue of $\mathbf{C}_b$. The observations are once again split into the basic and non-basic subsets with the exception that the number of observations placed into basic, r, subset is increased by one. As before the remaining observations are placed in the non-basic subset. This cycle is repeated until the basic subset contains $h = \lfloor (n + p + 1)/2 \rfloor$ observations. The final basic subset is then used to determine the final robust location and scale estimators.

Rocke and Woodruff (1996) stated that: "the algorithm ... breaks down if the contamination is extremely far away from the good data in the correct metric." Due to the use of the coordinatewise median this procedure is not affine equivariant.

3.3.2 Affine Equivariant Estimators

3.3.2.1 Convex Peeling

As the name hints, this procedure removes data points on the sample's convex hull. This is repeated until a sufficient number of points are removed (or peeled away). The remaining data points are then used in classical estimation techniques. A setback to this procedure is that it has a low breakdown point. This is due to the fact that each peeling step removes at least $p+1$ points from the sample, even though only one of these points might be considered to be an outlier. The peeling steps therefore reduce the number of good points in the data rapidly, which leads to the low breakdown point.

3.3.2.2 Transformed One-step Weighted dispersion Estimator

Ruiz-Gazen (1996) developed this robust dispersion estimator. An advantage to this method is that it is computationally inexpensive. The first step entails the calculation of an intermediate covariance matrix by a pre-determined robust location estimator $\hat{\boldsymbol{\mu}}$. The covariance matrix is defined as follows

$$\mathbf{V} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{z}_{y,i} - \hat{\boldsymbol{\mu}})(\mathbf{z}_{y,i} - \hat{\boldsymbol{\mu}})' . \tag{3.6}$$

The next step is to calculate a one-step covariance matrix using a kernel function that obtains the observations weights. The weighting function (with arguments that are proportional to the robust distances based on $(\hat{\boldsymbol{\mu}}, \mathbf{V})$) produces individual observation weights, defined as

$$w_i = K\left[\beta(\mathbf{z}_{y,i} - \hat{\boldsymbol{\mu}})'\mathbf{V}^{-1}(\mathbf{z}_{y,i} - \hat{\boldsymbol{\mu}})\right],$$

with K being a positive and decreasing function and β any non-negative scalar. These weights are used to obtain the one-step covariance matrix defined as

$$\mathbf{C} = \frac{\sum_{i=1}^{n} w_i \left(\mathbf{z}_{y,i} - \hat{\boldsymbol{\mu}}\right)\left(\mathbf{z}_{y,i} - \hat{\boldsymbol{\mu}}\right)'}{\sum_{i=1}^{n} w_i - 1}. \tag{3.7}$$

The final step consists of calculating the robust dispersion matrix via a transformation to attain a consistent estimator. The matrix is defined as

$$\mathbf{U} = \left(\mathbf{C}^{-1} - \beta \mathbf{V}^{-1}\right)^{-1}.$$

It is noted that in order to obtain $\mathbf{U}$ one must pre-define the initial location estimator $\hat{\boldsymbol{\mu}}$ (the coordinatewise median is recommended), the weighting function K (where $K(x) - \exp\left(\frac{-x}{2}\right)$) and the scalar β (where $\beta = \frac{26\varepsilon}{p}$). The ε represents the percentage of contamination found in the data set. It is further suggested that the choice of β plays a major role in whether the outlier observations are detected or not. The breakdown point of this procedure is suggested to be around 20%.

3.3.3 Affine Equivariant Estimators with High Breakdown Pionts

3.3.3.1 Stahel-Donoho Estimator

This first affine equivariant multivariate location estimator with a breakdown point of 50% was developed independently by Stahel (1981) and Donaho (1982). This estimator is referred to as the *outlyingness-weighted* mean. Rousseeuw and Leroy define this estimator as follows: For each observation $\mathbf{x}_i$, one looks for a one-dimensional projection that leaves it most exposed. Simply stated, the idea behind this estimator is that an outlier or high leverage point will be separated from the centre of the data cloud if viewed from

the right perspective. The robust multivariate location estimator is estimated in two stages. Firstly robust distances, u_i, are determined via a projection computation to determine the "outlyingness" of $\mathbf{x}_i$. With

$$u_i = \sup_{\|\mathbf{v}\|=1} \frac{\left|\mathbf{x}_i \mathbf{v}^t - \underset{j}{\text{med}}(\mathbf{x}_j \mathbf{v}^t)\right|}{\underset{k}{\text{med}}\left|\mathbf{x}_k \mathbf{v}^t - \underset{j}{\text{med}}(\mathbf{x}_j \mathbf{v}^t)\right|}, \tag{3.8}$$

where $\text{med}^j\left(\mathbf{x}_j \mathbf{v}^t\right)$ is the median of the projections of the data points $\mathbf{x}_i$ on the direction of the vector $\mathbf{v}$, and the denominator is the median absolute deviation of these projection.

These distances are then used as the arguments in the weight function

$$T(X) = \frac{\sum_{i=1}^{n} w(u_i)\mathbf{x}_i}{\sum_{i=1}^{n} w(u_i)}, \tag{3.9}$$

where $w(u)$ is a strictly positive and decreasing function of $u \geq 0$, such that $w(u)$ is bounded. This weighting function is then utilised to calculate the weighted mean vector and weighted covariance matrix. When $n > 2p+1$ and the data is in a general position, the estimator attains a breakdown point of 50%.

Rousseeuw and Zomeren proposed a shortcut to circumvent the computationally expensive problem that arises from taking the supremum over all possible directional vectors, in the original Stahel-Donaho estimator. This shortcut entails using only n directional vectors, one vector in the direction of each centred observation. The observations are centred by the coordinatewise median vector, starting from the origin. The robust distances are determined by these n directional vectors.

In general the robust distances are usually calculated after the robust location and dispersion estimators are determined. The Stahel-Donaho estimator the order of the calculations is done in reverse. Cook and Hawkins (1990) made the comment that this projection method can produce "outliers everywhere".

3.3.3.2 Minimum Volume Ellipsoid Estimator

Rousseeuw (1983, 1984) introduced this affine equivariant estimator by putting

$$T(X) = \text{centre of the minimal volume ellipsoid covering, at least, } h \text{ points of } X,$$

where $h = \lfloor n/2 \rfloor + 1$. This estimator was called the minimum volume ellipsoid estimator (MVE). Using this ellipsoid the $p \times 1$ location vector, **m** (also referred to as the $\mathbf{MVE}_1$), and the $p \times p$ covariance matrix, **C** (also referred to as the $\mathbf{MVE}_2$). These estimators are used as a robust benchmark to determine which observations are outliers and/or high leverage points. The problem with these estimators is that there is no closed form solution for obtaining the MVE estimators.

The first algorithm was proposed by Rousseeuw and Leroy (1987) in which, firstly, a large number of elemental subsets are drawn. The number of random subsets M that are needed is found by equating the probability that at least one of the M elemental sets will contain only good points to a value (near 1) that is acceptable. The value of 0.999 is often cited in the literature (ex. Rousseeuw and Leroy (1987)). The probability that at least one out of the M subsamples consist exclusively of "good" data points is

$$1 - \left(1 - (1 - \varepsilon)^{p+1}\right)^M,$$

where ε is the percentage of contamination in the data. If this probability is set equal to the value of 0.999 the following equation is obtained

$$0.999 = 1 - \left(1 - \left(1 - \varepsilon\right)^{p+1}\right)^{M}.$$

If $\varepsilon = 0.50$, the number of subsamples needed to start the process is

$$M = \frac{\ln(0.001)}{\ln\left(1 - (0.5)^{p+1}\right)}. \tag{3.10}$$

This step is often ignored and the researcher may choose to select an M equal to 500 or 1000 for convenience. The probabilistic argument guarantees the selection of at least one subset that contains only good observations with a "high probability". This subset of "good data" does is not necessarily represent the data accurately, which leads to the problem that a representative subset may not be obtained. This problem leads to the probabilistic argument underestimating the required number of subsets needed for this procedure.

We start by randomly selecting $p+1$ observations. These points will be indexed by $J = \{i_1, \ldots, i_{p+1}\}$. This number represents one more than the dimension of the estimation problem. This number ensures that a full rank covariance matrix can be estimated. There is more than one rational behind the selection of $p+1$ observations namely that it ensures that at least $p+1$ observations fall exactly on the boundary of the ellipsoid defined by the MVE.

The $p \times 1$ mean vector and the $p \times p$ covariance matrix for this sample are now calculated. Where the mean vector is defined by

$$\bar{\mathbf{z}} = \frac{1}{p+1} \sum_{i \in J} \mathbf{z}_{y,i} \tag{3.11}$$

and the covariance matrix by

$$\mathbf{C}_J = \frac{1}{p}\sum_{i \in J}(\mathbf{z}_{y,i} - \bar{\mathbf{z}}_J)(\mathbf{z}_{y,i} - \bar{\mathbf{z}}_J)', \tag{3.12}$$

where $\mathbf{C}_J$ is non-singular, whenever $\mathbf{z}_{i_1},\ldots,\mathbf{z}_{i_{p+1}}$ are in general positions. Next we calculate the set of m robust squared Mahanalobis distances, RD_i^2, and find the median, m_J^2, of them such that

$$m_J^2 = \underset{\forall i}{\text{med}}\, RD_i^2 = \underset{\forall i}{\text{med}}(\mathbf{z}_{y,i} - \bar{\mathbf{z}}_J)'\mathbf{C}_J^{-1}(\mathbf{z}_{y,i} - \bar{\mathbf{z}}_J).$$

We now have to find the minimum volume of all the ellipsoids formed by the each of the M subsets. The volume of the resulting ellipsoid, corresponding to $m_J^2\mathbf{C}_J$ is proportional, so that

$$\text{Volume} \propto (\det(\mathbf{C}_J))^{1/2}(m_J)p. \tag{3.13}$$

Thus, after each volume has been determined, if the volume is smaller than the previous volume determined, the current subset is stored as the current best subset. After all the subsets have been drawn and tested, the MVE estimators are based on the best subset found from the analyses and calculated as

$$\mathbf{m} = \bar{\mathbf{z}}_J$$

The location vector, **m**, is simply the average coordinate vector of the best elemental set and

$$\mathbf{C} = (\chi^2_{p,0.50})^{-1} m_J^2 \mathbf{C}_J,$$

where the covariance matrix of the best elemental set is multiplied by $\left(\chi^2_{p,0.50}\right)^{-1} m_J^2$ to expand (or shrink) the ellipsoid and ensure that half the data is contained, forming **C**.

Rousseeuw and Leroy (1987) further employed a one-step improvement to the MVE estimators. Firstly the Mahalanobis distance for each observation is calculated in the usual manner where the Mahalanobis distance is defined as

$$RD_i^2 = \left(\mathbf{z}_{y,i} - \mathbf{m}\right)' \mathbf{C}^{-1} \left(\mathbf{z}_{y,i} - \mathbf{m}\right), \quad i = 1,2,...,n. \tag{3.14}$$

A binary weighting scheme is now used to give weighting values to the distances, with the largest RD_i^2 values getting a zero weighting value due to their extremity in the regressor space. Small RD_i^2 values are an indication of a good observation and therefore receive a maximum weight. By interpreting the robust Mahalanobis distances as having a chi-square distribution with p degrees of freedom, the 0.975 quantile of this distribution becomes the critical value:

$$w_i = \begin{cases} 1, & if \quad RD_i^2 \le \chi^2_{p,0.975}, \\ 0, & otherwise. \end{cases}$$

The one-step improvement MVE estimators becomes

$$\mathbf{m} = \frac{\sum_{i=1}^{n} w_i \mathbf{z}_{y,i}}{\sum_{i=1}^{n} w_i} \tag{3.15}$$

and

$$\mathbf{C} = \frac{\sum_{i=1}^{n} w_i \left(\mathbf{z}_{y,i} - \mathbf{m}\right)\left(\mathbf{z}_{y,i} - \mathbf{m}\right)'}{\sum_{i=1}^{n} w_i - 1}. \tag{3.16}$$

It is worth noting that the results obtained from this random subsampling algorithm are only approximate values of the MVE. This approximation occurs since the exact MVE does not necessarily have to be defined by the sample mean or sample covariance matrix of some subset. The computation of the MVE is computationally very intensive. This is as a result of there being no closed form solution for the MVE.

Cook, Hawkins and Weisberg (1993) developed an exact iterative computational algorithm for the MVE. This algorithm requires that all C_h^n subsets or halfsets, as they contain h observations, are considered and for which the covariance matrix and its determinant are computed. As the algorithm proceeds, certain subsets lead to an iteration routine that allows the overall algorithm to converge to the exact solution for the MVE objective function. This approach is computationally very intensive even for a moderate sample size.

Hawkins (1993) developed the feasible solution algorithm (FSA) in order to circumvent this exhaustive evaluation of every possible halfset. This procedure uses N random starting halfsets. Each one of the h observation are at the outset (at iteration stage $k = 0$) assigned an equal weight, or $w_i^{(0)} = h^{-1}$. The weighted mean vector and associated covariance matrix is calculated. The robust Mahalanobis distances for each observation is determined (denoted by $D_i\left(w^{(k)}\right)$) and the maximum such distance is denoted by $D_{\max}\left(w^{(k)}\right)$. Now, if $D_{\max}\left(w^{(k)}\right) < p + \varepsilon$, where ε is a convergence bound, the iteration process stops. Otherwise the weights for the next iteration step, $k + 1$, is updated with the function

$$w_i^{(k+1)} = \frac{w_i^{(k)} D_i\left(w^{(k)}\right)}{p},$$

and we repeat the process starting with the calculation of the weighted mean and associated covariance matrix. This procedure is repeated for all the subsets with h observations. The smallest $D_{\max}\left(w^{(k)}\right)$ of the subsets are referred to as the feasible solutions.

Once again we note that the volume of an ellipsoid is proportional to the determinant of the weighted covariance matrix, the exact MVE estimators are the weighted mean vector and associated weighted covariance matrix which corresponds to the subset yielding the smallest such determinant.

3.3.3.3 Minimum Covariance Determinant Estimator

Rousseeuw (1983, 1984) proposed to generalise the least trimmed squares estimator to the multivariate location. This yielded

$T(X)$ = Mean of the h points of X for which the determinant of the covariance matrix is minimal.

As the name indicates the location and dispersion estimation is based on the determinant of the covariance matrix. It has been shown in the literature (Morrison (1990), Johnson and Wichern (1992)) that if a covariance matrix is a $n \times n$ symmetric positive definite matrix, then the p eigenvalues are positive. If a linear relationship (correlation) between the p variables exists then the eigenvalues that are obtained are close to zero. Now, the determinant of a covariance matrix can be determined by the product of the p eigenvalues. Therefore, if a matrix has a close to zero determinant value would indicate that some linear pattern exists between the variables.

As with the MVE, consider all C_h^n subsets, next compute the determinant of the covariance matrix for each subset. The subset with the smallest determinant is then used

to calculate the usual $p \times 1$ mean vector, $\mathbf{m}$ (or $\mathbf{MCD}_1$), and the corresponding $p \times p$ covariance matrix, $\mathbf{C}$ (or $\mathbf{MCD}_2$). These estimators are called the minimum covariance matrix determinant estimators (MCD's). It has been illustrated, by Hawkins (1994), that the MCD's are in fact the maximum likelihood estimators when h observations are from the correct multivariate normal distribution. The rest of the $n-h$ observations are from a different multivariate distributed, one in which the mean has been shifted. The MCD estimators are once again a robust benchmark for outlier and/or high leverage point detection.

As with the MVE algorithm, the computational processes in determining the MCD's are once again intensive, since by definition the exact MCD solution is found when all C_h^n subsets are analysed. This differs from the MVE algorithm in that at this stage iteration of the observation weights are still required.

Hawkins (1994), as in the MVE case, developed a feasible solution algorithm, which once again allows for a reduction in the computation process by once again using N random starts. Hawkins (1994) offers a probabilistic argument which is based on a simulation analysis, to hypothesis that only a small number of random subsets are required in order to obtain the exact MCD estimators with a high probability. The particular data structure along with the number of local minima that might be reached still dictates what the value of N might be. This would imply that the value of N should still be as large as realistically possible. It is also worth noting that Hawkins and Olive (1999) offer improved FSA algorithms for both the MCD and the MVE.

Both the MVE and MCD are affine equivariant. Both have a $[(n-p-1)/2]/n$ breakdown point, which is 50% asymptotically. They differ however in their rate of convergence, which is why the MCD is preferred to the MVE in terms of their asymptotic properties. The MCD reaches convergence at a rate of $n^{-1/2}$ while the MVE reaches convergence at a rate of $n^{-1/3}$ (Rousseeuw and Leroy (1987), Davies (1992)). Thus the MCD has a higher efficiency than the MVE does. Consistency for the MCD location and

dispersion estimator along with asymptotic normality was proven by Butler, Davies and Jhun (1993).

Chapter 4

High Breakdown Regression Estimators

4.1 Introduction

In the following sections we will discuss the first two high breakdown regression estimators, developed by Rousseeuw (1993), as well as two one-step improvements on two of the generalised *M*-estimators (Schweppe 1-step estimator and Mallows 1-step estimator). We next discuss regression procedures with high breakdown points. These procedures were developed as a result of the lack of ability of the *M* and *BI* estimators in handling large numbers of outliers or points of high leverage. Clusters of bad data may corrupt the methods and lead to poor estimators. High breakdown regression procedures have the ability to sift through data with large numbers of outliers and/or high leverage points (up to 50% for some procedures).

4.2 Least Median of Square

All of the regression methods that have been discussed up to now involved an Σ in the objective function. Rousseeuw (1984) proposed replacing the sum by a median. This change was essentially based on an idea by Hampell (1975, pg 380). He postulated that replacing the summation with a median would yield better results due to its more robust nature than the summation function. This yielded the least median of squares (LMS) estimator, with the objective function

$$\min_{\forall \mathbf{b}} \ \underset{i}{\text{med}}(y_i - \mathbf{x}_i'\mathbf{b})^2 \tag{4.1}$$

The resulting estimator proved to be very robust with respect to outliers in y as well as outliers in $\mathbf{x}$ (or high leverage points). As it turns out LMS has a 50% breakdown point, the highest breakdown point attainable. A setback to this estimator is that it converges at

an abnormally slow rate of $n^{-1/3}$, making its asymptotic efficiency against normal errors zero. Originally no closed form solution was attainable to calculate LMS in the regression setting, degenerate cases excluded. Stromberg (1993) provided an exact algorithm to calculate the LMS. The drawback to this was that his algorithm required that all C_{p+1}^{n} subsets be investigated. As this proves to be an almost impossible task, the general approach is to approximate LMS by one of the available subsampling algorithms.

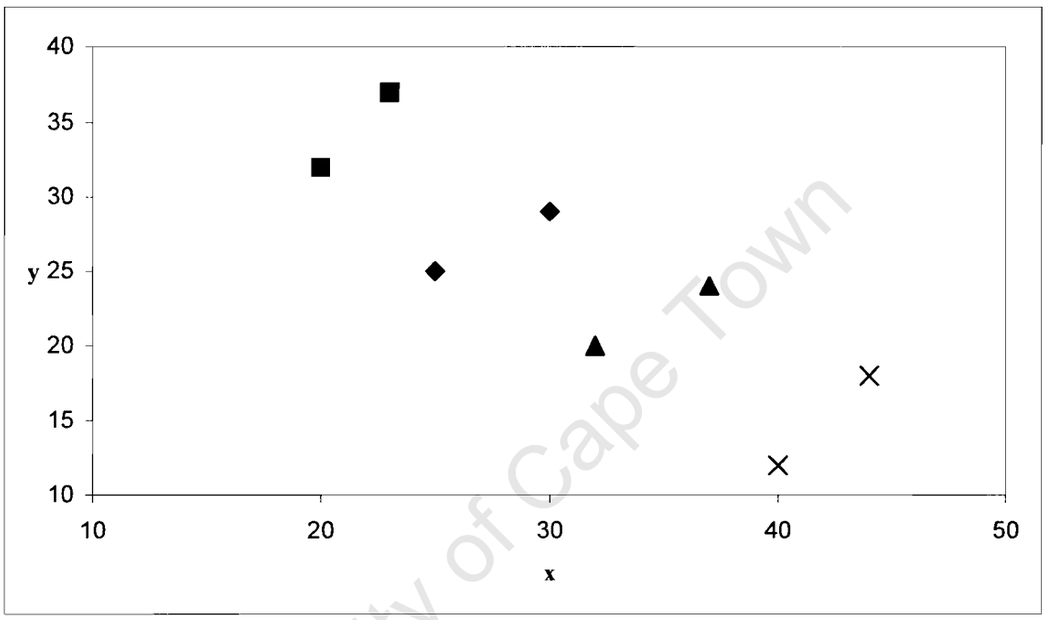
If a subsample of size p is considered, with p being the number of unknown coefficients. The subsample is referred to as the elemental set. This is due to the fact that assuming the reduced **X** matrix is of full rank, an exact fit can be obtained from these points and thus the objective function can be evaluated. The reduced matrix is the matrix formed by using only the p rows of **X** that correspond to the p subsampled observations.

One such subsampling algorithm, proposed by Rousseeuw and Leroy (1984), repeatedly drawing elemental sets and evaluating the objective function. The LMS estimate that is finally used corresponds to the estimate that had the smallest observed objective function. Rousseeuw and Basset (1991) showed that this estimate retains the high breakdown and convergence properties of the theoretical LMS. As with the MVE estimation, even if all C_{p}^{n} elemental sets are evaluated, the resulting LMS estimator obtained is generally not the true LMS estimator, but rather an estimation of the estimator. Furthermore, the calculations involved in obtaining the estimated regression coefficients are of the order n^{p+1} (Rousseeuw, 1993), which further gives an indication of the amount computation needed to determine the LMS estimator. This computational problem can be avoided by basing the selection criteria for the number of elemental subsets chosen on the probabilistic argument of how likely it is to obtain an elemental set of only good observations. This poses a problem in itself, in that, since the probability of obtaining an elemental set containing only good observations is less than 1, this could lead to a breakdown in the algorithm when calculating the high breakdown estimator. This seems to simply substitute one problem for another. Anyhow, the number of

elemental subsets needed is more or less $3 \cdot 2^p$ and $4.6 \cdot 2^p$ for 95% and 99% probabilities respectively, in order to obtain at least one good elemental subset.

Even though the probabilistic argument provides for the analysis of a strictly good elemental subset with high probability, there is no guarantee that the obtained subset reflects the general trend of the data. This in turn could lead to very misleading estimators.

Figure 4.1



Rousseeuw (1993) introduced an algorithm that reduced the order of computations required and eliminate the problem of algorithm breakdown as a result of the probability being less than 1 of obtaining a purely good elemental subset. Simply stated the data is randomly assigned into $2p-2$ sized blocks. Unassigned points are distributed as evenly as possible between the blocks. Within each the block, all possible subsets of size p are evaluated with the objective function. Once again the final LMS estimate corresponds to the estimate that had the smallest observed objective function. The advantage to this algorithm is theoretical, in practise it achieves a high breakdown point but the estimates

can be very misleading. This will be illustrated with the following example. Suppose a simple linear regression (SLR) model is posed for a data set with eight good observations. Applying the algorithm, the blocks contain $2(2)-2=2$ observations and there are 8/2 = 4 blocks. Suppose that figure 4.1 shows the data, randomly assigned into different blocks. Clearly the general slope of the data is negative, but the data in the blocks all produce have a positive slope estimates. The algorithm relies on asymptotic combinatory probabilities to overcome this problem, i.e. when *n* gets large the probability of obtaining such a block tends to zero.

A further, though modest, improvement on the LMS estimator obtained from the random subsampling algorithm can be achieved by adjusting the intercept. By viewing residuals as a univariate sample, the exact LMS algorithm for location estimation can be performed. The updated intercept is found as the "location LMS" of the residuals from the current LMS regression estimator. This procedure is guaranteed to reduce (not change though) the LMS objective function. Another positive result from the exact fit is that it also eliminates the condition of always having *p* zero residuals.

4.3 Least Trimmed Squares

Rousseeuw (1984) introduced Least Trimmed Squares to remedy the slow convergence rate $n^{-1/3}$ that hampers LMS. The LTS also possesses the maximum high breakdown point of 50%, but converges at a rate of $n^{-1/2}$. The objective function here is

$$\min_{\forall \mathbf{b}} \sum_{i=1}^{h} r_{[i]}^2, \tag{4.2}$$

this represents the sum of the smallest *h* smallest squared residuals. *h,* as before, is taken to be $[(n+p+1)/2]$. As with the LMS, no closed form solution exists (except for the location model) to construct the true LTS estimator. The algorithm used to approximate

the LMS can again be used to approximate the LTS, by simply changing the objective function being evaluated for each of the elemental sets.

The intercept adjustment discussed in the previous section, can again be utilised to further improve the LTS estimator. The "location LTS" estimate of the residuals from the current LTS regression estimator simply replaces the current intercept. Further discussions on the LTS algorithms are offered by Agullo (2001).

Asymptotically, both the LMS and LTS are inefficient estimators, with 1.39% and 7.14% asymptotic efficiency, respectively, versus the sample mean under normal errors. This lack of efficiency led to the development of one-step estimators. These estimators use LTS, due to its more rapid convergence rate, as an initial estimator and then do some improvement calculations. The idea is that the 1-step estimators inherit the high-breakdown properties of LTS and improve their lack of efficiency. Once again this leads to more problems, in that, the improvement step generally will require weights for the observations. These weights have to be robust weights in order to retain the high breakdown properties. This leads us back to popular choice amongst robust weighting schemes, the MVE-based Mallows weights.

In the remainder of this chapter we will be discussing two competing one-step estimators along with an overall high breakdown regression analysis of the stackloss case study.

4.4 1-Step Generalized *M*-Estimators

As a result of the poor efficiency coupled with a numerical sensitivity arising from both the random subsampling procedure and small internal movement of the data, found in both the LTS and LMS, other high breakdown regression techniques have been developed. In the following sections we will be discussing two such high breakdown regression techniques. Both fall under the 1-step generalized *M*-estimators. The one-step generalized *M*-estimator incorporates a high breakdown initial estimator combined with

the generalized *M*-estimator. This combination is utilised to remedy the problem of poor efficiency. The objective function in question has the same format as previously with the difference that the weighting function has been replaced by robust leverage weights. The solution to this problem is also solved through a different method, where previously it was solved via the IRLS, it is now solved using a one step Taylor series expansion of the objective function. We are able to write this estimator as the sum of two terms. Firstly we have the initial estimator, where the LTS has become the estimator of choice as a result of it converging more rapidly than the LMS. The second part consists of the one-step improvement calculations. In short this combination is beneficial because the *GM*-estimators inherit the high breakdown point of the LTS, while improving on the efficiency aspect.

The following points are applicable to both the Mallows 1-step estimator and the Schweppe 1-step estimator and are worth noting.

- as stated earlier, the initial estimator, $\hat{\boldsymbol{\beta}}_0$, is LTS,
- the residuals from the initial fit are denoted by $r_i\left(\hat{\boldsymbol{\beta}}_0\right) = y_i - \mathbf{x}_y^{'}\hat{\boldsymbol{\beta}}_0$,
- a diagonal matrix $\mathbf{W} = diag\left(w_i\left(\mathbf{x}_i\right)\right)$ of robust Mallows weights, with

$$w_i = \min\left[1, \left[\frac{\chi^2_{0.95,p-1}}{RD_i^2}\right]\right],$$

  is calculated using MVE estimates (based solely on the regressor space),
- The robust scale estimate, $\hat{\sigma}_0$, is based on the LMS estimate (Rousseeuw and Leroy, 1987, pg.202), and is found as

$$\hat{\sigma}_0 = 1.4826 \cdot \left(1 + \frac{5}{n-p}\right) \underset{\forall i}{med} \left|r_i\left(\hat{\boldsymbol{\beta}}_0\right)\right|$$

4.4.1 Mallows-1-Step Estimator

The first one-step generalized estimator under discussion, the Mallows-1-step (M1S) estimator, was introduced by Simpson, Ruppert and Carrol (1992). The 1-step improvement focuses on incorporating a leverage and outlier control term in its estimation of the $\hat{\beta}$ term. The M1S estimator is the solution to the altered normal equations

$$\sum_{i=1}^{n} w_i \psi\left(\frac{r_i\left(\hat{\beta}_0\right)}{\hat{\sigma}_0}\right)\mathbf{x}_i = \mathbf{0}. \tag{4.3}$$

The occurrence of outliers is counteracted by the ψ-function (the Huber ψ-function is used in this discussion), which is used to downweight large scaled residuals. High leverage points are downweighted as a result of the w_i's being robust Mallows weights. The altered normal equations are solved using the Newton-Rapson method, giving the estimator the form

$$\hat{\beta} = \hat{\beta}_0 + \mathbf{H}_0^{-1}\mathbf{g}_0,$$

where

$$\mathbf{g}_0 = \hat{\sigma}_0 \sum_{i=1}^{n} \psi\left(\frac{r_i\left(\hat{\beta}_0\right)}{\hat{\sigma}_0}\right) w_i \mathbf{x}_i$$

and

$$\mathbf{H}_0 = \sum_{i=1}^{n} w_i \mathbf{x}_i \mathbf{x}_i' \psi^{(1)}\left(\frac{r_i\left(\hat{\beta}_0\right)}{\hat{\sigma}_0}\right),$$

with $\psi^{(1)}$ being the first derivative of ψ. Rewriting this in matrix notation form simplifies the notation of the estimator into

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} + \left(\mathbf{X}'\mathbf{B}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}\psi\hat{\sigma}\ , \tag{4.4}$$

with ψ being a $n\times 1$ vector, $\mathbf{W}$ is the $n\times n$ weighting function defined in the previous section and $\mathbf{B}$ is the $n\times n$ diagonal matrix such that

$$\mathbf{B} = diag\left(w_i\psi^{(1)}\left(\frac{r_i\left(\hat{\beta}_0\right)}{\hat{\sigma}_0} \right) \right). \tag{4.5}$$

The diagonal elements of $\mathbf{B}$, due to the Huber ψ-function, can be defined as follows

$$b_{ii} = \begin{cases} w_i, & if \ \left| r_i\left(\hat{\beta}_0\right) \right| \le c\hat{\sigma}_0, \\ 0, & otherwise. \end{cases}$$

The Ψ vector elements are in turn calculated as

$$\psi_i = \begin{cases} -c, & if \ r_i\left(\hat{\beta}_0\right) < -c\hat{\sigma}_0, \\ \dfrac{r_i\left(\hat{\beta}_0\right)}{\hat{\sigma}_0}, & if \ \left| r_i\left(\hat{\beta}_0\right) \right| \le c\hat{\sigma}_0, \\ c, & if \ r_i\left(\hat{\beta}_0\right) > c\hat{\sigma}_0. \end{cases}$$

Next we need to calculate standard error terms for each of the coefficients to further the analysis beyond estimation only. Let the $p\times p$ matrix $\mathbf{M}_0$ be defined as

$$\mathbf{M}_0 = \hat{\sigma}_0^2 \sum_{i=1}^{n} w_i^2 \mathbf{x}_i \mathbf{x}_i' \psi^2 \left(\frac{r_i\left(\hat{\beta}_0\right)}{\hat{\sigma}_0} \right),$$

this leads to the estimated (asymptotic) covariance matrix for the parameter estimation having the form

$$Cov\left(\hat{\beta}\right) = \mathbf{H}_0^{-1} \mathbf{M}_0 \mathbf{H}_0^{-1}.$$

Now, if we define the matrix $\mathbf{V}$ as $\mathbf{V} = diag\left(w_i \psi \left(\frac{r_i\left(\hat{\beta}_0\right)}{\hat{\sigma}_0} \right) \right)$, the estimated covariance matrix can be written in the following matrix notation;

$$Cov\left(\hat{\beta}\right) = \hat{\sigma}_0^2 \left(\mathbf{X}'\mathbf{B}\mathbf{X}\right)^{-1} \left(\mathbf{X}'\mathbf{V}^2\mathbf{X}\right) \left(\mathbf{X}'\mathbf{B}\mathbf{X}\right)^{-1}. \tag{4.6}$$

The standard errors for the M1S coefficients are therefore determined by the square root of the diagonal elements of this estimated covariance matrix.

4.4.2 Schweppe-1-Step Estimator

We now take a look at another generalized *M*-estimator namely the Schweppe-1-step (S1S) that was introduced by Coakley and Hettmansperger (1993). Where the Mallows-1-Step estimator focused on controlling the affects of leverage points and outliers in the estimation of the $\hat{\beta}$ term, the Schweppe-1-Step estimator focuses on the selection of an appropriate weighting function. The altered normal equation of the S1S estimator differ from the M1S in that there is an added weight in the denominator of the ψ-function. Thus, the S1S estimator is the solution to the altered normal equations of the form

$$\sum_{i=1}^{n} w_i \psi\left(\frac{r_i\left(\hat{\beta}_0\right)}{\hat{\sigma}_0}\right)\mathbf{x}_i = \mathbf{0}\,. \tag{4.7}$$

As with the M1S estimator, these equations are of the same form as those for the BI regression. As with M1S the difference in the equations is in the definition of the weights, which are again the Mallows weights used in M1S instead of the hat diagonal-based Welch weights used in BI regression.

To solve this set of altered normal equations a Gauss-Newton approximation, using a first-order Taylor series expansion about the initial estimate $\hat{\beta}_0$, is used, which yields a 1-step improvement of the form

$$\hat{\beta} = \hat{\beta} + \left(\mathbf{X'BX}\right)^{-1}\mathbf{X'W}\psi\hat{\sigma}\ . \tag{4.8}$$

This is similar to the one-step improvement found in the M1S method, and uses the predefined weight matrix. The definition of the **B** matrix and ψ vector differ however. **B** is defined as follows

$$\mathbf{B} = diag\left(w_i\psi^{(1)}\left(\frac{r_i\left(\hat{\beta}_0\right)}{\hat{\sigma}_0 w_i}\right)\right),$$

where the diagonal elements of **B** are once again

$$b_{ii} = \begin{cases} w_i, & if\ \left|r_i\left(\hat{\beta}_0\right)\right| \le c\hat{\sigma}_0, \\ 0, & otherwise, \end{cases}$$

with the use of the Huber Ψ-function. The Ψ vector elements are in turn calculated as

$$\psi_i = \begin{cases} -c, & if \quad r_i\left(\hat{\beta}_0\right) < -c\hat{\sigma}_0 w_i, \\ \dfrac{r_i\left(\hat{\beta}_0\right)}{\hat{\sigma}_0 w_i}, & if \quad \left|r_i\left(\hat{\beta}_0\right)\right| \le c\hat{\sigma}_0 w_i, \\ c, & if \quad r_i\left(\hat{\beta}_0\right) > c\hat{\sigma}_0 w_i. \end{cases}$$

The estimated covariance matrix can be written in the matrix form

$$Cov\left(\hat{\beta}\right) = \hat{\sigma}_0^2 \left(\mathbf{X'BX}\right)^{-1} \left(\mathbf{X'V^2X}\right) \left(\mathbf{X'BX}\right)^{-1}, \tag{4.8}$$

by defining the matrix $\mathbf{V}$ as $\mathbf{V} = diag\left(w_i \psi\left(\dfrac{r_i\left(\hat{\beta}_0\right)}{\hat{\sigma}_0} \right) \right)$. The standard errors for the S1S coefficients are therefore determined by the square root of the diagonal elements of this estimated covariance matrix.

4.4.3 Case Study: Stackloss Data

In order to obtain S1S and M1S estimates we firstly have to estimate the LTS and MVE in order to provide high breakdown starting point and robust weights for improvement step. Using the repeated subsample (elemental set) algorithm with 50,000 iterations yields

$$\mathbf{MVE}_1\left(\mathbf{Z}\right) = \begin{bmatrix} 56.50 \\ 19.75 \\ 88.00 \end{bmatrix}$$

and

$$\mathbf{MVE}_2(\mathbf{Z}) = \begin{bmatrix} 99.6575 & 40.5274 & 31.8904 \\ 40.5274 & 38.2021 & 23.9178 \\ 31.8904 & 23.9178 & 82.3836 \end{bmatrix}.$$

These estimates define a minimum ellipsoid with a volume of 224.3446. This minimum volume corresponds to an objective function evaluation of 3.1797, the LTS estimator produces the fitted equation

$$\hat{y}_i = -37.031 + 0.734x_{1i} + 0.438x_{2i} + 0.000x_{3i}.$$

Given these preliminary calculations, along with the LTS-based scale estimate of $\hat{\sigma}_0 = 1.0793$, the high breakdown regression estimators M1S and S1S both yield the equation

$$\hat{y}_i = -40.8148 + 1.0443x_{1i} + 0.6805x_{2i} - 0.2133x_{3i}.$$

As shown in table 4.1, along with their respective asymptotic standard errors.

Table 4.1: High breakdown regression for stackloss data.

Parameter	LTS	M1S	M1S s.e.	S1S	S1S s.e.
Intercept	-37.031	-40.815	5.167	-40.815	5.167
x_1	0.734	1.044	0.289	1.044	0.289
x_2	0.438	0.681	0.197	0.681	0.197
x_3	0.000	-0.213	0.132	-0.213	0.132

The stackloss data has no high influence points in the data, only four outlier observations. Therefore it is not surprising that the two 1-step methods give the same estimates.

4.5 Computational Issues for High breakdown Regression

In summary, the object of high breakdown point regression is to ensure that outliers and high leverage points (up to 50% of the data) do not ruin the analysis. These proposed methods are not without problems either. The LMS suffers from a slow convergence rate and both the LMS and LTS are asymptotically inefficient. While M1S and S1S were developed to deal with the asymptotic inefficiency, they both require an initial estimate and a set of robust Mallows weights (based on the MVE estimators). In the estimation, using the LTS is where the first problem arises. The LTS is generally not the exact solution to the objective function but merely an estimation of the estimate. Hettmansperger and Sheather (1992) point out that small change in the centre of the regressor space lead to changes in LTS and LMS estimates. As a result of the objective function having local minima at various different locations leads to another problem encountered. The repeated sampling process can result in significantly different estimates for LTS. Along with the LTS, the MVE also yields significantly different results when the subsampling algorithm is repeated. Thus the robust weights may differ from one complete simulation to the next. Cook and Hawkins (1990) discuss this problem while trying to replicate the results obtained by Rousseeuw and von Zomeren (1990). It is further worth noting that Cook and Hawkins (1990) took nearly 60,000 iterations to achieve the true MVE for a data set containing 20 observations and has 5 regressor variables. This brings the validity of subsampling by the probabilistic argument, as proposed in Rousseeuw and Leroy (1987) and Rousseeuw (1993), in doubt.

One possible solution would be to use the FSA algorithm to obtain the MVE, as this algorithm is more stable with respect to random starts. The drawback is that the FSA algorithm increases the computation time.

In closing, using LTS-based initial estimates can lead to an internal breakdown. Furthermore, to obtain accurate results when estimating the various estimators the number of iterations suggested in the literature (less than 1000) may prove to be inadequate. Since S1S and M1S use LTS as an initial estimate these problems extend to these estimators as well. Thus, even if all the precautions are taken when an estimation procedure is used, every time the procedure is performed it may yield a different answer.

Chapter 5

The ARCH Model

5.1 Introduction

The ARCH model was proposed by Engle (1982) in an attempt to model the variance of the inflation in the United Kingdom. As is well documented in the literature the log returns of shares are not serially (or auto) correlated. Engle suggested that even though the returns are not serially correlated, the r_t^2 form a dependant polynomial. In figure 5.1 it can be seen that the log returns form a stationary and a not serially correlated time series. In figure 5.2 the dependant polynomial that is formed by squaring the log returns of Anglos, can be observed.

Figure 5.1

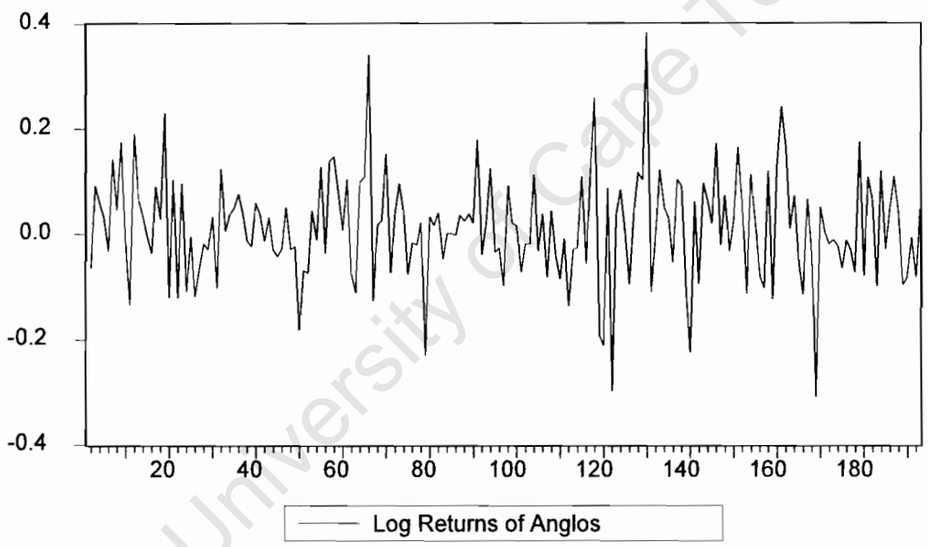
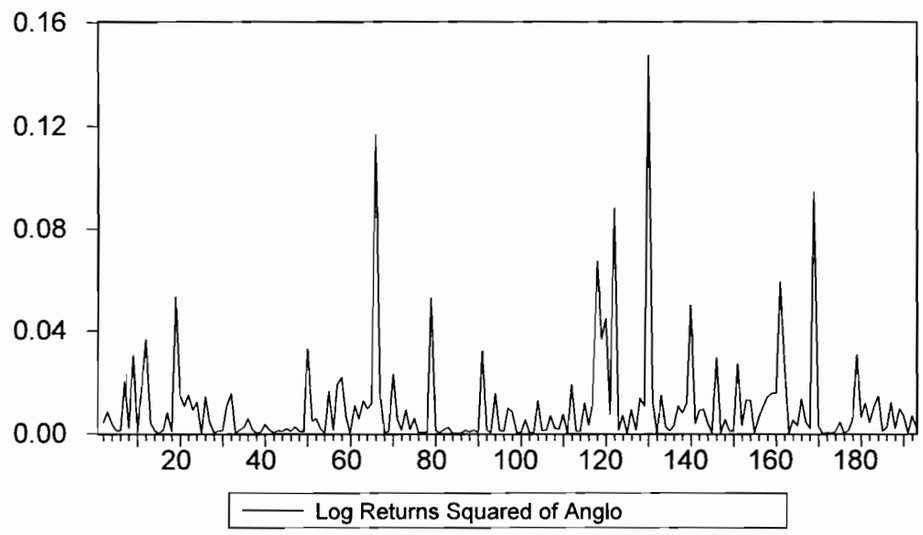


Figure 5.2



5.2 The ARCH(1) Model

The ARCH(1) models the conditional variance at time period t (denoted h_t) of the return of a share at time period t (denoted r_t) by using the square of the first lagged return. The ARCH(1) model can written as

$$R_t = \sqrt{h_t}\varepsilon_t \quad where \quad h_t = Var\left(R_t | I_{t-1}\right) = \alpha_0 + \alpha_1 R_{t-1}^2, \tag{5.1}$$

where ε_t is assumed to be the associated white noise process with a variance of 1. The white noise process ε_t is often modelled as a standard normal random variate or as a standardized student t distribution. I_{t-1} Represents the information set available at time $t-1$. In order to guarantee $h_0 \geq 0$, we assume that α_0 and $\alpha_1 \geq 0$.

As a result of the independence between ε_t and h_t, the conditional mean of R_t can be written as

$$E(R_t|I_{t-1}) = 0 \ \forall \ t, \tag{5.2}$$

while the conditional variance is

$$\begin{aligned} \text{var}(R_t|I_{t-1}) &= E\left(\left(\sqrt{h_t}\varepsilon_t\right)^2|I_{t-1}\right) \\ &= E\left(h_t\varepsilon_t^2|I_{t-1}\right) \\ &= E\left(h_t|I_{t-1}\right) \\ &= E\left(\alpha_0 + \alpha_1 R_{t-1}^2\right) \ \forall \ t. \end{aligned} \tag{5.3}$$

Another assumption that is made is that the unconditional mean as well as the variance are time invariant. As a result of this $\{R_t\}$ can be considered to be being weakly stationary. The unconditional variance of $\{R_t\}$ can now be shown as

$$\begin{aligned} \text{var}(R_t) &= \text{var}(R_{t-1}) \\ &= E\left(R_{t-1}^2\right) \end{aligned}$$

since $E(R_t) = 0$. This implies that the unconditional variance is

$$\text{var}(R_t) = \text{var}\frac{\alpha_0}{1-\alpha_1}.$$

In order to improve notation for the rest of this chapter all expectations and variances are conditional expectations and conditional variances, unless stated otherwise. The ARCH(1) model can be generalized to specify an ARCH(p) model. This ARCH(p) model can be shown to be

$$R_t = \sqrt{h_t}\varepsilon_t \quad \textit{where} \quad h_t = Var(R_t|I_{t-1}) = \alpha_0 + \alpha_1 R_{t-1}^2 + \alpha_2 R_{t-2}^2 + \ldots + \alpha_p R_{t-p}^2. \tag{5.6}$$

Bollerslev (1986) noted that in many financial time series models, the size of the residuals were apparently dependant on the size of recent residuals and this motivated this particular specification of heteroskedasticity. One can further surmise from equations (5.2) and (5.3) that the model implies that large prior returns will generate large conditional variances at time *t*, since the conditional variance at time *t* is dependant on the previous *p* lagged returns of a share. It is further worth noting that positive and negative returns are treated in the same manner as a result of the squaring of the lagged returns in order to model the conditional variance structure of the returns of a share.

It can be shown that when ε_t is normally distributed, the kurtosis of the ARCH(1) model is greater than 3. This value indicates that ARCH(1) model has heavier tails than the normal distribution.

A limitation on the ARCH models is the restrictions that are forced on the $\alpha_i's$, for all *i*. When examining the $E(R_t)$, this becomes evident. Assuming that $\{R_t\}$ is fourth order stationary $E(R_t)$ can be expressed as follows

$$E(R_t) = m_4 = \frac{3\alpha_0^2(1+\alpha_1)}{(1-\alpha_1)(1-3\alpha_1^2)}.$$

Since the $\text{var}(R_t) \geq 0$ and $\alpha_0 \geq 0$ implies that $0 \leq \alpha_1 \leq 1$, but since $m_4 \geq 0$ implies that $0 \leq \alpha_1 \leq \sqrt{\frac{1}{3}}$. In the next section we will discuss how to test for ARCH errors in a data set.

5.3 Testing for ARCH Errors

Engle (1982) provides a Lagrange multiplier (LM) test, this test highlights the presence of ARCH effects in the residuals of a time series. The null hypothesis, H_0, of the test

states that there are no ARCH terms up to order p in the residuals of the time series model. In order to test the hypothesis, a regression of the following form is run:

$$e_t^2 = \beta_0 + \beta_1 e_{t-1}^2 + \beta_2 e_{t-2}^2 + \ldots + \beta_p e_{t-p}^2, \tag{5.7}$$

where $\{e_t\}$ represents the residual series obtained from a time series model. This regression model represents the regression of the squared residuals on a constant and the lagged squared residuals up to order p. The LM test statistic is equal to nR^2, with the LM test statistic asymptotically χ_p^2 distributed, where n is equal to the number of observations and R^2 is equal to the coefficient of determination of the fitted time series model. The null hypothesis is rejected if $nR^2 \geq \chi_p^2$. Gourieroux (1997) supplies a complete proof of this test. Once it has been determined that an ARCH process is present, the next step is to determine whether the residual term, ε_t, is either normally distributed or follows a standardised student t distribution. Once an ARCH model has been chosen to model the conditional variance of a time series then $\hat{\varepsilon} = \frac{R_t}{\sqrt{\hat{h}_t}}$ should be normally distributed or have the standardised student t distribution. The distribution of $\hat{\varepsilon}_t$ depends on the initial assumption of the specification of the distribution of ε_t. The χ^2 goodness of fit statistic can be calculated to determine whether the estimated residuals follow the chosen distribution. Another possible testing method would be to utilize QQ plots in testing the error distribution assumption.

Another aspect in testing, is determining what order of the residuals should be tested up to. Define ε_t^*

$$\varepsilon_t^* = R_t^2 - \sigma_t^2, \tag{5.8}$$

where σ_t^2 is the true conditional variance at time t and ε_t^* is considered to uncorrelated with a mean of 0. If we now substitute $E(h_t)$, as an estimate for σ_t^2, equation (5.8) becomes

$$R_t^2 = \alpha_0 + \alpha_1 R_{t-1}^2 + \alpha_2 R_{t-2}^2 + \ldots + \alpha_p R_{t-p}^2 + \varepsilon_t^*. \tag{5.9}$$

From this equation the ARCH(p) specification can be regarded as an *AR(p)* process for R_t^2, the $\{\varepsilon_t^*\}$ is however not a *iid* sequence. Tsay and Tiao (1984) showed that the Box Jenkins methods as well as the Extended Sample Autocorrelation Function (ESACF) methodology can be applied to tentatively determine the order of the ARCH process. Next we will be looking at parameter estimation of the ARCH(p) model.

5.4 Parameter Estimation

Making use of the conditional log likelihood function, we are able to estimate the parameters of the ARCH(p). As stated previously, the ε_t is assumed to be either a standard normal variate or a standardised students t distribution. If ε_t is normally distributed, the conditional log likelihood can be shown to be

$$\log\left\{\prod_{t=p+1}^{T}\frac{1}{\sqrt{2\pi h_t}}e^{\left(-\frac{R_t^2}{2h_t}\right)}\right\} = \sum_{t=p+1}^{T}\left\{-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(h_t) - \frac{1}{2}\frac{R_t^2}{h_t}\right\}, \tag{5.9}$$

where

$$h_t = \alpha_0 + \alpha_1 R_{t-1}^2 + \alpha_2 R_{t-2}^2 + \ldots + \alpha_p R_{t-p}^2. \tag{5.10}$$

In order to maximise the conditional log likelihood function equation (5.9) should be evaluated iteratively for each observation. It can be shown that $R_t \sim N(0, h_t)$.

If the errors have a fat tailed distribution the standardised t distribution can be used to model the errors. This is done by transforming the error term to

$$\varepsilon_t = \frac{X}{\sqrt{\frac{v}{v-2}}},$$

for $v > 2$ where X has a students t distribution with v degrees of freedom. The pdf of ε_t can be shown to be

$$f_{\varepsilon_t}(\varepsilon_t|v) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)\sqrt{(v-2)\pi}}\left(1+\frac{\varepsilon_t^2}{v-2}\right)^{-\frac{(v+1)}{2}} \quad -\infty \le \varepsilon_t \le \infty \tag{5.11}$$

and thus the pdf of $R_t = \sqrt{h_t}\varepsilon_t$ is:

$$f_{R_t}(R_t|v) = \frac{\Gamma\left(\frac{(v+1)}{2}\right)}{\Gamma\left(\frac{v}{2}\right)\sqrt{(v-2)\pi h_t}}\left(1+\frac{R_t^2}{(v-2)h_t}\right)^{-\frac{(v+1)}{2}}. \tag{5.11}$$

The conditional log likelihood function that is used when performing parameter estimation is then

$$\sum_{t=p+1}^{T}\left\{\log\frac{\Gamma\left(\frac{(v+1)}{2}\right)}{\Gamma\left(\frac{v}{2}\right)\sqrt{(v-2)\pi h_t}} - \frac{v+1}{2}\log\left(1+\frac{R_t^2}{(v-2)h_t}\right) - \frac{1}{2}\log(h_t)\right\}.$$

Chapter 6

The GARCH Model

6.1 Introduction

A problem with the ARCH model is that it often requires relatively long lag structures in the conditional variance equation. Bollerslev (1986) developed the Generalised ARCH (GARCH) models to remedy this problem. As shown previously, the ARMA model is an extension of AR and MA models, similarly the GARCH model is an extension of the univariate ARCH model.

6.2 The GARCH(1,1) Model

Let R_t be the mean adjusted return of a share at time t and r_t is the return of a share at time t. Then R_t can be shown to be

$$R_t = r_t - \mu_t .$$

The GARCH(1,1) model can be specified as follows

$$R_t = \sqrt{h_t}\varepsilon_t \quad where \quad h_t = \alpha_0 + \alpha_1 r_{t-1}^2 + \beta_1 h_{t-1} \tag{6.1}$$

or

$$R_t^2 = \alpha_0 + (\alpha_1 + \beta_1)R_{t-1}^2 + w_t - \beta_1 w_{t-1} . \tag{6.2}$$

As with the ARCH model, we can estimate the parameters of an GARCH(1,1) model by using the conditional log likelihood function. Assuming the errors are $N(0,1)$ distributed then the conditional log likelihood function can be shown to be

$$\log\left\{\prod_{t=2}^{T}\frac{1}{\sqrt{2\pi h_t}}e^{\left(-\frac{R_t^2}{2h_t}\right)}\right\}=\sum_{t=2}^{T}\left\{-\frac{1}{2}\log(2\pi)-\frac{1}{2}\log(h_t)-\frac{1}{2}\frac{R_t^2}{h_t}\right\}. \qquad (6.3)$$

This equation is similar to the conditional log likelihood function used to estimate the parameters in the ARCH(p) model. They differ in their starting points of the summation. The ARCH(p) model starts at $t=p+1$ whereas the GARCH(1,1) summation starts at $t=2$. It is also worth noting that h_t depends on the first lagged conditional variance, this requires us to estimate h_1. The GARCH(1,1) model can be generalised to specify a GARCH(p,q) model.

6.3 The GARCH(p,q) Model

R_t Follows a GARCH(p,q) process if and only if

$$R_t=\sqrt{h_t}\varepsilon_t \quad where \quad h_t=\alpha_0+\sum_{i=1}^{p}\alpha_i R_{t-i}^2+\sum_{j=1}^{q}\beta_j h_{t-j}\,, \qquad (6.4)$$

where ε_t is assumed to be the associated white noise process with an $E(\varepsilon_t)=0$ and $Var(\varepsilon_t)=1$. The p and q in the GARCH(p,q) process is representative of the number of ARCH and GARCH terms respectively.

It can be shown that if R_t follows a GARCH(p,q) process then R_t^2 is an ARMA(max(p,q),q) process. This can be seen by rearranging the above equation as follows:

$$h_t = \alpha_0 + \left(\alpha_1 R_{t-1}^2 + \ldots + \alpha_p R_{t-p}^2\right) + \left(\beta_1 h_{t-1} + \ldots + \beta_q h_{t-q}\right).$$

Next we add R_t^2 to both sides of the equation and rearranging some of the terms we get

$$\begin{aligned} h_t + R_t^2 &= \alpha_0 + \left(\alpha_1 R_{t-1}^2 + \ldots + \alpha_p R_{t-p}^2\right) + \left(\beta_1 h_{t-1} + \ldots + \beta_q h_{t-q}\right) + R_t^2 \\ &= \alpha_0 - \beta_1\left(R_{t-1}^2 - h_{t-1}\right) - \beta_2\left(R_{t-2}^2 - h_{t-2}\right) - \ldots - \beta_q\left(R_{t-q}^2 - h_{t-q}\right) + \\ &\quad \beta_1 R_{t-1}^2 + \beta_2 R_{t-2}^2 + \ldots + \beta_q R_{t-q}^2 + \left(\alpha_1 R_{t-1}^2 + \ldots \alpha_p R_{t-p}^2\right) + R_t^2. \end{aligned}$$

Performing some rearranging again we get

$$\begin{aligned} R_t^2 = \alpha_0 &+ \left(R_t^2 - h_t\right) - \beta_1\left(R_{t-1}^2 - h_{t-1}\right) - \beta_2\left(R_{t-2}^2 - h_{t-2}\right) - \ldots - \beta_q\left(R_{t-q}^2 - h_{t-q}\right) \\ &+ \beta_1 R_{t-1}^2 + \beta_2 R_{t-2}^2 + \ldots + \beta_q R_{t-q}^2 + \alpha_1 R_{t-1}^2 + \alpha_2 R_{t-2}^2 + \ldots + \alpha_p R_{t-p}^2. \end{aligned}$$

So that finally we have

$$R_t^2 = \alpha_0 + \left((\alpha_1 + \beta_1) R_{t-1}^2 + \ldots + (\alpha_p + \beta_p) R_{t-p}^2\right) + \left(w_t - \beta_1 w_{t-1} - \ldots - \beta_q w_{t-q}\right),$$

where $w_t = R_{t-1}^2 - h_{t-1}$, so that

$$R_t^2 = \alpha_0 + \sum_{i=1}^{\max(p,q)} (\alpha_i + \beta_i) R_{t-i}^2 + \sum_{j=1}^{q} \beta_j w_{t-j}. \qquad (6.5)$$

6.4 GARCH Model Limitations

Since the GARCH model is a generalised ARCH model, the GARCH model inherits many of the model limitations associated with the ARCH model. The GARCH model can however model the conditional variance of a time series by using relatively few parameters. In general it is difficult to determine the p and q values in a GARCH(p,q)

model and therefore the GARCH(1,1) is most often used. Note that the GARCH model is a symmetric model with respect to the volatility, simply stated it treats positive and negative returns in the same manner since the square of lagged returns are utilised in modelling the conditional variance equation of shares returns.

This lead to the development of asymmetric volatility models which incorporates the sign of lagged returns into the specification of the variance equation. Two such volatility models are the Threshold ARCH (or TARCH) and the Exponential GARCH (or EGARCH) models. The Threshold ARCH model yields similar results to the exponential GARCH and will not be discussed further.

6.5 The Exponential GARCH

It was noted by both Black (1976) and Christie (1982) that there exists a negative correlation between current returns and future volatility of a share returns. Simply stated the volatility of a share tends to rise when bad news is received or perceived and falls when good news is received or perceived. Good news can be classified as when the return of a share is greater than the consensus view. As stated in the previous section, the GARCH model cannot incorporate this observation into its model as it models the conditional volatility as the sum of squared lagged returns and lagged conditional variance. This is because the GARCH model concentrates on the size of returns at each time period opposed to the sign of the return at that time period.

The exponential GARCH was developed by Nelson (1991) to incorporate the sign into the model; this allows for asymmetric effects between positive and negative share returns. The conditional variance can be modelled as

$$\log(h_t) = \alpha_t + \sum_{j=1}^{\infty} \beta_j g(\varepsilon_{t-j}), \tag{6.6}$$

where $\{\alpha_t\}_{t=(-\infty,\infty)}$ and $\{\beta_j\}_{j=[1,\infty)}$ are real, non-stochastic scalars and $g(.)$ is a function that incorporates both the sign and magnitude of a return during the time period *t*. Nelson (1991) proposed the following weighting scheme

$$g(\varepsilon_t) = \theta\varepsilon_t + \gamma\left[|\varepsilon_t| - E(|\varepsilon_t|)\right],$$

where θ and γ are real constants. The $E(|\varepsilon_t|) = 0$, since both the ε_t and $|\varepsilon_t| - E(|\varepsilon_t|)$ are zero mean *iid* sequences and continuous distributions. The $g(\varepsilon_t)$ can now be written as

$$g(\varepsilon_t) = \begin{cases} (\theta+\gamma)\varepsilon_t - \gamma E(|\varepsilon_t|) & \text{if } \varepsilon_t \geq 0, \\ (\theta-\gamma)\varepsilon_t - \gamma E(|\varepsilon_t|) & \text{if } \varepsilon_t < 0, \end{cases}$$

which illustrates asymmetric property of $g(.)$. This asymmetric property allows for the negative relationship between returns and conditional volatility (i.e. if $\gamma = 0$ and $\theta < 0$ then the change in the conditional variance is positive (negative) when $\varepsilon_t < 0$ $(\varepsilon_t > 0)$).

Chapter 7

Markowitzs Portfolio Selection

7.1 Markowitz Portfolio Theory

Consider a portfolio consisting of p stocks, with the vector of stock returns for the portfolio is written as

$$\mathbf{R} = \begin{pmatrix} R_1 \\ \vdots \\ R_p \end{pmatrix},$$

where the expected returns can be shown to be

$$E(\mathbf{R}) = \boldsymbol{\mu}$$

and R is defined as

$$R_i = \log P_t - \log P_{t-1}, \quad i = 1,\ldots,N, t = 1,\ldots,N. \tag{7.1}$$

Let P_t define the price of a stock or share at time t.

Furthermore, we assume that the covariances between the returns of the different shares are non-zero. The covariance matrix of the stock returns can be shown as

$$\Sigma = E(\mathbf{R} - \boldsymbol{\mu})(\mathbf{R} - \boldsymbol{\mu})',$$

$$\Sigma = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1p} \\ \vdots & & \vdots \\ \sigma_{p1} & \cdots & \sigma_{pp} \end{pmatrix},$$

where $\sigma_{ii} = \sigma_i^2$ is the variance of the *i'th* and σ_{ij} the covariance between the *i'th* and *j'th* stock.

Further assume that the returns, $\mathbf{R} \sim N(\mu, \Sigma)$, are multivariate normally distributed.

Purchasing a portfolio of stocks is akin to investing proportions of the funds available in different stocks. Let w_i denote the proportion (also known as the weight of stock *i*) of the cash invested in a stock. This leads to the equation

$$\sum_{i=1}^{p} w_i = 1.$$

Simply stated this equation means that one cannot invest more money than is available.

Let

$$\mathbf{W} = \begin{pmatrix} w_1 \\ \vdots \\ w_p \end{pmatrix}, \tag{7.2}$$

be the vector of investment weights in each of the p stocks, and P be the return on the portfolio, then

$$\begin{aligned} P &= \mathbf{W}'\mathbf{R} \qquad (7.3) \\ &= \sum_{i=1}^{p} w_i R_i , \end{aligned}$$

subject to

$$\sum_{i=1}^{p} w_i = 1.$$

We can now write the expected return of the portfolio as

$$\begin{aligned} E(P) &= \mathbf{W}' E(\mathbf{R}) \\ &= \mathbf{W}' \boldsymbol{\mu} \\ &= \sum_{i=1}^{p} w_i \mu_i \\ &= \mu_p, \end{aligned}$$

and the variance as

$$\begin{aligned} \text{var}(P) &= \mathbf{W}' \Sigma \mathbf{W} \\ &= \sum_{i=1}^{p} w_i w_j \sigma_{ij} \\ &= \sigma_p^2. \end{aligned}$$

Since we assumed that $\mathbf{R} \sim N(\boldsymbol{\mu}, \Sigma)$, this implies that $P \sim N(\mu_p, \sigma_p^2)$. It is worth noting that if the weights of the investment are changed, the expected return and variance of the portfolio changes.

The problem of portfolio optimization is two fold. Firstly one would like to maximize the return attained by a portfolio and secondly one would like to minimize the variance of the portfolio. The variance of a portfolio is referred to as the risk of a portfolio. The

portfolio problem can simply be stated as choosing the weights w_i, such that the expected return $E(P) = \mu_p$ is maximized and the variance (or risk) $\text{var}(P) = \sigma_p^2$ is minimized.

In equation form the portfolio problem cab be written as

$$\max_{w_i} E(P) = \mathbf{W}'\boldsymbol{\mu} \tag{7.4}$$

$$= \sum_{i=1}^{p} w_i \mu_i \ ,$$

$$\min_{w_i} \sigma_p^2 = \mathbf{W}' \sum \mathbf{W} \tag{7.5}$$

$$= \sum_{i=1}^{p} w_i w_j \sigma_{ij} \ ,$$

subject to

$$\sum_{i=1}^{p} w_i = 1 \ .$$

Markowitz further proposed to constrain the weights such that

$$0 \le w_i \le 1 \quad for \ \ i = 1, \ldots, p \ .$$

The Markowitz portfolio problem can then formally be written as

$$\max_{w_i} E(P) = \mathbf{W}'\boldsymbol{\mu} \tag{7.6}$$

$$= \sum_{i=1}^{p} w_i \mu_i \ ,$$

$$\min_{w_i} \sigma_p^2 = \mathbf{W}' \sum \mathbf{W} \tag{7.7}$$

$$= \sum_{i=1}^{p} w_i w_j \sigma_{ij} ,$$

subject to

$$\sum_{i=1}^{p} w_i = 1 ,$$

where $0 \leq w_i \leq 1 \quad for \ i = 1, \ldots, p$.

The above set of equations forms a non-linear (quadratic) programming problem, *QP-problem.* This problem can be solved by one of two methods, firstly fix the variance σ_p^2 and then maximize the return μ_p or fix the return μ_p and then minimize variance σ_p^2. This leads us to the following constrained *QP* problem

$$Min \ \sigma^2 = \mathbf{W}' \sum \mathbf{W} \tag{7.8}$$

$$= \sum_{i=1}^{p} w_i w_j \sigma_{ij} ,$$

subject to

$$\mu_p = \mathbf{W}' \boldsymbol{\mu} \tag{7.9}$$

$$= \sum_{i=1}^{p} w_i \mu_i$$

$$= E_k ,$$

where E_k is a fixed expected return, and

$$\sum_{i=1}^{p} w_i = 1,\ 0 \le w_i \le 1.$$

Applying Lagrange multipliers the equality constraints can be solved. The *QP* problem can now be written as

$$Min\ Z = \mathbf{W}' \sum \mathbf{W} - \phi\left(\sum_{i=1}^{p} w_i \mu_i - E_k\right) - \lambda\left(\sum_{i=1}^{p} w_i - 1\right), \tag{7.10}$$

subject to

$$0 \le w_i \le 1,\ \ i = 1, \ldots, p.$$

This form of a *QP* problem is referred to as a standard *QP* problem, with bounds $0 \le w_i \le 1,\ \ i = 1, \ldots, p$. The bounds of the *QP* problem can be generalized, so that any weight constraint can be imposed. One example of such a change in the boundary constraint is

$$L_i \le w_i \le U_i,\ \ i = 1, \ldots, N,$$

where $L_i \ge 0$, is a lower bound and U_i is an upper bound. The lower bound can be interpreted as not being allowed to invest less than the proportion L_i in a stock, while the upper bound implies that not more than the proportion U_i can be invested in a stock.

To this idea of the *QP* problem, Markowitz added the Efficient Frontier, which will be discussed in the following section.

7.2 The Efficient Frontier

A portfolio is defined to be efficient if it satisfies the following properties

- For a fixed (or predetermined) amount of risk, the expected return is maximised, or for a fixed (or predetermined) amount of return, the expected risk (or variance) is minimized.
- The portfolio is considered to be legitimate (i.e. no negative sales).

In order to compute this one must firstly generate a large number of portfolio returns, the values range between the largest and smallest returns for the portfolio. Secondly, each return is used to determine the weights which minimize the variance for a portfolio with that return. Lastly, these corresponding returns and variances are plotted. The boundary line on the graph is referred to as the efficient frontier. The efficient frontier can be defined as follows:

Let σ_p^2 represent the x-axis and E_k represent the y-axis on a Cartesian plane. The plot of all efficient portfolios on this set of axis is called the efficient frontier.

Chapter 8

Sharp Portfolio Selection

8.1 Sharp Single Index Model Theory

Let the log return of a particular stock be

$$R_t = \log P_t - \log P_{t-1},\ t = 1,\ldots N,$$

where t is the time and is sufficiently large enough for R_t to follow a Normal $N(\mu_r, \sigma_r^2)$ distribution. Secondly, let I_t be the return of the market proxy, then I_t can be written as

$$I_t = \log I_t - \log_{t-1},\ t = 1,\ldots, N,$$

with the market proxy also following a Normal distribution, $N(\mu_I, \sigma_I^2)$. Since both the R_t and I_t are normally distributed, combining the two distributions yields a bivariate normal distribution that can be written as

$$\begin{pmatrix} R_t \\ I_t \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_r \\ \mu_I \end{pmatrix}, \begin{pmatrix} \sigma_r^2 & \sigma_{rI} \\ \sigma_{Ir} & \sigma_I^2 \end{pmatrix}\right],$$

where $\sigma_{Ir} = \sigma_{rI}$ is the covariance between the return of stock r and the market proxy I.

The following properties of the bivariate normal distribution allow us to rewrite the model for R_t. The properties are

$$E(R_t / I_t) = \alpha + \beta I_t \qquad (8.1)$$

and

$$\text{var}(R_t / I_t) = \sigma_{rI}^2 = \sigma_I^2 (1 - \rho^2), \tag{8.2}$$

where ρ is the correlation between the stock return and the market proxy.

As a result the model can now be written as

$$R_t = \alpha + \beta I_t + e_t, \tag{8.3}$$

where $e_t \sim N(0, \sigma^2)$. The error terms are assumed to be independently distributed over time, that is

$$E(e_t e_s) = 0, \quad for \quad t \neq s. \tag{8.4}$$

This new model is known as the Sharpe index model.

The β (or beta) parameter is referred to as the systematic risk of a security. It can be used as a measure of the volatility of the security, relative to the market proxy. Simply stated, if β is greater than one, then, if the market rises, the return of the security will rise more rapidly than the return on the market. The opposite is also true, if the market falls then the return on the security falls more rapidly than the market. This implies that the security is more volatile than the market and hence more risky. The converse is also true if β is less than one.

It is worth noting some of the assumptions attached to the model in equation (8.1) that allow us to imply the Sharpe index model.

- The bivariate normal assumption (or multivariate normal for several securities and/or indices) appears to be well accepted in the literature.

- Fama, Fisher, Jensen and Roll showed that if the bivariate normality is not attainable, then the linearity assumption appears to be well satisfied.
- It is assumed that the beta coefficient is relatively stable over time. This assumption holds even truer if the time period under consideration increases. Thus the parameter can be estimated using the historical data.
- Further it is assumed that even if the beta coefficients change over time, the ranking of the security risk does not. This assumption plays an important role when the index model is used for portfolio optimization.
- It has been proven that the beta coefficient is a good measure of the inherent risk in a security.
- A fundamental characteristic of the firm relates directly to the value of beta, in any period.
- If the error terms e_t are not normally distributed they still posses the following characteristics; $E(e_t) = 0,\ E(e_t)^2 = \sigma^2$ and $E(e_t e_s) = 0,\ t \neq s$.

Taking the assumptions into account, a definition of the Sharpe index model can be given and is stated as

$$R_t = \alpha + \beta I_t + e_t,\ t = 1,\ldots,N, \tag{8.5}$$

with the following assumptions about the error term ;

$$E(e_t e_s) = 0,\ t \neq s = 1,\ldots,N \tag{8.6}$$

and

$$E(e_t^2) = \sigma_e^2.$$

If we take a portfolio of stocks, then the *i*'th stock can be written as

$$R_{it} = \alpha_i + \beta I_{it} + e_{it}, \quad i = 1,\ldots p; \quad t = 1,\ldots,N.$$

All the stocks are regressed against the same single index *I*, with the following assumptions about the error terms;

- $E(e_{it}^2) = \sigma_{ei}^2$, this assumption implies that each stock has a unique variance for the error term
- $E(e_{it}e_{is}) = 0, \quad t \neq s = 1,\ldots,N$, this assumption indicates that the error terms for each stock are independent over the period in question.
- $E(e_{it}I_t) = 0, \quad t = 1,\ldots,N$, the usual regressor assumption that error terms of each stock are un-correlated with the explanatory variable *I*.
- $E(e_{it}e_{jt}) = 0, \quad t = 1,\ldots,N$, this assumption implies that the error terms of the stocks are un-correlated; moreover the stocks are only related through their mutual relationship with the index *I*.

Let

$$E(I) = \mu_I \quad and \quad \text{var}(I) = \sigma_I^2$$

be the mean and variance of the index respectively.

The following holds for each stock, the expected value can be shown as

$$E_i = E(R_i) = \alpha_i + \beta_i \mu_I,$$

the variance as

$$\begin{aligned} \text{var}(R_i) &= \text{var}(\alpha_i + \beta_i I + e_i) \\ &= \beta_i^2 \sigma_I^2 + \sigma_{ei}^2 \\ &= \sigma_{ii} = \sigma_i^2, \end{aligned}$$

and finally the covariance is written as

$$\begin{aligned}\operatorname{cov}(R_i R_j) &= E\left[\left(R_i - E(R_i)\right)\left(R_j - E(R_j)\right)\right] \\ &= E\left[\left(\beta_i\left(I - \mu_I\right) + e_i\right)\left(\beta_j\left(I - \mu_I\right) + e_j\right)\right] \\ &= \beta_i \beta_j \sigma_I^2 .\end{aligned}$$

Our portfolio problem then becomes the maximisation of the objective function

$$\begin{aligned}Max\ Z &= \phi\mu_p - \sigma_p^2 \\ &= \phi \mathbf{W}'\mu - \mathbf{W}'\Sigma\mathbf{W}\end{aligned} \tag{8.7}$$

subject to $\sum_{i=1}^{p} w_i = 1$, where

$$\begin{aligned}\mu_p &= \sum_{i=1}^{p} w_i E_i \\ &= \sum_{i=1}^{p} w_i\left(\alpha_i + \beta_i \mu_I\right)\end{aligned} \tag{8.8}$$

and

$$\begin{aligned}\sigma_p^2 &= \mathbf{W}'\Sigma\mathbf{W} \\ &= \sum_{i=1}^{p}\sum_{j=1}^{p} w_i w_j \sigma_{ij}\end{aligned} \tag{8.9}$$

with

$$\begin{aligned}\sigma_{ii} &= \sigma_i^2 \\ &= \beta_i^2 \sigma_I^2 + \sigma_{ei}^2\end{aligned} \tag{8.10}$$

and

$$\sigma_{ij} = \beta_i \beta_j \sigma_I^2 . \tag{8.11}$$

In order to determine the efficient frontier the following variables have to be estimated. Firstly

$$\alpha_i, \ \beta_i, \ \sigma_{ei}^2, \ i = 1, \ldots, p$$

and secondly

$$\mu_I \ \text{ and } \ \sigma_I^2 .$$

It is worth noting that the total number of quantities that need to be estimated for the Sharp model is $3p+2$, compared to the Markowitz model which requires the estimation of $p + \frac{1}{2}p(p+1)$ quantities. As the value of p increases the number of quantities that need estimation in the Markoqitz model can become quite a significant task.

8.2 The Sharpe Multiple Index Model

The multi-index model can be written as

$$R_{it} = \alpha_i + \beta_{i1} I_1 + \beta_{i2} I_2 + \ldots + \beta_{iM} I_M + e_{it} ,$$
$$i = 1, \ldots, p, \ t = 1, \ldots, N .$$

Once again we make assumptions about the error terms, these assumptions are

$$E(e_{it}^2) = \sigma_{ei}^2, \tag{8.12}$$

$$E(e_{it}e_{is}) = 0, \quad t \neq s = 1,\ldots,N, \tag{8.13}$$

$$E(e_{it}I_{jt}) = 0, \quad j = 1,\ldots,M, \quad t = 1,\ldots,N, \tag{8.14}$$

$$E(e_{it}e_{jt}) = 0, \quad t = 1,\ldots,N, \tag{8.15}$$

$$E(I_{jt}I_{kt}) = c_{jk}, \quad k = 1,\ldots,M. \tag{8.16}$$

These assumptions are the same as for the single index model except we have the added assumption that Indices are dependant with covariances given by c_{jk} (equation 8.16). Furthermore we assume that the error term e_{it} is independent of the indices I_j, $j = 1,\ldots,M$.

Now, let

$$E_i = E(R_i) = \alpha_i + \beta_{i1}I_1 + \ldots + \beta_{iM}I_M, \quad i = 1,\ldots,p.$$

Thus, for a portfolio $P = W'R = \sum_{i=1}^{p} w_i R_i$ we have that

$$\mu_p = \sum_{i=1}^{p} w_i E_i$$

and

$$\begin{aligned}\sigma_p^2 &= \sum_{i=1}^{p} w_i w_j \sigma_{ij} \\ &= \sum_{i=1}^{p} w_i^2 \sigma_{ei}^2 \\ &\quad + \sum_{i=1}^{p} w_i^2 \sum_{k,l} \beta_{ik}\beta_{il}c_{kl}\end{aligned}$$

$$
\begin{aligned}
&+ \sum\sum_{i \neq j} w_i\, w_j \sum_{k,l} \beta_{ik} \beta_{jl} c_{kl} \\
&= \sum_{i=1}^{p} w_i^2 \sigma_{ei}^2 \\
&+ \sum_{i} \sum_{j} w_i w_j \sum_{k} \sum_{l} \beta_{ik} \beta_{jl} c_{kl} \\
&= \sum_{i=1}^{p} w_i^2 \sigma_{ei}^2 \\
&+ \sum_{k} \sum_{l} \sum_{i} \sum_{j} w_i \beta_{ik} w_j \beta_{jl} c_{kl} \\
&= \sum_{i=1}^{p} w_i^2 \sigma_{ei}^2 \\
&+ \sum_{k} \sum_{l} \left(\sum_{i} w_i \beta_{ik} \right) \left(\sum_{j} w_j \beta_{jl} \right) c_{kl} \\
&+ \sum_{i=1}^{p} w_i^2 \sigma_{ei}^2 + \sum_{k} \sum_{l} \beta_{pk} \beta_{pl} c_{kl}
\end{aligned}
$$

where

$$\beta_{pk} = \sum_{i=1}^{p} w_i \beta_{ik}, k = 1, \ldots, M$$

$$\beta_{pl} = \sum_{j=1}^{p} w_j \beta_{jl}, l = 1, \ldots, M .$$

As these two equations are the same, they only differ in their subscript notations, only one has to be computed.

The objective function of the Sharpe model is (minimize instead of maximize)

$$Min\ Z = -\Lambda \mu_p + \sigma_p^2, \tag{8.17}$$

subject to the following constraints

$$\begin{aligned}
\beta_{p1} &= \sum_{i=1}^{p} w_i \beta_{i1}, \\
\beta_{p2} &= \sum_{i=1}^{p} w_i \beta_{i2}, \\
&\vdots \\
\beta_{pM} &= \sum_{i=1}^{p} w_i \beta_{iM}, \\
\sum_{i=1}^{p} w_i &= 1 .
\end{aligned}$$

These constraints are not final and any other equality, in-equality or boundary constraint can be added.

Ignoring any other constraints and bounds the objective function becomes

$$\begin{aligned}
Min\ Z' = & -\Lambda \mu_p + \sigma_p^2 \\
& + \lambda_1 \left(\beta_{p1} - \sum_{i=1}^{p} w_i \beta_{i1} \right) \\
& + \lambda_2 \left(\beta_{p2} - \sum_{i=1}^{p} w_i \beta_{i2} \right) \\
& \vdots \\
& + \lambda_M \left(\beta_{pM} - \sum_{i=1}^{p} w_i \beta_{iM} \right) \\
& + \lambda_f \left(1 - \sum_{i=1}^{p} w_i \right).
\end{aligned}$$

Solving this equations requires that the partial derivative of Z' with respect to each variable be set to zero.

Thus, for each $i = 1,\ldots, p$,

$$\frac{\partial Z'}{\partial w_i} = 2\sigma_{ei}^2 - E_i\Lambda - \lambda_1\beta_{i1} - \ldots\lambda_M\beta_{iM} = 0.$$

For each $j = 1,\ldots, M$

$$\frac{\partial Z'}{\partial \beta_{pj}} = 2\beta_{p1}c_{j1} + 2\beta_{p2}c_{j2} + \cdots + 2\beta_{pM}c_{jM} + \lambda_j = 0.$$

For each $j = 1,\ldots, M$

$$\frac{\partial Z'}{\partial \lambda_j} = \beta_{pj} - \beta_{1j}w_1 - \cdots - \beta_{pj}w_p = 0.$$

For λ_f ,

$$\frac{\partial Z'}{\partial \lambda_f} = 1 - w_1 - \cdots - w_p = 0.$$

These equations lead to a system of linear equations $AX = B$ of the form

$$\begin{pmatrix} w_1 & \cdots & w_p & \beta_{p1} & \cdots & \beta_{pN} & \lambda_1 & \cdots & \lambda_M & \lambda_f \\ 2\sigma_{e1}^2 & \cdots & 0 & 0 & \cdots & 0 & -\beta_{11} & \cdots & -\beta_{1M} & -1 \\ 0 & \ddots & \vdots & \vdots & \cdots & 0 & \vdots & & \vdots & \ddots \\ 0 & \cdots & 2\sigma_{ep}^2 & 0 & \cdots & 0 & -\beta_{p1} & \cdots & -\beta_{pM} & -1 \\ 0 & \cdots & 0 & 2c_{11} & \cdots & 2c_{1M} & 1 & 0 & \cdots & 0 \\ \vdots & & & \vdots & \ddots & \vdots & & \ddots & & \vdots \\ 0 & \cdots & 0 & 2c_{M1} & \cdots & 2c_{MM} & 0 & \cdots & 1 & 0 \\ -\beta_{11} & & -\beta_{p1} & 1 & 0 & & & & & 0 \\ \vdots & & \vdots & 0 & \ddots & 0 & \cdots & & & 0 \\ -\beta_{1M} & & -\beta_{pM} & \vdots & & 1 & 0 & \cdots & & 0 \\ -1 & \cdots & -1 & 0 & \cdots & & & & & 0 \end{pmatrix} = \begin{pmatrix} 0 \\ \Lambda E_1 \\ \vdots \\ \Lambda E_p \\ 0 \\ \vdots \\ \\ \\ \\ 0 \\ -1 \end{pmatrix},$$

which has a solution of the form

$$\mathbf{X} = \mathbf{A}^{-1}\mathbf{B}.$$

If other equality, in-equality or bounds are imposed, these restrictions must be incorporated into set of linear equations.

Chapter 9

Innovations in the Classical Models

9.1 Generalisation of the Markowitz Formulation

As seen in Chapter 7, the Markowitz formulation of the portfolio theory problem is classically

$$\begin{aligned} Min\ \sigma^2 &= \mathbf{W}^{'} \sum \mathbf{W} \\ &= \sum_{i=1}^{p} w_i w_j \sigma_{ij}\ , \end{aligned}$$

subject to

$$\begin{aligned} \mu_p &= \mathbf{W}^{'} \boldsymbol{\mu} \\ &= \sum_{i=1}^{p} w_i \mu_i \\ &= E_k\ , \end{aligned}$$

where E_k is a fixed expected return, and

$$\sum_{i=1}^{p} w_i = 1,\ 0 \le w_i \le 1\ .$$

At the time of the Markowitz formulation development short sales (i.e. negative w_i) were not common in the market, which lead to the constraint $0 \le w_i \le 1$. This has changed over the decades and short sales area a common occurrence in the markets

around the world. Due to this fact, it is no longer necessary to restrict the weights of portfolios in order for the portfolio to be considered legitimate.

The generalised formulation for the Markowitz Model can now be written as

$$Min\ \sigma^2 = \mathbf{W}' \Sigma \mathbf{W} \qquad (9.1)$$
$$= \sum_{i=1}^{p} w_i w_j \sigma_{ij} ,$$

subject to

$$\mu_p = \mathbf{W}' \boldsymbol{\mu}$$
$$= \sum_{i=1}^{p} w_i \mu_i$$
$$= E_k ,$$

where E_k is a fixed expected return, and

$$\sum_{i=1}^{p} w_i = 1,$$

where

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}, \qquad (9.2)$$

where $\sigma_{ii} = \sigma_i^2$ represents the variance of the $i'th$ stock and σ_{ij} is the covariance between the $i'th$ and $j'th$ stock. This change in the restriction imposed on the weights, allows for the short sale of shares.

9.2 Innovations to the Sharpe Single Index Model

9.2.1 Bayesian estimates in the Sharpe single index model

Let I_f represent the posterior predictive mean and let

$$\mathbf{R}_f = \begin{pmatrix} R_{1f} \\ R_{2f} \\ \vdots \\ R_{pf} \end{pmatrix}$$

represent the future predictive returns of the p-stocks, then setting a prior distribution for

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

allows the derivation of the posterior distribution for

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix}.$$

This derivation of the posterior distribution allows in turn the derivation of the predictive distribution of R_f, which is dependant on the posterior predictive mean I_f.

9.2.2 Non-zero Covariance between Residuals of Shares

In the classical case the Sharpe single index model can be written as

$$R_{it} = \alpha_i + \beta I_{it} + e_{it}, \quad i = 1,\ldots p; \quad t = 1,\ldots,N, \tag{9.3}$$

with the following assumptions about the error terms

$$E\left(e_{it}^2\right) = \sigma_{ei}^2, \tag{9.4}$$

$$E(e_{it}e_{is}) = 0, \quad t \neq s = 1,\ldots,N \tag{9.5}$$

$$E(e_{it}I_t) = 0, \quad t = 1,\ldots,N \tag{9.6}$$

$$E(e_{it}e_{jt}) = 0, \quad t = 1,\ldots,N. \tag{9.7}$$

In vector notation equation (9.3) can be written as

$$\boldsymbol{R}_t = \boldsymbol{\alpha} + \boldsymbol{\beta} I_t + \boldsymbol{e}_t, \quad t = 1,\ldots,N$$

with

$$\boldsymbol{R}_t = \begin{pmatrix} R_{1t} \\ \vdots \\ R_{pt} \end{pmatrix}, \quad \boldsymbol{\alpha} = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_p \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \text{ and } \mathbf{e}_t = \begin{pmatrix} e_{1t} \\ \vdots \\ e_{pt} \end{pmatrix} \tag{9.8}$$

so that (when removing the index t)

$$E(\boldsymbol{R}) = \quad + \quad \mu_I \tag{9.9}$$

and

$$cov(\mathbf{e}) = \begin{pmatrix} \sigma_{e1}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{e2}^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & \sigma_{ep}^2 \end{pmatrix}. \tag{9.10}$$

This has the following implication

$$cov(\mathbf{R}) = \sigma_I^2 \boldsymbol{\beta}\boldsymbol{\beta}' + \begin{pmatrix} \sigma_{e1}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{e2}^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & \sigma_{ep}^2 \end{pmatrix}. \tag{9.11}$$

The assumption is made that all e_{it}, $i = 1,\ldots,p$, $t = 1,\ldots,N$ are independent. This assumption is now relaxed by assuming that there exists a correlation between the different error terms of the stocks, i.e. the error terms are dependant. This assumption leads to the following expectation

$$E(e_{it}e_{jt}) = \sigma_{ij} \quad i \neq j,$$
$$E(e_{it}e_{jt}) = \sigma_{ei}^2 \quad i = j,$$
$$= \sigma_{ii},$$

or

$$E(\mathbf{e}\mathbf{e}') = \boldsymbol{\Omega} = \begin{pmatrix} \sigma_{e1}^2 & \sigma_{12} & \cdots & 0 \\ \sigma_{21} & \sigma_{e2}^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{p1} & \cdots & \cdots & \sigma_{ep}^2 \end{pmatrix} \tag{9.12}$$

with

$$cov(\mathbf{R}) = \sigma_I^2 \beta\beta' + \Omega = \Phi. \tag{9.13}$$

Applying the non-zero covariance assumption to the portfolio $P = W'R$ we now have

$$E(P) = \mathbf{W}'(\alpha + \beta\mu_I) = \mu_p \tag{9.14}$$

and

$$var(P) = \mathbf{W}'(\sigma_I^2 \beta\beta' + \Omega)\mathbf{W} \tag{9.15}$$

$$= \mathbf{W}'\Phi\mathbf{W} = \sigma_p^2 \tag{9.16}$$

and we maximise

$$Z = \phi\mu_p - \sigma_p^2, \tag{9.17}$$

subject to the following constraint

$$\sum_{i=1}^{p} w_i = 1$$

plus any other equality or inequality constraints and bounds which the portfolio manager feels to be appropriate.

As before, this is a *QP* problem and can be solved by using the simultaneous set of equations generated by (9.1) with Φ replacing Σ.

The quantities that need to be estimated are:

$$\mu_I,\ \sigma_I^2,\ \alpha, \beta \text{ and } \Omega.$$

In order to estimate $\boldsymbol{\Omega}$ let

$$\hat{\mathbf{E}} = \begin{pmatrix} \hat{e}_{11} & \hat{e}_{12} & \cdots & \hat{e}_{1N} \\ \hat{e}_{21} & \hat{e}_{22} & \cdots & \hat{e}_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ \hat{e}_{p1} & \cdots & \cdots & \hat{e}_{pN} \end{pmatrix},$$

then we get that

$$\hat{\boldsymbol{\Omega}} = \frac{1}{N-2}\hat{\mathbf{E}}\hat{\mathbf{E}}'$$

and

$$\boldsymbol{\Phi} = \sigma_I^2 \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}' + \hat{\boldsymbol{\Omega}}$$

Chapter 10

Data Used in Analysis

The following shares were used in the analysis:

- Anglos
- JD Group
- Pick 'n Pay
- Remgro
- SAPPI
- SA-Eagle
- SASOL
- Tiger Brands
- Tongaat

There are a number of reasons for choosing these shares. The shares represent a cross section of the companies listed on the JSE. The shares posses a high market capitalisation and do not suffer from thin trading. Thin trading can have a negative effect on the estimation of the matrices, in that the matrices become singular and cannot therefore be inverted. Inverting the matrices is a vital step in determining both the efficient frontier and the optimal portfolio in the Markowitz portfolio theory.

For the Sharpe Single Index model the JSE All Share index is used as the market proxy.

The data runs from the end of July 1988 to the end of July 2004. Month end prices and index values were collected for this period. This period yielded a 193 data points from which a 192 log returns are calculated. These log returns are calculated using equation (7.1).

Chapter 11

Case Study

11.1 Introduction

In the following sections the robust methods described in the theory will be applied to portfolio selection. Programs developed by C.G. Troskie are used to determine both the Markowitz (MARKO Program) and Sharpe (SHARP Program) efficient frontier. A portfolio, to the value of R100,000.00, is constructed using the Minimum Volume Ellipsoid (MVE), Minimum Covariance Determinant (MCD) and T-Distribution (TDIST) robust estimation techniques to determine the covariance matrix. These matrices are then used as the covariance input into the Markowitz portfolio. The MARKO program was used to determine the efficient frontiers and the optimal portfolios. As a requirement a minimum of four shares had to be invested in. Therefore, if an optimal portfolio only contained three shares an additional portfolio was chosen which contained four shares. The performance of these portfolios are then analysed after a period of time. After the optimal portfolios have been determined and evaluated, the process was repeated to see how the results differ.

As a reference point, the optimal portfolio was determined using the classical covariance matrix as the input into the Markowitz portfolio. In all of the case studies bounds are imposed on the weights of the shares. The extent of these bounds will be discussed in each of the sections.

In the Sharpe portfolio analysis the inputs required are determined using the ARCH, GARCH and EGARCH methods to determine the inputs for the program. The bounds are discussed in the case study section.

11.2 Case Study 1

In this, the first case study, the data analysed runs from the end of July 1988 to the end of July 2003. It is assumed that an interest rate of 9% per annum could be obtained in a money market account during this time period. The weights have been bounded between zero and one in this case study, i.e. $0 \leq w_i \leq 1$. In Table 11.1 below the various investment percentages suggested for the different portfolios are shown.

Table 11.1

	Anglos	JD Group	Pick 'n Pay	Remgro	SA-Eagle
MVE Opt	0.00%	0.00%	7.60%	30.19%	37.39%
MVE	0.00%	0.00%	10.02%	17.68%	38.40%
MCD Opt	0.00%	0.61%	0.00%	0.00%	56.14%
MCD	0.00%	4.57%	2.81%	0.00%	50.11%
TDIST Opt	0.00%	0.00%	0.00%	24.88%	41.73%
TDIST	0.00%	0.00%	4.95%	18.41%	39.53%
Normal	0.00%	0.00%	9.04%	49.36%	23.83%

	SAPPI	SASOL	Tiger Brands	Tongaat
MVE Opt	0.00%	10.44%	14.38%	0.00%
MVE	0.00%	14.75%	17.74%	1.42%
MCD Opt	0.00%	18.10%	25.15%	0.00%
MCD	0.00%	17.53%	24.98%	0.00%
TDIST Opt	0.00%	17.17%	16.22%	0.00%
TDIST	0.00%	16.82%	20.29%	0.00%
Normal	0.00%	15.69%	2.08%	0.00%

From table 11.1 it can be seen that none of the portfolios invested in Anglos or SAPPI. Furthermore it can be seen that the robust portfolios invested heavily in SA-Eagle while the Normal estimation technique invested heavily in Remgro. Another difference in investment strategies are the amounts invested in Tiger Brands. The robust methods invested between seven and ten times more in this stock that the Normal portfolio. The performance of the portfolios suggested by the MARKO program is evaluated after an investment period of one year.

In table 11.2 the reduction in the minimum variance of the different portfolios (the measure of risk) is tabulated.

Table 11.2

	MVE	MCD	TDIST	Normal
Variance	0.04939	0.04827	0.04674	0.05867

The reduction in the variance can better be observed in figure 11.1. This figure illustrates the movement of the efficient frontiers to the left, which would indicate that the risk associated to the various port folios have decreased when being compared to the standard estimation technique. The graph illustrates that the MCD and TDIST methods decrease the risk in the portfolio the most.

In Table 11.3 the expected returns versus the achieved returns for the six month period in question are tabulated.

Table 11.3

	MVE Opt	MVE	MCD Opt	MCD	TDIST Opt	TDIST	Normal
Expected Return	16.98%	16.49%	15.24%	14.98%	14.93%	14.90%	16.08%
Realised Return	32.88%	34.24%	40.67%	40.24%	34.83%	34.45%	28.15%

It is evident from table 11.3 that all of the portfolios achieved a much higher realised rate of return, over the one year period, than was forecast by the models. It would appear however that the MVE portfolio did not perform as well as the rest of the portfolios, since it had the highest expected return but achieved the lowest realised return.

This procedure was repeated again in an attempt to duplicate the results. In tables 11.4 and 11.5 the different investment weights (as a percentage) are displayed and the expected returns versus the realised returns are tabulated respectively.

Figure 11.1 **Efficient Frontiers**

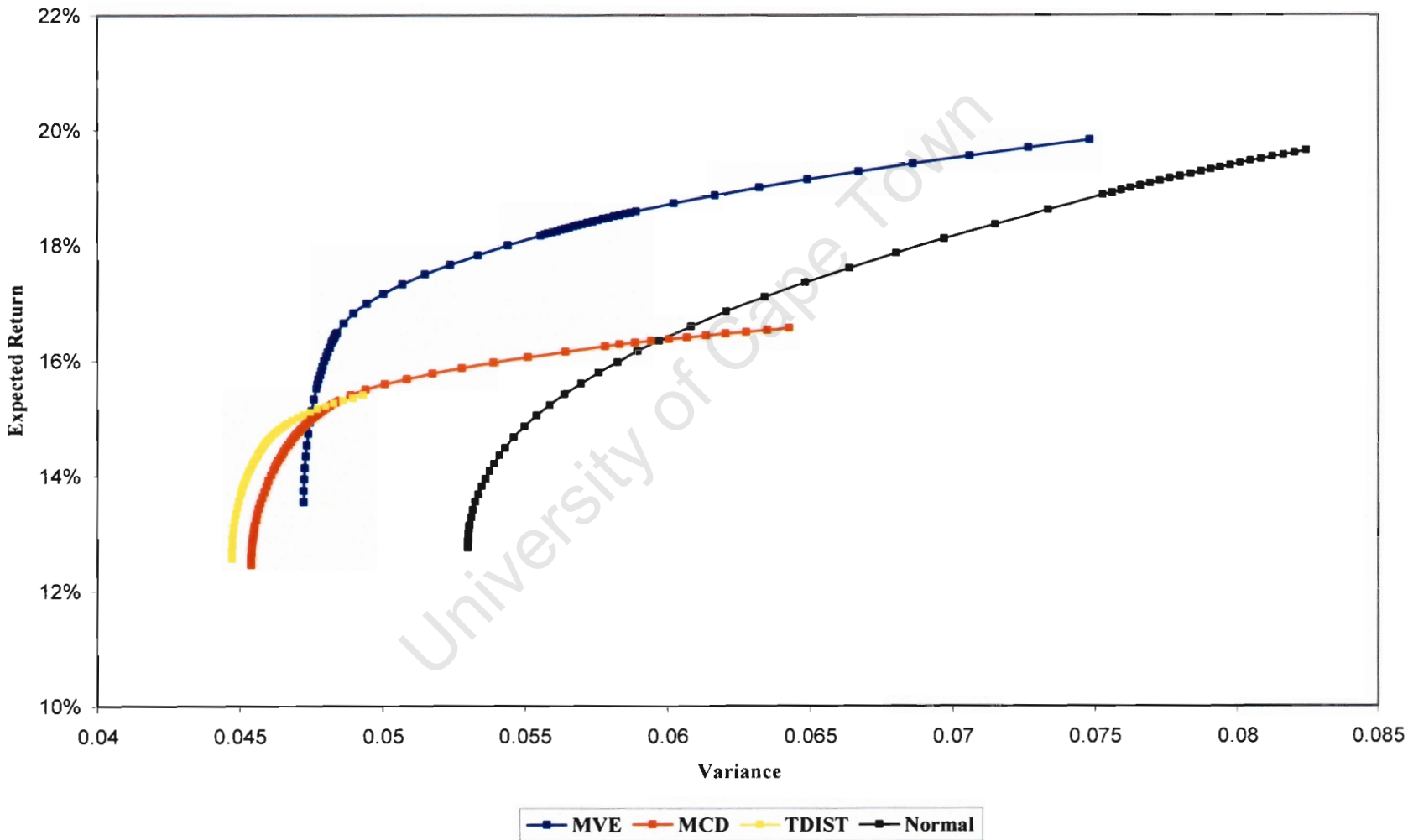


Table 11.4

	Anglos	JD Group	Pick 'n Pay	Remgro	SA-Eagle
MVE Opt	0.00%	0.00%	0.00%	0.00%	66.26%
MVE	0.00%	0.00%	4.33%	0.00%	57.91%
MCD Opt	0.00%	3.56%	0.00%	0.00%	61.76%
MCD	0.00%	6.37%	6.54%	0.00%	50.22%
TDIST Opt	0.00%	0.00%	0.00%	24.88%	41.73%
TDIST	0.00%	0.00%	4.95%	18.41%	39.53%
Normal	0.00%	0.00%	9.04%	49.36%	23.83%

	SAPPI	SASOL	Tiger Brands	Tongaat
MVE Opt	0.00%	25.08%	8.66%	0.00%
MVE	0.00%	21.37%	16.40%	0.00%
MCD Opt	0.00%	10.33%	24.34%	0.00%
MCD	0.00%	13.66%	23.21%	0.00%
TDIST Opt	0.00%	17.17%	16.22%	0.00%
TDIST	0.00%	16.82%	20.29%	0.00%
Normal	0.00%	15.69%	2.08%	0.00%

Table 11.5

	MVE Opt	MVE	MCD Opt	MCD	TDIST Opt	TDIST	Normal
Expected Return	17.70%	17.23%	15.20%	15.17%	14.93%	14.90%	16.08%
Realised Return	43.95%	41.20%	42.99%	40.71%	34.83%	34.45%	28.15%

As in the first run, the MVE, MCD and TDIST portfolios recommend investing in SA-Eagle while the Normal portfolio once again invested heavily in Remgro. In the second run the MCD, TDIST and Normal portfolios suggested more or less the same investment weights as in the first run. In the second run the MVE portfolio again has the highest expected return, but unlike the first run it has some of the highest realised returns. The investment weights for the MVE portfolio drastically change however. The sets of different weights are tabulated in table 11.6.

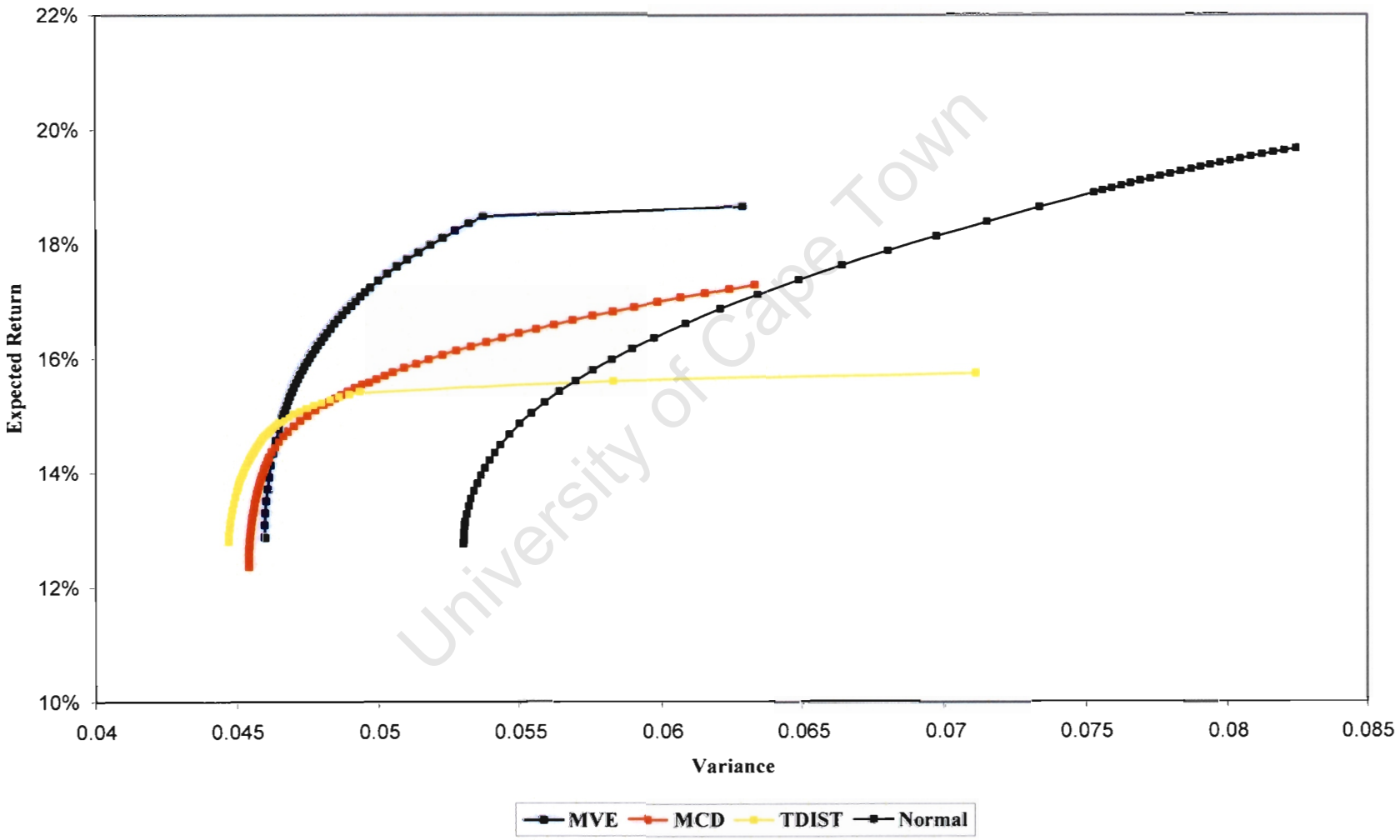
Table 11.6

	Anglos	JD Group	Pick 'n Pay	Remgro	SA-Eagle
MVE Opt 1	0.00%	0.00%	7.60%	30.19%	37.39%
MVE 1	0.00%	0.00%	10.02%	17.68%	38.40%
MVE Opt 2	0.00%	0.00%	0.00%	0.00%	66.26%
MVE 2	0.00%	0.00%	4.33%	0.00%	57.91%

	SAPPI	SASOL	Tiger Brands	Tongaat
MVE Opt 1	0.00%	10.44%	14.38%	0.00%
MVE 1	0.00%	14.75%	17.74%	1.42%
MVE Opt 2	0.00%	25.08%	8.66%	0.00%
MVE 2	0.00%	21.37%	16.40%	0.00%

This change in the investment weights can be contributed to the fact that, as stated in an earlier chapter, the method used to determine the MVE is only an approximation, since no closed form solution exists. This has the effect that this approximation of the MVE provides local minima for the objective function but not an absolute minimum. This change is best observed in figure 11.2, which illustrates the new efficient frontiers, constructed after the second run.

Figure 11.2 **Efficient Frontiers**



11.3 Case Study 2

In this case study the data runs from the end of July 1988 to end of July 2004. Unlike in the first case study, the various portfolios are evaluated after a six month period. It is assumed that an interest rate of 7% per annum can be obtained in a money market account during this time period. As in the first case study, the weights have been bounded between zero and one, i.e. $0 \leq w_i \leq 1$.

In table 11.7 below the various investment percentages suggested for the different portfolios are shown.

Table 11.7

	Anglos	JD Group	Pick 'n Pay	Remgro	SA-Eagle
MVE Opt	0.00%	0.00%	0.00%	0.00%	39.98%
MVE	0.00%	0.00%	0.00%	5.78%	38.89%
MCD Opt	0.00%	2.13%	0.00%	0.00%	64.58%
MCD	0.00%	6.94%	4.44%	0.00%	47.05%
TDIST Opt	0.00%	0.00%	0.00%	24.64%	51.10%
TDIST	0.00%	0.00%	4.13%	18.13%	40.51%
Normal	0.00%	0.52%	6.44%	57.74%	20.90%

	SAPPI	SASOL	Tiger Brands	Tongaat
MVE Opt	0.00%	27.30%	32.73%	0.00%
MVE	0.00%	22.89%	32.43%	0.00%
MCD Opt	0.00%	8.41%	24.88%	0.00%
MCD	0.00%	14.95%	26.62%	0.00%
TDIST Opt	0.00%	15.41%	8.85%	0.00%
TDIST	0.00%	15.90%	21.33%	0.00%
Normal	0.00%	14.39%	0.00%	0.00%

As in the first case study, the robust portfolios invested heavily in SA-Eagle while the Normal portfolio again invested heavily in Remgro. As before, the Normal method again invested a small percentage in Tiger Brands while the robust methods invest a substantial amount in the stock. In table 11.8 the reduction in the variance of the portfolios is tabulated.

Figure 11.3 **Efficient Frontiers**

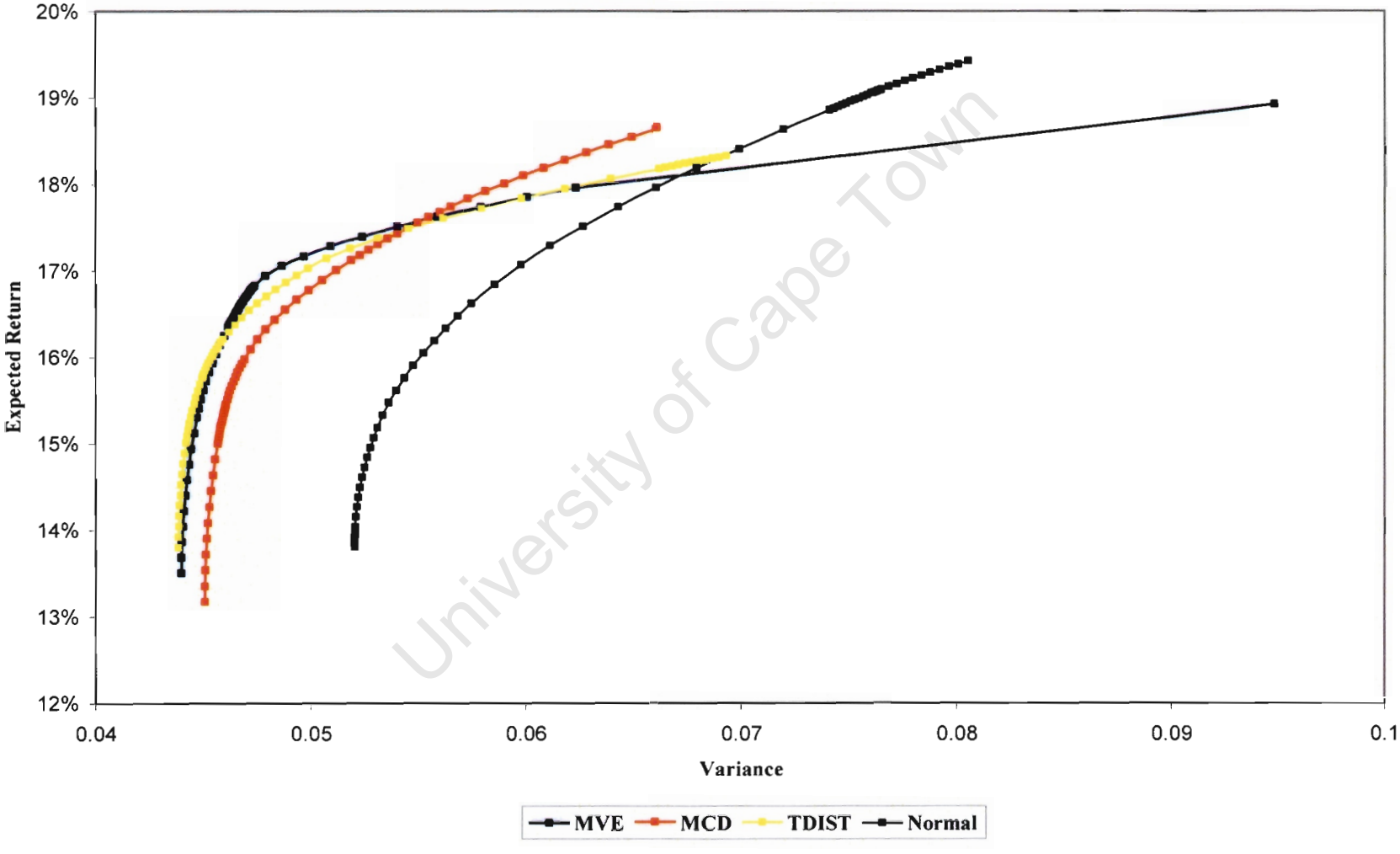


Table 11.8

	MVE	MCD	TDIST	Normal
Variance	0.04759	0.04974	0.0475	0.05988

The reduction in the variance is illustrated in figure 11.3. This figure illustrates the movement of the efficient frontiers to the left, which indicates that the risk associated to the various port folios have decreased when being compared to the standard estimation technique. The graph illustrates that the MVE and TDIST methods decrease the risk in the portfolio the most.

In table 11.9 the expected returns versus the achieved returns for the six month period in question are tabulated.

Table 11.9

	MVE Opt	MVE	MCD Opt	MCD	TDIST Opt	TDIST	Normal
Expected Return	8.49%	8.24%	7.62%	7.49%	7.46%	7.45%	8.54%
Realised Return	22.66%	22.76%	29.26%	27.37%	28.24%	25.25%	25.27%

From table 11.9 it can be seen that once again all of the portfolios achieved a much higher rate of return over the six month period than was forecast by the models. Although, it would appear that the MVE portfolio, again, did not perform as well as well as the rest of the portfolios in the first run.

As before, the estimation procedure was repeated in an attempt to duplicate the results. In table 11.10 the different investment weights (as a percentage) are displayed and in table 11.11 the expected returns versus the realised returns are tabulated.

Table 11.10

	Anglos	JD Group	Pick 'n Pay	Remgro	SA-Eagle
MVE Opt	0.00%	0.00%	0.00%	0.00%	68.53%
MVE	0.00%	0.00%	0.00%	0.00%	54.10%
MCD Opt	0.00%	2.01%	0.00%	0.00%	60.28%
MCD	0.00%	6.59%	6.40%	0.00%	46.42%
TDIST Opt	0.00%	0.00%	0.00%	24.64%	51.10%
TDIST	0.00%	0.00%	4.13%	18.13%	40.51%
Normal	0.00%	0.52%	6.44%	57.74%	20.90%

	SAPPI	SASOL	Tiger Brands	Tongaat
MVE Opt	0.00%	22.54%	8.93%	0.00%
MVE	0.00%	19.89%	23.19%	2.82%
MCD Opt	0.00%	6.69%	31.02%	0.00%
MCD	0.00%	14.31%	26.27%	0.00%
TDIST Opt	0.00%	15.41%	8.85%	0.00%
TDIST	0.00%	15.90%	21.33%	0.00%
Normal	0.00%	14.39%	0.00%	0.00%

Table 11.11

	MVE Opt	MVE	MCD Opt	MCD	TDIST Opt	TDIST	Normal
Expected Return	10.64%	10.11%	8.49%	8.26%	8.31%	8.11%	8.54%
Realised Return	30.62%	27.02%	27.84%	27.43%	28.24%	25.25%	25.27%

As in the previous case study, results of the second run indicate that the MVE method again has a higher expected return and like before, possesses the highest realised return. In table 11.12 the different investment weights, for the MVE portfolio, for the first and the second runs are tabulated. The difference can once again be attributed to the fact the values obtained for the MVE method are approximations.

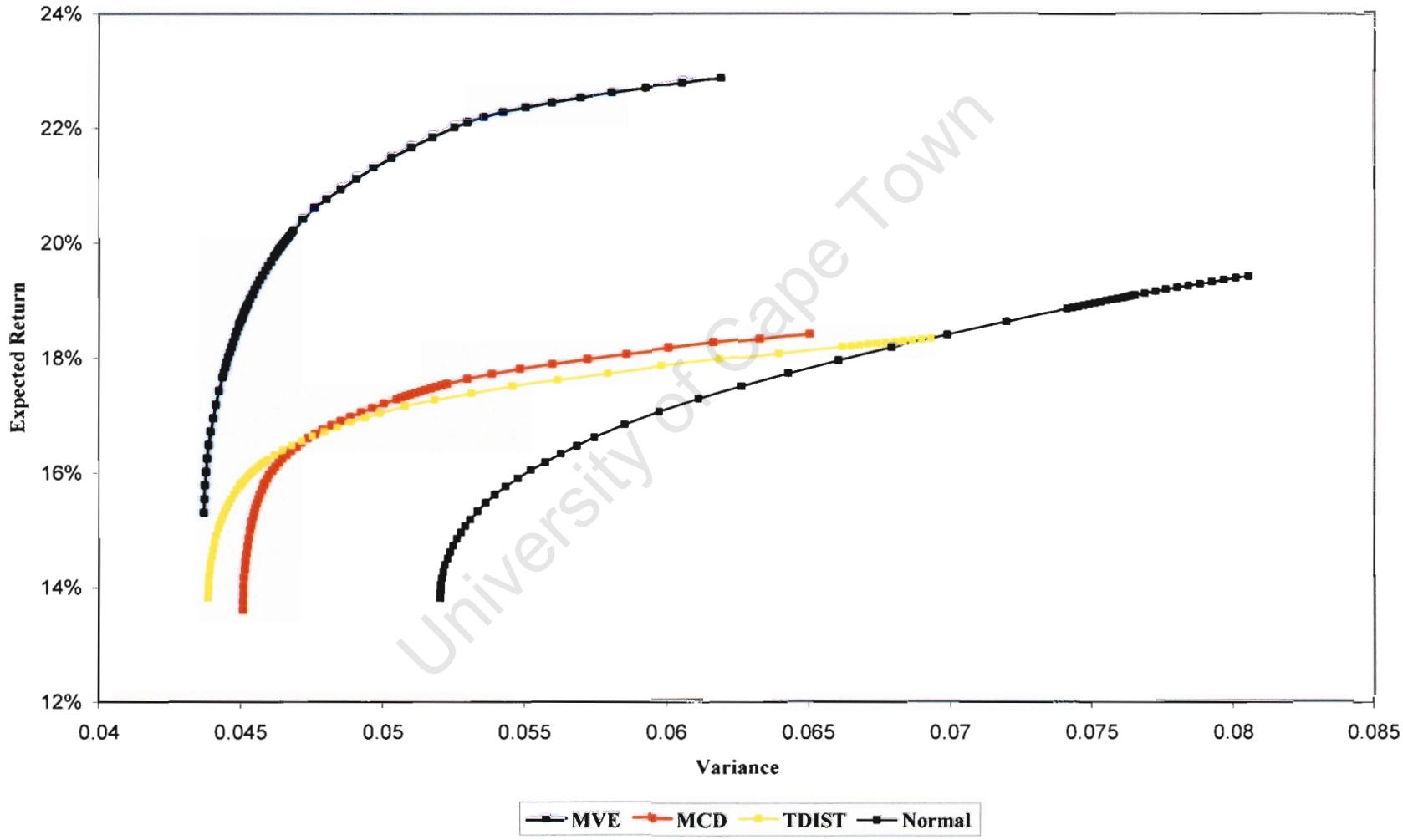
Table 11.12

	Anglos	JD Group	Pick 'n Pay	Remgro	SA-Eagle
MVE Opt 1	0.00%	0.00%	0.00%	0.00%	39.98%
MVE 1	0.00%	0.00%	0.00%	5.78%	38.89%
MVE Opt 2	0.00%	0.00%	0.00%	0.00%	68.53%
MVE 2	0.00%	0.00%	0.00%	0.00%	54.10%

	SAPPI	SASOL	Tiger Brands	Tongaat
MVE Opt 1	0.00%	27.30%	32.73%	0.00%
MVE 1	0.00%	22.89%	32.43%	0.00%
MVE Opt 2	0.00%	22.54%	8.93%	0.00%
MVE 2	0.00%	19.89%	23.19%	2.82%

Figure 11.4 illustrates the new efficient frontiers that were constructed after the second run. The different location of the efficient frontier can clearly be seen when comparing figure 11.4 to figure 11.3.

Figure 11.4 **Efficient Frontiers**



11.4 Case study 3

As in case study 2, the whole data set is utilised in this study. Again it is assumed that a return of 7% per annum is attainable in a money market. The difference is however in the bounds set on the portfolio. In this case study short sales are allowed. The weights are defined as $-0.25 \leq w_i \leq 0.75$.

In table 13 the suggested investment weights are tabulated.

Table 11.13

	Anglos	JD Group	Pick 'n Pay	Remgro	SA-Eagle
MVE	-3.96%	5.27%	5.91%	4.09%	37.48%
MCD	-5.39%	6.31%	7.82%	8.05%	38.94%
TDIST	-1.48%	3.14%	6.08%	11.40%	32.33%
Normal	-1.71%	0.46%	5.08%	18.27%	32.20%

	SAPPI	SASOL	Tiger Brands	Tongaat
MVE	8.33%	11.93%	22.30%	8.64%
MCD	5.21%	13.21%	19.58%	6.27%
TDIST	5.98%	11.56%	22.91%	8.08%
Normal	1.30%	14.15%	21.43%	8.83%

Each of the above portfolios is the optimal portfolio suggested by the MARKO program. The introduction of short sales has drastically changed the investment weights in the various portfolios. In the previous case studies none of the estimation techniques indicated that one should invest in all of the shares, where as in this case study investing in each of the shares is suggested.

In table 11.14 the expected and realised returns are tabulated.

Table 11.14

	MVE	MCD	TDIST	Normal
Expected Return	7.79%	7.07%	6.94%	6.90%
Realised Return	29.39%	32.86%	25.76%	27.16%

Once again the realised returns are much greater than the expected returns. This is, however, the lowest expected returns observed for each of the portfolios, in all of the case studies.

Table 11.15 tabulates the variances of the portfolio at the optimal investment point.

Table 11.15

	MVE	MCD Opt	TDIST	Normal
Variance	0.04364	0.04494	0.04382	0.05203

As in all of the previous cases, the robust estimation techniques decrease the variance in the portfolios. It is worth noting that the variances for the different portfolios are smaller than in the previous case studies. In the previous case studies none of the robust portfolios obtained a variance smaller than 0.0465, while the Normal portfolio never obtained a variance smaller than 0.0585.

11.5 Case Study 4

In this, the final case study, the full data set is utilised. The weights are bounded between zero and one, i.e. $0 \leq w_i \leq 1$.

In table 11.16 the various investment weights obtained using the SHARP program is tabulated. From table 11.16 it can be seen that all of the estimation techniques suggesed very similar investment weights. It is notable that, unlike in the first two case studies, the ARCH, GARCH and EGARCH estimation techniques assigned the highest investment weights to Remgro.

Table 11.16

	Anglos	JD Group	Pick 'n Pay	Remgro	SA-Eagle
ARCH 1	0.00%	9.68%	8.95%	69.96%	0.00%
ARCH 2	0.00%	8.26%	11.79%	56.57%	11.78%
GARCH 1	0.00%	9.00%	9.08%	69.02%	0.00%
GARCH 2	0.00%	7.66%	11.87%	55.88%	11.66%
EGARCH 1	0.00%	9.86%	10.91%	67.22%	0.00%
EGARCH 2	0.00%	8.60%	13.21%	55.82%	10.28%
Normal	0.00%	8.84%	7.21%	71.74%	0.00%

	SAPPI	SASOL	Tiger Brands	Tongaat
ARCH 1	0.00%	11.41%	0.00%	0.00%
ARCH 2	0.00%	11.60%	0.00%	0.00%
GARCH 1	0.00%	12.90%	0.00%	0.00%
GARCH 2	0.00%	12.93%	0.00%	0.00%
EGARCH 1	0.00%	12.01%	0.00%	0.00%
EGARCH 2	0.00%	12.08%	0.00%	0.00%
Normal	0.00%	12.21%	0.00%	0.00%

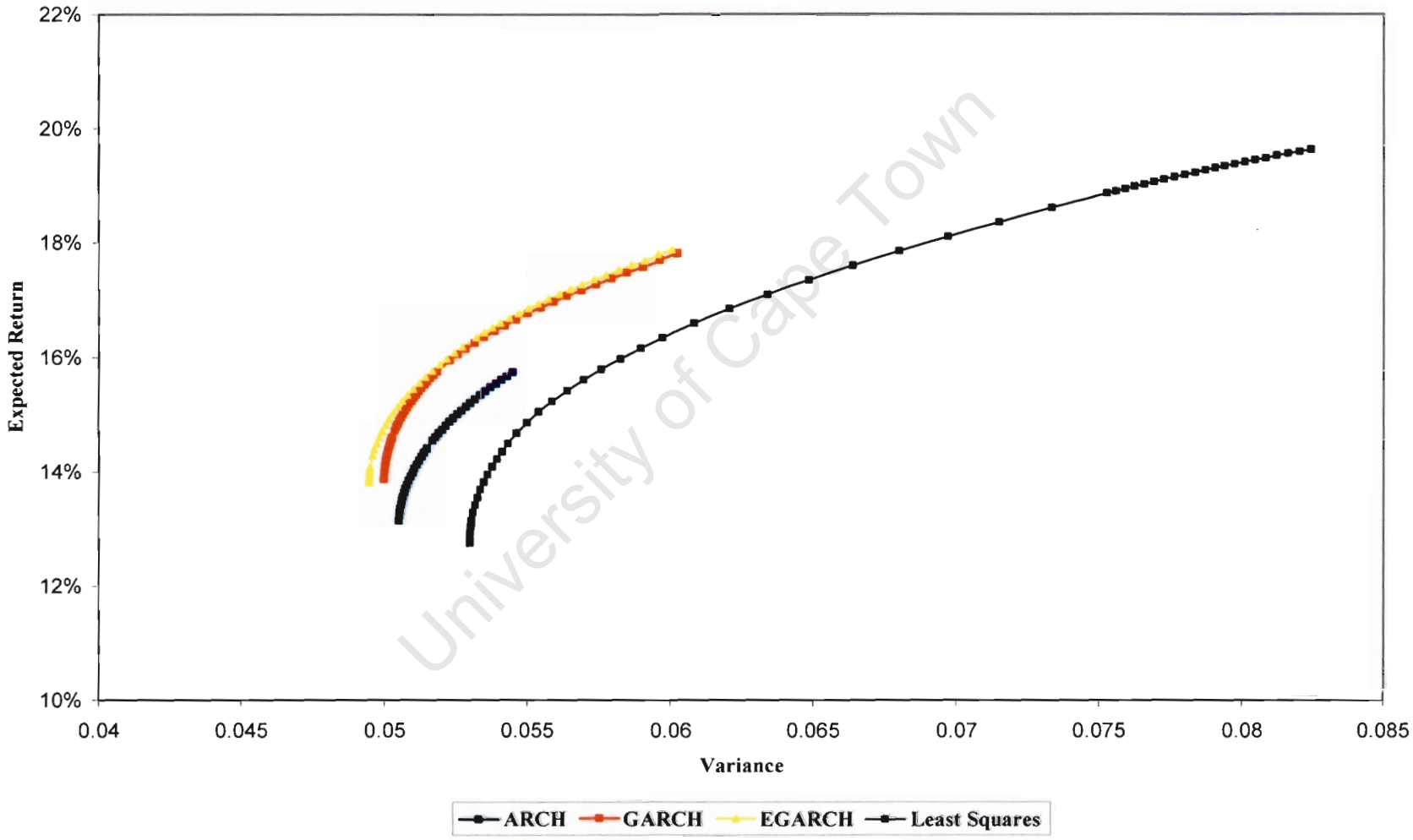
In table 11.17 the expected returns and the realised returns of the portfolios are listed.

Table 11.17

	ARCH 1	ARCH 2	GARCH 1	GARCH 2	EGARCH 1	EGARCH 2	Normal
Expected Return	9.70%	9.44%	9.68%	9.43%	9.70%	9.39%	9.70%
Realised Return	24.51%	26.31%	24.27%	26.08%	24.70%	26.24%	24.11%

As with the previous case studies, the realised returns are greater than the expected returns. In figure 11.5 the shift in the efficient frontiers can be seen.

Figure 11.5 **Efficient Frontiers**



Chapter 12

Conclusions

In all of the case studies it can be observed that making use of the robust estimators, to determine the input parameters in the Markowitz and Sharpe models, a significant reduction in the risk associated with the portfolios can be observed. These reductions are graphically illustrated in chapter 11. Even though in some of the case studies the robust procedures did not predict a higher expected return than the normal estimation procedures, in all of the case studies the robust procedures realised higher actual returns on the portfolios suggested. In some of the cases the robust procedures outperformed the normal estimation procedure by nearly 15% over the given time period.

The results obtained justify the usage of robust estimators in estimating the input parameters for the different portfolio models. There are however, areas that need improving. One such area is the estimating of the MVE estimator. Even though the MVE estimation technique produced some of the highest returns, it also produced some of the lowest returns. The disparity in the values obtained can be attributed to the fact that the values obtained for the MVE estimator are only approximations. Another area that needs development, is the algorithm used to determine the MCD estimator. In the case studies only nine shares were used and the MCD took a considerable greater amount of time to determine.

A final consideration that needs to be addressed is the variance obtained in the model. Due to the downweighting of outliers, the usage of robust estimation techniques in the Markowitz and Sharp portfolio models can lead to the model ultimately underestimating the portfolio risk. Another problem could be that the models underestimate the expected return. This could pose a problem if the expected return is used as a selection criterion, when deciding which portfolios to use.

References

Agulló, J. (2001), "New Algorithms For Computing the Least Trimmed Squares Regression Estimator", *Computational Statistics and Data Analysis*, 36, 425-439.

Birch, J. B. and Agard, D. B. (1993), "Robust inference in regression: A comparative study", *Communications in Statistics,* 22(1), 217-244.

Black, F. (1976). "Studies in stock price volatility changes", *Proceedings of The American StatisticalAssociation, Business and Economic Statistics Section*, pp.177-181

Butler, R. W., Davies, P. L. and Jhun, M. (1993), "Asymptotics for the minimum covariance determinant estimator", *Annals of Statistics*, 21, 1385-1400.

Bollerslev, T. (1986), "Generalised Autoregressive Conditional Heteroskedasticity", *Journal of Econometrics*, 31, 307-327.

Coakley, C. W. and Hettmansperger, T. P. (1993), "Exact iterative computation of the robust multivariate minimum volume ellipsoid estimator", *Statistics and Probability Letters*, 16, 213-218.

Cook, R. D., Hawkins, D. M. and Weisberg, S. (1993), "Exact iterative computation of the robust multivariate minimum volume ellipsoid estimator", *Statistics and Probability Letters,* 16, 213-218.

Christie, A. A. (1982), "The stochastic behaviour of common stock variances – value, leverage and interest rates effects", *Journal of Financial Economics,* 10, 407-432

Davies, P. L. (1992), "The asumptotics of Rousseeuw's minimum volume ellipsoid estimator", *The Annals of Statistics*, 20, 495-507.

Donoho, D. L. (1982), "Breakdown properties of multivariate location estimators", qualifying paper, Harvard University, Boston, MA.

Engle, R. F. (1982), " Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of U.K. Inflation", *Econometrics*, 50, 987-1008

Hadi, A. S., (1992), "Identifying multiple outliers in multivariate data", *Journal of the Royal Statistical Society,* Ser. B, 54, 761-771.

Hampel, F. R. (1975), "Beyond location parameters: robust concepts and methods", *In proceedings of the 40th Sessions of the ISI,* 46, 375-391.

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

Hawkins, D. M. (1993), "A feasible solution algorithm for the minimum volume ellipsoid estimator in multivariate data", *Computational Statistics and Data Analysis*, 8, 95-107.

Hawkins, D. M. (1994), "A feasible solution algorithm for the minimum covariance determinant estimator", *Computational Statistics and Data Analysis*, 17, 197-210.

Hawkins, D. M., Bradu, D. and Kass, G. V. (1984), "Location of Several outliers in multiple regression data using elemental subsets", *Technometrics*, 26, 197-208.

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

Hettmansperger, T. P. and Sheather, S. J. (1992), "A cautionary note on the method of least median of squares", *The American Statistician*, 46, 79-83.

Johnson, R. A. and Wichern, D. W. (1992), *Applied Multivariate Statistical Analysis*, Englewood Cliffs, N.J.:Prentice-Hall Inc.

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

Lopuhaa, H. P. (1992), "Highly efficient estimators of multivariate location with high breakdown point", *Annals of Statistics*, 20, 398-413.

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

Marazzi, A. (1993), *Algorithms Routines and S Functions for Robust Statistics*, Wadsworth and Brooks/Cole.

Markowitz, H. M. (1952), "Portfolio Selection", *Journal of Finance*, 7, 77-91.

Morrison, D. F. (1990), *Multivariate Statistical Methods,* New York:McGraw-Hill.

Nelson, D. B. (1991) "Conditional Heteroskedasticity in Asset Returns: A New Approach," *Econometrica*, 59, 347–370.

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

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

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

Rousseeuw, P. J. van Zomeren, B. C. (1990), "Unmasking multivariate outliers and leverage points", *Journal of the American Statistical Association*, 85, 633 -639.

Rousseeuw, P. J. (1993), "A resampling design for computing high-breakdown regression", *Statistics and Probability Letters*, 18, 125-128.

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

Ruiz-Gazen, A. (1996), "A very simple robust estimator of a dispersion matrix", *Computational Statistics and Data Analysis*, 21, 149-162.

Sharpe, W. F. (1963), "A simplified model for portfolio analysis", *Management Sceince*, 9, 277-93.

Simpson, D. G., Ruppert, D. and Carroll, R. J. (1992), "On one-step estimates and stability of inferences in linear regression", *Journal of the American Statistical Association,* 87, 439-450.

Stahel, W. A. (1981), "Robuste Schatzungen: Infinitesimale Optimalitat und Schatzungen von Kovarianzmatrizen", Ph.D. Thesis, ETH Zurich, Switzerland.

Venables, W. N. and Ripley, B. D. (1999), *Modern Applied Statistics with S-PLUS,* Third Edition. Springer.