SEQUENTIAL APPLICATION OF MULTIVARIATE OUTLIER TEST:
A ROBUST APPROACH

By
Sanjoy Kumar Sinha

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE AT DALHOUSIE UNIVERSITY HALIFAX, NOVA SCOTIA JULY 1997

© Copyright by Sanjoy Kumar Sinha, 1997
The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L’auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L’auteur conserve la propriété du droit d’auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.
To my Father
and
to the Memory of
my Mother
Contents

Chapter 1  Identification of Outliers ........................................ 1
  1.1 Introduction .................................................................. 1
  1.2 The Nature and Origin of Outliers ................................. 5
    1.2.1 What is an Outlier? ........................................... 5
    1.2.2 Origin of Outliers ............................................. 6
    1.2.3 Procedures for Handling Outliers ......................... 7

Chapter 2  Reviews of Relevant Literature on Outliers ................. 11
  2.1 Early Approaches ...................................................... 11
  2.2 Outliers in Simple Normal Samples .............................. 14
  2.3 Outliers in Multivariate Normal Samples ..................... 20

Chapter 3  Sequential Procedure for Testing Multivariate Outliers .... 31
  3.1 Identification of Multivariate Outliers ......................... 31
  3.2 Robust Estimation of Multivariate Location and Scale ...... 37
  3.3 Application of the Sequential Test ............................... 40

Chapter 4  Applications ..................................................... 46
  4.1 Computational Issues ............................................... 46
  4.2 Simulation Results .................................................. 47
  4.3 Some Examples ...................................................... 50

Discussion ........................................................................... 63

Appendix ............................................................................. 64

References ........................................................................... 65
Abstract

Identification of outliers in multivariate data is not trivial, especially when there exists several outliers in the data. The classical identification method based on the sample mean and sample covariance matrix cannot always find them, because the classical mean and covariance matrix are themselves affected by outliers. This problem is termed as masking, because the outliers get masked by each other. To avoid the masking effect, robust estimates of mean and covariance are suggested by many authors.

This thesis deals with the problem of identifying and testing a set of a number \( k \) of extreme sample points as significant outliers in a sample of size \( n \) drawn from a \( p \)-dimensional normal distribution with unknown parameters. A robust sequential procedure is suggested for the identification of multiple outliers in multivariate normal data. Chapter 1 gives a brief idea about outliers in statistical data. In Chapter 2 we review the rejection techniques for single and multiple outliers suggested by different authors. We also review some recent papers which deal with the problem of accommodation and identification of multiple outliers using robust procedures. In Chapter 3 we propose a test criterion for the detection of multivariate outliers based on the robust estimates of mean and covariance. Chapter 4 deals with the application of our proposed test procedure with different examples. A simulation study is carried out to support the good behaviour of the proposed sequential test when the data are multivariate normal. We also study the performance of the proposed sequential test in the presence of outliers.
Acknowledgements

It is, of course, a privilege and pleasure to acknowledge the help of others. I thank Dr. Chris Field and Dr. Lise Manchester for their valuable supervision of this thesis. I am also grateful to Dr. Brajendra Sutradhar of Memorial University of Newfoundland for helpful comments and suggestions on this thesis.
CHAPTER 1

Identification of Outliers

1.1 Introduction

There has been a concern about unrepresentative or outlying observations in sets of statistical data for many years. Comments by Bernoulli (1777) indicate that the practice of identifying extreme observations was commonplace 200 years ago. The history of the treatment of such observations is traced from these early comments to the present. Methods for the detection of outliers, sometimes rejection of them to restore the property of the data, or at least adopting methods of reducing their impact in statistical analysis have been studied by many authors from the earliest times to the present. Much of the relevant literature is surveyed by Hawkins (1980), Barnett and Lewis (1994). Other works which contain substantial review material are Rider (1933), Anscombe (1960), Grubbs (1969), Stigler (1973, 1975), Barnett (1978), Kale (1979), Prescott (1980), Beckman and Cook (1983).

Detection of outliers in univariate data is an old but important subject in statistical literature. Many authors have contributed to this area. Some of the important works are Grubbs (1950), Ferguson (1961), Tietjen and Moore (1972), Rosner (1975). However, much less work has been done on multivariate outliers. This may be due to
the fact that outliers in a multivariate point cloud can be hard to detect, especially when the dimension is greater than 2, because then we can no longer rely on visual inspection. Most of the proposed methods for the detection of multivariate outliers are graphical and informal, such as Healy (1968), Andrews (1972), Gnanadesikan and Kettenring (1972), Rohlf (1975), Bacon-Shone and Fung (1987). Wilks (1963) proposed a test which is based on the change of the determinant of the sample scatter matrix after some observations are eliminated from the sample. He defines $A$ as the matrix of sums of squares and cross-products of the sample observations drawn from a multivariate normal distribution with unknown parameters and calls $|A|$ the *internal scatter* of the sample and suggests that a sensible criterion for the declaration of an outlier is to choose that sample point whose omission leads to the least value for the so-called *one-outlier scatter ratio*

$$R_j = \frac{|A^{(j)}|}{|A|},$$

where $A^{(j)}$ is the sample scatter matrix without observation $j$. Wilks shows that $R_j$ are identically distributed as beta($\frac{n-p-1}{2}, \frac{p}{2}$) with a joint distribution symmetric over $\mathbb{R}^n$ subject to

$$\sum_{j=1}^{n} R_j = n \left(1 - \frac{p}{n-1}\right),$$

$$0 \leq R_j \leq 1 \quad (j = 1, \ldots, n).$$

A generalization of Wilks’ single-outlier test suitable for application to the many-outlier problem of detecting from 1 to $k$ outliers in a multivariate data is proposed by Caroni and Prescott (1992) following the approach proposed by Rosner (1983) for univariate samples.
Another classical method for the detection of multivariate outliers is to find the squared Mahalanobis distance

$$D_i = (x_i - \bar{x})'S^{-1}(x_i - \bar{x})$$

for each point $x_i (i = 1, \ldots, n)$. That observation $x_i$ which yields the maximum value $D_{(n)}$ is then a candidate for declaration as an outlier provided $D_{(n)}$ is unreasonably large in relation to the distribution of $D_{(n)}$ under the basic model. But it is well-known that the distance $D_i$ suffers from masking effect (Rousseeuw and van Zomeren 1990, Hadi 1992, Atkinson 1994, Rocke and Woodruff 1996). This is due to the fact that the sample mean $\bar{x}$ and sample covariance matrix $S$ are not robust. Atypical multivariate observations will tend to inflate variances. This will in general decrease the Mahalanobis distances for the outlying observations and distort the rest of the plot (Campbell 1980). Therefore, it is important to replace sample mean and sample covariance by robust estimates.

For the detection of multivariate outliers, Campbell (1980) suggests finding the Mahalanobis distances using robust $M$-estimates of means and of covariances, rather than the usual maximum likelihood estimates. If there is no outlier in the data, then robust $M$-estimates are similar to the usual estimates. Maronna (1976) developed an algorithm for calculating robust $M$-estimates of multivariate location and scatter. He also exhibits some small-sample Monte Carlo results on the properties of the estimators.

Rousseeuw and van Zomeren (1990) suggest the minimum volume ellipsoid ($MVE$) estimators of location and scatter introduced by Rousseeuw (1985). For location, they
use the center of the minimum volume ellipsoid covering half of the sample observations and scatter matrix is also determined from the same ellipsoid (multiplied by a correction factor to obtain consistency at multinormal distributions). They reject the $M$-estimates of Campbell (1980) in view of the low breakdown point and claim that $MVE$ estimators have a very high breakdown point which is almost one-half, that is, the $MVE$ estimators cannot be changed by an arbitrary amount without changing about half of the data.

A detailed study on the problem of identifying outliers in multivariate data is given in Rocke and Woodruff (1996). They describe significant improvements in methods for the detection of outliers, and from extensive simulation experiments, they claim that their hybrid algorithm proves as a package to be superior to other methods suggested for multivariate outlier identification. As they claim, the hybrid method can deal with even heavily contaminated data in high dimensions.

In their papers, Rousseeuw and van Zomeren (1990), Rocke and Woodruff (1996) suggest the robust Mahalanobis distance as test criterion for the detection of outliers. They assume that the robust Mahalanobis distance is asymptotically distributed as Chi-square with certain degrees of freedom, and use the percentage points of Chi-square variate as critical values for identifying outliers. But the Mahalanobis distances from a robust estimator of location and shape are very poorly approximated by Chi-square variates (comment by Rocke, through personal communication). Simulations reveal that use of Chi-square percentage points may lead to a very large number of erroneous rejections unless the sample size is very large. Rocke and Woodruff (1996) suggest some adjustments in the robust estimator of shape to overcome this difficulty.
Their approaches are reviewed later in some detail.

In this thesis, the proposed test statistic for identifying atypical observations is the maximum of the robust Mahalanobis distances, $\max_i RD_i$, where

$$RD_i = (x_i - \bar{x}^*)' S^{-1}(x_i - \bar{x}^*)$$

and $\bar{x}^*$ and $S^*$ are the robust estimators of location and shape, respectively. For the robust estimators of location and shape, Campbell's $M$-estimation method has been followed with some modifications. To perform the many-outlier test, a sequential application of the above test statistic is proposed.

1.2 The Nature and Origin of Outliers

1.2.1 What is an Outlier?

According to Barnett and Lewis (1994), an “outlier” is one that appears to deviate markedly from other members of the sample in which it occurs. Similar definition is pointed out by Beckman and Cook (1983): Observations that stand apart from the bulk of the data are termed as “outliers”, “discordant observations”, “contaminants”, “surprising values”, or “dirty data”. Investigators are concerned when such observations occur. An adequate method for handling such observations, or at least a firm understanding of the relative merits of available methods, is necessary.
1.2.2 Origin of Outliers

There are many ways in which outliers may arise. Hampel, Ronchetti, Rousseeuw and Stahel (1986) point out that many assumptions commonly made in statistics (such as normality, linearity, independence) are at most approximations to reality. The occurrence of gross errors, such as copying or keypunch errors, is the one reason for deviations from model assumptions. These errors usually show up as outliers, which are far away from the bulk of the data and are dangerous for many classical estimation procedures. Other reasons for deviations from idealized model assumptions include the empirical character of many models and the approximate character of many theoretical models. For example, when the central limit theorem is invoked, being a limit theorem, it can at most suggest approximate normality for real data.

Barnett and Lewis (1994), Anscombe (1960) arrange the causes of outliers in three different categories. These are inherent variability, measurement error, and execution error. Similar causes are pointed out by Beckman and Cook (1983). They arrange the causes of outliers into three broad and somewhat overlapping categories. These are global model weaknesses, local model weaknesses, and natural variability.

Outliers must be judged with some model, either implicit or explicit. From a statistical viewpoint, it is reasonable to think of the appearance of an outlier as being due to the inability of the model to provide an adequate fit or statistical explanation. Global model weaknesses may lead to the replacement of the present model with a new or revised model for the entire sample. Response variables in the wrong scale or frequently occurring outliers of a known nature may cause the global model weak-
nesses. The former cause may lead to a transformation of the response and the latter cause may lead to replacing the present model with a mixture.

Local model weaknesses are applied only on the outlying observations and not to the model as a whole. Isolated measurement and recording errors, and highly influential observations in regression due to remote points in the factor space may be considered as examples for local model weaknesses.

Finally, an outlier may be the result of natural variation over the population rather than any weakness of the model. It is uncontrollable and reflects the distributional properties of a correct basic model describing the generation of the data.

1.2.3 Procedures for Handling Outliers

Following different approaches to the processing of outliers, Barnett (1978), and Barnett and Lewis (1994) point out two distinct statistical methods. These are accommodation of outliers and discordancy tests.

In accommodation of outliers some statistical methods are designed to infer about the population from which a random sample has been drawn, and which will not be seriously distorted by the presence of outliers. That is, accommodation procedures are robust against the presence of outliers. In the general problem of statistical inference, robustness is a concept of much importance and is more specific than outlier study. In recent years major efforts have been made to obtain statistical techniques
which provide a protection against various types of uncertainty of knowledge of the data generating mechanism. These include robust methods for estimating or testing summary measures of the underlying distribution or more general inference procedures: where the estimation or test procedures retain desirable statistical properties over a range of different possible distributional forms. Huber asks, in his 1972 Wald Lecture Robust Statistics: A Review, 'What is a robust procedure?'. and goes on to say:

one never has a very accurate knowledge of the true underlying distribution; ... the performance of some of the classical tests or estimates is very unstable under small changes of the underlying distribution: ... some alternative tests or estimates ... lose very little efficiency for an exactly normal law, but show a much better and more stable performance under deviations from it.

While for years one had been concerned mostly with what was later called 'robustness of validity' (that the actual confidence levels should be close to, or at least on the safe side of the nominal levels), one realized now that 'robustness of performance' (stability of power, or of the length of confidence intervals) was at least as important ... (Huber. 1972)

In his book Robust Statistics. Huber says:

The word 'robust' is loaded with many ... connotations. We use it in a relatively narrow sense: for our purposes, robustness signifies insensitivity to small deviations from the assumptions. (Huber. 1981)

In terms of our current interest, an obvious area in which we seek the protection of robust statistical methods is where we encounter outliers in a set of data. For
example, if we are interested in estimating a parameter in an initial model, but concerned about the prospects of outliers, we would want to use an estimator which is not likely to be highly sensitive to such outliers. The study of accommodation of outliers is specifically concerned with robustness in the face of outliers, and methods of estimation and testing are derived with particular regard to the nature of any model which might be needed to explain the presence of outliers.

The another type of statistical method for handling outliers is that of testing an outlier with the prospect of rejecting it from the data set, or of identifying it as a feature of special interest. Although the philosophies underlying the notions of identification and accommodation seem distinct, they are often confused, since a method of accommodation may produce a method of identification as a by-product, or vice-versa. In general, one should not only detect and accommodate outliers, but also interpret (perhaps correct) them.

In small-size univariate samples, a visual inspection of the data is perhaps the most common method of identifying outliers. Observations on the extremes of the sample that are well separated from the bulk of the data are usually given special attention. In more complicated settings, a general idea behind methods to identify a single outlier is to transform the data into a set of $n$ univariate statistics, one for each observation in the data, that are inspected visually or, in some cases, used to construct significance tests. For example, the Studentized residuals are often used to identify outliers in analysis based on linear models. The observation with the largest absolute Studentized residual is usually given special attention and is taken as the observation most likely to be a contaminant.
For the detection of multivariate outliers, Barnett (1979) considers two general principles which are based on the basic model $F$ under the null hypothesis of no outliers and an alternative (contamination) model $\tilde{F}$ which is of a slippage type under which one observation is a contaminant. The two principles are:

**Principle A.** The most extreme observation is that one, $x_i$, whose omission from the sample $x_1, x_2, ..., x_n$ yields the largest incremental increase in the maximized likelihood under $F$ for the remaining data. If this increase is surprisingly large, declare $x_i$ to be an outlier.

**Principle B.** The most extreme observation is that one, $x_i$, whose assignment as the contaminant in the sense of $\tilde{F}$ maximizes the difference between the likelihoods of the sample under $\tilde{F}$ and $F$. If this difference is surprisingly large, declare $x_i$ to be an outlier.

We will see later that the test criterion for the detection of outliers in multivariate normal data based on these principles is nothing but the squared Mahalanobis distance. The observation with the largest distance is usually taken as the point most likely to be a contaminant.
CHAPTER 2

Reviews of Relevant Literature on Outliers

2.1 Early Approaches

The existence of the problem of doubtful or anomalous values has been recognized for a very long time, certainly since the middle of the eighteenth century. Bernoulli (1777) evidently questioned the assumption of identically distributed errors and condemned the widespread practice of discarding discordant observations in the absence of a priori information. Comments of Peirce (1852) indicate that the practice of rejecting extreme observations was still commonplace some 75 years after Bernoulli, but objective criteria for justifying the practice still had not been developed.

There is a stormy history behind the rejection of outlying observations. The first attempt at a rejection criterion based on some sort of probability reasoning was that of Peirce (1852). Three years later Gould (1855) constructed tables to help in the implementation of Peirce's criterion. However, the noted professor of astronomy Airy (1856), possibly motivated by an error in the published scale of longitudes for England caused by the retention of only the most accordant observations, totally rejected Peirce's idea. Airy argued that, while some observations were unlikely, any observation was possible and no observation should be preferred to another if there were
no specific reason to believe that unusual causes of error must have intervened in special observations. Realizing that the philosophy of never rejecting any observation no matter how large the error is contrary to common sense, he then rejected the mathematical theory of normal errors rather than the observation.

Wintlock (1856) found Peirce's criterion both correct and of value. He argued that Airy accepts, a posteriori, the magnitude of discordant observations when using the method of least squares. Peirce's argument was reproduced by Chauvenet (1863), who then gave a similar rule based on a simpler argument. Another proposal of the outright rejection of extreme observations was made by Stone (1868). He found the use of Peirce's criterion to be troublesome and, citing Airy, expressed doubts about the correctness of the mathematics.

The next significant step after Peirce and Chauvenet was apparently made by Wright (1884). After an excellent general discussion of the problems raised by the occurrence of outlying observations among astronomical readings, he suggests that the best rule for a computer to follow who is not the observer is to reject any observation whose residual exceeds in magnitude five times the probable error (i.e., 3.37 times the standard deviation). The reason given for this is that if the Gaussian law of error is truely satisfied, only about one observation in a thousand will be rejected, and therefore little damage will be done in any case.

While the techniques of outright rejection of atypical observations by Peirce, Chauvenet, and Stone generated much controversy, they did not seem to be widely used. Also, they worked with the model of Gaussian errors rather than with mixtures of
Gaussians, which would have made their arguments for the outright rejection of outliers more convincing.

Some alternatives to outright rejection of extreme values were also being considered. Several methods were proposed by different authors for the weighting of observations in calculating a sample mean. This can be regarded as a robust procedure for estimating a location parameter. It illustrates what is called the accommodation of outliers.

Glaisher (1872-73) was perhaps the first to publish such a weighting procedure. Glaisher's method was concerned with $n$ observations $x_1, \ldots, x_n$ from normal distributions, with a common mean $\mu$, and with unknown and unequal variances. He proposed estimating the mean iteratively by a weighted combination of the $x_i$ with weights determined from the squared deviations of the sample values. A few months later Stone (1873) criticised Glaisher's method and proposed an alternative weighting procedure, based in effect on maximizing the likelihood. This leads to a weighted mean $\mu^*$ given by the $(n - 1)$th degree equation

$$\sum_{i=1}^{n}(x_i - \mu^*)^{-1} = 0.$$

Another weighting method proposed at that time was by Newcomb (1886). He assumes that the sample of $n$ observations comes from a mixture of $r$ normal distributions, and evolves a final estimate of $\mu$ which is constructed as a weighted mean of $r^n$ different weighted means of the $x_i$. Rather interestingly, Newcomb refers in his paper to the "evil" of a value; this turns out to be the mean squared error of an estimate, and is an interesting early use of the concept of a loss function.
2.2 Outliers in Simple Normal Samples

The motivation for a statistical treatment of outliers came first from the problems of combining astronomical observations, and repeated measurements or determinations must always be one of the main contexts in which outlier problems arise. In many cases errors of measurement may plausibly be assumed to follow a normal distribution, whether purely as an empirical fact, or through the operation of the central limit theorem on contributory error components. It is natural, therefore, that the vast body of published methodology on outliers from the past to the present day rests on the working hypothesis of a normal distribution. In fact, outliers in exponential and other non-normal models have been specifically considered only from the last few decades.

Ogrodnikoff (1923), Jeffreys (1932), Dixon (1950), and Grubbs (1950) suggested mixture models to describe sampling distributions that produce outliers. Dixon (1950) produced two types of mixtures: the first of these was a mean-shift model (called model A by Ferguson 1961) where the distribution is a mixture of $N(\mu, \sigma^2)$ and $N(\mu + \lambda, \sigma^2)$ components, while the “variance-inflation model” (called model B by Ferguson 1961) is a mixture of $N(\mu, \sigma^2)$ and $N(\mu, a^2\sigma^2)$ components, $a^2 > 1$.

The use of such models led to the consideration of formal statistical tests for the rejection of atypical observations. Interestingly, the development of formal test statistics is more in line with the rejection of outliers for the purposes of special interest, although most authors cite improved estimation and accommodation as reasons for rejecting outliers. So the development of tests for the rejection of outliers still contin-
ues. The tests for the rejection of outliers led to a gap between those who advocate rejecting outliers and those who prefer a less severe weighting scheme. Most proponents of tests now do not mention estimation, while those interested in weighted means or robust estimators pay little attention to the rejection of outliers for the study of concomitant variables.

The early work of Peirce, Chauvenet, and Stone on formal tests for the rejection of outlying observations was continued in the 20th century. The gap between the first and second, and the second and third order statistics as a test statistic for the rejection of one or two outliers was proposed by Irvin (1925). By rejection of outliers he meant "the realization of the fact that the particular observations in question probably do not belong to the same homogeneous group". He found the critical values for these statistics based on a known variance. Similarly, Tippet (1925) proposed the use of the range for the detection of outliers assuming that the variance is known.

McKay (1935) proposed a different approach and found the distribution of the difference between the extreme observation and the sample mean scaled by a known standard deviation, that is. \((x_{(n)} - \bar{x})/\sigma\). Nair (1948) later tabulated the percentage points of the distribution of these variables, who also found the distribution of \((x_{(n)} - \bar{x})/s\), where \(s\) is an independent estimate of \(\sigma\).

Noting that in practice \(\sigma\) is estimated by the sample standard deviation. Thompson (1935) found the distribution of an arbitrary Studentized residual, \(T_i = (X_i - \bar{X})/\hat{\sigma}\). He showed that the statistic \((n - 2)^{1/2}T_i/((n - 1 - T_i^2)^{1/2})\) has a Student's \(t\) distribution with \(n - 1\) degrees of freedom. He found critical values by fixing the
frequency of rejection per sample. Pearson and Chandra Sekar (1936) pointed out
that although Thompson's criterion is both useful and practical for the detection of
one outlier, the Studentized residuals are not independent and that in practice someone
would look only at the Studentized residual with largest absolute value. They
considered \( (x_{(n)} - \bar{x})/s \) and \( (\bar{x} - x_{(1)})/s \) as criteria for rejecting individual observations
as significantly high and low outliers, respectively, where \( x_{(1)} < x_{(2)} < \ldots < x_{(n)} \) are
the order statistics of the sample \( x_1, x_2, \ldots, x_n \). In particular, they showed that the
upper tail of the distribution of \( (x_{(n)} - \bar{x})/s \) (or of \( (\bar{x} - x_{(1)})/s \)) has a density function
\( n f_n(u) \) on the interval \( (\sqrt{(n-2)/2}, \sqrt{n-1}) \), where \( f_n(u) \) is the probability density
function of \( (x_i - \bar{x})/s = u \), say. From this viewpoint they were able to obtain the
upper percentage points of the distribution of \( (x_{(n)} - \bar{x})/s \) (or of \( (\bar{x} - x_{(1)})/s \)) for
values of \( n \) ranging from 11 to 19.

Grubbs (1950) extended the work of Pearson and Chandra Sekar (1936) for indi-
vidual outliers by finding the exact distribution of \( (x_{(n)} - \bar{x})/s \) (or of \( (\bar{x} - x_{(1)})/s \))
in a sample from a normal distribution with unknown parameters. He tabulated the
upper percentage points of the distribution of \( (x_{(n)} - \bar{x})/s \) (or of \( (\bar{x} - x_{(1)})/s \)) for
all sample sizes \( n \leq 25 \). He also tabulated the lower percentage points of the distri-
bution of \( \sum_{i=1}^{n-1}(x_i - \bar{x}_n)^2/[(n-1)s^2] \) (or of \( \sum_{i=2}^{n}(x_i - \bar{x}_1)^2/[(n-1)s^2] \)) for all
\( n \leq 25 \), where \( \bar{x}_n \) is the mean of \( x_{(1)}, \ldots, x_{(n-1)} \) and \( \bar{x}_1 \) is the mean of \( x_{(2)}, \ldots, x_{(n)} \).
Grubbs also considered the case of two high (or two low) outliers, using as the cri-
terion of rejection \( \sum_{i=1}^{n-2}(x_i - \bar{x}_{n,n-1})^2/[(n-2)s^2] \) (or of \( \sum_{i=3}^{n}(x_i - \bar{x}_{1,2})^2/[(n-2)s^2] \)),
which \( \bar{x}_{n,n-1} \) is the mean of \( x_{(1)}, \ldots, x_{(n-2)} \) and \( \bar{x}_{1,2} \) is the mean of \( x_{(3)}, \ldots, x_{(n)} \). He
tabulated the lower percentage points of the distribution of these statistics for all
\( n \leq 20 \). Grubbs, however, was unable to obtain critical values for the test statistic
\[
\sum_{i=2}^{n-1} (x(i) - \bar{x}_{1,n})^2 / [(n - 2)s^2]
\]
for testing simultaneously the largest and smallest observations as outliers, where \( \bar{x}_{1,n} \) is the mean of \( x(2), ..., x(n-1) \).

Dixon (1951) has considered ratios of the form \((x(n) - x(n-j))/(x(n) - x(i))\) [or \((x(i+1) - x(i))/(x(n-i+1) - x(1))\)], \( i = 1, 2, 3; j = 1, 2 \), as criteria for testing extreme observations as outliers and he has tabulated the percentage points of the distributions of these statistics.

Various invariant test for outlier rejection was considered by Ferguson (1961). He showed that tests based on the sample skewness are locally best invariant tests for outliers with small positive shifts in the mean and that tests based on the sample kurtosis are locally best invariant tests for outliers with small shifts in the mean in either direction. Tests based on the sample kurtosis are also locally best invariant tests for testing small positive shifts in the variance. Although tests based on either kurtosis or skewness are locally optimal, no test is most powerful in determining outliers for small shifts in the mean. Furthermore, Ferguson showed that for samples of size \( n = 25 \) the Studentized residuals are more powerful for detecting intermediate to large (greater than 3\( \sigma \)) shifts in the mean.

The repeated applications of single-outlier tests for the rejection of multiple outliers is termed as a \textit{consecutive test} by Barnett and Lewis (1994). McMillan and David (1971) and McMillan (1971) studied powers of outlier tests comparing consecutive versions of maximum Studentized or standardized residuals with the block procedure for the detection of two outliers proposed by Murphy (1951). McMillan and David showed that when the variance is known, Murphy's test does only slightly
better than repeated applications of the normal residual test. However, McMillan showed that when the variance is unknown, the consecutive Studentized residual test tends to decrease in power as the two outlying observations move further from the mean in the same direction. This decrease in power for consecutive tests is termed as masking. Hawkins (1980) argued that the masking effect of two outliers tends to decrease as the sample size increases, but when the sample size increases the number of outliers may increase and hence more masking may occur.

Realizing that consecutive application of single-outlier tests leads to problems, several authors proposed tests for the simultaneous rejection of outliers. Tietjen and Moore (1972, 1979) proposed two statistics for the identification of multiple outliers. For testing \( k \) outliers on one side of the data they proposed the statistic

\[
L_k = \sum_{i=1}^{n-k} (X_{(i)} - \bar{X}_k)^2 / S^2.
\]

where \( \bar{X}_k = \sum_{i=1}^{n-k} X_{(i)}/(n-k) \). For \( k \) outliers on either side they proposed

\[
T_k = \sum_{i=k+1}^{n-k-1} (z_{(i)} - z_k)^2 / S^2.
\]

where \( z_{(i)} \) is the value of \( Z \) corresponding to the \( i \)th largest (or smallest) ordered absolute residual. Tietjen and Moore suggest that \( k \) be estimated by use of the largest gap in the order statistics when testing for outliers on one side of the data. But they do not give any indication of how to estimate \( k \) for testing outliers on either side of the data.

Tiku (1975) also proposed a test procedure for the simultaneous rejection of \( k \) outliers. He found as a test criterion the ratio of the ML estimator of \( \sigma \) from a
censored sample to that from an uncensored sample. The amount of censoring is estimated by the maximum gap in the order statistics, scaled by its expected value. Tiku conducted a power study, but he contaminated only the largest (or smallest) observations generated from a normal distribution.

In a significant paper, Rosner (1975) developed test statistics which do not suffer from masking and require only the knowledge of the maximum number of possible outliers, $k_u$. Using Rosner’s notation, consider the subsets $I_0, I_1, ..., I_n$, where $I_0 = \{x_1, x_2, ..., x_n\}$ and $I_{i+1}$ is formed by deleting from $I_i$ the point $x_i^t$ farthest from the mean of the points in $I_i$. Then consider $R_t = S(I_{t-1})$, where $S(I)$ is any statistic for detecting one outlier. Rosner suggested to find constants $\lambda_t(\beta)$ such that

$$P\left\{ \bigcup_{i=1}^{k_u} [R_i > \lambda_i(\beta)] \right\} = \alpha.$$ 

The test procedure declares $x_0, x_1, ..., x_{l-1}$ as outliers, when $l = \max\{i : R_i > \lambda_i(\beta)\}$. No outliers are declared by the test if $R_i < \lambda_i(\beta)$ for all $i$. Rosner studied the power of four outlier statistics and showed that the test procedure based on the extreme Studentized deviate seemed to be the most powerful. Power studies of the Rosner technique were conducted by Jain (1981) and Jain and Pingel (1981), but in each of these studies the Tiku method of contamination was used.

Hawkins (1978) and Cook and Beckman (1980) show that, while the Rosner procedure controls the $\alpha$ level for the rejection of the null hypothesis, it is prone to swamping, by which more outliers are declared than there are in the sample when the null hypothesis is rejected.
The most annoying issue in the case of multiple outlier tests is how to choose \( k \) or an estimate thereof. When multiple outliers are present, most techniques suffer from either masking or swamping or both. As Beckman and Cook (1983) suggests: “Masking, swamping, and the number of observations correctly identified need to be studied for these methods.”

### 2.3 Outliers in Multivariate Normal Samples

Although much work on univariate outliers has been done so far by many authors, much less work has been done on multivariate outliers. This is mainly due to the fact that outliers in a multivariate data can be hard to detect, especially when the dimension is greater than 2, because then we can no longer rely on visual inspection. A multivariate outlier no longer has a simple manifestation as an observation which sticks out at the end of the sample (Gnanadesikan and Kettenring 1972).

For the identification of a single outlier in a sample from a multivariate normal distribution with mean \( \mu \) and covariance matrix \( \Sigma \), Siotani (1959) proposes as test criterion the statistics

\[
T_i^2 = \max_i (X_i - \bar{X})' \Sigma^{-1} (X_i - \bar{X})
\]

and

\[
T_i^2 = \max_i (X_i - \bar{X})' S^{-1} (X_i - \bar{X}),
\]

where \( \Sigma \) is the known covariance matrix and \( S \) is an independent estimate of \( \Sigma \). If \( T_i^2 \) is significantly large, then the point corresponding to this \( T_i^2 \) will be declared as an outlier. Similar argument is true for \( T_i^2 \). The null distributions of these statistics are
not readily determined in exact form. Siotani discusses the problems associated with determining the percentage points of \( T^2 \) and presents approximate upper percentage points of this statistic for dimensions \( p = 2, 3, 4 \) and for different sample sizes. He also tabulates approximate percentage points for a studentized form of \( T^2 \) in which \( S \) is replaced by an external unbiased estimate \( S_\nu \) of \( \Sigma \) having a Wishert distribution with \( \nu \) degrees of freedom. These are of value for an informal test of discordancy of a single outlier in a multivariate sample, where \( \Sigma \) is not estimated from the sample itself but by means of such an external estimate.

The first detailed applications-oriented study of outlier detection in multivariate normal data was conducted by Wilks (1963). For identifying a single outlier in a multivariate data he suggests to choose that sample point as an outlier whose omission leads to the least value for the so-called one-outlier scatter ratio

\[
R_i = \frac{|A^{(i)}|}{|A|}
\]

where \( A \) is the matrix of sums of squares and cross-products of the observations, that is,

\[
A = \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})'
\]

and \( A^{(i)} \) is the restricted matrix obtained on omission of \( x_i \). Wilks shows that for any \( i \) the ratio \( R_i \) has the beta distribution \( \beta\left(\frac{n-p-1}{2}, \frac{p}{2}\right) \), where \( p \) is the dimension of the random variable \( X \) assumed to have a normal distribution. He points out that under the null hypothesis that all elements of the sample are independently drawn from a common \( p \)-dimensional normal distribution, the one-outlier scatter ratios \( R_1, \ldots, R_n \) are random variables having a distribution which is symmetric over the \( n \)-dimensional space of \( R_1, \ldots, R_n \) for which
\[
\sum_{i=1}^{n} R_i = n \left( 1 - \frac{p}{n-1} \right)
\]

\[0 \leq R_i \leq 1, (i = 1, 2, \ldots, n).\]

Wilks proposes the statistic \( R_{(1)} = \min_i R_i \) as test criterion for selecting and testing a single extreme observation as a significant outlier. As the joint distribution of \( R_{(1)}, \ldots, R_{(n)} \), or even of \( R_1, \ldots, R_n \) is very complicated, he uses Bonferroni inequalities to obtain an upper bound for the distribution function of \( R_{(1)} \), and hence lower bounds for the lower percentage points of \( R_{(1)} \). He also makes a comparison with exact results due to Grubbs (1950) for the case \( p = 1 \). Wilks tabulates lower bounds to the lower percentage points of \( R_{(1)} \) for dimensions \( p = 1, 2, \ldots, 5 \) and for different sample sizes.

Wilks adopts a similar approach for the simultaneous rejection of 2, 3, or 4 outliers in a multivariate data, by considering for the \( k \)-outlier case \( (k = 2, 3, 4) \) the \( k \)-outlier scatter ratios

\[
R_{i_1, i_2, \ldots, i_k} = \frac{|A^{(i_1, i_2, \ldots, i_k)}|}{|A|}
\]

where \( |A^{(i_1, i_2, \ldots, i_k)}| \) is the internal scatter when \( x_{i_1}, x_{i_2}, \ldots, x_{i_k} \) are omitted from the sample. The test criterion is obtained by taking the minimum of the \( \binom{n}{k} \) possible values of \( R_{i_1, i_2, \ldots, i_k} \). Wilks tabulates the critical values for \( k = 2 \), that is, for an outlying subset of two observations.

In fact, the statistics suggested by Siotani and Wilks are multivariate generalizations of work of Thompson (1935), Pearson and Chandra Sekar (1936), Nair (1948).
and Grubbs (1950). Fung (1988) discusses Wilks' two- and three-outlier tests in more detail. He points out that although the bounds for the Wilks' single-outlier test are very good, they are rather conservative for the two-outlier test. He determined the critical values for two and three outliers with dimensions $p = 2, 3, 4, 5$ and sample sizes up to 50 by simulation.

Bacon-Shone and Fung (1987) propose a graphical approach based on Wilks' statistic. They show that their graphical method is particularly valuable in avoiding the problems of masking and swamping. They illustrate the method with some examples and simulations.

Caroni and Prescott (1992) propose a generalization of Wilks' single-outlier test suitable for application to the many outlier problem of detecting from 1 to $k$ outliers in a multivariate data. They use the same principles as Rosner (1975, 1977) uses in the univariate cases. Rosner proposes sequential application of the generalized extreme Studentized deviate to univariate samples of reducing size, in which the type I error is controlled both under the hypothesis of no outliers and under the alternative hypothesis of $1, 2, ..., k$ outliers. Caroni and Prescott point out that critical values for the sequential application of Wilks' test to identify many outliers depend only on those for a single-outlier test which may be approximated by percentage points from the $F$-distributions as tabulated by Wilks. However, experience reveals that like consecutive tests of single outlier in univariate data, their sequential application of multivariate outlier test also suffers from masking effect.

A Bayesian approach for the detection of univariate as well as multivariate out-

23
liers is proposed by Guttman (1973). He considers that all observations are hopefully generated independently from the same normal distribution $N(\mu, \sigma^2)$, but fears that one of these may come from a spurious source, $N(\mu + a, \sigma^2)$. As in the univariate case, the posterior of the "shift parameter $a$" is shown to be a function of weights that indicate outlying observations. He also considers multivariate case that may be contrasted to the work of Wilks (1963).

In practice, for detecting multiple outliers in multivariate data the classical identification method does not always perform well, because it is based on the sample mean and covariance matrix that are themselves affected by the outliers. In order to overcome this problem, the robust estimates of location and scatter may be suggested. But in the case of multivariate data, the robust estimation is not straightforward. However, the development of computer packages has certainly facilitated the analysis of multivariate data.

The most commonly used test criterion for the identification of multivariate outliers is the squared Mahalanobis distance from a location and shape. that is.

$$MD_i = (X_i - \bar{X})' S^{-1} (X_i - \bar{X}).$$

To avoid the masking effect, several authors propose to compute the squared distances from the robust estimates of location and scatter. Maronna(1976) developed an algorithm for finding robust $M$-estimates of multivariate location and shape. He then exhibits some small-sample Monte Carlo results on the properties of the estimators.
Huber (1977) conducts a systematic study of the affine-invariant situation considered by Maronna (1976), with the aim of finding the most general form of $M$-estimator of covariance matrix.

For the detection of atypical observations from multivariate data sets, Campbell (1980) proposes Mahalanobis distances from robust $M$-estimates of means and covariances, rather than the usual maximum likelihood estimates. The weights associated with the robust estimation can also be used to indicate outlying observations. The robust estimates are defined such that for uncontaminated data they are similar to the usual classical estimates. Campbell suggests finding the $M$-estimators of location and scale by solving iteratively the simultaneous equations

$$
\overline{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}
$$

and

$$
S = \frac{\sum_{i=1}^{n} w_i^2 (x_i - \overline{x})(x_i - \overline{x})'}{\sum_{i=1}^{n} w_i^2 - 1}
$$

where

$$
w_i = w(d_i) = \omega(d_i)/d_i
$$

and

$$
d_i = \left\{ (x_i - \overline{x})'S^{-1}(x_i - \overline{x}) \right\}^{\frac{1}{2}}.
$$

The weight function $\omega(d)$ is commonly taken as a continuous function of bounded influence. Hampel (1973) suggests that the influence and hence the weight of an
extreme atypical observation should be zero, so that \( \omega(d) \) should redescend for sufficiently large values of \( d_i \). Campbell proposes the two-parameter form of \( \omega(d) \) as

\[
\begin{align*}
\omega(d) &= \begin{cases} 
  d & \text{if } d \leq d_0, \\
  d_0 \exp\left\{-\frac{1}{2}(d - d_0)^2/a_2^2\right\} & \text{if } d > d_0,
\end{cases}
\end{align*}
\]

with \( d_0 = \sqrt{p} + a_1/\sqrt{2} \), for suitable choice of positive quantities \( a_1 \) and \( a_2 \).

Devlin, Gnanadesikan, and Kettenring (1981) conduct a Monte Carlo study to compare the performances of several robust procedures for estimating a correlation matrix. They consider separate bivariate analyses, and estimation of the full correlation (or covariance) matrix using either (ellipsoidal) multivariate trimming or \( M \)-estimators. They show that the \( M \)-estimators stand up exceptionally well. However, the \( M \)-estimators can break down relatively easily in the case of higher dimensionality.

In a significant paper, Rousseeuw and van Zomeren (1990) develop an algorithm for the \textit{minimum volume ellipsoid (MVE) estimator} of location and scale which is introduced by Rousseeuw (1985). The purpose of this work is to detect outliers with the avoidance of the \textit{masking} by which multiple outliers do not necessarily have a large Mahalanobis distance. They propose the robust distance

\[
RD_i = \sqrt{(X_i - \bar{X}^*)' S^{-1} (X_i - \bar{X}^*)}
\]

as test criterion for identifying multiple outliers in multivariate data, where \( \bar{X}^* \) and \( S^* \) are the \textit{MVE} estimators of location and scale, respectively. Any point whose robust distance exceeds the cutoff value \( \sqrt{\chi^2_p(\alpha)} \) is declared as an outlier by their method.
Rousseeuw and van Zomeren find the $MVE$ estimators of location and scatter that are \textit{affine equivariant}. Consider a data set $X = (x_1, \ldots, x_n)$ of $n$ points in $p$ dimensions. Then the location estimator $T(X)$ and the shape estimator $C(X)$ are said to be affine equivariant if and only if

$$T(x_1A + b, \ldots, x_nA + b) = T(x_1, \ldots, x_n)A + b$$

and

$$C(x_1A + b, \ldots, x_nA + b) = A'C(x_1, \ldots, x_n)A$$

for any row vector $b$ and a nonsingular $p \times p$ matrix $A$.

The $MVE$ estimator is defined as the pair $(\bar{X}^*, S^*)$ such that the determinant of $S^*$ is minimized subject to

$$\# \left\{ i; (X_i - \bar{X}^*)'S^*-1(X_i - \bar{X}^*) \leq g^2 \right\} \geq h$$

where $h = [(n + p + 1)/2]$ in which $[m]$ is the integer part of $m$. The fixed constant $g^2$ can be chosen as $\chi^2_p(0.50)$ expecting that the majority of the data come from a normal distribution. For small samples, an adjustment factor $c_{n,p}^2$ is suggested which depends on $n$ and $p$.

Hadi (1992) criticizes the $MVE$ estimator of Rousseeuw and van Zomeren (1990) in the sense that it is "computationally expensive". He proposes a procedure for approximating the $MVE$ which is easy to compute, and at the same time, it has high breakdown point. Working separately, a similar procedure is suggested by Atkinson (1994). Their proposed algorithms start with an estimate of location and shape based on $(p+1)$ points and then select successively larger sets. Hadi uses the subset of
(p + 1) points to find the robust estimates of location and covariance matrix. He then computes the Mahalanobis distances of all points based on the location and scatter of the observations contained in the subset. In the next step, the size of the subset is increased by one observation until a stopping criterion (based on the distances) is met. The final subset is used to obtain the robust estimators of location and scale.

Hadi suggests using coordinatewise medians as an initial estimator of location and the covariance with that as center for a preliminary shape estimator. The initial set of p + 1 points consists of those whose distance from the initial location estimator is least, using the initial shape estimator. However, the coordinatewise median is not affine equivariant. On the other hand, Atkinson's method is affine equivariant. He suggests restarting the procedure many times with randomly selected sets of p + 1 points. For each trial, sequential addition is performed and for each stage in the sequential addition, the covariance matrix is calculated so that (n + p + 1)/2 points are included in the ellipsoid defined by the current location and shape. The estimate over all trials and over all stages of each trial in which the scaled shape matrix has minimum determinant can be taken as the robust estimate of the location and shape.

Another significant work on identification of multivariate outliers is done by Rocke and Woodruff (1996). They give new insights into why the problem of identifying multiple outliers can be more difficult with the increase of the dimension of the data. They designed a hybrid method which may extend the practical boundaries of outlier detection capabilities. An algorithm is developed by them to work with this method which they call hybrid algorithm.
The hybrid method is designed in two phases. In the first phase, the estimates of location and shape are calculated using the MCD. The MCD for any set of data is defined by the half sample whose covariance matrix has minimum determinant. It is convenient to search for MCD half-samples moving from half sample to half sample by removing one point in the current half sample and adding one point not currently in the half sample. Neighborhoods defined in this way can form the basis of a steepest descent to a local minimum. The hybrid method uses the MCD as a starting point for a sequential point addition algorithm (Atkinson 1994, Hadi 1992).

In the second phase, they standardize the shape matrix so that 1st ordered distance is equal to \( \sqrt{\chi^2_{p}(h/n)} \), where \( h = [(n+p+1)/2] \). Then a cutoff value \( T_\alpha \) is determined by simulation so that when multivariate normal data of size \( n \) and dimension \( p \) are submitted, a fraction \( \alpha \) of the points on the average lie beyond \( T_\alpha \). Then a new shape matrix is formed as \( c(p, \alpha)S \) with some adjustment factor \( c(p, \alpha) \), where \( S \) is the covariance matrix of the points whose distance in the previous stage is less than \( T_\alpha \). The location estimator is the average of those points. Then they suggest rejecting any point as outlier whose distance using the new location and shape is greater than \( \chi^2_{p}(\alpha) \).

However, the hybrid method is very tedious. Furthermore, the distances from the robust estimators of location and shape are poorly approximated by Chi-square. It may lead to a very large number of erroneous rejections unless the sample size is very large.

Several authors have suggested performing a principal component analysis on the data in the case of high dimensionality, and looking at sample values of the projection of the observations onto the principal components of different order. Campbell (1980)
suggests a procedure for robust principal component analysis which also indicates atypical observations and provides an analysis relatively little influenced by such observations. Gnanadesikan and Kettenring (1972) remark how the first few principal components are sensitive to outliers inflating variances or covariances (or correlations, if the principal component analysis is conducted in terms of the sample correlation matrix, rather than the sample covariance matrix). However, some authors are critical of the principal component analysis of outliers in view of the fact that outliers should be in the whole data, rather than a few principal components.
CHAPTER 3

Sequential Procedure for Testing Multivariate Outliers

3.1 Identification of Multivariate Outliers

Consider a sample of $n$ observations $x_1, ..., x_n$ of a $p$-dimensional random variable $X$ assumed to have a normal distribution $N_p(\mu, \Sigma)$. A possible alternative model which would account for a single contaminant is the slippage alternative, which can be of the following type:

$$X_i \sim N_p(\mu + \lambda, \Sigma) \text{ for some } i$$
$$X_j \sim N_p(\mu, \Sigma) \text{ (} j \neq i \text{).}$$

The parameters $\mu$ and $\Sigma$, the slippage parameter $\lambda$, and the index $i$ of an outlier are unknown.

The maximum of the log-likelihood under the basic model is (apart from the constant factor)

$$l_0(x) = -\frac{n}{2} \log |A|$$

where $A$ is the matrix of sums of squares and cross-products.
\[ A = \sum_{j=1}^{n} (x_j - \bar{x})(x_j - \bar{x})'. \]

Under the slippage alternative, the likelihood function is

\[ L_i(x) = L_i(\mu, \Sigma, \lambda|x) \]

\[
= \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{n}{2}}} \exp \left\{ -\frac{1}{2} \sum_{j=1}^{n} \left( (x_j - \mu)\Sigma^{-1}(x_j - \mu) - \frac{1}{2}(x_i - \mu - \lambda)\Sigma^{-1}(x_i - \mu - \lambda) \right) \right\}
\]

\[ = C \exp \left[ -\frac{1}{2} \left\{ \sum_{j=1}^{n} (x_j - \mu)'\Sigma^{-1}(x_j - \mu) + (x_i - \mu)'\Sigma^{-1}(x_i - \mu) - 2\lambda'\Sigma^{-1}(x_i - \mu) + \lambda'\Sigma^{-1}\lambda \right\} \right] \]

where

\[ C = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{n}{2}}}. \]

Assuming that \( \mu \) and \( \Sigma \) are fixed, the likelihood equation for estimating \( \lambda \) is

\[ 0 = \frac{\delta \log L_1(x)}{\delta \lambda} = \Sigma^{-1}(x_i - \mu) - \Sigma^{-1}\lambda \]

which implies that

\[ \hat{\lambda} = x_i - \mu. \]

Replacing \( \lambda \) by \( \hat{\lambda} \), the likelihood function becomes
From this likelihood function, the $ML$ estimators of $\mu$ and $\Sigma$ can be obtained (using the similar argument as in the estimation of mean and covariance in multivariate normal density) as

$$L_i(\mu, \Sigma, \hat{\lambda}|x) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{n}{2}}} \exp \left\{ -\frac{1}{2} \sum_{j=1}^{n} (x_j - \mu)' \Sigma^{-1} (x_j - \mu) \right\}.$$ 

Hence under the slippage alternative, the maximized log-likelihood is (apart from the constant factor)

$$l_i(x) = -\frac{n}{2} \log |A^{(i)}|$$

where $A^{(i)}$ is the restricted matrix with $x_i$ eliminated and $i$ is chosen to maximize

$$l_i(x) - l_0(x).$$

We can write

$$l_i(x) - l_0(x) = -\frac{n}{2} \log \frac{|A^{(i)}|}{|A|}.$$ 

So, equivalently, if we denote
the outlier is that observation corresponding with the smallest \( R_i \). Wilks (1963) termed the quantities \( R_1, ..., R_n \) as one-outlier scatter ratios of the sample \( x_1, ..., x_n \).

We can rewrite

\[
R_i = \frac{|A^{(i)}|}{|A|} 
\]

and so minimization of \( R_i \) is equivalent to maximization of the Mahalanobis distance

\[
MD_i = (x_i - \bar{x})'S^{-1}(x_i - \bar{x}).
\]

Wilks shows that for any \( i \), the scatter ratio \( R_i \) has the beta distribution beta \( \left( \frac{n-p-1}{2}, \frac{p}{2} \right) \). Caroni and Prescott (1992) point out that this may be shown by relating \( R_i \) to the Hotelling \( T_i^2 \) for comparing \( x_i \) with the means of the data \( \bar{x} \). We obtain

\[
T_i^2 = (n-2) \left( \frac{n}{n-1} \right) (x_i - \bar{x})'A^{(i)-1}(x_i - \bar{x})
\]

such that
\[
\frac{n-p-1}{p(n-2)}T_i^2 \sim F_{p,n-p-1}.
\]

Thus

\[
R_i = \left(1 + \frac{T_i^2}{n-2}\right)^{-1}
\]

is distributed as beta\(\left(\frac{n-p-1}{2}, \frac{p}{2}\right)\).

Wilks (1963) proposed the test statistic

\[
W_1 = \min_i R_i = \min_i \left|A^{(i)}\right| |A|
\]

for selecting and testing a single observation as a significant outlier. Percentage points of \(W_1\) were constructed using Bonferroni bounds obtained from the lower \(100\alpha/n\) percentage points of the above beta distribution. Caroni and Prescott (1992) indicate that the percentage points may also be obtained from the upper points of the \(F_{p,n-p-1}\) distribution, \(F_{p,n-p-1,1-\frac{\alpha}{n}}\) say, by using the transformation

\[
\left(1 + \frac{p}{n-p-1}F_{p,n-p-1,1-\frac{\alpha}{n}}\right)^{-1}.
\]

Wilks tabulated lower bounds to the lower 10, 5, 2.5 and 1 percent points of \(W_1\) for \(p = 1, \ldots, 5\) dimensions and for a selection of sample sizes up to 500.

A generalization of Wilks' single-outlier test for the application of many-outlier test in a multivariate data is proposed by Caroni and Prescott (1992). The authors
propose a sequential procedure for the detection of many outliers using Wilks’ single-outlier test statistic which is called a consecutive test by Barnett and Lewis (1994). But the consecutive test is found to decrease in power as the two outlying observations moved further from the mean in the same direction. This decrease in power for consecutive tests is called masking. In order to avoid masking, some authors suggest using robust procedure for the identification of outliers.

Following the principle of one-outlier test in multivariate data proposed by Wilks (1963), we propose the maximum robust distance, $Q_1$, defined by

$$Q_1 = Q_1(\bar{x}^*, S^*) = \max_i RD_i$$

$$= \max_i (x_i - \bar{x}^*)'(S^*)^{-1}(x_i - \bar{x}^*)$$

as test criterion for the detection of multivariate outliers. Here $\bar{x}^*$ and $S^*$ are the robust estimators of mean and covariance matrix, respectively. The robust estimators are obtained such that under the null hypothesis of no outliers, they are similar to the classical estimators in that the weights associated with the observations should be equal to or close to 1.

the outlier problem in some detail and develop a *hybrid* algorithm for identifying the outlying observations. We use some of these approaches for the robust estimation of location and shape. In the next section we discuss the robust estimation of mean and covariance matrix by means of which we find our test statistic $Q_1$.

### 3.2 Robust Estimation of Multivariate Location and Scale

The most difficult problem in robust inference is the estimation of location and scatter for a multivariate data. Much work has been done so far by many authors and the development of new approaches still continues. We reviewed some of the significant approaches for the robust estimation of location and shape in the previous chapter. Campbell (1980) proposes $M$-estimators for location and shape, which marked an important improvement in robust estimation. But, unfortunately, his $M$-estimators have low breakdown point in the sense that the $M$-estimation breaks down when the fraction of outliers is more than $1/(p + 1)$. Some authors consider estimators of multivariate location and covariance that have a high breakdown point (Rousseeuw and van Zomeren 1990, Rocke and Woodruff 1996). Although the performance of their proposed methods should be good in avoiding masking, these are computationally very expensive. We propose a procedure which is easy to compute, and works well even with heavily contaminated data. We use the $M$-estimation technique proposed by Campbell (1980) with some modifications.

We suggest finding the $M$-estimators by solving iteratively the simultaneous equations
and

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

and

$$S^* = \frac{\sum_{i=1}^{n} w_i^2 (x_i - \bar{x}^*)(x_i - \bar{x}^*)'}{\sum_{i=1}^{n} w_i^2 - 1}$$

where

$$w_i = \frac{\omega(d_i)}{d_i}$$

and

$$d_i = \left\{ (x_i - \bar{x}^*)' S^{*-1}(x_i - \bar{x}^*) \right\}^{\frac{1}{2}}.$$

The weight function $\omega(d)$ is defined as

$$\omega(d) = \begin{cases} 
  d & \text{if } d \leq d_0, \\
  d_0 \exp\left\{ -\frac{1}{2} (d - d_0)^2 / b^2 \right\} & \text{if } d > d_0,
\end{cases}$$

with $b = 1.25$ (as suggested by Hampel 1973) and

$$d_0 = \left\{ \frac{(n - 1)^2}{n} \left[ 1 - \left( 1 + \frac{p}{n - p - 1} F_{p,n-p-1,1-\frac{a}{n}} \right)^{-1} \right] \right\}^{\frac{1}{2}}.$$

The motivation behind this form of $d_o$ is that the Bonferroni bounds of $\max_i MD_i$ (discussed in the next section) can be obtained from the upper $100(1 - \frac{a}{n})$ percentage points of the $F_{p,n-p-1}$ distribution by using the transformation

$$\frac{(n - 1)^2}{n} \left[ 1 - \left( 1 + \frac{p}{n - p - 1} F_{p,n-p-1,1-\frac{a}{n}} \right)^{-1} \right].$$

38
Now, the important issue is finding a good starting point for the location and scatter, (i.e., a good initial estimator of $\bar{x}^*$ and $S^*$). For the initial estimators of location and scale, we follow the \textit{minimum volume ellipsoid (MVE)} estimators found by Rousseeuw and van Zomeren (1990). The authors developed an approximate algorithm for the \textit{MVE} estimators of location and scale. We have discussed the algorithm in our review section. Although the algorithm is a bit complicated, but it is very useful in avoiding the masking effect. And fortunately, computer software has been developed in \textit{Splus} for the \textit{MVE} estimators. Here notice that we do not suggest the approximate \textit{MVE} estimators as the final robust estimators, because these estimators do not guarantee the level of our consecutive test for the detection of outliers.

After obtaining primarily the estimates of location and scatter, we find the Mahalanobis distances $d_i$'s based on the location and scale estimators. Then we find the weight function $\omega(d)$, and hence the estimates of location and scatter $\bar{x}^*$ and $S^*$, respectively. This procedure should be repeated until we meet a stopping criterion. In practice, we have used 25 iterations for light contamination and 50 iterations for heavy contamination of the data. We also need to increase the number of iterations for high dimensionality. The robust estimates and distances have been computed by means of a Splus package.

In the next section, we have discussed the sequential application of the one-outlier test for the detection of many outliers in the multivariate data.
3.3 Application of the Sequential Test

After finding the statistic $Q_1$, our next problem is to find a critical value $q_1$ such that

$$P[Q_1 > q_1 | H_0] = \alpha.$$ 

In developing the critical value $q_1$ of $Q_1$ we assume that under $H_0$ (the null hypothesis of no outliers), $RD_i = MD_i$. Although this is an approximation, empirical evidence shows that it is likely to be quite accurate. The critical value $q_1$ can be obtained using Bonferroni bounds. We have

$$\alpha = P[Q_1 > q_1 | H_0]$$

$$= P \left[ \max_i (x_i - \bar{x})'S^{-1}(x_i - \bar{x}) > q_1 | H_0 \right]$$

$$= P \left[ \max_i (x_i - \bar{x})'S^{-1}(x_i - \bar{x}) > q_1 \right]$$

$$= P \left[ \max_i MD_i > q_1 \right]$$

$$= P[\text{at least one } MD_i > q_1]$$

$$\leq \sum_{i=1}^{n} P[MD_i > q_1]$$

$$= \alpha_1 + \alpha_2 + \ldots + \alpha_n \ (\text{say}).$$
If we take $\alpha_1 = \alpha_2 = \ldots = \alpha_n = \alpha/n$, then

$$P[Q_1 > q_1 | H_0] \leq \alpha.$$  

Thus we can find $q_1$ such that

$$\frac{\alpha}{n} = P[MD_i > q_1]$$

$$= P \left\{ (x_i - \bar{x})' S^{-1}(x_i - \bar{x}) > q_1 \right\}$$

$$= P \left[ 1 - \frac{n}{(n-1)^2} (x_i - \bar{x})' S^{-1}(x_i - \bar{x}) \leq 1 - \frac{n}{(n-1)^2} q_1 \right]$$

$$= P \left[ R_i \leq 1 - \frac{n}{(n-1)^2} q_1 \right]$$

$$= P \left[ \left( 1 + \frac{T_i^2}{n-2} \right)^{-1} \leq 1 - \frac{n}{(n-1)^2} q_1 \right]$$

$$= P \left[ \left( \frac{n-p-1}{p} \right) \frac{T_i^2}{n-2} > \left( \frac{n-p-1}{p} \right) \frac{n}{(n-1)^2} q_1 \right]$$

$$= P \left[ F_{p,n-p-1} > \left( \frac{n-p-1}{p} \right) \frac{n}{(n-1)^2} q_1 \right]$$

which implies that

$$\left( \frac{n-p-1}{p} \right) \frac{n}{(n-1)^2} q_1 \frac{n}{(n-1)^2} q_1 = F_{p,n-p-1,1-\frac{\alpha}{n}}$$

41
that is,
\[ q_1 = \frac{(n-1)^2}{n} \left[ 1 - \left( 1 + \frac{p}{n-p} F_{p,n-p-1,1-\frac{2}{n}} \right)^{-1} \right]. \]

Thus the critical value of \( Q_1 \) can be obtained using the \( F \)-distribution. If the observed value of \( Q_1 \) is significantly large, that is \( Q_1 > q_1 \), then the point corresponding to this \( Q_1 \) may be declared as an outlier.

So far we have discussed the test for a single outlier in a multivariate normal sample. Now, we shall discuss how this single-outlier test can be applied to a many-outlier test in samples of reduced size. Suppose, we have data consisting of \( n \) points in dimension \( p \). Initially, we find the statistic \( Q_1 \) and compare it with the corresponding critical value \( q_1 \). If a point is identified as an outlier, we set it aside and test for another outlier in the remaining sample of \( (n-1) \) observations. Here we find the maximum robust distance based on this sample of \( (n-1) \) points. This maximum distance may be denoted by \( Q_2 \). Then we find the corresponding critical value \( q_2 \) of \( Q_2 \) from the relation
\[ q_2 = \frac{(n-2)^2}{(n-1)} \left[ 1 - \left( 1 + \frac{p}{n-p-2} F_{p,n-p-2,1-\frac{2}{n-1}} \right)^{-1} \right]. \]

If \( Q_2 \) is significantly large, then the point corresponding to \( Q_2 \) should be declared as an outlier. Proceeding similarly, we can test the maximum number of outliers \( k \) by computing the series of test statistics \( Q_1, Q_2, \ldots, Q_k \) as described earlier and then comparing \( Q_k, Q_{k-1}, \ldots, Q_1 \) with the corresponding critical values \( q_k, q_{k-1}, \ldots, q_1 \). In fact,
\[ q_j = \frac{(n-j)^2}{(n-j+1)} \left[ 1 - \left( 1 + \frac{p}{n-p-j} F_{p,n-p-j,1-\frac{2}{n-j+1}} \right)^{-1} \right]. \]

42
for $j = 1, 2, \ldots, k$. The number of outliers is the highest value $r$ for which $Q_r > q_r$ ($r = 1, 2, \ldots$) is true. There is no outlier in the data if none of these inequalities is satisfied.

The entire sequential procedure for the detection of multiple outliers can be arranged as follows:

**Step 1.** On entry, we have data consisting of $n$ points in dimension $p$. Use the $MVE$ estimators of location and shape as a starting point for the $M$-estimation.

**Step 2.** Find the $M$-estimators of location and shape. (The $M$-estimation is described in Section 3.2.)

**Step 3.** Find the Mahalanobis distances of each point from the $M$-estimator of location, using the $M$-estimator of covariance matrix. Find the maximum distance $Q_1$.

**Step 4.** Compare the maximum distance $Q_1$ with the corresponding cutoff value $q_1$. If $Q_1$ is significantly large, declare the point corresponding to $Q_1$ as an outlier. If $Q_1$ is smaller than $q_1$, we may conclude that there is no outlier in the data. However, if a point is declared as an outlier, we may test for two outliers in the data using samples of reduced size.

**Step 5.** If a point is identified as an outlier, put it aside, and test for another outlier in the remaining $(n - 1)$ points. Find the maximum distance $Q_2$ based on the sample of $(n - 1)$ points using the similar arguments as in the steps 1-3. Compare
$Q_2$ with the corresponding critical value $q_2$. If $Q_2$ is found as significantly large, we may conclude that the point corresponding to $Q_2$ is also an outlier.

**Step 6.** If two points are identified as outliers, set them aside and test for another outlier in the remaining $(n - 2)$ points using the similar arguments. Continue the above process until all outliers are identified.

In this way, for maximum number of outliers $k$, we can compute the series of test statistics $Q_1, Q_2, ..., Q_k$ and then compare them with the corresponding cutoff values $q_1, q_2, ..., q_k$. The number of outliers is the highest value $r$ for which the inequality $Q_r > q_r$ ($r = 1, 2, ...$) is satisfied. If none of these inequalities is true, then we may conclude that there does not exist any outlier in the data.

**Remark.** In our sequential test we have found the critical values $q$'s such that

$$P \left[ Q_{s+1} > q_{s+1} | H_s \right] = \alpha$$

(for $s = 0, 1, ..., k - 1$), where $H_s$ is the hypothesis that there are $s$ outliers in the data which behaves like the null hypothesis when $Q_{s+1}$ is being tested.

Rosner (1983), Caroni and Prescott (1992) point out that we can also find the critical values $q$'s from the expression

$$P \left[ \bigcup_{j=s+1}^{k} (Q_j > q_j | H_s) \right] = \alpha',$$

which is more precise way of finding the critical values. However, we can assume that $\alpha'$ is close to $\alpha$. In fact.
\[ \alpha' = P \left( Q_j > q_j | H_s \right) \]

\[ = P \left( Q_{s+1} > q_{s+1} | H_s \right) \cup \left\{ \bigcup_{j=s+2}^{k} (Q_j > q_j | H_s) \right\} \]

\[ = P[Q_{s+1} > q_{s+1} | H_s] + P \left( Q_{s+1} \leq q_{s+1} | H_s \right) \cap \left\{ \bigcup_{j=s+2}^{k} (Q_j > q_j | H_s) \right\} \]

\[ = \alpha + P \left( Q_{s+1} \leq q_{s+1} | H_s \right) \cap \left\{ \bigcup_{j=s+2}^{k} (Q_j > q_j | H_s) \right\} . \]

It is obvious from the above equations that \( \alpha \) is smaller than \( \alpha' \), the difference being the probability of the event

\[ (Q_{s+1} \leq q_{s+1}) \cap \left\{ \bigcup_{j=s+2}^{k} (Q_j > q_j) \right\} \]

under \( H_s \), which is reasonably assumed to be small. Rosner (1983) considers this approximation for the univariate case while Caroni and Prescott (1992) consider it for the multivariate case. They found that for samples larger than 25 the approximation was very good.
CHAPTER 4

Applications

4.1 Computational Issues

From the viewpoint of applications, the $M$-estimators of location and scale can be considered as a simple modification of the classical estimators. This type of estimation gives full weight to observations assumed to come from the good data under the null hypothesis, but reduced weight or influence to atypical observations from the bad data under the slippage alternative. The Mahalanobis distances play a basic role in multivariate $M$-estimation. In practice, the influence of observations with unduly large Mahalanobis distances is downweighted.

Although we use the Mahalanobis distance as test criterion for the detection of outliers, the weights $w_i^2$ associated with the distances $d_i^2$ (as defined earlier) also indicate atypical observations. In each experiment of our consecutive test, a weight of less than unity indicates an atypical observation: a weight of unity is associated with a reasonable observation.

In the next section we do a simulation study to examine the level of our sequential test when the data are multivariate normal. We also study the power of the test under different alternatives to the null hypothesis that all observations in the sample come
from a \( p \)-dimensional normal distribution. For the power study, we consider different sets of contaminated data with \textit{shift outliers}, in which the outlying values are generated from a distribution with the same covariance matrix and a shifted mean. Rocke and Woodruff (1996) point out that shift outliers are the hardest kind of outliers to find. To contaminate some data with shift outliers they consider a situation in which there are good data drawn from a multivariate normal distribution and bad data drawn from the same distribution and then displaced. The same idea is considered here for the contamination of the data with that kind of outliers.

\section*{4.2 Simulation Results}

A simulation study is carried out to support the good behaviour of the proposed consecutive test when the data are multivariate normal. Each line of the Table 1 is based on 5000 sets of multivariate normal data which are generated using the Splus package. For each multivariate data set we test the null hypothesis that there is no outlier in the data against the alternative that there is at least one outlier. The third column is the fraction out of the total of 5000 data sets that are rejected as outliers. From these results it is easily seen that the risk declaring outliers when there are none is close to the specified Type I error for relatively large samples. We assume that the Type I error is the probability of finding 1 or more outliers when there are actually no outliers in the data. For small samples, however, the actual Type I error is more than the nominal Type I error. It is also clear from the results that for high dimensionality, the sample size should be increased for obtaining actual Type I error near the nominal rejection fraction.
Table 1: Actual Type I errors when the nominal Type I error is .05.

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Sample Size</th>
<th>Fraction Rejected</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p$</td>
<td>$n$</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>0.0616</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>0.0566</td>
</tr>
<tr>
<td>2</td>
<td>25</td>
<td>0.0506</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>0.0502</td>
</tr>
<tr>
<td>3</td>
<td>15</td>
<td>0.1698</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>0.0626</td>
</tr>
<tr>
<td>3</td>
<td>25</td>
<td>0.0518</td>
</tr>
<tr>
<td>3</td>
<td>30</td>
<td>0.0507</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>0.1312</td>
</tr>
<tr>
<td>4</td>
<td>25</td>
<td>0.0546</td>
</tr>
<tr>
<td>4</td>
<td>30</td>
<td>0.0564</td>
</tr>
<tr>
<td>4</td>
<td>40</td>
<td>0.0504</td>
</tr>
<tr>
<td>5</td>
<td>25</td>
<td>0.1376</td>
</tr>
<tr>
<td>5</td>
<td>30</td>
<td>0.0584</td>
</tr>
<tr>
<td>5</td>
<td>40</td>
<td>0.0530</td>
</tr>
</tbody>
</table>

To assess the appropriateness of our proposed sequential test in the presence of outliers, a simulation study was carried out with data of different sample sizes, and dimensions $p = 5$ and $p = 10$. The data were contaminated by adding a slippage
vector to individual observations. That is, the outliers were produced by adding a given amount (slippage) to each dimension.

Table 2: Performance of the consecutive test in the presence of outliers, at the nominal 5% level.

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Sample Size</th>
<th>No. of Contaminants</th>
<th>Slippage</th>
<th>Success Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>25</td>
<td>3</td>
<td>4</td>
<td>80</td>
</tr>
<tr>
<td>5</td>
<td>25</td>
<td>3</td>
<td>8</td>
<td>78</td>
</tr>
<tr>
<td>5</td>
<td>50</td>
<td>3</td>
<td>4</td>
<td>92</td>
</tr>
<tr>
<td>5</td>
<td>50</td>
<td>3</td>
<td>8</td>
<td>94</td>
</tr>
<tr>
<td>5</td>
<td>50</td>
<td>5</td>
<td>4</td>
<td>96</td>
</tr>
<tr>
<td>5</td>
<td>50</td>
<td>5</td>
<td>8</td>
<td>92</td>
</tr>
<tr>
<td>5</td>
<td>80</td>
<td>10</td>
<td>4</td>
<td>98</td>
</tr>
<tr>
<td>10</td>
<td>50</td>
<td>3</td>
<td>3</td>
<td>86</td>
</tr>
<tr>
<td>10</td>
<td>50</td>
<td>3</td>
<td>6</td>
<td>92</td>
</tr>
<tr>
<td>10</td>
<td>100</td>
<td>3</td>
<td>3</td>
<td>92</td>
</tr>
<tr>
<td>10</td>
<td>100</td>
<td>3</td>
<td>6</td>
<td>94</td>
</tr>
<tr>
<td>10</td>
<td>50</td>
<td>5</td>
<td>3</td>
<td>94</td>
</tr>
<tr>
<td>10</td>
<td>50</td>
<td>5</td>
<td>6</td>
<td>84</td>
</tr>
<tr>
<td>10</td>
<td>100</td>
<td>10</td>
<td>3</td>
<td>98</td>
</tr>
<tr>
<td>10</td>
<td>100</td>
<td>10</td>
<td>6</td>
<td>92</td>
</tr>
<tr>
<td>10</td>
<td>150</td>
<td>20</td>
<td>3</td>
<td>88</td>
</tr>
<tr>
<td>10</td>
<td>150</td>
<td>20</td>
<td>6</td>
<td>94</td>
</tr>
</tbody>
</table>
Table 2 shows the results for different number of outliers in the data with a nominal 5% level. Each line of the table is based on 50 sets of standard normal data. To contaminate each data set, a certain amount (slippage) was added to each dimension of a certain number of observations. For example, in the first line of Table 2, we have drawn a set of 50 random samples each of size 25 in dimension 5. Then we contaminate the last three points of each sample by adding the amount 4 (slippage) to each dimension of these points. Here we have used the same technique as Rocke and Woodruff (1996) used in contamination of a data set with shift outliers. And our results are also similar to their results. The last column is the success rate which represents the percentage of the data sets in which all contaminated points were declared as outliers correctly and no other points were identified as outliers. From the results it is easily seen that the success rate is high when the sample size is large.

4.3 Some Examples

Example 1. As a simple example, let us consider the constructed data in Table 3 having 25 observations of the two-dimensional random variable X. We have constructed the data by contaminating the last five points by adding a slippage amount 8 to each dimension of these points. Our purpose was to find if there was any masking effect in the contaminated data and, if so, how good was the performance of our test procedure in avoiding the masking. We were also interested in finding the effect of masking in classical approach of identification.
Table 3: Constructed data with some contamination.

<table>
<thead>
<tr>
<th>Observation</th>
<th>$X_1$</th>
<th>$X_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.6496</td>
<td>0.3224</td>
</tr>
<tr>
<td>2</td>
<td>-0.3497</td>
<td>0.7260</td>
</tr>
<tr>
<td>3</td>
<td>1.2366</td>
<td>-1.0805</td>
</tr>
<tr>
<td>4</td>
<td>-0.7887</td>
<td>0.1085</td>
</tr>
<tr>
<td>5</td>
<td>1.0545</td>
<td>1.3387</td>
</tr>
<tr>
<td>6</td>
<td>-2.1357</td>
<td>-1.6633</td>
</tr>
<tr>
<td>7</td>
<td>-0.7011</td>
<td>0.5172</td>
</tr>
<tr>
<td>8</td>
<td>-1.6238</td>
<td>-0.4432</td>
</tr>
<tr>
<td>9</td>
<td>-0.6055</td>
<td>0.9121</td>
</tr>
<tr>
<td>10</td>
<td>-0.9255</td>
<td>-0.2339</td>
</tr>
<tr>
<td>11</td>
<td>1.4735</td>
<td>0.9712</td>
</tr>
<tr>
<td>12</td>
<td>1.8996</td>
<td>0.9257</td>
</tr>
<tr>
<td>13</td>
<td>0.5378</td>
<td>-0.2982</td>
</tr>
<tr>
<td>14</td>
<td>-0.6636</td>
<td>-0.3663</td>
</tr>
<tr>
<td>15</td>
<td>0.6417</td>
<td>-0.4822</td>
</tr>
<tr>
<td>16</td>
<td>-0.9995</td>
<td>-1.6703</td>
</tr>
<tr>
<td>17</td>
<td>1.1638</td>
<td>0.1800</td>
</tr>
<tr>
<td>18</td>
<td>0.7322</td>
<td>-0.2544</td>
</tr>
<tr>
<td>19</td>
<td>-2.3298</td>
<td>0.4268</td>
</tr>
<tr>
<td>20</td>
<td>-0.6704</td>
<td>0.4753</td>
</tr>
<tr>
<td>21</td>
<td>7.6260</td>
<td>7.8482</td>
</tr>
<tr>
<td>22</td>
<td>8.6317</td>
<td>8.3571</td>
</tr>
</tbody>
</table>
Table 3: Constructed data with some contamination.

<table>
<thead>
<tr>
<th>Observation</th>
<th>$X_1$</th>
<th>$X_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>6.4212</td>
<td>7.3422</td>
</tr>
<tr>
<td>24</td>
<td>8.6213</td>
<td>7.6830</td>
</tr>
<tr>
<td>25</td>
<td>9.1867</td>
<td>8.9677</td>
</tr>
</tbody>
</table>

In Table 3, we have data of 25 observations in dimension 2. Our purpose is to identify if there are any, and if so how many, outliers in the data. Figure 1 illustrates the distinction between classical and robust estimates (MVE robust estimates proposed by Rousseeuw and van Zomeren 1990). We start with the concept of "tolerance ellipse" which is used in Rousseeuw and van Zomeren (1990). We see that the 97.5% tolerance ellipse obtained from the classical estimates (dashed line) is blown up by the outliers, and contains all the sample points. The tolerance ellipse based on the MVE is much narrower (solid line) and does not contain the outliers.

Now, let us apply our sequential procedure to the data in Table 3. The results are given in Table 4. We see that in the 1st stage the point 25 has the maximum robust distance 128.42 which is highly significant. So we declare it as an outlier. We put it aside and test for another outlier in the remaining 24 observations. Here we see that the point 22 is also declared as an outlier. Proceeding similarly, the identified outliers, from largest to the smallest, are the points 25, 22, 24, 21 and 23. Here notice that if we use the classical sequential method, no point is declared as an outlier because of masking.
Figure 1: Plot of the constructed data in Table 1. with 97.5% tolerance ellipse based on the classical mean and covariance (dashed line) and on the robust estimators of mean and covariance (solid line).
Table 4: Results of the sequential procedure for the identification of multiple outliers using data in Table 3.

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Selected Point</th>
<th>Maximum Robust Distance</th>
<th>5% Critical Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>25</td>
<td>128.42</td>
<td>9.94</td>
</tr>
<tr>
<td>24</td>
<td>22</td>
<td>111.94</td>
<td>9.80</td>
</tr>
<tr>
<td>23</td>
<td>24</td>
<td>98.69</td>
<td>9.65</td>
</tr>
<tr>
<td>22</td>
<td>21</td>
<td>96.21</td>
<td>9.48</td>
</tr>
<tr>
<td>21</td>
<td>23</td>
<td>81.29</td>
<td>9.31</td>
</tr>
<tr>
<td>20</td>
<td>6</td>
<td>5.00</td>
<td>9.13</td>
</tr>
</tbody>
</table>

Example 2. Here we consider the famous Hawkins, Bradu, and Kass (1984) data, which is a constructed data consisting of 75 points in 3 explanatory variables plus a response variable. We are interested in examining if there are any leverage points in the data of three predictor variables. Rousseeuw and van Zomeren (1990) find that there are 14 leverage points in the data, but only two points show up as such if ordinary Mahalanobis distances are used. The same conclusion is obtained from our sequential analysis of identification of outliers. Table 5 exhibits the outcomes of our sequential analysis. Here we see that the leverage points, from largest to the smallest, are the points 14, 12, 13, 11, 4, 5, 9, 3, 10, 7, 6, 2, 8 and 1, agreeing with the conclusions of Rousseeuw and van Zomeren (1990), Rocke and Woodruff (1996).

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Selected Point</th>
<th>Maximum Robust Distance</th>
<th>5% Critical Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>75</td>
<td>14</td>
<td>1688.50</td>
<td>15.55</td>
</tr>
<tr>
<td>74</td>
<td>12</td>
<td>1440.59</td>
<td>15.51</td>
</tr>
<tr>
<td>73</td>
<td>13</td>
<td>1362.91</td>
<td>15.46</td>
</tr>
<tr>
<td>72</td>
<td>11</td>
<td>1342.37</td>
<td>15.42</td>
</tr>
<tr>
<td>71</td>
<td>4</td>
<td>1079.92</td>
<td>15.37</td>
</tr>
<tr>
<td>70</td>
<td>5</td>
<td>1041.86</td>
<td>15.33</td>
</tr>
<tr>
<td>69</td>
<td>9</td>
<td>1021.04</td>
<td>15.28</td>
</tr>
<tr>
<td>68</td>
<td>3</td>
<td>1017.03</td>
<td>15.23</td>
</tr>
<tr>
<td>67</td>
<td>10</td>
<td>957.46</td>
<td>15.18</td>
</tr>
<tr>
<td>66</td>
<td>7</td>
<td>941.31</td>
<td>15.13</td>
</tr>
<tr>
<td>65</td>
<td>6</td>
<td>935.70</td>
<td>15.08</td>
</tr>
<tr>
<td>64</td>
<td>2</td>
<td>912.37</td>
<td>15.03</td>
</tr>
<tr>
<td>63</td>
<td>8</td>
<td>888.01</td>
<td>14.98</td>
</tr>
<tr>
<td>62</td>
<td>1</td>
<td>866.85</td>
<td>14.92</td>
</tr>
<tr>
<td>61</td>
<td>53</td>
<td>6.33</td>
<td>14.87</td>
</tr>
</tbody>
</table>

**Example 3.** Here we consider the data used by Caroni and Prescott (1992) involved the cost of transportation of milk from farm to dairy plants taken from Johnson and Wichern (1982). The data consist of 36 observations with three variables, fuel costs, repair, and capital costs. We are interested in identifying possible outliers in the
data. The sequential analysis of this data is presented in Table 6. We see that the identified outliers are the points 9 and 21, agreeing with the conclusions of Caroni and Prescott (1992) where they use classical sequential application of multiple outlier test.

Table 6: Sequential analysis of the data used in Caroni and Prescott (1992).

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Selected Point</th>
<th>Maximum Robust Distance</th>
<th>5% Critical Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>36</td>
<td>9</td>
<td>47.92</td>
<td>12.96</td>
</tr>
<tr>
<td>35</td>
<td>21</td>
<td>23.34</td>
<td>12.85</td>
</tr>
<tr>
<td>34</td>
<td>36</td>
<td>9.42</td>
<td>12.74</td>
</tr>
</tbody>
</table>

Example 4. In this example, we consider the data taken from Johnson and Wichern (1982) which is called "Battery-Failure Data". The data consist of 20 points in 5 explanatory variables plus a response variable. Our purpose is to identify if there are any leverage points in the data of 5 predictor variables. The data are reproduced in Table 7 omitting the response variable.
Table 7: Battery-Failure Data (Johnson et al 1982).

<table>
<thead>
<tr>
<th>Observation</th>
<th>$X_1$ Charge rate</th>
<th>$X_2$ Discharge rate</th>
<th>$X_3$ Depth of discharge</th>
<th>$X_4$ Temperature</th>
<th>$X_5$ End of charge voltage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.375</td>
<td>3.13</td>
<td>60.0</td>
<td>40</td>
<td>2.00</td>
</tr>
<tr>
<td>2</td>
<td>1.000</td>
<td>3.13</td>
<td>76.8</td>
<td>30</td>
<td>1.99</td>
</tr>
<tr>
<td>3</td>
<td>1.000</td>
<td>3.13</td>
<td>60.0</td>
<td>20</td>
<td>2.00</td>
</tr>
<tr>
<td>4</td>
<td>1.000</td>
<td>3.13</td>
<td>60.0</td>
<td>20</td>
<td>1.98</td>
</tr>
<tr>
<td>5</td>
<td>1.625</td>
<td>3.13</td>
<td>43.2</td>
<td>10</td>
<td>2.01</td>
</tr>
<tr>
<td>6</td>
<td>1.625</td>
<td>3.13</td>
<td>60.0</td>
<td>20</td>
<td>2.00</td>
</tr>
<tr>
<td>7</td>
<td>1.625</td>
<td>3.13</td>
<td>60.0</td>
<td>20</td>
<td>2.02</td>
</tr>
<tr>
<td>8</td>
<td>0.375</td>
<td>5.00</td>
<td>76.8</td>
<td>10</td>
<td>2.01</td>
</tr>
<tr>
<td>9</td>
<td>1.000</td>
<td>5.00</td>
<td>43.2</td>
<td>10</td>
<td>1.99</td>
</tr>
<tr>
<td>10</td>
<td>1.000</td>
<td>5.00</td>
<td>43.2</td>
<td>30</td>
<td>2.01</td>
</tr>
<tr>
<td>11</td>
<td>1.000</td>
<td>5.00</td>
<td>100.0</td>
<td>20</td>
<td>2.00</td>
</tr>
<tr>
<td>12</td>
<td>1.625</td>
<td>5.00</td>
<td>76.8</td>
<td>10</td>
<td>1.99</td>
</tr>
<tr>
<td>13</td>
<td>0.375</td>
<td>1.25</td>
<td>76.8</td>
<td>10</td>
<td>2.01</td>
</tr>
<tr>
<td>14</td>
<td>1.000</td>
<td>1.25</td>
<td>43.2</td>
<td>10</td>
<td>1.99</td>
</tr>
<tr>
<td>15</td>
<td>1.000</td>
<td>1.25</td>
<td>76.8</td>
<td>30</td>
<td>2.00</td>
</tr>
<tr>
<td>16</td>
<td>1.000</td>
<td>1.25</td>
<td>60.0</td>
<td>0</td>
<td>2.00</td>
</tr>
<tr>
<td>17</td>
<td>1.625</td>
<td>1.25</td>
<td>43.2</td>
<td>30</td>
<td>1.99</td>
</tr>
<tr>
<td>18</td>
<td>1.625</td>
<td>1.25</td>
<td>60.0</td>
<td>20</td>
<td>2.00</td>
</tr>
<tr>
<td>19</td>
<td>0.375</td>
<td>3.13</td>
<td>76.8</td>
<td>30</td>
<td>1.99</td>
</tr>
<tr>
<td>20</td>
<td>0.375</td>
<td>3.13</td>
<td>60.0</td>
<td>20</td>
<td>2.00</td>
</tr>
</tbody>
</table>
For this data, the maximum robust distance is obtained as 7.23 which is insignificant (the critical value is 12.62). So no points are identified as leverage points in this data. Here notice that the maximum of the simple Mahalanobis distances based on the classical estimates of mean and covariance is also 7.23 indicating no leverage points in the data.

Example 5. As a final example, we consider the following constructed data in Table 8 which consist of 50 observations with 8 variables. Our purpose was to examine the performance of the proposed sequential test in the case of high dimensionality. We have constructed the data by contaminating the last 8 points by adding a slip-page vector to each of these observations. Then we apply the sequential procedure of many-outlier test for this data.
Table 8: Constructed data consisting of 50 observations in 8 dimensions.

<table>
<thead>
<tr>
<th>Observation</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$X_5$</th>
<th>$X_6$</th>
<th>$X_7$</th>
<th>$X_8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.96</td>
<td>2.89</td>
<td>1.91</td>
<td>2.78</td>
<td>1.69</td>
<td>1.61</td>
<td>2.24</td>
<td>2.21</td>
</tr>
<tr>
<td>2</td>
<td>2.22</td>
<td>2.54</td>
<td>2.25</td>
<td>3.36</td>
<td>1.88</td>
<td>1.67</td>
<td>1.67</td>
<td>2.07</td>
</tr>
<tr>
<td>3</td>
<td>2.82</td>
<td>2.02</td>
<td>2.90</td>
<td>3.56</td>
<td>2.37</td>
<td>2.72</td>
<td>3.24</td>
<td>2.77</td>
</tr>
<tr>
<td>4</td>
<td>1.44</td>
<td>0.86</td>
<td>1.98</td>
<td>0.89</td>
<td>1.84</td>
<td>1.50</td>
<td>0.77</td>
<td>0.32</td>
</tr>
<tr>
<td>5</td>
<td>1.86</td>
<td>2.19</td>
<td>1.90</td>
<td>1.01</td>
<td>1.84</td>
<td>2.94</td>
<td>2.67</td>
<td>1.82</td>
</tr>
<tr>
<td>6</td>
<td>1.28</td>
<td>0.12</td>
<td>0.95</td>
<td>3.30</td>
<td>1.74</td>
<td>0.50</td>
<td>1.33</td>
<td>1.83</td>
</tr>
<tr>
<td>7</td>
<td>1.40</td>
<td>2.41</td>
<td>1.87</td>
<td>1.66</td>
<td>3.39</td>
<td>2.22</td>
<td>2.82</td>
<td>3.22</td>
</tr>
<tr>
<td>8</td>
<td>2.19</td>
<td>2.14</td>
<td>2.96</td>
<td>2.88</td>
<td>2.71</td>
<td>0.89</td>
<td>2.77</td>
<td>1.03</td>
</tr>
<tr>
<td>9</td>
<td>3.15</td>
<td>1.31</td>
<td>2.03</td>
<td>1.20</td>
<td>1.98</td>
<td>1.80</td>
<td>1.40</td>
<td>0.35</td>
</tr>
<tr>
<td>10</td>
<td>1.33</td>
<td>2.20</td>
<td>1.62</td>
<td>2.01</td>
<td>2.58</td>
<td>2.09</td>
<td>1.33</td>
<td>2.62</td>
</tr>
<tr>
<td>11</td>
<td>2.68</td>
<td>2.60</td>
<td>2.28</td>
<td>2.43</td>
<td>1.71</td>
<td>1.29</td>
<td>1.55</td>
<td>2.47</td>
</tr>
<tr>
<td>12</td>
<td>2.99</td>
<td>3.16</td>
<td>2.32</td>
<td>2.59</td>
<td>2.67</td>
<td>3.94</td>
<td>1.92</td>
<td>2.69</td>
</tr>
<tr>
<td>13</td>
<td>2.87</td>
<td>2.56</td>
<td>1.16</td>
<td>2.27</td>
<td>1.13</td>
<td>2.80</td>
<td>1.83</td>
<td>0.99</td>
</tr>
<tr>
<td>14</td>
<td>1.97</td>
<td>2.31</td>
<td>2.80</td>
<td>3.36</td>
<td>2.98</td>
<td>1.44</td>
<td>2.06</td>
<td>1.60</td>
</tr>
<tr>
<td>15</td>
<td>2.28</td>
<td>2.06</td>
<td>1.14</td>
<td>3.08</td>
<td>2.28</td>
<td>0.83</td>
<td>2.41</td>
<td>0.03</td>
</tr>
<tr>
<td>16</td>
<td>0.81</td>
<td>0.50</td>
<td>0.89</td>
<td>0.83</td>
<td>2.21</td>
<td>0.54</td>
<td>0.84</td>
<td>1.89</td>
</tr>
<tr>
<td>17</td>
<td>2.11</td>
<td>1.66</td>
<td>3.00</td>
<td>2.63</td>
<td>2.85</td>
<td>2.89</td>
<td>1.28</td>
<td>1.79</td>
</tr>
<tr>
<td>18</td>
<td>1.37</td>
<td>1.95</td>
<td>1.05</td>
<td>1.33</td>
<td>1.89</td>
<td>0.47</td>
<td>2.10</td>
<td>1.91</td>
</tr>
<tr>
<td>19</td>
<td>3.11</td>
<td>2.52</td>
<td>2.85</td>
<td>1.52</td>
<td>2.31</td>
<td>3.43</td>
<td>2.43</td>
<td>1.25</td>
</tr>
</tbody>
</table>
Table 8: Constructed data consisting of 50 observations in 8 dimensions.

<table>
<thead>
<tr>
<th>Observation</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$X_5$</th>
<th>$X_6$</th>
<th>$X_7$</th>
<th>$X_8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>2.36</td>
<td>1.55</td>
<td>3.55</td>
<td>3.45</td>
<td>3.02</td>
<td>2.88</td>
<td>3.55</td>
<td>2.94</td>
</tr>
<tr>
<td>21</td>
<td>1.94</td>
<td>1.94</td>
<td>2.46</td>
<td>2.08</td>
<td>1.22</td>
<td>3.96</td>
<td>2.75</td>
<td>1.00</td>
</tr>
<tr>
<td>22</td>
<td>2.18</td>
<td>2.24</td>
<td>2.24</td>
<td>1.27</td>
<td>3.40</td>
<td>2.41</td>
<td>0.65</td>
<td>2.68</td>
</tr>
<tr>
<td>23</td>
<td>3.62</td>
<td>3.58</td>
<td>4.00</td>
<td>2.86</td>
<td>3.43</td>
<td>2.96</td>
<td>4.48</td>
<td>3.85</td>
</tr>
<tr>
<td>24</td>
<td>4.34</td>
<td>4.09</td>
<td>2.91</td>
<td>3.70</td>
<td>2.96</td>
<td>4.58</td>
<td>3.26</td>
<td>2.32</td>
</tr>
<tr>
<td>25</td>
<td>1.84</td>
<td>1.48</td>
<td>2.62</td>
<td>1.94</td>
<td>2.17</td>
<td>2.24</td>
<td>1.95</td>
<td>2.90</td>
</tr>
<tr>
<td>26</td>
<td>3.46</td>
<td>2.30</td>
<td>2.75</td>
<td>2.62</td>
<td>1.96</td>
<td>1.82</td>
<td>2.58</td>
<td>0.89</td>
</tr>
<tr>
<td>27</td>
<td>1.39</td>
<td>1.07</td>
<td>0.10</td>
<td>2.70</td>
<td>1.28</td>
<td>0.68</td>
<td>2.47</td>
<td>2.00</td>
</tr>
<tr>
<td>28</td>
<td>4.24</td>
<td>2.18</td>
<td>2.97</td>
<td>3.94</td>
<td>2.41</td>
<td>3.63</td>
<td>3.68</td>
<td>2.08</td>
</tr>
<tr>
<td>29</td>
<td>0.16</td>
<td>0.95</td>
<td>1.30</td>
<td>1.11</td>
<td>0.39</td>
<td>1.43</td>
<td>1.48</td>
<td>1.58</td>
</tr>
<tr>
<td>30</td>
<td>1.64</td>
<td>1.14</td>
<td>0.35</td>
<td>0.13</td>
<td>1.37</td>
<td>0.19</td>
<td>1.18</td>
<td>1.31</td>
</tr>
<tr>
<td>31</td>
<td>1.22</td>
<td>2.80</td>
<td>0.92</td>
<td>0.74</td>
<td>1.26</td>
<td>3.08</td>
<td>1.28</td>
<td>1.18</td>
</tr>
<tr>
<td>32</td>
<td>1.49</td>
<td>2.71</td>
<td>2.68</td>
<td>2.35</td>
<td>3.13</td>
<td>2.58</td>
<td>3.54</td>
<td>2.65</td>
</tr>
<tr>
<td>33</td>
<td>1.64</td>
<td>1.01</td>
<td>1.98</td>
<td>1.91</td>
<td>0.85</td>
<td>1.19</td>
<td>0.88</td>
<td>0.36</td>
</tr>
<tr>
<td>34</td>
<td>1.76</td>
<td>1.66</td>
<td>1.76</td>
<td>3.07</td>
<td>1.25</td>
<td>2.22</td>
<td>2.07</td>
<td>1.75</td>
</tr>
<tr>
<td>35</td>
<td>1.12</td>
<td>0.24</td>
<td>1.29</td>
<td>0.69</td>
<td>0.41</td>
<td>1.41</td>
<td>0.09</td>
<td>0.14</td>
</tr>
<tr>
<td>36</td>
<td>2.40</td>
<td>0.53</td>
<td>1.85</td>
<td>2.19</td>
<td>2.72</td>
<td>1.90</td>
<td>3.08</td>
<td>4.34</td>
</tr>
<tr>
<td>37</td>
<td>1.90</td>
<td>0.26</td>
<td>2.94</td>
<td>1.89</td>
<td>1.88</td>
<td>1.24</td>
<td>2.03</td>
<td>1.50</td>
</tr>
<tr>
<td>38</td>
<td>3.70</td>
<td>2.61</td>
<td>3.18</td>
<td>2.60</td>
<td>2.77</td>
<td>2.91</td>
<td>2.99</td>
<td>3.24</td>
</tr>
</tbody>
</table>
Table 8: Constructed data consisting of 50 observations in 8 dimensions.

<table>
<thead>
<tr>
<th>Observation</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$X_5$</th>
<th>$X_6$</th>
<th>$X_7$</th>
<th>$X_8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>39</td>
<td>3.51</td>
<td>3.28</td>
<td>2.43</td>
<td>1.82</td>
<td>3.05</td>
<td>2.06</td>
<td>1.82</td>
<td>2.93</td>
</tr>
<tr>
<td>40</td>
<td>2.18</td>
<td>2.68</td>
<td>1.47</td>
<td>2.81</td>
<td>0.87</td>
<td>3.33</td>
<td>2.42</td>
<td>1.53</td>
</tr>
<tr>
<td>41</td>
<td>1.11</td>
<td>0.99</td>
<td>2.48</td>
<td>2.06</td>
<td>3.06</td>
<td>2.31</td>
<td>1.54</td>
<td>1.07</td>
</tr>
<tr>
<td>42</td>
<td>2.25</td>
<td>2.52</td>
<td>3.30</td>
<td>2.18</td>
<td>1.66</td>
<td>2.72</td>
<td>4.68</td>
<td>3.39</td>
</tr>
<tr>
<td>43</td>
<td>2.25</td>
<td>8.19</td>
<td>1.68</td>
<td>4.49</td>
<td>7.01</td>
<td>7.97</td>
<td>6.91</td>
<td>2.37</td>
</tr>
<tr>
<td>44</td>
<td>2.05</td>
<td>8.37</td>
<td>2.51</td>
<td>6.94</td>
<td>7.08</td>
<td>8.65</td>
<td>8.91</td>
<td>4.63</td>
</tr>
<tr>
<td>45</td>
<td>0.75</td>
<td>7.34</td>
<td>2.98</td>
<td>4.08</td>
<td>6.96</td>
<td>9.31</td>
<td>7.16</td>
<td>1.81</td>
</tr>
<tr>
<td>46</td>
<td>2.43</td>
<td>6.28</td>
<td>2.67</td>
<td>5.57</td>
<td>5.63</td>
<td>8.81</td>
<td>7.67</td>
<td>4.43</td>
</tr>
<tr>
<td>47</td>
<td>2.71</td>
<td>7.87</td>
<td>3.34</td>
<td>4.17</td>
<td>5.81</td>
<td>8.94</td>
<td>7.47</td>
<td>3.86</td>
</tr>
<tr>
<td>48</td>
<td>3.51</td>
<td>9.06</td>
<td>4.14</td>
<td>5.08</td>
<td>7.69</td>
<td>8.50</td>
<td>8.79</td>
<td>6.13</td>
</tr>
<tr>
<td>49</td>
<td>2.95</td>
<td>9.13</td>
<td>4.80</td>
<td>5.44</td>
<td>7.73</td>
<td>9.95</td>
<td>7.70</td>
<td>4.84</td>
</tr>
<tr>
<td>50</td>
<td>2.26</td>
<td>7.03</td>
<td>2.48</td>
<td>4.30</td>
<td>6.76</td>
<td>8.58</td>
<td>7.67</td>
<td>3.79</td>
</tr>
</tbody>
</table>

The results of the sequential analysis of multiple outlier test is given in Table 9. From the table, we see that there are 8 identified outliers in the data. These outliers, from largest to the smallest, are the points 45, 44, 43, 49, 50, 48, 46, and 47. Here notice that no points show up as outliers if the classical method of sequential test is used.
Table 9: Sequential analysis of the constructed data in Table 8.

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Selected Point</th>
<th>Maximum Robust Distance</th>
<th>5% Critical Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>45</td>
<td>243.27</td>
<td>21.57</td>
</tr>
<tr>
<td>49</td>
<td>44</td>
<td>224.66</td>
<td>21.45</td>
</tr>
<tr>
<td>48</td>
<td>43</td>
<td>208.87</td>
<td>21.33</td>
</tr>
<tr>
<td>47</td>
<td>49</td>
<td>176.01</td>
<td>21.20</td>
</tr>
<tr>
<td>46</td>
<td>50</td>
<td>172.16</td>
<td>21.07</td>
</tr>
<tr>
<td>45</td>
<td>48</td>
<td>153.68</td>
<td>20.94</td>
</tr>
<tr>
<td>44</td>
<td>46</td>
<td>135.64</td>
<td>20.80</td>
</tr>
<tr>
<td>43</td>
<td>47</td>
<td>132.45</td>
<td>20.66</td>
</tr>
<tr>
<td>42</td>
<td>36</td>
<td>15.85</td>
<td>20.51</td>
</tr>
</tbody>
</table>
Discussion

The purpose of this thesis was to find a easier technique for identifying outlying observations in multivariate data. The proposed consecutive test can be applied very easily on a multivariate sample using Splus package. We have shown by simulation that our proposed sequential test procedure controls the $\alpha$ level for the rejection of the null hypothesis. It can also avoid the masking effect even in the case of heavy contamination. Detection of outliers turns out to be harder when $n/p$ is relatively small. Roughly, we recommend applying our procedure when there are more than 5 observations per dimension, so that $n/p > 5$. Although this thesis mainly focuses on the identification of outliers, but at the same time we also get as a by-product the robust estimates of location and scatter which can accommodate the outliers in such a manner that the weight of an extreme atypical observation (i.e., an extreme outlier) should be zero.

No attempt has been made to compare the power of our consecutive test with other test procedures for identification of multivariate outliers. This would be a much more extensive investigations than studying only the consecutive test presented in this thesis. Such a study remains to be done.
Appendix: S-Plus Code for the Sequential Test

```splus
> distance.rb
function(x, iter, alpha)
{
  # x is the data matrix
  # iter is the number of iterations
  # alpha is the level of the test
  mm <- cov.mve(x)  # for "mve" estimation of location and shape
  xbar <- mm$center  # mve estimator of mean
  xvar <- mm$cov  # mve estimator of covariance
  w <- 1
  for(i in 1:iter) {
    # for M-estimation of location and shape
    d <- mahalanobis(x, xbar, xvar, inverted = FALSE)
    dis <- sqrt(d)
    nu <- ncol(x)
    b2 <- 1.25
    d0 <- percentile(alpha, nu, nrow(x))
    do <- sqrt(d0$c$value)
    for(i in 1:nrow(x)) {
      w[i] <- ifelse(dis[i] <= d0, dis[i], d0 * exp((-(1/2) * (dis[i] - d0)^2/b2^2)))
      w[i] <- w[i]/dis[i]
    }
    w <- matrix(w, nrow(x), 1)
    s0 <- matrix(c(0), ncol(x), ncol(x))
    xx <- cov.wt(x, wt = c(w))
    xbar <- xx$center
    for(i in 1:nrow(x)) {
      s <- cbind(w[i] * (x[i, ] - xbar)) %*% (w[i] * (x[i, ] - xbar))
      s0 <- s0 + s
    }
    xvar <- s0/(sum(w^2) - 1)
  }
  maxdis <- max(d)  # maximum robust distance
  list(wt = c(w), robust.mean = xbar, robust.cov = xvar, robust.dis = d,
       max.dis = maxdis, cutoff.val = d0^2)
}

> percentile
# subprogram for the cutoff value
function(alpha, p, n)
{
  prob <- qf(1 - alpha/n, p, n - p - 1)
  pl <- 1/(1 + p/(n - p - 1) * prob)
  ppl <- ((1 - pl) * (n - 1)^2)/n
  list(c.value = ppl)
}
```

64
REFERENCES


Nair, K. R. (1948), "The Distribution of the Extreme Deviate From the Sample Mean and Its Studentized Form", Biometrika, 35, 118-144.


