THE DETECTION AND TESTING OF MULTIVARIATE OUTLIERS

by

RICHARD ALAN WHITE

B.Sc., The University of British Columbia, 1990

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE

THE FACULTY OF GRADUATE STUDIES
THE DEPARTMENT OF STATISTICS

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

July 1992

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of STATISTICS

The University of British Columbia
Vancouver, Canada

Date JULY 17 1992

Abstract

The classical estimators of multivariate location and scatter for the normal model are the sample mean and sample covariance. However, if outliers are present in the data, the classical estimates can be very inaccurate and robust estimates should be used in their place. Most multivariate robust estimators are very difficult if not impossible to compute, thus limiting their use. I will present some simple approximations that make these estimators computable.

Robust estimation down weighs or completely ignores the outliers. This is not always best because the outliers can contain some very important information about the population. If they can be identified, the outliers can be further investigated and an appropriate action can be taken based on the results. To detect outliers, a sequential multivariate scale-ratio test is proposed. It is based on a non-robust estimate and a robust estimate of scatter and is applied in a forward fashion, removing the most extreme point at each step, until the test fails to indicate the presence of outliers. We will show that this procedure has level α when applied to an uncontaminated sample, is uneffected by swamping or masking and is accurate in detecting outliers. Finally, we will apply the scale-ratio test to several data sets and compare it to the sequential Wilk's outlier test as proposed by C. Caroni and P. Prescott in 1992.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank Dr. Ruben Zamar and Dr. Harry Joe for their help, guidance and support during my Masters program at U.B.C.. Both have played important roles in my approach to statistic. I would also like to thank Dr Hyunshik Lee of Statistic Canada for supporting me during the start of this research.

1 Introduction

The outlier problem has been around for many years. Interest in the subject has been like a roller-coaster ride ranging from periods of intense research to periods of no research. Intuitively, an outlier is an observation which deviates from the the rest of the data to such an extent that it arouses suspicion about its underlying distribution. The non-outlying data will be referred to as the core data. It is assumed that the core data contains more than 50% of the observations in the sample. I will further classify an outlier into at least one of two possible categories, an extreme observation and a contaminant. An extreme observation is a point that lies on the convex hull of the data set. There are two extreme observations in one dimension, the maximum and the minimum. In higher dimensions, the number of extreme observations is quite arbitrary. A contaminant is an observation that is generated from a different distribution than that of the core data. With this definition, it is possible for a contaminant to lie in the middle of the core data but we will assume that all points that are more extreme than a contaminant, relative to some suitable standardization, are themselves contaminants. Our main goal is to detect only those outliers that are contaminants.

In lower dimensions, graphical techniques can be used to detect potential outliers. In univariate samples, the boxplot identifies an outlier as an observation that lies beyond some cutoff point based on the median and inter-quartile range. In two and three dimensions, outliers can be detected by scatterplots and spin plots respectively but the degree of outlyingness is based on the judgement of the observer. Unfortunately, once the dimension is greater than three, no graphical tool exists to identify outliers.

Multivariate outliers appear in two forms, the gross outlier and the more subtle structural outlier. The gross outlier is an observation that appears to be outlying in one or more of the original variables. If boxplots are made of these variables then the

gross outliers will be detected. The structural outlier may not appear to be an outlier in any of the original variables but is outlying relative to the covariance structure of the core data. It may not appear in any of the pairwise scatterplots or three-way spin plots either. In fact, it may only be noticeable if all dimensions of the data are considered simultaneously. This is illustrated for three dimensions in Figure 1.

By suitably rotating the data, a structural outlier can be converted into a gross outlier. Therefore a structural outlier in higher dimensions may be graphically detected if the data are rotated by an orthogonal transformation before the plotting is done. If the true underlying distribution is multivariate normal or elliptically contoured, then the eigenvectors of the true covariance structure for the core data will be the best transformation to use, but this assumes that the outliers are already known. Estimating the covariance structure may help if the outliers are not too damaging, but in certain situations the outliers can distort the classical estimates in such a fashion that the eigenvectors can actually disguise the outliers instead of revealing them. See Figure 2 for an example of this. Random orthogonal transformations can be used in the hope that at least one of them will reveal the outliers, but this approach is inefficient because the outliers may only appear in a small fraction of the transformations and therefore, a large number would have to be generated to ensure a high probability of selecting an appropriate projection.

Given that outliers cannot always be detected through graphical means, one must resort to other methods. There are two general approaches to the problem, diagnostic tests and robust methods of analysis. They attack the problem from opposite points of view and, oddly enough, the advantages of one method tend to be the disadvantages of the other.

The main advantage of a diagnostic test is that it detects the outliers and allows the

investigator to decide how the observation should be dealt with. This is a great advantage because not all outliers are spurious observations. Sometimes the most important information about a sample is contained in the outliers and discarding or ignoring them does more harm than good. Unfortunately, the accuracy of diagnostic tests is very suspect and these inaccuracies can seriously affect their performance. The inaccuracies occur because most diagnostic tests require an advanced knowledge of the number of outliers present in the data. This is mainly due to the masking and swamping effects that outliers can have on the testing procedures. Masking is caused when several outliers are close enough together that the removal of some but not all of them results in little improvement in the scatter estimate for the sample. Therefore, if the test is conducted for fewer than the true number of outliers present in the data, the test may not detect any outliers at all. When the test is conducted for more than the true number of outliers, masking cannot occur. However, under these conditions, the test may still be significant and therefore all the good points that are falsely included with the outliers will be incorrectly labelled as outliers. This is known as the swamping effect. Applying a diagnostic test in a sequential manner will not help either, because a forward procedure cannot avoid the masking problem and a backward procedure cannot avoid the swamping problem.

Robust procedures are designed for the case of several outliers. They can be applied when only an upper bound for the number of outliers is known and they are not affected by the masking effect. They usually result in good estimates for the core data when the sample is contaminated but usually lack efficiency when the sample is uncontaminated. The main problem with robust procedures is that they down-weight or completely ignore the outliers. This takes some of the control out of the investigators hands because the robust procedures decide how the outliers will be dealt with. Furthermore, any

information contained in the outliers is lost.

Since each method is strong where the other is weak, the two approaches to the outlier problem should be combined to produce a diagnostic test that is immune to the effects of masking. This test could then be applied sequentially in a forward fashion to not only detect the outliers but to indicate the number present as well. Furthermore, the test should only have to be applied until it fails to detect the presence of an outlier because it should not be fooled by masked outliers.

In this thesis, I will propose a robust diagnostic tool for detecting outliers in a multivariate context. This tool, which I call the multivariate scale-ratio test, is based on the relative size of two multivariate estimates of scatter, one sensitive to outliers the other highly robust. The size of a matrix can be measured in several ways. The most common is the determinant but others such as the largest eigenvalue will also be considered. Another problem facing this procedure is the computation of a robust multivariate estimates of scatter. Several proposed estimators exist but all that possess a high breakdown point, are computationally prohibitive. To deal with this problem, I propose approximations to these estimators that are relatively easy to compute and appear to cost very little in terms of performance. The properties of the multivariate scale-ratio test will be investigated through several Monte Carlo simulations. These will provide some evidence about the asymptotic distribution, power and other properties of the test. Finally the multivariate scale-ratio test will be compared to its main competitor, the multivariate Wilk's outlier test. For comparative purposes, the Wilk's outlier test will be applied in the sequential fashion proposed by Caroni and Prescott [3] in 1992. The comparison will be made by applying the two tests to several real and synthetic data sets in order to evaluate their relative performances.

Figure 1: Outliers in 3-dimensions that are Hidden in 2-dimensions

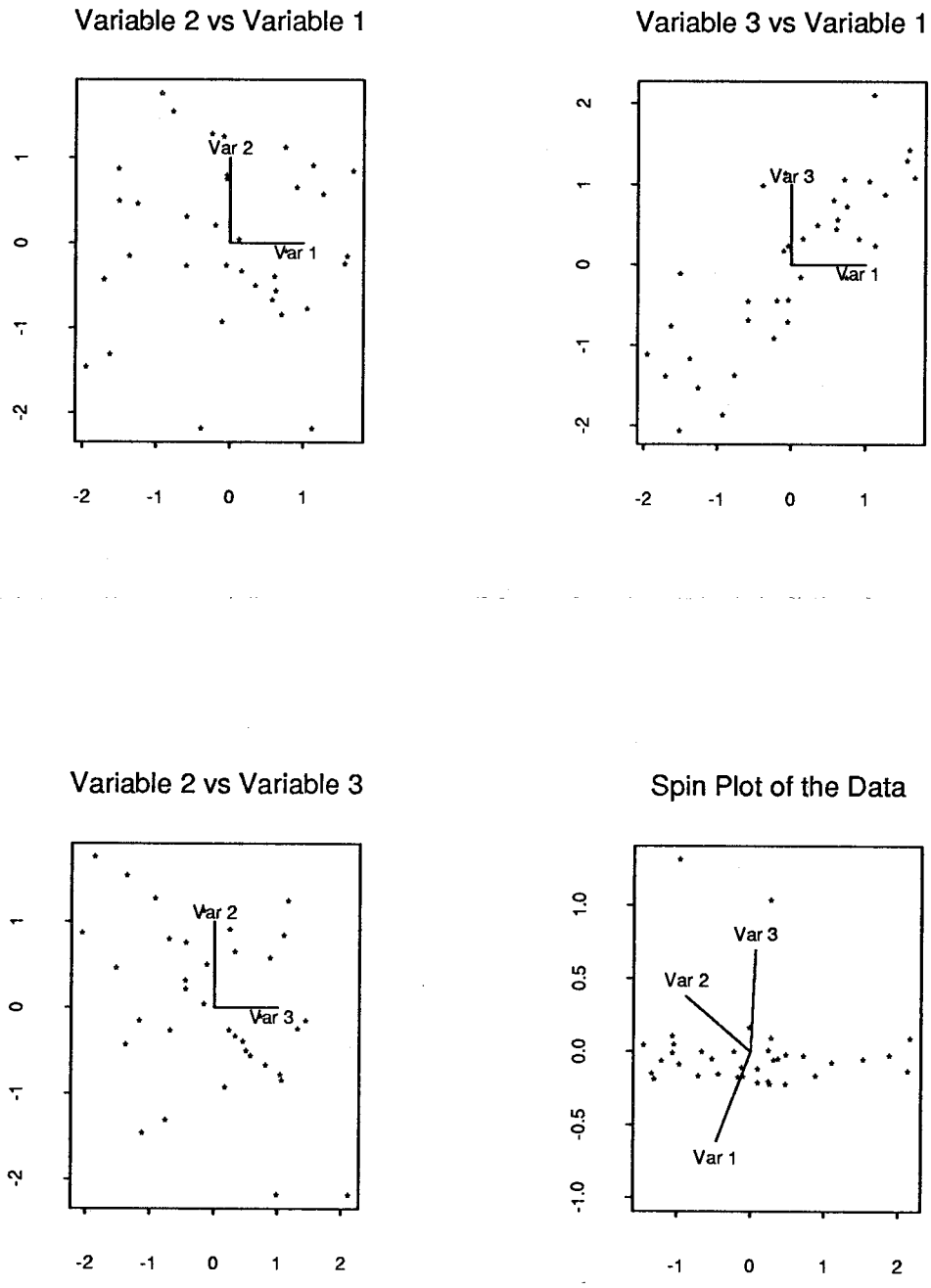
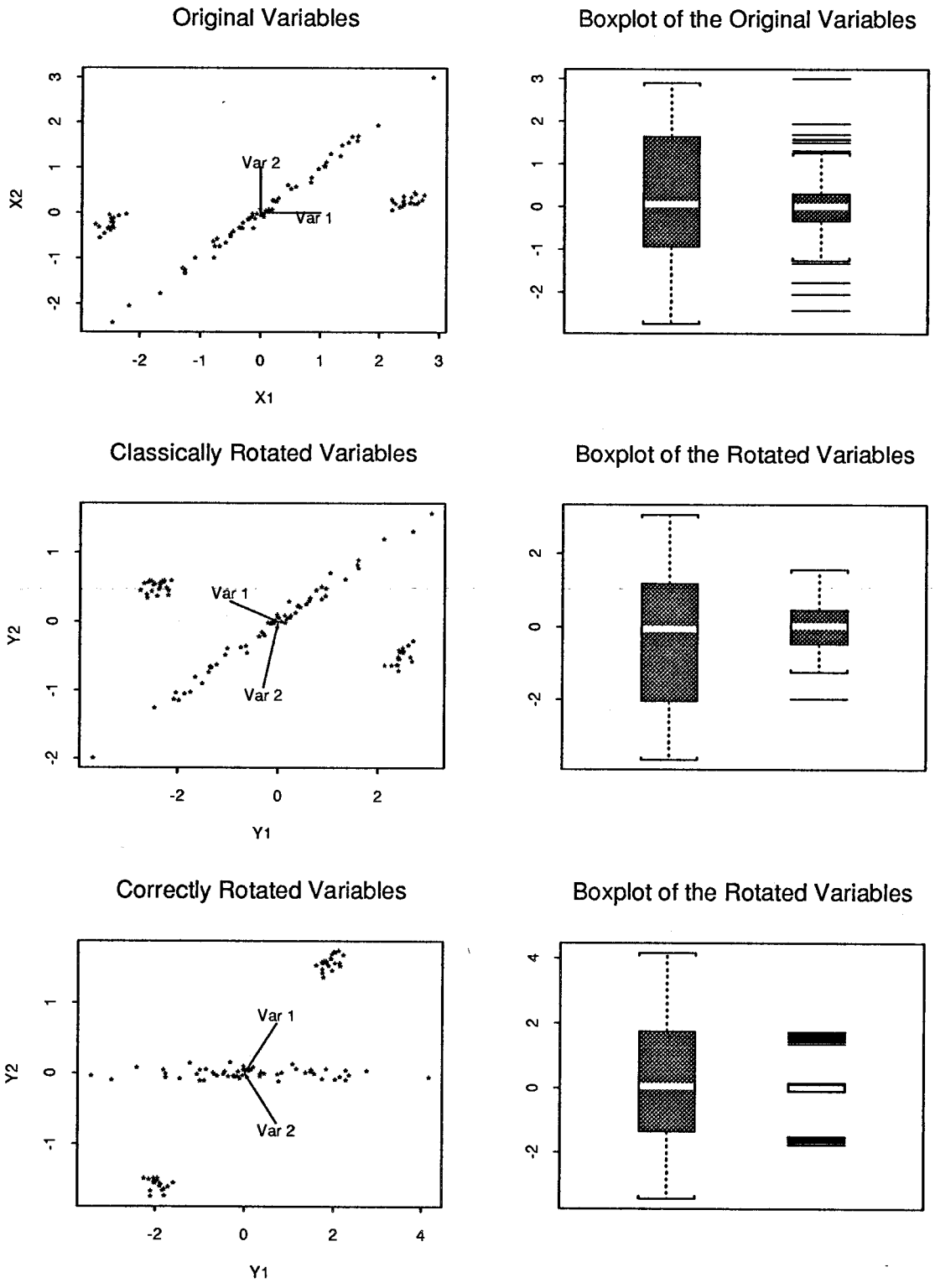


Figure 2: Distortion of the Eigenvectors in 2-dimensions



2 Concepts, Definitions and Notation

In this section, we will fix the notation used in the thesis and define a few basic concepts that will be used later. The notation is summarized in Table 1 while the concepts are presented in formal definitions.

2.1 Equivariance and Invariance

We begin by defining several forms of equivariance and invariance.

Definition 1 *Suppose we have* $\mathbf{x}_1, \ldots, \mathbf{x}_n$. *Let* $\mathbf{a}$ *be a vector in* $\mathcal{R}^p$, c *be a constant,* $\mathbf{\Gamma}$ *be an othogonal* $p \times p$ *matrix and* $\mathbf{A}$ *be a nonsingular* $p \times p$ *matrix. Then a location estimator* T *is said to be*

1. *location equivariant if* $T(\mathbf{x}_1 + \mathbf{a}, \ldots, \mathbf{x}_n + \mathbf{a}) = T(\mathbf{x}_1, \ldots, \mathbf{x}_n) + \mathbf{a}$;
2. *scale equivariant if* $T(c\mathbf{x}_1, \ldots, c\mathbf{x}_n) = cT(\mathbf{x}_1, \ldots, \mathbf{x}_n)$;
3. *orthogonal equivariant if* $T(\mathbf{\Gamma}\mathbf{x}_1 + \mathbf{a}, \ldots, \mathbf{\Gamma}\mathbf{x}_n + \mathbf{a}) = \mathbf{\Gamma}T(\mathbf{x}_1, \ldots, \mathbf{x}_n) + \mathbf{a}$;
4. *affine equivariant if* $T(\mathbf{A}\mathbf{x}_1 + \mathbf{a}, \ldots, \mathbf{A}\mathbf{x}_n + \mathbf{a}) = \mathbf{A}T(\mathbf{x}_1, \ldots, \mathbf{x}_n) + \mathbf{a}$;

and a scale estimator S *is said to be*

1. *location invariant if* $S(\mathbf{x}_1 + \mathbf{a}, \ldots, \mathbf{x}_n + \mathbf{a}) = S(\mathbf{x}_1, \ldots, \mathbf{x}_n)$;
2. *scale equivariant if* $S(c\mathbf{x}_1, \ldots, c\mathbf{x}_n) = |c|S(\mathbf{x}_1, \ldots, \mathbf{x}_n)$;
3. *orthogonal equivariant if* $S(\mathbf{\Gamma}\mathbf{x}_1 + \mathbf{a}, \ldots, \mathbf{\Gamma}\mathbf{x}_n + \mathbf{a}) = \mathbf{\Gamma}S(\mathbf{x}_1, \ldots, \mathbf{x}_n)\mathbf{\Gamma}'$;
4. *affine equivariant if* $S(\mathbf{A}\mathbf{x}_1 + \mathbf{a}, \ldots, \mathbf{A}\mathbf{x}_n + \mathbf{a}) = \mathbf{A}S(\mathbf{x}_1, \ldots, \mathbf{x}_n)\mathbf{A}'$.

We want an estimator to possess these properties because the parameters that they estimate also possess them. However, as we will see later, sometimes it may be necessary to relax the affine equivariance condition in order to make the estimator computable.

Table 1: Notation

Symbol	Meaning
$\mathcal{R}^p$	p-dimensional Euclidean space.
$\mathbf{V}$ or $\mathbf{\Sigma}$	matrices or vector valued random variables.
$\mathbf{x}$ or $\vec{\mu}$	p-dimensional vectors.
$\mathbf{x}'$ or $\mathbf{\Sigma}'$	the transpose of a vector ot matrix.
$\mathbf{V}^{1/2}$	a square root matrix such that $\mathbf{V}^{1/2}\mathbf{V}^{1/2} = \mathbf{V}$
$\lvert\mathbf{V}\rvert$	the deteminant of the matrix $\mathbf{V}$.
$\lfloor y \rfloor$	the greatest integer less that or equal to y.
Φ	the standard normal distribution function (univariate).
ϕ	the standard normal density function (univariate).
χ^2_p	a chi-square distribution with p degrees of freedom.
$\beta_{n,m}$	a beta distribution with parameters n and m.
F_α	the upper α quantile from a specified distribution.
$1_A(x)$	an indicator function of the set A.
$\overset{\text{iid}}{\sim}$	independent and identically distributed from
$\text{med}(\mathbf{x}_i)$	median of $\mathbf{x}_1, \ldots \mathbf{x}_n$
$\text{mad}(\mathbf{x}_i)$	median absolute deviation from the median of $\mathbf{x}_i$
MVE	minimum volume ellipsoid
MCD	minimum covariance determinant

2.2 Concepts of Robustness

Next, we turn our attention to robust statistics. We begin by defining what we mean by a contaminated sample. A contaminated sample is drawn from a mixture distribution $F_{\epsilon,H} = (1-\epsilon)F+\epsilon H$ where F is a p-dimensional distribution with mean $\vec{\mu}$ and covariance $\mathbf{\Sigma}$, H is any other p-dimensional distribution and $\epsilon < 1/2$ is a positive constant.

Definition 2 *A sample* $\mathbf{x}_1, \ldots, \mathbf{x}_n$ *is said to be contaminated if* $\mathbf{x}_i$, $i = 1, \ldots, n$, $\overset{\text{iid}}{\sim} F_{\epsilon,H}$ *and both* H *and* ϵ *are unknown.*

If H has moments to the second order or $\epsilon = 0$ then $F_{\epsilon,H}$ has moments to the second order. If the latter is true then $F_{\epsilon,H} = F$ and the moments are $\vec{\mu}$ and $\mathbf{\Sigma}$, otherwise if H has moments then the moments of $F_{\epsilon,H}$ are given in Fact 1.

Fact 1 *Suppose* F *is a* p*-dimensional distribution with mean* $\vec{\mu}$ *and covariance* $\mathbf{\Sigma}$ *and* H *is any other* p*-dimensional distribution with mean* $\vec{\theta}$ *and covariance* $\mathbf{\Gamma}$, $\epsilon \in [0, 1/2)$ *is a constant and* $\mathbf{X}$ *is distributed as* $F_\epsilon = (1-\epsilon)F + \epsilon H$, *then*

$(1) E\mathbf{X} = (1-\epsilon)\vec{\mu} + \epsilon\vec{\theta}$

$(2) Var(\mathbf{X}) = (1-\epsilon)\mathbf{\Sigma} + \epsilon\mathbf{\Gamma} + \epsilon(1-\epsilon)(\vec{\mu}-\vec{\theta})(\vec{\mu}-\vec{\theta})'.$

Proof

$(1)\ \int_{-\infty}^{\infty} \mathbf{x} dF_\epsilon(\mathbf{x}) = \int_{-\infty}^{\infty} \mathbf{x} d[(1-\epsilon)F(\mathbf{x}) + \epsilon H(\mathbf{x})]$

$= (1-\epsilon)\int_{-\infty}^{\infty} \mathbf{x} dF(\mathbf{x}) + \epsilon \int_{-\infty}^{\infty} \mathbf{x} dH(\mathbf{x}) = (1-\epsilon)\vec{\mu} + \epsilon\vec{\theta}$

$(2)\ \int_{-\infty}^{\infty} \mathbf{x}\mathbf{x}' dF_\epsilon - E\mathbf{x}E\mathbf{x}' = \int_{-\infty}^{\infty} \mathbf{x}\mathbf{x}' d[(1-\epsilon)F(\mathbf{x}) + \epsilon H(\mathbf{x})] - ((1-\epsilon)\vec{\mu} + \epsilon\vec{\theta})((1-\epsilon)\vec{\mu} + \epsilon\vec{\theta})'$

$= (1-\epsilon)\int_{-\infty}^{\infty} \mathbf{x}\mathbf{x}' dF(\mathbf{x}) + \epsilon \int_{-\infty}^{\infty} \mathbf{x}\mathbf{x}' dH(\mathbf{x}) - (1-\epsilon)^2\vec{\mu}\vec{\mu}' - \epsilon^2\vec{\theta}\vec{\theta}' - 2\epsilon(1-\epsilon)\vec{\mu}\vec{\theta})'$

$= (1-\epsilon)(\int_{-\infty}^{\infty} \mathbf{x}\mathbf{x}' dF(\mathbf{x}) - \vec{\mu}\vec{\mu}') + \epsilon(\int_{-\infty}^{\infty} \mathbf{x}\mathbf{x}' dH(\mathbf{x}) - \vec{\theta}\vec{\theta}') + (1-\epsilon)\vec{\mu}\vec{\mu}' + \epsilon\vec{\theta}\vec{\theta}'$

$$-(1-2\epsilon+\epsilon^2)\vec{\mu}\vec{\mu}' - \epsilon^2\vec{\theta}\vec{\theta}' - 2\epsilon(1-\epsilon)\vec{\mu}\vec{\theta}'$$

$= (1-\epsilon)\mathbf{\Sigma} + \epsilon\mathbf{\Gamma} + \epsilon(1-\epsilon)(\vec{\mu}\vec{\mu}' + \vec{\theta}\vec{\theta}' - 2\vec{\mu}\vec{\theta}')$

$= (1-\epsilon)\mathbf{\Sigma} + \epsilon\mathbf{\Gamma} + \epsilon(1-\epsilon)(\vec{\mu}-\vec{\theta})(\vec{\mu}-\vec{\theta})'$

If H does not have moments and $\epsilon > 0$ then $F_{\epsilon,H}$ will not have moments either. This implies that for any level of contamination $\epsilon > 0$, the moments of $F_{\epsilon,H}$ can be changed to any value, even infinite or undefined, by choosing an appropriate contaminating distribution H.

Using F_ϵ, we can define several measures of robustness for a given estimator T.

Definition 3 *The maximum bias of T for a given level of contamination ϵ is defined by*

$$B(\epsilon; T) = \sup_{F_{\epsilon,H}} ||T(F_{\epsilon,H}) - T(F)||$$

Plotting $B(\epsilon; T, F_\epsilon)$ verses ϵ produces the *maximum bias curve* [16]. This curve carries a lot of information about the robust properties of the estimator. Two important robustness quantities that can be derived from $B(\epsilon, T)$ are the breakdown point and the gross error sensitivity of an estimator.

Definition 4 *The breakdown point of T is defined by*

$$\epsilon^*(T) = \inf \{\epsilon : B(\epsilon, T) = \infty\}$$

In other words, the breakdown point is the smallest level of contamination that can cause $T(F_{\epsilon,H})$ to be arbitrarily far from $T(F)$.

Definition 5 *The gross error sensitivity is defined by*

$$GES = \left. \frac{dB(\epsilon; T)}{d\epsilon} \right|_{\epsilon=0}$$

It measures the maximum effect that an infinitesimal amount of contamination can have on T. It can also be used as a linear approximation to maximum bias when ϵ is near zero.

To measure the sensitivity of T to an infinitesimal amount of contamination from an arbitrary distribution H, we use the influence function.

Definition 6 *The influence function of T for a contaminating distribution H is given by*

$$IF(H;T,F) = \lim_{\epsilon \to 0} \frac{T(F_{\epsilon,H}) - T(F)}{\epsilon}$$

for distributions H for which the limit exists. It can be seen that

$$\sup_H ||IF(H;T,F)|| = GES.$$

A larger value of $IF(H;T,F)$ means that the estimator T is more greatly influenced by contamination with H.

3 Robust Estimation of Multivariate Location and Scatter

To estimate the location and scatter of a multivariate distribution F, one uses the sample mean vector $\bar{\mathbf{x}}$ and sample covariance matrix $\mathbf{S}$ defined as

$$\bar{\mathbf{x}} = \sum_{i=1}^{n} \mathbf{x}_i / n \tag{1}$$

$$\mathbf{S} = \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' / (n-1). \tag{2}$$

Unfortunately, if the sample contains some contaminating elements, then $\bar{\mathbf{x}}$ and $\mathbf{S}$ can perform very poorly. For example consider the contaminated distribution $F_{\epsilon,H}$ as defined in Section 2. For this case $E(\mathbf{X}) = (1-\epsilon)\vec{\mu} + \epsilon\vec{\theta}$ and $\mathrm{Var}(\mathbf{X}) = (1-\epsilon)\mathbf{\Sigma} + \epsilon\mathbf{\Gamma} + \epsilon(1-\epsilon)(\vec{\mu} - \vec{\theta})(\vec{\mu} - \vec{\theta})'$. So, if one wishes to estimate the location and scatter of F and one observes $\mathbf{x}_1, \ldots, \mathbf{x}_n \overset{\text{iid}}{\sim} F_\epsilon$, then $\bar{\mathbf{x}}$ and $\mathbf{S}$ can be extremely inaccurate. Therefore, these estimators are unreliable and a different method is needed to ensure accurate estimates.

Suppose that k of the n points in the sample come from the distribution H and the exact k points are known. Then one simply removes these points and estimates the parameters of F directly, using $\bar{\mathbf{x}}$ and $\mathbf{S}$ on the remaining points. However, the underlying distribution of each point as well as the exact number of contaminants is usually unknown and therefore diagnostic tools must be used to estimate these values. See Section 4 for details on how this is done in a multivariate context.

A second method of estimating the location and scatter of F is to use robust techniques. The main advantage of this method over the previous is that only an upper bound on k need be known for the techniques to work. The disadvantage is that some of the desirable properties of the estimate have to be sacrificed in order that others may be maintained. In the univariate context the trade-off occurs between robustness and

efficiency. In the multivariate situation, one must also consider computational efficiency and the equivariance of the estimator. If a desirable property must be sacrificed, which one should it be? There is no clear solution here and different people will argue for different estimators.

3.1 M-Estimators

The first robust estimators were called M-estimates because they were based on a generalized maximum likelihood score function. The univariate M-estimate of location was proposed in 1964 by Peter Huber. M-estimates were later generalized to estimate scale and location and scale simultaneously. To define the univariate M-estimate, one begins with maximum likelihood estimation. Suppose we have $x_1, \ldots, x_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$, then the maximum likelihood estimate of location when the scale is known is the root t of the equation

$$\frac{1}{n}\sum_{i=1}^{n} \psi_{LS}\left(\frac{x_i - t}{\sigma}\right) = 0,$$

where $\psi_{LS}(x) = x$, the maximum likelihood estimate of scale when the location is known is the root v to the equation

$$\frac{1}{n}\sum_{i=1}^{n} \rho_{LS}\left(\frac{x_i - \mu}{v}\right) = b,$$

where $\rho_{LS}(x) = x^2$ and $b = 1$, and the maximum likelihood estimate of location and scale is the vector (t, v) that satisfies

$$\frac{1}{n}\sum_{i=1}^{n} \psi_{LS}\left(\frac{x_i - t}{v}\right) = 0 \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} \rho_{LS}\left(\frac{x_i - t}{v}\right) = b \tag{3}$$

simultaneously. M-estimates are a generalization of the maximum likelihood estimates to a class of functions $\psi(x)$ and $\rho(x)$ which satisfy the following properties:

1. $\psi(x)$ is odd with at most a finite number of discontinuities;

2. $\rho(x)$ is even, differentiable almost everywhere and $\rho(0) = 0$;

3. $\rho(x)$ is strictly increasing on $[0, c)$ and constant on $[c, \infty)$;

4. $b = E\left[\rho\left(\frac{x-\mu}{\sigma}\right)\right] > 0$.

M-estimates of location are unbiased at the uncontaminated model and have a breakdown point of 1/2 if $\sup|\psi(x)| < \infty$, zero otherwise. They can be very efficient but this usually carries a heavy price in terms of the maximum bias curve. M-estimates of location are asymptotically normally distributed with a $\sqrt{n}$ rate of convergence. The M-estimate of scale has a breakdown point of b/M against outliers and $1 - b/M$ against inliers where $M = \sup \rho(x)$. They also converge to a normal distribution at a $\sqrt{n}$ rate and are consistent at the uncontaminated model, but they cannot obtain good efficiency while still maintaining a high breakdown point against outliers. M-estimates of location and scale will have a positive breakdown point against outliers only if both $\psi(x)$ and $\rho(x)$ are bounded.

As seen above $\bar{x}$ and s^2 are special cases of M-estimators. They are efficient but have a breakdown point of zero. The estimators $t = \text{med}(x_i)$ and $v = \text{mad}(x_i) = \text{med}|x_i - t|/\Phi^{-1}(3/4)$ are the other extreme for M-estimators. These are obtained by $\psi_{med}(x) = \text{sgn}(x)$, $\psi_{med}(0) = 0$ and

$$\rho_{mad}(x) = \begin{cases} 0 & \Phi^{-1}(1/4) < x < \Phi^{-1}(3/4), \\ 1/2 & x = \Phi_{1/4} \text{ or } x = \Phi_{3/4}, \\ 1 & \text{otherwise.} \end{cases}$$

One can easily verify that $b = \int \rho_{mad}(x)\phi(x)dx = 1/2$. The median and the mad are highly robust in terms of the maximum bias curve and have a breakdown point of 1/2 but are very inefficient at the normal model. To achieve an M-estimator that is robust and whose location estimator is relatively efficient, a compromise is needed between the

two extremes. This is done by defining ψ and ρ so that they behave like ψ_{LS} and ρ_{LS} for small values of x but become constant like ψ_{med} and ρ_{mad} when x is large. Huber [9] combined the extremes to obtain the following class of functions

$$\psi_c^H(x) = \begin{cases} x & \text{if } |x| \le c, \\ c \operatorname{sgn}(x) & \text{otherwise,} \end{cases} \quad \text{and} \quad \rho_k^H(x) = \begin{cases} x^2 & \text{if } |x| \le k, \\ k^2 & \text{otherwise,} \end{cases} \tag{4}$$

where $0 < c, k < \infty$. The balance between efficiency and robustness is controlled by the constants c and k. Note that $(t, v) \to (\bar{x}, s^2)$ as $c, k \to \infty$ and $t \to \text{med}(x)$ as $c \to 0$.

Some people prefer to have $\psi = \rho'$ because this holds for the maximum likelihood estimate. A set of functions that satisfy this criterion are Tukey's biweight functions defined by

$$\psi_c^T(x) = \begin{cases} x(1 - 2(x/c)^2 + (x/c)^4) & \text{if } |x| \le c, \\ 0 & \text{otherwise.} \end{cases} \quad \text{and}$$

$$\rho_k^T(x) = \begin{cases} x^2(3 - 3(x/k)^2 + (x/k)^4) & \text{if } |x| \le k, \\ k^2 & \text{otherwise.} \end{cases}$$

Like Huber's function, Tukey's function has the maximum likelihood estimator as a special case, but for any $c < \infty$, the function ψ_c^T redescends to zero.

M-Estimates can be generalized to higher dimensions and, in 1976, Maronna [15] did just that. M-estimators of multivariate location and scatter are defined in a similar fashion as in the univariate context. They are the solution $(\hat{\mathbf{t}}, \hat{\mathbf{V}})$ to the following system of equations

$$u_1\left(\left\{(\mathbf{x}_i - \mathbf{t})'\mathbf{V}^{-1}(\mathbf{x}_i - \mathbf{t})\right\}^{1/2}\right)(\mathbf{x}_i - \mathbf{t}) = \mathbf{0}, \tag{5}$$

$$\sum_{i=1}^{n} u_2\left((\mathbf{x}_i - \mathbf{t})'\mathbf{V}^{-1}(\mathbf{x}_i - \mathbf{t})\right)(\mathbf{x}_i - \mathbf{t})(\mathbf{x}_i - \mathbf{t})' - \mathbf{V} = \mathbf{0}, \tag{6}$$

where u_1 and u_2 are functions defined for $s \ge 0$ satisfying, for $\mathbf{V} = \mathbf{SS}'$,

$$Eu_1(|\mathbf{S}^{-1}(\mathbf{x}_i - \mathbf{t})|)\mathbf{S}^{-1}(\mathbf{x}_i - \mathbf{t}) = \mathbf{0},$$

$$Eu_2(|\mathbf{S}^{-1}(\mathbf{x}_i - \mathbf{t})|^2)[\mathbf{S}^{-1}(\mathbf{x}_i - \mathbf{t})][\mathbf{S}^{-1}(\mathbf{x}_i - \mathbf{t})]' = \mathbf{I}_p.$$

In addition,

1. $u_1(s)$ and $u_2(s)$ are nonnegative, nonincreasing, and continuous for $s \geq 0$;
2. $\psi_i(s) = su_i(s)$ is bounded for $i = 1, 2$;
3. $\psi_2(s)$ is strictly increasing in the interval where $\psi_2(s) < K = \sup_{s \geq 0} \psi_2(s)$ and constant in the regions where $\psi_2(s) = K$;
4. $\exists\ s_0$ such that $\psi_2(s_0^2) > p$, and $u_1(s) > 0$ for $s \leq s_0$;
5. $\exists\ a > 0$ such that for every hyperplane H, $P_F(H) \leq 1 - p/K - a$.

Maronna shows, under certain conditions, that the estimates are unique, consistent, asymptotically normal, affine equivariant and relatively easy to compute. However, the breakdown point is less than $(p+1)^{-1}$. Therefore these estimates do not perform well in higher dimensions if the level of contamination is large.

3.2 S-Estimators

S-estimates were originally developed, in the regression context, by Rousseeuw and Yohai in 1984 [23]. To describe these estimators, it is best to consider the special case of estimating univariate location and scale. Suppose we have $x_1, \ldots, x_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$ then the S-estimates of location and scale is the vector (t, v) such that v is minimized subject to the constraint,

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{x_i - t}{v}\right) = E\left[\rho\left(\frac{x - \mu}{\sigma}\right)\right], \tag{7}$$

where $\rho(x)$ and b are defined as in the case of M-estimates of scale. In fact, for a fixed location, an S-estimate is exactly an M-estimate of scale and, like the M-estimate, has the maximum likelihood estimate as a special case by defining $\rho_{LS}(x) = x^2$. The

other extreme is obtained when $\rho_{mve} = 1 - 1_{[\Phi.25,\Phi.75]}(x)$ and $b = 1 - (\lfloor n/2 \rfloor + 1)/n$ are used. If the location is fixed to be the median, then the corresponding M-estimate of scale will be the mad. However, as an S-estimate, ρ_{mve} defines the SHORTH which is the most robust univariate estimator of location and scale with respect to minimizing the maximum bias among all M-estimators of scale [17]. The distribution of the scale estimate converges weakly to a normal distribution at a rate of $\sqrt{n}$ but the distribution of the location estimate converges at a rate of $\sqrt[3]{n}$ to a non-normal distribution. Both estimates are extremely inefficient at the normal model.

The properties of S-estimates are similar to the properties of M-estimates of scale. They are consistent at the uncontaminated model, possess a breakdown point of b/M against outliers and $1 - b/M$ against inliers, where $M = \sup \rho(x)$ and they cannot achieve a high breakdown point and good efficiency at the same time. The other main problem with S-estimates is their asymptotic distribution. If ρ is not smooth enough, the estimates can converge at a slower rate to a non-normal distribution as demonstrated by the SHORTH. An asymptotically normal estimate is preferable but only if the breakdown point of ρ_{mve} can be maintained. The boundedness of ρ is necessary and sufficient to maintain a positive breakdown point and $\rho(x)$ being twice continuously differentiable is a sufficient condition for asymptotic normality at a $\sqrt{n}$ rate. Tukey's biweight function,

$$\rho_c(x) = \begin{cases} (x/k)^2(3 - 3(x/k)^2 + (x/k)^4) & \text{if } |x| \leq k, \\ 1 & \text{otherwise,} \end{cases} \tag{8}$$

possesses all these properties.

Davies [5] generalized S-estimates to the multivariate context in 1987. They are defined as the vector $\hat{\mathbf{t}}$ and the positive definite matrix $\hat{\mathbf{V}}$ that minimize $|\mathbf{V}|$ subject to

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left[\left\{(\mathbf{x}_i - \mathbf{t})'\mathbf{V}^{-1}(\mathbf{x}_i - \mathbf{t})\right\}^{1/2}\right] = b, \tag{9}$$

where $\rho : [0, \infty) \rightarrow [0, \infty)$ is strictly increasing on $[0, c)$ and constant on $[c, \infty)$ and $b = E\rho\left[\left\{(\mathbf{x} - \vec{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x} - \vec{\mu})\right\}^{1/2}\right]$ [12].

Multivariate S-estimates are affine equivariant, consistent at the uncontaminated model, possess a breakdown point of b/M against outliers and $1 - b/M$ against inliers, where $M = \sup \rho(x)$, but, like their univariate counterparts, they cannot achieve a high breakdown point and good efficiency at the same time and they onverge at a slower rate to a non-normal distribution if ρ is not smooth enough. Again, one seeks a ρ function which is bounded above as well as twice continuously differentiable. Tukey's biweight function, defined in (8), does the job. Unfortunately, the distribution of $y = (\mathbf{x} - \vec{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x} - \vec{\mu})$ depends on the dimension of $\mathbf{x}$ thus causing the tuning constant c to change as well. This can cause problems if the distribution of y is difficult to compute. Fortunately, the normal case is one in which the calculation can be done because $y \sim \chi^2_p$ when $\mathbf{x} \sim \mathrm{N}_p(\vec{\mu}, \mathbf{\Sigma})$. Substituting y into ρ_c and letting $k = c^2$, (8) becomes

$$\rho_k(y) = \begin{cases} (y/k)(3 - 3y/k + (y/k)^2) & \text{if } y \leq k \\ 1 & \text{otherwise} \end{cases} . \tag{10}$$

The expected value for a given k is

$$E\left(\rho_k(y)\right) = \int_0^k \left(\frac{3z}{k} - \frac{3z^2}{k^2} + \frac{z^3}{k^3}\right) d\chi^2_p + \int_k^\infty d\chi^2_p$$

Setting $E\left(\rho_k(y)\right) = b$ and solving for k, we get k such that

$$1 + \int_0^k \left(\frac{3z}{k} - \frac{3z^2}{k^2} + \frac{z^3}{k^3} - 1\right) d\chi^2_p = b. \tag{11}$$

Table 2 contains the tuning constant $k = c^2$ for certain values of b and p.

In 1989, Lopuhaä [12] showed that an S-estimate can also be defined as the solution to the equations

$$u\left[\left\{(\mathbf{x}_i - \mathbf{t})'\mathbf{V}^{-1}(\mathbf{x}_i - \mathbf{t})\right\}^{1/2}\right](\mathbf{x}_i - \mathbf{t}) = \mathbf{0}, \tag{12}$$

Table 2: Constant k for Tukey's biweight function

	dimension p									
b	1	2	3	4	5	6	7	8	9	10
.10	26.86	55.86	84.8	113.8	142.8	171.8	200.8	229.7	258.7	287.7
.25	8.63	19.62	30.6	41.5	52.5	63.4	74.3	85.3	96.2	107.2
.50	2.40	7.08	11.9	16.8	21.6	26.5	31.4	36.2	41.1	46.9

$$\sum_{i=1}^{n} pu\left[\left\{(\mathbf{x}_i-\mathbf{t})'\mathbf{V}^{-1}(\mathbf{x}_i-\mathbf{t})\right\}^{1/2}\right](\mathbf{x}_i-\mathbf{t})(\mathbf{x}_i-\mathbf{t})'-v\left[\left\{(\mathbf{x}_i-\mathbf{t})'\mathbf{V}^{-1}(\mathbf{x}_i-\mathbf{t})\right\}^{1/2}\right]\mathbf{V}=\mathbf{0}, \tag{13}$$

where $u(x)=\psi(x)/x$, $v(x)=x\psi(x)-\rho(x)+b$ and $\psi(x)$ is the derivative of $\rho(x)$. These equations are very similar to (5) and (6), which means that there is a solution which has a breakdown point that is at most $1/(p+1)$. However, there is at least one other solution which has a breakdown point of b/M, where $M=\sup\rho$. The global minimum of (9) is one of the high breakdown solutions. Therefore, one should minimize (9) rather than solve (12) and (13) because the former guarantees the solution to be one of high breakdown if the global minimum is reached.

Unfortunately, S-estimates are computationally expensive because they involve the minimization of an implicitly defined, non-convex function. If the function is convex then the breakdown point of the S-estimate is zero. The possibility of multiple minima means that the minimum obtained need not be the best (although it will have a high breakdown point). A good starting point can increase the probability of reaching the global minimum but this would require a robust estimate of location and scatter which is exactly what we seek in the first place.

In 1983, Rousseeuw proposed the first S-estimate of multivariate location and scatter

with a positive breakdown point called the minimum volume ellipsoid estimator or MVE. The MVE is the multivariate version of the SHORTH and is based on an ellipsoid of minimal volume which covers at least $\lfloor n/2 \rfloor + 1$ points of the data. The centre of the ellipsoid is the location estimate, while the rescaled ellipsoid is the corresponding covariance estimate. The MVE is affine equivariant with a breakdown point of $(\lfloor n/2 \rfloor - p + 1)/n$. Unfortunately, the MVE is not computable for any reasonable sample size because of the combinatorial explosion. For example, a sample size of 50 in 2 dimensions has over 10^{14} ways to choose 27 distinct point from the sample. The MVE is computed by finding the one that yields the ellipsoid of minimal volume. A second way to compute the MVE is to minimize $\rho_{mve}(z) = 1 - 1_{[0,\chi^2_{p,.5}]}(z)$, where $z = (\mathbf{x} - \vec{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x} - \vec{\mu})$ and $\chi^2_{p,\alpha} = \{x : \chi^2_p(x) = 1 - \alpha\}$. Asymptotically $b = 1/2$, however for finite samples $b = 1 - \lfloor (n + p + 1)/2 \rfloor / n$ to guarantee a solution. To find a local minimum, let alone the global minimum, is very difficult because the function to be minimized is discontinuous and nonconvex. However, the MVE can be approximated through the following resampling scheme:

1. Select a subsample that contains exactly $p + 1$ distinct points indexed by $J = \{i_1, \ldots, i_{p+1}\}$;

2. Compute

$$\bar{\mathbf{x}}_J = \frac{1}{p+1} \sum_{i \in J} \mathbf{x}_i \quad \text{and} \quad \mathbf{C}_J = \frac{1}{p} \sum_{i \in J} (\mathbf{x}_i - \bar{\mathbf{x}}_J)'(\mathbf{x}_i - \bar{\mathbf{x}}_J);$$

3. Set $\mathbf{V}_J = m_J \mathbf{C}_J$ where $m_J = \text{med}(\mathbf{x}_i - \bar{\mathbf{x}}_J)\mathbf{C}_J^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}_J)'$ for $i = 1, \ldots, n$;

4. Compute $|\mathbf{V}_J| = m_J^p |\mathbf{C}_J|$.

The resampling is repeated for many J after which the one with the smallest determinant $|\mathbf{V}_J|$ is retained; call this subsample I. The final estimates are $\bar{\mathbf{x}}_I$ and $k\mathbf{V}_I$, where $k = \chi^2_{p,.5}$ makes the scatter estimate consistent at the multivariate normal model.

The total number of subsamples drawn is m, which is selected to ensure a high probability that at least one subsample will not contain any contaminants. Since the subsamples are drawn independently, the probability that there is at least one good subsample in m is given by $1-(1-\beta)^m$, where

$$\beta = \frac{((1-\epsilon)n)!(n-p-1)!}{n!((1-\epsilon)n-p-1)!}$$

is the probability of drawing a good sample from the data and ϵn is the number of contaminants in the data. Observe that $\beta = P(X=0)$, where X follows a hypergeometric distribution with parameters $n, (1-\epsilon)n, p+1$. It is easily shown that $\beta \to (1-\epsilon)^{p+1}$ as $n \to \infty$. Furthermore, the breakdown point for the MVE implies that $\epsilon < 1/2$, otherwise the estimate will break. Therefore, the probability of getting a good sample when the level of contamination is unknown is approximately equal to $1-(1-1/2^{p+1})^m$. From setting this probability to 99% and solving for m, one obtains a lower bound on the number of subsamples needed. To verify that this is, in fact a lower bound, one must consider the exact breakdown point, $\epsilon = 1-(\lfloor n/2 \rfloor - p+1)/n$ and β to calculate the true value of m, say m_0, then show that $m_0 \leq m$ for all $n < \infty$ and $\lim_{n\to\infty} m_0 = m$. Table 3 contains m_0 for several sample sizes and dimensions as well as the asymptotic approximation or lower bound.

Like the SHORTH, the MVE does not behave well asymptotically. Its distribution converges weakly to a non-normal distribution at the slow rate of $\sqrt[3]{n}$. To improve this rate, Rousseeuw also proposed the minimum covariance determinant estimator or MCD. The calculation of the MCD is identical to the MVE except that the final estimator is the mean vector and rescaled covariance matrix of the $\lfloor n/2 \rfloor + 1$ points inside the MVE. The MCD has all the robust properties of the MVE and it converges to a normal distribution at the rate of $\sqrt{n}$.

The resampling scheme used to calculate the MVE can also be used to calculate

Table 3: # subsamples such that P(# good samples ≥ 1) = 99%

sample	dimension									
size	1	2	3	4	5	6	7	8	9	10
25	13	21	41	54	100	112	203	192	341	270
100	15	32	60	119	211	414	702	1376	2236	4377
1000	16	35	71	143	283	566	1116	2232	4373	8731
10000	16	35	72	145	292	585	1171	2343	4679	9356
100000	17	35	72	146	293	587	1177	2355	4710	9422
∞	17	35	72	146	293	588	1176	2356	4714	9430

other S-estimates. To see this, one simply writes the MVE in the form of an S-estimate, using the function ρ_{mve} as defined in Section 3.2. Furthermore, by letting $\hat{\mathbf{V}} = \hat{s}^2\hat{\mathbf{C}}$ such that $|\hat{\mathbf{C}}| = 1$, the MVE becomes the vector $\hat{\mathbf{t}}$ and the positive definite symmetric matrix $\hat{s}^2\hat{\mathbf{C}}$ that minimize s subject to

$$\frac{1}{n}\sum_{i=1}^{n}\rho_{mve}\left(\frac{\sqrt{(\mathbf{x}_i-\mathbf{t})'\mathbf{C}^{-1}(\mathbf{x}_i-\mathbf{t})}}{s}\right) = 1 - \lfloor(n+p+1)/2\rfloor/n \qquad (14)$$

In this form, the resampling approximation to the MVE becomes the mean and rescaled covariance of the subsample that minimizes s. Furthermore, since any S-estimate can be written in this form by replacing ρ_{mve} by the appropriate function ρ and $1-\lfloor(n+p+1)/2\rfloor/n$ by the appropriate value b, the resampling scheme can be used to approximate any S-estimate. Computationally, this is not very attractive because a non-linear function must be solved m times in order to compute the estimate. But this can be avoided by modifying the algorithm in a way similar to that proposed by Yohai and Zamar [29]

in the case of regression estimates. Consider

$$h(s|\mathbf{t},\mathbf{C}) = \frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{\sqrt{(\mathbf{x}_i-\mathbf{t})'\mathbf{C}^{-1}(\mathbf{x}_i-\mathbf{t})}}{s}\right) \tag{15}$$

It is easy to verify that $h(s|\mathbf{t},\mathbf{C})$ is a decreasing function. Therefore, if $h(s_0|\mathbf{t},\mathbf{C}) > b$ then s_0 will have to be increased. This implies that s need not be calculated for each subsample because if our current minimum value for s is s_0 and $h(s_0|\mathbf{t},\mathbf{C}) > b$ then s_0 will not be updated because the new value of s will be greater than s_0. Using this fact, the following modification is proposed for the resampling algorithm.

1. Calculate an initial value s_0, $\hat{\mathbf{t}}$ and $\hat{\mathbf{V}}$.

2. For J in $1,\ldots,m$,

   (a) select subsample J containing exactly $p+1$ distinct points,

   (b) compute

   $$\mathbf{t}_J = \frac{1}{p+1}\sum_{i\in J}\mathbf{x}_i \quad \text{and} \quad \mathbf{C}_J = \frac{1}{p}\sum_{i\in J}(\mathbf{x}_i-\bar{\mathbf{t}}_J)'(\mathbf{x}_i-\bar{\mathbf{t}}_J),$$

   (c) rescale $\mathbf{C}_J$ so that $|\mathbf{C}_J| = 1$,

   (d) if $h(s_0|\mathbf{t}_i,\mathbf{C}_i) < b$ then:

      i. $s_0 = \{s : h(s|\mathbf{t}_J,\mathbf{C}_J) = b\}$;

      ii. $\hat{\mathbf{t}} = \mathbf{t}_J$;

      iii. $\hat{\mathbf{V}} = s_0^2\mathbf{C}_J$;

3. return $\hat{\mathbf{t}}$ and $\hat{\mathbf{V}}$.

With this modification, the expected number of times that s_0 needs to be updated is of the order $\log(m)$. To show this, let

$$Y_j = \begin{cases} 1 & \text{if } h(s_0|\mathbf{t}_j,\mathbf{C}_j) < b \\ 0 & \text{otherwise} \end{cases}.$$

Observe that the number of times that s_0 needs to be updated is η, where

$$\eta = \sum_{1}^{m} Y_J.$$

Using the properties of expected values it is clear that $E(\eta) = \sum_1^m E(Y_j)$. Since each subsample is drawn independently and

$$s_0 = \min_s h(s|\mathbf{t}_i, \mathbf{C}_i) = b, \text{ for } i = 1, \ldots, j,$$

$P(Y_j = 1) \leq 1/j$ with equality as $n \to \infty$, which implies that $E(Y_j) \leq 1/j$, and, therefore, $E(\eta) \leq \sum_1^m (1/j) \approx \log(m)$.

The only possible difficulty with this algorithm is the updating of s_0. However, I claim that this will be a relatively easy task because it is a function of a single variable, the solution is bounded below by zero and above by s_0 and, if one imposes the sufficient condition for asymptotic normality, the function will be twice continuously differentiable. Therefore, a Newton-Raphson routine can be used to find the solution. Furthermore, it will converge quite rapidly because the solution is unique (see Section 3.1), and bounded. The uniqueness is guaranteed because $s_0|\mathbf{t}, \mathbf{C}$ is an M-estimate of scale.

To use this method, a starting value is needed for s_0. The usual approach is to set $s_0 = \{s : h(s|\mathbf{t}_1, \mathbf{C}_1) = b\}$. I suggest using $s_0 = \{s : h(s|\bar{\mathbf{x}}, \mathbf{S}/|\mathbf{S}|^{1/p}) = b\}$ because it forces the S-estimate to beat the classical estimate. If it does not then the location estimate becomes $\bar{\mathbf{x}}$ and the scatter becomes $s_0^2\mathbf{S}/|\mathbf{S}|^{1/p}$. Furthermore, if the probability of beating the classical estimate is β then the covariance of the location estimator becomes $(1-\beta)S_{\bar{\mathbf{x}}} + \beta S_{\mathbf{t}}$. Since $\bar{\mathbf{x}}$ is the most efficient estimator for the uncontaminated model, the efficiency of the location estimator has been improved. The estimated scatter is also slightly improved because the estimated correlation structure receives the same benefits as the location estimator. Unfortunately, the scale estimate does not because it is an M-estimate and, therefore, not very efficient unless a low breakdown point is used.

This starting point should also work if the sample is contaminated because there is a high probability that at least one of the subsamples will contain all good points. This subsample alone should cause s_0 to be updated thus changing the location, correlation and scale estimates.

3.3 τ-Estimators

A drawback of S-estimates is that they cannot achieve a high breakdown point and high efficiency at the same time. Yohai and Zamar [28] addressed this problem in 1988 with the introduction of τ-estimates. A τ-estimate is an adaptive multiple of an M-estimate of scale. In the simple case of univariate location and scale, let

$$\tau^2(t) = \frac{s^2(t)}{nb_2} \sum_{i=1}^{n} \rho_2 \left(\frac{x_i - t}{s(t)} \right) \tag{16}$$

where $s(t)$ is an M-estimate of scale defined for a given t, as the solution to

$$\frac{1}{n} \sum_{i=1}^{n} \rho_1 \left(\frac{x_i - t}{s} \right) = b_1,$$

ρ_2 defines another M-estimate which is usually different from $s(t)$ and b_2 is the normalizing constant for ρ_2. Instead of minimizing $s(t)$ over t, $\tau(t)$ is minimized. The value $\hat{t}$ that minimizes $\tau(t)$ is the location estimate and $\tau(\hat{t})$ is the scale estimate. τ-estimates of regression can be defined in an analogous way.

In 1990, Lopuhaä generalized τ-estimates to the multivariate context. He defined them to be the vector $\hat{\mathbf{t}}$ and matrix

$$\hat{\mathbf{V}} = \frac{\hat{\mathbf{C}}}{nb_2} \sum_{i=1}^{n} \rho_2 \left[\left\{ (\mathbf{x}_i - \hat{\mathbf{t}})' \hat{\mathbf{C}}^{-1} (\mathbf{x}_i - \hat{\mathbf{t}}) \right\}^{1/2} \right], \tag{17}$$

where $\hat{\mathbf{t}}$ and $\hat{\mathbf{C}}$ are the vector and positive definite symmetric matrix that minimize

$$|\mathbf{C}| \left\{ \sum_{i=1}^{n} \rho_2 \left[\left\{ (\mathbf{x}_i - \mathbf{t})' \mathbf{C}^{-1} (\mathbf{x}_i - \mathbf{t}) \right\}^{1/2} \right] \right\}^p \tag{18}$$

subject to

$$\frac{1}{n}\sum_{i=1}^{n}\rho_1\left[\left\{(\mathbf{x}_i-\mathbf{t})'\mathbf{C}^{-1}(\mathbf{x}_i-\mathbf{t})\right\}^{1/2}\right]=b_1. \tag{19}$$

τ-estimates are a generalization of S-estimates. Indeed, if $\rho_1=\rho_2$ and $b_1=b_2$ then (17) reduces to a multivariate S-estimate as defined in Section (3.2). This implies that the classical mean and covariance as well as the MVE are special cases of τ-estimates. However, in practice ρ_1 and b_1 are taken to be quite different from ρ_2 and b_2 so that high breakdown and good efficiency can be combined. ρ_1 controls the breakdown properties of the τ-estimate provided that

$$2\rho_2(x)-x\rho_2'(x)>0,\ \text{for } x>0 \tag{20}$$

and ρ_2 is bounded. In fact, it can be shown that for this case, the breakdown point against outliers is given by b_1/M_1 where $M_1=\sup\rho_1$. (20) is needed to guarantee that (18) is strictly increasing in the magnitude of C. ρ_2 alone controls the efficiency of the estimate. To see this, consider the univariate case with ρ_1 defining any estimate with a 50% breakdown point and let $\rho_2=\rho_k^H$ as defined in (4) then

$$\lim_{k\to\infty}\tau^2=\lim_{k\to\infty}\frac{s^2(t)}{n}\sum_{i=1}^{n}\rho_k^H\left(\frac{x_i-t}{s(t)}\right)^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-t)^2$$

which defines the classical estimates.

The other properties of τ-estimates depend on ρ_1, b_1, ρ_2 and b_2 simultaneously. If b_1 and b_2 are chosen so that ρ_1 and ρ_2 each defines a consistent S-estimates then the resulting τ-estimate will also be consistent. As for the asymptotic distribution of the τ-estimate, one must consider the multivariate S-estimate for $\hat{\mathbf{t}}$ and $\hat{\mathbf{C}}$ as defined by (12) and (13), where the functions

$$u(x)=\frac{A\psi_1(x)+B\psi_2(x)}{x}\quad\text{and}\quad v(x)=A\psi_1(x)+B\psi_2(x)-2b_2\{\rho_1(x)-b_1\}$$

depend on the data through

$$A=\frac{1}{n}\sum_{1}^{n}2\rho_2(\mathbf{y}_i)-\mathbf{y}_i\psi_2(\mathbf{y}_i)\quad\text{and}\quad B=\sum_{1}^{n}\mathbf{y}_i\psi_1(\mathbf{y}_i)/n$$

where $\mathbf{y}_i = \sqrt{(\mathbf{x}_i - \mathbf{t})'\mathbf{C}^{-1}(\mathbf{x}_i - \mathbf{t})}$. The asymptotic behaviour of the actual τ-estimators $\hat{\mathbf{t}}$ and $\hat{\mathbf{V}}$ can be obtained from that of $\hat{\mathbf{t}}$ and $\hat{\mathbf{C}}$ [12].

The τ-estimate is computationally more difficult than the S-estimate because it is the solution of a constrained minimization problem. However, a resampling scheme similar to the one derived for S-estimators also works for τ-estimators. Consider

$$h_1(s|\mathbf{t}, \mathbf{C}) = \frac{1}{n}\sum_{i=1}^{n} \rho_1 \left(\frac{\sqrt{(\mathbf{x}_i - \mathbf{t})'\mathbf{C}^{-1}(\mathbf{x}_i - \mathbf{t})}}{s} \right) \tag{21}$$

and

$$h_2(s|\mathbf{t}, \mathbf{C}) = \frac{1}{n}\sum_{i=1}^{n} \rho_2 \left(\frac{\sqrt{(\mathbf{x}_i - \mathbf{t})'\mathbf{C}^{-1}(\mathbf{x}_i - \mathbf{t})}}{s} \right), \tag{22}$$

where $|\mathbf{C}| = 1$. With this notation the τ-estimator becomes the vector $\hat{\mathbf{t}}$ and matrix

$$\hat{\mathbf{V}} = s^2 \hat{\mathbf{C}} h_2(s|\hat{\mathbf{t}}, \hat{\mathbf{C}})/b_2$$

where $\hat{\mathbf{t}}$ and $\hat{\mathbf{C}}$ are the vector and positive definite symmetric matrix that minimize $s^2 h_2(s|\mathbf{t}, \mathbf{C})$ subject to $h_1(s|\mathbf{t}, \mathbf{C}) = b_1$. Therefore for each subsample of size $p + 1$, the mean $\mathbf{t}_i$ and rescaled covariance $\mathbf{C}_i$ such that $|\mathbf{C}_i| = 1$ are used to compute $k_i = s^2 h_2(s|\mathbf{t}_i, \mathbf{C}_i)$ subject to $h_1(s|\mathbf{t}_i, \mathbf{C}_i) = b_1$. The subsample that produces the smallest k_i is used to approximate the τ-estimate.

Like the algorithm for the S-estimate, the equations need not be solved for each subsample because of the monotonicity of $h_1(s|\mathbf{t}, \mathbf{C})$ and $s^2 h_2(s|\mathbf{t}, \mathbf{C})$, see (20). Suppose our current minimium is $k_0 = s_0^2 h_2(s_0|\mathbf{t}_0, \mathbf{C}_0)$ and we have $\mathbf{t}$ and $\mathbf{C}$ from the next subsample drawn. If $h_1(s_0|\mathbf{t}, \mathbf{C}) > b_1$ and $s_0^2 h_2(s_0|\mathbf{t}, \mathbf{C}) > k_0$ then the subsample cannot produce the minimum because s_0 must be increased to satisfy (19), $h_1(s|\mathbf{t}, \mathbf{C})$ is a decreasing function of s, guaranteeing that $s^2 h_2(s|\mathbf{t}, \mathbf{C}) > k_0$ because it is an increasing function of s. Therefore the resampling algorithm for the τ-estimate works as follows.

1. Calculate an initial value for $k_0 = s_0^2 h_2(s_0|\mathbf{t}_0, \mathbf{C}_0)$.

2. For J in $1, \ldots, m$,

   (a) select subsample J containing exactly $p+1$ distinct points,

   (b) compute

$$\mathbf{t}_J = \frac{\sum_{i \in J} \mathbf{x}_i}{p+1} \quad \text{and} \quad \mathbf{C}_J = \frac{1}{p} \sum_{i \in J} (\mathbf{x}_i - \bar{\mathbf{x}}_J)'(\mathbf{x}_i - \bar{\mathbf{x}}_J),$$

   (c) rescale $\mathbf{C}_J$ so that $|\mathbf{C}_J| = 1$,

   (d) if $h_1(s_0|\mathbf{t}_J, \mathbf{C}_J) < b$ or $s_0^2 h_2(s_0|\mathbf{t}_J, \mathbf{C}_J) < k_0$ then:

      i. $s_* = s : h(s|\mathbf{t}_i, \mathbf{C}_i) = b$;

      ii. $\mathbf{t}_* = \mathbf{t}_i$;

      iii. $\mathbf{C}_* = \mathbf{C}_i$;

      iv. if $s_*^2 h_2(s_*|\mathbf{t}_*, \mathbf{C}_*) < k_0$ then:

         A. $s_0 = s_*$;

         B. $\mathbf{t}_0 = \mathbf{t}_*$;

         C. $\mathbf{C}_0 = \mathbf{C}_*$;

         D. $k_0 = s_*^2 h_2(s_*|\mathbf{t}_*, \mathbf{C}_*)$;

3. $\mathbf{V}_0 = \mathbf{C}_0 k_0 / b_2$,

4. return $\mathbf{t}_0$ and $\mathbf{V}_0$.

The expected number of times that (18) needs to be solved, appears to be quite difficult to compute and is left as the subject of further research.

3.4 Stahel-Donoho Estimator

Another multivariate estimator of location and scatter with a high breakdown point was obtained in 1981 by Stahel [27] and independently by Donoho [7] in 1982.

To calculate this estimator, called "outlyingness-weighted mean and covariance", one begins by defining a measure of "outlyingness" for $\mathbf{x}_i \in \mathcal{R}^p$

$$u_i = \sup_{\|\mathbf{v}\|=1} \frac{|\mathbf{v}'\mathbf{x}_i - \text{med}(\mathbf{v}'\mathbf{x}_i)|}{\text{mad}(\mathbf{v}'\mathbf{x}_i)}. \tag{23}$$

Next, one defines a weight function $w : [0, \infty) \to [0, \infty)$ such that $w(x)$ is decreasing and $xw(x)$ is bounded for $x > 0$. Finally one combines $w_i = w(u_i)$ and $\mathbf{x}_i$ to calculate the following weighted mean vector and covariance matrix

$$\mathbf{t} = \frac{\sum_{i=1}^n w_i \mathbf{x}_i}{\sum_{i=1}^n w_i}, \tag{24}$$

$$\mathbf{V} = \frac{\sum_{i=1}^n w_i^2 (\mathbf{x}_i - \mathbf{t})(\mathbf{x}_i - \mathbf{t})'}{\sum_{i=1}^n w_i^2}. \tag{25}$$

Donoho [7] shows that (24) is affine equivariant and has a breakdown point of $([(n+1)/2] - p)/n$ which tends to $1/2$ as $n \to \infty$, provided that no more than p points of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ lie in a $(p-1)$ dimensional affine subspace. Similarly, one can show that the same properties hold for (25). The asymptotic behaviour of the estimators such as rate of convergence, limiting distribution and consistency have not been investigated yet.

The motivation for the definition of u_i is the classical Mahalanobis distance defined by

$$m_i = \sup_{\|\mathbf{v}\|=1} \frac{|\mathbf{v}'\mathbf{x}_i - \bar{\mathbf{v}}'\mathbf{x})|}{\text{SD}(\mathbf{v}'\mathbf{x}_i)} = \sqrt{(\mathbf{x}_i - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})} \tag{26}$$

Unfortunately, u_i cannot be expressed in a form that is easy to compute because unlike the SD, there is no known multivariate covariance estimator whose univariate equivalent is the mad. Therefore, all directions in $\mathcal{R}^p$ must be searched in order to compute a single u_i. Obviously this cannot be done.

In 1990, Patak [20] proposed a modification to Stahel-Donoho estimator that is easy to compute. The method is iterative with the weights for the $(k+1)^{st}$ iteration being updated by the principal components of $\hat{\mathbf{S}}_k$, as calculated in the k^{th} iteration of the

algorithm. Suppose we have $\hat{\mathbf{S}}_k$ and weights W_i^k then the $(k+1)^{st}$ step of the algorithm works as follows:

1. Set $\mathbf{a}_j, \quad j = 1, \ldots, p$ equal to the eigenvectors of $\hat{\mathbf{S}}_k$;
2. Let $w_{ij} = w(u_{ij})$, where $w : [0, \infty) \to [0, \infty)$ is such that $w(x)$ is decreasing, $xw(x)$ is bounded for $x > 0$ and $u_{ij} = |\mathbf{a}_j'\mathbf{x}_i - \text{med}(\mathbf{a}_j'\mathbf{x}_i)|/\text{mad}(\mathbf{a}_j'\mathbf{x}_i)$;
3. Let $W_i^{k+1} = \prod_{j=1}^{p} w_{ij}$;
4. If $W_i^{k+1} \geq W_i^k$ then $W_i^{k+1} = W_i^k$;
5. $\mathbf{t}_{k+1} = \sum_{i=1}^{n} W_i^{k+1}\mathbf{x}_i / \sum_{i=1}^{n} W_i^{k+1}$;
6. $\hat{\mathbf{S}}_{k+1} = \sum_{i=1}^{n}(W_i^{k+1})^2(\mathbf{x}_i - \mathbf{t}_{k+1})(\mathbf{x}_i - \mathbf{t}_{k+1})' / \sum_{i=1}^{n}(W_i^{k+1})^2$.

The weights, after convergence, are used to compute the final estimates $\hat{\mathbf{t}}$ and $\hat{\mathbf{V}}$ as defined in (24) and (25). The convergence of the algorithm is guaranteed because the weights are decreasing functions bounded below by zero. There could be a problem if all the weights converge to zero but Patak showed that at least $p + 1$ points that do not lie in a lower dimensional hyperplane will have weights greater than zero so the estimators will not collapse or become singular. $\hat{\mathbf{t}}$ is consistent, orthogonal equivariant, and possesses a breakdown point of $([n/2] - p)/n$. $\hat{\mathbf{V}}$ is also orthogonal equivariant, possesses a breakdown point of $([n/2] - p)/n$ but is only consistent to within a constant k that depends on the underlying distribution and the particular weight function used. k can be quite difficult to compute in general so I recommend estimating it from the data. Let ρ define an M-estimate of scale, then k can be estimated by $\hat{k}$ which is the solution to

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{\sqrt{(\mathbf{x}_i - \mathbf{t})'\mathbf{V}^{-1}(\mathbf{x}_i - \mathbf{t})}}{k}\right) = b \tag{27}$$

It follows from the properties of M-estimates that $\hat{k}\mathbf{V}$ is consistent.

The above algorithm describes how $\hat{\mathbf{S}}_{k+1}$ is computed from $\hat{\mathbf{S}}_k$. However one needs an initial estimate of covariance $\hat{\mathbf{S}}_0$, which can be calculated through the following algorithm.

1. Set $W_i^0 = 1, \quad i = 1, \ldots, n$;

2. For $k = 1$ to m,

   (a) randomly generate $\mathbf{a}_1 \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$ and normalize;

   (b) calculate $\mathbf{a}_2, \ldots, \mathbf{a}_p \in \mathcal{R}^p$ such that $\mathbf{a}_i'\mathbf{a}_j = 1_{i=j}$, $1 \leq i < j \leq p$;

   (c) compute $w_{ij} = w(u_{ij})$;

   (d) let $W_i^k = \prod_{j=1}^p w_{ij}$

   (e) if $W_i^k < W_i$ then $W_i = W_i^k$;

3. $\mathbf{t}_0 = \sum_{i=1}^n W_i \mathbf{x}_i / \sum_{i=1}^n W_i$;

4. $\hat{\mathbf{S}}_0 = \sum_{i=1}^n W_i^2 (\mathbf{x}_i - \mathbf{t}_0)(\mathbf{x}_i - \mathbf{t}_0)' / \sum_{i=1}^n W_i^2$.

The number of iterations (m) is quite arbitrary. It is used to control computational efficiency. Small m means a very efficient estimate from a computational point of view but the starting point can be extremely poor. If we let m get large, the starting point will be excellent. In fact

$$\lim_{m \to \infty} \mathbf{t}_0 = \mathbf{t} \quad \text{and} \quad \lim_{m \to \infty} \hat{\mathbf{S}}_0 = \mathbf{V}$$

as defined in (24) and (25) respectively, but these estimators are computationally prohibitive. Therefore m must be chosen to optimize the algorithm in terms of speed and quality of the initial estimates. In his thesis, Patak sets $m = 10p$. However it might be more realistic to allow the data to determine the number of iterations. For example, if

the weights do not change significantly after a predetermined number of iterations then stop and use the current value of S_0.

Another modification to this procedure is to centre the data first by some location estimate $\tilde{\mathbf{t}}$. This location estimator should be independent of the covariance estimator and possess all the properties of $\hat{\mathbf{t}}$. Thus, it must be consistent, orthogonal equivariant and have a breakdown point of 1/2 asymptotically. The L1 location estimator [12], defined by

$$\text{argmin}_{\mathbf{T}} \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{T}||$$

has all the required properties. When the modified procedure is carried out, the data are assumed to be centred and therefore, $\mathbf{t}_k = 0$ and $\text{med}(\mathbf{a}_j'\mathbf{x}_i) = 0$ by assumption.

This estimate was originally proposed in the context of principal component analysis and for that purpose, orthogonal equivariance is enough. However, as an estimator of multivariate location and scatter, one would prefer an affine equivariant estimate. $\mathbf{S}_k$ is not affine equivariant because it depends on the principal components of $\mathbf{S}_{k-1}$. $\mathbf{S}_0$ is not affine equivariant either because it uses an orthogonal basis at each iteration. Even if the calculation of $\mathbf{S}_0$ is modified to skip this step, it is still not affine equivariant because, in general,

$$\frac{|\mathbf{a}'\mathbf{x}_i - \text{med}(\mathbf{a}'\mathbf{x}_j)|}{\text{mad}(\mathbf{a}'\mathbf{x}_i)} \neq \frac{|\mathbf{a}'\mathbf{A}\mathbf{x}_i - \text{med}(\mathbf{a}'\mathbf{A}\mathbf{x}_j)|}{\text{mad}(\mathbf{a}'\mathbf{A}\mathbf{x}_i)}$$

for an arbitrary affine transformation $\mathbf{A}$. In order to obtain affine equivariance, the direction $\mathbf{a}$ must be transform to a direction $\mathbf{b}$ such that

$$\frac{|\mathbf{a}'\mathbf{x}_i - \text{med}(\mathbf{a}'\mathbf{x}_j)|}{\text{mad}(\mathbf{a}'\mathbf{x}_i)} = \frac{|\mathbf{b}'\mathbf{A}\mathbf{x}_i - \text{med}(\mathbf{b}'\mathbf{A}\mathbf{x}_j)|}{\text{mad}(\mathbf{b}'\mathbf{A}\mathbf{x}_i)}.$$

This implies that $\mathbf{b} = (\mathbf{A}^{-1})'\mathbf{a}$. So, the algorithm can be made affine equivariant if the transformation $\mathbf{A}$ is known. This is usually not the case so we attempt to estimate it.

The covariance structure of the data is central to estimating a transformation $\mathbf{A}$. If the structure is known before and after the transformation then one can estimate $\mathbf{A}$ by $\mathbf{S}_y^{1/2}\mathbf{S}_x^{-1/2} = \mathbf{A}\mathbf{S}_x^{1/2}\mathbf{S}_x^{-1/2} = \mathbf{A}$ where $\mathbf{y} = \mathbf{A}\mathbf{x}$. When $\mathbf{S}_x$ is unknown, it can assumed to be $\mathbf{I}$ and $\mathbf{A}$ becomes $\mathbf{S}_y^{1/2}$.

This method works great if the data are uncontaminated, but certain types of contamination can destroy the estimated transformation in such a fashion that the direction of the outliers has a lower probability of being searched. Thus, there appears to be no easy way to make the algorithm affine equivariant and robust.

Example 1 *The sample* $\mathbf{y}_1, \ldots, \mathbf{y}_{40}$, *with* $\bar{\mathbf{y}} = (500, 0)$ *and* $\mathbf{S}_y = \mathbf{I}$ *is added to the sample* $\mathbf{x}_1, \ldots, \mathbf{x}_{60}$, *with* $\bar{\mathbf{x}} = (0, 0)$ *and* $\mathbf{S}_x = \mathbf{I}$. *The final covariance structure is*

$$\mathbf{S} = \begin{bmatrix} 60000 & 0 \\ 0 & 1 \end{bmatrix} \quad \textit{and} \quad \mathbf{S}^{-1/2} = \begin{bmatrix} 0.00408 & 0 \\ 0 & 1 \end{bmatrix}$$

If $\mathbf{a} = (0.8, 0.6)$ *then the outliers appear to be approximately 400 standard units away from the centre. However* $\mathbf{S}^{-1/2}\mathbf{a} = (0.00327, 0.6)$ *and the outliers now appear to be anound 2.72 standard units from the centre.*

4 Detection and Testing of Outliers

For the above estimators to work, one only needs a bound on the amount of contamination in the sample. If this bound is not exceeded then the estimators will perform reasonably well. However, these estimators yield no information about the sample itself. They do not estimate ϵ nor do they indicate which points are the outliers if any are present at all. If the sample is clean and the underlying distribution is normal then the best estimators of location and scatter are the sample mean and covariance. Therefore it would be wise to check if the sample is contaminated and use the appropriate estimator based on the results. The decision can be made through various techniques of detecting and testing for the presence of outliers.

In order to detect a multivariate outlier, we must have some notion of "extreme". One such notion is the statistical norm $\mathcal{X}^2$ defined by $\mathcal{X}^2 = (\mathbf{x} - \vec{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x} - \vec{\mu})$, where $\mathbf{x}$ comes from a distribution with mean $\vec{\mu}$ and covariance $\mathbf{\Sigma}$. If the statistical norm is too large then the point is considered to be an outlier. Too large depends on the distribution of $\mathcal{X}^2$. For example if $\mathbf{x}$ is normally distributed then $\mathcal{X}^2$ follows a χ^2_p distribution and too large is is defined as $\mathcal{X}^2 > \chi^2_{p,\alpha}$ where $\chi^2_{p,\alpha} = \{x : \chi^2_p(x) = 1 - \alpha\}$. Unfortunately, the statistical norm assumes the true location and scale are known.

When the true location and scale are unknown, the statistical norm can be estimated by replacing the parameters with accurate estimates. Suppose, we have $\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{y} \overset{\text{iid}}{\sim} F$, $\sum_1^n \mathbf{x}_i/n = \bar{\mathbf{x}}_x$ and $\sum_1^n (\mathbf{x}_i - \bar{\mathbf{x}}_x)'(\mathbf{x}_i - \bar{\mathbf{x}}_x)/(n-1) = \mathbf{S}_x$. Then the sample Mahalanobis distance $D^2 = (\mathbf{y} - \bar{\mathbf{x}}_x)'\mathbf{S}_x^{-1}(\mathbf{y} - \bar{\mathbf{x}}_x)$ can be used to approximate $\mathcal{X}^2$. If F is a normal distribution then D^2 follows a scaled $\mathcal{F}_{p,n-p}$ distribution with scaling factor $k = (n^2 - 1)p/(n-p)n$. Therefore, to detect all the outliers in a sample, $\mathbf{z}_1, \ldots, \mathbf{z}_{n+1}$, one can apply the above procedure to the i^{th} of $n+1$ partitions of the data defined by removing the i^{th} point from tha sample, considering that observation to be $\mathbf{y}$ while the remaining

n points play the role of $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Mathematically, the i^{th} partition is defined as $\mathbf{y} = \mathbf{z}_i$ and

$$\mathbf{x}_j = \begin{cases} \mathbf{z}_j & j = 1, \ldots, i-1, \\ \mathbf{z}_{j+1} & j = i, \ldots, n. \end{cases} \tag{28}$$

Next, one calculates D_i^2 for the i^{th} partition and declare those points such that $D_i^2/k > \mathcal{F}_{p,n-p,\alpha}^{-1}$ to be outliers. The reason for excluding $\mathbf{y}$ from the calculation of $\bar{\mathbf{x}}$ and $\mathbf{S}_x$ is to ensure that the $\mathcal{F}$ distribution holds. If $D_*^2 = (\mathbf{y} - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{y} - \bar{\mathbf{x}})$ is used instead of D^2, where $\bar{\mathbf{x}}$ is the mean of the entire sample and $\mathbf{S}$ is the covariance of the entire sample, then the distribution of D_*^2 is unknown and will have to be derived in order to establish an accurate bound for outlier detection.

Now assume that $\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{y} \overset{\text{iid}}{\sim} \mathrm{F}_\epsilon$, and we wish to check if $\mathbf{y}$ is an outlier. D^2 does not work anymore because we have inaccurate estimates of $\vec{\mu}$ and $\mathbf{\Sigma}$. Using robust estimates, say $\mathbf{t}_x$ and $\mathbf{V}_x$, to calculate $\mathcal{D}^2 = (\mathbf{y} - \mathbf{t}_x)'\mathbf{V}_x^{-1}(\mathbf{y} - \mathbf{t}_x)$, one can solve this problem except that the distribution of $\mathcal{D}^2$ must be derived. Robust estimates work because they are, to a great degree, independent of the outliers in the sample. If one wishes to detect the outliers for an entire sample, then one can calculate $\mathcal{D}_i^2$ for each partition as defined by (28) and declare the points for which $\mathcal{D}_i^2$ is beyond a certain cutoff point, to be outliers. However, the time needed to calculate n robust values of $\mathbf{t}_x$ and $\mathbf{V}_x$ can be quite substantial. Therefore, $\mathbf{t}$ and $\mathbf{V}$ should be calculated only once, using the entire sample, and $\mathcal{D}_*^2 = (\mathbf{y} - \mathbf{t})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{t})$ used instead of $\mathcal{D}^2$. This works because the distributions of the statistic needs to be derived before either can be used and removing an outlier from the sample before the robust estimates are calculated will not change them by much.

Outlier detection does not differentiate between extreme observations and contaminants. In fact, for any fixed cutoff point, we can guarantee the detection of at least one outlier in a clean sample, by increasing the sample size. For example, in a clean sample

of size 200 with $\alpha = .01$, the probability of at least one observation being detected as an outlier is approximately $1 - 0.99^{200} = 0.866$. The advantage of testing a sample for outliers is that regardless of sample size, a good observation will only be rejected with a specified level α. However, the disadvantage is that a contaminant is only rejected with a probability β, which depends on the particular type of contamination and test being used.

4.1 Tests based on Wilk's Lambda

The classical approach to testing and detecting multivariate outliers is based on the Wilk's Lambda statistic used in a MANOVA set up. Suppose $\mathbf{x}_1, \ldots, \mathbf{x}_n \overset{\text{iid}}{\sim} F_\epsilon$ such that k of the n points in the sample come from H. Suppose the sample is permuted so that these points are $\mathbf{x}_1, \ldots, \mathbf{x}_k$. One wishes to test the null hypothesis,

$$H_0 : \mathbf{x}_i \overset{\text{iid}}{\sim} F \text{ for } i = 1, \ldots, n,$$

against the alternative

$$H_1 : \mathbf{x}_i \overset{\text{iid}}{\sim} \begin{cases} H \text{ for } i = 1, \ldots, k, \\ F \text{ for } i = k+1, \ldots, n. \end{cases}$$

Under H_0, the sufficient statistic is $\mathbf{A} = (n-1)\mathbf{S}$ where $\mathbf{S}$ is the classical sample covariance matrix defined on page 12, however if H_1 is true then the sufficient statistic for F is $\mathbf{A}_{(k)} = \sum_{i=k+1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}}_{(k)})(\mathbf{x}_i - \bar{\mathbf{x}}_{(k)})'$ where $\bar{\mathbf{x}}_{(k)} = \sum_{i=k+1}^{n} \mathbf{x}_i/(n-k)$. There are several possible test statistic, all of them based on the eigenvalues of $\mathbf{A}\mathbf{A}_{(k)}^{-1}$ where the i^{th} eigenvalue of $\mathbf{A}\mathbf{A}_{(k)}^{-1}$ is $(1+\lambda_i)$ for $i = 1, \ldots, p$. Wilk's lambda defined as $\Lambda = \prod_1^p (1+\lambda_i)^{-1} = |\mathbf{A}\mathbf{A}_{(k)}^{-1}|^{-1}$ is the generalized likelihood ratio test. Λ can also be defined as $|\mathbf{A}_k|/|\mathbf{A}|$ and in this form it can be seen that Λ rejects H_0 when it is too small because

$$\mathbf{A} = \mathbf{A}_{(k)} + \frac{n-k}{n}\sum_{i=1}^{k}(\mathbf{x}_i - \hat{\mathbf{x}}_{(k)})(\mathbf{x}_i - \hat{\mathbf{x}}_{(k)})'.$$

Λ is known to have good power when H is a $N_p(\mu^*, \Sigma)$ distribution and the null distribution is either known or can be closely approximated. The null distribution depends on the sample size and number of suspected outliers. Let $\Lambda_{n,p,k}$ denote the value of Λ for a p dimensional sample of size n with a k-outlier configuration. The distributions of $\Lambda_{n,p,1}$ and $\Lambda_{n,p,2}$ are very straight forward. $\Lambda_{n,p,1}$ follows a $\beta_{(n-p-1)/2,p/2}$ distribution while $\Lambda_{n,p,2}$ is distributed like a $\beta_{n-p-2,p}$ random variable. More generally, $\Lambda_{n,p,k}$ can be expressed in terms of the product of $(k+1)/2$ beta random variables [8].

Other possible test statistics such as Roy's largest root defined as $\lambda = \max(\lambda_i)$ and Hotellings $T_0^2 = \sum_{i=1}^{p} \lambda_i$ tend to be less powerful than Λ and have unknown null distributions.

When the true outlier configuration is unknown but the exact number present is known, an outlier test can still be performed. This is done by calculating Λ for all $M = n!/k!(n-k)!$ possible configurations and using the one that minimizes Λ to identify the possible outliers. Unfortunately the distribution of $\min(\Lambda_k)$ is unknown and the p-value is therefore based on a conservative bound given by the Bonferroni inequality. Another disadvantage of this method is the substantial time needed to calculate Λ for all possible outlier configurations. For example, if $n = 50$ and $k = 10$, the total number of outlier configurations possible is over ten billion, of which only one will contain all $n-k$ good points in a single partition. In the univariate case, this problem can be avoided because the sample can be ordered resulting in only $k+1$ of the M configurations as possible candidates for the true outlier configuration. Unfortunately, there is no accurate way to order a multivariate sample unless the true location and scatter are known.

The assumption that k is known is another serious weakness of this method because in practice it usually is unknown. An upper bound on k can always be assumed because if $k > \lfloor n/2 \rfloor$ then F would be considered to be the contaminating distribution and

H would be considered to be the true distribution. Now assume that one knows that $k \leq m$ then an outlier test can be performed by considering all $\sum_1^m n!/k!(n-k)!$ possible outlier configurations. For each configuration, the p-value is calculated for $\Lambda_{n,p,k}$ and the most statistically significant configuration is used to identify the number of outliers, and the particular points that are considered as possible outliers. Whether or not these are true outliers is again decided using the Bonferroni inequality.

In 1992, Caroni and Prescott [3] suggested that Wilk's outlier test be applied sequentially assuming that there are at most m outliers in the data. Starting with a single outlier, $\Lambda^i_{n,p,1}$ is computed by removing the i^{th} point from the sample. The statistic $D_1 = \min(\Lambda^i_{n,p,1})$ is saved and the observation that corresponds to the minimum, $\mathbf{x}^{(1)}$, is removed from the sample. The procedure is then applied to the reduced sample $\mathbf{x}_1^{(1)}, \ldots, \mathbf{x}_{n-1}^{(1)}$. After m repetitions of the procedure, one produces a sequence of statistics $D_m, D_{m-1}, \ldots, D_1$, where $D_i = \min \Lambda_{n+1-i,p,1|\mathbf{x}_1^{(i-1)},\ldots,\mathbf{x}_{n-i+1}^{(i-1)}}$ and there corresponding observations $\mathbf{x}_{(m)}, \ldots, \mathbf{x}_{(1)}$. Comparing $D_m, D_{m-1}, \ldots, D_1$ against the appropriate critical values $\lambda_m, \ldots, \lambda_1$, where $\lambda_i = \beta_{\frac{n-p-1}{2}, \frac{p}{2}, \frac{\alpha}{n+1-i}}$, the number of outliers declared by the sequential procedure is the highest value r such that $D_r < \lambda_r$ or zero if none are significant. Caroni and Prescott show that the sequential Wilk's outlier test has a type I error which is only slightly conservative for the sought level α if testing for the $(k+1)^{st}$ outlier when only k exist in the data. Furthermore, the test will be reasonably powerful when only a single outlier remains because at that point, the test becomes equivalent to the generalized likelihood ratio test. Applying the test in a forward fashion makes it immune to the effects of swamping. However, since it is based on classical methods, it will be very vulnerable to the effects of masking. This vulnerability exposes the main weakness with the sequential Wilk's outlier test. If the outliers are masking one another, the sequential method may detect several good data points and ignore the true outliers

altogether. The error is amplified even further if the test actually declares some of these points to be true outliers. See Section 6 for an example of the effect of masking on this test.

4.2 The Scale-Ratio Test

The scale-ratio test was developed in the univariate set up by Le and Zamar ([13]) in 1991. The test statistic is a ratio of two scale estimates, $\tilde{\sigma}$, which is sensitive to outliers, and $\hat{\sigma}$, which is highly robust. The test statistic is defined as $v = \tilde{\sigma}/\hat{\sigma}$ and the null hypothesis is rejected when v is too large. Le and Zamar derive the asymptotic distribution of the scale-ratio test as well as conditions on the estimates, $\tilde{\sigma}$ and $\hat{\sigma}$ so that a Pitman efficient test is obtained.

There are several possible extensions of the scale-ratio test to the multivariate setup. One such generalization is based on the ratio of determinants. For example, suppose, we have two estimates of scatter for the same sample, $\tilde{\Sigma}$, which is highly sensitive to outliers, and $\hat{\Sigma}$ which is highly robust. Summarizing the size of each matrix by its determinant, we define the multivariate scale ratio test to be

$$V_0 = |\tilde{\Sigma}|/|\hat{\Sigma}| \tag{29}$$

It should be noted that V_0 reduces to v^2 when $p = 1$.

Another measure of the relative size of a matrix are its eigenvalues. Suppose $\tilde{\Sigma}$ has eigenvalues $\tilde{\lambda}_1 > \cdots > \tilde{\lambda}_p$ with corresponding eigenvectors $\tilde{\mathbf{e}}_1, \ldots, \tilde{\mathbf{e}}_p$ and $\hat{\Sigma}$ has eigenvalues $\hat{\lambda}_1 > \cdots > \hat{\lambda}_p$ with corresponding eigenvectors $\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_p$ then a test statistic in this case can be defined as

$$V^* = \max \tilde{\lambda}_i/\hat{\lambda}_i \tag{30}$$

However $V_{\max}$ may lack power because $\tilde{\mathbf{e}}_i$ and $\hat{\mathbf{e}}_i$ need not point in the same direction. We will illustrate this through an example.

Example 2

$$\hat{\Sigma} = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \quad \tilde{\Sigma}_1 = \begin{bmatrix} 8 & 0 \\ 0 & 1 \end{bmatrix} \quad \tilde{\Sigma}_2 = \begin{bmatrix} 2 & 0 \\ 0 & 4 \end{bmatrix}$$

Suppose we have a two dimensional sample with covariance structure $\hat{\Sigma}$. *A single outlier is added to expand the scale of one of the co-ordinates by a factor of four. If the contamination occurs in the first co-ordinate, the covariance structure becomes* $\tilde{\Sigma}_1$ *and* $V^* = 4$. *However if the contamination occurs in the second co-ordinate, the covariance structure becomes* $\tilde{\Sigma}_2$ *and* $V^* = 2$. *Therefore* V^* *can lack power under certain situations.*

A third possible measure of the relative size of two matrices is

$$V^*_{\max} = \max_{||\mathbf{a}||=1} \frac{\tilde{\sigma}^2(\mathbf{a}'\mathbf{x}_i)}{\hat{\sigma}^2(\mathbf{a}'\mathbf{x}_i)} \tag{31}$$

where $\tilde{\sigma}$ and $\hat{\sigma}$ are univariate estimates of scale. If $\tilde{\sigma}^2$ and $\hat{\sigma}^2$ are the univariate versions of $\tilde{\Sigma}$ and $\hat{\Sigma}$, respectively and these covariance estimators are affine equivariant then $V^*_{\max}$ reduces to

$$V_{\max} = \max_{||\mathbf{a}||=1} \frac{\mathbf{a}'\tilde{\Sigma}\mathbf{a}}{\mathbf{a}'\hat{\Sigma}\mathbf{a}} = \max_{||\mathbf{b}||=1} \mathbf{b}'\hat{\Sigma}^{-1/2}\tilde{\Sigma}\hat{\Sigma}^{-1/2}\mathbf{b} \tag{32}$$

where $\mathbf{b} = \hat{\Sigma}^{1/2}\mathbf{a}/||\hat{\Sigma}^{1/2}\mathbf{a}||$. Therefore, $V^*_{\max}$ is equal to $\ddot{\lambda}_1$ which is equal to the largest eigenvalue of $\ddot{\Sigma} = \hat{\Sigma}^{-1/2}\tilde{\Sigma}\hat{\Sigma}^{-1/2}$. If at least one of these estimators is not affine equivariant then $V^*_{\max} \approx V_{\max}$ only. It should also be noted that there are some univariate estimators, such as the mad, that do not have a multivariate equivalent. Therefore, $V^*_{\max}$ defines a larger class of test statistics than $V_{\max}$. However, in general $V^*_{\max}$ cannot be computed directly if an equivalent $V_{\max}$ does not exist because all directions of $\mathcal{R}^p$ would have to be searched in order to compute it. Therefore, one should use $V_{\max}$ to approximate $V^*_{\max}$ wherever possible.

The best statistic to use for the multivariate scale-ratio test will depend on the asymptotic distribution and local power versus a specific alternative hypothesis. Another

problem is to choose the specific estimators $\tilde{\Sigma}$ and $\hat{\Sigma}$. In the univariate case, Le and Zamar show that if $\tilde{\sigma}$ and $\hat{\sigma}$ are M-estimates of scale such that $\tilde{\rho} - \hat{\rho} = \beta(x^4 - 6x^2)$, for some $\beta \neq 0$, then the test is Pitman efficient.

Unfortunately, multivariate M-estimates of scale have a breakdown point of at most $1/(p+1)$. Therefore a different estimator must be used. Any of the multivariate estimators proposed in section (3) will work, however, the best one is an area of further research.

The scale-ratio test tests the hypothesis

$$H_0 : \mathbf{x}_i \overset{\text{iid}}{\sim} \mathrm{N}_p(\vec{\mu}, \Sigma) \quad \text{for } i = 1, \ldots, n$$

$$\text{vs} \quad H_1 : \exists \text{ at least one } \mathbf{x}_i \sim H \tag{33}$$

However, if the test rejects, there is no indication of how many or which points are contaminants. This problem is solved by applying the scale-ratio test in a forward sequential fashion. If the test reject H_0 then the point with the largest robust Mahalanobis distance $D_i^2 = (\mathbf{x}_i - \mathbf{t})'\hat{\mathbf{\Sigma}}^{-1}(\mathbf{x}_i - \mathbf{t})$, where $\mathbf{t}$ is the location estimate that corresponds to $\hat{\mathbf{\Sigma}}$, is removed and the test is re-applied to the remaining points. This procedure is repeated until the test does not reject H_0 or 50% of the points have been removed. If the latter happens then one should question the assumption of normality in the null hypothesis.

If each test in the sequence is performed at level α then the overall level of the sequential scale-ratio test is also α. This is obvious since the probability of not rejecting the first point is $1-\alpha$ and no more tests will be conducted if this happens. Furthermore, the probability of rejecting at least one good point when there are k outliers in the data is less than or equal to α, assuming that $\max D_i^2$ corresponds to an outlier if one is still present. This assumption will be verified in Section 5 through simulation. A formal proof is the subject of further research. However, under this assumption, the above

statement holds because the probability of rejecting a good point after all the outliers have been removed is α and the probability of removing all the outliers is less than or equal to one.

Another possible hypothesis that can be tested by the scale ratio test is

$$H_0 : \mathbf{x}_i \overset{\text{iid}}{\sim} \mathrm{N}_p(\vec{\mu}, \boldsymbol{\Sigma})$$

$$\text{vs} \quad H_1 : \mathbf{x}_i \overset{\text{iid}}{\sim} H \tag{34}$$

where H is any non-normal distribution. The test works in this case because the robust covariance matrix is tuned and the cutoff points are set for a normal distribution. If the data are non-normal then the robust estimate will tend to under-estimate the true covariance of a heavy tailed distribution and over-estimate the covariance of a short tailed distribution. The bias in the robust estimates is large enough to cause the test to reject if H is sufficiently different from a normal distribution. The test is applied only once for this alternative and if the null hypothesis is rejected then the only conclusion is that the $\mathbf{x}_i$'s are not normally distributed.

The scale-ratio test can also be modified to work for other distributions besides the normal. To do this, one must calculate proper cutoff points for the test statistic. It may also help to re-tune the robust estimator of scale to fit the true distribution of the data.

Some information about the computer implementation of the scale-ratio test, sequential Wilk's outlier test as well as the approximation algorithms of the robust estimators is available on page 70 in the appendix.

5 Properties of the Scale-Ratio Test

Before the scale-ratio test can be applied, the estimators $\tilde{\mathbf{\Sigma}}$ and $\hat{\mathbf{\Sigma}}$ must be chosen. In all cases $\tilde{\mathbf{\Sigma}} = \mathbf{S}$, the classical sample covariance matrix defined by (2). For $\hat{\mathbf{\Sigma}}$ we will consider two possible estimators, (a) $\hat{\mathbf{\Sigma}}_s$, an S-estimate with $\rho(x)$ defined by (8) and a breakdown point of 50%, and (b) $\hat{\mathbf{\Sigma}}_{pp}$, a Stahel-Donoho estimator, approximated by Patak's algorithm, with

$$w(x) = \begin{cases} 1 & \text{if } x \leq \chi^2_{p,\min(\frac{n-1}{n},0.99)} \\ 0 & \text{otherwise} \end{cases}$$

subject to $\sum w(x_i) < n$ and $\hat{k}$ defined by (8) with $b = 1/2$. Furthermore, both V_0 and $V_{\max}$ as given by (29) and (32) respectively, will be considered as possible test statistics.

The first order of business is to calculate the cutoff points for the scale-ratio test. This is done by computing the appropriate test statistic for 10000 $\mathrm{N}_p(\mathbf{0}, \mathbf{I})$ samples of size n and estimating the cutoff points by the appropriate order statistic. The mean can be $\mathbf{0}$ and the covariance can be $\mathbf{I}$ without loss of generality because the estimators are equivariant to rotations and translations. The results are tabulated in Tables 4-7 for various sample sizes up to 300, $p = 2, 3, 4$ and $\alpha = .05, .025, .01$. Other estimated cutoff points are presented with the specific application in Section 6. It should be noted that the theory of order statistics tells us that the estimated cutoff points $\{\hat{c}_i\}$ are normally distributed with parameters

$$\mu_i = F^{-1}(1-\alpha_i), \quad \sigma_i^2 = \frac{\alpha_i(1-\alpha_i)}{nf^2\left[F^{-1}(1-\alpha_i)\right]} \quad \text{and} \quad \sigma_{ij} = \frac{\alpha_i(1-\alpha_j)}{nf\left[F^{-1}(1-\alpha_i)\right]f\left[F^{-1}(1-\alpha_i)\right]}$$

where $\alpha_i > \alpha_j$, F is the true distribution of the test statistic and f is the density of the test statistic.

Both $\hat{\mathbf{\Sigma}}_s$ and $\hat{\mathbf{\Sigma}}_{pp}$ are random variables for a given data set because both estimates contain a random element. The S-estimate uses a random subset of all possible sub-

samples of size $p+1$ while the Stahel-Donoho estimate is based on random projection in $\mathcal{R}^p$. The variability of the estimates is of interest because it can have a significant effect on their performance. If the $cv = s/\bar{x}$, coefficient of variation, is too large then the approximation will be too imprecise and possibly lack power because of this added variability. Simulations indicate that the cv for $\hat{\Sigma}_{pp}$ is quite a bit smaller that the cv for $\hat{\Sigma}_s$. Both are acceptable when the sample is uncontaminated but the cv for $\hat{\Sigma}_s$ appears to increase dramatically with the level and the magnitude of the contamination. Also, asymmetric contamination appear to be more damaging than symmetric. The cv for $\hat{\Sigma}_{pp}$ is only mildly affected by the presence of contamination. The cv for each estimator is summarized in Tables 8 and 9 for $p = 2$, $\alpha = 0, .05, .25$, $n = 40, 80, 120$ and $H = \mathrm{N}(3, .01), \mathrm{N}(6, .01), \mathrm{N}(0, 3), \mathrm{N}(0, 6)$.

Next, I consider the asymptotic distribution of the test statistics. This is done through a Monte Carlo simulation of 5000 replicates under the null hypothesis. For various sample sizes, a probability histogram is plotted with an appropriate normal curve overlaid. The parameters for the normal curve are estimated from the sample itself. The plots for a sample size of 1000 in 2 dimensions are shown in Figure 3 for the four test statistics considered in this section. All of them appear to follow the normal curve quite closely. Therefore, for larger sample sizes, the cutoff points need not be approximated. Instead, a p-value can be computed based on the appropriate normal curve. Computing the asymptotic mean and variance of the test statistic, the rate of convergence as well as a formal proof of the asymptotic distribution are areas of further research. However, the parameters of the normal curve can be estimated more accurately with fewer replicates than the appropriate order statistics. An estimated p-value is another advantage of the normal approximation.

Finally, we consider the power of the scale-ratio test. The test is applied 1000 times

to random samples of various sizes, dimensions, contamination types and contamination levels in order to estimate the power under these conditions. Keeping all but one factors fixed, the power as function of the changing factor is produced. A plot of this function for each test statistic will help us decide which is best for a given situation. Plotting power as a function of dimension is not done because the contamination type also depends on the dimension. Therefore, the results presented in this section are for the 2-dimensional case only. However, similar results appear to hold in higher dimensions as well. The power curves appear in Figure 4 are for a significance level of 0.01.

Six situations are considered for power calculations. First, a single outlier is added to 99 2-dimensional normal observations to make a sample of size 100. The magnitude of the outlier is increased and the power for a significance level 0.01, is plotted in the top left corner of figure 4. The top right plot corresponds to a sample of size n independent and identically distributed observations that come from $F(x,y) = (1-e^{-|x|})(1-e^{-|y|})$. The power is plotted as a function of sample size. This situation is considered as an example of the sensitivity of the scale-ratio test to non-normal distributions. If a short tailed distribution is used is substituted for the double exponential distribution, the scale-ratio test appears to retain its sensitivity. The middle two plots explore the power of the scale-ratio test when the the sample contains some symmetric contamination. Power as a function of contamination level and sample size are considered here. The bottom two plots are the same as the middle two except that the contamination is asymmetric.

The plots indicate that the test defined by Σ_{pp} is more sensitive than the test defined by Σ_s. However, for Σ_{pp}, the two test statistics V_0 and $V_{\max}$ perform similarly. It is not surprising that V_0 is more powerful when facing symmetric contamination and less powerful for asymmetric, but V_0 also appears to be more sensitive to the normal

assumption. Another interesting feature is that V_0 performs slightly better than $V_{\max}$ when there is only one or two outliers and their deviations are not too extreme.

Table 4: Cutoff Points for V_0 using $\hat{\Sigma}_{pp}$

Sample	Dimension and level								
	2			3			4		
Size	.05	.025	.01	.05	.025	.01	.05	.025	.01
25	1.709	1.972	2.316	1.627	1.842	2.104	1.541	1.756	2.073
30	1.570	1.757	2.045	1.508	1.673	1.913	1.436	1.621	1.890
35	1.484	1.638	1.873	1.453	1.603	1.807	1.382	1.522	1.698
40	1.445	1.582	1.778	1.399	1.530	1.686	1.376	1.513	1.682
45	1.405	1.526	1.696	1.366	1.468	1.643	1.316	1.419	1.568
50	1.374	1.490	1.638	1.324	1.421	1.552	1.287	1.388	1.525
55	1.340	1.445	1.574	1.292	1.382	1.509	1.260	1.342	1.439
60	1.329	1.410	1.529	1.287	1.379	1.509	1.265	1.336	1.445
80	1.256	1.329	1.411	1.231	1.294	1.368	1.205	1.266	1.341
100	1.218	1.277	1.345	1.192	1.247	1.316	1.174	1.231	1.290
120	1.199	1.248	1.316	1.174	1.221	1.286	1.150	1.197	1.259
140	1.179	1.225	1.283	1.161	1.200	1.260	1.148	1.192	1.239
160	1.166	1.206	1.255	1.143	1.184	1.230	1.131	1.167	1.209
180	1.158	1.194	1.244	1.142	1.179	1.223	1.122	1.156	1.196
200	1.147	1.181	1.229	1.129	1.165	1.208	1.118	1.147	1.186
220	1.139	1.174	1.225	1.122	1.156	1.192	1.111	1.140	1.176
240	1.134	1.165	1.204	1.117	1.148	1.184	1.109	1.134	1.166
260	1.121	1.154	1.189	1.112	1.141	1.174	1.101	1.126	1.158
280	1.120	1.151	1.187	1.107	1.133	1.166	1.097	1.121	1.149
300	1.114	1.143	1.179	1.106	1.133	1.164	1.094	1.118	1.147

Table 5: Cutoff Points for $V_{\max}$ using $\hat{\Sigma}_{pp}$

Sample	Dimension and level								
	2			3			4		
Size	.05	.025	.01	.05	.025	.01	.05	.025	.01
25	2.410	2.806	3.469	2.998	3.497	4.400	3.445	4.081	5.085
30	2.071	2.354	2.828	2.439	2.801	3.352	2.704	3.129	3.689
35	1.887	2.098	2.399	2.119	2.382	2.755	2.347	2.626	3.019
40	1.734	1.929	2.161	1.901	2.083	2.362	2.029	2.250	2.533
45	1.645	1.810	2.034	1.792	1.972	2.183	1.878	2.072	2.287
50	1.566	1.688	1.867	1.689	1.825	1.999	1.741	1.876	2.068
55	1.512	1.630	1.803	1.603	1.722	1.904	1.646	1.761	1.907
60	1.462	1.569	1.711	1.552	1.646	1.795	1.590	1.691	1.861
80	1.336	1.403	1.498	1.375	1.454	1.532	1.397	1.469	1.563
100	1.266	1.326	1.397	1.288	1.343	1.405	1.299	1.353	1.413
120	1.233	1.277	1.332	1.248	1.293	1.354	1.249	1.291	1.340
140	1.201	1.242	1.302	1.220	1.261	1.309	1.216	1.250	1.289
160	1.187	1.224	1.267	1.193	1.222	1.266	1.197	1.229	1.270
180	1.172	1.203	1.246	1.177	1.211	1.248	1.176	1.199	1.232
200	1.160	1.188	1.227	1.161	1.187	1.229	1.162	1.186	1.215
220	1.149	1.179	1.216	1.148	1.173	1.206	1.150	1.173	1.204
240	1.141	1.167	1.199	1.142	1.164	1.194	1.142	1.161	1.189
260	1.131	1.153	1.182	1.134	1.156	1.179	1.133	1.152	1.178
280	1.128	1.149	1.181	1.127	1.146	1.173	1.122	1.141	1.162
300	1.121	1.141	1.168	1.121	1.139	1.162	1.117	1.135	1.157

Table 6: Cutoff Points for V_0 using $\hat{\Sigma}_s$

Sample	Dimension and level								
	2			3			4		
Size	.05	.025	.01	.05	.025	.01	.05	.025	.01
25	1.521	1.726	2.061	1.266	1.403	1.598	1.152	1.237	1.363
30	1.442	1.596	1.843	1.254	1.364	1.508	1.163	1.251	1.377
35	1.379	1.519	1.699	1.234	1.320	1.449	1.170	1.253	1.358
40	1.336	1.456	1.619	1.236	1.318	1.419	1.164	1.245	1.327
45	1.311	1.413	1.550	1.224	1.309	1.419	1.175	1.255	1.355
50	1.297	1.392	1.530	1.222	1.295	1.401	1.178	1.251	1.343
55	1.281	1.362	1.480	1.212	1.289	1.391	1.159	1.220	1.306
60	1.258	1.339	1.457	1.206	1.274	1.361	1.156	1.214	1.283
80	1.220	1.280	1.361	1.185	1.237	1.309	1.155	1.208	1.272
100	1.195	1.245	1.321	1.163	1.213	1.276	1.138	1.187	1.249
120	1.178	1.226	1.288	1.150	1.195	1.247	1.130	1.165	1.208
140	1.165	1.209	1.261	1.139	1.177	1.219	1.126	1.161	1.208
160	1.154	1.192	1.238	1.132	1.168	1.217	1.118	1.151	1.193
180	1.144	1.181	1.232	1.125	1.160	1.199	1.112	1.141	1.183
200	1.139	1.175	1.215	1.118	1.147	1.187	1.106	1.132	1.169
220	1.131	1.164	1.199	1.118	1.146	1.188	1.102	1.130	1.158
240	1.122	1.152	1.190	1.110	1.141	1.173	1.098	1.124	1.155
260	1.121	1.149	1.182	1.106	1.131	1.162	1.097	1.119	1.146
280	1.114	1.143	1.181	1.106	1.133	1.160	1.093	1.116	1.146
300	1.112	1.139	1.172	1.101	1.128	1.160	1.090	1.113	1.140

Table 7: Cutoff Points for $V_{\max}$ using $\hat{\Sigma}_s$

Sample	Dimension and level								
	2			3			4		
Size	.05	.025	.01	.05	.025	.01	.05	.025	.01
25	3.000	3.802	4.771	1.974	3.247	5.019	1.038	1.061	1.108
30	2.470	2.995	3.731	1.100	1.385	3.005	1.039	1.060	1.091
35	2.114	2.537	3.175	1.079	1.114	1.635	1.040	1.058	1.080
40	1.926	2.287	2.887	1.074	1.102	1.153	1.041	1.057	1.079
45	1.752	2.111	2.512	1.073	1.099	1.138	1.041	1.058	1.079
50	1.630	1.966	2.399	1.071	1.093	1.130	1.042	1.057	1.077
55	1.479	1.745	2.182	1.067	1.090	1.120	1.038	1.051	1.069
60	1.423	1.750	2.142	1.061	1.079	1.105	1.038	1.050	1.068
80	1.186	1.419	1.706	1.058	1.073	1.093	1.036	1.047	1.060
100	1.123	1.225	1.486	1.053	1.067	1.085	1.034	1.044	1.056
120	1.100	1.153	1.378	1.048	1.063	1.079	1.032	1.040	1.052
140	1.092	1.125	1.265	1.045	1.057	1.073	1.031	1.040	1.049
160	1.083	1.109	1.176	1.042	1.052	1.067	1.029	1.037	1.046
180	1.076	1.101	1.159	1.041	1.052	1.064	1.026	1.034	1.043
200	1.071	1.089	1.121	1.039	1.048	1.062	1.026	1.032	1.041
220	1.064	1.082	1.107	1.038	1.046	1.056	1.025	1.031	1.037
240	1.061	1.079	1.102	1.036	1.044	1.054	1.024	1.030	1.038
260	1.059	1.073	1.094	1.034	1.042	1.053	1.023	1.029	1.035
280	1.057	1.072	1.094	1.033	1.041	1.050	1.022	1.028	1.035
300	1.055	1.068	1.090	1.032	1.040	1.050	1.022	1.027	1.033

Table 8: Coefficient of Variation for 50% biweight S-estimate

Contamination		Sample Size		
Type	Level	40	80	120
None		0.007	0.002	0.002
N(3,.01)	.05	0.019	0.020	0.016
N(6,.01)	.05	0.054	0.052	0.052
N(3,.01)	.25	0.042	0.034	0.025
N(6,.01)	.25	0.108	0.115	0.114
N(0,3)	.05	0.010	0.005	0.001
N(0,6)	.05	0.028	0.023	0.018
N(0,3)	.25	0.027	0.012	0.007
N(0,6)	.25	0.049	0.025	0.019

Table 9: Coefficient of Variation for Stahel-Donoho Estimate

Contamination		Sample Size		
Type	Level	40	80	120
None		0.004	0.002	0.001
N(3,.01)	.05	0.002	0.002	0.001
N(6,.01)	.05	0.001	0.001	0.001
N(3,.01)	.25	0.007	0.009	0.002
N(6,.01)	.25	0.003	0.002	0.001
N(0,3)	.05	0.003	0.002	0.001
N(0,6)	.05	0.005	0.001	0.001
N(0,3)	.25	0.006	0.003	0.002
N(0,6)	.25	0.003	0.002	0.001

Figure 3: Normal Approximations to the Test Statistics of the Scale-Ratio Test

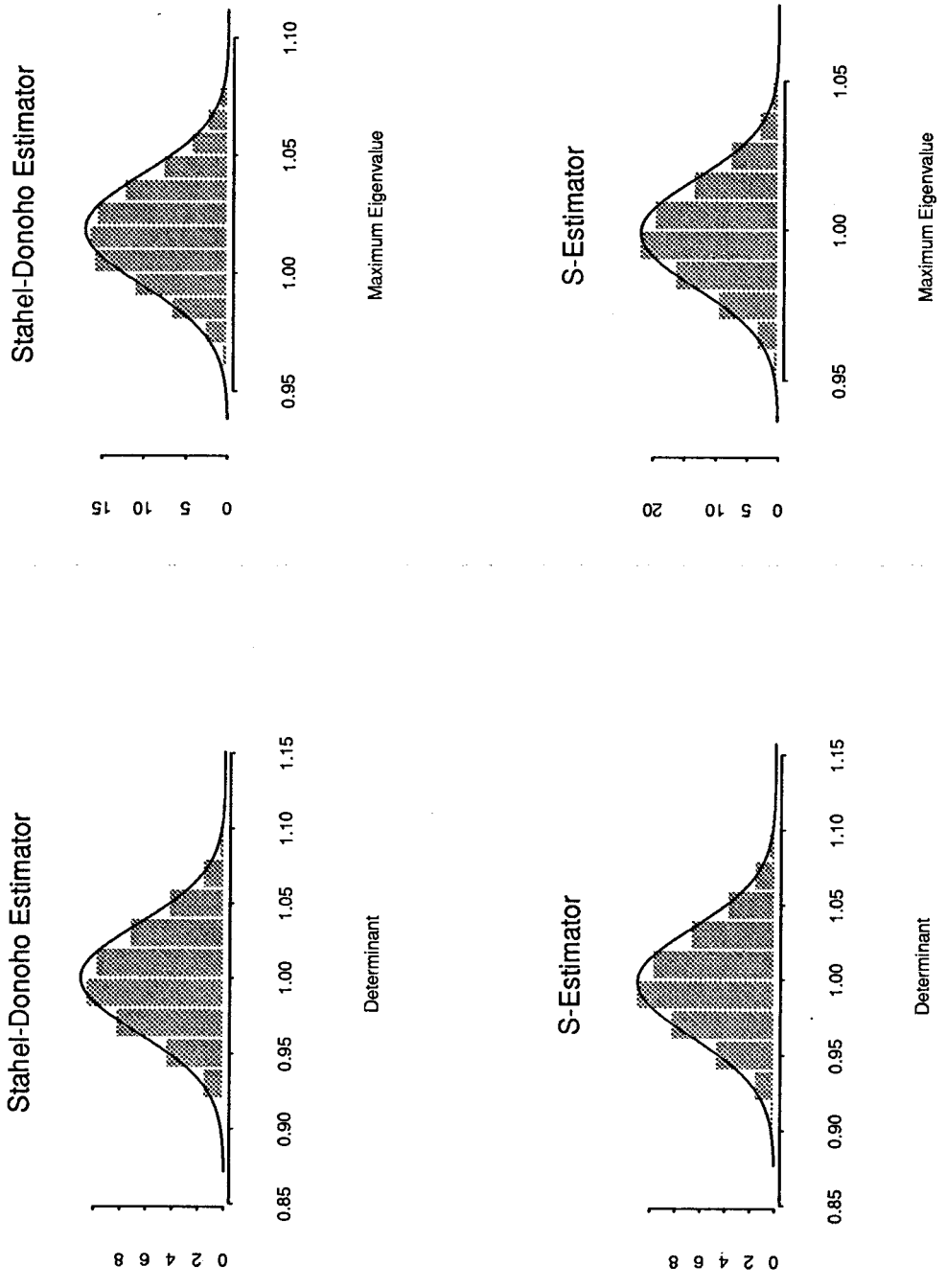
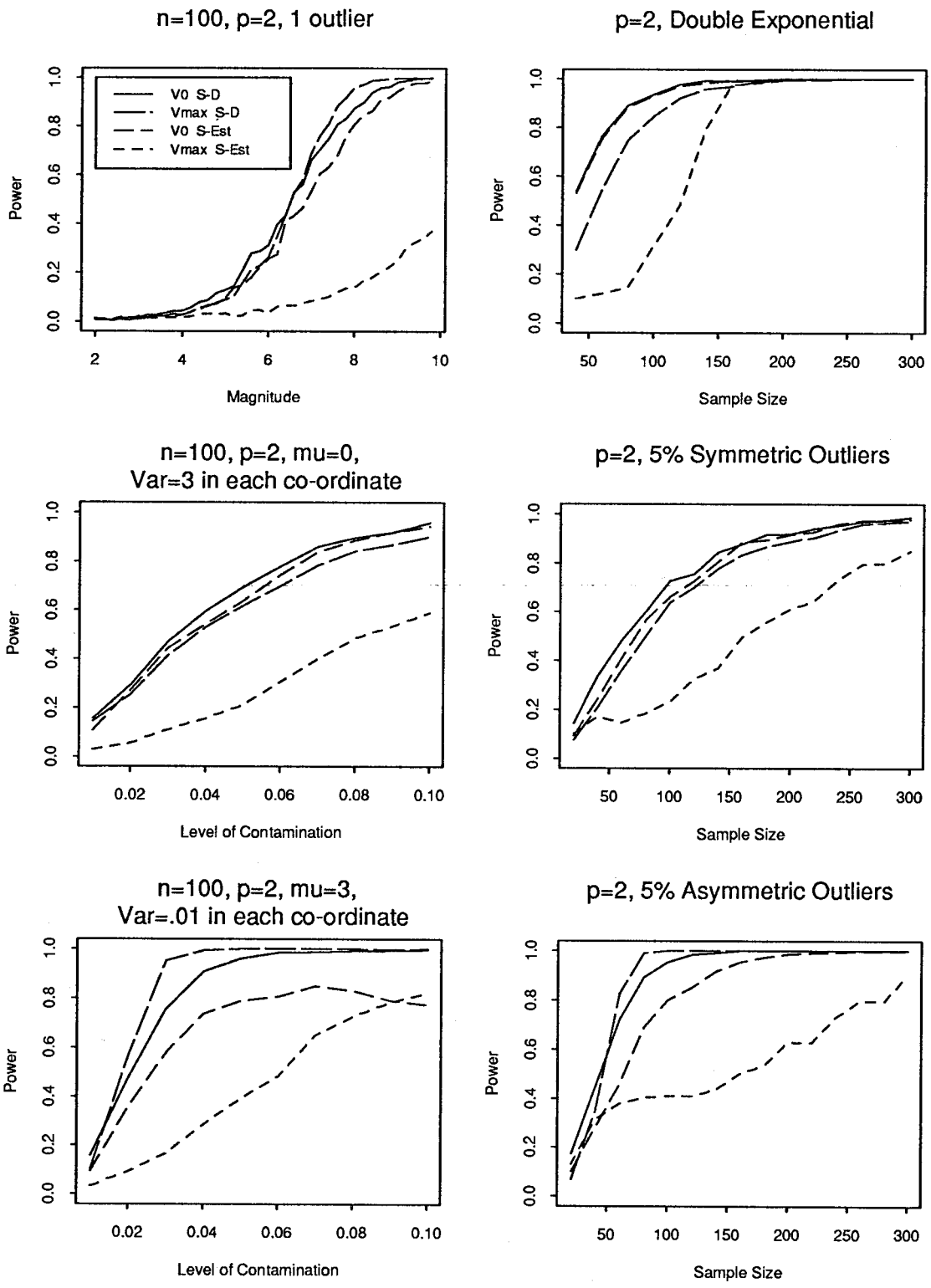


Figure 4: Power Curves for the Scale-Ratio Test



6 Applications of the Scale-Ratio Test

In this section, the scale-ratio test will be applied to several data sets for the purpose of outlier detection. This includes an estimate of the number of outliers present as well as the identity of the bad observations. I will also compare the scale-ratio test to the sequential Wilk's outlier test. For the applications considered, I will use the scale-ratio test defined by Σ_{pp} and V_0 as the robust estimator and test statistic, respectively . This one is chosen because it appears to have reasonable power against most alternatives.

The comparison begins by applying the scale-ratio test to the transportation cost data given in Johnson and Wishern [11]. These data are considered because Caroni and Prescott [3] use them in their paper. The purpose of this example is to show that the scale-ratio test produces similar results to the Wilk's outlier test in situations where the Wilk's test works well. The data appear in Table 10, the pairwise plots along with a few spin plots are in Figure 5, the results for the sequential Wilk's outlier test appear in Table 11 and the results for the scale-ratio test appear in Table 12.

The results of the two tests are very similar. The extreme points are selected in the same order. Observation 9 is the most extreme followed by observation 21 and then observation 36. Both tests agree that that observation 9 is a definite contaminant and observation 36 is not. However, they disagree about observation 21. The Wilk's outlier test rejects this point at .025 level but not at .01 while the scale-ratio test rejects at .10 but not .05. The plots indicate that observation 21 is an outlier and therefore, the scale-ratio test does not appear to be quite as powerful as the Wilk's test for this sample.

The next application for outlier detection comes from the Eastern Lake Survey [6]. The data are extracted for the state of Pennsylvania and the chloride concentration, flouride concentration and sulfate concentration are considered in the log scale. The nitrate concentration is not used because it contains too many missing values. The

Figure 5: Scatterplots and Spin Plots for the Transportation Data

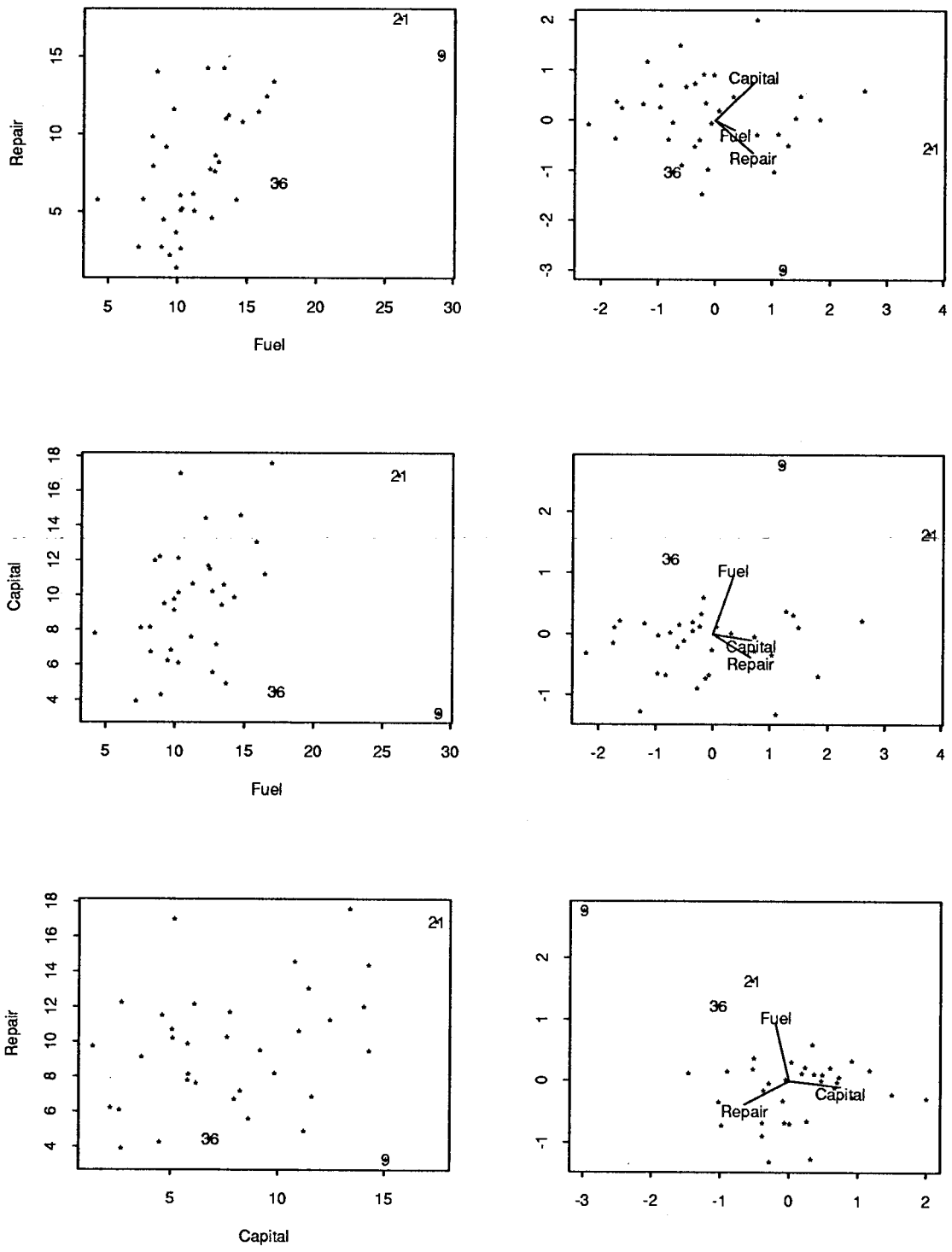


Table 10: Transportation Cost Data: U.S. dollars per mile

Obs	Fuel	Repair	Capital	Obs	Fuel	Repair	Capital
1	16.44	12.43	11.23	19	12.34	7.73	11.68
2	7.19	2.70	3.92	20	8.51	14.02	12.01
3	9.92	1.35	9.75	21	26.16	17.44	16.89
4	4.24	5.78	7.78	22	12.95	8.24	7.18
5	11.20	5.05	10.67	23	16.93	13.37	17.59
6	14.25	5.78	9.88	24	14.70	10.78	14.58
7	13.50	10.98	10.60	25	10.32	5.16	17.00
8	13.32	14.27	9.45	26	8.98	4.49	4.26
9	29.11	15.09	3.28	27	9.70	11.59	6.83
10	12.68	7.61	10.23	28	12.72	8.63	5.59
11	7.51	5.80	8.13	29	9.49	2.16	6.23
12	9.90	3.63	9.13	30	8.22	7.95	6.72
13	10.25	5.07	10.17	31	13.70	11.22	4.91
14	11.11	6.15	7.61	32	8.21	9.85	8.17
15	12.17	14.26	14.39	33	15.86	11.42	13.06
16	10.24	2.59	6.09	34	9.18	9.18	9.49
17	10.18	6.05	12.14	35	12.49	4.57	11.49
18	8.88	2.70	12.23	36	17.32	6.86	4.44

Table 11: Wilks Outlier Test Applied to the Transportation Data

Sample size	Point selected	Wilk's statistic	Critical Values .01	.025	.05	.10
36	9	0.481	0.558	0.592	0.619	0.648
35	21	0.577	0.548	0.583	0.611	0.640
34	36	0.706	0.539	0.574	0.602	0.632

Table 12: Scale-ratio Test Applied to the Transportation Data

Sample size	Point selected	Scale-ratio statistic	Approximate Critical Values .01	.025	.05	.10
36	9	2.167	1.787	1.578	1.435	1.296
35	21	1.321	1.818	1.616	1.449	1.308
34	36	0.968	1.917	1.638	1.466	1.319

Table 13: Wilk's Outlier Test Applied to the Lakes Anion Data

Sample size	Point selected	Scale-ratio statistic	Approximate Critical Values .01	.025	.05	.10
106	33	0.786	0.813	0.828	0.840	0.852
105	98	0.814	0.811	0.827	0.839	0.851
104	69	0.875	0.810	0.826	0.838	0.850

sample size for this subset is 106. A spin plot of the data indicates that observation 33 and 98 are probably outliers with observation 69, 29, 61 and 77 as possible candidates as well. The $Q-Q$ plots of the marginal variables indicate that the most serious departure from normality occurs in the chloride ion which appears to be heavy tailed. The other two appear to follow the normal distribution quite closely. The scatterplots and a few spin plots appear in Figure 6, the results for the Wilk's outlier test are in Table 13 and the results for the scale-ratio test are in Table 14.

Again, the two tests seem to be in close agreement. The only difference is that the Wilk's outlier test only rejects observations 33 and 98 as outliers where as the scale-ratio test also includes 69 in this category.

The next example is used to demonstrate the vulnerability of the Wilk' outlier test to the effects of masking. It also serves as a demonstration that the scale-ratio test does not have the same weakness. The data are fictitious with obvious but severely masked outliers. The data appear in Table 15, the results of the Wilk's outliers test are in Table 16 and the results for the scale-ratio test are in Tablesmask.

Knowing that there is at most 5 outliers in the data, the Wilk's Outlier test is applied 6 times to the data, the points detected as potential outliers are, in order of detection, $\mathbf{x}_{12}, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_{11}, \mathbf{x}_8$ and $\mathbf{x}_{13}$. By looking at the data it is clear that the true outliers are

Figure 6: Scatterplots and Spin Plots for the Lakes Data

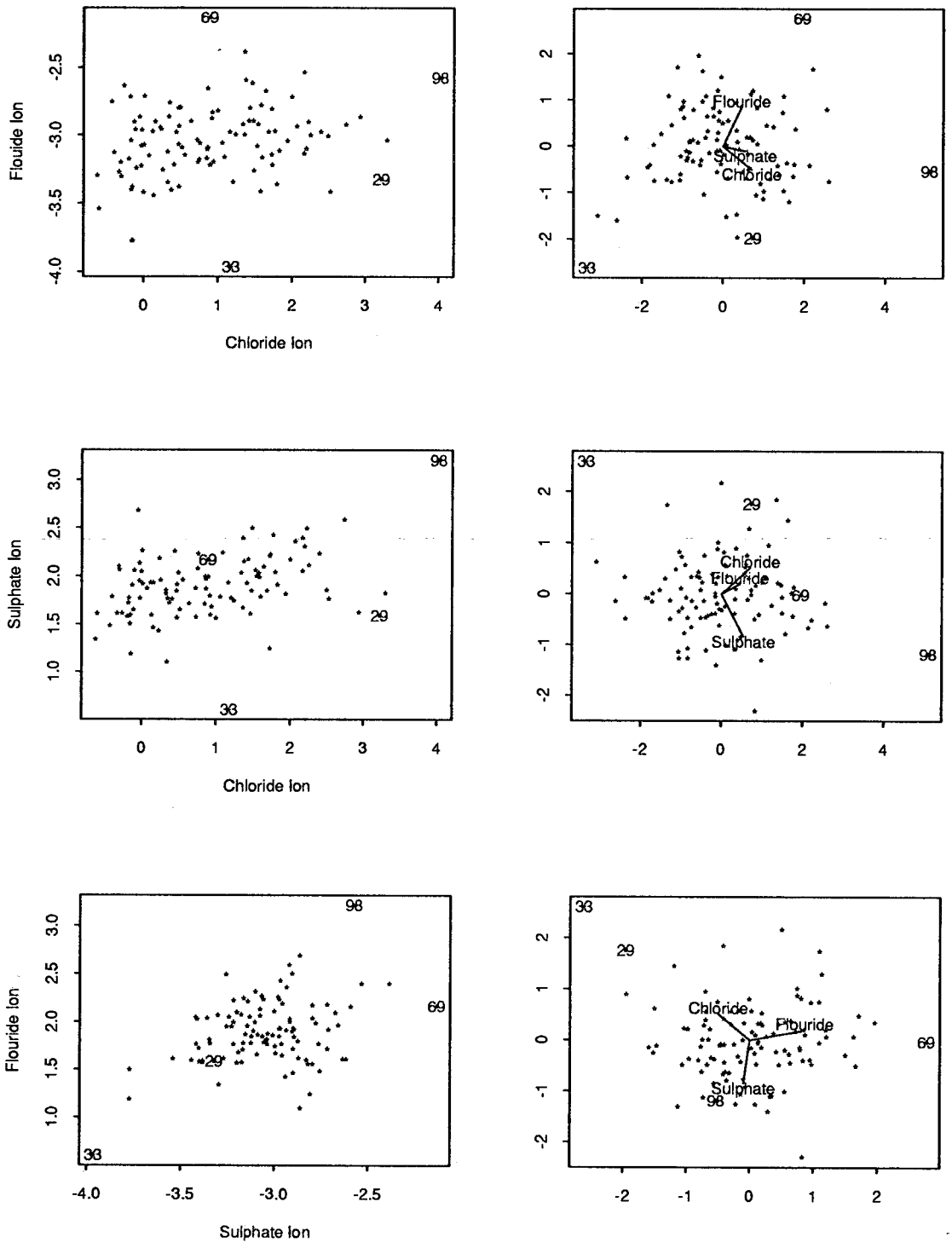


Table 14: Scale-ratio Test Applied to the Lakes Anion Data

Sample size	Point selected	Scale-ratio statistic	Approximate Critical Values .01	.025	.05	.10
106	33	1.632	1.314	1.236	1.182	1.130
105	98	1.387	1.318	1.243	1.188	1.135
104	69	1.232	1.217	1.245	1.191	1.133
103	29	1.185	1.314	1.247	1.192	1.136

Table 15: Masked Outlier Data

obs	x	y	obs	x	y	obs	x	y
1	0.393	-2.193	10	0.136	-0.287	19	0.827	0.516
2	-1.721	1.086	11	1.160	-1.042	20	-0.953	0.436
3	0.700	0.035	12	2.023	-2.296	21	1000.044	1000.079
4	0.050	-0.028	13	-0.516	0.728	22	999.912	999.971
5	-0.383	-0.696	14	0.693	-0.638	23	1000.121	1000.158
6	-1.605	0.434	15	-0.304	-0.913	24	999.951	1000.079
7	-1.292	0.451	16	-0.558	0.177	25	1000.158	1000.194
8	0.586	-0.879	17	0.430	0.536			
9	-0.327	0.886	18	-1.012	0.363			

Table 16: Wilk's Outlier Test Applied to the Masked Outlier Data

Sample size	Point selected	Wilk's statistic	Critical Values .01	.025	.05	.10
25	12	0.665	0.491	0.533	0.568	0.605
24	1	0.788	0.477	0.520	0.555	0.593
23	2	0.773	0.461	0.505	0.542	0.581
22	11	0.748	0.445	0.490	0.527	0.567
21	8	0.808	0.427	0.473	0.511	0.552
20	13	0.771	0.409	0.455	0.494	0.536

$\mathbf{x}_{21}, \ldots, \mathbf{x}_{25}$. None of the detected points are significant but this still demonstrates a severe lack of power when the outliers mask each other. If Wilk's outlier test is applied in a non-sequential fashion, the correct 5 points are detected as the potential outliers and $\Lambda_5 = 0.000000736$ is highly significant. The problem with applying Wilk's Outlier test directly comes from the swamping effect. See Caroni and Prescott [3] for an example of this.

When the scale-ratio test is applied to the same data, the outliers, in order of detection, are $\mathbf{x}_{25}, \mathbf{x}_{23}, \mathbf{x}_{21}, \mathbf{x}_{24}$ and $\mathbf{x}_{22}$. The test is highly significant for these 5 points and declares them all to be contaminants but when the test is applied to the remaining 20 observation $\mathbf{x}_1$ is detected as the extrme observation but it is not rejected as a contaminant. For this extreme example, the scale-ratio test yields a correct result while the Wilk's outlier test is completely fooled.

The final example is designed to show the sensitivity of the Wilk's Outlier test to the number of suspected outliers. The data consist of 50 observations in 4 dimensions. Observations 46 through 50 are structural outliers that do not appear in any of the 3-

Table 17: Scale-ratio Test Applied to the Masked Outlier Data

Sample size	Point selected	Scale-ratio statistic	Approximate Critical Values .01	.025	.05	.10
25	25	342515.530	2.376	1.962	1.711	1.492
24	23	382872.985	2.447	2.018	1.754	1.512
23	21	399269.468	2.473	2.085	1.796	1.529
22	24	319646.961	2.565	2.123	1.848	1.553
21	22	254922.959	2.749	2.201	1.889	1.589
20	1	1.042	2.724	2.245	1.908	1.600

dimensional spin plots or 2-dimensional scatterplots. Figure 7 shows the 6 scatterplots. The results of the Wilk's outlier test are in Table 18 while the results for the scale-ratio are in Table 19.

The Wilk's Outlier test is affected by masking in this situation but not completely fooled by it. However, a single application of the Wilk's outlier test indicates no outliers. Therefore, if the level of contamination is unknown, an investigator may apply the test once then stop because it yields an insignificant result. The scale-ratio test, on the other hand, is unaffected by the masking and indicates an outlier problem with the first application of the test. This may not seem very serious because the Wilk's test is very significant after 4 applications but there could be several layers of masked outliers in the data resulting in a swing from insignificant to significant results as the test works through each layer of contamination.

The above examples demonstrate the performance of the scale-ratio test. It appears to work as well as the Wilk's outlier test when the outliers are not masked and dramatically out performs the Wilks outlier test when the outliers are masked.

Figure 7: Scatterplots for a 4-dimensional Data Set with hidden Outliers

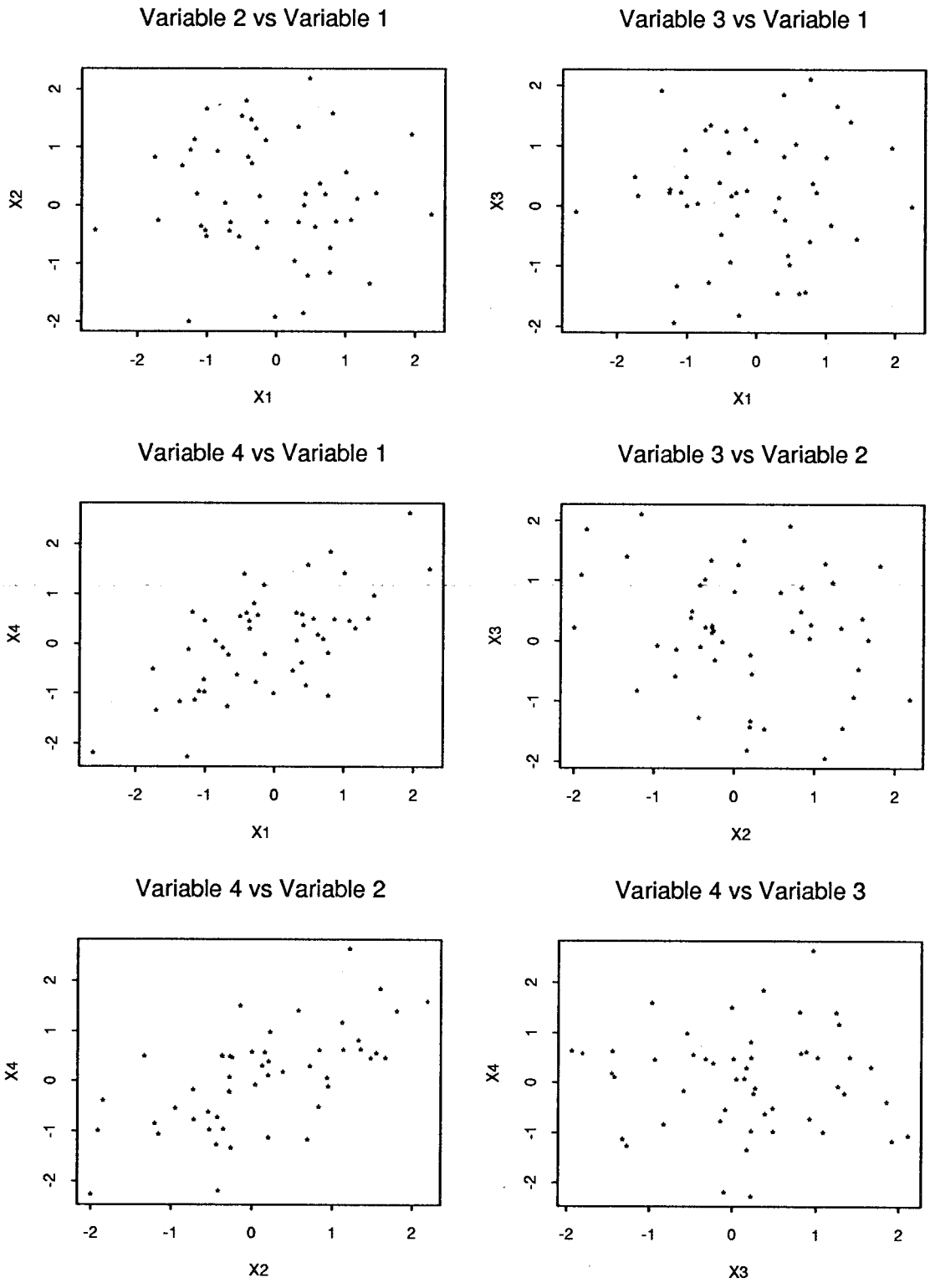


Table 18: Wilks Outlier Test Applied to the Hidden Outlier Data

Sample size	Point selected	Wilk's statistic	Critical Values .01	.025	.05	.10
50	47	0.727	0.619	0.647	0.669	0.692
49	50	0.666	0.614	0.642	0.664	0.687
48	48	0.620	0.607	0.636	0.658	0.682
47	49	0.446	0.601	0.630	0.653	0.677
46	46	0.008	0.594	0.624	0.647	0.671
45	34	0.771	0.588	0.617	0.641	0.665

Table 19: Scale-ratio Test Applied to the Hidden Outlier Data

Sample size	Point selected	Scale-ratio statistic	Approximate Critical Values .01	.025	.05	.10
50	50	272.123	1.522	1.385	1.292	1.197
49	48	246.926	1.510	1.376	1.295	1.199
48	47	213.242	1.533	1.389	1.289	1.193
47	49	158.282	1.532	1.401	1.298	1.185
46	46	88.168	1.550	1.401	1.304	1.198
45	34	0.920	1.546	1.414	1.309	1.202

7 Conclusions and Recommendations

The limitation of graphical techniques for detecting outliers illustrates the need for other methods of outlier detection in higher dimensions. This doesn't mean that the outliers tests should be blindly applied in higher dimensions. In fact, exploratory data analysis should always be carried out before any tests are conducted. This will help in the detection of obvious outliers as well as revealing some basic symmetries in the data such as univariate normality or elliptical contours in the bivariate margins.

After the exploratory analysis is done, outlier tests can be conducted for the structural outliers. The assumption of multivariate normality for the Wilk's outlier test and the scale-ratio test is not very limiting because the univariate margins can be converted to near normality before either test is applied. This does not even guarantee bivariate normality let alone p-variate normality but non-normal univariate margins does guarantee non-normal margins in higher dimensions.

Robust estimation plays an important role in the scale-ratio test because the robust estimator is responsible for the resistance of the test against the masking effect. I have never suggested that robust estimation should be used in place of classical estimation. In fact, I suggest that the robust estimates be used only as part of a diagnostic tool to detect the outliers. Once the outliers have been detected, investigated and dealt with properly, classical estimates can be used based on the new information. This should result in a highly efficient and robust method of dealing with outliers.

The sequential application of the scale-ratio test or the Wilk's outlier tests is responsible for the resistance to the swamping effect. The resistance comes from the removal of the outliers before the next step is conducted. When no outliers remain in the data, the test should no longer be significant.

The specific test to use is not clear. I recommend using the scale-ratio test over

the Wilk's or sequential Wilk's outlier test because it is unaffected by swamping or masking. However, the best scale-ratio test is an area of further research. Of the four combinations I tried, I found V_0 based on Σ_{pp} as defined in Section 5 to be the best because it appears to have greater power than the others in most situation an it appears to be asymptotically normally distributed.

The cutoff point need to be investigated further. They seem to decrease as the sample size and dimension increases. I believe a smooth curve will fit through the cutoff point for a given α level allowing them to be estimated by a simple formula. The asymptotic normality also allows for the modelling of the mean and variance of the cutoff points as another way of estimating their values for larger sample sizes.

If the critical values are ignored, the scale-ratio test can be used as a method of ordering multivariate data in terms of extremeness from the centre. Once the data has been ordered, one can investigate a percentage of the most or least extreme points.

As for the approximation to the multivariate robust estimators, I think these appear to work quite well as they are but further research could be used to improve them. For example, the Stahel-Donoho estimator can be used to increase the probability of selecting a subsample that contains all good point, for the calculation of an S-estimate or a τ-estimate, by sampling each point based on the weight assigned by the Stahel-Donoho estimator.

References

[1] Afifi, A. A. and Azen, A. P. (1972). *Statistical Analysis A Computer Oriented Approach,* New York: Academic Press.

[2] Barnett, V. and Lewis, T. (1984). *Outliers in Statistical Data* (2ed ed.), Toronto: John Wiley & Sons.

[3] Caroni, C. and Prescott, P. (1992). "Sequential Applications of Wilk's Multivariate Outlier Test", *Applied Statistics,* **41** 355-364.

[4] Das, R. Sinha, B. K. (1986). "Detection of Multivariate Outliers with Dispersion Slippage in Elliptically Symmetric Distributions", *The Annals of Statistics,* **14**, 1619-1624.

[5] Davies, P. L. (1987). "Asymptotic behaviour of S-estimates of Multivariate Location and Dispersion Matrices" *The Annals of Statistics,* **15** 1269-1292.

[6] Delempady, M. and Douglas, A. (1990). "Eastern Lake Survey - Phase 1. Documentation for the Data and the Derived Data Sets." *Sims Technical Report,* **160** The University of British Columbia.

[7] Donoho, D. L. (1982). Breakdown Properties of Multivariate Location Estimators. Ph.D. Qualifying Paper. Harvard University.

[8] Hawkins, D. M. (1980). *Identification of Outliers* , New York: Chapman and Hall.

[9] Huber, P. J. (1964). "Robust estimation of a Location Parameter" *The Annals of Mathematical Statistics,* **35** 73-101.

[10] Huber, P. J. (1981). *Robust Statistics,* New York: John Wiley & Sons.

[11] Johnson, R. A. and Wichern, D. W. (1988). *Applied Multivariate Statistical Analysis,* Englewood Cliffs, New Jersey: Prentice-Hall.

[12] Lopuhaä, R. (1990). *Estimation of Location and Covariance with High Breakdown Point,* Ph.D. Thesis, Delft University of Technology, Netherlands.

[13] Le, N. D. and Zamar, R. H. (1991). " A Global Test for Effects in a 2^k Factorial Design Without Replicates" *The Journal of Statistical Computing and Simulation,* **41** 41-54.

[14] Lind, J. C. (1979). *A Comparison of Some Robust Estimates of Correlations in the Presence of Asymmetry,* Ph.D. Thesis, University of British Columbia.

[15] Maronna, R. A. (1976). Robust M-estimates of Multivariate Location and Scatter. *Annals of Statistics* **4** 51-67.

[16] Martin, R. D., Yohai, V. J. and Zamar, R. H. (1989). "Min-max Bias Robust Regression." *Annals os Statistics,* **17** 1608-1630.

[17] Martin, R. D. and Zamar R. H. (1992). Bias Robust Estimation of Scale when Location is Unknown. Technical Report.

[18] Mood, A. M., F. A. Graybill, and D. C. Boes (1974). *Introduction to the Theory of Statistics* (3rd ed.), Montreal: McGraw-Hill Book Co.

[19] Moussa M. A. A. (1983). "Detection and Accommodation of Multivariate Statistical Outliers" *Computer Programs in Biomedicine,* **16**, 217-230.

[20] Patak, Z. S. (1990) *Robust Principal Component Analysis via Projection Pursuit,* Master Thesis, University of British Columbia.

[21] Press, W. H. et al, (1988) *Numerical Recipes in C : The Art of Scientific Programming,* Cambridge, England: Cambridge University Press.

[22] Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection,* Toronto: John Wiley & Sons.

[23] Rousseeuw, P. J. and Yohai, V. (1984) "Robust Regression by Mears of S-Estimators. *Robust and Nonlinear Time Series Analysis* Lecture Notes in Statistics **26** 256-272. Springer Verlag, New York.

[24] Serfling, R. J. (1980) *Approximation Theorems os Mathematical Statistics,* Toronto John Wiley & Sons.

[25] Sinha, B. K. (1984). "Detection of Multivariate Outliers in Elliptically Symmetric Distributions", *The Annals of Statistics,* **12**, 1558-1565.

[26] Schwager, S. J. and Margolin, B. H. (1982). "Detection of Multivariate Normal Outliers", *The Annals of Statistics,* **10**, 943-954.

[27] Stahel, W. A. (1981). Robust Estimation: Infinitesimal Optimality and Covariance Matrix Estimators. Ph.D. Thesis (in German), ETH, Zurich.

[28] Yohai, V. J. and Zamar, R. (1988). "High breakdown-point of Estimates of Regression by Means of the Minimization of an Efficient Scale", *Journal of the American Statistical Association,* **83**, 406-413.

[29] Yohai, V. J. and Zamar, R. (1990). Discussion of Paper No. 90 SM 493-7 PWRS.

Appendix : Computer Implementation

Most of the software used in this thesis was written by the author in the C language and run on a Sun workstation. The model of workstation used was either a Sparc2, Sparc1, or SparcELC. The one exception to this was the Wilk's sequential outlier test routine which was written in Splus3.0.

The core routines manipulate matrices in one way or another. Since several procedures require the eigenvalues of a matrix, any procedure that could be calculated from the spectral decomposition of a matrix was calculated this way. The spectral decomposition routine comes from Numerical Recipes in C [21]. The specific routines are TRED2 and TQLI. The uniform deviates were generated by the RANDOM function built in to the standard C library on a Sun computer, a Newton-Raphson routine was used to solve for the constants in Table 2 and for the scale in the S-estimate approximation algorithm, Romberg integration was used for the expected value of the biweight function for a given k, and the quicksort sorting algorithm was used to sort any data that needed sorting. The Newton-Raphson, Romberg and quicksort routines were written by the author.

The code is available upon request. However, it is undocumented at this time and it may not compile properly on any machine other than a Sun workstation.