Results in Finding an Unknown Number of Multivariate Outliers in Large Data Sets

Marco Riani[1*], Anthony C. Atkinson[2†] and Andrea Cerioli[1‡]
[1]Dipartimento di Economia, Università di Parma, Italy
[2]The London School of Economics, London WC2A 2AE, UK

April 25, 2007

Abstract

We use the forward search to provide parameter estimates for Mahalanobis distances used to detect the presence of outliers in a sample of multivariate normal data. Theoretical results on order statistics and on estimation in truncated samples provide the distribution of our test statistic. Comparisons of our procedure with tests using other robust Mahalanobis distances show the good size and high power of our procedure. We also provide a unification of results on correction factors for estimation from truncated samples.

Keywords: forward search; graphics; logistic plots; Mahalanobis distance; order statistics; power comparisons; truncated distributions; very robust methods

1 Introduction

The normal distribution, perhaps following data transformation, has a central place in the analysis of multivariate data. Mahalanobis distances provide the standard test for outliers in such data. However, the performance of the test depends crucially on the subset of observations used to estimate the parameters of the distribution.

It is well known that the estimates of the mean and covariance matrix using all the data are extremely sensitive to the presence of outliers. In this paper we use the forward search to provide an adaptively selected sequence of subsets of the data from which the parameters are estimated. We compare the resulting Mahalanobis distances as an outlier test with a variety of robust procedures, all of which can be described as using estimates based on one or two subsets. We show that our procedure has superior power as well as good size and so is to be recommended.

Mahalanobis distances and the forward search are introduced in §2. In §3 we exhibit bootstrap envelopes for the distribution of distances in the forward search. Theoretical results on the distribution are in §4. In particular, §4.1 uses results on order statistics to find

*e-mail: mriani@unipr.it
†e-mail: a.c.atkinson@lse.ac.uk
‡e-mail: andrea.cerioli@unipr.it

the distribution of ordered Mahalanobis distances. In §4.2 we use a result of Tallis (1963) to adjust for the bias caused by estimation of the covariance from a subset of observations.

We use Mahalanobis distances to develop a test for the presence of one or more outliers in a sample. Our procedure for this form of outlier detection is described in §5 with two examples in the following section. Several established robust procedures for the detection of individual outlying observations, such as those of Davies and Gather (1993), Rousseeuw and Van Driessen (1999) and Hardin and Rocke (2005) are introduced in §7. Some of these methods use reweighted estimates and so are based on two subsamples of the data. To adapt these tests to determining whether there are any outliers in a sample, we introduce in §8 a Bonferroni correction to allow for simultaneity. This allows us to develop two new versions of reweighted Mahalanobis distances. The comparisons of size in §9.1 show that our procedure has better size than many competitors. In §9.2 we use logistic plots of power to provide simple comparisons of tests with markedly different sizes. The results show the superior performance of our procedure.

Examples of analyses of individual sets of data are in §10. The first appendix discusses the importance of careful numerical procedures in the calculation of extreme values of order statistics and the second draws a connection between the results of Tallis and the distribution of observations in a truncated univariate normal distribution.

Our procedure provides the most powerful test for outliers amongst those in our comparisons. It can be further enhanced by use of the rich variety of information that arises from monitoring the forward search.

2 Distances

The main tools that we use are the values of various Mahalanobis distances. The squared distances for the sample are defined as

$$d_i^2 = \{y_i - \hat{\mu}\}^T \hat{\Sigma}^{-1} \{y_i - \hat{\mu}\}, \tag{1}$$

where $\hat{\mu}$ and $\hat{\Sigma}$ are the unbiased moment estimators of the mean and covariance matrix of the n observations and y_i is $v \times 1$.

In the methods compared in this paper the parameters μ and Σ are estimated from a subset of m observations, yielding estimates $\hat{\mu}(m)$ with $\hat{\mu}(m)_j = \bar{y}_j$ and $\hat{\Sigma}(m)$ with $\hat{\Sigma}(m)_{jk} = (y_j - \bar{y}_j)^T (y_k - \bar{y}_k)/(m-1)$. Note that here y_j and y_k are $m \times 1$. From this subset we obtain n squared Mahalanobis distances

$$d_i^2(m) = \{y_i - \hat{\mu}(m)\}^T \hat{\Sigma}^{-1}(m) \{y_i - \hat{\mu}(m)\}, \qquad i = 1, \ldots, n. \tag{2}$$

The single subsets used for each MCD-based method are defined in §7. In the forward search we use many subsets for outlier detection, rather than one. The difference is between viewing a movie and a single snapshot.

In the forward search we start with a subset of m_0 observations which grows in size during the search. When a subset $S^*(m)$ of m observations is used in fitting we order the squared distances and take the observations corresponding to the $m+1$ smallest as the new subset $S^*(m+1)$. Usually this process augments the subset by one observation, but sometimes two or more observations enter as one or more leave. To start the procedure we find a starting subset $S^*(m_0)$ that is not outlying in any two-dimensional projection of the data (Atkinson et al. 2004, §2.13).

In our examples we look at forward plots of quantities derived from the distances $d_i(m)$ in which the parameters are estimated from the observations in $S^*(m)$. These distances for $i \notin S^*(m)$ tend to decrease as n increases. If interest is in the latter part of the search we may use **scaled** distances

$$d_i^{\text{sc}}(m) = d_i(m) \times \left(|\hat{\Sigma}(m)| / |\hat{\Sigma}(n)| \right)^{1/2v}, \tag{3}$$

where $\hat{\Sigma}(n)$ is the estimate of Σ at the end of the search.

To detect outliers all methods compare selected Mahalanobis distances with a threshold. We examine the minimum Mahalanobis distance amongst observations not in the subset

$$d_{\text{min}}(m) = \min d_i(m) \quad i \notin S^*(m), \tag{4}$$

or its scaled version $d_{\text{min}}^{\text{sc}}(m)$. If this ordered observation $[m+1]$ is an outlier relative to the other m observations, this distance will be large compared to the maximum Mahalanobis distance of observations in the subset. Because we potentially make many comparisons, one for each value of m, the form of our threshold needs to allow for simultaneity, so that we have a test with size α for the presence of at least one outlier. Adjustment for simultaneity in the other procedures is discussed in §7.

3 The Structure of Forward Plots and the Importance of Envelopes: Swiss Banknotes

Our purpose is to provide methods for relatively large data sets. Here we present a brief analysis of a smaller example, which illustrates the use of forward plots with thresholds that are pointwise envelopes. In this example the bootstrap envelopes are found by simulating the search 10,000 times. For larger examples we use the theoretical results of §4

Flury and Riedwyl (1988, pp. 4–8) introduce 200 six-dimensional observations on Swiss banknotes. Of these, units 101 to 200 are believed to be forgeries. The left-hand panel of Figure 1 shows a forward plot of the (unscaled) minimum Mahalanobis distances for the forgeries. There is a large peak at m = 85, indicating that there are at least 15 outliers. The peak occurs because the outliers form a loose cluster. Once one of these observations has been included in $S^*(m)$, the parameter estimates are slightly changed, making less remote the next outlier in the cluster. At the end of the search the distances increase again when the remaining observations not in $S^*(m)$ are somewhat remote from the cluster of outliers. Large distances at the end of the search are, as shown in Figure 5, typical of data with unclustered outliers.

An important feature of Figure 1 is that the plot goes outside the upper envelope when m is slightly less than 85. This is because, if we have a sample of 85 observations from the normal distribution, the last few distances will be relatively large and the envelope will curve upwards as it does in the plots for m a little less than 100. If we remove the 15 observations that form the outlying group and superimpose the new envelope for n = 85, we can see whether all outliers have been identified.

The curve for scaled distances in the right-hand panel of the figure lies below the envelopes in the earlier part of the search because scaling is by the estimate $\hat{\Sigma}(n)$ from the end of the search, which is inflated by the presence of the outliers. Hence the distances

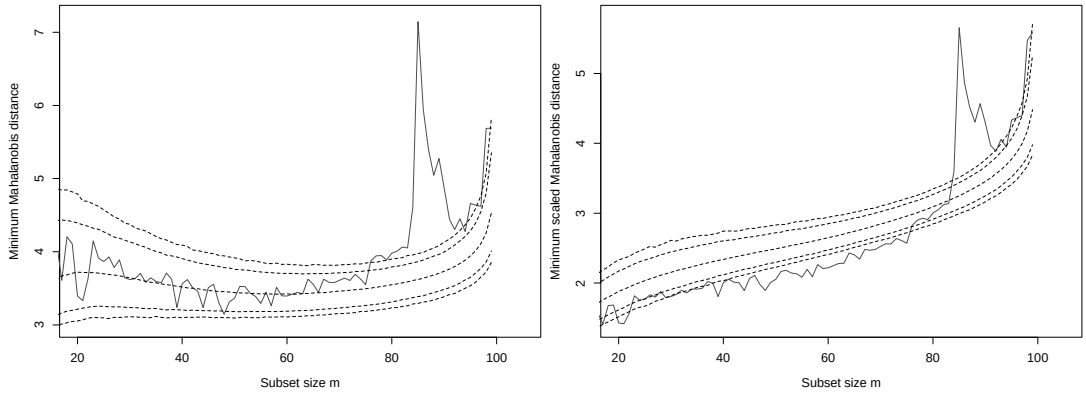

Figure 1: Swiss banknotes, forgeries (n = 100): forward plot of minimum Mahalanobis distance with superimposed 1, 5, 95, and 99% bootstrap envelopes using 10000 simulations. Left panel unscaled distances, right panel scaled distances. There is a clear indication of the presence of outliers which starts around $m = 84$.

are shrunken. This plot shows that scaled distances may yield a procedure with low power if several outliers are present. However, we avoid extensive simulations by first finding theoretical envelopes for scaled distances and then converting them to those that are unscaled. Ease of computation is of particular importance if we have to superimpose a series of envelopes for different subsample sizes.

4 Envelopes from Order Statistics

4.1 Scaled Distances

We now use order statistics to find good, fast approximations to our bootstrap envelopes. For the moment we take μ and Σ as known, so our results apply to both scaled and unscaled distances. The test statistic (4) is the $m+1$st ordered value of the n Mahalanobis distances. We can therefore use distributional results to obtain approximate envelopes for our plots. Since these envelopes do not require simulation in their calculation, we can use them for much more extreme points of the distribution than would be possible for bootstrap intervals without massive simulations.

Let $Y_{[m+1]}$ be the $(m+1)$st order statistic from a sample of size n from a univariate distribution with c.d.f. $G(y)$. Then the c.d.f of $Y_{[m+1]}$ is given exactly by

$$P\{Y_{[m+1]} \leq y\} \quad = \quad \sum_{j=m+1}^{n} \left(\begin{array}{c} n \\ j \end{array} \right) \{G(y)\}^j \{1-G(y)\}^{n-j}. \tag{5}$$

See, for example, Lehmann (1991, p. 353). Further, it is well known that we can apply properties of the beta distribution to the RHS of (5) to obtain

$$P\{Y_{[m+1]} \leq y\} = I_{G(y)}(m+1, n-m), \tag{6}$$

where

$$I_p(A,B) = \int_0^p \frac{1}{\alpha(A,B)} u^{A-1}(1-u)^{B-1} du$$

is the incomplete beta integral. From the relationship between the F and the beta distribution it is possible to rewrite equation (6) as

$$P\{Y_{[m+1]} \leq y\} = P\left\{F_{2(n-m),2(m+1)} > \frac{1-G(y)}{G(y)}\frac{m+1}{n-m}\right\} \tag{7}$$

where $F_{2(n-m),2(m+1)}$ is the F distribution with $2(n-m)$ and $2(m+1)$ degrees of freedom (Guenther 1977). Thus, the required quantile of order γ of the distribution of $Y_{[m+1]}$ say $y_{m+1,n;\gamma}$ can be obtained as

$$y_{m+1,n;\gamma} = G^{-1}\left(\frac{m+1}{m+1+(n-m)x_{2(n-m),2(m+1);1-\gamma}}\right) \tag{8}$$

where $x_{2(n-m),2(m+1);1-\gamma}$ is the quantile of order $1-\gamma$ of the F distribution with $2(n-m)$ and $2(m+1)$ degrees of freedom. The argument of $G^{-1}(.)$ in (8) becomes extremely close to one at the end of the search, that is as $m \to n$, particularly for large n and extreme γ. Consequently, care needs to be taken to ensure that the numerical calculation of this inverse distribution is sufficiently accurate. Details of one case are in §10.3

We now consider the choice of $G(x)$. If we knew both μ and Σ, $G(x)$ would be χ^2_v. When both μ and Σ are estimated using maximum likelihood on the whole sample, the squared distances have a scaled beta distribution. But, in our case, we estimate from a subsample of m observations that do not include the observation being tested. Atkinson, Riani, and Cerioli (2004, p. 43-4) derive distributional results for such deletion Mahalanobis distances. In the present case we estimate Σ on $m-1$ degrees of freedom. If the estimate of Σ were unbiased the null distribution of this squared distance would be

$$d^2_{(i)} \sim \frac{n}{(n-1)}\frac{v(m-1)}{(m-v)}F_{v,m-v}. \tag{9}$$

The superiority of the F-approximation is shown in Figure 2 for the case $n = 100$ and $v = 6$, values for which asymptotic arguments are unlikely to hold. The left-hand panel of the figure shows that the χ^2 approximation is poor, the envelopes being systematically too low.

Unfortunately, the estimate of Σ that we use is biased since it is calculated from the m observations in the subset that have been chosen as having the m smallest distances. However, in the calculation of the scaled distances (3) we approximately correct for this effect by multiplication by a ratio derived from estimates of Σ. So the envelopes for the scaled Mahalanobis distances are given by

$$V_{m,\gamma} = \sqrt{\frac{n}{(n-1)}}\sqrt{\frac{v(m-1)}{(m-v)}}\sqrt{y_{m+1,n;\gamma}}. \tag{10}$$

4.2 Approximations for Unscaled Distances

Unscaled distances cannot take advantage of the beneficial cancellation of bias provided by the ratio $|\hat{\Sigma}(m)|/|\hat{\Sigma}(n)|$ in (3). However, an approximate correction factor for the envelopes of unscaled squared Mahalanobis distances (2) can be obtained from results on elliptical truncation in the multivariate normal distribution. Suppose that $y_i \sim N(\mu, \Sigma)$ is restricted to lie in the subspace

$$0 \leq (y_i - \mu)^T\Sigma^{-1}(y_i - \mu) \leq b(m), \tag{11}$$

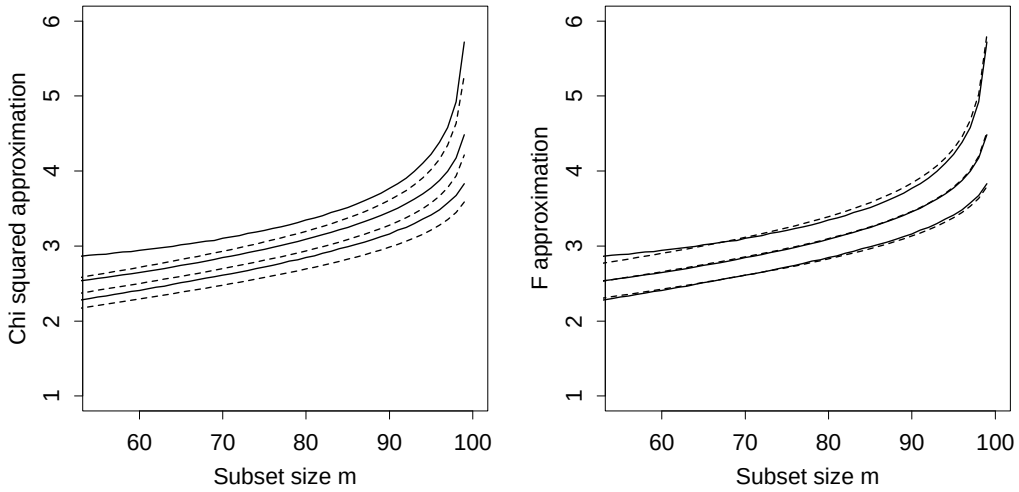

Figure 2: Comparison of 1%, 50% and 99% asymptotic envelopes for scaled distances: $n = 100$, $v = 6$. Left-hand panel: χ^2; right-hand panel: scaled F distribution. Continuous lines, envelopes found by simulation

where $b(m)$ is an arbitrary positive constant. Then it follows from the results of Tallis (1963) that

$$\mathrm{E}(y_i) = \mu \qquad \text{and} \qquad \mathrm{var}(y_i) = k(m)\Sigma,$$

where

$$k(m) = \frac{P\{\chi^2_{v+2} < b(m)\}}{P\{\chi^2_v < b(m)\}}.$$

Our estimate of Σ at step m is calculated from the m observations y_i that have been chosen as having the m smallest (squared) Mahalanobis distances. If we ignore the sampling variability in this truncation we can take $b(m)$ as the limiting value of the m-th order statistic in a sample of n squared Mahalanobis distances. Hence $c_{FS}(m) = k(m)^{-1}$ is the inflation factor for $\hat{\Sigma}(m)$ to achieve consistency at the normal model. In large samples

$$c_{FS}(m) = \frac{m/n}{P(\chi^2_{v+2} < X^2_{v,m/n})}, \tag{12}$$

where $X^2_{v,m/n}$ is the m/n quantile of χ^2_v. Our envelopes for unscaled distances are then obtained by scaling up the values of the order statistics

$$V^*_{m,\gamma} = c_{FS}(m) V_{m,\gamma}.$$

The bound $\sqrt{b(m)}$ in (11), viewed as a function of m, is sometimes called a radius for trimming size $(n-m)/n$. García-Escudero and Gordaliza (2005) studied the asymptotic behaviour of its empirical version when μ and Σ are replaced by consistent robust estimators, such as the MCD-based estimators of §7.2. There we take $m = h$, where h, defined in (14), is a carefully selected half of the data. Then $c_{FS}(m)$ is equal to the consistency factor (16) derived for the MCD scatter estimator by Butler, Davies, and Jhun (1993) and Croux and Haesbroeck (1999). A corollary of the results of Tallis, relating the truncated univariate normal distribution and χ^2_3 is given in Appendix 2.

4.3 Asymptotic Results for Very Large Samples

For very large n we use the asymptotic normality of order statistics to provide a satisfactory approximation to (5), once more for known μ and Σ. The asymptotic expectation of $Y_{[m+1]}$ is (Cox and Hinkley 1974, p.470) approximately

$$\xi_{m+1,n} = G^{-1}\{(m+1-3/8)/(n+1/4)\}.$$

If we let $p_\xi = (m+1-3/8)/(n+1/4)$ and $\xi_{m+1,n} = G^{-1}(p_\xi)$, the variance of $\xi_{m+1,n}$ (Stuart and Ord 1987, p.331) is

$$\sigma_\xi^2 = p_\xi(1-p_\xi)/\{nG^2(\xi_{m+1,n})\}.$$

Thus, replacing G with the scaled F distribution (9) yields the asymptotic $100\alpha\%$ point of the distribution of the scaled squared distance as

$$\xi_{m+1,n} + \sigma_\xi \Phi^{-1}(\alpha), \tag{13}$$

where $\Phi(z)$ is the c.d.f. of the standard normal distribution.

For scaled distances (13) replaces (10). To obtain approximations for the unscaled distance we again need to apply the results of §4.2.

4.4 A Comparison of Some Bootstrap and Order-Statistic Based Envelopes

We now present plots illustrating the quality of our order-statistic approximations to the envelopes.

The right-hand panels of Figure 3 show bootstrap envelopes (solid line) and the order-statistic approximation of §4.1 for scaled distances when $n = 200$ and $v = 5$ and 10. Agreement with the results of 10,000 simulations is very good virtually throughout the whole range of m. The plots in the left-hand panels of the figure are for unscaled distances using the results of §4.2. Although the approximation is not perfect, as we shall see the bounds are adequate for outlier detection where we look at the upper boundaries typically in the last one third of the search.

Figure 4 is a similar plot for $n = 600$. Here the approximations for the unscaled distances are improved compared with those in Figure 3: the effect of increased v is reduced and the agreement in the upper envelope extends at least to $n/2$.

5 The Forward Search for Outlier Detection

5.1 Motivation

If there are a few large outliers they will enter at the end of the search, and their detection is not a problem. As an instance Figure 5 shows an example with two appreciable outliers with mean shift 2.1 in a sample of 200 six-dimensional observations. At the end of the search there are two observations lying clearly outside the 99.9% envelope. Here the forward search has no difficulty in detecting these anomalous observations. The same is true for many other outlier detection procedures.

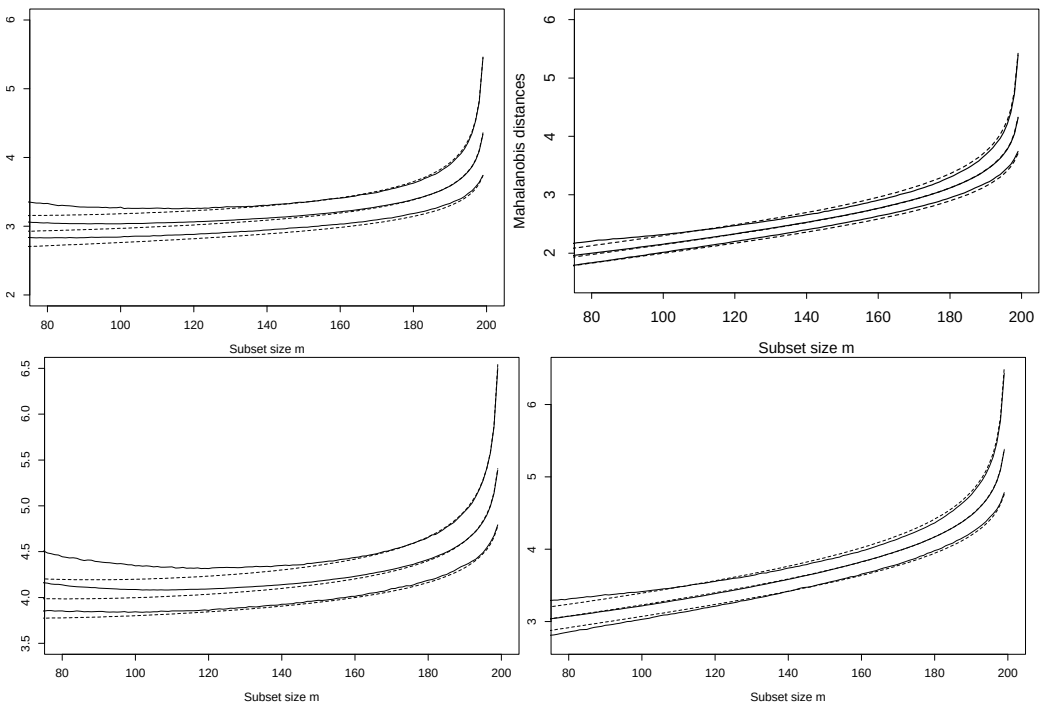

Figure 3: Agreement between bootstrap envelopes (solid line) and the order-statistic approximation of §4 when $n = 200$. Left-hand panels: unscaled distances, right-hand panels: scaled distances. Top panels: $v = 5$, bottom panels $v = 10$.

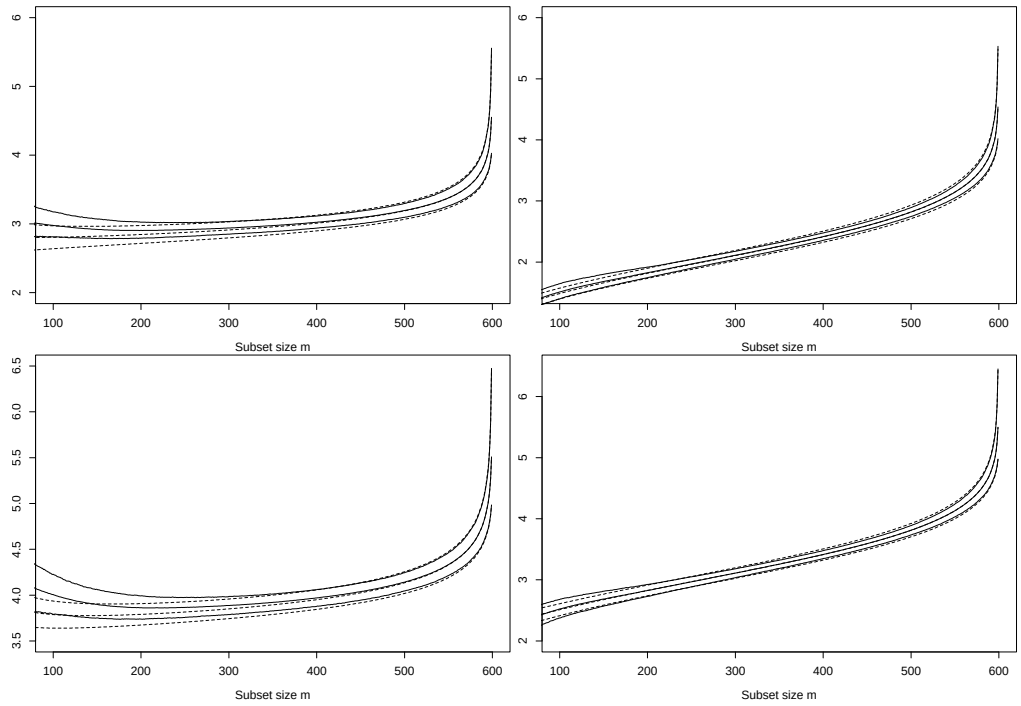

Figure 4: Agreement between bootstrap envelopes (solid line) and the order-statistic approximation of §4 when $n = 600$. Left-hand panels: unscaled distances, right-hand panels: scaled distances. Top panels: $v = 5$, bottom panels $v = 10$.

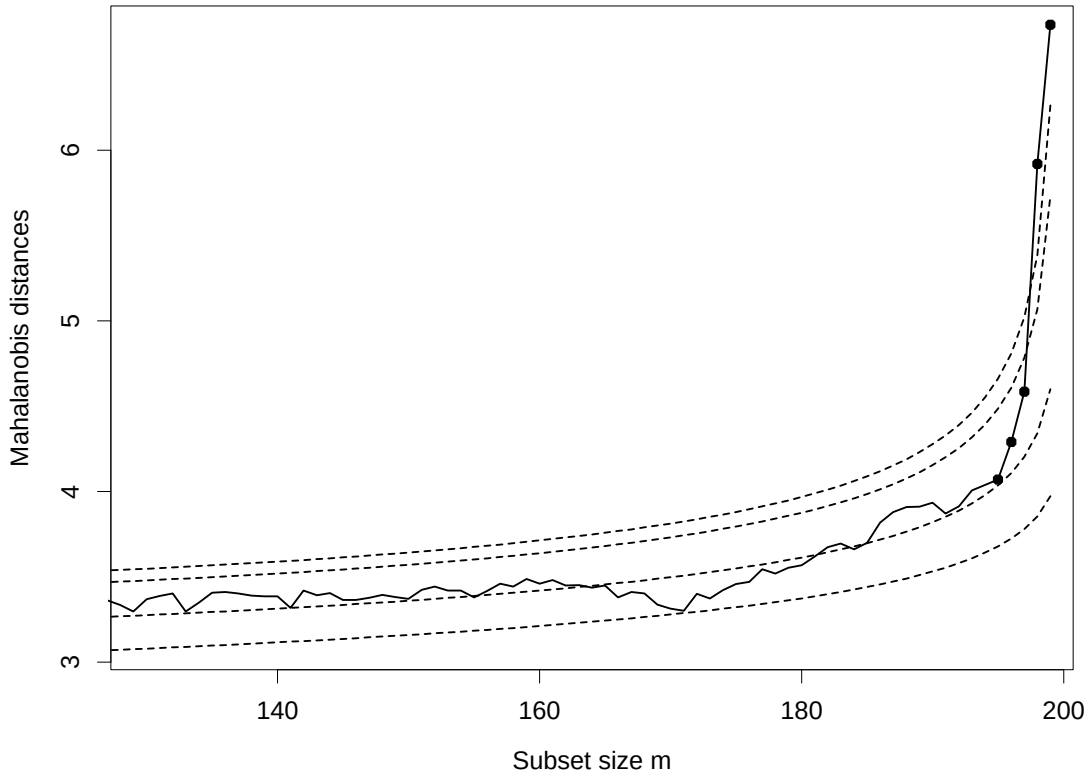

Figure 5: Easily detected outliers. There are two contaminated units in this sample with $n = 200$ and $v = 6$ that are clearly revealed at the end of the search; 1%, 50%, 99% and 99.9% envelopes

Even relatively small clusters of outliers can however be more difficult to identify. Figure 1 of the Swiss banknote data shows that the forward plot of distances twice goes outside the 99%envelope and twice returns within it. These events indicate two instances of masking. At the end of the search the Mahalanobis distance d_{min} for $m = n-1$ is 5.691 and lies just below the 99% envelope from bootstrap simulations for which the value is 5.834. Only if two observations are deleted do these few outliers become apparent: when $m = n - 3$ the 99% bootstrap value of d_{min} is 4.62; the observed value is 4.77, which lies well outside the envelope. However, there is also a cluster of outliers and the search has a central peak, around $m = 85$ in Figure 1, before a series of lower values of the distance. In more extreme cases with a cluster of outliers masking may cause the plot to return inside the envelopes at the end of the search. An example is in our second set of simulated data in Figure 10. Methods of using the forward search for the formal detection of outliers have to be sensitive to these two patterns - a few "obvious" outliers at the end and a peak earlier in the search caused by a cluster of outliers.

5.2 Procedure

To use the envelopes in the forward search for outlier detection we propose a two stage process. In the first stage we run a search on the data, monitoring the bounds for all n observations until we obtain a "signal" indicating that observation $m^{\dagger}$, and therefore succeeding observations, may be outliers, because it lies beyond our threshold. In the second part we superimpose envelopes for values of n from this point until the first time we intro-

duce an observation we recognise as an outlier.

The thresholds need to be chosen to avoid the problem of simultaneity. We require a procedure that combines high power with a size of α for declaring the sample to contain at least one outlier. In our exposition and examples we take $\alpha = 1\%$.

We can expect the occasional observation to fall outside the bounds during the search even if there are no outliers. If we ignore the correlation in adjacent distances induced by the ordering imposed by the search, each observation can be taken to have a probability $\gamma = 1 - \alpha$ of falling above the α point of the pointwise envelope. If γ is small, say 1%, and $n = 1,000$ the number of observations outside the envelope will have approximately a Poisson distribution with mean 10. The probability that no observations fall above the envelope will then be e^{-10}, a very small number. We need to be able to distinguish these random occurrences during the search from the important peaks illustrated in the two figures.

The envelopes shown in Figures 3 and 4 consist roughly of two parts; a flat "central" part and a steeply curving "final" part. Our procedure FS1 for the detection of a "signal" takes account of these two parts and is as follows:

- In the central part of the search we require 3 consecutive values of $d_{\text{min}}(m)$ above the 99.99% envelope or 1 above 99.999%;
- In the final part of the search we need two consecutive values of $d_{\text{min}}(m)$ above 99.9% and 1 above 99%;
- $d_{\text{min}}(n-2) > 99.9\%$ envelope;
- $d_{\text{min}}(n-1) > 99\%$ envelope.

The final part of the search is defined as:

$$m \geq n - \left[13\,(n/200)^{0.5}\right],$$

where here [] stands for rounded integer. For $n = 200$ the value is slightly greater than 6% of the observations.

The purpose of, in particular, the first point is to distinguish real peaks from random fluctuations. Once a signal takes place (at $m = m^{\dagger}$) we start superimposing 99% envelopes taking $n = m^{\dagger} - 1, m^{\dagger}, m^{\dagger} + 1, \ldots$ until the final, penultimate or antepenultimate value are above the 99% threshold or, alternatively, we have a value of $d_{\text{min}}(m)$ for any $m > m^{\dagger}$ which is greater than the 99.9% threshold.

Some slight variations of the former procedure are possible. Here are two. If we failed to detect any outliers in FS1 but had an incontrovertible signal:

- FS2. Three consecutive values of $d_{\text{min}}(m)$ above the 99.999% threshold, or
- FS3. Ten values of $d_{\text{min}}(m)$ above the 99.999% threshold,

we then decide that outliers are present.

Some features of this procedure may seem arbitrary. However, as we see in §7, there are likewise arbitrary decisions in the MCD based procedures in the definition of the subset of m observations that are used in the final calculation of Mahalanobis distances and in the reference distributions used for testing these distances.

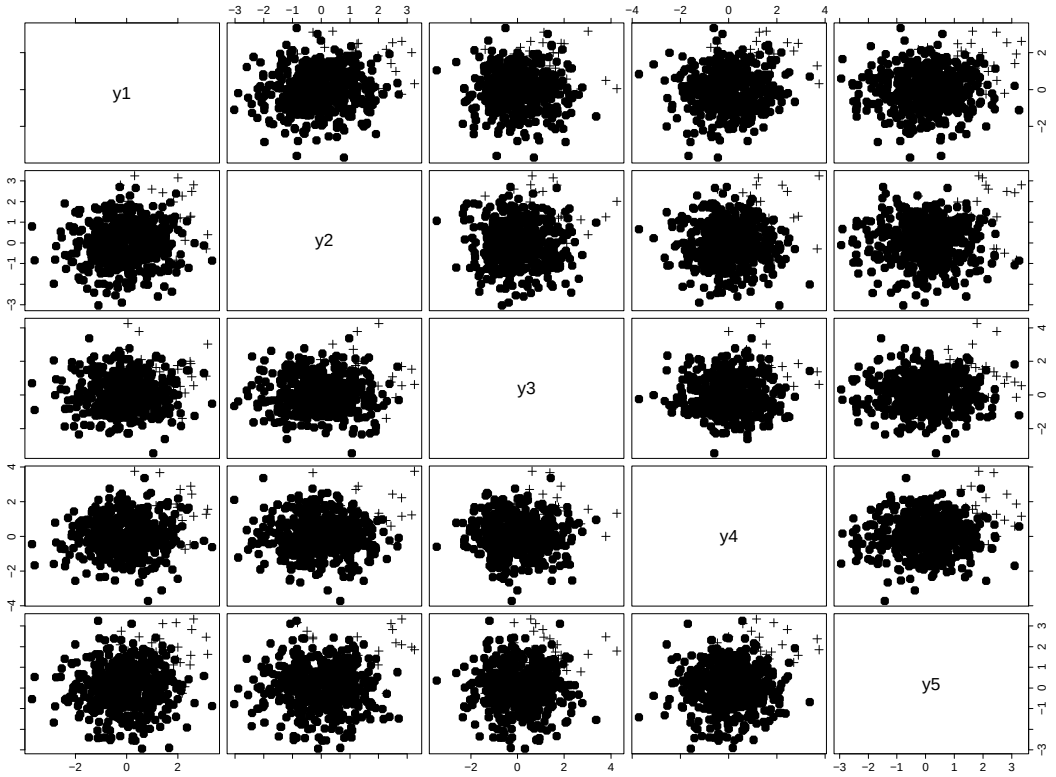

Figure 6: Slight contamination. Scatterplot matrix of data yielding Figure 7: $n = 500$, $v = 5$; 5% of the units are contaminated. Level shift = 1.4 for each dimension. Original units, •; contaminated units $+$

6 Two Examples

6.1 Slight Contamination

The purpose of this example is to show in practice how the procedure works in the presence of slight contamination and a small number of contaminated units.

There are 500 observations and $v = 5$. There is a shift contamination of 1.4 in all dimensions applied to 5% of the units, those numbered 1–25. The scatterplot matrix of the data is in Figure 6, with the forward plot of minimum Mahalanobis distances in Figure 7. This plot shows that there is a series of large values around $m = 480$, even though the value at the end of the search is below the 99% envelope. There is thus visual evidence of the presence of around 20 masked outliers.

More formally, we now apply our rule FS1 and find that a signal occurs when $m = 479$ because, for this value we have two consecutive values of $d_{\text{min}}(m)$ above the 99.9% threshold and, in addition, one other value above 99%. In particular the threshold levels are:

$$d_{\text{min}}(479) > 99.9\%, \quad d_{\text{min}}(480) > 99.9\% \quad \text{with } d_{\text{min}}(478) > 99\%.$$

We receive the signal at $m = 479$ because this is the first point at which we have an observation above the 99.9% threshold.

We now proceed to the second part of our outlier detection process and superimpose envelopes for a series of increasing sample sizes until we identify the outliers signalled

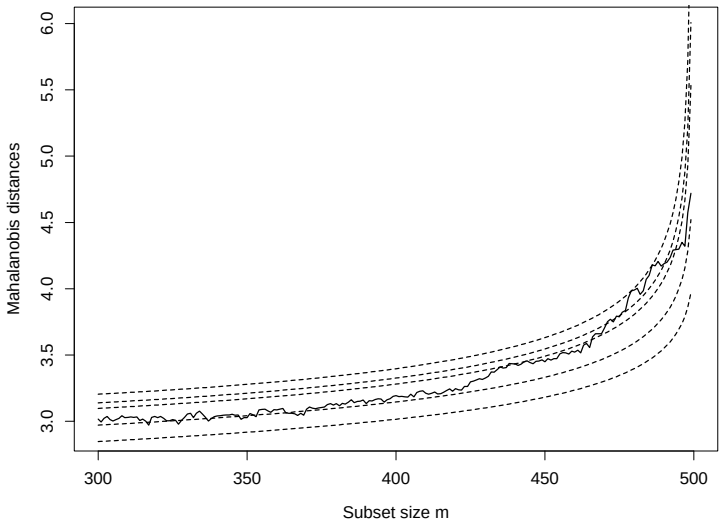

Figure 7: Slight contamination. Forward plot of minimum Mahalanobis distances for the data of Figure 6. The first 3 bands are 1%, 50% and 99%. The highest 2 are 99.9% and 99.999%. The extreme envelope has been superimposed just to show that this sample contains a strong evidence of not being homogeneous.

in the first stage of the process. In this example we start with $n = 479$. Figure 8 shows the envelopes and forward plot of minimum Mahalanobis distances for several values of n. When $n = 482$ the curve lies well within the envelopes. Around $n = 490$ the observed curve starts to become closer to the 99% envelope. When $n = 494$ some values are close to the 99.9% envelope. The first time the observed values go out of the 99.9% envelope is when $n = 495$.

The procedure of resuperimposing envelopes stops when $n = 495$, the first time in which we have a value of $d_{\text{min}}(m)$ for $m \leq m^{\dagger}$ greater than the 99.9% threshold. The group can therefore be considered as homogeneous up to when we include 494 units. In these data the shifted observations are units 1 - 25. The last six units included in the search plotted in Figure 7 are numbered 2, 343, 6, 16, 23, 1, so that five out of these six are indeed contaminated units.

6.2 Appreciable Contamination

We now consider an example with the same structure but with an appreciable number of contaminated units, although the mean shift itself is not large. We take $n = 200$ and $v = 5$, with 30% contamination from a mean shift of 1.2 in each dimension. The original observations again have a standard independent multivariate normal distribution. The scatterplot matrix of the data is in Figure 9 with the forward plot of minimum Mahalanobis distances in Figure 10.

With 30% contamination and 200 observations there are 60 outliers. Figure 10 shows a peak around $m = 130$ followed by a trough a little after $m = 140$, which therefore come roughly where we would expect. The peak is a little early because of the overlapping nature of the two groups. The trough is caused by the effect of the inclusion of outliers on the estimate of Σ which becomes too large, giving small Mahalanobis distances.

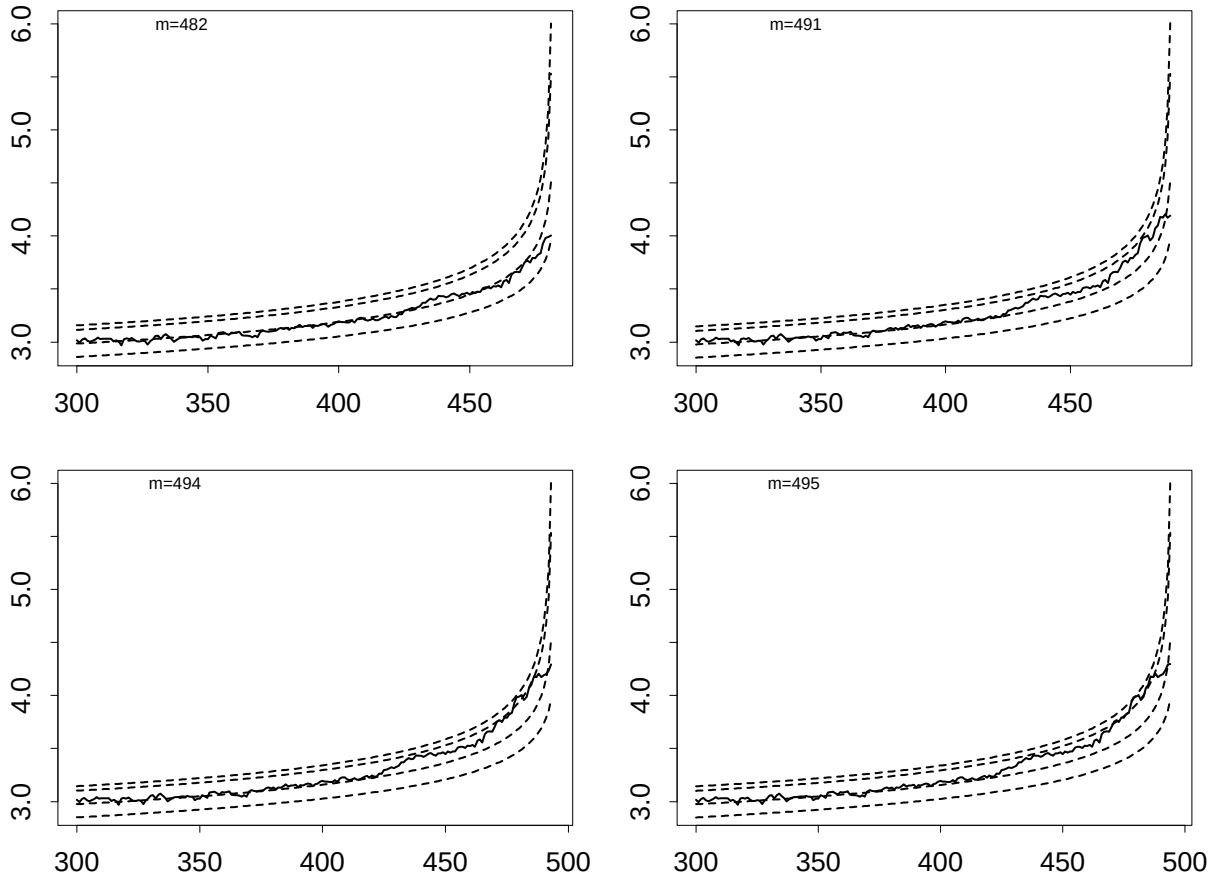

Figure 8: Slightly contaminated data. When $n = 482$ the curve lies well within the envelopes. Around $n = 490$ the observed curve starts getting closer to the 99% envelope and when $n = 494$ some values are close to the 99.9% envelope. The first time the curve goes above the 99.9% envelope is step $n = 495$.

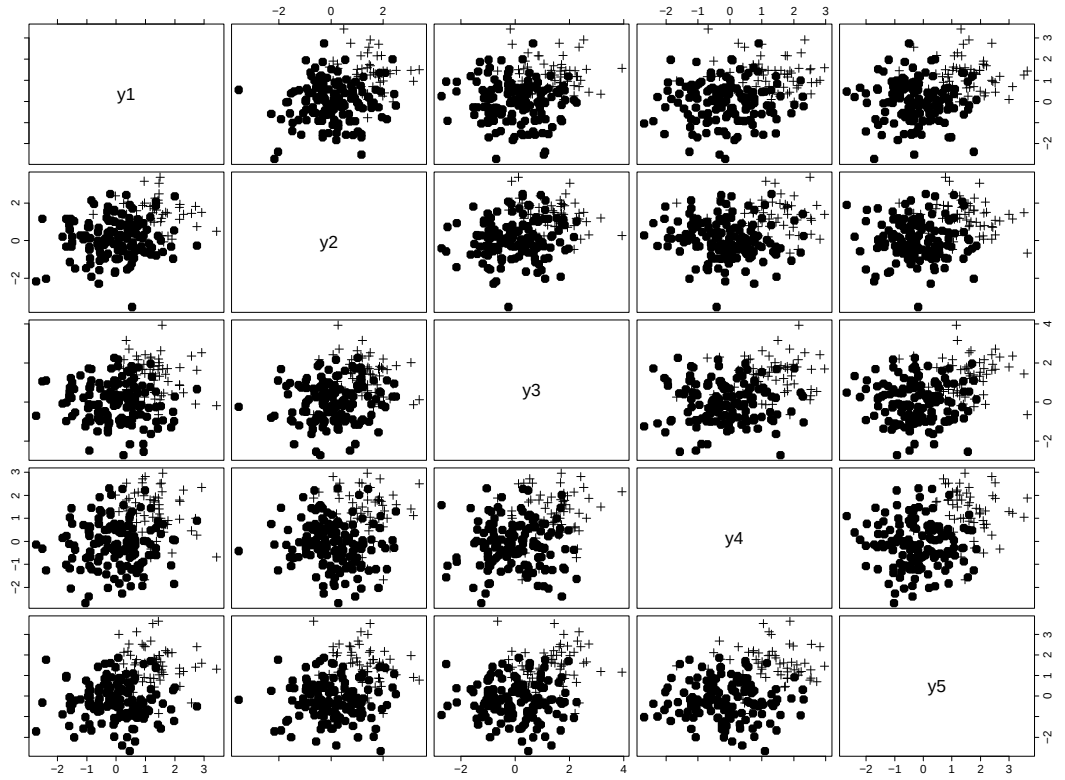

Figure 9: Scatterplot matrix of appreciably contaminated data with $n = 200$ and $v = 5$ yielding Figure 10: 30% of the units are contaminated. Level shift =1.2 for each dimension. Original units, •; contaminated units $+$

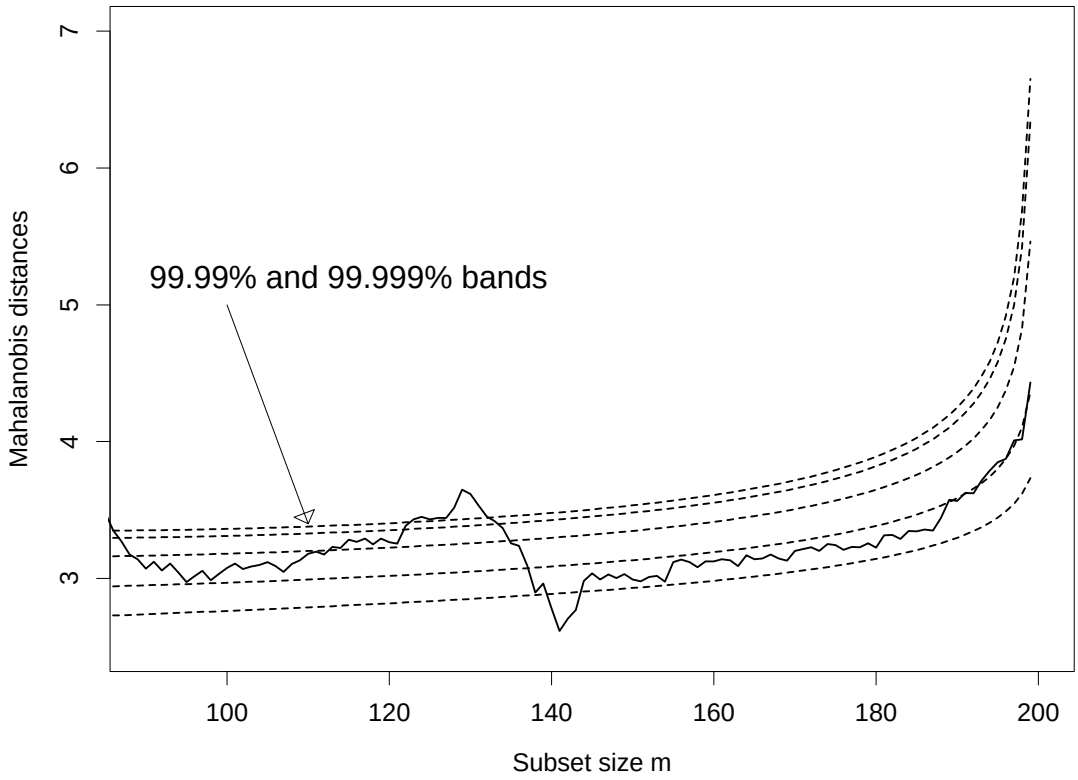

Figure 10: Appreciably contaminated data; minimum Mahalanobis distance with superimposed 1%, 50%, 99%, 99.99% and 99.999% confidence bands.

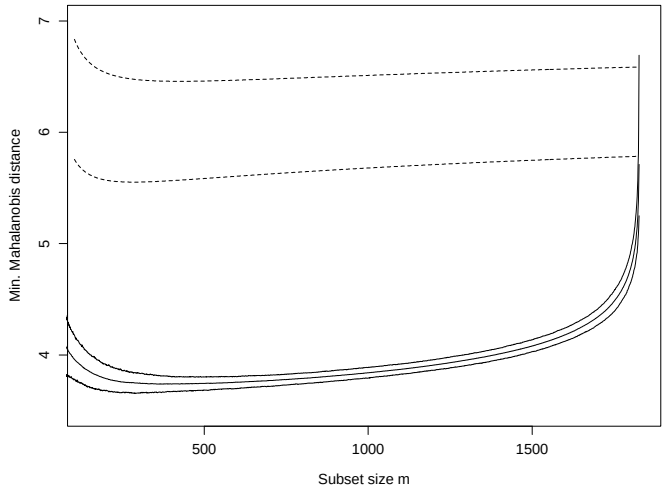

Figure 11: 2,000 normal observations, $v = 10$: forward plot of 90% and 99% envelopes of minimum Mahalanobis distances with superimposed Bonferroni bounds including Hadi's correction

If we apply our rules to this plot we find that, from $m = 122$ to $m = 133$ the consecutive values of $d_{\text{min}}(m)$ are greater than the 99.99% envelope while from $m = 123$ to $m = 132$ they are all greater than the 99.999% envelope. FS2 is therefore satisfied and we do not need to confirm the outliers by successively superimposing bounds. The figure shows how masking will cause the failure of procedures that look at only the largest values of the distances, or that try to detect outliers by backwards deletion. The structure of a peak, followed by a dip, in the plot of Figure 10 is further evidence of the presence of a cluster of outliers that can only be obtained from the forward search. However we do not here make use of this as a procedure for detecting outliers, concentrating instead solely on upper exceedances of the bounds.

7 Other Outlier Detection Procedures

7.1 Bonferroni Bounds

The statistic (2) provides the basis for our test of the outlyingness of observation $[m + 1]$. Hadi (1994) uses a Bonferroni bound to allow for the ordering of the distances during his forward search and compares a slightly scaled version of (2) with the percentage points of $\chi^2_{v,(\alpha/n)}$, the scaling being to allow for the estimation of Σ.

Since the test is for an outlier in a sample of size $m + 1$, it seems appropriate to use the Bonferroni bound $\chi^2_{v,\{\alpha/(m+1)\}}$ rather than $\chi^2_{v,(\alpha/n)}$. Figure 11 shows the resulting 95 and 99% bounds superimposed on a forward plot of bootstrap envelopes for $n = 2000$ and $v = 10$. These bounds were calculated using the empirical scaling in §2 of Hadi (1994). They are unrelated to the true distribution, except for the last step of the search; due to the low correlation of the distances the bound is almost exact when $m = n - 1$. Earlier in the search the bounds are far too large, because $\hat{\Sigma}(m)$, despite Hadi's rescaling, is treated as an estimate from a full sample, rather than from the truncated sample that arises from the ordering of the distances.

Wisnowski et al. (2001, p. 360) report that the related procedure of Hadi and Simonoff

(1993) for regression has a low detection rate for moderate and small outliers and an abnormally low false alarm rate. Similar properties for multivariate data can be inferred from Figure 11.

7.2 Distances for Outlier Detection

In this section we describe a number of variants of the Mahalanobis distance that have been recommended for outlier detection. These vary in the subset or subsets of observations used for parameter estimation. When robust estimates are used, there are several possible adjustments to obtain consistent estimators of Σ. There is also a choice of reference distribution against which to assess the observed distances. We leave until §8 the adjustments made for simultaneous inference which introduce further subsets of the data to be used for estimation.

• **MD and MDK.**

The Mahalanobis distance (1), with parameters estimated from all the data was long suggested as an outlier test, for example by Wilks (1963). As is well known, it is exceptionally sensitive to masking. However, we include it in some of our comparisons to illustrate just how sensitive it is.

If the values of the parameters μ and Σ were known, the distribution of the distance would be χ^2_v. As an outlier test we call this **MDK** with **MD** the test based on the same distances but referred to the correct scaled Beta distribution. Section 2.6 of Atkinson et al. (2004) gives this distribution; §2.16 gives references to the repeated rediscovery of related distributional results.

Robust distances. The customary way to detect multivariate outliers is to compute robust estimates of μ and Σ based on one or two carefully chosen subsets of the data (Rousseeuw and van Zomeren 1990). Mahalanobis distances from this robust fit are then compared with the $\alpha\%$ cut-off value of the reference distribution, with α usually between 0.01 and 0.05, and unit i is nominated as an outlier if its distance exceeds the threshold. The distribution of squared Mahalanobis distances depends on the robust estimators at hand, but it has been proven that asymptotically it is either exactly or proportional to χ^2_v in many situations; see e.g. Davies (1992), Butler, Davies, and Jhun (1993), Lopuhaä (1999) and Maronna, Martin, and Yohai (2006). We list four robust distances, versions of which are used in our comparisons.

• **MCD.**

We consider the minimum covariance determinant (MCD) estimator described in Rousseeuw and Leroy (1987, p. 262) and some of its variants. In the MCD approach, the estimators of Σ and μ, say $\hat{\mu}_{\rm MCD}$ and $\hat{\Sigma}_{\rm MCD}$, are defined to be the mean and the covariance matrix of the subset of

$$h = \lfloor \frac{n+v+1}{2} \rfloor \tag{14}$$

observations for which the determinant of the covariance matrix is minimal, where $\lfloor \cdot \rfloor$ denotes the integer part. The resulting breakdown value is then

$$\frac{\lfloor (n-v+1)/2 \rfloor}{n}. \tag{15}$$

The MCD is used because it has rate of convergence $n^{-1/2}$, unlike the minimum volume ellipsoid estimator (Davies 1992) for which convergence is at rate $n^{-1/3}$.

Rousseeuw and Van Driessen (1999) developed a fast algorithm for computing $\hat{\mu}_{\mathrm{MCD}}$ and $\hat{\Sigma}_{\mathrm{MCD}}$, which has been implemented in different languages, including `R`, `S-Plus`, `Fortran` and `Matlab`. Software availability and faster rate of convergence with respect to other high breakdown estimators have made the MCD approach a popular choice in applied robust statistics.

A crucial issue with the MCD scatter estimator $\hat{\Sigma}_{\mathrm{MCD}}$ is that it tends to underestimate Σ even in large samples. With breakdown value (15), the appropriate large-sample correction factor for $\hat{\Sigma}_{\mathrm{MCD}}$ was derived by Butler, Davies, and Jhun (1993) and by Croux and Haesbroeck (1999) as

$$c_{\mathrm{MCD}}(h,n,v) = \frac{h/n}{P(\chi^2_{v+2} < X^2_{v,h/n})}. \tag{16}$$

However, although consistent at the normal model, the estimator

$$c_{\mathrm{MCD}}(h,n,v)\hat{\Sigma}_{\mathrm{MCD}}$$

is still biased for small sample sizes. Pison, Van Aelst, and Willems (2002) showed by Monte-Carlo simulation the importance of applying a small sample correction factor to $c_{\mathrm{MCD}}(h,n,v)\hat{\Sigma}_{\mathrm{MCD}}$. Let $s_{\mathrm{MCD}}(h,n,v)$ be this factor for a specific choice of n and v and breakdown value (15). The resulting robust Mahalanobis distances are then

$$d_{(\mathrm{MCD})i} = \sqrt{k_{\mathrm{MCD}}(y_i - \hat{\mu}_{\mathrm{MCD}})^T \hat{\Sigma}^{-1}_{\mathrm{MCD}}(y_i - \hat{\mu}_{\mathrm{MCD}})} \qquad i = 1,\ldots,n, \tag{17}$$

where $k_{\mathrm{MCD}} = \{c_{\mathrm{MCD}}(h,n,v)s_{\mathrm{MCD}}(h,n,v)\}^{-1}$.

• **HR.**

The exact finite-sample distribution of the robust Mahalanobis distances (17) is unknown, but Hardin and Rocke (2005) proposed a scaled F approximation which, in small and moderate samples, outperforms the asymptotic χ^2_v approximation of MCD.

• **RMCD-C.**

To increase efficiency, a reweighted version of the MCD estimators is often used in practice. These reweighted estimators, $\hat{\mu}_{\mathrm{RMCD}}$ and $\hat{\Sigma}_{\mathrm{RMCD}}$, are computed by giving weight 0 to observations for which $d_{(\mathrm{MCD})i}$ exceeds a cutoff value. Thus a first subset of h observations is used to select a second subset from which the parameters are estimated. The default choice (Rousseeuw and Leroy 1987, Rousseeuw and Van Driessen 1999) for this cutoff value is

$$\sqrt{X^2_{v,0.025}}. \tag{18}$$

Both the consistency (Croux and Haesbroeck 2000) and the small sample (Pison, Van Aelst, and Willems 2002) correction factors $c_{\mathrm{RMCD}}(h,n,v)$ and $s_{\mathrm{RMCD}}(h,n,v)$ can be applied to $\hat{\Sigma}_{\mathrm{RMCD}}$, when the robust Mahalanobis distances become

$$d_{(\mathrm{RMCD-C})i} = \sqrt{k_{\mathrm{RMCD-C}}(y_i - \hat{\mu}_{\mathrm{RMCD}})^T \hat{\Sigma}^{-1}_{\mathrm{RMCD}}(y_i - \hat{\mu}_{\mathrm{RMCD}})} \qquad i = 1,\ldots,n, \tag{19}$$

where $k_{\mathrm{RMCD-C}} = \{c_{\mathrm{RMCD}}(h,n,v)s_{\mathrm{RMCD}}(h,n,v)\}^{-1}$.

RMCD.

The original MCD literature (Rousseeuw and Leroy 1987, Rousseeuw and Van Driessen 1999) did not suggest use of the consistency correction factor $c_{\mathrm{RMCD}}(h,n,v)$. The robust Mahalanobis distances arising from this basic reweighted MCD estimator, $d_{(\mathrm{RMCD})i}$, are then computed as in equation (19), but with $k_{\mathrm{RMCD}} = s_{\mathrm{RMCD}}(h,n,v)^{-1}$ replacing $k_{\mathrm{RMCD-C}}$.

8 Simultaneity and Bonferronisation

The published literature describing the properties of robust Mahalanobis distances for multivariate outlier detection is mainly concerned with rejection of the single null hypothesis

$$H_0 : y_i \sim N(\mu, \Sigma) \tag{20}$$

at level α. On the contrary, in our procedure of §4 the test statistic (4) is the $m+1$st ordered value of the n Mahalanobis distances. Therefore, its distribution involves the joint distribution of all the n Mahalanobis distances $d_i^2(m)$, so that the null hypothesis of interest becomes the intersection hypothesis

$$H_0 : \{y_1 \sim N(\mu, \Sigma)\} \cap \{y_2 \sim N(\mu, \Sigma)\} \cap \ldots \cap \{y_n \sim N(\mu, \Sigma)\} \tag{21}$$

that there are no outliers in the data. The Forward Search α is the size of the test of (21), i.e. the probability that at least one of the individual hypotheses (20) is rejected for some m when (21) is true. In our approach, we are willing to tolerate a wrong conclusion in $(100\alpha)\%$ of *data sets* without outliers, while under (20) one should be prepared to declare $(100\alpha)\%$ of *observations* as outliers in any application.

We let α have the same interpretation in MCD procedures by comparing all the individual statistics $d_{(\mathrm{MCD})i}$, $d_{(\mathrm{RMCD})i}$ and $d_{(\mathrm{RMCD-C})i}$, $i = 1, \ldots, n$, with the $\alpha^* = \alpha/n$ cutoff value of their reference distributions. A Bonferroni approach is appropriate in this context because extreme observations are approximately independent of the MCD estimators $\hat{\mu}_{\mathrm{MCD}}$ and $\hat{\Sigma}_{\mathrm{MCD}}$, as shown by Hardin and Rocke (2005). Hence the intersection between multiple tests of (20), sharing the same MCD estimates, should be negligible, at least when H_0 is rejected. Gather, Pawlitschko, and Pigeot (1997) discuss properties of multiple tests and provide further references.

This Bonferroni procedure applies to the level at which we say that at least one outlier is present. We can, in addition, apply the Bonferroni argument to selection of observations to be used in parameter estimation for the reweighted distances. We suggest two such modifications.

• **RMCD-B.**

We set $\alpha = 0.01$ in all our simulations. The default cutoff value for excluding observations in the computation of reweighted MCD estimators is given by (18). However, this cutoff is inappropriate when testing the intersection hypothesis (21), as individual outlier tests are now performed with size $\alpha^* = 0.01/n$. We accordingly calculate a modified version of the reweighted estimators, say $\hat{\mu}_{\mathrm{RMCD-B}}$ and $\hat{\Sigma}_{\mathrm{RMCD-B}}$, where observations are given weight 0 if $d_{(\mathrm{MCD})i}$ exceeds

$$\sqrt{X^2_{v,\alpha^*}}. \tag{22}$$

Substituting these modified estimators into (19), we obtain the Bonferroni-adjusted reweighted distances

$$d_{(\mathrm{RMCD-B})i} = \sqrt{k_{\mathrm{RMCD}}(y_i - \hat{\mu}_{\mathrm{RMCD-B}})' \hat{\Sigma}^{-1}_{\mathrm{RMCD-B}} (y_i - \hat{\mu}_{\mathrm{RMCD-B}})} \qquad i = 1, \ldots, n, \tag{23}$$

• **RMCD-D.**

An alternative Bonferroni-adjusted reweighted-MCD distance is obtained by substituting $k_{\mathrm{RMCD-C}}$ for k_{RMCD} in equation (23), thus including the consistency factor as we did in the definition of RMCD-C.

The correction factors in these Bonferroni-adjusted versions of RMCD include the small sample correction $s_{\mathrm{RMCD}}(h, n, v)$ which was derived without allowance for simultaneous inference. The appropriate small-sample factor for RMCD-B and RMCD-D is not available in the MCD literature.

A summary of the Mahalanobis distance outlier tests considered in our simulations is given in Table 1.

Table 1: Mahalanobis distance outlier tests to be compared with the Forward Search

Acronym	Description
MDK	Squared non-robust distances d_i^2 asymptotic χ_v^2 distribution
MD	Squared non-robust distances d_i^2 Exact scaled Beta distribution
MCD	Squared MCD distances $d^2_{(\mathrm{MCD})i}$ asymptotic χ_v^2 distribution
RMCD	Squared reweighted-MCD distances $d^2_{(\mathrm{RMCD})i}$ asymptotic χ_v^2 distribution
RMCD-C	Squared reweighted-MCD distances with consistency correction $d^2_{(\mathrm{RMCD-C})i}$ asymptotic χ_v^2 distribution
RMCD-B	Squared Bonferroni-adjusted reweighted-MCD distances $d^2_{(\mathrm{RMCD-B})i}$ asymptotic χ_v^2 distribution
RMCD-D	Squared Bonferroni-adjusted reweighted-MCD distances with consistency correction $d^2_{(\mathrm{RMCD-D})i}$ asymptotic χ_v^2 distribution
HR	Squared MCD distances $d^2_{(\mathrm{MCD})i}$ scaled F distribution of Hardin and Rocke (2005)

9 Size and Power

9.1 Size

To compare the performance of the various outlier tests we need them to have at least approximately the same size. To establish the size we performed each nominal 1% test on 10,000 sets of simulated multivariate normal data for four values of n from 100 to 1,000 and with dimension $v = 5$ and 10. The result was considered significant if at least one outlier was detected.

We summarise our findings in Table 2. For the first eight tests, based on various Mahalanobis distances, we use the Bonferroni correction to obtain a test with nominal size of 1%. The first entry in the table is for the standard Mahalanobis distance with reference

Table 2: Size of the nominal 1% test based on 10,000 simulations ($v = 5$ first entry and $v = 10$ second entry in each cell): classical Mahalanobis distances, the six MCD-based procedures of Table 1 and our three proposals

	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
MDK	0.28%	0.42%	0.70%	0.79%
	0.06%	0.44%	0.52%	0.89%
MD	1.12%	0.97%	0.97%	0.89%
	1.04%	1.21%	0.99%	1.19%
MCD	62.43%	32.91%	8.81%	3.71%
	88.59%	49.21%	11.76%	4.72%
RMCD	30.04%	10.95%	3.78%	3.02%
	61.78%	16.37%	5.15%	3.64%
RMCD-C	10.13%	3.39%	1.70%	1.16%
	32.25%	6.04%	2.15%	1.77%
RMCD-B	4.94%	1.94%	1.16%	1.03%
	12.45%	3.33%	1.61%	1.40%
RMCD-D	3.41%	1.64%	1.09%	1.01%
	8.11%	2.90%	1.51%	1.36%
HR	2.41%	2.53%	1.17%	0.97%
	5.28%	2.34%	1.09%	1.17%
FS1	1.02%	1.14%	1.13%	1.15%
	1.16%	1.26%	1.15%	1.19%
FS2	1.03%	1.15%	1.14%	1.15%
	1.53%	1.27%	1.17%	1.20%
FS3	1.04%	1.16%	1.15%	1.16%
	1.54%	1.31%	1.18%	1.20%

values from asymptotic χ^2 distribution that ignores the effect of estimating the parameters. The results are surprisingly bad: for $n = 100$ and $v = 10$ the size is 0.06% rather than 1%. Even when $n = 1,000$, a value by which asymptotics are usually expected to be a good guide, the size is only 0.79% when $v = 5$. There is a sharp contrast with the results using the correct Beta distribution, when the sizes correctly fluctuate between 0.89 and 1.21%. These results provide a measure of the fluctuation to be found in our simulation results. They also confirm that our Bonferroni correction does indeed provide a test with power close to 1%. Despite the correct size of the test, our simulations in §9.2 quantify what is well known in general, that the standard Mahalanobis distance can have very low power when used as an outlier test.

The next two sets of results are for the MCD and the RMCD. These results, especially for $n = 100$ are exceptionally bad, with sizes of up to 89%, clearly rendering the test unusable for 'small' samples of 100. As n increases, the asymptotically based correction factor improves the size. But even when $n = 1,000$, the sizes are between 3 and 5%. In view of this performance, we do not need to consider these tests any further.

The following four tests are versions of the MCD but with better size that improves as we go down the table. For RMCD-C, that is reweighted MCD with a consistency correction in the reweighting, the size is around 10% when $n = 100$ and $v = 5$. When $v = 10$ it rises to over 32%. For this and the other three reweighted MCD rules the size decreases with n, being close to the hoped-for value when $n = 500$. In RMCD-B we extend RMCD-C by including Bonferroni reweighting to obtain sizes around 5% when $n = 100$ and $v = 5$; for $v = 10$ the value is 12.5%. The version of RMCD-B with consistency correction, which we call RMCD-D, has sizes of 3.4% and 8.1% when $n = 100$, with all sizes less than those for RMCD-B. The sizes for HR when $n = 100$ are also too large, although throughout the table this test has values amongst the best for all values of n. The three versions of the forward search have satisfactory sizes for all values of n in the range studied, although the values are slightly above 1%.

As a result of this preliminary exploration we decided to focus our investigation on the properties of four outlier detection procedures: MD, RMCD-B, HR and FS3. Other procedure are sometimes included if some particular property is thereby revealed.

9.2 Power

To begin our comparisons of power, Table 3 shows the results of 10,000 simulations of samples with $n = 200$, $v = 5$ and with 5% of shifted observations, for a shift in all dimensions from 1 to 2.4; the first line of the table, in which the shift is zero, serves as a reminder of the size. The results are the percentage of samples in which at least one outlier was detected. In this table we have included all three of the Forward Search rules of §5.2. For this example there is nothing to choose between these three and, in the rest of the paper, we only give results for FS3.

The general conclusion from this table is that the FS rules behave slightly better than RMCD-B, which has a larger size, but lower power for level shifts above 1.6. The HR rule behaves less well than these, with MD by far the worst, despite its excellent size.

Table 4 repeats the results of Table 3 but with 30% of the observations shifted. The broad conclusions from the two tables are similar, but more extreme for the more contaminated data. The best rule is FS3. Unlike HR, RMCD-B loses appreciable power in this more heavily contaminated setting. Most sensationally of all, masking is so strong that MD

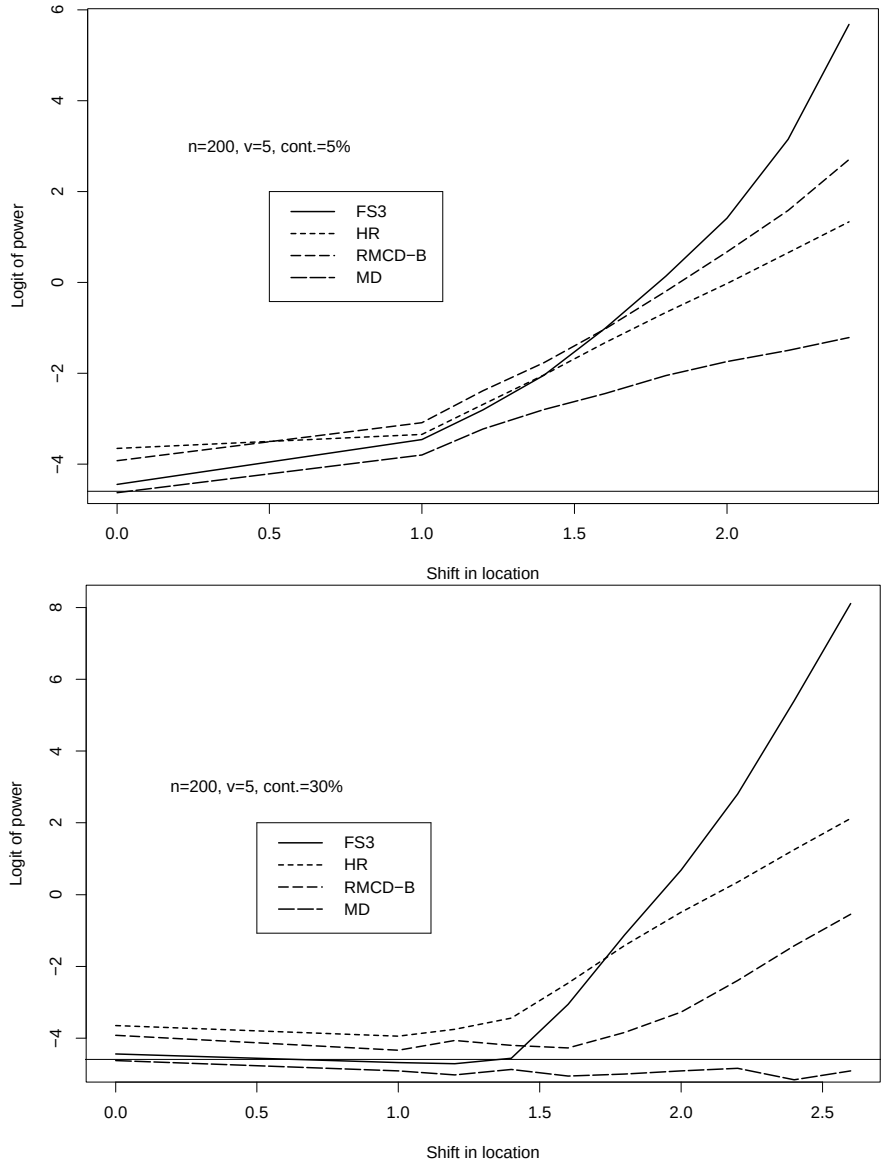

Figure 12: Power comparisons $n = 200$, $v = 5$. Logit of power: upper panel 5% contamination, lower panel 30% contamination. The lower horizontal line corresponds to a power of 1%, the nominal size of the tests

indicates that less than 1% of the samples contain outliers.

These comparisons are made simpler by the plots of Figure 12. It is customary to plot the power directly, on a scale going from 0 to 100%. However, such plots are not usually informative, since virtually all procedures start with a size near zero and finish up with a power near one. The eye is drawn to the less informative region of powers around 50%. Accordingly, we instead plot the logit of the power. That is, if the power of the procedure is p, we plot $y = \log p/(1-p)$, an unbounded function of p. An additional advantage of such plots is that we are able to make useful comparisons of tests with different actual sizes although the nominal sizes may be the same.

The upper panel of Figure 12, for 5% contamination, shows that initially FS3 has a more nearly correct size than the robust procedures RMCD-B and HR and that, as the shift in means increases, FS3 gradually becomes the most powerful procedure. The conclusions

Table 3: Power comparisons - %: $n = 200$, $v = 5$; 5% shifted observations

Shift	FS1	FS2	FS3	HR	RMCD-B	MD
0	1.14	1.15	1.16	2.53	1.94	0.97
1	3.00	3.04	3.05	3.41	4.36	2.2
1.2	5.68	5.69	5.71	6.41	8.47	3.82
1.4	11.43	11.46	11.47	11.54	14.60	5.74
1.6	26.61	26.64	26.65	20.95	26.43	8.00
1.8	53.39	53.41	53.42	34.15	45.16	11.42
2	80.42	80.43	80.44	49.38	66.32	14.94
2.2	95.87	95.88	95.89	65.90	83.03	18.33
2.4	99.64	99.65	99.66	79.12	93.73	22.89

Table 4: Power comparisons - %: $n = 200$, $v = 5$; 30% shifted observations

Shift	FS3	HR	RMCD-B	MD
0	1.16	2.53	1.94	0.97
1	0.92	1.90	1.29	0.73
1.2	0.89	2.28	1.68	0.65
1.4	1.03	3.09	1.47	0.76
1.6	4.47	7.85	1.37	0.63
1.8	24.48	19.49	2.10	0.67
2	66.39	37.95	3.64	0.73
2.2	94.27	58.66	8.36	0.78
2.4	99.55	77.73	19.31	0.57
2.6	99.97	89.35	36.55	0.73

from the lower panel for 30% contamination are similar. For large displacements not only is FS3 again the most powerful procedure, but it is comparatively more powerful than the other procedures as the shift increases. Robust tests of the correct size could be found by subtracting a constant from the logits to bring the curves down to the 1% line, when the curve for FS3 would lie together with or above those for the other curves, showing the superiority of the forward search procedure for these configurations. Such a procedure however does not provide operational tests as a simulation is needed to establish the required adjustment to the logits.

Two minor points are also of interest. One is that the lower panel reveals the complete masking associated with the non-robust MD. The other is that, for 30% contamination, all procedures require an appreciable shift before the high proportion of outliers can be detected.

Tables 5 and 6 present the results for $v = 10$, with the powers plotted in Figure 13. Table 2 shows that the sizes of HR and RMCD-B are too large. This shows in the plot by the lines for these two procedures being the highest for small contamination. However, as the mean shift increases the power curve for FS3 rises more rapidly revealing it again as

the most powerful procedure.

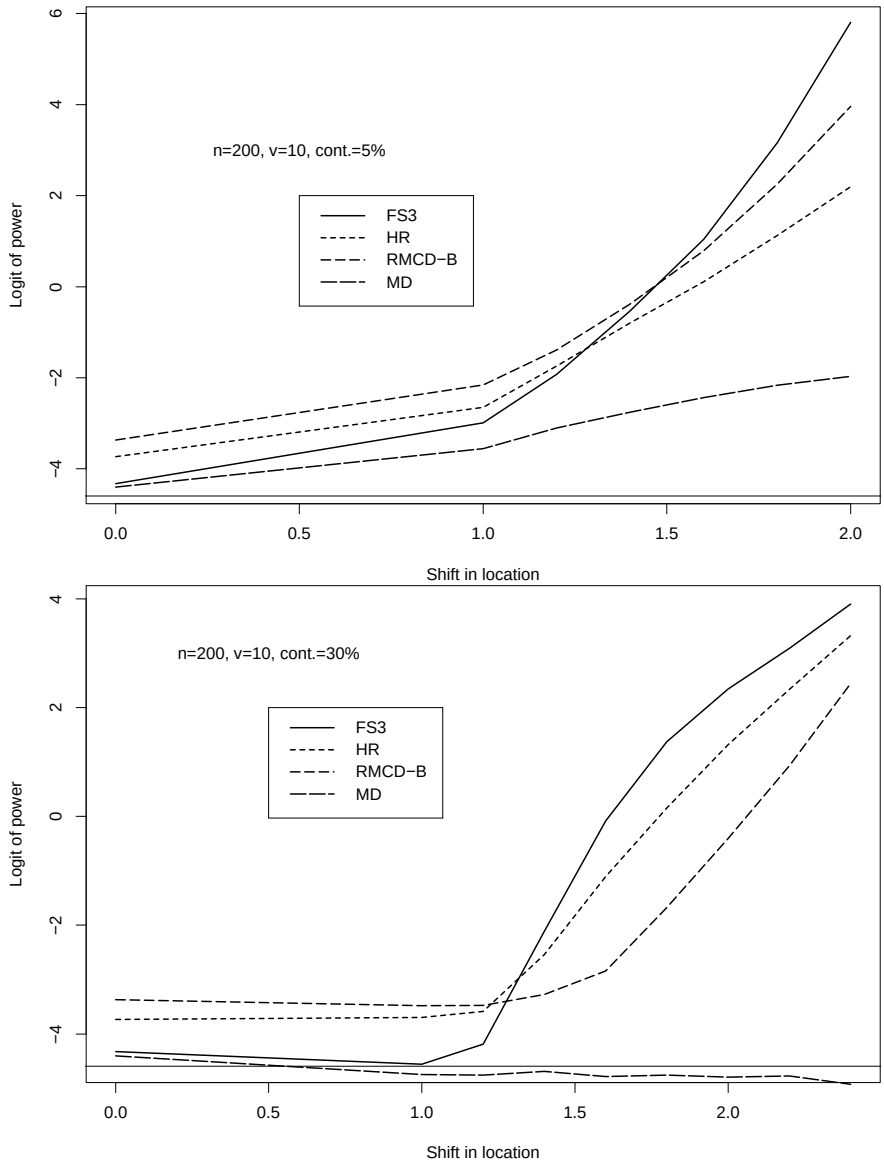

Figure 13: Power comparisons $n = 200$, $v = 10$. Logit of power: upper panel 5% contamination, lower panel 30% contamination. The lower horizontal line corresponds to a power of 1%, the nominal size of the tests

As a final comparison we look at some results for much larger samples, $n = 1000$, with $v = 5$ and 5% contamination. The results are in Table 7. The first four comparisons are those of the procedures FS3, HR, RMCD-B and MD that we have already compared for $n = 200$. The results are plotted in the upper panel of Figure 14. Now, as we know from Table 2, all procedures have virtually the correct size, so the plots of power start close together. As we have seen before, FS3 has the highest power for larger shifts in mean.

Also included in Table 7 are the results for three further versions of reweighted MCD distances. These power curves are plotted, together with that for FS3 in the lower panel of Figure 14. The plot for RMCD, the reweighted estimator without the consistency factor and lacking the Bonferroni adjustment in the reweighting, lies above the very similar curve for RMCD-C, the version of RMCD in which the consistency correction was included.

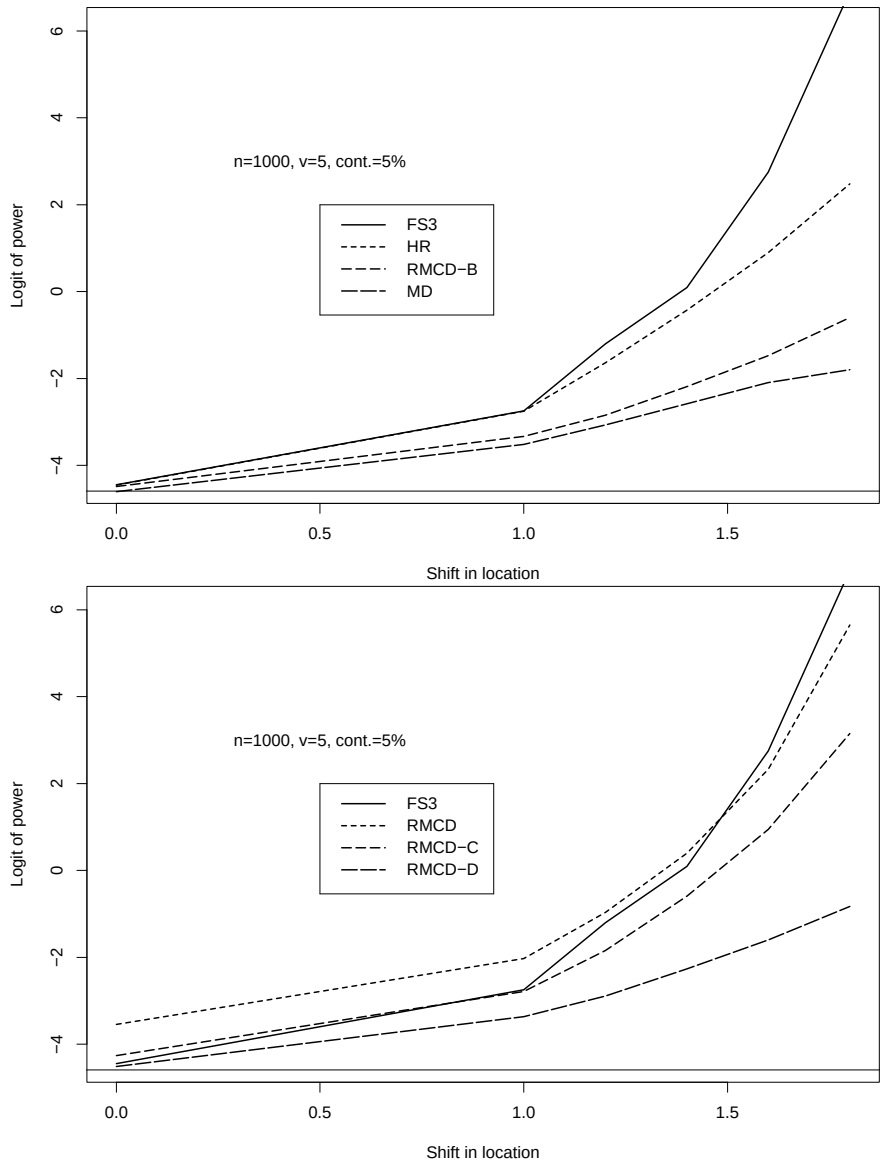

Figure 14: Power comparisons $n = 1,000$, $v = 10$ with 5% contamination. Logit of power: upper panel FS3, HR, RMCD-B and the nonrobust MD. Lower panel FS3 and three further reweighted versions of MCD. The lower horizontal line corresponds to a power of 1%, the nominal size of the tests

Both have lower power than FS3 for large mean shifts. The final version of these robust distances, RMCD-D, is the version of RMCD in which we use a Bonferroni adjustment in the reweighting of RMCD with additional consistency correction. This has the poorest performance of all, apart from the non-robust MD. The conclusion is that, once adjustment is made for size, RMCD has much the same properties as HR and RMCD-C. Here HR is better than RMCD-B, as it is for the datasets with $n = 200$ and 30% contamination. There is little to choose between these two for $n = 200$ and 5% contamination when adjustment is made for size. In all comparisons FS3 has the highest power, combined with good size.

10 Examples

Our results show the good size and superior power of our forward search procedures. In this section we conclude by revisiting our three examples and suggest why our procedure has greater power than that of MCD derived approaches.

10.1 Slight Contamination

In this example there were 500 observations, $v = 5$ and the first 25 units were contaminated. The forward plot of minimum Mahalanobis distances in Figure 7 exhibited a series of large values around $m = 480$, but with the value at the end of the search below the 99% envelope. Such masked behaviour can be expected to cause difficulties for methods that test only the largest value of a robust Mahalanobis distance based on an arbitrary or non-adaptive subset of the data. We now look at a variety of distances for each unit.

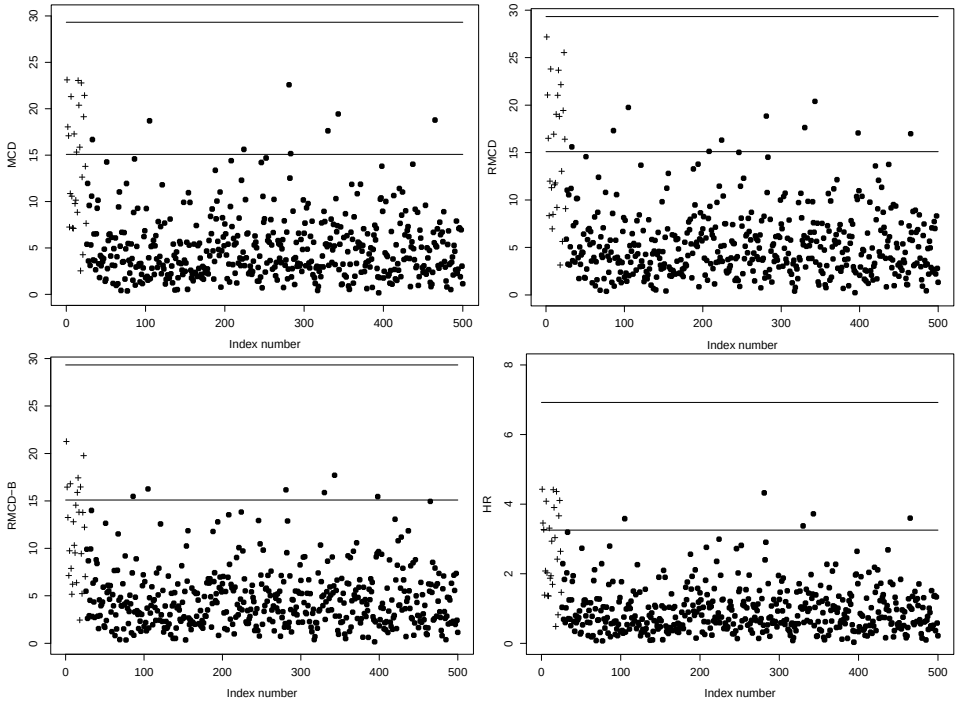

Figure 15: Slightly contaminated data: output from MCD, RMCD, RMCD-B and HR. Distances against unit number, the original units are represented by black dots, the 25 contaminated units by the symbol $+$. Upper line, 1% Bonferroni limit; lower line, 1% limit of individual distribution

The top left-hand panel of Figure 15 shows a plot of MCD distances against observation number, with the first 25 units, shown by crosses, being those that are contaminated. The lower horizontal line on the plot is the 1% point of the nominal distribution of the individual statistics. As the figure shows, for this and all other distances, there are several uncontaminated units above this threshold as well as, in this case, half of the contaminated units. A Bonferroni limit is clearly needed. However, in this and all other panels, imposition of the Bonferroni limit, the upper horizontal bound, fails to reveal any outliers. This much structure of the four panels is common. However, the figure does show that RMCD comes nearest to revealing the presence of outliers, with seven of the first 25 units forming

the largest distances. Although the Bonferroni bound is so large that these distances are not significant, the bound is not, in general, conservative. For instance, the size of RMCD with $n = 500$ and $v = 5$ is almost 4%.

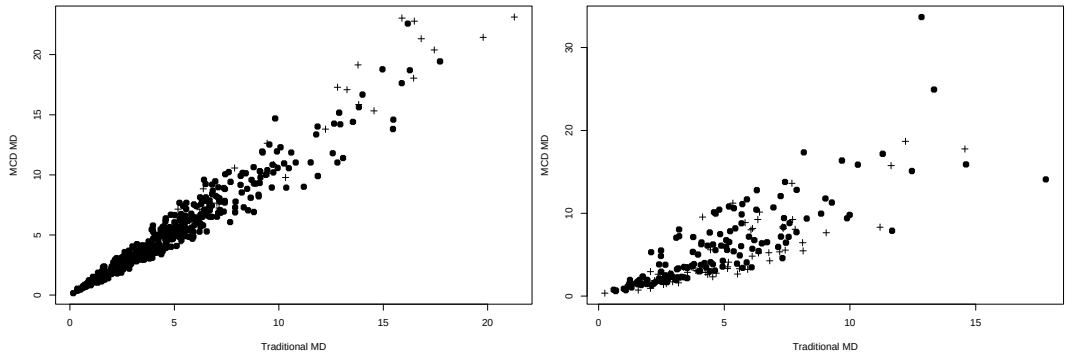

Figure 16: Robust Mahalanobis distances MCD against non-robust Mahalanobis distances MD. The original units are represented by black dots, the contaminated units by the symbol $+$. Left-hand panel, slightly contaminated data, right-hand panel, appreciably contaminated data

It is frequently suggested, for example by Rousseeuw and van Zomeren (1990), that a useful diagnostic for the presence of outliers is to compare robust and non-robust analyses. Accordingly, in the left hand-panel of Figure 16 we show a scatter plot of distances from the MCD against the full-sample Mahalanobis distances. This plot is unrevealing. Although there is a preponderance of contaminated units in the upper-right hand corner of the plot, the two sets of distances form a wedge pattern, with no obvious differences between the two. The structure is basically linear, with scatter increasing with magnitude. There is no diagnosis of the presence of outliers.

10.2 Appreciable Contamination

We now repeat this analysis but for the appreciably contaminated data of §6.2; there are 200 observations with $v = 5$, but there is 30% contamination which is in units 1-60 caused by a mean shift of 1.2 in each dimension. The forward plot of minimum Mahalanobis distances was given in Figure 10. This again has a peak well before the end, with 'good' behaviour after $m = 145$.

The panels of Figure 17 repeat those of Figure 15, showing plots of four robust distances against observation number, now with the first 60 units, shown by crosses, being those that are contaminated. The lower horizontal line on the plot is the 1% point of the nominal distribution of the individual statistics.

These plots make rather different points from those of Figure 15. Again, there would be a large number of outliers if the limit for the individual statistics were used and virtually none if the Bonferroni limit is employed. In fact, only the MCD would lead to detection of an 'outlier'. However, as the upper-left panel of the figure shows, this observation is not an outlier. More surprisingly, none of the four procedures shows a contaminated observation as having the largest distance. With this relatively small sample size, the procedures are completely failing to detect the 30% of contamination in the data. This might seem puzzling given the large size of the procedures revealed in Table 2. However, the latter part of the forward plot of Figure 10 shows that if the parameters are estimated from a subset including

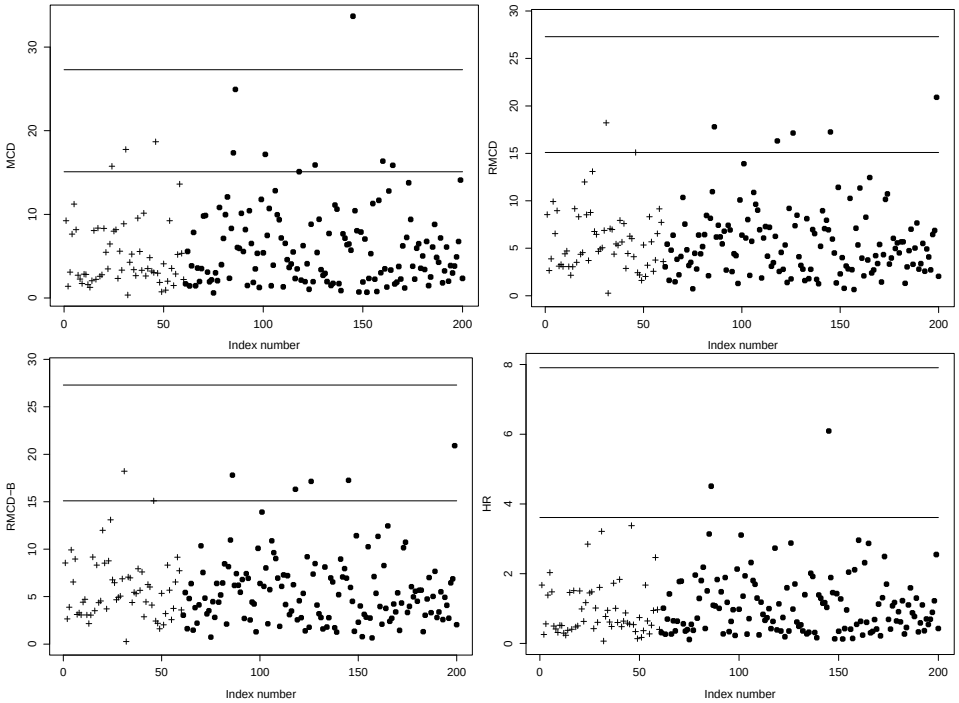

Figure 17: Appreciably contaminated data: output from MCD, RMCD, MCD-B and HR. Distances against unit number, the original units are represented by black dots, the 60 contaminated units by the symbol $+$. Upper line, 1% Bonferroni limit; lower line, 1% limit of individual distribution

contaminated observations, the resulting over-estimation of Σ leads to small distances and a failure to detect outliers.

Finally, in the right hand-panel of Figure 16 we show the scatter plot of distances from the MCD against the standard non-robust Mahalanobis distances. The structure is similar to that for the lightly contaminated data in the left-hand panel, but certainly no more informative about the existence of the 60 outliers. In fact, the plot might even be thought to be misleading; the majority of observations with larger distances lying away from the centre of the wedge shape are uncontaminated units.

10.3 Swiss Banknotes

Application of FS1 to the forward plot of distances in Figure 1 yields a value of 84 for $m^{\dagger}$. Figure 18 shows the successive superimposition of envelopes from this value. There is no evidence of any outliers when $n = 84$ and 85, but when $n = 86$ we obtain clear evidence of a single outlier with observation [86] well outside the 99% envelope. When $n = 87$ we have even stronger evidence of the presence of outliers. As a result we conclude that there are 15 outlying observations in the data on forged banknotes.

For these data the four robust methods we have compared on other sets of data also all reveal the presence of outliers. As the index plots of Mahalanobis distances in Figure 19 show, the 1% Bonferroni level for MCD, RMCD and RMCD-B all reveal the 15 outliers without any false positives. However, the sizes of these procedures when $n = 100$ are totally unacceptable, namely 62%, 30% and 5%. Only HR is too conservative, indicating just 5 outliers, from a test with size 2.4%. Since HR is a rescaled version of MCD, the

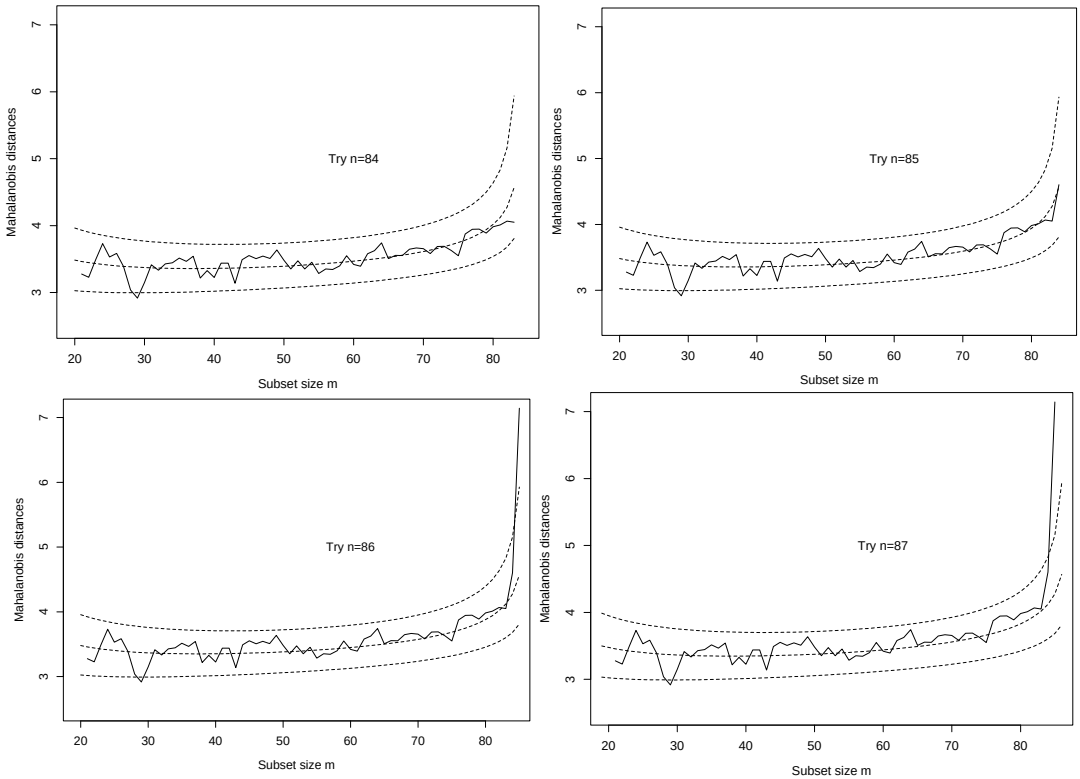

Figure 18: Swiss Banknotes: forward plot of minimum Mahalanobis distance. When $n = 84$ and 85, the observed curve lies within the 99% envelope, but there is clear evidence of an outlier when $n = 86$. The evidence becomes even stronger when another observation is included.

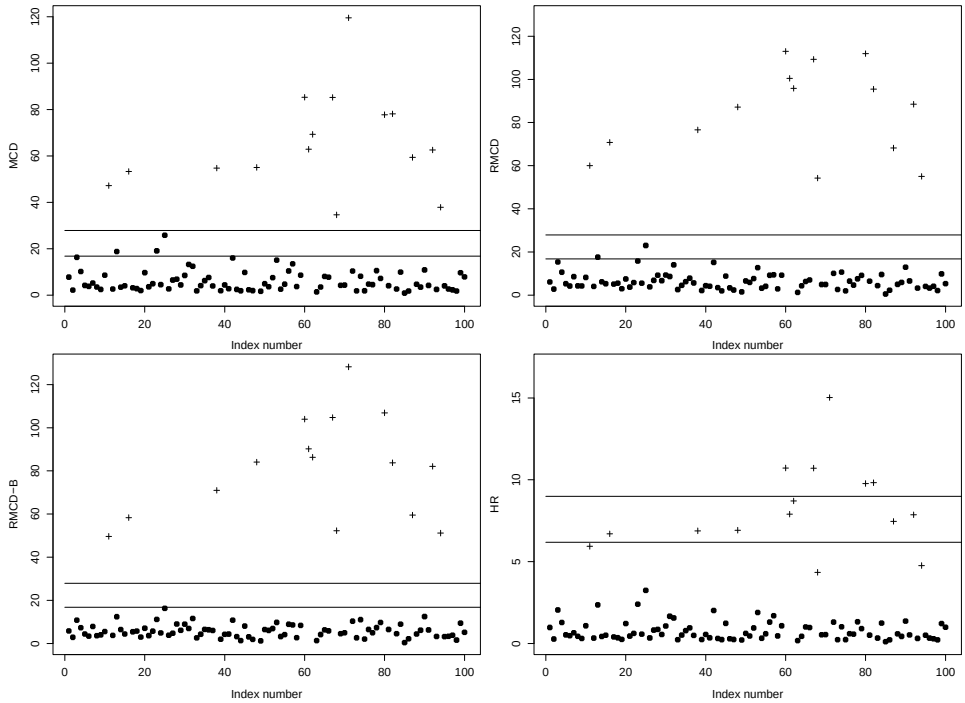

Figure 19: Swiss banknote data: output from MCD, RMCD, MCD-B and HR. Distances against unit number; the 15 outlying units are represented by the symbol $+$. Upper line, 1% Bonferroni limit; lower line, 1% limit of individual distribution

figure confirms that the 15 outliers do indeed have the largest distances for HR. The plot of robust against non-robust distances in Figure 20 also revels the 15 outliers, which fall into a separate group from the other 85 observations.

These plots serve to make the point that our comparisons have been solely of whether the methods identify at least one outlier in a sample. The comparison of methods for the number of outliers can be problematic. Consider our canonical example of a multivariate normal sample, some observations of which have a mean shift. If the shift is sufficiently large, the outliers will be evident and most methods will detect them. However, if as in our paper, the shift is slight, the two groups will overlap and the number of 'outliers' will not be as great a the number of shifted observations. Comparisons of methods then require a two-way table of counts for each procedure in which the factors are whether or not the observation was shifted and whether it was identified as outlying.

Appendix 1: Numerical

In §4.1 we mentioned that care is needed in evaluating the integral in (8) for large n as $m \to n$. For example, when $n = 1,000$ and $v = 10$, in the final step of the search we have $m = n - 1 = 999$, $x_{2,2000;0.01} = 0.01005$ and $F(y_{2000,2000;0.99}) = 0.9999899497$. This implies that we have to find the quantile of an F distribution with 10 and 989 degrees of freedom associated with probability 0.9999899497; in Fortran the IMSL function `DFIN` gave a value of 4.1985, the same value as the S-Plus function `qf`. Using this number we obtain a value of 6.512259 in equation (10). After dividing by the consistency factor we obtain a final value of 6.520. Note that the Bonferroni value is 6.426 and the coefficient

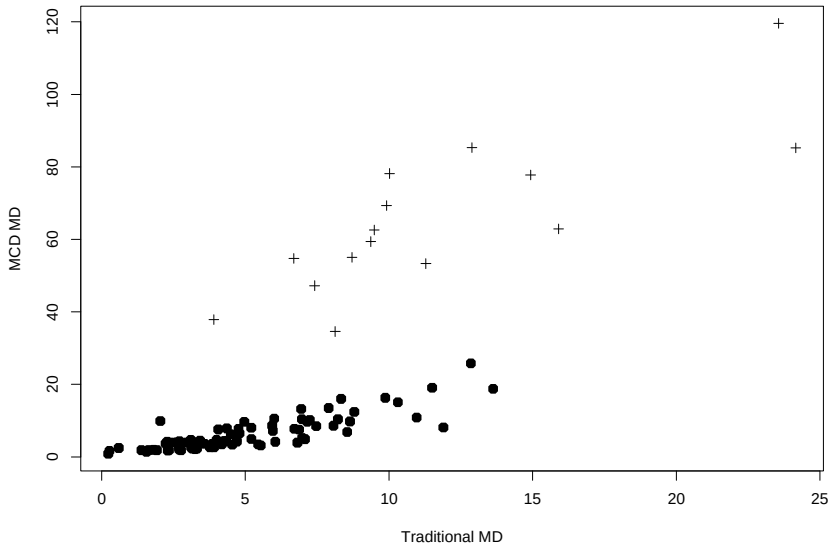

Figure 20: Swiss banknote data: robust Mahalanobis distances MCD against non-robust Mahalanobis distances MD. The 15 outlying units are represented by the symbol $+$

obtained by Hadi using simulations is 6.511. From 30,000 simulations using Gauss the value we obtained was 6.521, very close to our final value coming from the theoretical arguments leading to (10).

Appendix 2: The χ^2_3 c.d.f. as a Function of the Standard Normal Distribution

The application of standard results from probability theory shows that the variance of the truncated normal distribution containing the central m/n portion of the full distribution is

$$\sigma^2_T(m) = 1 - \frac{2n}{m}\Phi^{-1}\left(\frac{n+m}{2n}\right)\phi\left\{\Phi^{-1}\left(\frac{n+m}{2n}\right)\right\},$$

where $\phi(.)$ and $\Phi(.)$ are respectively the standard normal density and c.d.f. See, for example, Johnson, Kotz, and Balakrishnan (1994, pp. 156-162). On the other hand the results from elliptical truncation due to Tallis (1963) that we used in §4.2 show that this variance can be written as

$$\sigma^2_T(m) = \frac{n}{m}F_{\chi^2_3}\left\{F^{-1}_{\chi^2_1}\left(\frac{m}{n}\right)\right\}$$

After some algebra it appears that

$$F^{-1}_{\chi^2_1}\left(\frac{m}{n}\right) = \left\{\Phi^{-1}\left(\frac{m+n}{2n}\right)\right\}^2,$$

when, rearranging terms, we easily obtain that

$$F_{\chi^2_3}(x^2) = \frac{m}{n} - 2x\phi(x)$$

where $x = \Phi^{-1}\{(m+n)/(2n)\}$. This result links the c.d.f of the χ^2_3 in an unexpected way to the density and c.d.f. of the standard normal distribution.

References

Atkinson, A. C., M. Riani, and A. Cerioli (2004). *Exploring Multivariate Data with the Forward Search*. New York: Springer–Verlag.

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

Cox, D. R. and D. V. Hinkley (1974). *Theoretical Statistics*. London: Chapman and Hall.

Croux, H. and G. Haesbroeck (1999). Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *Journal of Multivariate Analysis 71*, 161–190.

Croux, H. and G. Haesbroeck (2000). Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika 87*, 603–618.

Davies, L. (1992). The asymptotics of Rousseeuw's minimum volume ellipsoid estimator. *The Annals of Statistics 20*, 1828–1843.

Davies, L. and U. Gather (1993). The identification of multiple outliers (with discussion). *Journal of the American Statistical Association 88*, 782–801.

Flury, B. and H. Riedwyl (1988). *Multivariate Statistics: A Practical Approach*. London: Chapman and Hall.

García-Escudero, L. A. and A. Gordaliza (2005). Generalized radius processes for elliptically contoured distributions. *Journal of the American Statistical Association 100*, 1036–1045.

Gather, U., J. Pawlitschko, and I. Pigeot (1997). A note on invariance of multiple tests. *Statistica Neerlandica 51*, 366–372.

Guenther, W. C. (1977). An easy method for obtaining percentage points of order statistics. *Technometrics 19*, 319–321.

Hadi, A. S. (1994). A modification of a method for the detection of outliers in multivariate samples. *Journal of the Royal Statistical Society, Series B 56*, 393–396.

Hadi, A. S. and J. S. Simonoff (1993). Procedures for the identification of multiple outliers in linear models. *Journal of the American Statistical Association 88*, 1264–1272.

Hardin, J. and D. M. Rocke (2005). The distribution of robust distances. *Journal of Computational and Graphical Statistics 14*, 910–927.

Johnson, N. L., S. Kotz, and N. Balakrishnan (1994). *Continuous Univariate Distributions - 1, 2nd Edition*. New York: Wiley.

Lehmann, E. (1991). *Point Estimation, 2nd edition*. New York: Wiley.

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

Maronna, R. A., R. D. Martin, and V. J. Yohai (2006). *Robust Statistics: Theory and Methods*. Chichester: Wiley.

Pison, G., S. Van Aelst, and G. Willems (2002). Small sample corrections for LTS and MCD. *Metrika 55*, 111–123.

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

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

Rousseeuw, P. J. and B. C. van Zomeren (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association 85*, 633–9.

Stuart, A. and K. J. Ord (1987). *Kendall's Advanced Theory of Statistics, Vol.1, 5th Edition*. London: Griffin.

Tallis, G. M. (1963). Elliptical and radial truncation in normal samples. *Annals of Mathematical Statistics 34*, 940–944.

Wilks, S. S. (1963). Multivariate statistical outliers. *Sankhya A 25*, 407–426.

Wisnowski, J. W., D. C. Montgomery, and J. R. Simpson (2001). A comparative analysis of multiple outlier detection procedures in the linear regression model. *Computational Statistics and Data Analysis 36*, 351–382.

Table 5: Power comparisons - %: $n = 200$, $v = 10$; 5% shifted observations

Shift	FS3	HR	RMCD-B	MD
0	1.31	2.34	3.33	1.21
1	4.79	6.59	10.34	2.78
1.2	12.75	14.93	19.93	4.31
1.4	36.96	31.01	40.61	5.96
1.6	73.98	52.83	68.78	8.04
1.8	95.93	75.47	90.47	10.31
2	99.70	89.96	98.13	12.23

Table 6: Power comparisons - %: $n = 200$, $v = 10$; 30% shifted observations

Shift	FS3	HR	RMCD-B	MD
0	1.31	2.34	3.33	0.0121
1	1.04	2.42	2.98	0.086
1.2	1.50	2.69	3.00	0.085
1.4	10.79	7.30	3.64	0.091
1.6	47.91	24.85	5.51	0.083
1.8	79.79	53.85	15.75	0.085
2	91.21	78.93	40.00	0.082
2.2	95.65	91.18	71.79	0.084
2.4	98.02	96.52	92.00	0.072
2.6	99.17	98.13	97.67	0.073

Table 7: Power comparisons for seven rules on large samples - %: $n = 1000$, $v = 5$; 5% shifted observations

Shift	FS3	HR	RMCD-B	MD	RMCD	RMCD-C	RMCD-D
0	1.16	1.15	1.11	0.99	2.80	1.39	1.08
1	6.02	5.98	3.45	2.88	11.63	5.79	3.33
1.2	23.00	16.27	5.49	4.44	27.47	13.69	5.26
1.4	52.40	39.47	10.07	7.04	59.75	35.60	9.40
1.6	94.00	71.11	18.58	10.97	91.11	71.96	16.77
1.8	99.90	92.26	35.39	14.20	99.65	95.90	30.31