Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2009)
71, *Part* 2, *pp.*

Finding an unknown number of multivariate outliers

Marco Riani,

Università di Parma, Italy

Anthony C. Atkinson

London School of Economics and Political Science, UK

and Andrea Cerioli

Università di Parma, Italy

[Received September 2007. Revised July 2008]

Summary. We use the forward search to provide robust Mahalanobis distances 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. We also introduce several new robust distances with associated distributional results. 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; Minimum covariance determinant; Order statistics; Power comparisons; Simultaneous inference; 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, it is well known that the estimates of the mean and covariance matrix found by using all the data are extremely sensitive to the presence of outliers. When there are many outliers the parameter estimates may be so distorted that the outliers are 'masked' and the Mahalanobis distances fail to reveal any outliers, or indicate as outlying observations that are not in fact so. Accordingly, several researchers have suggested the use of robust parameter estimates in the calculation of the distances.

For example, Rousseeuw and van Zomeren (1990) used minimum volume ellipsoid estimators of both parameters in calculation of the Mahalanobis distances. More recent work such as Pison *et al.* (2002) or Hardin and Rocke (2005) uses the minimum covariance determinant (MCD) estimator, whereas Atkinson *et al.* (2004), chapter 3, employed the series of robust estimators that is provided by the forward search to explore the structure of the Mahalanobis distances. The evidence that these procedures work for finite samples is little more than anecdotal—outliers that are generally believed, from other analyses, to be present are found as are outliers that are introduced into simulated data sets, or formed by perturbation of existing data. There are very

Address for correspondence: Anthony C. Atkinson, Department of Statistics, London School of Economics and Political Science, Houghton Street, London, WC2A 2AE, UK.
E-mail: a.c.atkinson@lse.ac.uk

 1369–7412/09/71000

few discussions of size and power. Cook and Hawkins (1990) suggested that the procedure of Rousseeuw and van Zomeren (1990) may find 'outliers everywhere'. The implication is that the size of the outlier test may be very much larger than the nominal 5% or 1%.

The purpose of our paper is to introduce a new outlier test using the forward search and to compare its size and power with the best existing tests. Many of these methods are designed to test whether individual observations are outlying. As do Becker and Gather (1999), we, however, stress the importance of multiple outlier testing and focus on simultaneous tests of outlyingness. We develop methods that are intended, when the samples are multivariate normal, to find outliers in $\alpha\%$ of the data sets. We show that our procedure has superior power as well as good size and so is to be recommended.

We find theoretical boundaries for our forward search procedure that allow for simultaneous inference. For the other tests, which use robust Mahalanobis distances, we use Bonferroni arguments to control for simultaneity. Our comparisons include both small sample correction factors and consistency corrections, some of which are newly derived. But for most existing tests even these corrections, combined with Bonferronization, cannot control the size of the global test. In fact, we find that, even when simultaneity is taken into account in these procedures, the actual size of the global outlier test may be much larger than the nominal 1% (in one case even 88.6%). The findings of Cook and Hawkins (1990) are not due solely to the omission of correction factors but are indicative of a wider problem with this class of procedures especially when the standard χ^2 reference distribution is employed.

Mahalanobis distances and the forward search are introduced in Section 2. In Section 3 we exhibit bootstrap envelopes for the distribution of distances in the forward search. Theoretical results on the distribution are in Section 4. In particular, Section 4.1 uses results on order statistics to find the distribution of ordered Mahalanobis distances. In Section 4.2 we use the results of Tallis (1963) on elliptically truncated multivariate normal distributions to adjust for the bias that is caused by estimation of the covariance from a subset of observations. This result also provides a simple unexploited route to the consistency corrections that are required for distances based on the MCD.

Our procedure for testing for the presence of one or more outliers in a sample is described in Section 5. Several established robust procedures for the detection of individual outlying observations, such as those of Rousseeuw and Van Driessen (1999) and Hardin and Rocke (2005), are recalled in Section 6. Some of these methods use reweighted estimates and so are based on two subsamples of the data. To adapt these tests to the detection of outliers in a sample, we introduce in Section 6.3 a Bonferroni correction to allow for simultaneity. This allows us to develop two new versions of reweighted Mahalanobis distances. The comparisons of size in Section 7.1 show that our procedure has better size than many competitors. In Section 7.2 we evaluate power both for the detection of samples containing outliers and for determining the actual number of outlying observations. The results show the superior performance of our procedure.

The paper concludes with brief comments on the relationship of outlier detection to cluster analysis. Appendix A discusses the importance of careful numerical procedures in the calculation of extreme values of order statistics and Appendix B draws a connection between the results of Tallis (1963) and the distribution of observations in a truncated univariate normal distribution. To keep our paper to a reasonable length, numerical details of several of our examples, together with plots, are given in Riani *et al.* (2007).

Our procedure provides the most powerful test for outliers among 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

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

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 that are 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 $\hat{\Sigma}(m)_{jk} = (y_j - \bar{y}_j)^{\mathrm{T}}(y_k - \bar{y}_k)/(m-1)$. Note that here y_j and y_k are $m \times 1$. From this subset we obtain n squared Mahalanobis distances

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

The single subsets that are used for each MCD-based method are defined in Section 6. 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), section 2.13).

In our examples we look at forward plots of quantities that are 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

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

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 among observations that are not in the subset

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

or its scaled version $d_{\min}^{\mathrm{sc}}(m)$. If this ordered observation $[m+1]$ is an outlier relative to the other m observations, this distance will be 'large' compared with the maximum Mahalanobis distance of observations in the subset.

In uncalibrated use of the distances $d_{\min}(m)$ to detect outliers the decision whether a difference in distances is 'large' is subjective, without reference to any null distribution. Examples include the data analyses in chapter 3 of Atkinson *et al.* (2004). Even if the asymptotic distribution of the individual distances is determined, as in Clarke and Schubert (2006), we still need to calibrate the complete search to provide an outlier test of ascertainable properties.

To calibrate the forward search and so to provide an objective basis for decisions about the number of outliers in a sample requires the distribution of $d_{\min}(m)$ in the forward search. One possibility is to use bootstrap simulations. We simulate the search a large number of times, perhaps 10000. Then, for each value of m that is of interest, we have 10000 values of $d_{\min}(m)$ from which we determine, for example, the 99% point of the distribution for each m. The upper

99% envelope is then the collection of these pointwise values which we use as a threshold in the detection of outliers. We now illustrate the use of such envelopes.

3. The structure of forward plots and the importance of envelopes: Swiss banknotes

Flury and Riedwyl (1988), pages 4–8, introduced 200 six-dimensional observations on Swiss banknotes withdrawn from circulation, which an expert has classified as genuine or forged. Units 1–100, the genuine notes, form a distinct cluster. We analyse units 101–200, the forgeries, which are less homogeneous. Fig. 1 is a plot of the minimum unscaled Mahalanobis distances during the forward search with envelopes from 10000 simulations. 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 that are not in $S^*(m)$ are somewhat remote from the cluster of outliers. Large distances at the end of the search are typical of data with unclustered outliers.

An important feature of Fig. 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.

To illustrate this point, Fig. 2 shows the successive superimposition of envelopes from $n = 84$. There is no evidence of any outliers when $n = 84$ and $n = 85$, but when $n = 86$ we obtain clear evidence of a single outlier with observation 86 well outside the 99% envelope. When $n = 87$ we

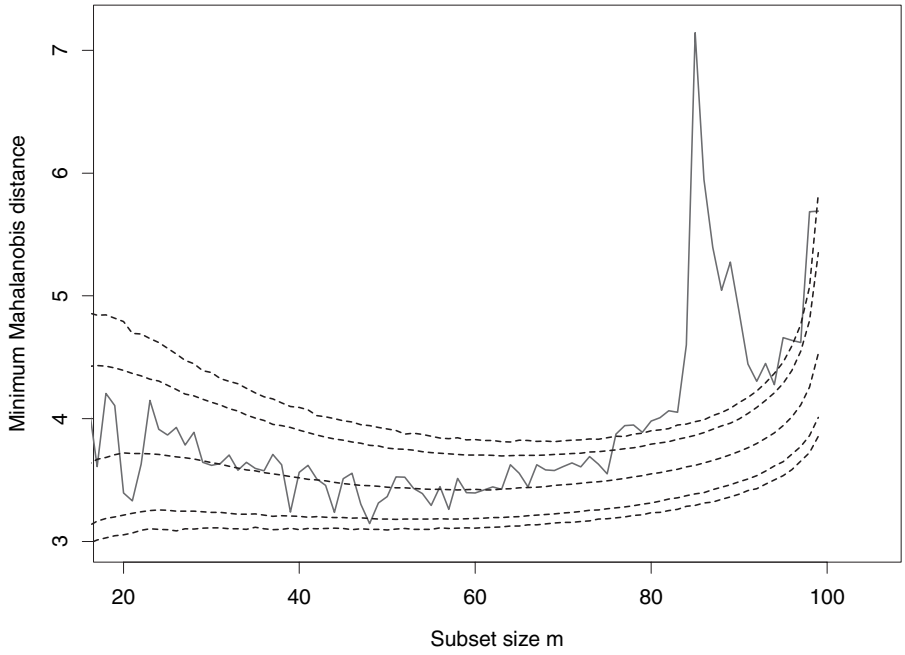

Fig. 1. Swiss banknotes, forgeries ($n = 100$)—forward plot of the minimum unscaled Mahalanobis distance with superimposed 1%, 5%, 95% and 99% bootstrap envelopes from 10000 simulations: there is a clear indication of the presence of outliers which starts around $m = 84$

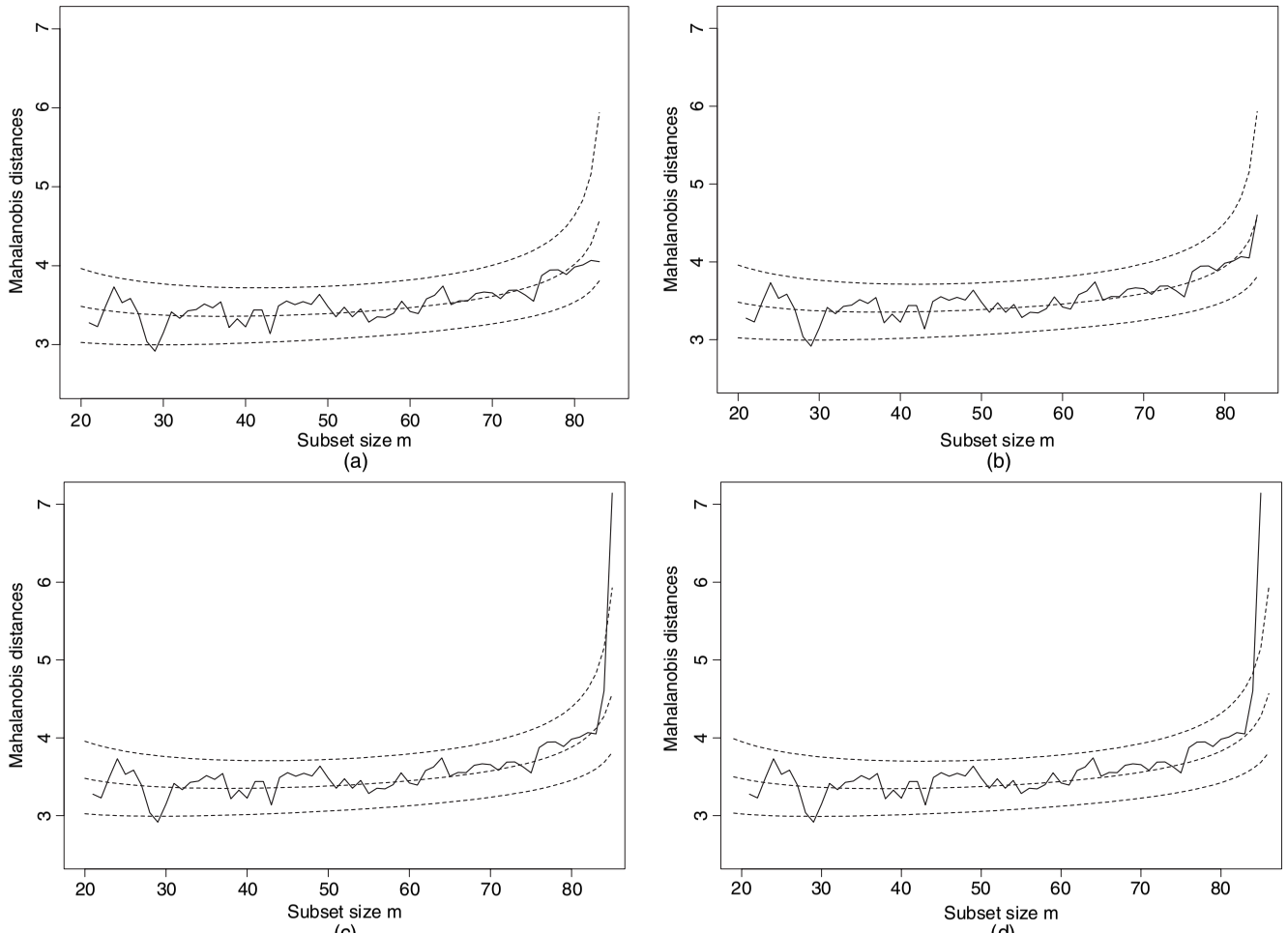

Fig. 2. Swiss banknotes—forward plot of the minimum Mahalanobis distance (when $n = 84$ and $n = 85$, the observed curve lies within the 99% bootstrap envelope, but there is clear evidence of an outlier when $n = 86$; the evidence becomes even stronger when another observation is included): (a) try $n = 84$; (b) try $n = 85$; (c) try $n = 86$; (d) try $n = 87$

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. The scatter plot in Fig. 3.52 of Atkinson *et al.* (2004) shows that these observations form a loose cluster, perhaps indicative of the activities of a second forger.

The outlier detection method that we introduce in Section 5 is an automatic version of procedures such as that illustrated here. An important feature is that we may need envelopes for several values of n as outliers are identified. However, we avoid extensive simulations by use of the theoretical results of Section 4 to provide the required envelopes. We first find theoretical envelopes for scaled distances and then convert them to those that are unscaled.

In using the forward search we potentially make many comparisons, one for each value of m. The interpretation of the threshold and so our detection rule need to allow for simultaneity, so that we have a test with size α for the presence of at least one outlier. For the effect of simultaneity on the size of forward search tests in regression see Atkinson and Riani (2006).

4. Envelopes from order statistics

4.1. Scaled distances

We now use order statistics to find good, fast approximations to the bootstrap envelopes that were mentioned in the previous section. 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 $(m+1)$th 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 $(m+1)$th order statistic from a sample of size n from a univariate distribution with cumulative distribution function (CDF) $G(y)$. Then the CDF of $Y_{[m+1]}$ is given exactly by

$$P(Y_{[m+1]} \leqslant y) = \sum_{j=m+1}^{n} \binom{n}{j} G(y)^j \{1 - G(y)\}^{n-j}. \tag{5}$$

See, for example, Casella and Berger (2002), page 228. Further, it is well known that we can apply properties of the beta distribution to the right-hand side of equation (5) to obtain

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

where

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

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

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

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 $G^{-1}(\cdot)$ in equation (8) becomes extremely close to 1 at the end of the search, i.e. 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 Appendix A.

We now consider the choice of $G(x)$. If we knew both μ and Σ, $G(x)$ would be χ^2_v distributed. When both μ and Σ are estimated by 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 *et al.* (2004), pages 43–44, derived 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

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

Plots in Riani *et al.* (2007) show the superiority of the F-approximation when $n=100$ and $v=6$, values for which asymptotic arguments are unlikely to hold. The χ^2-approximation is poor, with 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 that includes $\hat{\Sigma}(n)$ which is unbiased in the absence of outliers (Atkinson *et al.* (2004), page 66). So the envelopes for the scaled Mahalanobis distances are given by

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

4.2. Approximations for unscaled distances

Unscaled distances cannot take advantage of the beneficial cancellation of bias that is provided by the ratio $|\hat{\Sigma}(m)|/|\hat{\Sigma}(n)|$ in equation (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

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

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

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

where

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

and the variable X^2_ν has a χ^2-distribution on ν degrees of freedom. 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 mth order statistic in a sample of n squared Mahalanobis distances. Hence $c_{\mathrm{FS}}(m) = k(m)^{-1}$ for the forward search is the inflation factor for $\hat{\Sigma}(m)$ to achieve

consistency at the normal model. In large samples

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

with $\chi^2_{v,m/n}$ 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

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

The bound $\sqrt{b(m)}$ in condition (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 for fixed m, such as the MCD-based estimators of Section 6.2. A corollary of the results of Tallis (1963), relating the truncated univariate normal distribution and χ^2_3, is given in Appendix B.

Plots for typical values of n, 200 and 600, with $v = 5$ and $v = 10$ show that the approximation for scaled distances agrees well with the results of 10000 simulations and is very good virtually throughout the whole range of m. The approximation for unscaled distances is not perfect but, as we shall see, the bounds are adequate for outlier detection where we look at the upper boundaries typically in the last third of the search.

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 condition (5), once more for known μ and Σ. The asymptotic expectation of $Y_{[m+1]}$ is (Cox and Hinkley (1974), page 470) approximately

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

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

$$\sigma^2_\xi = \frac{p_\xi(1-p_\xi)}{n\, G^2(\xi_{m+1,n})}.$$

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 CDF of the standard normal distribution.

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

5. The forward search for outlier detection

If there are a few large outliers they will enter at the end of the search, and their detection is not a problem. However, even relatively small clusters of outliers can be more difficult to identify. In the Swiss banknote data the search had a central peak around $m = 85$. 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. Methods of using the forward search for the formal detection of outliers must

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.

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 that is 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 $n = 1000$ 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 $\exp(-10)$, which is a very small number. We need to be able to distinguish these random occurrences during the search from important peaks and to control for 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 $\alpha = 1\%$.

To use the envelopes in the forward search for outlier detection we accordingly 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 the value of the statistic lies beyond our threshold. In the second part we superimpose envelopes for values of n from this point until the first time that we introduce an observation we recognize as an outlier. The envelopes that are shown in Figs 1 and 2 consist roughly of two parts; a flat 'central' part and a steeply curving 'final' part. Our procedure FS for the detection of a signal takes account of these two parts.

5.1. Step 1—detection of a signal

There are four conditions, the fulfilment of any one of which leads to the detection of a signal:

(a) in the central part of the search we require three consecutive values of $d_{\min}(m, n)$ above the 99.99% envelope or one above 99.999%;
(b) in the final part of the search we need two consecutive values of $d_{\min}(m, n)$ above 99.9% and one above 99%;
(c) $d_{\min}(n-2, n)$ above the 99.9% envelope;
(d) $d_{\min}(n-1, n)$ above the 99% envelope. In this case a single outlier is detected and the procedure terminates.

The final part of the search is defined as

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

where here [ ] stands for rounded integer. For $n = 200$ the value is slightly greater than 6% of the observations. Here we have extended the notation to make clear both the subset size m for the minimum distance and the size n of the sample and so of the envelopes against which it is to be assessed.

5.2. Step 2—confirmation of a signal

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 check whether the signal is true. If $d_{\min}(m^\dagger, m^\dagger + 1)$ is below the 1% envelope, we decide that the signal is false, increment m and return to step 1.

5.3. Step 3—Identification of outliers

With a true signal we start superimposing 99% envelopes taking $n^* = m^\dagger - 1, m^\dagger, m^\dagger + 1, \ldots$ until

the final, penultimate or antepenultimate value is above the 99% threshold or, alternatively, we have a value of $d_{\min}(m, n^*)$ for any $m > m^\dagger$ which is greater than the 99.9% threshold.

In addition if we have an incontrovertible signal,

(a) three consecutive values of $d_{\min}(m, n)$ above the 99.999% threshold or
(b) 10 values of $d_{\min}(m, n)$ above the 99.999% threshold,

we then decide that outliers are present and proceed directly to step 3.

Some features of this procedure may seem arbitrary. However, as we see in Section 6, 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 that are used for testing these distances. In addition, most existing robust procedures rely heavily on simulation-based calibration factors that become essential for simultaneous outlier testing.

6. Other outlier detection procedures

6.1. Bonferroni bounds in the forward search

The statistic (2) provides the basis for our test of the outlyingness of observation $[m+1]$. Hadi (1994) used a Bonferroni bound to allow for the ordering of the distances during his forward search and compared a slightly scaled version of expression (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 $m+1$, we use the Bonferroni bound $\chi^2_{v,\{\alpha/(m+1)\}}$ rather than $\chi^2_{v,(\alpha/n)}$. Fig. 3 shows the resulting 95% and 99% bounds super-

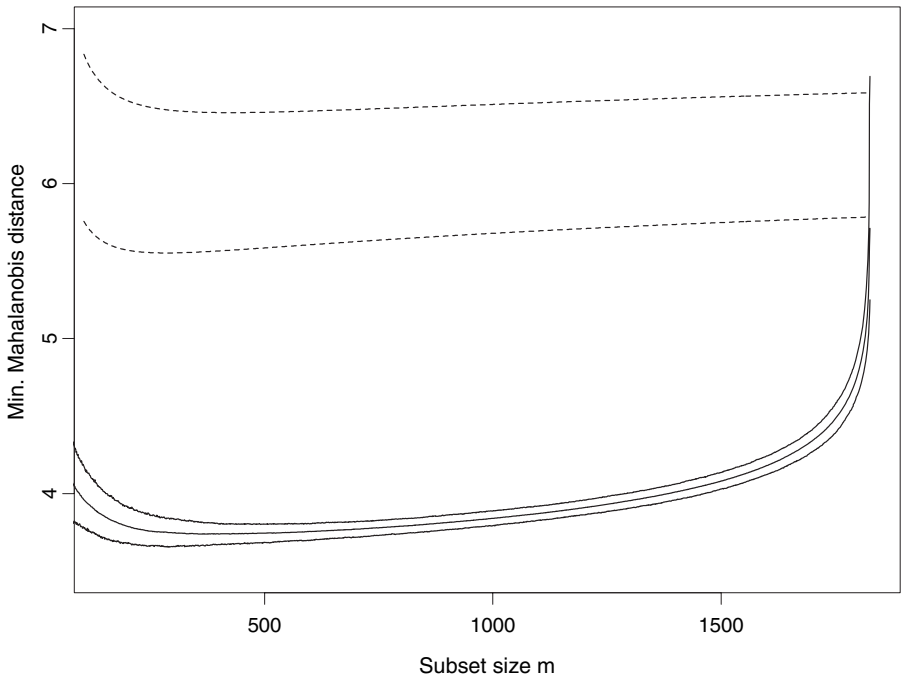

Fig. 3. 2000 normal observations, $v = 10$: forward plot of 90% and 99% envelopes of minimum Mahalanobis distances with superimposed Bonferroni bounds including Hadi's correction

imposed on a forward plot of bootstrap envelopes for $n = 2000$ and $v = 10$. These bounds were calculated by using the empirical scaling in section 2 of Hadi (1994) which is an improvement of Hadi (1992). They are unrelated to the true distribution, except for the last step of the search; owing 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), page 360, reported 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 Fig. 3.

6.2. Distances for outlier detection

In this section we describe some variants of the Mahalanobis distance that have been recommended for outlier detection. These vary in the subset or subsets of observations that are used for parameter estimation. When robust estimates are used, there are several possible adjustments to obtain consistent and possibly unbiased estimators of Σ. There is also a choice of reference distribution against which to assess the observed distances. We leave until Section 6.3 the adjustments that are made for simultaneous inference which introduce further subsets of the data to be used for estimation.

6.2.1. Tests MD and MDK

The Mahalanobis distance (1), with parameters estimated from all the data, has long been suggested as an outlier test, e.g. 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 that is based on the same distances but referred to the correct scaled beta distribution. Section 2.6 of Atkinson *et al.* (2004) gives this distribution; section 2.16 gives references to the repeated rediscovery of related distributional results.

6.2.2. Test MCD

One popular robust estimator is the MCD estimator that was described in Rousseeuw and Leroy (1987), page 262. Fix an integer h such that

$$\left\lfloor \frac{n+v+1}{2} \right\rfloor \leqslant h < n,$$

where $\lfloor \cdot \rfloor$ denotes the integer part. The preferred choice of h for outlier detection is its lower bound, which yields the breakdown value

$$\frac{\lfloor (n-v+1)/2 \rfloor}{n}. \tag{14}$$

Let $\hat{\mu}_{\mathrm{MCD}}$ and $\hat{\Sigma}_{\mathrm{MCD}}$ be the mean and the covariance matrix of the subset of h observations for which the determinant of the covariance matrix is minimal. $\hat{\mu}_{\mathrm{MCD}}$ is defined to be the MCD estimator of μ, whereas the MCD estimator of Σ is proportional to $\hat{\Sigma}_{\mathrm{MCD}}$. The proportionality constant is chosen to achieve consistency at the normal model. It was derived by Butler *et al.* (1993) and by Croux and Haesbroeck (1999) as

$$c_{\mathrm{MCD}}(h,n,v)=\frac{h/n}{P(X^2_{v+2}<\chi^2_{v,h/n})}. \tag{15}$$

For MCD-based estimators with $h=m$, the consistency factor is then equal to $c_{\mathrm{FS}}(m)$ defined in equation (12). 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 $n^{-1/3}$. Another reason is the fast algorithm of Rousseeuw and Van Driessen (1999), which has been implemented in many languages.

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 *et al.* (2002) showed by Monte Carlo simulation the importance of applying a small sample correction factor to it. Let $s_{\mathrm{MCD}}(h,n,v)$ be this factor for a specific choice of n and v and breakdown value (14). The resulting (squared) robust Mahalanobis distances are then

$$d^2_{(\mathrm{MCD})i}=k_{\mathrm{MCD}}(y_i-\hat{\mu}_{\mathrm{MCD}})^{\mathrm{T}}\hat{\Sigma}^{-1}_{\mathrm{MCD}}(y_i-\hat{\mu}_{\mathrm{MCD}}), \qquad i=1,\ldots,n, \tag{16}$$

where $k_{\mathrm{MCD}}=\{c_{\mathrm{MCD}}(h,n,v)\,s_{\mathrm{MCD}}(h,n,v)\}^{-1}$. These distances are compared with the α% cut-off value of their asymptotic χ^2_v-distribution, with α in most published reports between 0.01 and 0.05.

6.2.3. *Test HR*

The exact finite sample distribution of the robust Mahalanobis distances (16) 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.

6.2.4. *Test 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 cut-off value. Thus a first subset of h observations is used to select a second subset from which the parameters are estimated. The default choice for this cut-off is

$$\sqrt{\chi^2_{v,0.025}}. \tag{17}$$

Both the consistency (Croux and Haesbroeck, 2000) and the small sample (Pison *et al.*, 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 squared robust Mahalanobis distances become

$$d^2_{(\mathrm{RMCD\text{-}C})i}=k_{\mathrm{RMCD\text{-}C}}(y_i-\hat{\mu}_{\mathrm{RMCD}})^{\mathrm{T}}\hat{\Sigma}^{-1}_{\mathrm{RMCD}}(y_i-\hat{\mu}_{\mathrm{RMCD}}), \qquad i=1,\ldots,n, \tag{18}$$

with $k_{\mathrm{RMCD\text{-}C}}=\{c_{\mathrm{RMCD}}(h,n,v)\,s_{\mathrm{RMCD}}(h,n,v)\}^{-1}$. The reweighted distances are again compared with their asymptotic χ^2_v-distribution (Lopuhaä, 1999).

6.2.5. *Test 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, $d^2_{(\mathrm{RMCD})i}$, are then computed as in equation (18), but with $k_{\mathrm{RMCD}} = s_{\mathrm{RMCD}}(h, n, v)^{-1}$ replacing $k_{\mathrm{RMCD\text{-}C}}$.

6.2.6. Test PP

An alternative approach to robust multivariate estimation is based on projecting the sample points onto a set of univariate directions. Peña and Prieto (2001) suggested considering the set of $2v$ directions that are obtained by maximizing and minimizing the kurtosis coefficient of the projected data. They also proposed an iterative algorithm where each observation is repeatedly tested for outlyingness in these directions, using subsamples of decreasing size with potential outliers removed. Their final robust estimates, $\hat{\mu}_{\mathrm{PP}}$ and $\hat{\Sigma}_{\mathrm{PP}}$, are computed by using the observations which are considered not to be outliers at the end of the iterations. A calibration factor k_{PP} is still required to allow for bias in estimation of Σ. The resulting (squared) robust Mahalanobis distances

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

are compared with the $\{v(n-1)/(n-v)\}F_{v,n-v}$-distribution. The calibration factor k_{PP} and the thresholds that are required to define potential outliers were chosen by simulation.

6.3. Simultaneity and Bonferronization

To be compared, different testing procedures must look at the same null hypothesis. With only a few exceptions, such as Becker and Gather (1999), the published literature on robust Mahalanobis distances has been concerned with detection of individual outliers, i.e. with repeated testing of the null hypotheses

$$H_0 : y_i \sim N(\mu, \Sigma), \qquad i = 1, \ldots, n, \tag{20}$$

each at level α. In that framework the per comparison error rate α is interpreted as the proportion of *observations* that one should be prepared to declare as outliers in any application. On the contrary, in our procedure of Section 4 the test statistic (4) is the $(m+1)$th ordered value of the n Mahalanobis distances. Its distribution involves the joint distribution of all the distances, so 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. A special case of hypothesis (21) was called 'the general outlier problem' by Schwager and Margolin (1982), who derived the corresponding locally best invariant test as a function of the multivariate sample kurtosis. However, the invariant test is not robust and does not address the problem of specifying the number and the identity of the outliers when the null hypothesis is rejected, a goal which is achieved by the third stage of our procedure in Section 4; superimposing envelopes serves the purpose of precise identification of how many and which of the n individual hypotheses in expression (21) should be rejected once a signal has been confirmed.

The forward search α is the size of the test of hypothesis (21), i.e. the probability that at least one of the individual hypotheses (20) is rejected for some m when hypothesis (21) is true. By controlling this error rate under the simultaneous hypothesis, we are willing to tolerate a wrong conclusion in $(100\alpha)\%$ of *data sets* without outliers. We let α have a similar interpretation in the robust procedures of Section 6.2 by comparing the individual statistics $d_{(\mathrm{MCD})i}$, $d_{(\mathrm{RMCD})i}$, $d_{(\mathrm{RMCD\text{-}C})i}$ and $d_{(\mathrm{PP})i}$, $i = 1, \ldots, n$, with the $\alpha^* = \alpha/n$ cut-off value of their reference distribu-

tions. A Bonferroni approach is appropriate in the MCD context because, as was shown by Hardin and Rocke (2005), extreme observations are approximately independent of the estimators $\hat{\mu}_{\mathrm{MCD}}$ and $\hat{\Sigma}_{\mathrm{MCD}}$. Hence the intersection between multiple tests of hypothesis (20), sharing the same MCD estimates, should be negligible, at least when H_0 is rejected. Although the extent of dependence among the projection-based distances $d_{(\mathrm{PP})i}$ is unknown, we note that a Bonferroni approach was adopted also by Becker and Gather (1999) in their study of robust distances for simultaneous outlier identification.

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 by the MCD of observations that are to be used in parameter estimation for the reweighted distances. We suggest two such modifications.

6.3.1. *Test RMCD-B*

The default cut-off value for excluding observations in the computation of reweighted MCD estimators is given by expression (17). However, this cut-off is not coherent with 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 $\hat{\mu}_{\mathrm{RMCD\text{-}B}}$ and $\hat{\Sigma}_{\mathrm{RMCD\text{-}B}}$, where observations are given weight 0 if $d_{(\mathrm{MCD})i}$ exceeds

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

Substituting these modified estimators into expression (18), we obtain the Bonferroni-adjusted (squared) reweighted distances

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

6.3.2. *Test RMCD-D*

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

Table 1. Mahalanobis distance outlier tests to be compared with the forward search

Test	*Description*
MDK	Squared non-robust distances d_i^2; asymptotic χ^2_v-distribution
MD	Squared non-robust distances d_i^2; exact scaled beta distribution
MCD	Squared MCD distances $d^2_{(\mathrm{MCD})i}$; asymptotic χ^2_v-distribution
RMCD	Squared reweighted MCD distances $d^2_{(\mathrm{RMCD})i}$; asymptotic χ^2_v-distribution
RMCD-C	Squared reweighted MCD distances with consistency correction $d^2_{(\mathrm{RMCD\text{-}C})i}$; asymptotic χ^2_v-distribution
RMCD-B	Squared Bonferroni-adjusted reweighted MCD distances $d^2_{(\mathrm{RMCD\text{-}B})i}$; asymptotic χ^2_v-distribution
RMCD-D	Squared Bonferroni-adjusted reweighted MCD distances with consistency correction $d^2_{(\mathrm{RMCD\text{-}D})i}$; asymptotic χ^2_v-distribution
HR	Squared MCD distances $d^2_{(\mathrm{MCD})i}$; scaled F-distribution of Hardin and Rocke (2005)
PP	Squared projection-based robust distances of Peña and Prieto (2001); scaled F-distribution

The correction factors in these Bonferroni-adjusted versions of test 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 tests RMCD-B and RMCD-D is not available in the MCD literature.

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

7. Size and power

7.1. 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 10000 sets of simulated multivariate normal data for four values of n from 100 to 1000 and with dimension $v = 5$ and $v = 10$. The result was considered significant if at least one outlier was detected.

We summarize our findings in Table 2. For the first eight tests, based on various Mahalanobis distances, and for test PP we use the Bonferroni correction to obtain a test with nominal size 1%. The first entry in Table 2 is for the standard Mahalanobis distance with reference 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 $n = 1000$, a value by which asymptotics are usually expected to be a good guide, the size is only 0.79%

Table 2. Size of the nominal 1% test based on 10000 simulations: classical Mahalanobis distances, the six MCD-based procedures of Table 1, the projection method PP and our proposal FS†

Test	*Sizes (%) for the following values of n:*			
	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
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
PP	1.13	1.04	1.07	1.12
	1.07	1.00	0.99	1.03
FS	1.04	1.16	1.15	1.16
	1.54	1.31	1.18	1.20

†The first entry in each cell is for $v = 5$ and the second entry is for $v = 10$.

when $v = 5$. There is a sharp contrast with the results by 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 Section 7.2 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 test. 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 $n = 1000$, 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 Table 2. For test 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 test RMCD-B we extend test 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 test 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 test HR when $n = 100$ are also too large, although throughout Table 2 this test has values that are among the best for all values of n. The last two entries of Table 2 show that the projection method PP (after Bonferronization) and our proposal test FS both have satisfactory sizes for all values of n in the range that was studied, although the values are slightly above 1%.

7.2. Power

We performed numerous simulations to compare the power of the various procedures. We start with the simultaneous hypothesis (21) that the sample contains at least one outlier. These simulation results are presented as plots. 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 0 and finish with a power near 1. The eye is drawn to the less informative region of powers around 50%. Accordingly, we instead plot the logit of the power, i.e., if the power of the procedure is p, we plot $y = \log\{p/(1-p)\}$, which is an unbounded function of p. An additional advantage of such plots is that we can make useful comparisons of tests with different actual sizes although the nominal sizes may be the same.

We again consider the multivariate location–shift model with constant contamination on all variables. Fig. 4(a) shows results for $n = 200$ and $v = 5$ with 5% contamination calculated from 10000 simulations. The initial part of the plot reflects the larger size of tests HR and RMCD-B that was a feature of Table 2. The power of all the procedures increases with the shift in location. Once adjustment has been made for the differences in size, tests HR and RMCD-B have better power than MD and PP, which behave very similarly. Test FS is revealed as the most powerful procedure.

Fig. 4(b) is for 30% contamination. The results are similar to those for 5% contamination with test FS again the most powerful procedure. The PP procedure completely fails, although our simulations show that it can detect outliers for larger shifts. Further results on rejection of the simultaneous hypothesis are in section 9.2 of Riani *et al.* (2007). These findings show the clear superiority of our algorithm in detecting at least one outlier in contaminated data.

When the simultaneous hypothesis has been rejected it remains to determine which of the observations actually come from the contaminated part of the sample. We interpret our results by using 'average power' (Benjamini and Hochberg (1995), section 4.2), i.e. the average number

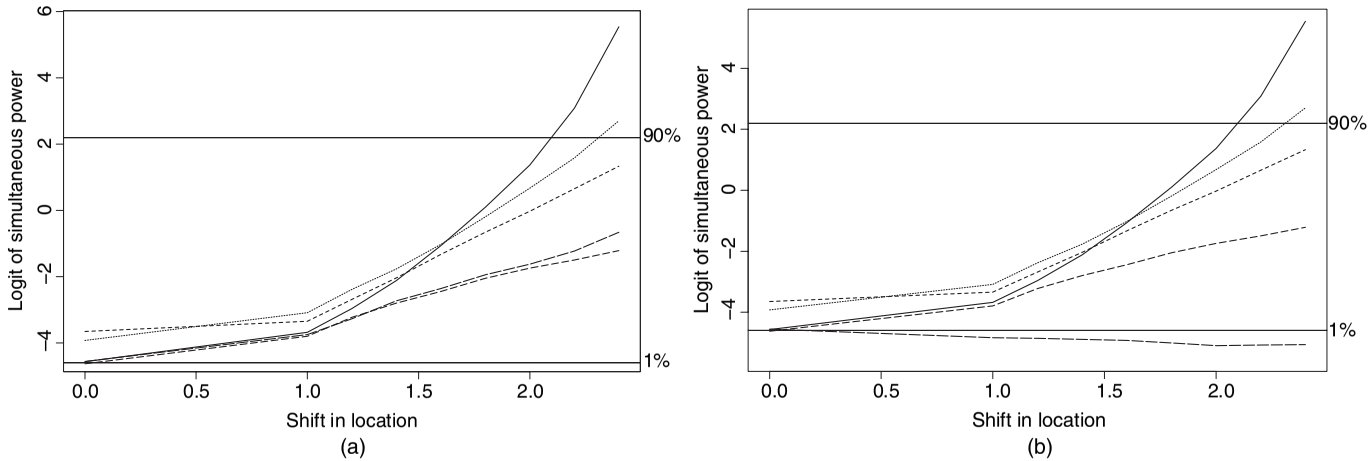

Fig. 4. Power comparisons, $n = 200$ and $v = 5$—logit of simultaneous power (the lower horizontal line corresponds to a power of 1%, the nominal size of the tests; the upper line to 90%; ———, FS; · · · · · · ·, RMCD-B; -------, HR; – – –, MD; — —, PP): (a) 5% contamination; (b) 30% contamination

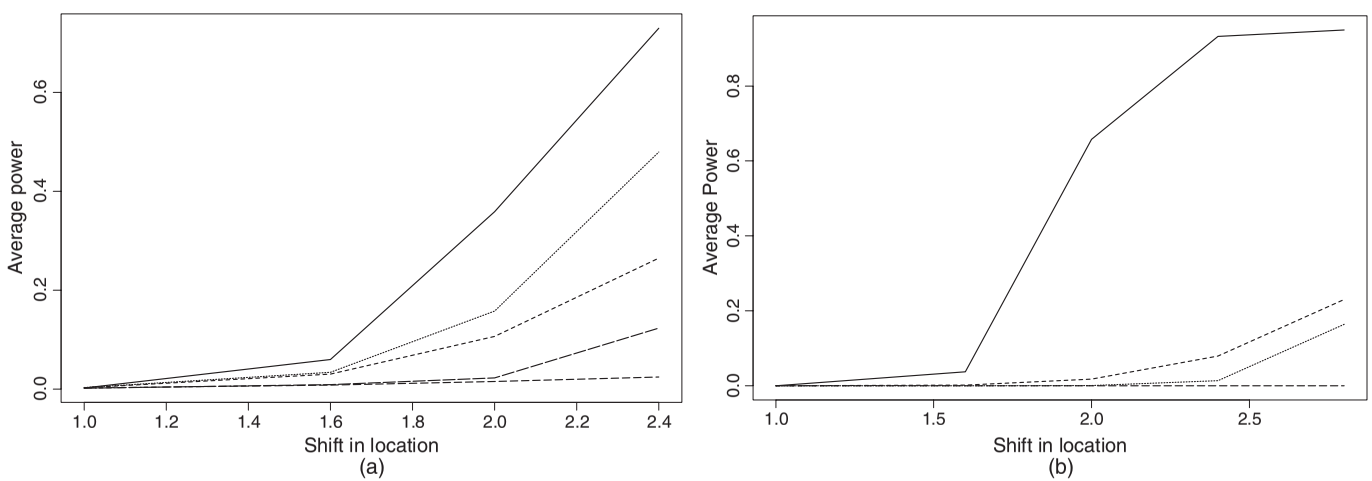

Fig. 5. Power comparisons, $n = 200$ and $v = 5$—average power: (a) 5% contamination (———, FS; · · · · · · ·, RMCD-B; -------, HR; – – –, MD; — —, PP); (b) 30% contamination (———, FS; · · · · · · ·, RMCD-B; -------, HR; – – –, MD and PP)

of contaminated observations correctly identified. Unlike the previous power calculations, the size of this test tends to zero as the shift decreases until, in the limit, there are no contaminated observations, so they cannot be identified. We accordingly plot the results of our power calculations directly, rather than as logits.

The comparisons are in Fig. 5 and are even more advantageous for the FS test than those for simultaneous power. Fig. 5(a), for 5% contamination, shows that virtually no outliers are correctly identified when the shift in location is 1. As the shift increases, the average power increases for all procedures except MD. The power of the FS test is highest, with the other procedures in the same order as for the higher shifts in Fig. 4. The MD test identifies one or no outliers, the simultaneous power of 0.224 in Fig. 4 being divided by 10 to give the low value here.

The final plot, Fig. 5(b), shows how excellently test FS behaves relatively to all other procedures when there is appreciable contamination. Tests MD and PP are completely disabled by masking, HR and RMCD-B are little better, but FS identifies an increasingly large number of outliers as the shift increases above 1.5. This outstanding performance is due to the ability to respond to the peak in the middle of the search that was caused by a cluster of outliers that is a feature of Fig. 1.

8. Discussion

Our results show the good size and superior power of our FS procedure. Riani *et al.* (2007) supplement the analysis of size and simultaneous power with detailed analysis of individual data sets. These include plots of the robust distances of Table 1 against observation number with percentage points of both the nominal distribution of the individual statistics and of the Bonferroni limit. Such plots are informative about the relationship between the various distances and the observations that they label as outliers. Another tool is the scatter plot of robust against non-robust distances (Rousseeuw and van Zomeren, 1990). This is informative for the banknote data, but not for some other data structures with a less obvious structure of outliers. As the scatter plot of the fourth and sixth variables of the banknote data in Fig. 3.52 of Atkinson *et al.* (2004) shows, the 15 observations that were identified as outliers are well separated from the rest of the data. Any plausible outlier detection procedure should be able to detect this structure.

Finally, we note that MCD-based methods require that at least half the data come from a single normal distribution. There is, however, no such constraint on the forward search, which starts from very small subsets of the data; outlier identification then merges seamlessly with cluster detection. Atkinson and Riani (2007) illustrated the use of random-start forward searches in exploratory clustering. We hope that the methods of the present paper will provide a basis for the statistical modelling of this process.

Acknowledgements

We are grateful to the Joint Editor, Associate Editor and referees whose comments led to an appreciable clarification of the argument of our paper.

Our work was supported by the grants 'Metodi statistici multivariati per la valutazione integrata della qualità dei servizi di pubblica utilità: efficacia-efficienza, rischio del fornitore, soddisfazione degli utenti' and 'Metodologie statistiche per l'analisi di impatto e la valutazione della regolamentazione' of the Ministero dell'Università e della Ricerca—Progetti di Ricerca di Interesse Nazionale 2006. Financial support of the Departments of Economics at the University of Parma and of Statistics at the London School of Economics and Political Science is also gratefully acknowledged.

Appendix A: Numerical details

In Section 4.1 we mentioned that care is needed in evaluating the integral in equation (8) for large n as $m \to n$. For example, when $n = 1000$ and $v = 10$, in the final step of the search we have $m = n - 1 = 999$, $x_{2,2000;0.01} = 0.01005$ and $F(y_{2000,2000;0.99}) = 0.999\,989\,949\,7$. This implies that we must find the quantile of an F-distribution with 10 and 989 degrees of freedom associated with probability $0.999\,989\,949\,7$; in Fortran the International Mathematical and Statistical Libraries 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 that was obtained by Hadi using simulations is 6.511. From 30 000 simulations using Gauss the value that we obtained was 6.521, which is very close to our final value coming from the theoretical arguments leading to equation (10).

Appendix B: The χ^2_3 cumulative distribution function 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 $\phi(\cdot)$ and $\Phi(\cdot)$ are respectively the standard normal density and CDF. See, for example, Johnson *et al.* (1994), pages 156–162. However, the results from elliptical truncation due to Tallis (1963) that we used in Section 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 CDF of the χ^2_3-distribution in an unexpected way to the density and CDF of the standard normal distribution.

References

Atkinson, A. C. and Riani, M. (2006) Distribution theory and simulations for tests of outliers in regression. *J. Computnl Graph. Statist.*, **15**, 460–476.

Atkinson, A. C. and Riani, M. (2007) Exploratory tools for clustering multivariate data. *Computnl Statist. Data Anal.*, **52**, 272–285.

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

Becker, C. and Gather, U. (1999) The masking breakdown point of multivariate outlier identification rules. *J. Am. Statist. Ass.*, **94**, 947–955.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B, **57**, 289–300.

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

Casella, G. and Berger, R. L. (2002) *Statistical Inference*, 2nd edn. Pacific Grove: Duxbury.

Clarke, B. R. and Schubert, D. D. (2006) An adaptive trimmed likelihood algorithm for identification of multivariate outliers. *Aust. New Zeal. J. Statist.*, **48**, 353–371.

Cook, R. D. and Hawkins, D. M. (1990) Comment on Rousseeuw and van Zomeren (1990). *J. Am. Statist. Ass.*, **85**, 640–644.

Cox, D. R. and Hinkley, D. V. (1974) *Theoretical Statistics*. London: Chapman and Hall.
Croux, H. and Haesbroeck, G. (1999) Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *J. Multiv. Anal.*, **71**, 161–190.
Croux, H. and Haesbroeck, G. (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. *Ann. Statist.*, **20**, 1828–1843.
Flury, B. and Riedwyl, H. (1988) *Multivariate Statistics: a Practical Approach*. London: Chapman and Hall.
García-Escudero, L. A. and Gordaliza, A. (2005) Generalized radius processes for elliptically contoured distributions. *J. Am. Statist. Ass.*, **100**, 1036–1045.
Guenther, W. C. (1977) An easy method for obtaining percentage points of order statistics. *Technometrics*, **19**, 319–321.
Hadi, A. S. (1992) Identifying multiple outliers in multivariate data. *J. R. Statist. Soc.* B, **54**, 761–771.
Hadi, A. S. (1994) A modification of a method for the detection of outliers in multivariate samples. *J. R. Statist. Soc.* B, **56**, 393–396.
Hadi, A. S. and Simonoff, J. S. (1993) Procedures for the identification of multiple outliers in linear models. *J. Am. Statist. Ass.*, **88**, 1264–1272.
Hardin, J. and Rocke, D. M. (2005) The distribution of robust distances. *J. Computnl Graph. Statist.*, **14**, 910–927.
Johnson, N. L., Kotz, S. and Balakrishnan, N. (1994) *Continuous Univariate Distributions*, vol. 1, 2nd edn. New York: Wiley.
Lopuhaä, H. P. (1999) Asymptotics of reweighted estimators of multivariate location and scatter. *Ann. Statist.*, **27**, 1638–1665.
Peña, D. and Prieto, F. J. (2001) Multivariate outlier detection and robust covariance matrix estimation (with discussion). *Technometrics*, **43**, 286–310.
Pison, G., Van Aelst, S. and Willems, G. (2002) Small sample corrections for LTS and MCD. *Metrika*, **55**, 111–123.
Riani, M., Atkinson, A. C. and Cerioli, A. (2007) Results in finding an unknown number of multivariate outliers in large data sets. *Research Report 140*. Department of Statistics, London School of Economics and Political Science, London.
Rousseeuw, P. J. and Leroy, A. M. (1987) *Robust Regression and Outlier Detection*. New York: Wiley.
Rousseeuw, P. J. and Van Driessen, K. (1999) A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, **41**, 212–223.
Rousseeuw, P. J. and van Zomeren, B. C. (1990) Unmasking multivariate outliers and leverage points. *J. Am. Statist. Ass.*, **85**, 633–639.
Schwager, S. J. and Margolin, B. H. (1982) Detection of multivariate normal outliers. *Ann. Statist.*, **10**, 943–954.
Stuart, A. and Ord, K. J. (1987) *Kendall's Advanced Theory of Statistics*, vol. 1, 5th edn. London: Griffin.
Tallis, G. M. (1963) Elliptical and radial truncation in normal samples. *Ann. Math. Statist.*, **34**, 940–944.
Wilks, S. S. (1963) Multivariate statistical outliers. *Sankhya* A, **25**, 407–426.
Wisnowski, J. W., Montgomery, D. C. and Simpson, J. R. (2001) A comparative analysis of multiple outlier detection procedures in the linear regression model. *Computnl Statist. Data Anal.*, **36**, 351–382.