Monitoring Random Start Forward Searches for Multivariate Data

Anthony C. Atkinson[1], Marco Riani[2], and Andrea Cerioli[2]

[1] Department of Statistics, London School of Economics
London WC2A 2AE, UK, *a.c.atkinson@lse.ac.uk*

[2] Dipartimento di Economia, Università di Parma
43100 Parma, Italy, *mriani@unipr.it*, *andrea.cerioli@unipr.it*

Abstract. During a forward search from a robustly chosen starting point the plot of maximum Mahalanobis distances of observations in the subset may provide a test for outliers. This is not the customary test. We obtain distributional results for this distance during the search and exemplify its use. However, if clusters are present in the data, searches from random starts are required for their detection. We show that our new statistic has the same distributional properties whether the searches have random or robustly chosen starting points.

Keywords: clustering, Mahalanobis distance, order statistic, outlier detection, robustness

1 Introduction

The forward search is a powerful general method for detecting systematic or random departures from statistical models, such as those caused by outliers and the presence of clusters. The forward search for multivariate data is given book-length treatment by Atkinson, Riani and Cerioli (2004). To detect outliers they study the evolution of Mahalanobis distances calculated during a search through the data that starts from a carefully selected subset of observations. More recently Atkinson and Riani (2007) suggested the use of many searches starting from random starting points as a tool in the detection of clusters. An important aspect of this work is the provision of bounds against which to judge the observed values of the distances. Atkinson and Riani (2007) use simulation for this purpose as well as providing approximate numerical values for the quantiles of the distribution.

These theoretical results are for the minimum Mahalanobis distance of observations not in the subset used for fitting when the starting point of the search is robustly selected. In this paper we consider instead the alternative statistic of the maximum Mahalanobis distance amongst observations in the subset. We derive good approximations to its distribution during the forward search and empirically compare its distribution to that of the minimum distance, both for random and robust starts. We find for the maximum distance, but not for the minimum, that the distribution of the distance does not depend on how the search starts. Our ultimate purpose is a more automatic method of outlier and cluster identification.

We start in §2 with an introduction to the forward search that emphasises the importance of Mahalanobis distances in outlier detection. Some introductory theoretical results for the distributions of distances are in §3. Section 4 introduces the importance of random start searches in cluster detection. Our main theoretical results are in §5 where we use results on order statistics to derive good approximations to the distribution of the maximum distance during the search. Our methods are exemplified in §6 by the analysis of data on horse mussels. The comparison of distributions for random and elliptical starts to the search is conducted by simulation in §7.

2 Mahalanobis distances and the forward search

The tools that we use for outlier detection and cluster identification are plots of various Mahalanobis distances. The squared distances for the sample are defined as

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

where $\hat{\mu}$ and $\hat{\Sigma}$ are estimates of the mean and covariance matrix of the n observations.

In the forward search the parameters μ and Σ are replaced by their standard unbiased estimators from a subset of m observations, yielding estimates $\hat{\mu}(m)$ and $\hat{\Sigma}(m)$. From this subset we obtain n squared Mahalanobis distances

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

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 $S(m+1)$. In what we call 'normal progression' this process augments the subset by one observation, but sometimes two or more observations enter as one or more leave.

In our examples we look at forward plots of quantities derived from the distances $d_i(m)$. These distances tend to decrease as n increases. If interest is in the latter part of the search we may use **scaled** distances

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

where v is the dimension of the observations y and $\hat{\Sigma}(n)$ is the estimate of Σ at the end of the search.

To detect outliers Atkinson et al. (2004) and Atkinson and Riani (2007) examined the minimum Mahalanobis distance amongst observations not in the subset

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

or its scaled version $d^{\mathrm{sc}}_{\min}(m)$. In either case let this be observation $i_{\min}(m)$. If observation $i_{\min}(m)$ is an outlier relative to the other m observations, the distance (4) will be large compared to its reference distribution.

In this paper we investigate instead the properties of the maximum Mahalanobis distance amongst the m observations in the subset

$$d_{\max}(m) = \max d_i(m) \quad i \in S(m), \tag{5}$$

letting this be observation $i_{\max}(m)$. Whether we monitor $d_{\max}(m)$ or $d_{\min}(m)$ the search is the same, progressing through the ordering of $d_i^2(m)$.

3 Minimum and maximum Mahalanobis distances

We now consider the relationship between $d_{\min}(m)$ and $d_{\max}(m)$ as outlier tests. This relationship depends on the subsets $S(m)$ and $S(m+1)$.

Let the kth largest ordered Mahalanobis distance be $d_{[k]}(m)$ when estimation is based on the subset $S(m)$. In normal progression

$$d_{[m+1]}(m) = d_{\min}(m) \tag{6}$$

and $S(m+1)$ is formed from $S(m)$ by the addition of observation $i_{\min}(m)$. Likewise, in normal progression this will give rise to the largest distance within the new subset, that is

$$d_{[m+1]}(m+1) = d_{\max}(m+1) = d_{i\min(m)}(m+1). \tag{7}$$

The distance $d_{i\min(m)}(m+1)$ is that for the new observation $i_{\min}(m)$ when the parameters are estimated from $S(m+1)$. The consequence of (7) is that both $d_{\max}(m+1)$ and $d_{\min}(m)$ are tests of the outlyingness of observation $i_{\min}(m)$.

Although both statistics are testing the same hypothesis they do not have the same numerical value and should be referred to different null distributions. In §5 we discuss the effect of the ranking of the observations on these distributions as well as the consequence of estimating μ and Σ from a subset of the observations. For estimates using all n observations $d_{\max}(n)$ is one of the distances in (1). Standard distributional results in, for example, Atkinson et al. (2004, §2.6) show that

$$d_i^2(\hat{\mu}, \hat{\Sigma}) = d_i^2(n) \sim \frac{(n-1)^2}{n} \mathrm{Beta}\left(\frac{v}{2}, \frac{n-v-1}{2}\right). \tag{8}$$

On the other hand, $d^2_{\min}(n-1)$ is a deletion distance in which the parameters are estimated, in general, with the omission of observation i. The distribution of such distances is

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

although the distribution of $d^2_{\text{min}}(n-1)$ depends on the order statistics from this distribution. For moderate n the range of the distribution of $d^2_i(n)$ in (8) is approximately $(0, n)$ rather than the unbounded range for the F distribution of the deletion distances. As we shall see, the consequence is that the distribution of $d^2_{\text{max}}(m)$ has much shorter tails than that of $d^2_{\text{min}}(m)$, particularly for small m.

Our argument has been derived assuming normal progression. This occurs under the null hypothesis of a single multivariate normal population when there are no outliers or clusters in the data, so that the ordering of the observations by closeness to the fitted model does not alter appreciably during the search. Then we obtain very similar forward plots of $d_{\text{max}}(m)$ and $d_{\text{min}}(m)$, even if they have to be interpreted against different null distributions. In fact, we do not need the order to remain unchanged, but only that $i_{\text{min}}(m)$ and $i_{\text{max}}(m+1)$ are the same observation and that the other observations in $S(m+1)$ are those that were in $S(m)$. Dispersed outliers likewise do not appreciably affect the ordering of the data. This is however affected by clusters of observations that cause appreciable changes in the parameter estimates as they enter $S(m)$. A discussion of the ordering of observations within and without $S(m)$ is on pp. 68-9 of Atkinson et al. (2004).

4 Elliptical and random starts

To find the starting subset for the search Atkinson et al. (2004) use the robust bivariate boxplots of Zani, Riani and Corbellini (1998) to pick a starting set $S^*(m_0)$ that excludes any two-dimensional outliers. The boxplots have elliptical contours, so we refer to this method as the elliptical start. However, if there are clusters in the data, the elliptical start may lead to a search in which observations from several clusters enter the subset in sequence in such a way that the clusters are not revealed. Searches from more than one starting point are then needed to reveal the clustering structure. If we start with an initial subset of observations from each cluster in turn, the other clusters are revealed as outliers. However, such a procedure is not suitable for automatic cluster detection. Atkinson and Riani (2007) therefore instead run many forward searches from randomly selected starting points, monitoring the evolution of the values of $d_{\text{min}}(m)$ as the searches progress. Here we monitor $d_{\text{max}}(m)$.

As the search progresses, the examples of Atkinson and Riani (2007) show that the effect of the starting point decreases. Once two searches have the same subsets $S(m)$ for some m, they will have the same subsets for all successive m. Typically, in the last third of the search the individual searches from random starts converge to that from the elliptical start. The implication is that the same envelopes can be used, except in the very early stages of the search, whether we use random or elliptical starts. If we are looking for a few outliers, we will be looking at the end of the search. However, the envelopes for $d_{\min}(m)$ and $d_{\max}(m)$ will be different.

5 Envelopes from order statistics

For relatively small samples we can use simulation to obtain envelopes for $d_{\max}(m)$ during the search. For larger samples we adapt the method of Riani, Atkinson and Cerioli (2007) who find very good approximations to the envelopes for $d_{\min}(m)$ using order statistics and a result of Tallis (1963) on truncated multivariate normal distributions.

Let $Y_{[m]}$ be the mth order statistic from a sample of size n from a univariate distribution with c.d.f. $G(y)$. From, for example Lehmann (1991, p. 353) and Guenther (1977), the required quantile of order γ of the distribution of $Y_{[m]}$ say $y_{m,n;\gamma}$ can be obtained as

$$y_{m,n;\gamma} = G^{-1}\left(\frac{m}{m+(n-m+1)x_{2(n-m+1),2m;1-\gamma}}\right), \tag{10}$$

where $x_{2(n-m+1),2m;1-\gamma}$ is the quantile of order $1-\gamma$ of the F distribution with $2(n-m+1)$ and $2m$ degrees of freedom. Riani et al. (2007) comment that care needs to be taken to ensure that the numerical calculation of this inverse distribution is sufficiently accurate as $m \rightarrow n$, particularly for large n and extreme γ.

We now consider our choice of $G(x)$, which is different from that of Riani et al. (2007). We estimate Σ on $m-1$ degrees of freedom. The distribution of the m distances in the subset can, from (8), be written as

$$d_i^2(m) \sim \frac{(m-1)^2}{m} \text{Beta}\left(\frac{v}{2}, \frac{m-v-1}{2}\right), \qquad i \in S(m). \tag{11}$$

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 derived from $d_{\max}(m)$ are given by

$$V_{m,\gamma} = \sqrt{\frac{(m-1)^2}{m}} \sqrt{y_{m,n;\gamma}}, \tag{12}$$

with G the beta distribution in (11).

For **unscaled** distances we need to correct for the bias in the estimate of Σ. We follow Riani et al. (2007) and consider elliptical truncation in the multivariate normal distribution. From the results of Tallis (1963) they obtain the large-sample correction factor

$$c_{FS}(m) = \frac{m/n}{P(X^2_{v+2} < \chi^2_{v,m/n})}, \tag{13}$$

with $\chi^2_{v,m/n}$ the m/n quantile of χ^2_v and X^2_{v+2} a chi-squared random variable on $v+2$ degrees of freedom. Envelopes for unscaled distances are then obtained by scaling up the values of the order statistics

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

Figure 1 shows the agreement between simulated envelopes (continuous lines) and theoretical envelopes (dotted lines) for $d_{\max}(m)$ when $n = 1000$. Scaled distances are in the upper panel; agreement between the two sets of envelopes is excellent throughout virtually the whole range. Agreement for the unscaled distances in the lower panel of the figure is less good, but is certainly more than satisfactory for inferences about outliers at least in the last half of the search.

Unfortunately, the inclusion of $\hat{\Sigma}(n)$ in the expression for scaled distances (3) results in small distances in the presence of outliers, due to the inflation of the variance estimate and to consequent difficulties of interpretation. For practical data analysis we have to use the unscaled distances, which are less well approximated.

6 Horse mussels

As an example of the uses of elliptical and random starts in the analysis of multivariate data we look at measurements on horse mussels from New Zealand introduced by Cook and Weisberg (1994, p. 161) who treat them as regression with muscle mass, the edible portion of the mussel, as response. They focus on independent transformations of the response and of one of the explanatory variables. Atkinson et al. (2004, §4.9) consider multivariate normality obtained by joint transformation of all five variables.

There are 82 observations on five variables: shell length, width, height and mass and the mass of the mussels' muscle, which is the edible part.

We begin with an analysis of the untransformed data using a forward search with an elliptical start. The left-hand panel of Figure 2 monitors $d_{\min}(m)$, whereas the right-hand panel monitors $d_{\max}(m)$. The two sets of simulation envelopes were found by direct simulation of 5,000 forward searches. The figure shows how very different the two distributions are at the beginning of the search. That in the left-hand panel for $d_{\min}(m)$ is derived from the unbounded F distribution (9) whereas that for $d_{\max}(m)$ in the right-hand panel is derived from the beta distribution (11).

The two traces are very similar once they are calibrated by the envelopes. They both show appreciable departure from multivariate normality in the last one third of the search. Since we are selecting observations by their closeness to the multivariate normal model, we expect departure, if any, to be at the end of the search. Even allowing for the scaling of the two plots, the maximum distances seem to show less fluctuation at the beginning of the search. For

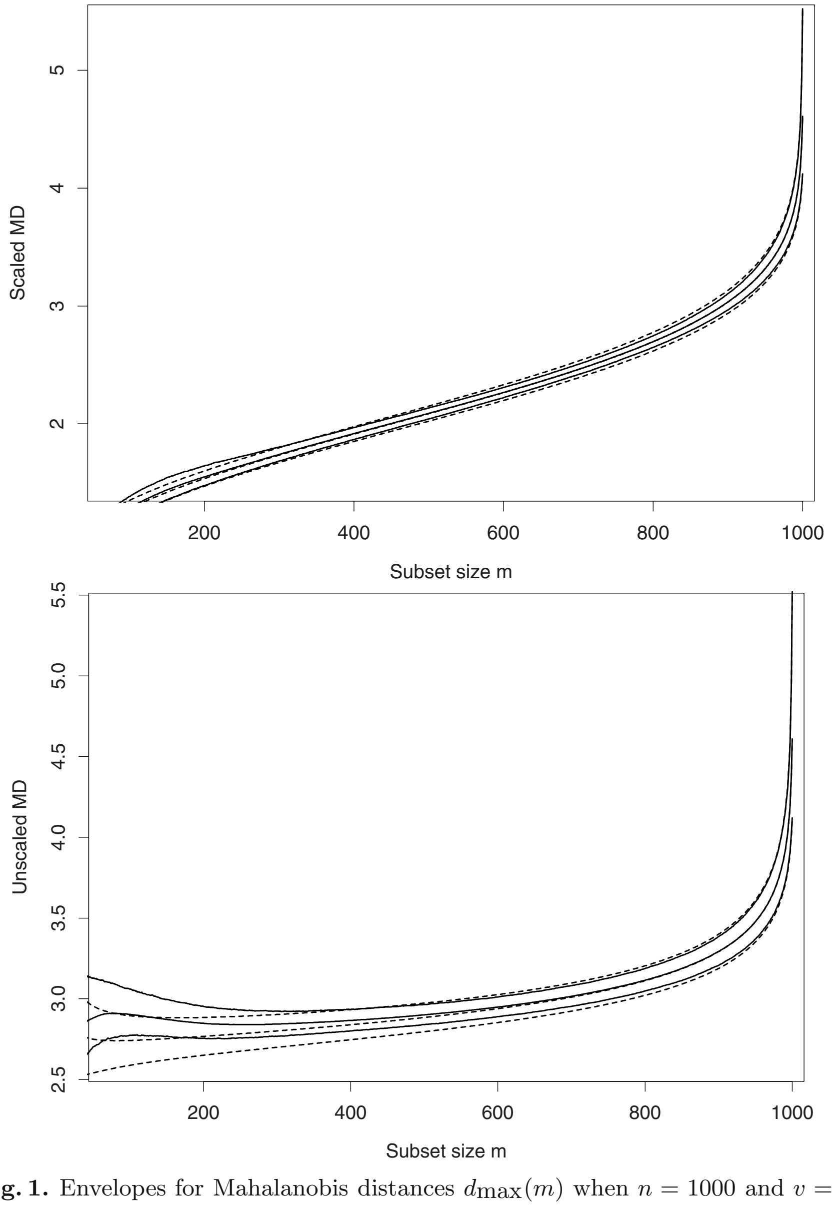

Fig. 1. Envelopes for Mahalanobis distances $d_{\max}(m)$ when $n = 1000$ and $v = 5$. Dotted lines from order statistics, continuous lines from 5,000 simulations. Upper panel scaled distances, lower panel unscaled distances. Elliptical starts. 1%, 50% and 99% points.

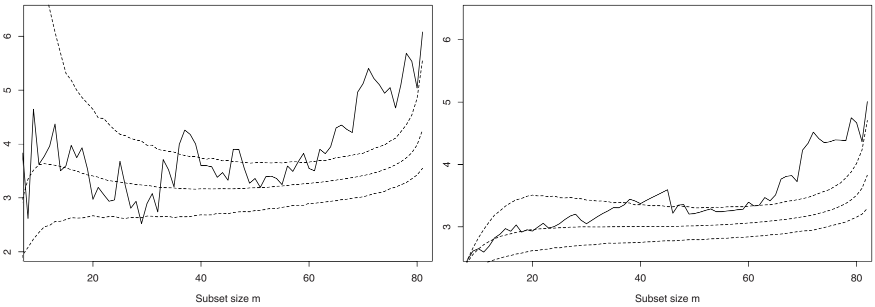

Fig. 2. Horse mussels: forward search on untransformed data. Left-hand panel $d_{\mathrm{min}}(m)$, right-hand panel $d_{\mathrm{max}}(m)$. Elliptical starts; 1%, 50% and 99% points from 5,000 simulations.

much of the rest of the paper we focus on plots of the maximum Mahalanobis distances $d_{\mathrm{max}}(m)$.

We now analyse the data using a multivariate version of the parametric transformation of Box and Cox (1964). As a result of their analysis Atkinson et al. (2004) suggest the vector of transformation parameters $\lambda = (0.5, 0, 0.5, 0, 0)^T$; that is, the square root transformation for y_1 and y_3 and the logarithmic transformation for the other three variables. We look at forward plots of $d_{\mathrm{max}}(m)$ to see whether this transformation yields multivariate normality.

The upper-left panel of Figure 3 shows the maximum distance for all $n = 82$ observations for the transformed data. The contrast with the right-hand panel of Figure 2 is informative. The plot still goes out of the 99% envelope at the end of the search, but the number of outliers is much smaller, now only around 5.

The last five units to enter are those numbered 37, 16, 78, 8 and finally 48. The plot of the maximum distance in the upper-right panel of Figure 3 shows that, with these five observations deleted, the last value just lies below the 99% point of the distribution. We have found a multivariate normal sample, after transformation, with five outliers. That there are five outliers, not four, is confirmed in the lower panel of Figure 3 where observation 37 has been re-included. Now the plot of maximum distances goes outside the 99% envelope at the end of the search.

The limits in figures like 3 have been simulated to have the required pointwise level, that is they are correct for each m considered independently. However, the probability that the observed trace of values of $d_{\mathrm{max}}(m)$ exceeds a specific bound at least once during the search is much greater than the pointwise value. Atkinson and Riani (2006) evaluate such simultaneous probabilities; they are surprisingly high.

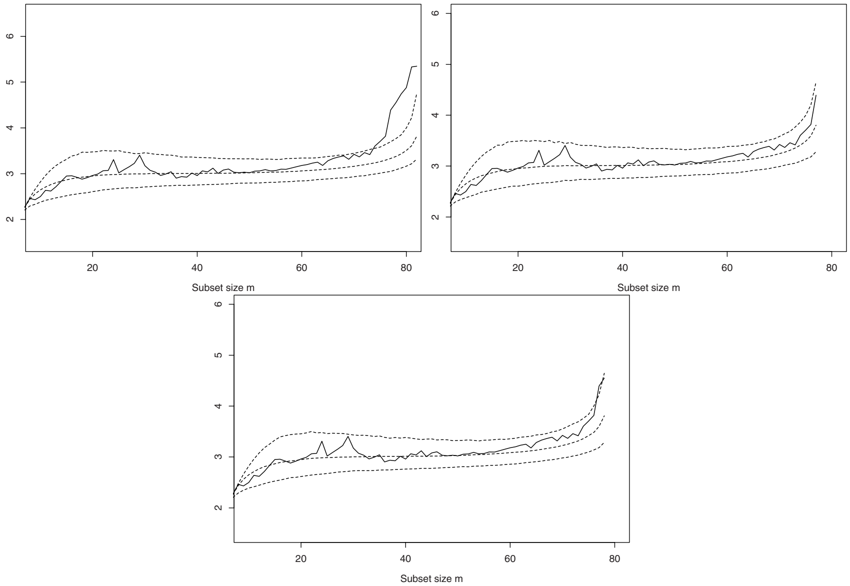

Fig. 3. Horse mussels: forward search on transformed data. Plots of $d_{\max}(m)$ for three different sample sizes. Upper-left panel $n = 82$; upper-right panel, observations 37, 16, 78, 8 and 48 removed ($n = 77$). Lower panel, observation 37 re-included ($n = 78$). Elliptical starts; 1%, 50% and 99% points from 5,000 simulations. There are five outliers.

7 Elliptical and random starts

We have analysed the mussels data and investigated the properties of $d_{\max}(m)$ and $d_{\min}(m)$ using searches with elliptical starts. Finally we look at the properties of plots of both distances when random starts are used, for example to aid in the identification of clusters.

The upper panel of Figure 4 presents a comparison of simulated envelopes for random and elliptical starts for $d_{\max}(m)$ from data with the same dimensions as the mussels data. For this small data set there is no operationally important difference between the two envelopes. The important conclusion is that, for larger data sets, we can use the approximations of §5 based on order statistics whether we are using random or elliptical starts.

The surprising conclusion that we obtain the same envelopes for searches from elliptical or random starts however does not hold when instead we monitor $d_{\min}(m)$. The lower panel of Figure 4 repeats the simulations for $d_{\min}(m)$. Now there is a noticeable difference, during the first half of the search, between the envelopes for random and those from elliptical starts.

We now consider the implications of this difference on the properties of individual searches. The left-hand panel of Figure 5 repeats the simulated envelopes for elliptical starts from the upper panel of Figure 4 and adds 250

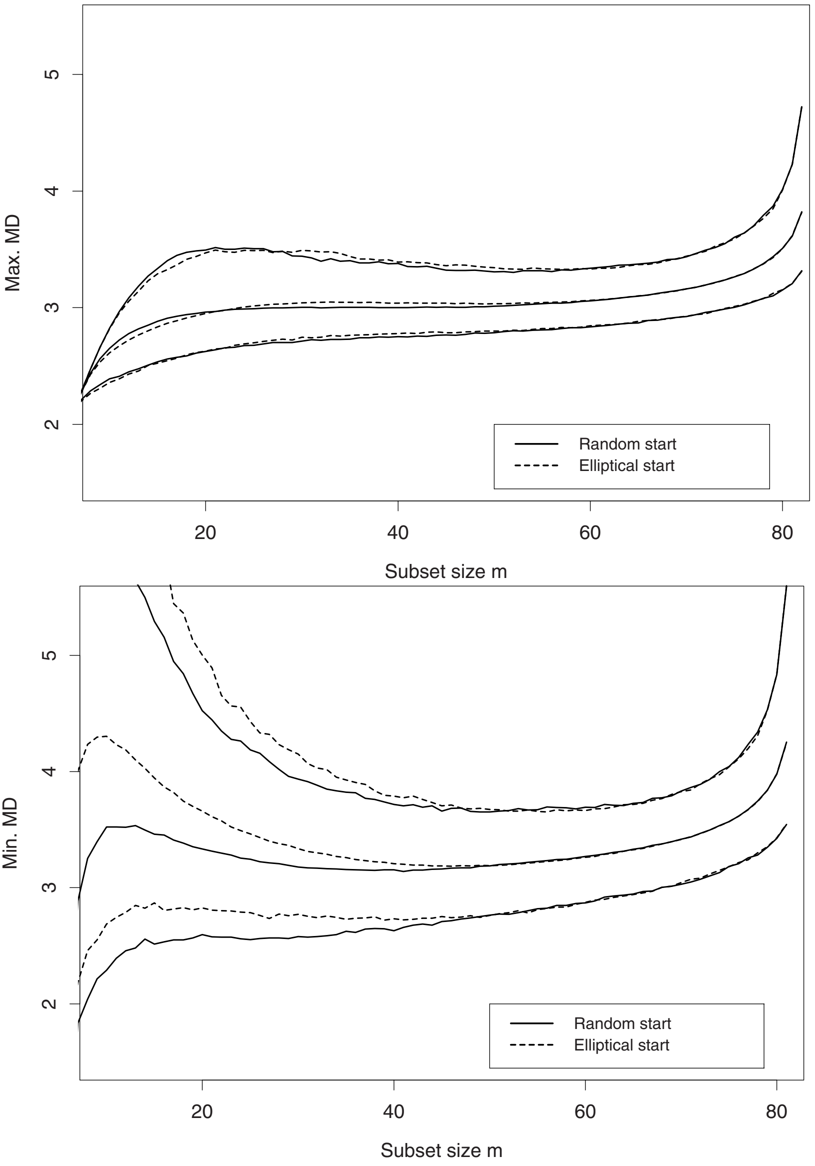

Fig. 4. Simulated envelopes for Mahalanobis distances when $n = 82$ and $v = 5$. Upper panel, $d_{\text{max}}(m)$, lower panel $d_{\text{min}}(m)$ Dotted lines from elliptical starts, continuous lines from random starts. 1%, 50% and 99% points from 5,000 simulations.

trajectories of distances for $d_{\max}(m)$ for simulations from random starting points. The simulated values nicely fill the envelope, although there are a surprising number of transient peaks above the envelope. The right-hand panel of the figure repeats this process of envelopes from elliptical starts and trajectories from random starts but for the minimum distances $d_{\min}(m)$. Now, as we would expect from Figure 4, the simulated values sit a little low in the envelopes. If a subset contains one or more outliers, these will give rise to a too large estimate of Σ. As a consequence, some of the distances of units not included in the subset will be too small and the smallest of these will be selected as $d_{\min}(m)$. On the contrary, if outliers are present in $S(m)$ when we calculate $d_{\max}(m)$, the distance that we look at will be that for one of the outliers and so will not be shrunken due to the too-large estimate of Σ.

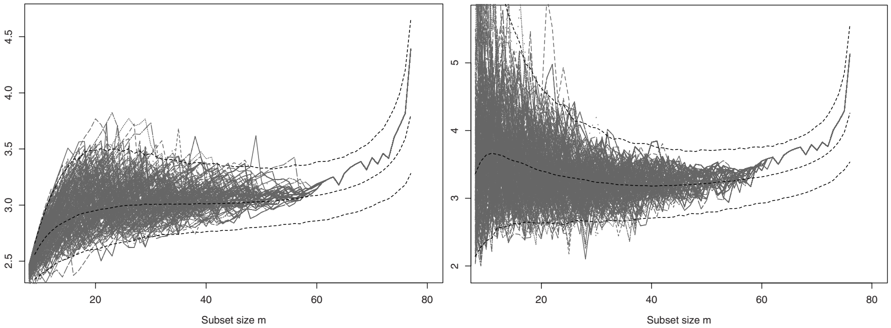

Fig. 5. Simulated envelopes from elliptical starts for Mahalanobis distances when $n = 82$ and $v = 5$ with 250 trajectories from random starts. Left-hand panel $d_{\max}(m)$, right-hand panel $d_{\min}(m)$.

Our justification for the use of random start forward searches was that searches from elliptical starts may not detect clusters in the data if these start from a subset of units in more than one cluster. We have however analysed the mussels data using values of $d_{\max}(m)$ from elliptical starts. Our conclusion, from Figure 3, was that after transformation there were 77 units from a multivariate normal population and five outliers. We checked this conclusion using random start forward searches with $d_{\max}(m)$ and failed to detect any clusters.

The purpose of this paper has been to explore the properties of the maximum distance $d_{\max}(m)$. We have found its null distribution and obtained good approximations to this distribution for use in the forward search.The lack of dependence of this distribution on the starting point of the search is an appealing feature. However, we need to investigate the properties of this measure when the null distribution does not hold. One particular question is whether use of $d_{\max}(m)$ provides tests for outliers and clusters that are as powerful as those using the customary minimum distance $d_{\min}(m)$.

References

ATKINSON, A.C. and RIANI, M. (2000): *Robust Diagnostic Regression Analysis.* Springer-Verlag, New York.

ATKINSON, A.C. and RIANI, M. (2006): Distribution theory and simulations for tests of outliers in regression. *Journal of Computational and Graphical Statistics 15, 460–476.*

ATKINSON, A.C. and RIANI, M. (2007): Exploratory tools for clustering multivariate data. *Computational Statistics and Data Analysis 52, 272-285.* doi:10.1016/j.csda.2006.12.034.

ATKINSON, A.C., RIANI, M. and CERIOLI, A (2004): *Exploring Multivariate Data with the Forward Search.* Springer-Verlag, New York.

BOX, G.E.P. and COX, D.R. (1964): An analysis of transformations (with discussion). *Journal of the Royal Statistical Society, Series B 26, 211-246.*

COOK, R.D. and WEISBERG, S. (1994): *An Introduction to Regression Graphics.* Wiley, New York.

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

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

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.

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

ZANI, S., RIANI, M. and CORBELLINI, A. (1998): Robust bivariate boxplots and multiple outlier detection. *Computational Statistics and Data Analysis 28, 257-270.*