Random Start Forward Searches with Envelopes for Detecting Clusters in Multivariate Data

Anthony Atkinson[1]
Department of Statistics
London School of Economics
a.c.atkinson@lse.ac.uk

Marco Riani & Andrea Cerioli [2]
Dipartimento di Economia
Università di Parma
mriani@unipr.it or statec1@ipruniv.cce.unipr.it

Abstract: During a forward search the plot of minimum Mahalanobis distances of observations not in the subset provides a test for outliers. However, if clusters are present in the data, their simple identification requires that there are searches that initially include a preponderance of observations from each of the unknown clusters. We use random starts to provide such searches, combined with simulation envelopes for precise inference about clustering.

Keywords: Classification, diagnostics, elliptical start, Mahalanobis distance, outliers, random start

1 Introduction

The forward search is a powerful general method for detecting unidentified subsets and multiple masked outliers and for determining their effect on models fitted to the data. The search for multivariate data is given book length treatment by Atkinson *et al.* (2004). To detect clusters they use forward searches starting from subsets of observations in tentatively identified clusters. The purpose of this paper is to demonstrate the use of randomly selected starting subsets for cluster detection that avoid any preliminary data analysis. The goal is a more automatic method of cluster identification.

2 Mahalanobis Distances and the Forward Search

The main tools that we use are plots 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 estimates of the mean and covariance matrix of the n observations.

[1]London WC2A 2AE, UK
[2]43100 Parma, Italy

In the forward search the parameters μ and Σ are estimated by maximum likelihood applied to 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 $S(m+1)$. Usually 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 we examine 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 $d_{\min}(m)^{\mathrm{sc}}(m)$. In either case let this be observation $i_{\min}$. If observation $i_{\min}$ is an outlier relative to the other m observations, the distance (4) will be large compared to the maximum Mahalanobis distance of the m observations in the subset.

3 Minimum and Ordered Mahalanobis Distances

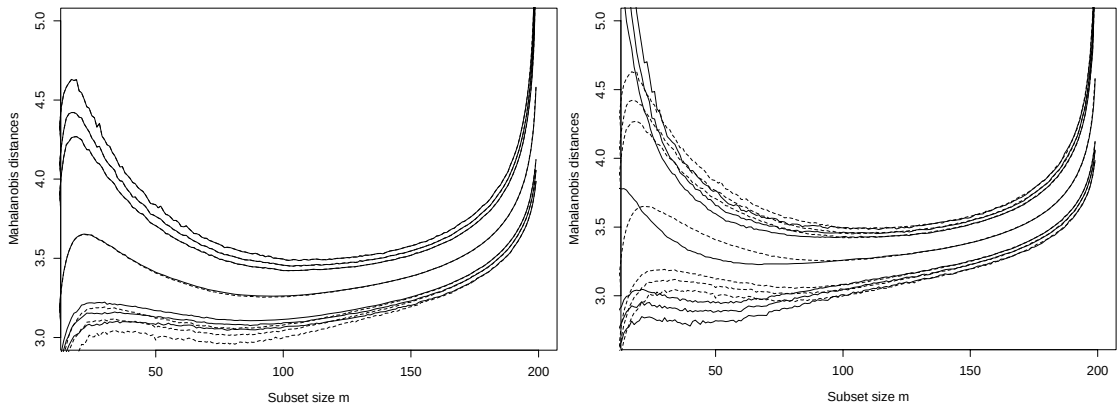
Now consider the ordered Mahalanobis distances with $d_{[k]}(m)$ the kth largest distance when estimation is based on the subset $S(m)$. In many, but not necessarily all, steps of the search

$$d_{[m+1]}(m) = d_{\min}(m). \tag{5}$$

Instead of using $d_{\min}(m)$ as an outlier test, we could use the value of $d_{[m+1]}(m)$. In this section we describe when the difference in the two distances can arise and what the lack of equality tells us about the presence of outliers or clusters in the data. We then use simulation to compare the null distribution of tests in the forward search based on the two distances.

Lack of equality in (5) can arise because the observations in $S(m)$ come from ordering the n distances $d_i(m-1)$ based on $S(m-1)$ not on $S(m)$. The effect is most easily understood by considering the case when the observation added in going from $S(m-1)$ to $S(m)$ is the first in a cluster of outliers. In that case the parameter estimates $\hat{\mu}(m)$ and $\hat{\Sigma}(m)$ may be sufficiently different from $\hat{\mu}(m-1)$ and $\hat{\Sigma}(m-1)$

Figure 1: *Envelopes from 10,000 simulations of forward searches with multivariate normal data when $n = 200$ and $v = 6$. Left-hand panel - elliptical starts: continuous lines, the order statistic $d_{[m+1]}(m)$; dotted lines, $d_{min}(m)$, the minimum distance amongst observations not in the subset. Right-hand panel - plots of $d_{min}(m)$: dotted lines, elliptical starts as in the left-hand panel; continuous lines, random starts. 1, 2.5, 5, 50, 95, 97.5 and 99 % envelopes*



that the other observations in the cluster will seem less remote. Indeed, some may have smaller distances than some of those in the subset. More formally, we will have

$$d_{\min}(m) < d_{[k]}(m) \quad k \leq m, \tag{6}$$

for one or more values of k. Then the difference

$$g_1(m) = d_{\min}(m) - d_{[m]}(m) \tag{7}$$

will be negative, whereas when (5) holds, which it typically does in the absence of outliers,

$$g_2(m) = d_{[m+1]}(m) - d_{[m]}(m) = d_{\min}(m) - d_{[m]}(m) \tag{8}$$

is positive. The forward plot of $g_1(m)$ and $g_2(m)$ is called a gap plot, appreciable differences between the two curves indicating the entry of a group of outliers or of a new cluster of observations into the subset. At such moments interchanges may occur when one or more of the observations in $S(m)$ leave the subset as two or more enter to form $S(m+1)$. A more detailed discussion of the ordering of observations within and without $S(m)$ is on pp. 68-9 of Atkinson *et al.* (2004). The gap plot for the Swiss banknote data, which contains two clusters, is on p.118.

The above argument suggests that, for a single multivariate population with no outliers, (5) will hold in most steps of the search and that use of $d_{[m+1]}(m)$ or of $d_{\min}(m)$ as an outlier test will give identical results. To demonstrate this we show in the left-hand panel of Figure 1 forward plots of simulated percentage points of the empirical distribution of the unscaled versions of the two quantities from 10,000 simulations of 200 observations from a six-dimensional normal distribution. The continuous curves are for $d_{[m+1]}(m)$, whereas $d_{\min}(m)$ is represented by dotted lines. There is no discernible difference over the whole search in the median and upper percentage points of the distribution. There is some difference in the lower

percentage points where the average values of $d_{\min}(m)$ are slightly lower. This is explained because, in the earlier stages of the search there are a few samples in which the observations are not well ordered and the subset is unstable, so that condition (6) holds. That the difference between the two distributions is only in the lowest tails shows that such behaviour is comparatively rare. Since we use the upper tails of the distribution for detection of outliers, the figure confirms that the test is indifferent to the use of $d_{[m+1]}(m)$ or of $d_{\min}(m)$. In the remainder of this paper we only consider the minimum distances $d_{\min}(m)$.

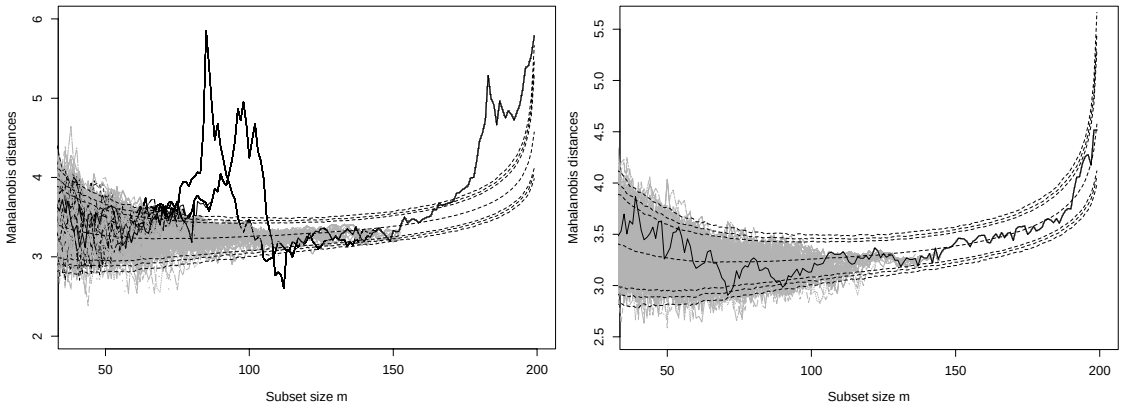
4 Elliptical and Random Starts

When the observations come from a single multivariate normal population with some outliers, these outlying observations enter at the end of the search. To start the search under these conditions Atkinson *et al.* (2004) use the robust bivariate boxplots of Zani *et al.* (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 in fact needed to reveal the clustering structure. Typically it is necessary to start with an initial subset of observations from each cluster in turn, when the other clusters are revealed as outliers. An example using the data on Swiss banknotes is in Chapter 1 of Atkinson *et al.* (2004). In this example finding initial subsets in only one of the two clusters requires a preliminary analysis of the data. Such a procedure is not suitable for automatic cluster detection. We therefore instead run many forward searches from randomly selected starting points, monitoring the evolution of the values of $d_{\min}(m)$ as the searches progress.

In order to interpret the results of such plots we again need simulation envelopes. The right-hand panel of Figure 1 repeats, in the form of dotted lines, the envelopes for $d_{\min}(m)$ from the left-hand panel, that is with elliptical starts. The continuous lines in the figure are for the values of $d_{\min}(m)$ from random starts. At the start of the search the random start produces some very large distances. But, almost immediately, the distances for the random start are smaller, over the whole distribution, than those from the elliptical start. This is because the elliptical start leads to the early establishment of subsets $S(m)$ from the centre of the distribution. But, on the other hand, the subsets $S_{\mathrm{R}}(m)$ from the random start may contain some observations not from the centre of the distribution. As a consequence, the estimate of variance will be larger than that from the elliptical start and the distances to all units will be smaller. As the search progresses, this effect decreases as the $S_{\mathrm{R}}(m)$ for individual searches converge to the $S(m)$ from the elliptical start. As the figure shows, from just below $m = 100$ there is no difference between the envelopes from the two searches. Further, for appreciably smaller values of m inferences about outliers from either envelope will be similar.

The results of this section lead to two important simplifications in the use of envelopes in the analysis of multivariate normal data. One is that procedures based on either $d_{[m+1]}(m)$ or on $d_{\min}(m)$ are practically indistinguishable. The other is that the same envelopes can be used, except in the very early stages of the search,

Figure 2: *Forward plots of $d_{min}(m)$ for 500 searches with random starting points. Left-hand panel, Swiss banknote data showing two groups and outliers; the searches shown in grey always contain units from both groups. Right-hand panel, Swiss heads data, a homogeneous sample. An arbitrarily selected search is shown in black. 1, 2.5, 5, 50, 95, 97.5 and 99 % envelopes from 10,000 simulations*



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. If we are detecting clusters, their confirmation involves searches of only the cluster members so that, as we see in §6, we are again looking only at the end of the search.

5 Swiss Banknotes and Swiss Heads

There are two hundred observations in the Swiss banknote data. The notes have been withdrawn from circulation and contain 100 notes believed to be genuine and 100 probable forgeries, on each of which six measurements were made. The left-hand panel of Figure 2 contains the results of 500 forward searches from randomly selected starting subsets with $m_0 = 10$. For each search we have plotted the outlier test $d_{\min}(m)$, the minimum unscaled Mahalanobis distance amongst observations not in the subset. Also included in the plot are 1, 2.5, 5, 50, 95, 97.5 and 99 % simulation envelopes for $d_{\min}(m)$ when the observations come from a single six-dimensional normal distribution.

The first feature of the plot is that, from m around 150, all searches follow the same trajectory, regardless of starting point. This is empirical justification of the assertion of Atkinson *et al.* (2004) that the starting point is not of consequence in the latter part of the search. The end of the search shows a group of 20 outliers, most of which, in fact, come from Group 2, the forgeries (there seem to have been two forgers at work). The peak around $m = 98$ is for searches containing only units from Group 1. At these values of m the outliers from Group 1 and observations from Group 2 are all remote and have large distances. Because of the larger number of outliers from Group 2, the peak for this cluster comes earlier, around $m = 85$. The searches that do not give rise to either peak always contain units from both clusters and are non-informative about cluster structure. They are shown in grey in the figure.

This plot shows the clear information that can be obtained by looking at the data from more than one viewpoint. It also shows how quickly the search settles down: the first peak contains 70 searches and the second 62. Fewer searches than this will have started purely in one cluster; because of the way in which units are included and excluded from the subset, the searches tend to produce subsets located in one or other of the clusters.

The left-hand panel of the figure can indeed be interpreted as revealing the clusters. But we also need to demonstrate that we are not finding structure where none exists. The right-hand panel of the figure is again a forward plot of the minimum distance of observations not in the subset, but this time for the 200 observations of six-dimensional data on the size of Swiss heads also analysed by Atkinson *et al.* (2004). This plot shows none of the structure of clustering that we have found in the banknote data. It however does show again how the search settles down in the last one third, regardless of starting point.

The plot in the left-hand panel of Figure 2 leads to the division of the data into two clusters, the units in the subsets just before the two peaks. Once the data have been dissected in this way, the procedures described in Atkinson *et al.* (2004) can be used to explore and confirm the structure. For example, their Figure 3.30 is a forward plot of all 200 Mahalanobis distances when the search starts with 20 observations on genuine notes; in Figure 3.35 the search starts with 20 forgeries. In both these plots, which are far from identical, the structure of two groups and some outliers is evident. However, in their Figure 3.28, in which the search starts with a subset of units from both groups, there is no suggestion of the group structure.

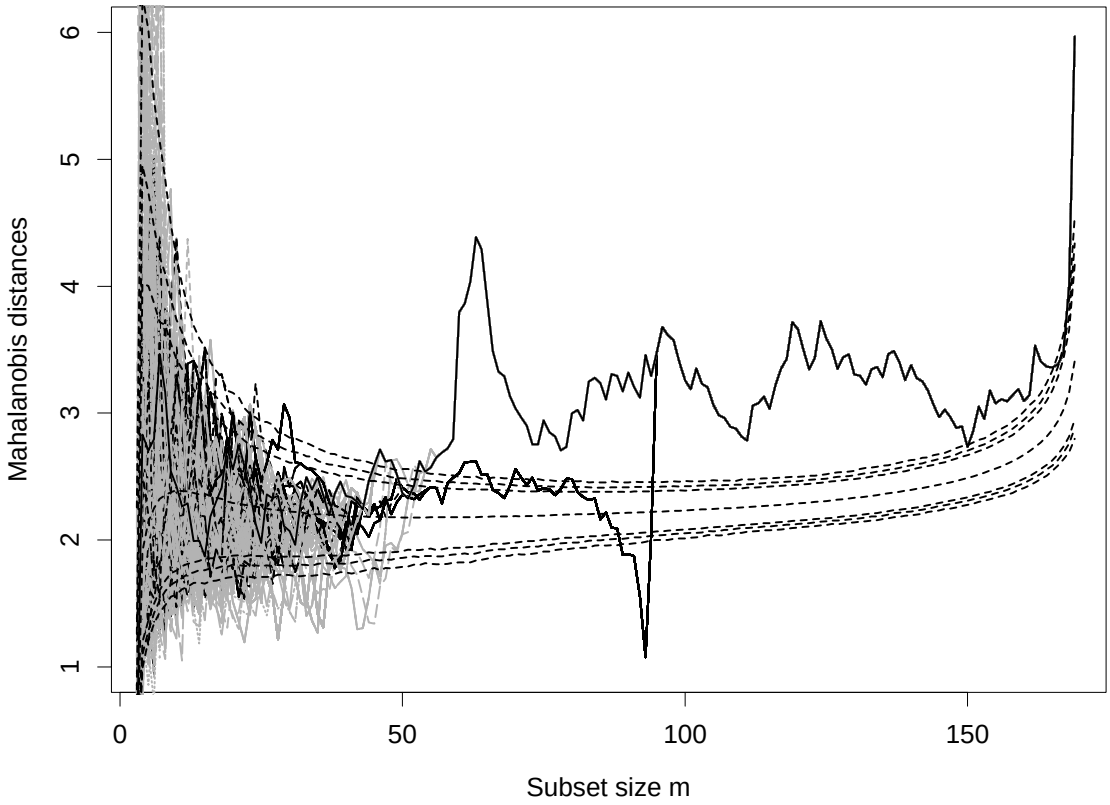
6 Bridge Data

In their §7.5 Atkinson *et al.* (2004) introduce the "bridge" data; 170 two-dimensional observations that consist of a dispersed cluster of 80 observations, a separate tight cluster of 60 observations and an intermediate bridge joining the two groups consisting of 30 observations. The data are plotted in their Figure 7.18. An important feature of these data without the bridge is that the cluster structure is not detected by very robust statistical methods which fit to a subsample coming from both clusters and so fail to reveal the clustered nature of the data. We first use these data as a second example of the power of random starts to indicate clustering. We then show how the repeated use of envelopes for varying sample sizes n can lead to the virtually exact determination of cluster membership.

Figure 3 shows plots of $d_{\min}(m)$ for 500 searches with random starting points. The general structure is that, from around $m = 50$, there are two trajectories. The upper one, which at this point contains units from the compact group of 60 observations, has a peak at $m = 61$. The lower trajectory initially contains units from the dispersed group of 80 observations and then, for larger m, neighbouring units from the bridge are included. There follows a large interchange of units when most of those from the dispersed group are removed from the subset and, from $m = 95$, both trajectories are the same; the subset subsequently grows by inclusion of units from the dispersed group.

We now consider a careful analysis of the trajectory of $d_{\min}(m)$ using subsets of the data of increasing sizes identified from Figure 3 as giving the upper trajectory,

Figure 3: *"Bridge" data: $d_{min}(m)$ for 500 searches with random starting points. The peak at $m = 61$ comes from searches that contain only units in the compact group. The other main trajectory is for searches based on the dispersed group*

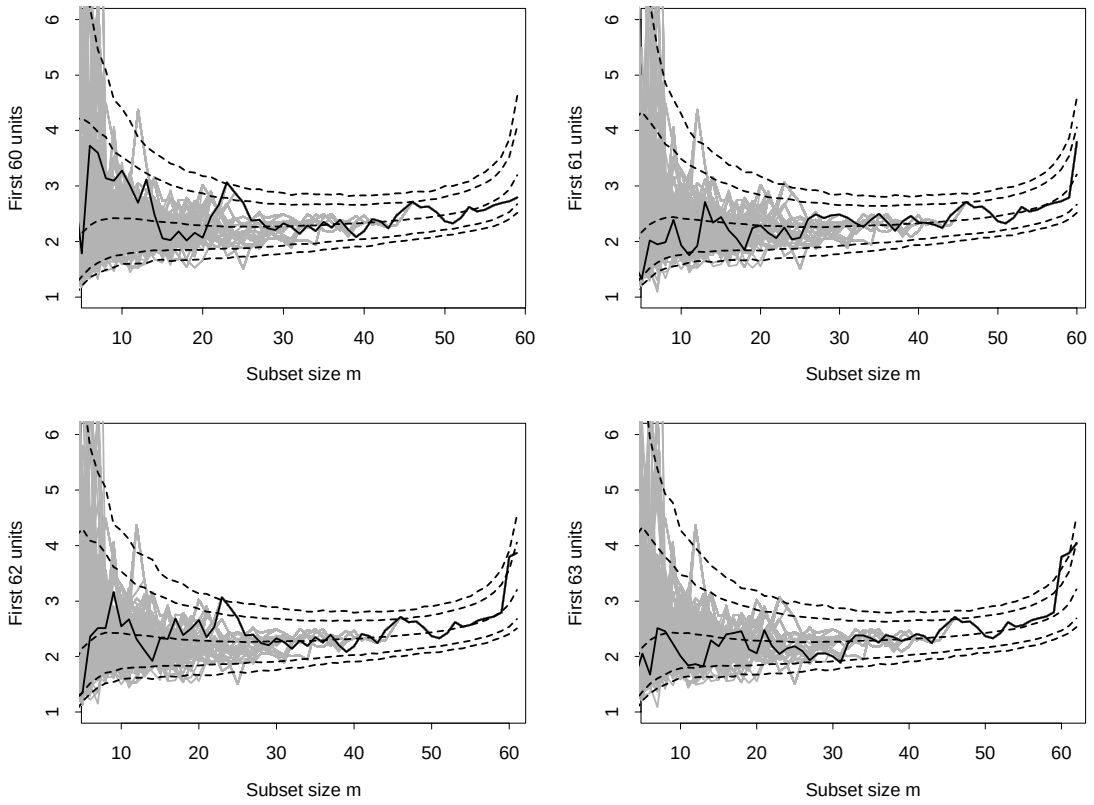


that is seemingly coming from the compact group. To discuss individual observations we use the ordering imposed by the forward search, notated as observation $[i]$. The top-left panel of Figure 4 is for 500 searches with random starts using the first 60 units to enter the subset, so the envelopes, which stop at $m = 59$ are found by simulations with $n = 60$. The trajectory lies within the simulated envelopes; there is no evidence of any outlier. In fact the trajectory is a little too flat at the end, as though a large, but not outlying, observation or two has been incorrectly excluded.

In fact, the first 60 observations in the search consist of 59 from Group 1 and one from the bridge. Of course, because the data are simulated with random error, group membership can be overlapping. The upper-right panel in the figure is for $n = 61$, with new simulation envelopes from this value of n. It is important in the exact detection of outliers that the upwards curve towards the end of the search is sensitive to sample size. Addition of observation [61] (observation 118 in Table A.15 of Atkinson *et al.*, 2004) causes an upwards jump in the trajectory, although not a sufficiently large jump to take the trajectory outside the envelopes.

Observation [61] is the last from Group 1. Addition of observation [62] (161), shown in the lower-left panel of the figure, takes observation [61] partially outside the envelopes, although observation [62] remains inside. Finally, the plot for the first 63 observations shows observation [61] well outside all envelopes, with the trajectory returning inside the envelope.

Figure 4: *"Bridge" data: $d_{min}(m)$ for 500 searches with random starting points for n = 60, 61, 62 and 63 observations giving the upper trajectory in Figure 3. The first 60 observations are shown to belong to a homogenous group. 1, 5, 50, 95, and 99 % envelopes*



The behaviour of the trajectory in the lower plots is typical of the effect of adding a cluster of different observations from those in the subset earlier in the search, which was discussed in §3. What these plots do show is that observation [61] is indeed an outlier and that the first 60 units form a homogeneous group. The next stage in the analysis would be to remove these sixty observations and to run further series of searches with random starts to identify any remaining structure.

7 The Importance of Envelopes

The analysis in this paper has depended crucially on the use of simulation envelopes in the forward search, a feature missing from our books Atkinson and Riani (2000) and Atkinson *et al.* (2004). The envelopes used here are similar to those described by Riani and Atkinson (2007) for testing for outliers in multivariate normal data. Since they are looking for outliers from a single population, they only use elliptical starts, rather than the random starts we use here to detect clusters. For large samples in high dimensions, the repeated simulation of envelopes for increasing sample sizes, as in §6, can be excessively time consuming. Riani and Atkinson (2007) describe methods for numerical approximation to the envelopes, particularly for the

scaled distances (3). They also give theoretical results, based on order statistics from scaled F distributions, that give excellent approximations to the envelopes, even for moderate n and v.

References

Atkinson A.C. and Riani M. (2000) *Robust Diagnostic Regression Analysis*, Springer–Verlag, New York.

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

Riani M. and Atkinson A.C. (2007) Finding an unknown number of multivariate outliers in larger data sets, (Submitted).

Zani S., Riani M. and Corbellini A. (1998) Robust bivariate boxplots and multiple outlier detection, *Computational Statistics and Data Analysis*, 28, 257–270.