Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 52 (2007) 272–285
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Exploratory tools for clustering multivariate data

A.C. Atkinson[a,*], M. Riani[b]

[a] *Department of Statistics, The London School of Economics, London WC2A 2AE, UK*
[b] *Sezione di Statistica e Informatica, Dipartimento di Economia, Università di Parma, Italy*

Available online 28 December 2006

Abstract

The forward search provides a series of robust parameter estimates based on increasing numbers of observations. The resulting series of robust Mahalanobis distances is used to cluster multivariate normal data. The method depends on envelopes of the distribution of the test statistics in forward plots. These envelopes can be found by simulation; flexible polynomial approximations to the envelopes are given. New graphical tools provide methods not only of detecting clusters but also of determining their membership. Comparisons are made with `mclust` and *k*-means clustering.

Keywords: Forward search; Graphics; *k*-means; `mclust`; Mahalanobis distance; Random start; Simulation envelopes

1. Introduction

We describe procedures based on robust Mahalanobis distances that identify and confirm separated and overlapping clusters in multivariate normal data. Our procedure is related to outlier detection in multivariate data since, from the perspective of a model fitted to one cluster, members of other clusters appear as outliers.

The use of robust methods for the detection of outliers was popularised by Rousseeuw and Leroy (1987). Rousseeuw and van Zomeren (1990) describe a robust method with exploratory graphics for the detection of outliers in multivariate normal data and Rousseeuw and Van Driessen (1999) provide a fast algorithm for the calculation of the robust distances. The robust estimates of means and covariances that we use in the calculation of the Mahalanobis distances in this paper come from the forward search, in which subsamples of increasing size are used for parameter estimation. The resulting Mahalanobis distances can be plotted as a function of sample size. Such a 'forward' plot provides a powerful method of diagnosing a wide variety of structures in multivariate data. Versions of the forward search for the detection of multiple outliers were introduced by Hadi (1992) and Atkinson (1994). Many applications of the forward search in the analysis of multivariate data are given by Atkinson et al. (2004). Their analyses typically rely on forward plots of only the robust distances. However, in order to provide sensitive inferences about clusters, it is necessary to augment their plots with envelopes of the distributions of the statistics being plotted. These envelopes can be found by simulation, which will be time consuming if outliers are dropped sequentially, so that envelopes are required for a series of sample sizes, as they are in our procedure for establishing cluster membership in Section 7. Approximations to the envelopes may therefore be required.

* Corresponding author.
E-mail addresses: a.c.atkinson@lse.ac.uk (A.C. Atkinson), mriani@unipr.it (M. Riani).

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.12.034

For outlier detection Atkinson et al. (2004) use single forward searches from a robustly chosen starting point. The forward search and the outlier test are the subject of Section 2; we give plots of the envelopes for the distribution of the test statistic in Section 3 and find polynomial approximations to the envelopes in Section 4. For cluster identification, as opposed to outlier identification, several searches are needed, the most informative being those that start in individual clusters and continue to add observations from the cluster until most observations in that cluster have been used in estimation. There is then usually a clear change in the Mahalanobis distances as units from other clusters enter the subset used for estimation. This strategy seemingly requires that we know the clusters, at least approximately, before running the searches. But we instead use many searches with random starting points to provide information on cluster existence and membership. We discuss random starts in Section 5 and argue that one set of simulation envelopes is appropriate both for searches with a random start and those in which the starting point is selected robustly.

We then employ our polynomial approximation to the envelopes to analyse two examples. We introduce our exploratory approach to clustering in Section 6 on a simulated example with 1000 observations. We use our envelopes for varying sample sizes in Section 7 in a more detailed analysis of data on the Old Faithful geyser and, in Section 8, compare our results with two other methods of clustering, `mclust` and k means. Like our method, `mclust` fits a mixture of normal distributions to the data. See, for example McLachlan and Peel (2000). However, unlike these standard methods, the forward search leads immediately to procedures for checking cluster membership.

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 of n v-dimensional observations 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 forward search the parameters μ and Σ are estimated from a subset $S(m)$ of m of the n observations $Y^{n \times v}$, with element y_{ij}. The parameter estimates are $\hat{\mu}(m)$ with

$$\hat{\mu}(m)_j = \sum_{i \in S(m)} y_{ij}/m, \quad j = 1, \ldots, v \tag{2}$$

and $\hat{\Sigma}(m)$ where

$$\hat{\Sigma}(m)_{jk} = \sum_{i \in S(m)} \{y_{ij} - \hat{\mu}(m)_j\}\{y_{ik} - \hat{\mu}(m)_k\}/(m-1), \quad j, k = 1, \ldots, v. \tag{3}$$

From this subset we obtain n squared Mahalanobis distances

$$d_i^2(m) = \{y_i - \hat{\mu}(m)\}^{\mathrm{T}} \hat{\Sigma}^{-1}(m) \{y_i - \hat{\mu}(m)\}, \quad i = 1, \ldots, n, \tag{4}$$

for each value m.

The steps of the search are:

(1) *Choice of the initial subset.* To start the search when the observations are assumed to come from a single multivariate normal population with some outliers, 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. Other robust methods can be used to identify the starting point. Then one search is run from this unique robust starting point.

(2) *Adding observations.* The subset $S(m)$ grows in size during the search. From the subset $S(m)$ of m observations, $m_0 \leqslant m \leqslant n-1$, we obtain the n squared Mahalanobis distances $d_i^2(m)$ given by (4). We order these 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. This is particularly the case when one cluster is completely fitted and observations from a second cluster have to be included as the search progresses and m increases (Atkinson et al., 2004, Section 7.3).

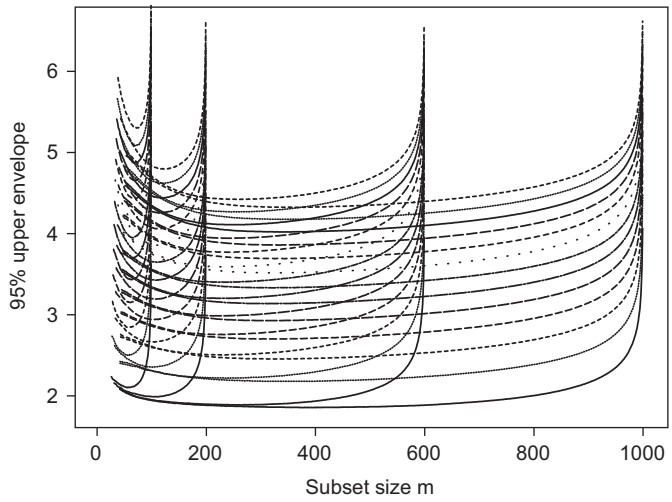

Fig. 1. Minimum Mahalanobis distances: 95% point for $n = 100$, 200, 600 and 1000 as v goes from 1 to 13. Small v at the bottom of the plot.

(3) *Outlier detection.* To detect outliers we examine the minimum Mahalanobis distance amongst observations not in the subset

$$d_{\min}(m) = \min d_i(m), \quad i \notin S(m). \tag{5}$$

If this observation is an outlier relative to the other m observations, this distance will be large compared to the maximum Mahalanobis distance of observations in the subset. All other observations not in the subset will, by definition, have distances greater than $d_{\min}(m)$ and will therefore also be outliers.

3. The structure of forward plots

Fig. 1 is a forward plot of simulation envelopes for minimum Mahalanobis distances from 10,000 simulations for samples sizes $n = 100$, 200, 600 and 1000. Throughout this section we consider only searches with a single robust start. The envelope given is the 95% point of the empirical distribution of the minimum Mahalanobis distance amongst observations not in the subset for v from 1 to 13. There is clearly some common structure as n and v vary. The curves for $n = 1000$ are virtually horizontal in the centre of the plot. In general the plot looks like a series of superimposed prows of viking longships, although the results for $n = 100$ are more curved. We exploit this common structure to find methods of interpolating and extrapolating to conditions for which we do not have simulations.

For moderate $n(< 1000)$ and v, we find a parametric form for each individual curve. Since the parameter estimates vary in a smooth way with v and n, we fit a second-order response surface to the estimated parameter values; interpolation of neighbouring estimates allows excellent interpolation of curves.

4. Parameterising a polynomial curve

To find a parametric form for the curves in Fig. 1 we take as known the last value for the quantile α of interest. Call this $d_{\min,\alpha}(n-1)$. To ensure the correct curvature we force our curve through this point. To do this let

$$x = \frac{n-1}{n} - \frac{m}{n} = \frac{n-m-1}{n}.$$

Thus, $x = 0$ when $m = n - 1$. We need a nonlinear term to approximate the curvature for small x and take

$$\tilde{d}_{\min,\alpha}(m) = d_{\min,\alpha}(n-1) + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 \exp(\beta_5 x), \tag{6}$$

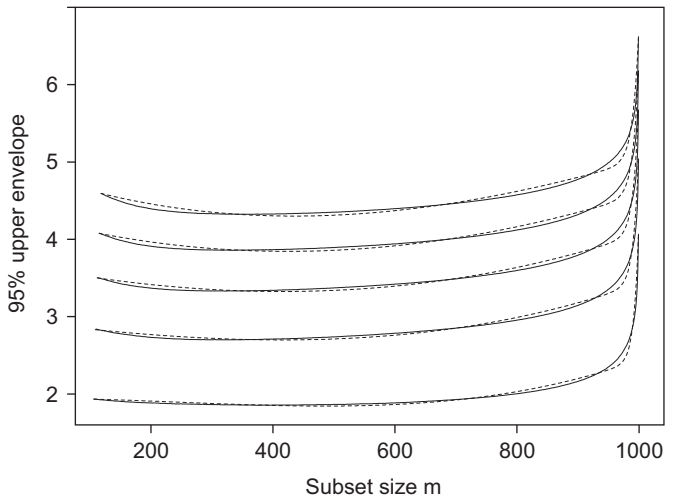

Fig. 2. Minimum Mahalanobis distances; dotted lines, 95% points fitted using the parameterisation (6) and, continuous lines, empirical curves for $n = 1000$; reading from the bottom, $v = 1, 4, 7, 10$ and 13.

for $m = m_0, \ldots, n - 1$. From a statistical data-analytical standpoint it may seem surprising that so many polynomial terms are needed. But the error in the curves is only due to fluctuations in the simulations, rather than to observational error.

We used nonlinear least squares to fit a separate curve (6) for each value of v from 1 to 13 and for n from 100 to 1000 in steps of 100 for $n \leqslant 400$ and in steps of 200 thereafter. Fig. 2 shows the empirical curves and the fitted values when $n = 1000$ for selected values of v. The dotted fitted lines agree well with the observed values. Although the fit is slightly less good in the earlier steps of the search, for most values of m the agreement is excellent and the curves provide a more than adequate guide to the significance of forward plots calculated from data.

The next stage is to smooth the parameter estimates. Plots, not given here, of the five parameter estimates, for each curve, show that they vary smoothly with n and v. We fit a second-order response surface, including interaction, to smooth the variation of each coefficient with the two variables. The smoothed values are then used in (6). The stages of the procedure are thus

(1) Determine the percentage point 100α required: 1, 2.5, 5, 50, 95, 97.5 or 99%.
(2) For a specific α interpolate linearly in v and n in stored tables of the values at $m = n - 1$. These are being made available in the Rfwdmv package downloadable from CRAN.
(3) Calculate the values of the five estimated parameters in (6) using the coefficients of the response surface, also similarly available.
(4) Calculate the envelope and compare with the data.

Fig. 3 shows the empirical curves and the fitted values from this procedure for nine combinations of n and v, with n increasing along rows. The fits are all more than satisfactory for our purpose of cluster identification.

An important aspect of this method is that it is local. If we only require envelopes for a small range of values of either v or n we can simulate envelopes over this smaller range and apply our procedure to this reduced set of results.

5. Robust or random starts

The envelopes and approximations in Sections 3 and 4 come from searches with robust starts. For cluster detection we instead run many forward searches from randomly selected starting points. In order to interpret the plots of distances we again need simulation envelopes as we do for searches with robust starts.

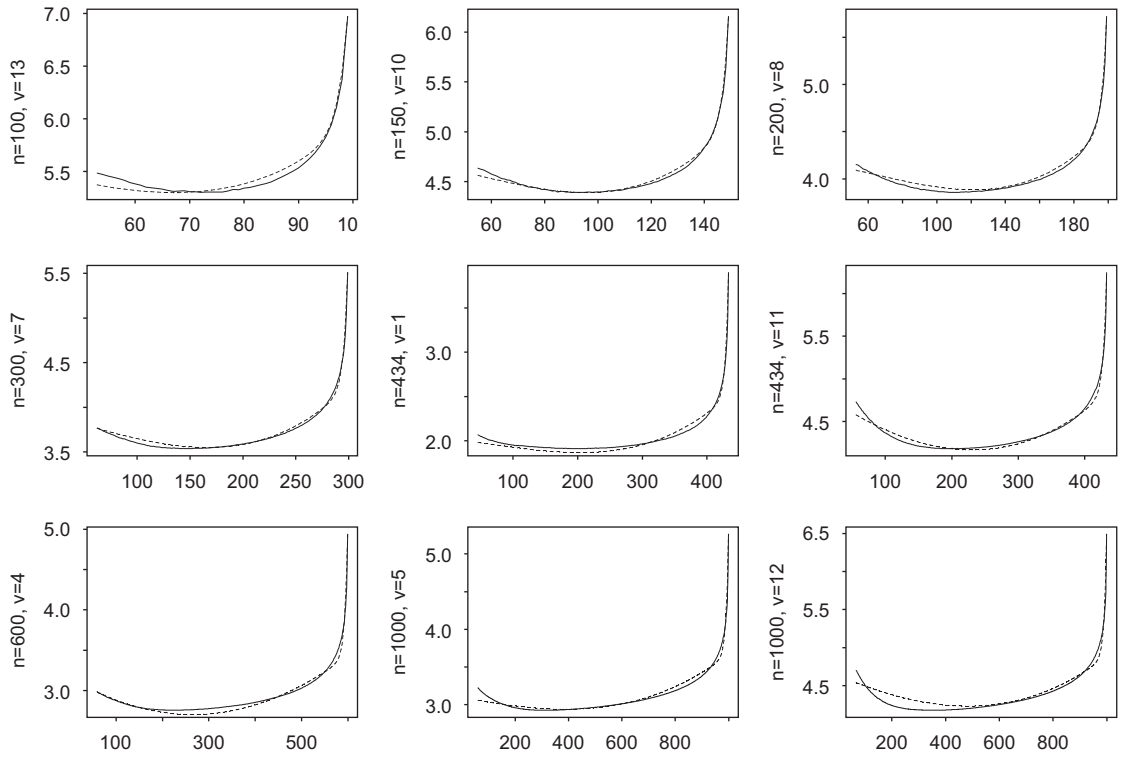

Fig. 3. Minimum Mahalanobis distances; fitted and empirical 95% curves for nine parameter combinations.

Atkinson et al. (2006) compared simulation envelopes from robust and random starts when the observations come from a single normal population, which is the null case for comparison. At the start of the search the two procedures are very different. However, they found that as the search progressed, this difference rapidly decreased as the subsets $S_R(m)$ for individual random searches converge to the $S(m)$ from the robust start. Their Fig. 1 shows that, from just below $m = 100$ in their example with $n = 200$, 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. Other comparisons, not reported here, likewise show little evidence of difference between the two sets of envelopes. Our Figs. 5, 10, 12 and 13 all show the convergence of the subsets $S_R(m)$ to a single trajectory. For our examples we accordingly use the polynomial approximation of Section 4 to the envelopes for robust starts while ignoring the envelope in the initial stages. The advantage is that one set of envelopes serves for both cluster identification and outlier detection. For sets of data appreciably smaller than those in this paper, direct simulation can be used and will be computationally efficient.

6. The detection of clusters

In this section we look at a synthetic example with $n = 1000$. We show how random start forward searches combined with envelope plots of forward Mahalanobis distances lead to the detection of clusters. We then interrogate the forward plots to obtain an initial idea of membership of the clusters. We use our first example to introduce our new procedures for cluster identification, giving a more detailed analysis of the Old Faithful example in the next section. We indicate how the information gained from these initial forward searches can be used as the basis of a definitive clustering. In both examples we start with $m_0 = v + 1$, the smallest possible size and that which gives the highest probability of leading to a subset consisting solely of observations from one cluster.

Fig. 4 is a scatter plot matrix of a sample of 1000 five-dimensional normal observations, 500 each from two populations that have the same independent error structure; as the boxplots on the diagonal of the plot show, the two clusters differ

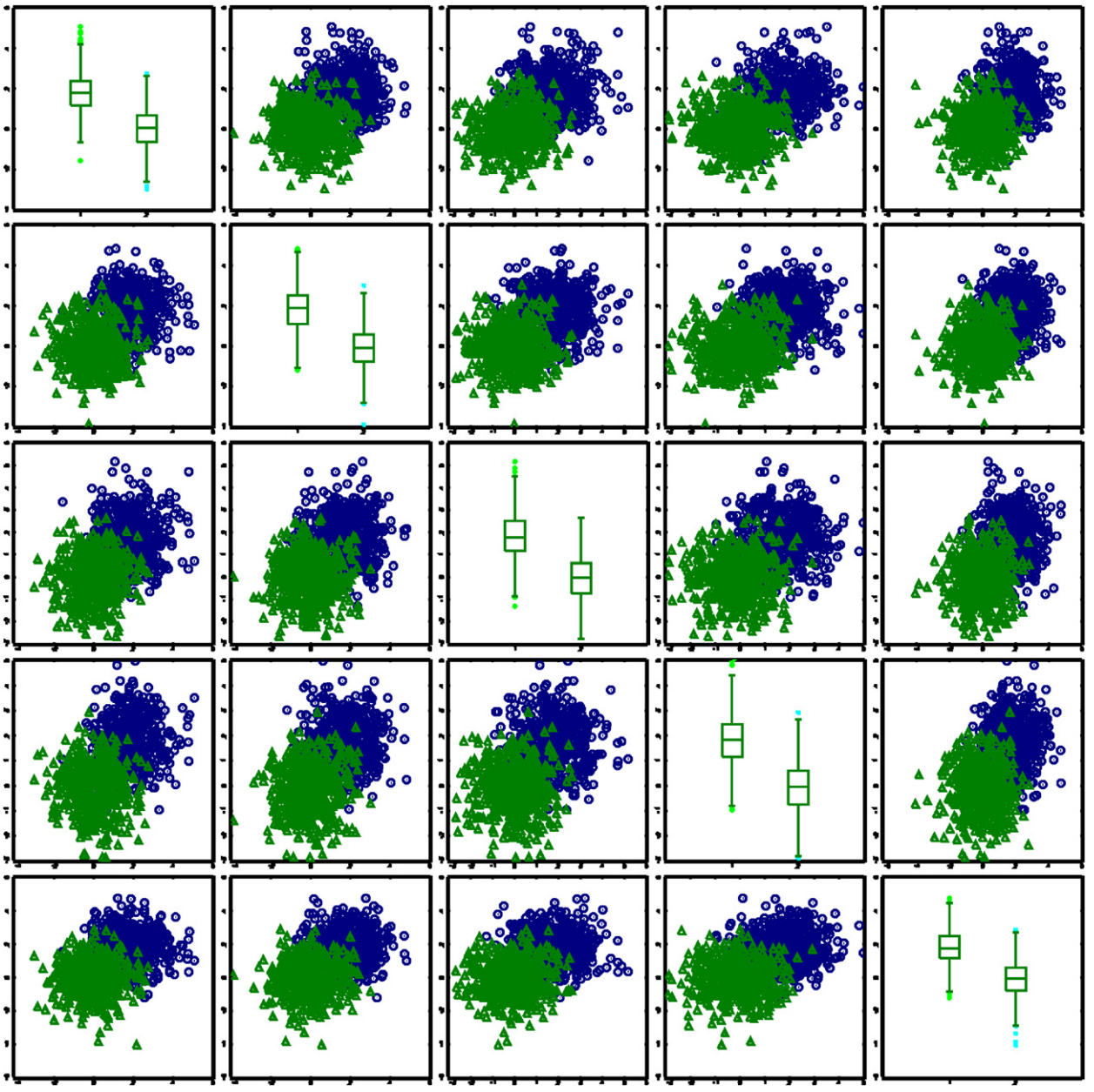

Fig. 4. Two clusters of independent normal variables: scatter plot matrix.

slightly in mean, but there is appreciable overlapping. The two samples were generated in Gauss 7 using the code

```
seed =498765;
x =rndns(1000,5,seed);
x[1:500,.] =x[1:500,.] + 1.8;
```

We used 200 random start forward searches to elucidate the structure of the data. The results are shown in Fig. 5. The forward searches fall into three classes: those that start in Group 1, those that start in Group 2 and those that, from the beginning of the search, include observations from both groups. These are shown in grey on the plot. From around $m = 300$ the searches with observations from only one group start to lie outside the envelopes. As the curves in Fig. 1 show, the plots for smaller n and given m rise above those with larger n as m increases. The curves here are behaving

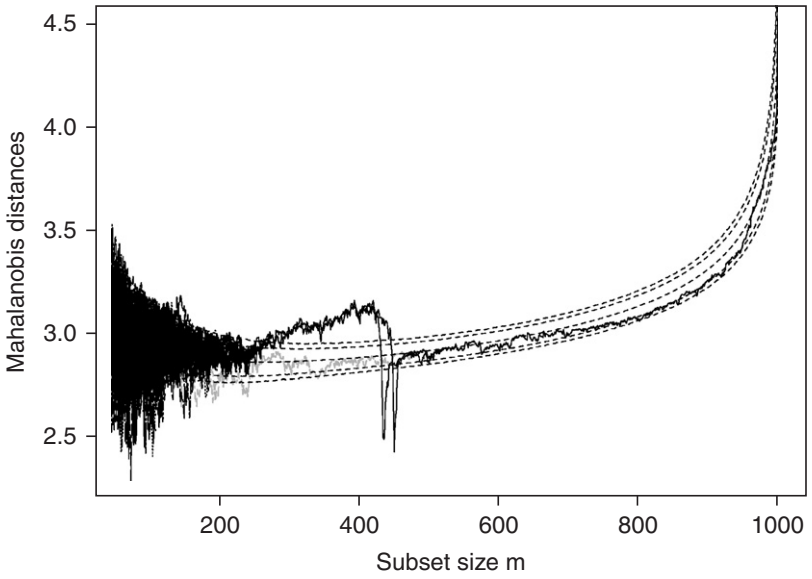

Fig. 5. Two clusters of independent normal variables: forward plot of minimum Mahalanobis distances from 200 random starts with 1%, 5%, 50%, 95% and 99% envelopes. Two clusters are evident around $m = 400$. Trajectories in grey always include units from both groups.

as those from samples of around 500. However, around $m = 440$ both curves suddenly dip below the envelopes as relatively remote observations from the other group enter the subset. Shortly thereafter there is a single forward plot, in which a common mean and common covariance matrix are calculated from the observations in the two groups. As a result the estimated covariance matrix is slightly too large and the plotted distances are slightly small; they lie rather low in the envelope in Fig. 5.

With 500 observations in each group we would like the dips below the envelopes in Fig. 5 to occur near $m = 500$. That they occur earlier is a result of the overlapping nature of the clusters; about 880 observations are identified as coming from one or other of the clusters. The remaining 120 observations require more careful analysis; with overlapping clusters it may be impossible to assign them unambiguously to a cluster.

We now move from Fig. 5 to cluster membership. Fig. 5 shows that, in the second half of the search, all 200 searches have converged in the sense that, for each m, there is one common set of observations $S(m)$ and one value of $d_{\min}(m)$. Once two searches have the same value of $S(m)$ they will continue to follow the same path, wherever they started from, producing identical values of $d_{\min}(m)$. Fig. 5 shows that initially there are many different values of $d_{\min}(m)$.

How the number of different values of $d_{\min}(m)$ decreases with m is shown in Fig. 6. Up to $m = 95$ there are 200 trajectories. The left-hand panel of the figure shows that the number then decreases rapidly, reaching 1 at $m = 524$. We are interested in the subsets $S(m)$ for these trajectories where there is evidence of a cluster structure. From Fig. 5 this is around $m = 400$. The right-hand panel of Fig. 6 shows that the number of distinct values of $d_{\min}(m)$ has decreased to five at this value. To find the clusters we interrogate Fig. 5 at this point to find the subsets giving rise to the larger values of $d_{\min}(m)$.

Fig. 7 is a plot of the frequency distribution of the values of $d_{\min}(m)$ when $m = 400$. The vertical lines in the plot correspond to the 1%, 50% and 99% points of the envelope at $m = 400$. There are only five values of $d_{\min}(m)$ and so only five residual trajectories. The largest value occurs 31 times and the two next highest 42 and 40 times. Only the largest two values lie above the 99% point and so can be expected to be indicative of clusters. The membership of the subsets for these values can be illustrated using an 'entry' plot.

The entry plot is a way of representing the membership of $S(m)$ as a search progresses. For each m those observations included in the subset are plotted with a symbol, so that the plot becomes darker as m increases. Such plots are discussed in Atkinson et al. (2004, Section 7.3.3). Here we need to combine information from several searches.

Fig. 8 is the entry plot for one of the 31 searches with the most extreme value of $d_{\min}(m)$ in Fig. 7. Since the searches have converged at $m = 400$, all will have the same residual trajectory, so it does not matter which of the 31 we choose to plot. For $m < 400$ we select randomly from one of the 31 searches to obtain a typical plot. Fig. 8 shows clearly

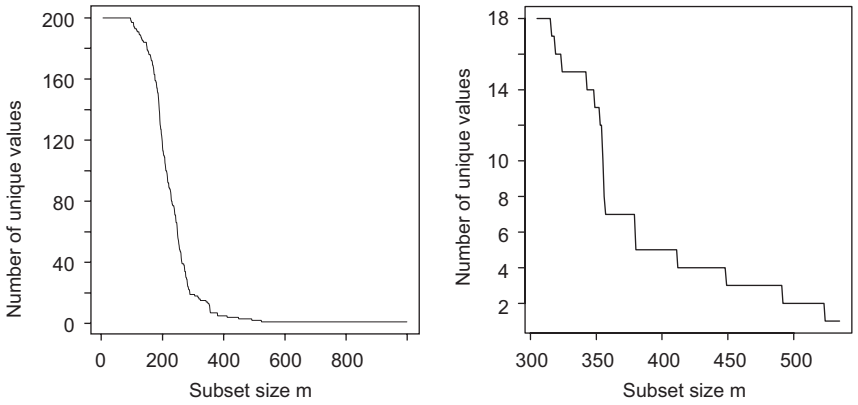

Fig. 6. Two clusters of independent normal variables: forward plots of number of unique minimum Mahalanobis distances from 200 random starts. Left-hand panel, from 200 to 1; right-hand panel zoom of plot where clusters become apparent; there are five distinct values at $m = 400$.

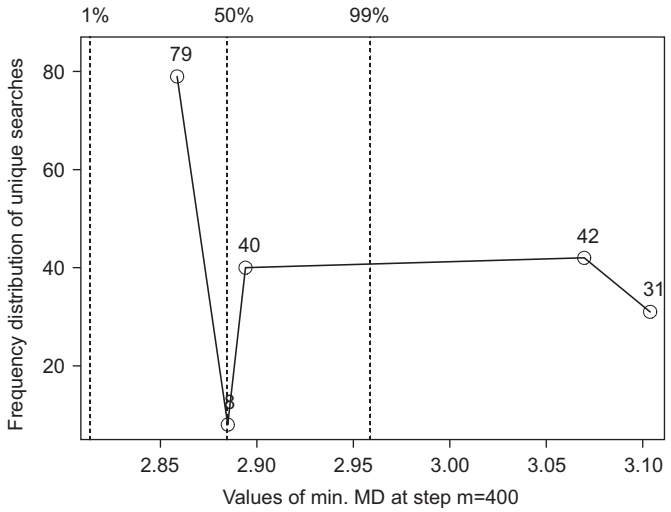

Fig. 7. Two clusters of independent normal variables: frequency distribution of $d_{\min}(400)$ in Figs. 5 and 6 from 200 random starts. The vertical lines are the 1%, 50% and 99% points at $m = 400$ of the envelope in Fig. 5.

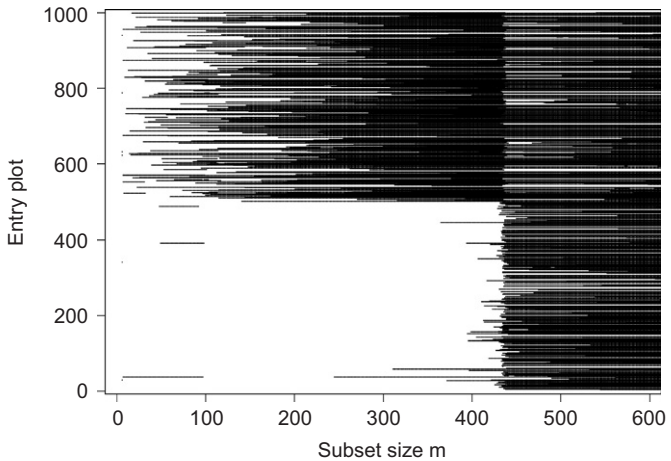

Fig. 8. Two clusters of independent normal variables: entry plot for a trajectory yielding the highest value of $d_{\min}(400)$ in Fig. 7.

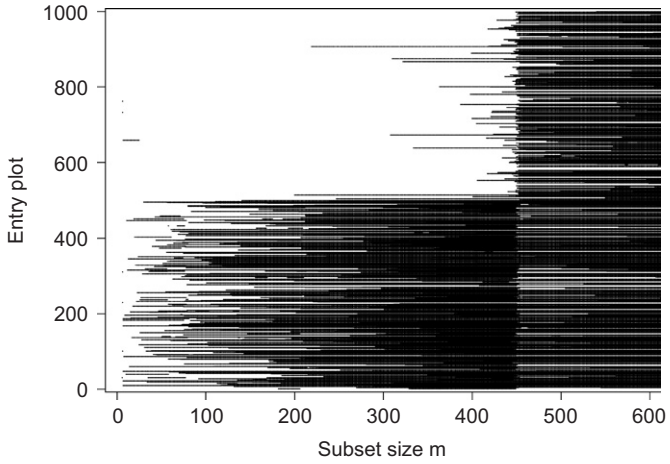

Fig. 9. Two clusters of independent normal variables: entry plot for a trajectory yielding the second highest value of $d_{\min}(400)$ in Fig. 7.

that we have found the trajectories that include observations from those numbered 501–1000. Many members of this cluster are unambiguously identified. Likewise, Fig. 9 shows the complementary cluster of observations numbered 1–500. The third highest value of $d_{\min}(m)$ in Fig. 7 gives rise to an entry plot, not given here, with no evident structure, observations from both clusters being in the subsets throughout. This is in line with the interpretation of Fig. 5 where the third highest trajectory at $m = 400$ is within the envelopes.

The results shown in Figs. 8 and 9 clearly indicate that our cluster identification procedure works. In addition, the plots show the overlapping nature of the clusters as units from both are in the subset well before $m = 500$. We have truncated the plots at a value of m slightly above 400, as the plots for larger m grow steadily darker without providing any further information. Since the clusters correspond to the natural order of the units, we are able to check that our method has achieved what is required. Although permutation of the labels of the units renders the plots uninformative, we do not use this form of information in the more detailed analyses in the next section.

7. Eruptions of Old Faithful

As a non-simulated example we cluster data on successive eruptions of the 'old faithful' geyser in Yellowstone National Park, Wyoming. The data are taken from the MASS library (Venables and Ripley, 2002). There are 272 observations with x_{1i} the duration of the ith eruption and x_{2i} the waiting time to the start of that eruption from the start of eruption $i - 1$. There are several similar data sets in the literature. That literature and the physics of the problem are discussed by Azzalini and Bowman (1990) who employ a time series analysis. Here we use clustering to establish whether one or more than one, mechanism is present and, if so, how many.

Fig. 10 shows forward plots of minimum Mahalanobis distances from 300 random starts. There are two clear maxima in this plot, one at $m = 98$ and the other at $m = 177$, suggesting the existence of two clusters. The total of these two numbers is 275, greater than 272, the total number of units; the peaks are caused by units that are far from the clusters already established by the search. To obtain initial clusters we can interrogate this plot at values of m around these. The left-hand panel of Fig. 11, calculated at $m = 100$, shows that 68 searches pass through this first peak. The right-hand panel, for $m = 177$ shows that a much larger number, 228, of the searches form the second peak.

To investigate this cluster structure we take the 100 units forming the first peak and run 100 random start searches just on these units. The resulting forward plot of minimum Mahalanobis distances is in Fig. 12(a). There are clearly some outliers in this cluster. We drop the last unit to enter and run the search with $n = 99$. The results are similar, although a little less extreme. Fig. 12(b) shows the plot for the searches when $n = 98$; there is still some indication of outlying observations at the end of the search. The successive plots for $n = 97$ and 96 in Figs. 12(c) and (d) suggests that the cluster contains 97 units, the plot for large values of m lying entirely within the envelopes.

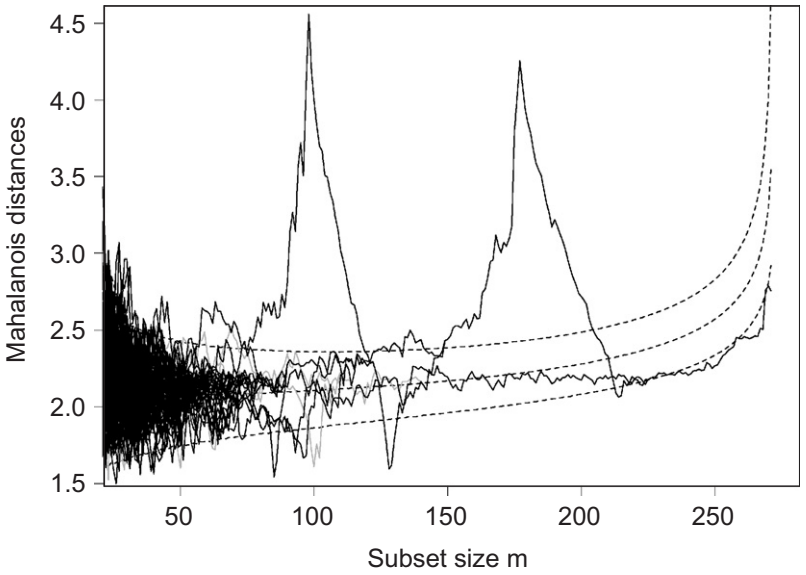

Fig. 10. Old Faithful data: forward plot of minimum Mahalanobis distances from 300 random starts with 1%, 50% and 99% envelopes. Two clusters are evident around $m = 99$ and 178. The few trajectories in grey always include units from both groups.

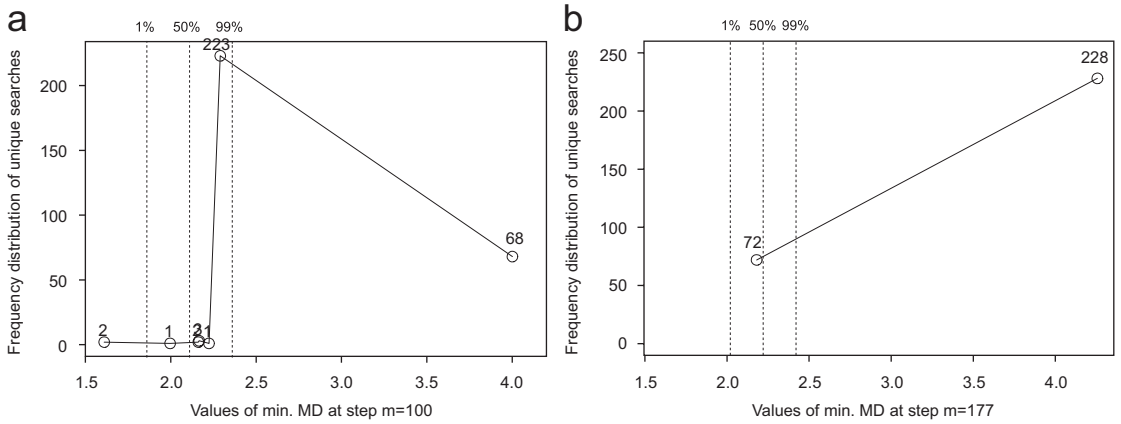

Fig. 11. Old Faithful data: frequency distributions of (a) $d_{\min}(100)$ and (b) $d_{\min}(177)$ in Fig. 10. The vertical lines are the 1%, 50% and 99% points at m of the envelope in Fig. 10.

Fig. 13 shows the results of applying the same procedure to the grouping suggested by the peak just before $m = 180$. We take these 180 observations and run forward searches on decreasing sample sizes. The indication, in Fig. 13, is that the cluster contains 177 observations.

The preliminary classification based on this procedure is shown in Fig. 14. We have two clusters, one of 97 units and the other of 177, making 274 units in all, two more units than there are in the data. As the figure shows, there are no outliers and two units could belong to either cluster. A more subtle analysis of the classification can be obtained through a forward search in which two clusters are fitted, as illustrated for three clusters and different data in Section 7.5.4 of Atkinson et al. (2004).

There are several points to be made about our general procedure:

(1) The analysis of the tentative clusters leading to Figs. 12 and 13 required envelopes for a series of sample sizes. Individual simulation of these would be prohibitively time consuming and approximations like those of Section 4 greatly speed the process of data analysis;

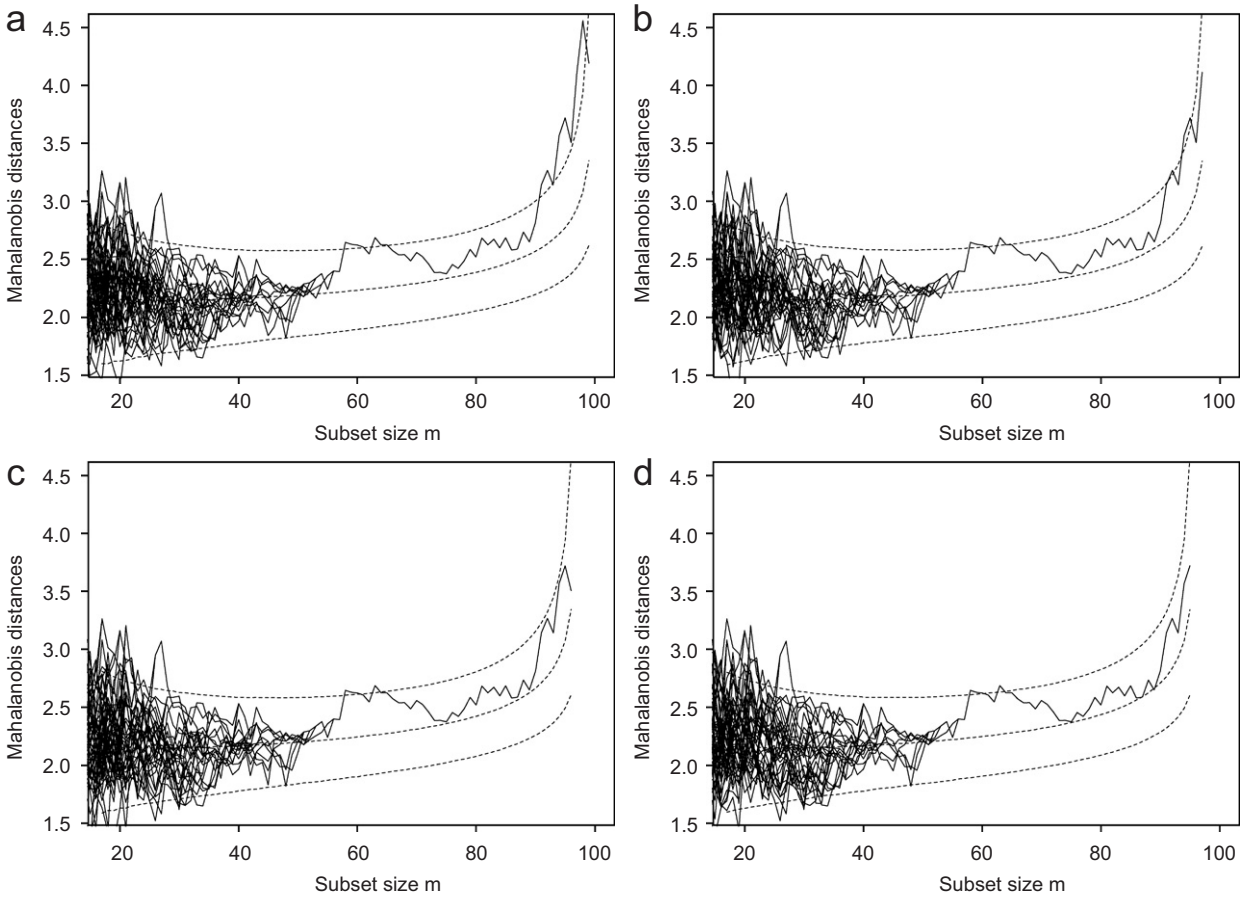

Fig. 12. Old Faithful data: forward plots of minimum Mahalanobis distances from 100 random starts with 1%, 50% and 99% envelopes for the data forming the first cluster in Fig. 10: (a) $n = 100$; (b) $n = 98$; (c) $n = 97$; (d) $n = 96$. A cluster of 97 units is indicated.

(2) By using random start searches for the individual clusters we are able to establish that these clusters are homogeneous and do not require any further division;
(3) The content of the envelopes found by simulation are pointwise for each m. Atkinson and Riani (2006) show how to use the simulations to provide simultaneous bounds on the probability of slight deviations such as that around $m = 60$ in the searches shown in Fig. 12. The probability of such a slight departure once in a search of this size is surprisingly high, often well over 50% depending on the precise feature to be identified. However, the probability of the large peaks in Fig. 10 is negligible;
(4) The two peaks in Fig. 10 are both approximately the same height as the upper envelope at the end of the search. This is in line with the simulations of Fig. 1 and suggests the two clear clusters without outliers that we have found. On the contrary, the peaks in Fig. 5, although evident, are slight. When we apply the method of individual clusters used in this section we find that either cluster can grow appreciably. Methods that fit more than one cluster are necessary for achieving some separation of such overlapping groups, arbitrary though the separation must to some extent be.

8. Other clustering analyses

In this section we briefly compare our analyses with those produced by two standard methods: the `mclust` library (Fraley and Raftery, 2003, 2006) and k-means (Anderberg, 1973, p. 166), which can be run on either standardized

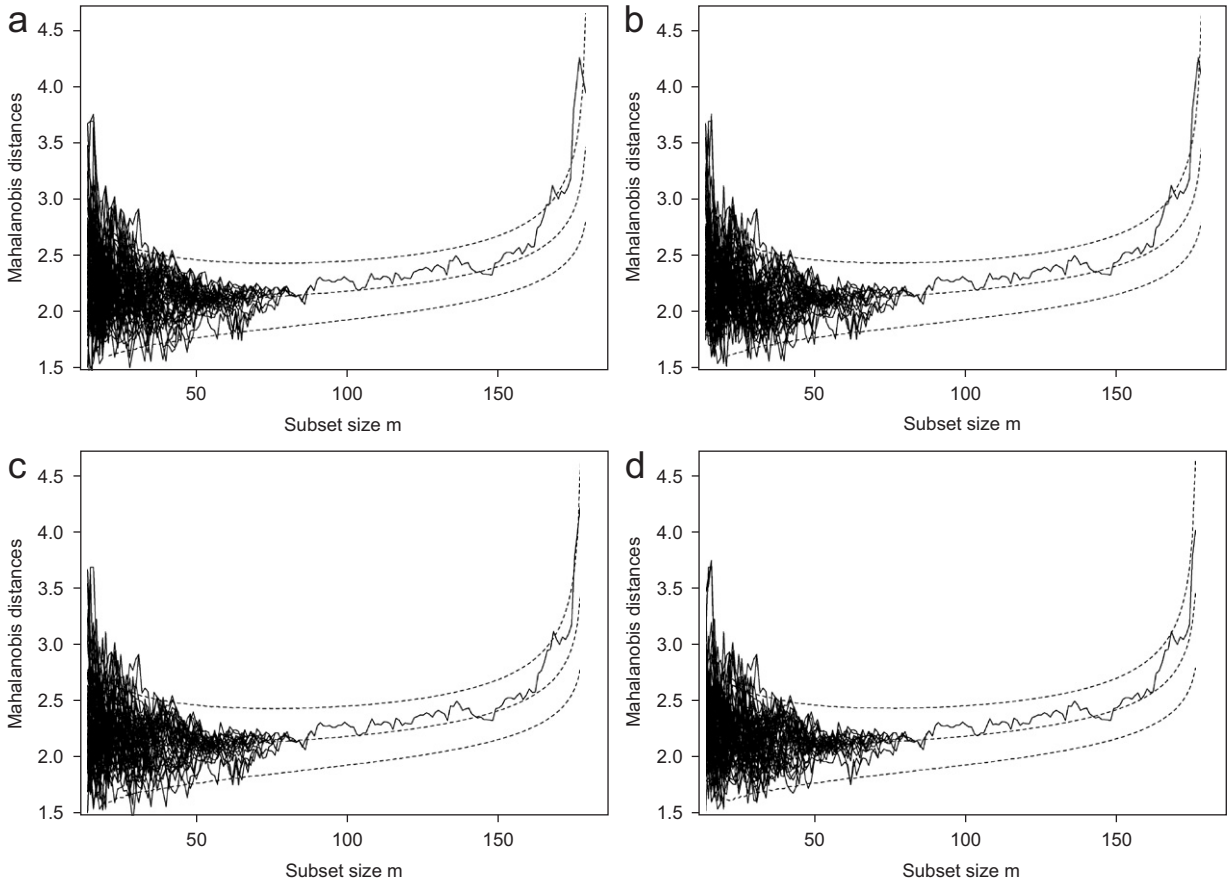

Fig. 13. Old Faithful data: forward plots of minimum Mahalanobis distances from 100 random starts with 1%, 50% and 99% envelopes for the data forming the second cluster in Fig. 10: (a) $n = 180$; (b) $n = 179$; (c) $n = 178$; (d) $n = 177$. A cluster of 177 units is indicated.

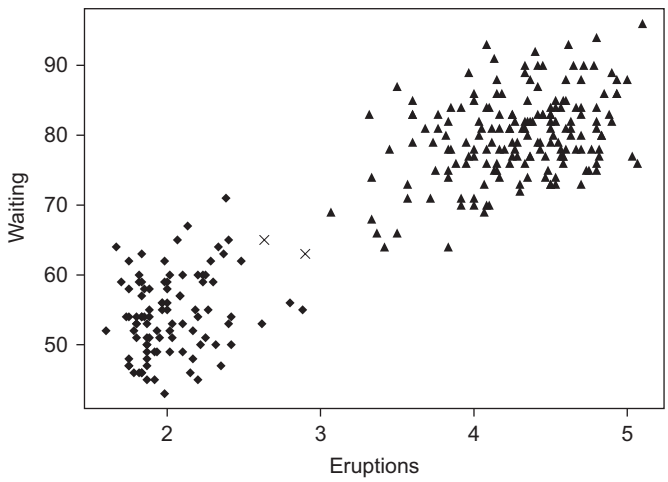

Fig. 14. Old Faithful data: scatterplot matrix showing the two clusters. The units marked × could lie in either cluster.

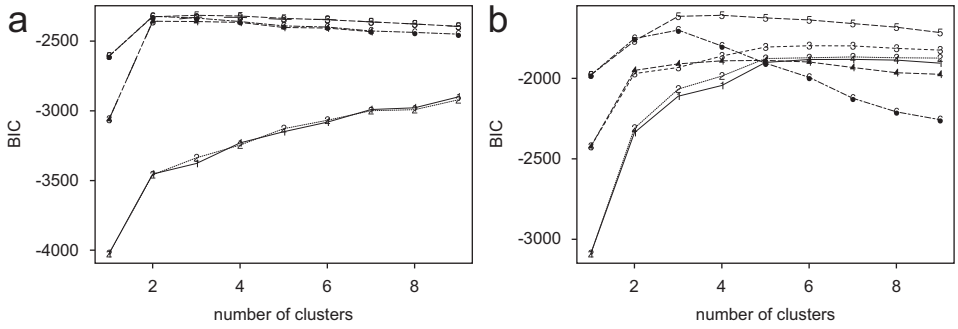

Fig. 15. BIC plots from `mclust`: (a) Old Faithful data; (b) Swiss banknote data.

or non-standardized data. For our selection criterion we use the Calinski–Harabasz index (Calinski and Harabasz, 1974).

Our simulated data in Fig. 4 consisted of two spherical normal clusters. Like our method, both comparative methods found two clusters. In the case of k means, whether we used standardized or unstandardized variables, there was a sharp indication of two clusters. The indication from `mclust` was also of two clusters, with the same spherical normal distribution. However, the BIC criterion declined very slowly as the number of groups increased, in stark contrast to the strong clustering information given by Fig. 10.

The behaviour of the methods for the geyser data was not at all the same. The maximum value of the BIC criterion for `mclust` was for three clusters with similarly shaped covariance matrices, although they were differently scaled and oriented. The next best model also had this covariance structure, but with four clusters; the third best was the model we have found, with two clusters and unconstrained covariance matrices. The left-hand panel of Fig. 15 shows how very flat the BIC criterion is as the number of groups increases. The three group solution splits the larger cluster in Fig. 14 into two almost circular sets of units, those in the west and those in the east. See Fig. 2 of Atkinson and Riani (2007).

The k-means solution for the geyser data depends on whether or not the variables are standardised. The plot of the index for the standardised variables indicates six clusters, while the unstandardized distances increase steadily to 10; we did not explore higher values. Although it is known that k-means behaves poorly with non-spherical clusters it was a surprise to us that it behaved so poorly in this simple example.

To conclude our comparisons we briefly extend them to consideration of the data on Swiss banknotes (Flury and Riedwyl, 1988; Flury, 1997) in which there are two hundred observations on notes withdrawn from circulation. The data contain 100 notes believed to be genuine and 100 probable forgeries, on each of which six measurements were made. They are extensively analysed by Atkinson et al. (2004) who show that the forgeries fall into two clusters, making three clusters in all and a few outliers. This structure is evident in the forward plot of random start forward searches given by Atkinson et al. (2006). The plot of the BIC criterion for these data forms the right-hand panel of Fig. 15. The `mclust` procedure again finds one more cluster than we do (the line numbered 5), with again information matrices of the same shape, orientation and size. The plot for this structure decreases very slowly. On the other hand, the line for the second-best solution (numbered 6) decreases relatively rapidly as the number of clusters increases. However, this line is associated with a completely different model which uses a distinct volume, shape and orientation of the ellipsoid for each group. This specification also suggests three groups, but with the corresponding values of BIC appreciably lower than those for the structure preferred by `mclust`.

The behaviour of k means is more satisfactory; for standardised data three groups are preferred, whereas with unstandardised data, three and five groups have virtually the same index value.

It is standard that the use of AIC in model choice tends to overfit. One implication of our comparisons is that BIC for cluster choice also tends to overfit, producing too many clusters. Diagnostic use of the forward search methods demonstrated here, but starting from the clusters indicated by `mclust`, should provide a useful diagnostic method of clustering.

A final, general, point is that the random start forward search provides a robust method of establishing cluster numbers and membership. On the contrary, `mclust` is not robust. Our numerical experiments show that small changes in the Old Faithful data lead to changes in the number of clusters for which BIC is maximized.

References

Anderberg, M.R., 1973. Cluster Analysis for Applications. Academic Press, New York.
Atkinson, A.C., 1994. Fast very robust methods for the detection of multiple outliers. J. Amer. Statist. Assoc. 89, 1329–1339.
Atkinson, A.C., Riani, M., 2006. Distribution theory and simulations for tests of outliers in regression. J. Comput. Graph. Statist. 15, 460–476.
Atkinson, A.C., Riani, M., 2007. Discussion of paper by Handcock, Raftery and Tantrum. J. Roy. Statist. Soc. Ser. B 69 (In press).
Atkinson, A.C., Riani, M., Cerioli, A., 2004. Exploring Multivariate Data with the Forward Search. Springer, New York.
Atkinson, A.C., Riani, M., Cerioli, A., 2006. Random start forward searches with envelopes for detecting clusters in multivariate data. In: Zani, S., Cerioli, A., Riani, M., Vichi, M. (Eds.), Data Analysis, Classification and the Forward Search. Springer, Berlin, pp. 163–171.
Azzalini, A., Bowman, A., 1990. A look at some data on the Old Faithful geyser. Appl. Statist. 39, 357–365.
Calinski, T., Harabasz, J., 1974. A dendrite method for cluster analysis. Commun. Statist. — Theory Methods 3, 1–27.
Flury, B., 1997. A First Course in Multivariate Statistics. Springer, New York.
Flury, B., Riedwyl, H., 1988. Multivariate Statistics: A Practical Approach. Chapman & Hall, London.
Fraley, C., Raftery, A.E., 2003. Enhanced model-based clustering, density estimation and discriminant analysis: MCLUST. J. Classification 20, 263–286.
Fraley, C., Raftery, A.E., 2006. `MCLUST` version 3: an R package for normal mixture modeling and model-based clustering. Technical Report 504, University of Washington, Department of Statistics, Seattle, WA.
Hadi, A.S., 1992. Identifying multiple outliers in multivariate data. J. Roy. Statist. Soc. Ser. B 54, 761–771.
McLachlan, G., Peel, D., 2000. Finite Mixture Models. Wiley, New York.
Rousseeuw, P.J., Leroy, A.M., 1987. Robust Regression and Outlier Detection. Wiley, New York.
Rousseeuw, P.J., Van Driessen, K., 1999. A fast algorithm for the minimum covariance determinant estimator. Technometrics 41, 212–223.
Rousseeuw, P.J., van Zomeren, B.C., 1990. Unmasking multivariate outliers and leverage points. J. Amer. Statist. Assoc. 85, 633–639.
Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S. Fourth ed. Springer, New York.
Zani, S., Riani, M., Corbellini, A., 1998. Robust bivariate boxplots and multiple outlier detection. Comput. Statist. Data Anal. 28, 257–270.