To appear in the *Journal of Applied Statistics*
Vol. 00, No. 00, Month 20XX, 1–19

Cluster Detection and Clustering with Random Start Forward Searches

Anthony C. Atkinson[a]* Marco Riani and Andrea Cerioli[b]

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

(*Received 00 Month 20XX; accepted 00 Month 20XX*)

The forward search is a method of robust data analysis in which outlier free subsets of the data of increasing size are used in model fitting; the data are then ordered by closeness to the model. Here the forward search, with many random starts, is used to cluster multivariate data. These random starts lead to the diagnostic identification of tentative clusters. Application of the forward search to the proposed individual clusters leads to the establishment of cluster membership through the identification of non-cluster members as outlying. The method requires no prior information on the number of clusters and does not seek to classify all observations. These properties are illustrated by the analysis of 200 six-dimensional observations on Swiss banknotes. The importance of linked plots and brushing in elucidating data structures is illustrated.

We also provide an automatic method for determining cluster centres and compare the behaviour of our method with model-based clustering. In a simulated example with 8 clusters our method provides more stable and accurate solutions than model-based clustering. We consider the computational requirements of both procedures.

Keywords: *Keywords:* brushing; data structure; forward search; graphical methods; linked plots; Mahalanobis distance; MM estimation; outliers; S estimation; Tukey's biweight.

AMS Classification codes: 62H30; 62P20; 62F35

1. Introduction

The purpose of cluster analysis is to use observations on a set of units (plants, schools, countries, ...) so that units in the same cluster are more similar to each other than to those in other clusters. Clustering is a well-established, much studied statistical problem, for example Kaufman and Rousseeuw [18], to which there is no definitive solution, although there are many algorithms, such as that of Caliński and Harabasz [7]. The focus here is on samples from populations which are believed to have multivariate normal distributions. However the clusters can be of disparate shapes, the number of clusters is unknown and there may also be outliers, so that not all units have to be assigned to a cluster.

A standard approach to clustering, for example McLachlan and Peel [22], is to fit a mixture of multivariate normal distributions to the data. Information criteria such as that due to Akaike [1] (AIC) or the Bayesian criterion of Schwarz [28] (BIC) penalize the loglikelihood by a function of the number of estimated parameters and are often used to choose the number of clusters. Such methods assume that all the observations can be assigned to clusters. To allow for outliers, Fraley and Raftery [13] suggest including a

*Corresponding author. Email: a.c.atkinson@lse.ac.uk uniformly distributed component which Coretto and Hennig [10] extend to an improper constant density. However, protection may also be required against other kinds of departure. For example, the effective robust clustering procedure described and exemplified by García-Escudero *et al.* [15] allows for outliers through the specification of a level of trimming of the data, as well as requiring that the number of clusters is known. Atkinson *et al.* [6] apply the methodology of random start forward searches to the 2,000 observation example of García-Escudero *et al.* [15]. An advantage of this method is that they obtain a structure of clusters and outliers without any pre-specification of trimming level or cluster number.

One purpose of the present paper is to detail and exhibit the properties of the random start forward search for clustering multivariate data when coupled with the power of graphical methods, including the brushing of linked plots. As an example we reanalyse the Swiss banknote data [12] with these recently developed tools. Comparisons of clustering methods applied to these data include Morris and McNicholas [23], who find between 2 and 4 clusters and Klinke [19] who finds up to five. An important distinction between our method and others such as those, is that we are able to see how inferences depend on the properties of each observation, rather than relying solely on aggregate statistics such as means and estimated covariance matrices.

A second purpose of our paper is to introduce an automatic method, based on the forward search, for determining the number of clusters. We use a simulated example with 8 clusters and 400 observations to compare our new method with the use of AIC and BIC. We notice that our new procedure gives more stable solutions, even in the absence of outliers. Finally, we argue that our new method is not appreciably more computationally intensive than standard non-robust methods using information criteria.

The Forward Search (FS) for multivariate data is described by Atkinson *et al.* [3]. Data which are generated by a single model often contain multiple outliers, when the methods of robust statistics, such as those surveyed by Maronna *et al.* [21], are required to obtain good parameter estimates and to detect outliers. The forward search, in which the model is fitted to data subsets of increasing size, is a competitor of these methods. The subsets provide unbiased estimates of the parameters and lead to the detection of outliers.

Such robust procedures usually assume that the greater part of the observations are generated by a single model; the FS starts from a carefully selected initial subset of observations, chosen to lie near the centre of the data. If the data come from more than one population, starting with a subset of observations in one of the clusters leads to observations from other clusters being identified as outliers. In the general clustering problem examined here, there is no information about approximate cluster membership, nor even of the number of clusters. There is then no simple robust path to finding an informative initial subset. One solution is the 'random-start' forward search introduced by Atkinson *et al.* [4] which uses many randomly selected initial subsets of size m_0. In this paper we show how monitoring the behaviour of these numerous forward searches provides estimates of the number of clusters and their membership.

The paper is structured as follows. The FS for outlier detection is introduced in §2, together with some algebraic background. The Swiss banknote data are introduced in §3.1; there are 200 observations on 6 continuous measurements of pictorial aspects of the design on the notes. A preliminary classification by a human expert of notes withdrawn from circulation led to 100 of the notes being classified as genuine and 100 as forgeries. In order to exemplify the FS for outlier detection, we analyse the genuine notes in §3.2 and illustrate the information that can be provided by brushing linked plots. The analysis of the forgeries is in §3.3; we obtain markedly different analyses for the two groups. These analyses agree with those from standard robust methods for which we choose S and MM-estimation. The random start FS is introduced in §4.1 and, in §4.2, applied to the banknote data, ignoring the labels from the preliminary classification. We recover the clustering of the human expert, together with the identification of the outliers obtained in §3.3. Section 5 demonstrates the necessity of using a robust method specifically intended to identify clusters. We analyse the combined data of 200 observations and show that the FS starting from a single subset fails, like S and MM estimation, to detect the clustered structure of the data.

The example with eight clusters is used in §6 to illustrate our method for determining the number of clusters. There are six well-separated groups of data, two of which contain two nearby clusters. We use extreme envelopes in the random start FS in §6.1 to identify and remove nearly all the observations in these clusters. Our method identifies all eight clusters. Comparisons with the use of AIC and BIC in fitting mixtures of normal distributions using the Matlab routine `gmdistribution.fit` are in §6.2. These show that the solutions obtained depend upon the number of iterations in the optimization procedure. In our example the number of clusters found, despite the absence of outliers, ranges from 8 to 15. In §6.3 we discuss recent developments that have improved the numerical efficiency of the FS.

In the final section we discuss some other analyses of the banknote data. We stress the importance for cluster detection of informed fitting of different models to different subsets of the data. The purpose is to present a method of data analysis which, unlike either customary robust methods or those based on information criteria, provides rich insight into the structure of more complicated data structures. Unlike the latter methods, our procedure can also handle outliers.

2. The Forward Search

2.1 *Outlier Detection*

The method described by Riani *et al.* [25] provides outlier tests with specified size and good power when the sample comes from a single multivariate normal population, maybe with outliers. Both theoretical and data analytical properties of the FS are discussed in Atkinson *et al.* [5]. The search for a single population starts from a subset of m_0 observations, robustly chosen. The size of the subset is increased from m to $m+1$ by forming the new subset $S^*(m+1)$ from those observations with the $m+1$ smallest squared Mahalanobis distances. Thus, some observations in $S^*(m)$ may not be included in $S^*(m+1)$. For each m ($m_0 \leq m \leq n-1$), the test for the presence of outliers is based on the observation outside the subset with the smallest squared Mahalanobis distance.

2.2 *Mahalanobis Distances*

The parameters μ and Σ of the v-dimensional multivariate normal distribution of y are estimated in the forward search by the standard unbiased estimators from a subset of m observations providing estimates $\hat{\mu}(m)$ and $\hat{\Sigma}(m)$. Using these estimates we calculate n squared Mahalanobis distances

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

To detect outliers we use the minimum Mahalanobis distance amongst observations not in the subset

$$d_{\min}(m) = \min d_i(m) \quad i \notin S^*(m). \tag{2}$$

Testing for outliers requires a reference distribution for $d_i^2(m)$ in (1) and hence for $d_{\min}(m)$ in (2). When Σ is estimated from all n observations, the squared statistics have a scaled beta distribution. However, the estimate $\hat{\Sigma}(m)$ in the search uses the central m out of n observations, so that the variability is underestimated. Results of Tallis [29] on truncated distributions provide a consistency factor

$$c(m,n) = \frac{n}{m} C_{v+2}\{C_v^{-1}(m/n)\} \tag{3}$$

where C_r is the c.d.f. of the χ^2 distribution on r degrees of freedom. Then $c(m,n)\hat{\Sigma}(m)$ is an approximately unbiased estimate of Σ.

2.3 *Envelopes and Multiple Testing*

Atkinson *et al.* [3, pp 43–44] give results on the distribution of deletion Mahalanobis distances from a sample of size n. These results, applied to a randomly chosen subset of size $m+1$, show that their distribution is

$$\frac{v(m^2-1)}{m(m-v)} F_{v,m-v}. \tag{4}$$

The distribution of the rescaled minimum Mahalanobis distance $c(m,n)d^2_{\min}(m)$ constructed with the centroid and covariance matrix of the subset estimated using the units with the m smallest Mahalanobis distances is then the distribution of the $(m+1)$th order statistic from (4).

Results, for example Guenther [16], on the order statistics $Y_{[1]}, Y_{[2]}, \cdots, Y_{[n]}$ from a sample of size n from a distribution with CDF $G(y)$, show that

$$P\{Y_{[m+1]} \le y\} = P\left\{F_{2(n-m),2(m+1)} > \frac{1-G(y)}{G(y)} \times \frac{m+1}{n-m}\right\}. \tag{5}$$

Given that in our case $G(y)$ is the CDF of the $F_{v,m-v}$, we can rewrite equation (5) as

$$\begin{aligned}\gamma = {} & P\{d^2_{\min}(m) \le d^*\} = \\ & 1 - F_{2(n-m),2(m+1)}\left(\left[\frac{1}{F_{v,m-v}\left\{\frac{m(m-v)}{v(m^2-1)}c(m,n)d^*\right\}} - 1\right]\frac{m+1}{n-m}\right),\end{aligned}$$

where $F_{a,b}(y)$ is the CDF of the F distribution with a and b degrees of freedom evaluated at y.

In addition to formal estimation and testing, it can be extremely helpful to look at forward plots of quantities of interest such as $d_{\min}(m)$ during the search and to compare them with the envelopes from several values of γ. The envelopes in Figures 3 and 9 show that there is appreciable curvature in the plots as $m \to n$; the envelopes increase rapidly, as, in the absence of outliers, large distances occur at the end of the search. To clarify visual presentation we use a pointwise normal-score transformation of the envelopes, and of the observed distances, in order to give plots with horizontal envelopes. The plot in normal coordinates uses the inverse of the cdf of the standard normal distribution, that is $\Phi^{-1}(\gamma)$, which, of course, does not change the statistical properties of the plot.

During the search, we perform a series of outlier tests, one for each $m \ge m_0$. We thus perform many tests. To allow for multiple testing, we follow Riani *et al.* [25] and use a rule which depends on the sample size which determines the relationship between the envelopes calculated for the distribution of the test statistic and the significance of the observed values. This rule has a relatively simple form only in the last part of the search, indicated by a vertical line in Figure 1.

3. Outliers and the Data on Swiss Banknotes

3.1 *Swiss Banknotes*

The data introduced by Flury and Riedwyl [12, pages 4–8] are 200 six-dimensional observations on Swiss banknotes withdrawn from circulation. The notes are the 1,000 Franc of the second series (see `www.snb.ch`), introduced in 1911 and withdrawn during 1956–7. As the website comments "this represents an extraordinarily long life span". It is therefore not surprising, especially given the huge value of the note, that it attracted the attention of forgers. The reverse of the note shows a scene in a foundry. An expert, on the basis of this scene, classified a selection of notes as genuine or forgeries. Flury and Riedwyl present measurements of aspects of the foundry scene and its relationship to the border of the banknote. They give readings for 100 each of notes believed to be genuine and those believed to be forgeries. Analyses of these two sets of readings, separately and combined, provide an excellent illustration of the properties of FS methods for outlier detection and clustering.

3.2 *The Genuine Notes*

The FS starts from a subset of m_0 observations chosen not to be outlying in any two-dimensional subset of the variables. After univariate outliers have been removed, the co-ordinate wise median of each pair of variables is found and the covariance matrix of the two-dimensional subset estimated around this centre. The resulting elliptical contours of Mahalanobis distances are then jointly scaled so that a specified number m_0 of observations lies within all of them. These central observations form the initial subset. We accordingly call this the elliptical start. The search then moves forward increasing the number of observations in the subset, until the nearest observation to those already in the subset appears to be an outlier, as judged by an appropriate envelope of the distribution of the test statistic. We call this a "signal".

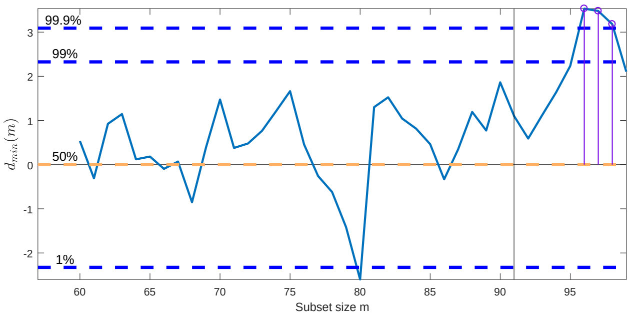

Figure 1. Genuine notes. Normal scores forward plot of minimum Mahalanobis distance $d_{\text{min}}(m, n)$. There is no signal at $m = 80$, but there is at $m = 97$ since $d_{\text{min}}(96, 100)$, $d_{\text{min}}(97, 100)$ and $d_{\text{min}}(98, 100)$ are all above the 99.9% envelope. The vertical line at $m = 91$ indicates the final part of the search.

We calibrate the FS envelopes used to identify outliers using the two-stage procedure of Riani *et al.* [25]. In the first stage of the search we monitor the bounds for all n

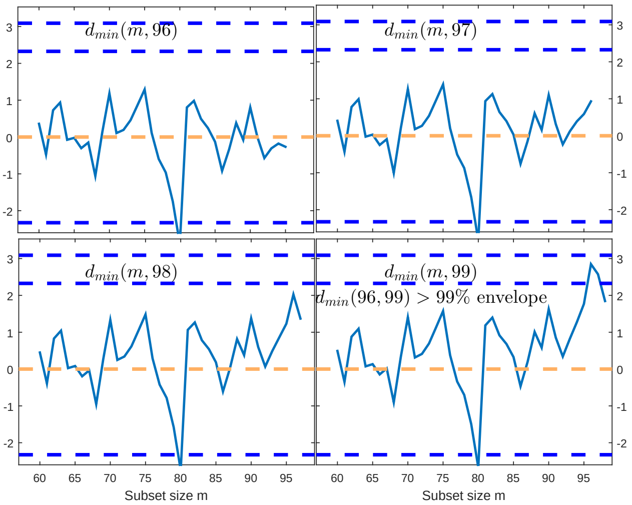

Figure 2. Genuine notes; resuperimposition of envelopes. Normal scores forward plot of $d_{\min}(m, n)$ for $n^\dagger = 96$, 97, 98 and 99. The first outlier is identified at $n^\dagger = 99$. 1%, 50%, 99% and 99.9% bands. There are two outliers

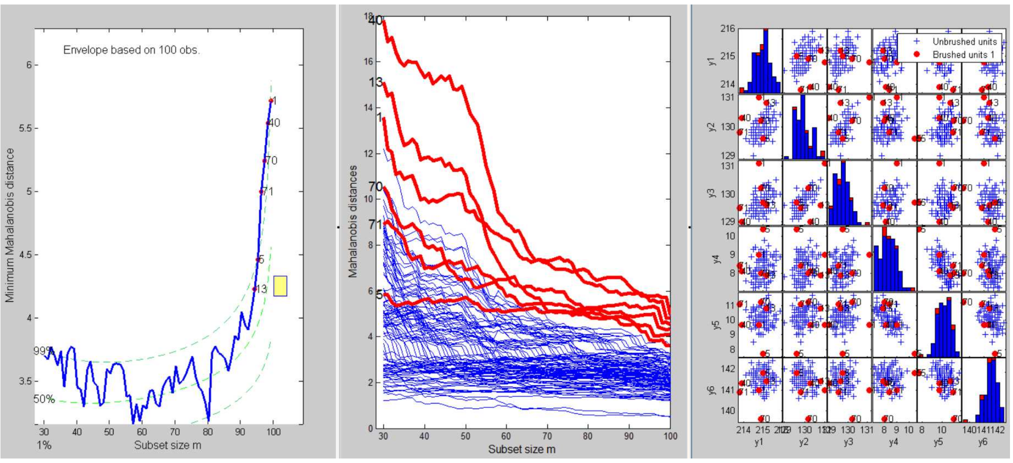

Figure 3. Genuine notes. Left-hand panel: minimum Mahalanobis distance in the original scale with brushing of the last six units to enter the search. Centre panel: trajectories of the Mahalanobis distances of these six units. Right hand panel: the scatterplot matrix showing the last six units, including the two outliers, observations 1 and 40, best seen in the plot of y_1 against y_2

observations. If at some point, the value of the test statistic lies above the threshold we obtain a signal indicating that observation $m^\dagger$, and the remaining observations not in the subset, may be outliers. The hypothesis is that the sample contains some non-outlying observations; this number may be less than n. If so, we need to judge the values of the statistics against envelopes from appropriately smaller population sizes. The second stage of the analysis consists of superimposing envelopes for a series of smaller sample sizes $n^\dagger$, starting from $m^\dagger - 1$ onwards, until the first introduction of an observation recognised as an outlier. The procedure moves forward from the elliptical start to find a value of $m^\dagger$, and then proceeds with the superimposition of envelopes until an outlier is identified and the outlier free sample size is established. We use the extended notation $d_{\min}(m, n)$ to indicate the sample size for which the envelopes were computed.

In summary, the two steps of the procedure are:

(1) Monitoring using bounds for all n observations. Detection of a "signal" at $m^\dagger$ indicates the presence of one or more outliers;
(2) Identification of individual outliers through the resuperimposition of envelopes for a series of sample sizes starting from $m^\dagger - 1$.

Figure 1 shows the procedure of Riani *et al.* [25] for the 100 genuine notes. In interpreting this plot it is important that the significance levels of the envelopes are for pointwise exceedances. Simulations for the null case of no outliers reported in Table 1 of Atkinson and Riani [2] show that the probability of at least a single exceedance of the 1% band in the last half of a search with 100 regression observations is 0.196. The peak at $m = 80$ is of such a type. The purpose of the two-stage procedure is to provide a test for the whole sample of a specified size, in our example 1%. With only a few outliers, these will enter the subset at the end of the search, being revealed by large values of the minimum Mahalanobis distance. One part of the rule for this final part of the search indicates a signal when $m = 97$, since $d_{\text{min}}(96, 100)$, $d_{\text{min}}(97, 100)$ and $d_{\text{min}}(98, 100)$ are all above the 99.9% pointwise envelope. We then start resuperimposing from $n = 96$. The four panels of Figure 2 show that, with these new envelopes, the trajectory of values of $d_{\text{min}}(m)$ lies inside the new envelopes until $n = 99$. There are therefore two outliers, observations numbered 1 and 40.

The three panels of Figure 3 illustrate the use of linked plots in interpreting the results of the FS. The left-hand panel repeats Figure 1 in the original scale, with brushing of the last six units to enter the search. These are, from least to most remote, units 13, 5, 71, 70, 40 and 1. The small box indicates the last unit to be brushed. The centre panel shows forward plots of Mahalanobis distances for all units; the trajectories of the brushed units are highlighted. Initially units 1, 13 and 40 have the largest Mahalanobis distances, with unit 40 the most remote. Although the ordering of the most remote units at the end of the FS changes during the search, they are most remote for the last 30% of the FS. Comparison with similar, but unlinked, plots in Atkinson *et al.* [3] shows that such generally decreasing curves are indicative of a well-behaved normal population with a few mild outliers. The right-hand panel shows the scatterplot matrix of the data with the last six units highlighted. The outlying nature of units 1 and 40 is most apparent in the plot of y_1 against y_2. A final comment on linked graphs is that clicking on a curve provides a pop-up box with information on the unit number, the value of the Mahalanobis distance and when the unit enters the subset.

The FS is one of several methods of detecting outliers in multivariate data. We compare our analyses with results from using both MM and S estimation in which extreme observations are downweighted by a function ρ.

In estimation of Mahalanobis distances, as in (1), the estimate of the mean μ does not depend on the estimate of Σ. However, this is not the case in such robust methods as MM and S estimation. These are derived from M estimation in which the downweighting function ρ is used with the variance assumed known. S and MM differ in the estimation of Σ. In S estimation this estimate is found as the minimizing solution of an equation including a ρ function which is then used in a second equation with, possibly another, ρ function to estimate μ. The estimator of μ is called an S-estimator because it is derived from a scale statistic, albeit in an implicit way.

The results of Riani *et al.* [26, §2.2] establish an asymptotic relationship between the breakdown point and efficiency of S estimators; as one increases, the other decreases. In an attempt to break out of this relationship, Yohai [31] introduced MM estimation, which extends S estimation. In the first stage the breakdown point of estimation of $\Sigma = \sigma^2\Gamma$ is set at 0.5, thus providing a high breakdown point. This fixed estimate of σ^2 is then used in the estimation of μ and Γ with high efficiency. Details of the methods are in Maronna *et al.* [21, §6.4]. In both cases we used Tukey's biweight ρ function and chose settings giving very robust procedures; an asymptotic breakdown point of 50% for S estimation and 95% nominal efficiency for MM estimation.

We now analyse the genuine notes using both MM and S estimation. Figure 4 shows index plots of the resulting robust squared Mahalanobis distances, with those for S estimation in the upper panel. The plots in the two panels are virtually indistinguishable. The two largest distances, lying above the 99.99% pointwise band, are for observations 1 and 40, which were also indicated by the FS. Three further observations have distances lying above the 99.9% band: 70, 71 and 5. However, since we are testing each of the 100 distances, we need to allow for multiple testing. The 99.99% bound corresponds to the Bonferroni approximation to the samplewise 99% band, so that just two outlying observations are identified at this level.

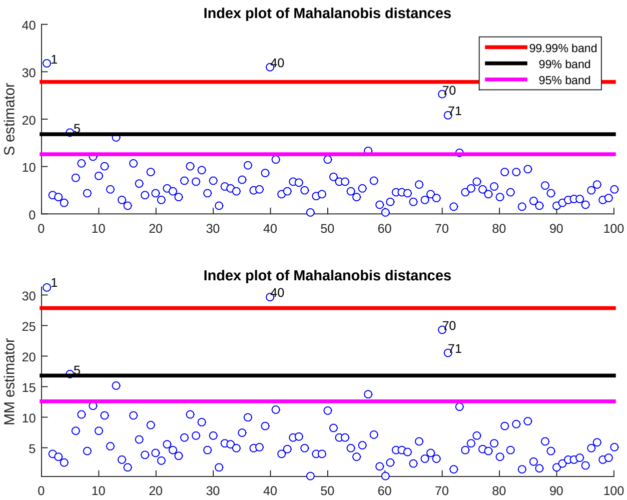

Figure 4. Genuine notes. Plots of squared robust Mahalanobis distances. Upper panel, S-estimation; lower panel MM

It is interesting to compare these findings with those in Atkinson *et al.* [3, Figure 3.44], which gives boxplots of the individual responses for the genuine notes. Observations that are particularly outlying are 1, especially in y_3, and 70 in y_6. However, as Figure 3 shows, observations 1 and 40 are the only two which lie away from the main group in two dimensions, although they are only marginally outlying in either y_1 (40) or y_2 (1). This is a simple example of the inadequacy of one-dimensional plots, such as boxplots or histograms, to indicate outliers in higher dimensions. As before, we can also link the robust distances in Figure 4 to other plots to extract such information.

Some comparisons of the FS with other robust methods for single populations are presented by Riani *et al.* [25]. We can expect that the comparisons with MM and S estimation given here depend on our chosen extreme values of efficiency and breakdown point. We do not pursue this matter further, but note that the monitoring of robust estimates for regression in Riani *et al.* [26] shows that such estimates are often stable over wide ranges of values of efficiency and breakdown point and then suddenly switch to quite different values.

3.3 *The Forgeries*

Figure 5 shows the forward plots of Mahalanobis distances when the same procedure is applied to the forgeries. A signal is obtained at $m = 84$, suggesting more outliers than in Figure 1. The large value of $d_{\min}(84, 100)$ indicates that this observation, and the others not included in the subset $S^*(83)$, lie far from the observations so far included in the subset. The curve of values of minimum Mahalanobis distances in the figure also shows a return to values close to the centre of the distribution as the search continues. This “masking” is caused by the eventual inclusion into $S^*(m)$ of observations from a cluster of outliers. As a consequence, the parameter estimates are corrupted and the remaining

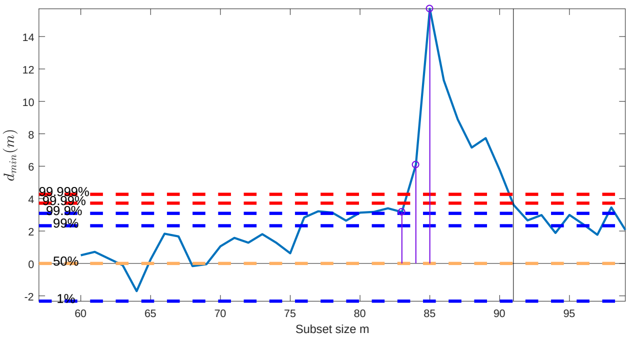

Figure 5. Forgeries. Normal scores forward plot of $d_{\text{min}}(m, n)$. There is a signal at $m = 84$ since $d_{\text{min}}(84, 100)$ is above the 99.999% envelope

outliers seem increasingly less remote. For another example of masking in the FS for multivariate data see Atkinson *et al.* [3, §3.2].

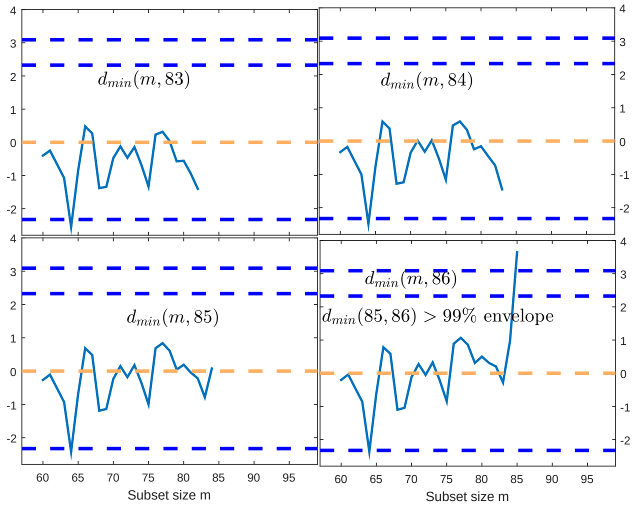

Figure 6. Forgeries; resuperimposition of envelopes. Normal scores forward plot of $d_{\text{min}}(m, n)$ for $n^{\dagger} = 83$, 84, 85 and 86. The first outlier is identified at $n^{\dagger} = 86$. 1%, 50%, 99% and 99.9% bands. There are 15 outliers

We start the confirmatory resuperimposition of envelopes from $n = 83$. Figure 6 shows the trajectories. There are no outliers for $n = 83, 84$ and 85. However, for $n = 86$, $d_{\text{min}}(85, 86) > 99\%$ envelope. Thus we conclude that there are 15 outliers.

Figure 7 shows the scatterplot matrix with the 15 outliers plotted with a different symbol. In some scatterplots, such as that of y_4 against y_6, these 15 observations seem to form a separate, approximately ellipsoidal, cluster. It is tempting to conclude that there were two forgers at work, or perhaps one whose workshop had to be moved and consequently recalibrated.

The index plots of Mahalanobis distances from the two robust procedures in Figure 8 are, like those in Figure 4, again similar to each other. This time they also show 15 distances with values above the 99.99% band. The conclusion is that our procedure for the detection of outliers and that based on robust versions of Mahalanobis distances agree, in this example. However, the appropriate choice of the parameters of the robust procedure is crucial, the choice depending on the specific set of data. For data with appreciable contamination, an emphasis on too high an efficiency can lead to a procedure which fails to deliver any robustness.

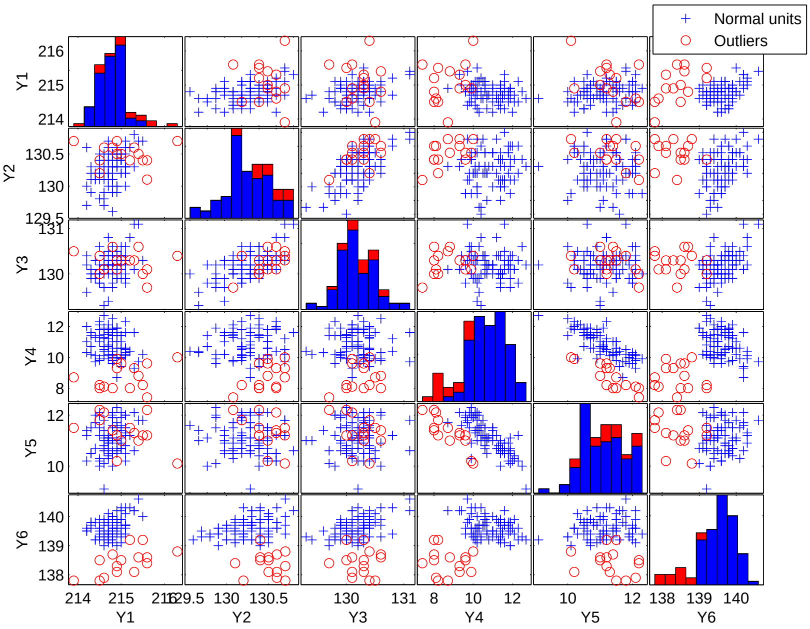

Figure 7. Forgeries. Scatterplot matrix showing the 15 outliers. In some bivariate plots, such as y_4 against y_6, the data seem to form two elliptical clusters

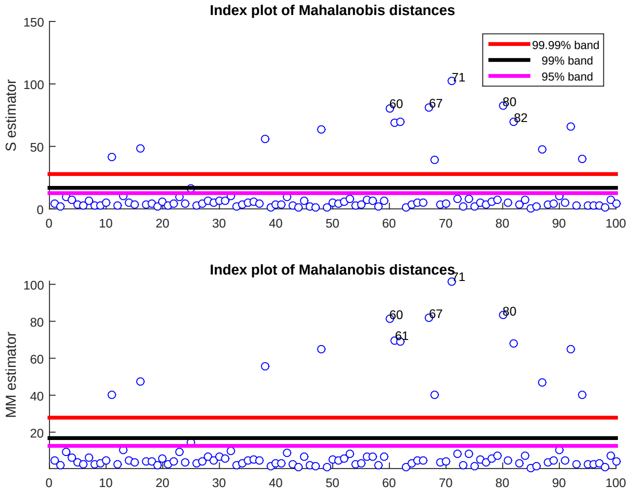

Figure 8. Forgeries. Plots of squared robust Mahalanobis distances. Upper panel, S-estimation; lower panel MM. These units are again numbered from 1 to 100

4. Clustering the Banknote Data

4.1 *The Random Start Forward Search*

The method illustrated here for single populations starts from a subset of m_0 observations, chosen in a robust manner. However, for data containing clusters, the initial subset $S^*(m_0)$ may lead to a search in which observations from several clusters enter the subset in a manner that does not reveal the clusters. The clustering structure is only revealed through searches from more than one starting point. We accordingly run many forward searches, each from a randomly selected starting point, and monitor the evolution of the values of $d_{\min}(m)$ as the individual searches progress. A random start can produce some very large distances in the initial stages of the search. However, the search can drop units from the subset as well as adding them. Some searches are consequently attracted to cluster centres. As the searches continue, the various random start trajectories converge to have subsets containing the same units. Converged trajectories cannot diverge again. Figure 9 is typical of those for many data structures. It shows that the initial distinct searches are rapidly reduced to a relatively few trajectories, some of which show marked peaks. These peaks provide information on the number and membership of the clusters. Usually, in the last one quarter to one third of the search, all trajectories have converged. Once tentative clusters have been identified, starting the FS with a subset of observations from a specific cluster results in the identification of observations from other clusters as outliers. The procedure for outlier detection may then be used to confirm cluster membership.

4.2 *The Banknote Data*

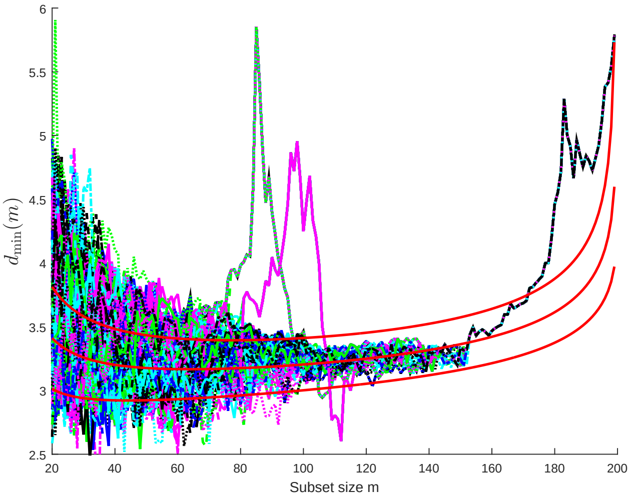

Figure 9. Swiss banknotes. Forward plot of $d_{\min}(m)$ from 500 random starts. The first peak is formed by 123 trajectories, the second by 189

Unlike Atkinson *et al.* [3], we now ignore the expert's classification of the notes into genuine and forged, and cluster the 200 observations using the random start FS. Figure 9 shows the forward plot of minimum Mahalanobis distances from 500 random starts with initial subsets of $v + 1 = 7$ observations. The plot shows two early peaks, one at $m = 85$ and the other at $m = 98$. There is also a third peak at $m = 182$, mainly caused by the outliers from the forgeries evident towards the end of the search in Figure 5.

From this plot we identify two tentative groups. If we take those observations in the subsets at the maxima of the two trajectories, we obtain groups with 86 and 99 units; units 85, 103 and 125 belong to both groups, being classified as belonging to one or the other in different searches. In the search leading to the group with 86 observations the three units enter in steps 86, 82 and 85. For the other search they enter in steps 83, 97 and 99. These observations thus enter both searches just at the end of the inclusion of units from single populations; they can be expected to be relatively remote from both cluster centres. Once the two tentative groups have been indicated, we use the procedure for outlier detection on the 99 units tentatively assigned to cluster 1 and to the 86 units assigned to cluster 2. In neither case do we obtain a signal, indicating that the clusters are homogeneous.

Figure 10 is the scatterplot matrix for the classification we have found, which recovers that found from the separate analyses of the genuine notes and the forgeries. As well as the two major clusters, we show the outlying units not included in either cluster. The

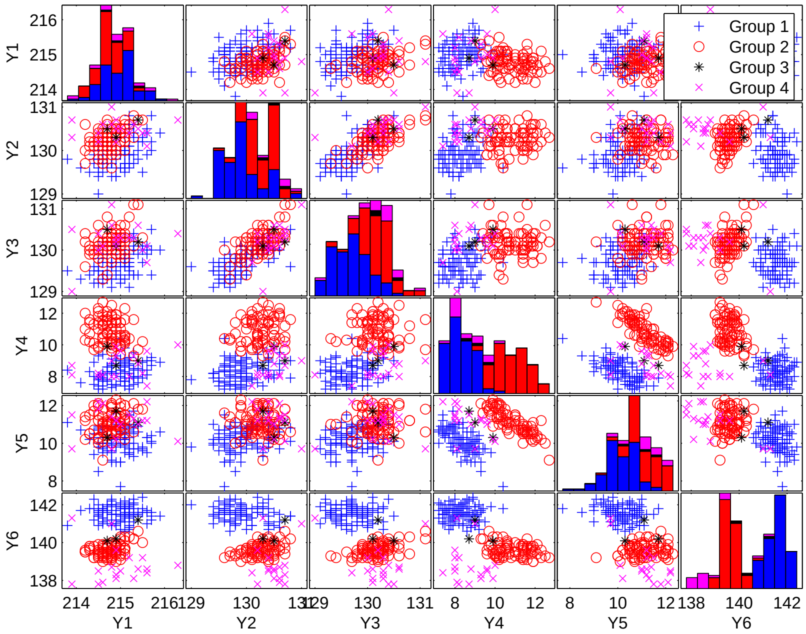

Figure 10. Swiss banknotes: final classification. Scatterplot matrix showing the two main clusters of data: $+$ - the genuine notes and $\circ$ - the forgeries. Also x - the unclustered units and * - the units that could belong to either cluster; see the plot for y_4 and y_5

majority of these are forged. However, the panel for y_1 and y_2 shows the two units from the genuine notes that remain unclassified. We also plot the three units that could belong to either group. As is to be expected, in many panels they lie between the two groups; an example is the panel for y_4 against y_5.

There are two computational parameters in the random start FS algorithm. The first is the number of observations m_0 in the initial subset. We have taken the minimum number, that is $v+1$, in order to maximise the number of searches that are attracted to individual cluster centres. Although the search can both add and delete observations from a subset, larger initial subsets reduce the probability of finding a subset of units from just one cluster. The second parameter is the number of random starts. Here we used 500. Of these, 123 gave trajectories passing through the first peak and 189 through the second in Figure 9. Clearly, in this case a smaller number of random starts would have revealed the structure. In §6.1 we use 300. Our experience has been that such numbers do not lead to excessive computation and enable the detection of even small clusters which could be overlooked with too few searches.

5. The Failure of Single Population Models

The random start FS succeeds in identfying clusters because it allows different models to be fitted to various subsets of the data. In this section we briefly exhibit the failure of standard robust methods that assume a single population, plus unstructured outliers, to identify clusters of observations.

Figure 11 shows index plots of squared robust Mahalanobis distances when S and MM estimation are applied to all 200 observations of the banknote data. In this plot the genuine notes are numbered from 1 to 100 and the forgeries from 101 to 200. For comparability with the samplewise bounds used in the FS, Bonferroni intervals are used in this plot.

The upper panel shows that, at the 1% level, S estimation detects 12 outliers; ten out

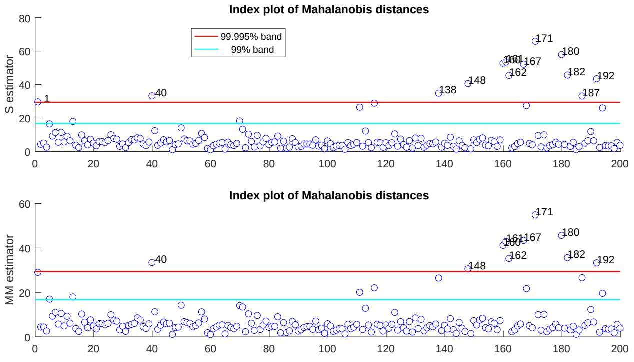

Figure 11. Swiss banknotes. Index plots of squared robust Mahalanobis distances; Bonferroni bounds

of the 15 outliers for the forgeries and units 1 and 40 from the genuine notes. This plot is virtually a concatenation of the index plots for the genuine notes and the forgeries from Figures 4 and 8. For MM estimation in the lower panel slightly fewer outliers are detected; nine out of the 15 for the forgeries and just unit 40.

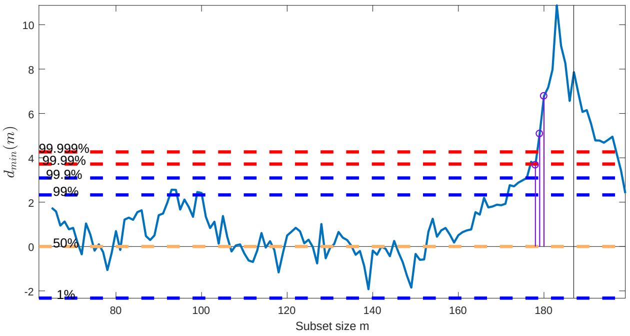

Figure 12. Swiss banknotes. Normal scores forward plot of $d_{\text{min}}(m, n)$. There is no indication of the two main clusters

The normal scores plot of Mahalanobis distances from the FS is in Figure 12. The FS detects slightly more outliers than the two methods summarised in Figure 11, namely 14 out of 15 outliers from the forgeries and both units 1 and 40. However, this is not the point. The right-hand portion of Figure 12 is similar to Figure 5; there is again no indication of the two groups identified by the random start FS. The failures arise because the single population methods find a centroid between the centres of the two main groups, so that neither group is particularly outlying. A detailed description is in §§7.3 and 7.4 of Atkinson *et al.* [3].

6. An Example with Many Clusters

6.1 *Automatic Cluster Detection with the Forward Search*

In order to provide a fully automatic method of cluster identification we slightly modify the above procedure. As before, we use an appreciable number of random starts (in our numerical example 300). With several clusters we obtain a figure like that for the Swiss banknotes in Figure 9 but with many peaks. We then sequentially delete the observations corresponding to the first peak, again running 300 forward searches after each deletion of a group of observations.

Because we are focused on the detection of clusters, we omit the second step, that of resuperimposition, outlined in §3.2. We thus do not expect that our method will initially identify all the observations in a cluster. Rather, we require an automatic method of determining cluster centres. Once they have been determined, single population forward searches can proceed from each centre, identifying all observations in the cluster and classifying the remaining observations as outliers, in the sense of not belonging to the cluster being studied.

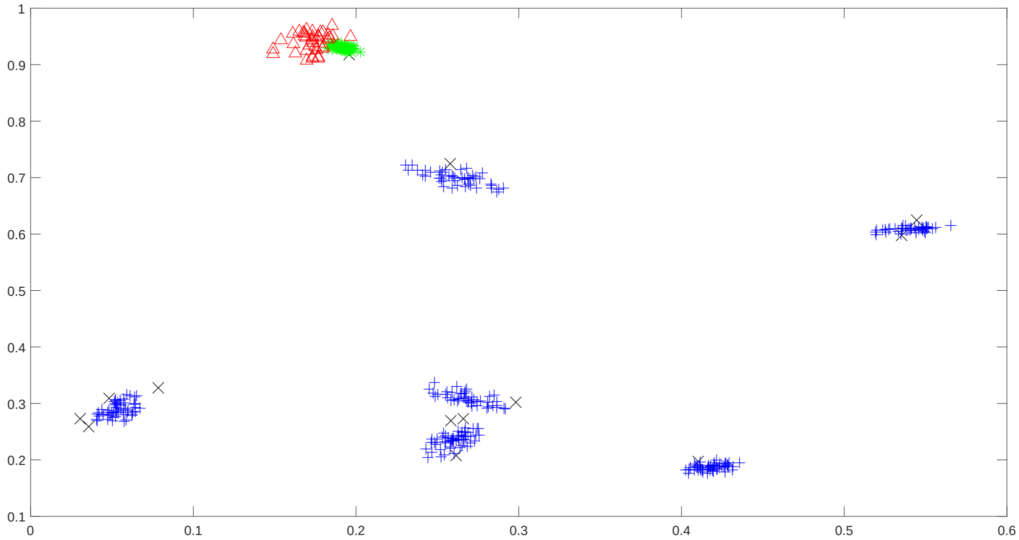

Figure 13. Data with eight clusters. Identification of cluster centres with repeated random start forward searches: + units included in clusters; **X** units not included at this stage. For the pairs of nearby clusters labelled 1 and 2 in Figure 14, the method correctly identifies the two clusters in each pair

In our example there are 400 two-dimensional observations divided into eight clusters. As figure 13 shows, there are four clusters well separated from each other and from two pairs of nearby clusters. The figure also shows the results of our automatic identification of cluster centres. This was achieved using exceedances over the 99.9999% envelope. All observations up to, but not including the first excedance were taken to form the first cluster. These observations were then deleted and the random start FS repeated to identify the next cluster, the observations forming which were then also deleted and so forth. As the figure shows, the more extreme members of the clusters are, at this stage, identified as outliers. Once the cluster centres have been established, the FS can be used, with the inclusion of Step 2, to determine cluster membership for each outlier. Since this procedure of cluster confirmation is the same as that exemplified in Atkinson *et al.* [6] we do not repeat it here.

We also used the automatic procedure on the banknote data where we obtained two groups plus outliers. These were the smaller group of forgeries plus, as would be expected from the extreme envelope used in Step 1, the more outlying observations from the other two clusters. Step 2 searches with resuperimposition are again needed to establish precise clustering.

6.2 *Automatic Cluster Detection with Information Criteria*

The method of model-based clustering described by Fraley and Raftery [13] and McLachlan and Peel [22] is to fit a mixture of multivariate normal distributions to the data. In comparison to the random start FS this is a seemingly deterministic procedure. However the likelihood surface contains many local maxima and numerical maximization in the Matlab routine `gmdistribution.fit` requires many iterations from randomly selected starting points. The number of these affects the results obtained.

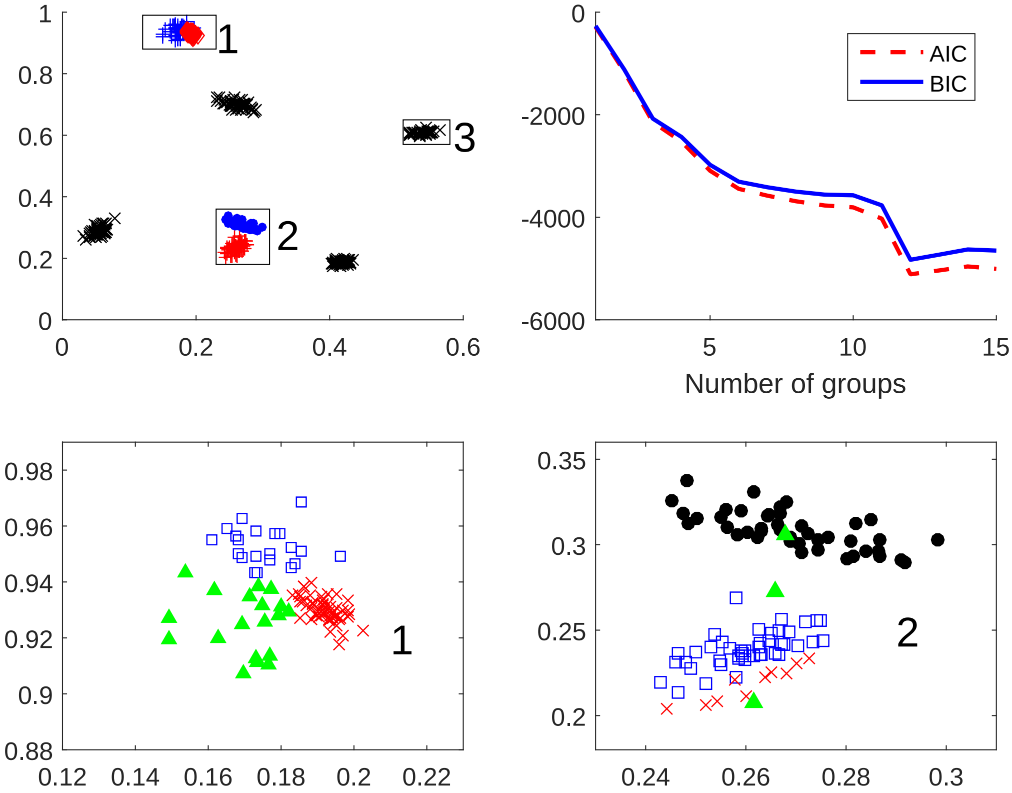

Figure 14. Data with eight clusters. Clustering with multivariate normal distributions using Matlab routine `gmdistribution.fit`; 50,000 starting values. Upper-left panel, the 11 indicated clusters. Lower-left panel, Cluster pair 1 split into three groups. Lower-right panel, Cluster pair 2 split into four groups. Upper-right panel, plots of AIC and BIC against number of groups.

Figure 14 shows the results with 50,000 starting values; different numerical searches are required for each postulated number of groups, which we allowed to increase to 15. The plot of AIC and BIC in the upper-right panel of the figure indicates that there are 12 clusters. However, after allocation of units to clusters there remain the 11 shown in the top-left panel. Highlight 1 shows that the two original clusters have been split into three; Highlight 2 shows that, in this case, the two original clusters have been split into four, one of which contains three collinear observations, plotted as (green) triangles.

The tendency of the method to split clusters into subclusters is well documented, for example Fraley *et al.* [14]. A second difficulty addressed by [14] is the growth in the computational burden with increasing sample size and number of clusters.

We therefore repeated our model-based clustering with 5,000 iterations. The results in Figure 15 again show 11 clusters, but with some differences. The top-left panel shows, for both AIC and BIC, a clear indication for 11 groups rather than the 12 in Figure 14. However, the values of BIC at these points are rather different. The minimum value for 50,000 iterations is over 200 less than that for 5,000 iterations. With the six additional parameters in the model from introducing a new two-dimensional cluster, this value is highly significant when compared with χ^2_6. The use of 50,000 iterations has indeed found a better solution to maximizing the likelihood. It is interesting that the values of BIC for the two iteration sizes are indistinguishable up to this number of groups.

The clusters are also not identical. The three unhighlighted clusters in Figure 14 are correctly identified. However, Highlight 1 in Figure 15 shows that the two groups found

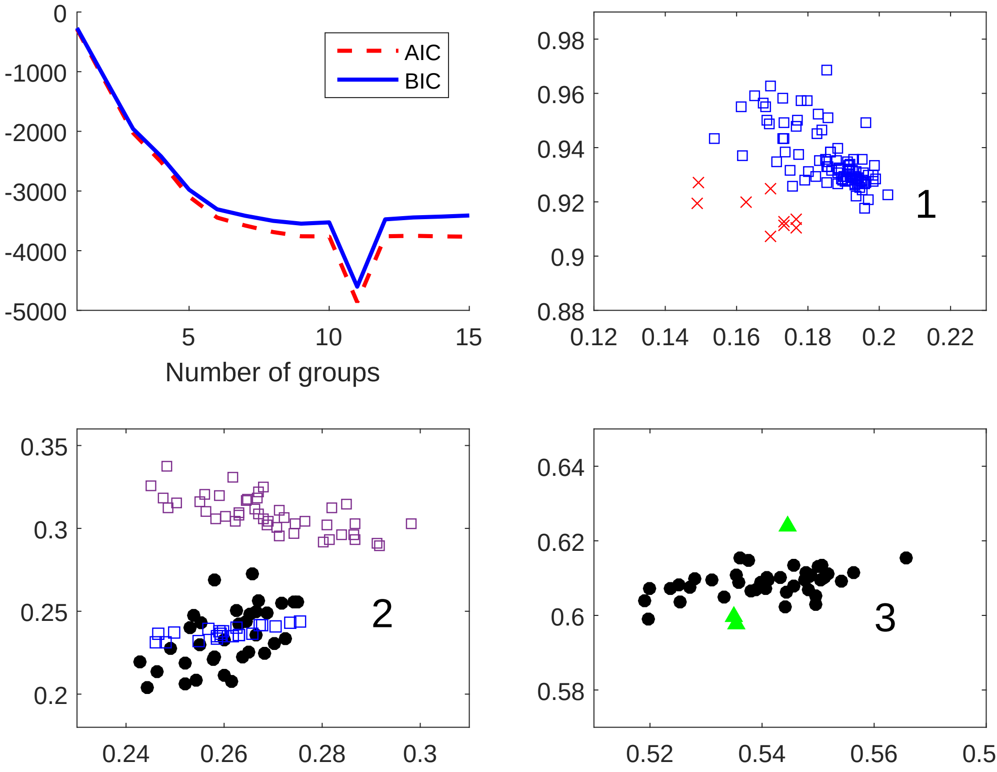

Figure 15. Data with eight clusters. Clustering with multivariate normal distributions using Matlab routine `gmdistribution.fit`; 5,000 starting values. Upper-left panel, plots of AIC and BIC against number of groups; 11 groups are indicated. Upper-right panel, cluster pair 1 split into two other groups. Lower-left panel, Cluster pair 2 split into three other groups. Lower-right panel, Highlight 3, the cluster split into two groups

are not the original two clusters identified by the FS and shown in Figure 13. Highlight 2 shows 3 clusters, rather than the 4 in the same highlight for Figure 14; there are correctly 2. Finally Highlight 3 shows that one cluster has been split into two parts, one of which contains three virtually collinear observations. Virtually collinear clusters are a general problem in cluster analysis. García-Escudero *et al.* [15] find it necessary, in their robust procedure, to apply an eigenvalue constraint to avoid such configurations.

These differences in solutions are rather disconcerting. In this particular example they contrast poorly with the behaviour of our FS method for cluster identification. The use of the single population FS could, for example, be used to examine these cluster and coalesce some of them, a kind of post processing that we need following our FS-based identification of clusters. Further, we note that this variability occurs in an algorithm which, unlike the FS, is not adapted to the presence of outliers and is therefore totally non-robust.

6.3 *Numerically Efficient Forward Search*

Repeated iteration of the random start FS to identify the number of clusters obviously requires efficient computational algorithms, especially when the sample size and the number of clusters are moderate to large. The same is true when the number of random initializations has to be increased, e.g. in order to better explore the space of possible subsamples (Hawkins and Olive [17]). For this reason our implementation takes advantage of the fast updating rules for parameter estimation developed by Riani *et al.* [27]. These rules are recursive, exploiting at each step of the FS the information coming from the previous step, and do not require data sorting, matrix inversion or QR decomposition.

Our updating rules yield fast computation of likelihood contributions whose running time increases almost linearly with the sample size. Therefore, we do not see application to large data sets to be more challenging for our method than for alternative approaches of similar complexity, such as the non-robust model-based clustering approach considered in §6.2. Popular data mining algorithms for cluster analysis, such as k-means clustering, instead trade statistical sophistication and efficiency for computational time. They achieve applicability on huge masses of data, but at the expense of a great simplification in the data generating process (e.g., by assuming spherical covariances in all the groups) and of the potentially extreme influence of a few contaminant observations. The development of robust multivariate methods suitable for high dimensional data, when $n < v$, appears to be a challenging and largely unsolved research goal (Wilms and Croux [30]).

7. Discussion

The FS analysis of the banknote data has led to the discovery of three clusters, rather than the two in the original description of the data. The random start FS analysis of the complete data set, ignoring the two categories, recovers the division into genuine notes and forgeries, as well as identifying the two clusters in the forgeries, most clearly seen in the plot of y_4 against y_6.

Our results in §3 showed that, for both the genuine notes and the forgeries, the robust multivariate methods using either S or MM-estimation led to the same conclusions as the FS. However, application of these methods to all 200 observations fails to indicate the presence of two or three groups. The index plots of robust squared Mahalanobis distances from a single fit to all the data are somewhat like Figures 4 and 8 plotted next to each other. Outliers are identified, particularly those from the second group, the forgeries. But, in order to detect the presence of clustered data, robust fits are needed to the identifiable subsets into which the data fall. The random start FS enables identification of such potential clusters.

Our modification of this procedure to provide an automatic method for determining cluster centres behaved well in our simulated example and appears promising. We look forward to a more extended investigation of its properties than is appropriate to report here.

There have been several other analyses of the banknote data, in addition to those mentioned in §1. Some, such as Pison *et al.* [24], use the forgeries to illustrate the behaviour of a particular method, in their case robust factor analysis, in the presence of outliers. Others, for example Croux and Joossens [11], use the complete data set, including the expert's categorisation, to illustrate robust discriminant analysis. The analysis of Cook and Yin [9] is more informative about the structure of the data. They use recently developed methods of visualization and dimension reduction in discriminant analysis to show that the problem has two dimensions; their Figure 7 reveals a structure similar to that in our plot of y_4 against y_6 in Figure 10.

The analysis of the Swiss banknotes exemplifies the power of the random start FS algorithm for clustering in the presence of outliers and the absence of precise knowledge of the number of clusters. However, if clusters overlap, it is often not possible to ascribe units with certainty to a specific cluster. A probabilistic allocation may be needed (McLachlan and Peel [22]). One approach, combining the FS and normal mixture models, is suggested by Calò [8].

The calculations for this paper used the FSDA Matlab toolbox (`http://www.riani.it/MATLAB`) which includes tools for brushing and linking plots. For Example 2, the centroids and covariance matrices were generated using the FSDA routine `MixSim.m`, derived from that of Maitra and Melnykov [20], imposing a maximum expected overlap of 0.01. The data were generated using routine `simdataset.m`. The seed for the random number generator was `rng(948)`.

References

[1] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723.

[2] 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.

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

[4] Atkinson, A. C., Riani, M., and Cerioli, A. (2006). Random start forward searches with envelopes for detecting clusters in multivariate data. In S. Zani, A. Cerioli, M. Riani, and M. Vichi, editors, *Data Analysis, Classification and the Forward Search*, pages 163–171. Springer-Verlag, Berlin.

[5] Atkinson, A. C., Riani, M., and Cerioli, A. (2010). The forward search: theory and data analysis (with discussion). *Journal of the Korean Statistical Society*, **39**, 117–134. doi:10.1016/j.jkss.2010.02.007.

[6] Atkinson, A. C., Cerioli, A., Morelli, G., and Riani, M. (2014). Finding the number of disparate clusters with background contamination. In B. Lausen, S. Krolak-Schwerdt, and M. Böhmer, editors, *Data Science, Learning by Latent Structures, and Knowledge Discovery*, pages 29–42. Springer-Verlag, Heidelberg.

[7] Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, **3**, 1–27.

[8] Calò, D. (2008). Mixture models in forward search methods for outlier detection. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, and R. Decker, editors, *Data Analysis, Machine Learning and Applications. Springer series in "Studies in Classification, Data Analysis and Knowledge Organization"*, pages 103–110. Springer Verlag, Berlin.

[9] Cook, R. D. and Yin, X. (2001). Dimension reduction and visualization in discriminant analysis (with discussion). *Australian and New Zealand Journal of Statistics*, **43**, 147–199.

[10] Coretto, C. and Hennig, C. (2017). Robust improper maximum likelihood: Tuning, computation, and a comparison with other methods for robust Gaussian clustering. *Journal of the American Statistical Association*, **112**, (In press).

[11] Croux, C. and Joossens, K. (2005). Influence of observations on the misclassification probability in quadratic discriminant analysis. *Journal of Multivariate Analysis*, **96**, 384 – 403.

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

[13] Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, **97**, 611–631.

[14] Fraley, C., Raftery, A. E., and Wehrens, R. (2005). Incremental model-based clustering for large datasets with small clusters. *Journal of Computational and Graphical Statistics*, **14**, 529–546.

[15] García-Escudero, L. A., Gordaliza, A., Matrán, C., and Mayo-Iscar, A. (2011). Exploring the number of groups in model-based clustering. *Statistics and Computing*, **21**, 585–599.

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

[17] Hawkins, D. M. and Olive, D. J. (2002). Inconsistency of resampling algorithms for high-breakdown regression estimators and a new algorithm (with discussion). *Journal of the American Statistical Association*, **97**, 136–159.

[18] Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data. An Introduction to Cluster Analysis*. Wiley, New York.

[19] Klinke, S. (1997). *Data Structures for Computational Statistics*. Physica-Verlag, Heidelberg.

[20] Maitra, R. and Melnykov, V. (2010). Simulating data to study performance of finite mixture modeling and clustering algorithms. *J Comput Graph Stat*, **19**, 354–376.

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

[22] McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York.

[23] Morris, K. and McNicholas, P. D. (2013). Dimension reduction for model-based clustering via mixtures of shifted asymmetric Laplace distributions. *Statistics and Probability Letters*, **83**, 2088–2093.

[24] Pison, G., Rousseeuw, P. J., Filzmoser, P., and Croux, C. (2003). Robust factor analysis. *Journal of Multivariate Analysis*, **84**, 145 – 172.

[25] Riani, M., Atkinson, A. C., and Cerioli, A. (2009). Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society, Series B*, **71**, 447–466.

[26] Riani, M., Cerioli, A., Atkinson, A. C., and Perrotta, D. (2014). Monitoring robust regression. *Electronic Journal of Statistics*, **8**, 642–673.

[27] Riani, M., Perrotta, D., and Cerioli, A. (2015). The forward search for very large datasets. *Journal of Statistical Software*, **66**, Code Snippet 4.

[28] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

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

[30] Wilms, I. and Croux, C. (2015). Sparse canonical correlation analysis from a predictive point of view. *Biometrical Journal*, **57**, 834–851.

[31] Yohai, V. J. (1987). High breakdown-point and high efficiency estimates for regression. *The Annals of Statistics*, **15**, 642–656.