Finding the Number of Disparate Clusters with Background Contamination

Anthony C. Atkinson, Andrea Cerioli, Gianluca Morelli, and Marco Riani

Abstract The Forward Search is used in an exploratory manner, with many random starts, to indicate the number of clusters and their membership in continuous data. The prospective clusters can readily be distinguished from background noise and from other forms of outliers. A confirmatory Forward Search, involving control on the sizes of statistical tests, establishes precise cluster membership. The method performs as well as robust methods such as TCLUST. However, it does not require prior specification of the number of clusters, nor of the level of trimming of outliers. In this way it is "user friendly".

1 Introduction

It is now widely recognized that contamination can strongly affect the results of clustering methods if it is not properly taken into account. The first attempts in this direction mainly aimed at protecting against *contamination from noise*. In fact, the inclusion of noise variables can have dramatic masking consequences on the data structure recovered by distance-based clustering methods. Pioneering studies of such effects were presented by Milligan (1980) and Fowlkes et al. (1988). In a model-based clustering framework, noise can be described through a uniformly distributed component and Fraley and Raftery (2002) suggest a unified approach for dealing with it; see also Coretto and Hennig (2010). However, uniform background noise is not the only type of departure from "ideal" conditions against which one may want to protect. Typical instances that need to be accommodated in practice include both *extreme outliers*, e.g. anomalous observations due to undetected changes in the physical process generating the data or to measurement errors, and *intermediate outliers*, e.g. observations not firmly belonging to any of the

A.C. Atkinson
Department of Statistics, London School of Economics, London WC2A 2AE, UK
e-mail: a.c.atkinson@lse.ac.uk

A. Cerioli (✉) • G. Morelli • M. Riani
Dipartimento di Economia, Università di Parma, Parma, Italy
e-mail: andrea.cerioli@unipr.it; gianluca.morelli@unipr.it; mriani@unipr.it

B. Lausen et al. (eds.), *Data Science, Learning by Latent Structures, and Knowledge Discovery*, Studies in Classification, Data Analysis, and Knowledge Organization, DOI 10.1007/978-3-662-44983-7_3

established groups (Riani et al. 2014). Robust clustering methods, for a thorough review of which we refer to Gallegos and Ritter (2009) and to García-Escudero et al. (2010), aim at addressing all these issues, by neutralizing the effect of both types of outliers and that of noise.

Robust clustering is often defined in a model-based framework in which observations are assumed to come from distinct multivariate populations. Formally, let $y = \{y_1, \ldots, y_n\}$ denote a sample of v-variate observations $y_i = (y_{i1}, \ldots, y_{iv})'$, for $i = 1, \ldots, n$. In case of no contamination, a popular approach is to search for the partition into K groups that maximizes the "classification likelihood" function

$$\prod_{k=1}^{K} \prod_{i=1}^{n} [\phi(y_i, \theta_k)]^{z_{ik}}, \tag{1}$$

where $\phi(y_i, \theta_k)$ is a v-variate density depending on the multidimensional parameter θ_k, and z_{ik} is the indicator variable taking the value 1 if y_i belongs to group k and 0 otherwise. The value of K, the number of groups, is assumed to be known when fitting the classification likelihood function.

When outliers and noise are present, a trimmed version of (1) has been proposed by García-Escudero et al. (2008). Their robust approach, called TCLUST, is based on maximizing the function

$$\prod_{k=1}^{K} \prod_{i \in R_\alpha} [\pi_k \phi(y_i, \theta_k)]^{z_{ik}}, \tag{2}$$

where $\phi(\cdot)$, z_{ik} and K have the same meaning as in (1), $\pi_k \in [0, 1]$ is an unknown weight taking into account the specific size of group k, $0 \leq \alpha \leq 0.5$ is the *trimming level* and R_α is a subset of the set of indices $\{1, \ldots, n\}$ whose cardinality is $\lfloor n(1-\alpha)\rfloor + 1$. TCLUST clustering is currently implemented in an R package (Fritz et al. 2012) and in the FSDA toolbox for Matlab (Riani et al. 2012).

The TCLUST methodology has good theoretical properties and has proven to be effective in practical applications (see, e.g., Cerioli and Perrotta 2014). Being a model-based approach, it requires many *a priori* choices on behalf of the user. First, it is necessary to specify the form of the v-variate density $\phi(y_i, \theta_k)$. Although alternative distributions have been recently introduced (see Lee and McLachlan 2013 and the subsequent discussion), the typical choice in (2) is the multivariate normal density function, which is also at the heart of the non-robust proposal (1). Therefore, θ_k corresponds to the mean vector μ_k and to the covariance matrix Σ_k for group k. These parameters are estimated in TCLUST through a complex iterative EM-like procedure, also involving the restrictions

$$\frac{\max_{l=1,\ldots,v} \max_{k=1,\ldots,K} \lambda_l(\hat{\Sigma}_k)}{\min_{l=1,\ldots,v} \min_{k=1,\ldots,K} \lambda_l(\hat{\Sigma}_k)} \leq c. \tag{3}$$

where $\lambda_l(\hat{\Sigma}_k)$ is a generic eigenvalue of the estimated covariance matrix of group k and $c \geq 1$ is a fixed constant which defines constraints on the shape and, implicitly, on the size of the K clusters. For example, the value $c = 1$ corresponds to imposing spherical groups of constant variance, while the limiting case $c \to \infty$ gives unconstrained heterogeneous clusters. The specific value of c is a tuning parameter that must be chosen by the user.

Similarly, both K and α are assumed to be known when fitting the objective function (2). García-Escudero et al. (2011) suggest the use of the so-called classification trimmed likelihood curves (TLC), as useful exploratory tools for selecting K and α from the data. These curves are based on (2) and measure the gain achieved by allowing a unitary increase in the number of groups for a given value of α. However, there might be applications where the evidence provided by TLC is not so clear-cut (Morelli 2013). The curves also depend on the selected value of c. Therefore, there is a crucial interplay among the choices made by the user with respect to the different tuning parameters required for fitting the trimmed likelihood function (2). The development of a unified data-driven framework for assisting these choices is indeed an area of active research.

The goal of this contribution is to compare the performance of TCLUST, with respect both to the selection of the tuning parameters and to cluster recovery, with that obtained through the Forward Search (FS). The FS is a robust diagnostic method which has powerful outlier detection properties (Riani et al. 2009), but also good potential for clustering purposes (Atkinson et al. 2006; Atkinson and Riani 2007). Since it does not aim at finding the maximum of a complex objective function, exploratory use of the FS does not require an explicit choice of a tuning constant like c in (3). Also the density $\phi(y_i, \theta_k)$ is only used for confirmatory purposes, when formal hypothesis testing is required, and not for cluster assignment at an exploratory stage. In fact, the technique adopts some of the classical tools developed for the customary normal-based multivariate model, like Mahalanobis distances, but in its exploratory form the user has the ability for visual detection of departures from such a model. The FS assumes the existence of one or more elliptical populations for the "good" part of the data, which are taken as a reference benchmark against which the observations are compared. Therefore, although computations require specification of a multivariate normal density, as in TCLUST, the diagnostic use of the technique can effectively lead to detection of many alternative structures, such as extreme and intermediate outliers, or skew distributions. Even more importantly, the FS provides simple data-driven rules for the choice of K and α that do not rely on fitting the trimmed likelihood function (2). Therefore, it may be considered as a practical "user-friendly" alternative to complex robust model-based clustering methods. We will see in a difficult example that it can provide powerful information on the data structure, being virtually as illuminating as TCLUST when the latter is appropriately tuned, while requiring only minimal intervention from the user.

The structure of the paper is as follows. Section 2 introduces the FS, with Sect. 2.3 describing the random start FS that we use to identify cluster structure. In Sect. 3 this methodology is applied to a 2,000 observation example from

García-Escudero et al. (2011). Initial cluster identification from the search with random starts is in Sect. 3.1. Section 3.2 uses plots of individual Mahalanobis distances to illustrate the structure of the clusters. Confirmation of this structure, using tests of specified size, is in Sect. 3.3. The paper concludes with a comparison of the analysis of the same data using the robust TCLUST procedure.

2 The Forward Search

2.1 *Outlier Detection and Clustering*

The main tool in clustering multivariate data with the forward search (FS) is the detection of outliers. For a sample believed to come from a single multivariate normal population, perhaps with outliers, the method described by Riani et al. (2009) has good statistical properties (size and power of the outlier test). As described, it does however require intervention of the data analyst. In this paper we introduce the automatic version of this procedure extended to groups.

The method for a single population starts from a robustly chosen subset of m_0 observations. The subset is increased from size m to $m+1$ by forming the new subset from the observations with the $m+1$ smallest squared Mahalanobis distances. For each m ($m_0 \le m \le n-1$), we test for the presence of outliers, using the observation outside the subset with the smallest Mahalanobis distance.

With data coming from two or more populations, starting with a subset of observations in one of the clusters results in some observations from other clusters being identified as outliers. Atkinson et al. (2004, Sect. 3.4) illustrate this point in an analysis of 200 observations on Swiss banknotes. In that case the banknotes had already been preliminarily classified into two groups, "genuine" and "forgeries". However, in the general clustering problem considered here, the number of clusters is not known, let alone any approximate cluster membership, so that there is no simple robust path to an initial subset. This difficulty was overcome by Atkinson et al. (2006), who suggested a "random-start" forward search in which many initial subsets of size m_0 are selected at random. Monitoring the behaviour of the resulting forward searches leads to an indication of the number and membership of clusters, which can then be refined by the automatic outlier detection procedure. The details are described in Sect. 2.3.

2.2 *Mahalanobis Distances*

In the forward search we estimate the parameters μ and Σ of the v-dimensional multivariate normal distribution of y by the standard unbiased estimators from a subset of m observations, yielding estimates $\hat{\mu}(m)$ and $\hat{\Sigma}(m)$. From this subset we

obtain n squared Mahalanobis distances

$$d_i^2(m) = \{y_i - \hat{\mu}(m)\}' \hat{\Sigma}^{-1}(m)\{y_i - \hat{\mu}(m)\}, \qquad i = 1, \dots, n. \tag{4}$$

Let $S^*(m)$ be the subset of size m found by the search. To detect outliers we use the minimum Mahalanobis distance amongst observations not in the subset

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

In order to test for outliers we need a reference distribution for $d_i^2(m)$ in (4) and hence for $d_{\min}(m)$ in (5). If we estimated Σ from all n observations, the statistics would have an F distribution. However, in the search we select the central m out of n observations to provide the estimate $\hat{\Sigma}(m)$, so that the variability is underestimated. To allow for estimation from this truncated distribution, Riani et al. (2009) provide a consistency factor to make the estimate approximately unbiased. They also provide an order-statistic argument for the distribution of $d_{\min}(m)$ which obviates the need for simulations in the calculation of the reference distribution. As the search progresses, we perform a series of outlier tests, one for each $m \ge m_0$. To allow for the problem of multiple testing, we use the outlier detection rule of Riani et al. (2009) which depends on the sample size and on the calculated envelopes for the distribution of the test statistic. The results of the FS are conveniently presented graphically through forward plots of quantities of interest as functions of m. We illustrate such plots in Sect. 3.3.

The expected values of the minimum Mahalanobis distances increase rapidly towards the end of the search, since we are looking at the distribution of the largest ordered Mahalanobis distances. In some cases, more informative plots, which we do not use for testing, come from the use of the scaled Mahalanobis distances

$$d_i^{sc}(m) = d_i(m) \times \left(|\hat{\Sigma}(m)| / |\hat{\Sigma}(n)| \right)^{1/2v}, \tag{6}$$

where $\hat{\Sigma}(n)$ is the estimate of Σ at the end of the search. Examples are in Sect. 3.2 with, again, the details in Riani et al. (2009).

2.3 *Random Start Forward Searches*

If there are clusters in the data, the robustly chosen initial subset m_0 may lead to a search in which observations from several clusters enter the subset haphazardly in such a way that the clusters are not revealed. Searches from more than one starting point are necessary to reveal the clustering structure. We therefore instead run many forward searches, say R (500 in our example), from randomly selected starting points, monitoring the evolution of the values of $d_{\min}(m, j)$ for each search j, ($j = 1, \dots, R$). The criterion used for moving from step m to $m + 1$ is also reminiscent of the Mahalanobis Fixed Point Clusters procedure of Hennig and Christlieb (2002), where, however, a χ^2_v threshold is used for defining the fitting subset.

At the beginning of the search, a random start produces some very large distances. But, because the search can drop units from the subset as well as adding them, some searches are attracted to cluster centres. As the searches progress, the various random start trajectories converge, with subsets containing the same units. Once trajectories have converged, they cannot diverge again. As we see in Fig. 2, which is typical of those for many data structures, the search is rapidly reduced to only a few trajectories. It is these that provide information on the number and membership of the clusters. Typically, in the last third of the search all trajectories have coalesced into one. Interpreting such figures requires intervention from the data analyst. However, once prospective clusters have been identified, the procedure for cluster confirmation is automatic.

3 An Example of García-Escudero et al.

3.1 *Random Start Forward Search*

García-Escudero et al. (2011) provide a simulated data test case of 1,800 observations simulated from two-dimensional normal distributions, with 200 outliers generated from a uniform distribution. The data are plotted in Fig. 1. The plot appears to show one clear tight cluster, one moderately clearly defined cluster, a

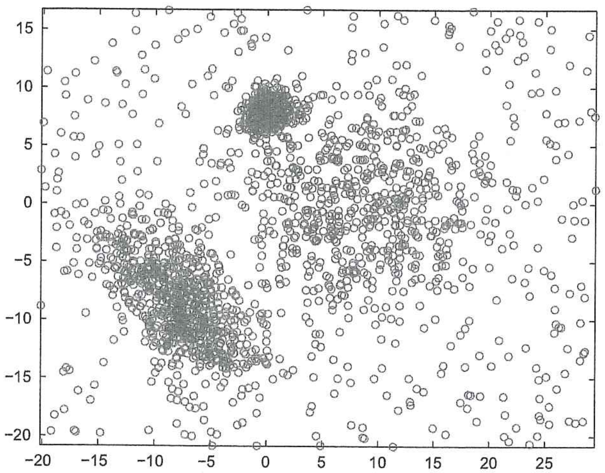

Fig. 1 The 2,000 simulated observations of García-Escudero et al. (2011), which include background contamination

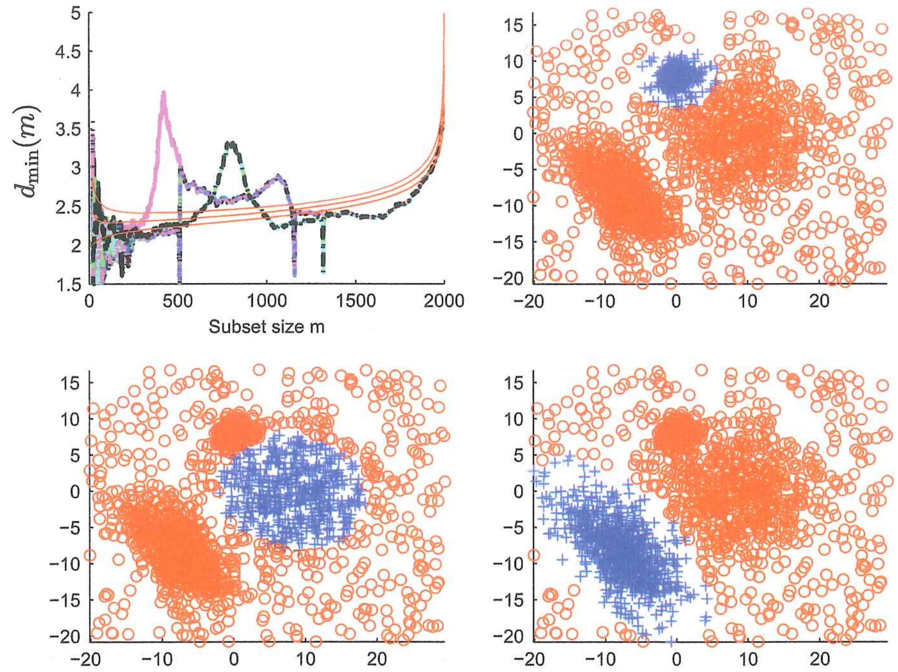

Fig. 2 Preliminary cluster identification. *Top left-hand panel*, the trajectories of minimum Mahalanobis distances $d_{\min}(m, j)$ from 500 random start forward searches; the vertical trajectories occur when clusters merge. *Top right-hand panel*, scatterplot, with preliminary Group 1 highlighted; the first 420 units from the sharpest peak. *Bottom left-hand panel*, preliminary Group 2; the first 490 units from the lowest trajectory. *Bottom right-hand panel*, preliminary Group 3, the first 780 units from the central peak. Highlighted group in each panel indicated +

third more dispersed cluster and a background scatter. The clusters are thus of disparate sizes, orientations and shapes (as measured by the eigenvalue ratio for each cluster) and there is background contamination. A strange feature is that this background contamination seems to contain a hole for large values of y_2 and slightly smaller values of y_1.

A traditional robust clustering method, such as PAM (Partitioning Around Medoids, Kaufman and Rousseeuw 1990) when $K = 3$ finds three large clusters, which include all the observations. A strong advantage of the FS is that we do not have to cluster all observations, but can declare a data-chosen number of them to be outliers.

The top left-hand panel of Fig. 2 shows the forward plot of minimum Mahalanobis distances from 500 random start forward searches. The structure is exemplary in its clarity. By the time $m = 300$ there are just three distinct trajectories, which we examine to see if they correspond to clusters. The vertical trajectories occur at values of m when two clusters merge. At this stage in the analysis of cluster structure we are not concerned about using tests of the correct size, but only in the detection of structure. Tests with carefully controlled statistical properties are, however, important in the confirmation of cluster membership (Sect. 3.3).

We start with the first, highest, peak, which we call Group 1. The peak is caused by $d_{\min}(m)$ indicating the remoteness of the nearest outlier to the fitted cluster. Eventually, as more remote observations are included in the subset of size m, the parameter estimates are sufficiently corrupted that the next nearest observation does not seem so far away, and the peak declines. We "interrogate" the search just before or after the peak. Here we choose $m = 420$, which is before the three trajectories collapse into two. The top right-hand panel of Fig. 2 shows a scatterplot of those units in the subset at this value of m for searches with this trajectory. These searches have identified the small, tight cluster. With a small dispersion matrix, even observations which are close to the cluster in the Euclidean norm have appreciable Mahalanobis distances. That we continued a little after the peak in the search is shown in the plot by the slightly more remote sample data points surrounding the central core.

We now repeat the procedure for the other two trajectories. At $m = 490$ the lowest trajectory yields the preliminary clustering shown in the bottom left-hand panel of Fig. 2. This relatively dispersed group does not have a clear boundary, unlike Group 1. We do not get a peak in the trajectory because Groups 1 and 2 merge shortly after $m = 490$. Because Group 2 has a larger dispersion matrix than Group 1, units in Group 1 do not, as the search progresses, seem particularly remote from Group 2. Finally, we look at the third trajectory at $m = 780$. This is a relatively isolated group, the trajectory for which rises to a clear peak before declining. Its membership is shown in the bottom right-hand panel of the figure.

There remains one final feature of interest in Fig. 2. In a forward plot of distances calculated from a sample from a normal population, the curves rise at the end, as the envelopes in Fig. 2 show. However, when all observations are fitted, there is no evidence of any outliers; the sample trajectory lies within the envelope. More importantly, for the last third or more, the trajectory lies appreciably below the envelopes, suggesting that the tails of the single multivariate distribution being fitted are too short; here, this is an indication of uniform outliers over a fixed region.

3.2 *Plots of Individual Scaled Distances*

Further insight into the structure of the data and the workings of the FS can be obtained from the forward plots of individual Mahalanobis distances, that is the trajectory of the distances for each observation. For this we use the scaled distances (6) which, unlike the distances in the top-left panel of Fig. 2, do not increase appreciably towards the end of the search.

With clustered data, there are many possible series of plots. We divide the observations into four groups based on the classification shown in Fig. 2 and see how the different groups behave during a search starting in tentative Group 1. Figure 3 shows the very different behaviour of these distances. The top left-hand panel is for the members of Group 1. Initially the distances have a distribution ranging upwards

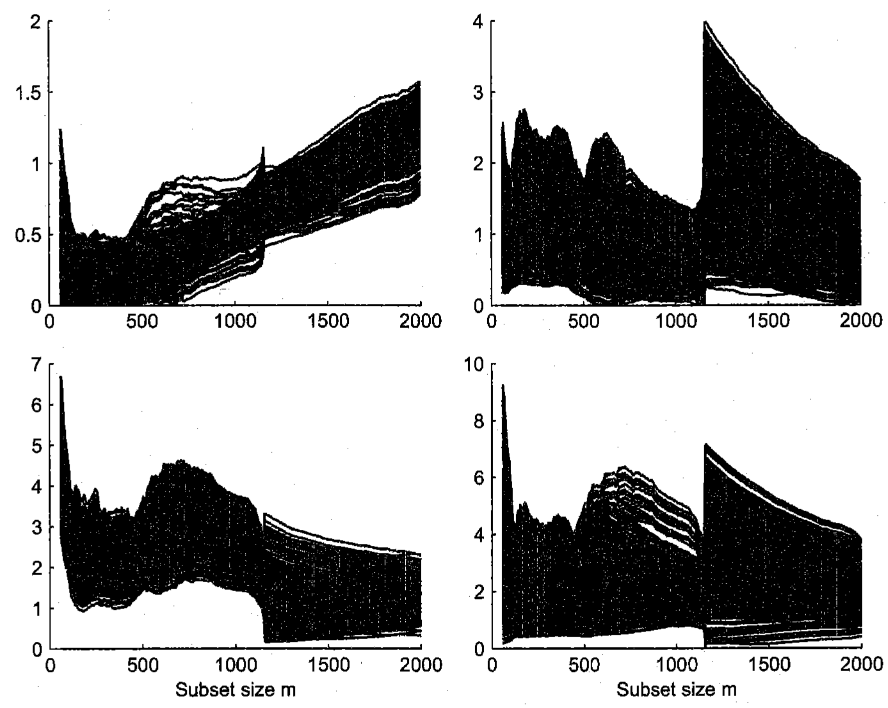

Fig. 3 Forward plots of individual scaled Mahalanobis distances $d_i^{sc}(m)$ from the preliminary classification shown in Fig. 2 when the FS starts in Group 1. Reading across, Groups 1, 2, 3 and zero (the outliers). Note the differing vertical scales in the panels

from zero, whereas the other two groups and the outliers, all remote from the cluster centre, have distances that start away from zero. There is an abrupt change, shown by the virtually vertical line in the top left panel of Fig. 2 around $m = 510$, when the centre of the fitted single population switches near to that of Group 2. The observations in Group 1 become increasingly remote. There is another dramatic change at around $m = 1{,}150$ when Groups 2 and 3 begin to combine. At this point, the outliers, with distances plotted in the bottom right-hand panel, also become less remote. An interesting feature of the last part of the search is that there are virtually no very small Mahalanobis distances; the cluster centre is lying between groups so that all distances are larger than they should be for a normal population. Only, towards the very end of the search, are there some small values in Group 2. A QQ-plot of the squared distances from fitting all observations might, therefore, indicate that all was not well, without being informative about the precise departure from assumptions.

3.3 *Cluster Confirmation and Automatic Outlier Detection*

We employ an automatic version of the two-stage procedure of Riani et al. (2009) to calibrate the FS envelopes used to confirm the clusters indicated by the random start FS. In the first stage we run a search on the data, monitoring the bounds for all n observations until we obtain a "signal" indicating that observation $m^\dagger$, and therefore succeeding observations, may be outliers, because the value of the statistic lies beyond our threshold. The hypothesis is that cluster j contains an unknown number n_j of observations. All that we know is that n_j is much less than n. We therefore need to judge the values of the statistics against envelopes from appropriately smaller population sizes. In the second part we accordingly superimpose envelopes for values of n from $m^\dagger - 1$ onwards, until the first time we introduce an observation we recognize as an outlier. The details of the rule are in Riani et al. (2009). Here we use an automatic version that, starting from a set of observations in each tentative cluster, finds $m_j^\dagger$, and then proceeds with the superimposition of envelopes until an outlier is identified and the cluster size n_j is established.

We start with the tentative Group 1. The upper left-hand panel of Fig. 4 shows the first-stage search with the sharp peak we have seen before around $m = 400$. The first outlier, the signal, is identified at $m^\dagger = 244$. The automatic procedure therefore starts with $n = 243$. These new envelopes, for a much smaller sample size, are broader than those for $n = 2{,}000$ and curve upwards at the end. There is no sign of any outlier for $n = 243$, so the sample size is augmented to 244 and the procedure repeated. Finally, as the bottom left-hand panel of Fig. 4 shows, there is no outlier when $n = 390$, although there is one above the 99.9 % envelope at $n = 397$. The first group therefore contains 396 observations. A similar procedure for Group 3 leads to a cluster of 768 observations.

The analysis for Group 2 needs more care. Because this group is relatively dispersed compared to Group 1, an FS starting from Group 2 will absorb many units from the compact group. We proceed by removing the observations confirmed as belonging to Groups 1 and 3, leaving 836 units. The left-hand panel of Fig. 5 shows the FS for these units. There is a signal at $m = 543$. The automatic procedure accordingly starts at $n = 542$ and proceeds to $n = 661$, as shown in the right-hand panel of Fig. 5. At this point, an outlier is indicated at $m = 597$, the very last part of the data trace lying below the lower threshold, indicating the presence of a few units from a different group. When these four units are removed we have 656 units in the group.

Figure 6 shows the scatterplot of the final clusters, together with the outliers. The histogram in the right-hand panel of the figure shows the outliers with the next-to-darkest colour. Importantly, there are very few outliers in the central part of the plot; virtually all have been classified as being in either Groups 1 or 3. This illustrates the general point that it is not possible to distinguish between clusters of observations and background contamination with the same values of y_1 and y_2.

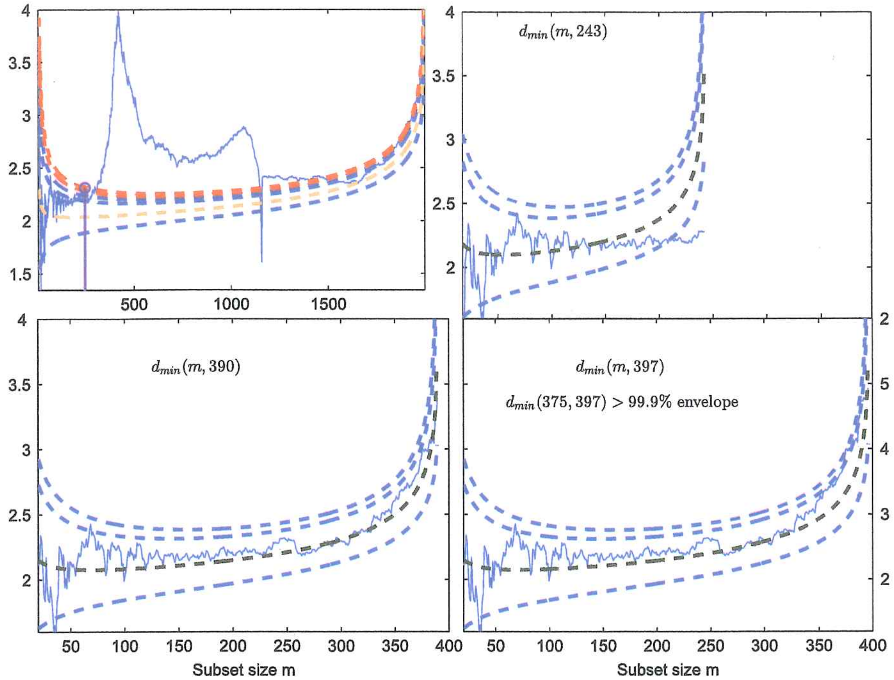

Fig. 4 Confirming Group 1. *Top left-hand panel*, forward plot of minimum Mahalanobis distances $d_{\min}(m)$ starting with units believed to be in Group 1; signal at $m^\dagger = 244$. Succeeding panels, distances for $n = 243$, 390 and 397. 396 units are assigned to the group

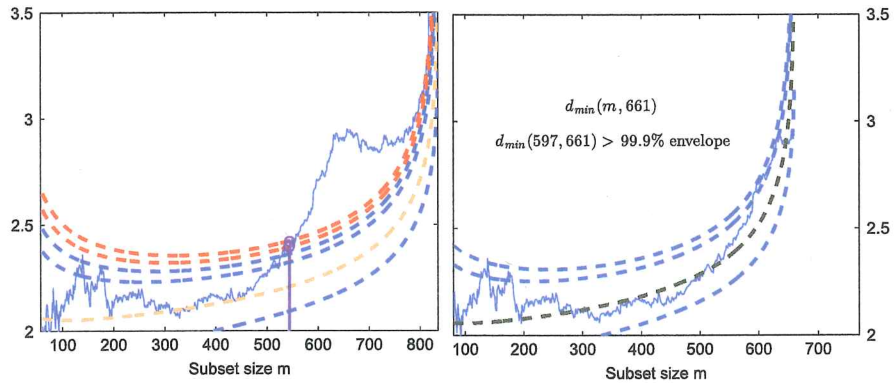

Fig. 5 Confirming Group 2 from 840 unassigned units. *Left-hand panel*, forward plot of minimum Mahalanobis distances $d_{\min}(m)$ starting with units believed to be in Group 2; signal at $m^\dagger = 543$. *Right-hand panel*, distances for $n = 661$. 656 units are ultimately assigned to the group

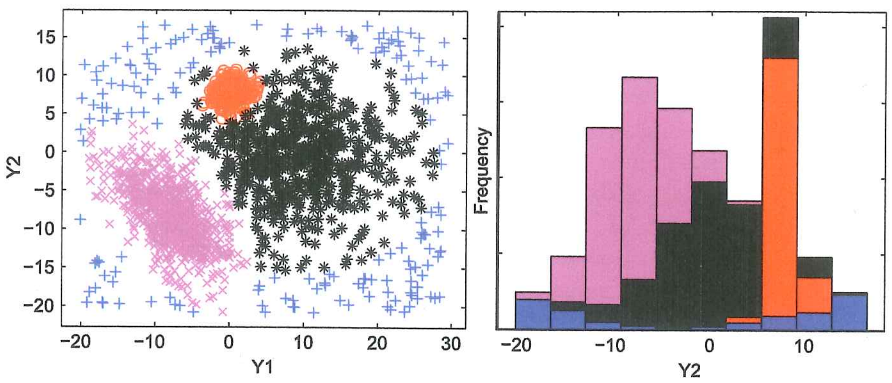

Fig. 6 Final FS clustering. *Left-hand panel*, scatterplot of the three groups and background contamination. *Right-hand panel*, histogram of classification of y_2 values: the contamination is shown in the next-to-darkest colour (*blue* in the pdf version). For central values of y_2, virtually all these observations have been included in one of the groups

4 Comparisons and Discussion

We now compare our clustering results with those of García-Escudero et al. (2011). Unlike our method, theirs requires some input parameters. We take these from the clustering we have found, namely: number of clusters $K = 3$, percentage of background contamination (outliers) 9 % and ratio of maximum to minimum eigenvalue over all clusters 100. With these parameters the agreement between TCLUST and FS clustering is very strong, the modified Rand index being 0.947. Table 1 gives a summary of the clustering performance of the two methods, together with the structure of the simulated data.

Given the choices for TCLUST in the previous paragraph which were derived from FS clustering, it is not perhaps surprising that both procedures produce clusters that differ from the data in a similar way. Clusters 1 and 3, which we established separately, are both larger than they should be, perhaps partly from the absorption of units from the background contamination; Cluster 0 has an identically low number of units for both methods. We were concerned that the relatively dispersed Cluster 2 would include some of the units of Cluster 1. In the event, it is slightly smaller than it should be.

Our results convincingly illustrate the use of the random start FS for discovering the number and identity of clusters. Refinement by the automatic version of the multivariate outlier detection method of Riani et al. (2009) leads to clustering in the presence of disparate clusters and background contamination. The output from the FS procedure can also be used to provide input to other robust clustering algorithms. A full two-step procedure, where confirmation follows exploration along the lines originally described in Atkinson et al. (2004, Sect. 7), could also help to handle overlapping clusters.

Table 1 Comparison of Forward Search (FS) and TCLUST (TCL)

	Numbers			Percentages		
Cluster	Data	FS	TCL	Data	FS	TCL
0	200	180	180	10	9.0	9.0
1	360	396	378	18	19.8	18.9
2	720	656	678	36	32.8	33.9
3	720	768	764	36	38.4	38.3

Numbers and percentages of observations in the groups in the original data and after clustering

Acknowledgements We are very grateful to Berthold Lausen and Matthias Bömher for their scientific and organizational support during the European Conference on Data Analysis 2013. We also thank an anonymous reviewer for careful reading of an earlier draft, and for pointing out the reference to Hennig and Christlieb (2002). Our work on this paper was partly supported by the project MIUR PRIN "*MISURA—Multivariate Models for Risk Assessment*".

References

Atkinson, A. C., & Riani, M. (2007). Exploratory tools for clustering multivariate data. *Computational Statistics and Data Analysis, 52*, 272–285.
Atkinson, A. C., Riani, M., & Cerioli, A. (2004). *Exploring multivariate data with the forward search*. New York: Springer.
Atkinson, A. C., Riani, M., & Cerioli, A. (2006). Random start forward searches with envelopes for detecting clusters in multivariate data. In S. Zani, A. Cerioli, M. Riani, & M. Vichi (Eds.), *Data Analysis, Classification and the Forward Search* (pp. 163–171). Berlin: Springer.
Cerioli, A., & Perrotta, D. (2014). Robust clustering around regression lines with high density regions. *Advances in Data Analysis and Classification, 8*, 5–26.
Coretto, P., & Hennig, C. (2010). A simulation study to compare robust clustering methods based on mixtures. *Advances in Data Analysis and Classification, 4*, 111–135.
Fowlkes, E. B., Gnanadesikan, R., & Kettenring, J. R. (1988). Variable selection in clustering. *Journal of Classification, 5*, 205–228.
Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association, 97*, 611–631.
Fritz, H., García-Escudero, L. A., & Mayo-Iscar, A. (2012). TCLUST: An R package for a trimming approach to cluster analysis. *Journal of Statistical Software, 47*, 1–26.
Gallegos, M. T., & Ritter, G. (2009). Trimming algorithms for clustering contaminated grouped data and their robustness. *Advances in Data Analysis and Classification, 3*, 135–167.
García-Escudero, L. A., Gordaliza, A., Matrán, C., & Mayo-Iscar, A. (2008). A general trimming approach to robust cluster analysis. *Annals of Statistics, 36*, 1324–1345.
García-Escudero, L. A., Gordaliza, A., Matrán, C., & Mayo-Iscar, A. (2010). A review of robust clustering methods. *Advances in Data Analysis and Classification, 4*, 89–109.
García-Escudero, L. A., Gordaliza, A., Matrán, C., & Mayo-Iscar, A. (2011). Exploring the number of groups in model-based clustering. *Statistics and Computing, 21*, 585–599.
Hennig, C., & Christlieb, N. (2002). Validating visual clusters in large datasets: Fixed point clusters of spectral features. *Computational Statistics and Data Analysis, 40*, 723–739.
Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data. An introduction to cluster analysis*. New York: Wiley.
Lee, S. X., & Mclachlan, G. J. (2013). Model-based clustering and classification with non-normal mixture distributions. *Statistical Methods and Applications, 22*, 427–454.

Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika, 45*, 325–342.
Morelli, G. (2013). *A comparison of different classification methods*. Ph.D. dissertation, Università di Parma.
Riani, M., Atkinson, A. C., & Cerioli, A. (2009). Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society, Series B, 71*, 447–466.
Riani, M., Perrotta, D., & Torti, F. (2012). FSDA: A MATLAB toolbox for robust analysis and interactive data exploration. *Chemometrics and Intelligent Laboratory Systems, 116*, 17–32.
Riani, M., Atkinson, A. C., & Perrotta, D. (2014). A parametric framework for the comparison of methods of very robust regression. *Statistical Science, 29*, 128–143.