Robust Clustering for Performance Evaluation

Anthony C. Atkinson, Marco Riani, and Andrea Cerioli

Abstract The evaluation of the effectiveness of organisations can be aided by the use of cluster analysis, suggesting and clarifying differences in structure between successful and failing organisations. Unfortunately, traditional methods of cluster analysis are highly sensitive to the presence of atypical observations and departures from normality. We describe a form of robust clustering using the forward search that allows the data to determine the number of clusters and so allows for outliers. An example is given of the successful clustering of customers of a bank into groups that are decidedly non-normal.

1 Introduction

The evaluation of the effectiveness of organisations has become an important strategic element in both the public and private sectors. Successful organisational structures need to be studied and emulated, whilst those that are failing need to be identified as early as possible so that preventive measures can be put in place and the waste of resources minimized. If organisations can be appropriately classified into homogeneous groups their differences in structure become more certainly identifiable and the number of special cases that has to be studied is dramatically reduced. The clustering of data is being increasingly used as a method of evaluation in public administration, see Peck (2005), and as a strategic element of political and administrative action, partly because it falls within the range of methods which has been deemed appropriate by the EU and the OECD (see for example the working papers contained in the web site http://www.oecd.org).

There are many statistical methods for the classification of multivariate observations such as those that describe the properties of an organisation. But, as is well known, at least to statisticians, the traditional methods of cluster analysis are highly sensitive to the presence of atypical observations and to incorrectly specified

A. Cerioli (✉)
Dipartimento di Economia, University of Parma, Via Kennedy 6, Italy
e-mail: andrea.cerioli@unipr.it

F. Palumbo et al. (eds.), *Data Analysis and Classification*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-03739-9_43,

structures. Despite this sensitivity, robust statistical methods that are unaffected by outliers and model-misspecification are little used. It is the purpose of the present paper to extend and apply robust cluster analysis using the forward search as introduced in Chapter. 7 of Atkinson et al. (2004). This graphics-rich robust approach to clustering uses the data to identify the number of clusters, to confirm cluster membership and to detect outlying observations that do not belong to any cluster. More specifically, our analyses rely on forward plots of robust Mahalanobis distances. In order to provide sensitive inferences about the existence of clusters it is necessary to augment such graphs with envelopes of the distributions of the statistics being plotted. Examples of such envelopes and their use in the forward search for clustering moderate sized data sets are presented by Atkinson et al. (2006) and Atkinson and Riani (2007), in which the largest example has 1,000 observations. The theoretical results of Riani et al. (2009) provide the tools for extending our methodology to larger data sets, where indeed inspection of the trajectory of a single minimum Mahalanobis distance, defined in (3), greatly simplifies the cluster identification process. In Bini et al. (2004) we applied earlier versions of these methods to the analysis of a complicated set of data on the performance of Italian universities. Here we exemplify our method with a simpler example from banking. Other successful applications of the forward search to classification problems with several clusters and outliers are described by Cerioli et al. (2006) and Riani et al. (2008).

2 Mahalanobis Distances and the Forward Search

The main tools that we use are plots of Mahalanobis distances. The squared distances for the sample of n v-dimensional observations are defined as

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

where $\hat{\mu}$ and $\hat{\Sigma}$ are 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

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

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

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

From this subset we obtain n squared Mahalanobis distances

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

To start the search for cl
$v + 1$ observations, the mi
this subset to be as small a
of $S(m_0)$ come from the
size during the search in s
with high probability. Whe
order the squared distance
smallest as the new subset
one observation, but someti
To detect outliers we
observations not in the sub

d_n

If this observation is an
will be "large" compared t
the subset. All other obser
greater than $d_{\min}(m)$ and
For small datasets we c
the threshold of our statist
we can instead use the pc
For larger samples, Atkir
by simulation to have the
(2009) use arguments fro
obtain envelopes without
For cluster definition,
needed, the most informa
tinue to add observations
been used in estimation.
as units from other cluste
ingly requires that we kn
searches. But we, as do A
random starting points to

3 Example

To illustrate our method
from a bank operating in

y_1: Direct debts to th
y_2: Assigned debts fr
y_3: Amount of funds
y_4: Total amount inv

ıods that are unaffected by
the purpose of the present
he forward search as intro-
ics-rich robust approach to
rs, to confirm cluster mem-
elong to any cluster. More
Mahalanobis distances. In
e of clusters it is necessary
ions of the statistics being
forward search for cluster-
et al. (2006) and Atkinson
observations. The theoret-
xtending our methodology
tory of a single minimum
s the cluster identification
ıs of these methods to the
of Italian universities. Here
banking. Other successful
lems with several clusters
ıni et al. (2008).

Search

s distances. The squared
are defined as

(1)

the mean and covariance

nated from a subset $S(m)$
parameter estimates are

$\ldots, v$

), $j, k = 1, \ldots, v.$

es

$i = 1, \ldots, n.$ (2)

To start the search for cluster identification we take a random sample of $m_0 = v + 1$ observations, the minimum size for which Σ can be estimated. We require this subset to be as small as possible to maximize the probability that all members of $S(m_0)$ come from the same cluster. This subset of m_0 observations grows in size during the search in such a way that non-cluster members will be excluded with high probability. When a subset $S(m)$ of m observations is used in fitting we order the squared distances and take the observations corresponding to the $m + 1$ smallest as the new subset $S(m + 1)$. Usually this process augments the subset by one observation, but sometimes two or more observations enter as one or more leave.

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

$$d_{\min}(m) = \min d_i(m) \quad i \notin S(m). \tag{3}$$

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.

For small datasets we can use envelopes from bootstrap simulations to determine the threshold of our statistic during the forward search. For moderate sized datasets we can instead use the polynomial approximations of Atkinson and Riani (2007). For larger samples, Atkinson et al. (2007) rescale a paradigmatic curve obtained by simulation to have the correct sample size and number of variables. Riani et al. (2009) use arguments from order statistics and estimation in truncated samples to obtain envelopes without requiring simulation.

For cluster definition, 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 all observations in that cluster have been used in estimation. There is then 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, as do Atkinson and Riani (2007), instead use many searches with random starting points to provide information on cluster existence and definition.

3 Example

To illustrate our methodology we look at an example with a dataset of customers from a bank operating in Italy. The variables that we consider are:

y_1: Direct debts to the bank;
y_2: Assigned debts from third parties;
y_3: Amount of funds deposited;
y_4: Total amount invested in government securities.

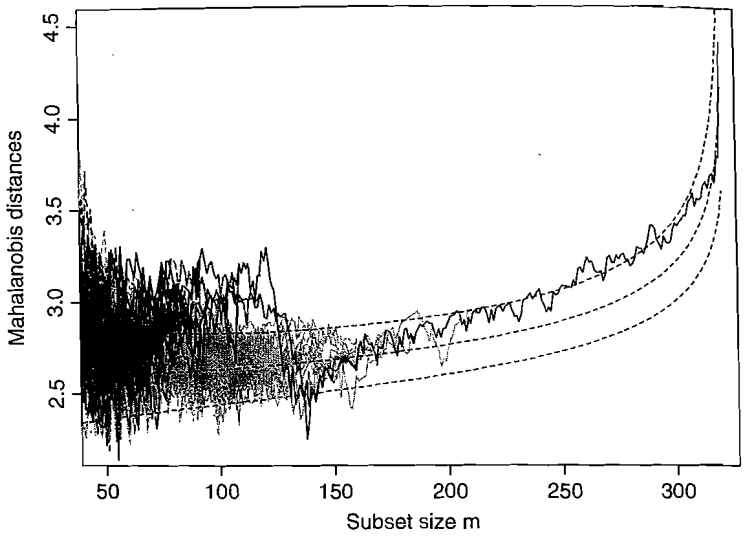


Fig. 1 Logged Banking data. Forward plot of minimum Mahalanobis distances, indicating two clusters; the trajectories in grey always include units from both of our final groups

The bank under study had just undertaken a thorough restructuring of all its activities. The purpose of the data analysis was to classify into homogeneous groups only those customers who had positive values for these four variables, of whom there were 322. Because the data were highly asymmetric, logs were taken to achieve approximate symmetry. In order to avoid singularity problems the logged data were also slightly jittered by adding a small normal noise.

Figure 1 shows a forward plot of minimum Mahalanobis distances from 200 random starts with 1 and 99% bounds. The structure of this plot is similar to that seen in Fig. 5 of Atkinson and Riani (2007), in which the simulated data consisted of two overlapping clusters.

As m increases the number of different subsets found by the forward search decreases, as is shown in the panels of Fig. 2. For m greater than 215 all searches follow the same trajectory. Earlier, around $m = 110$–130, there are two sets of trajectories lying clearly outside the envelopes (the black lines in the figure) and a large number of trajectories, represented in grey, within or close to the envelopes. The two sets of black trajectories in this range correspond to searches in which all the units in the subset are likely to come from a single cluster. If we identify the units in the subsets at $m = 118$ we obtain two initial clusters of observations. The largest value of $d_{\min}(m)$ gives a cluster with 118 observations and the second largest value a cluster of 115 observations, once three observations that might be in either cluster are removed. At this point we have preliminary clusters with a total of 233 observations and 89 observations to be clustered.

The scatterplot of the values of y_3 and y_4 for these two initial clusters are shown in the left-hand panel of Fig. 3. The two groups are clearly separated as they are

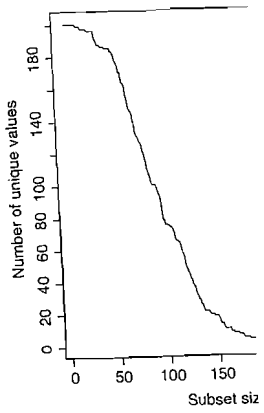


Fig. 2 Logged Banking data. 200 random starts. *Left-hand p* become apparent

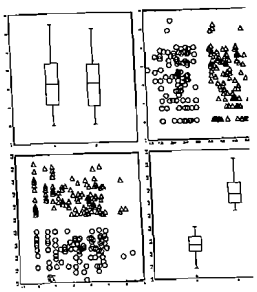


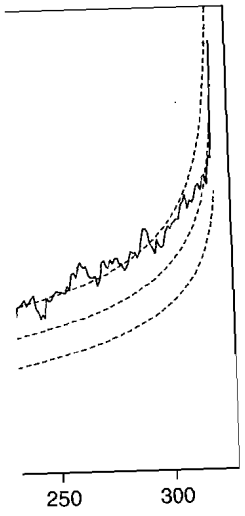
Fig. 3 Logged banking dat observations found from Fig. and y_2

in the centre panel, whi in the right-hand panel, clear clusters, which pla 89 observations which m or that may be unstructu

To explore these pos starting with the cluster now assess two Mahalan

$$d_i^2(l,m) = \{y_i$$

where $\hat{\mu}_l(m)$ and $\hat{\Sigma}_l(m$ on the observations in g of observations in the s



anobis distances, indicating two
our final groups

restructuring of all its activ-
o homogeneous groups only
ır variables, of whom there
logs were taken to achieve
)blems the logged data were

lanobis distances from 200
)f this plot is similar to that
he simulated data consisted

)und by the forward search
reater than 215 all searches
30, there are two sets of tra-
nes in the figure) and a large
se to the envelopes. The two
ırches in which all the units
we identify the units in the
)ervations. The largest value
second largest value a clus-
ıight be in either cluster are
h a total of 233 observations

wo initial clusters are shown
learly separated as they are

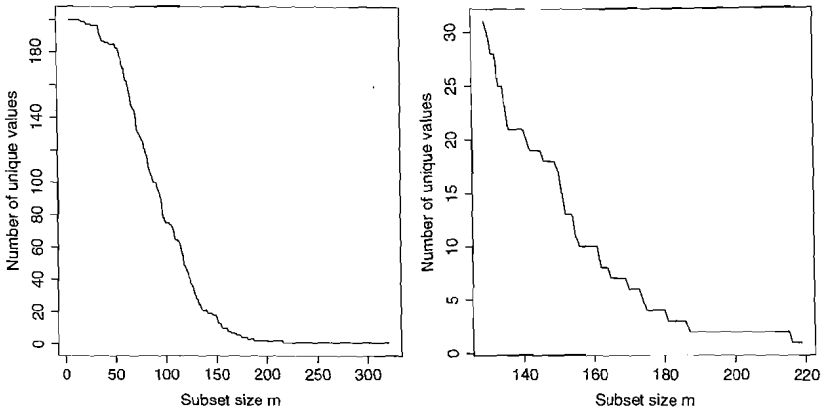


Fig. 2 Logged Banking data. Forward plots of number of unique Mahalanobis distances from 200 random starts. *Left-hand panel*, from 200 to 1; *right-hand panel*, zoom of plot where clusters become apparent

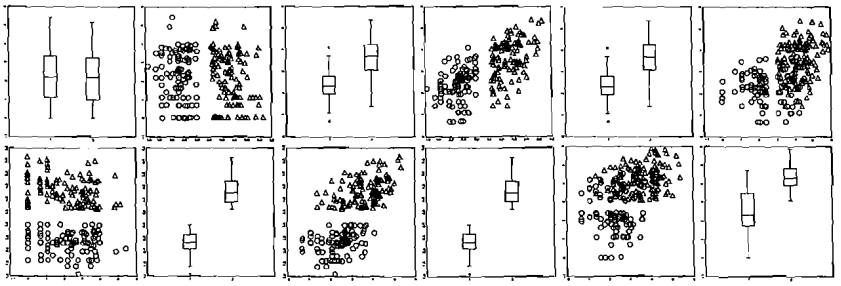


Fig. 3 Logged banking data: Scatterplot matrices of the two initial clusters of 118 and 115 observations found from Fig. 1. Reading across: y_3 and y_4, y_2 and y_4 and, *right-hand panel*, y_1 and y_2

in the centre panel, which is the scatterplot of y_2 and y_4. However they overlap in the right-hand panel, the scatter plot for y_1 and y_2. We have thus found two clear clusters, which plausibly have a multivariate normal structure, together with 89 observations which may perhaps belong to one of the groups, or to other groups, or that may be unstructured outliers.

To explore these possibilities we now run a forward search with two clusters starting with the cluster centres we have already found. In an extension of (2) we now assess two Mahalanobis distances for each unit

$$d_i^2(l,m) = \{y_i - \hat{\mu}_l(m)\}^T \hat{\Sigma}_l^{-1}(m)\{y_i - \hat{\mu}_l(m)\}, \qquad (l = 1, 2), \qquad (4)$$

where $\hat{\mu}_l(m)$ and $\hat{\Sigma}_l(m)$ are the estimates of the mean and covariance matrix based on the observations in group l, $l = 1$ or 2, and $m = m_1 + m_2$ is the total number of observations in the subsets for both groups. As before we start with a subset of

$m_0 = m_{01} + m_{02}$ observations. But now we want to preserve the cluster structure we have already established. So, for each m, we only consider the properties of the $2(n - m_0)$ squared Mahalanobis distances for the units that are not in the initial subset. We repeat the process several times for increasing values of m_0 that we take as 75% of the numbers of units which are indicated as correctly classified.

For each value of m we can use the values of $d_i^2(l, m)$ to allocate each unit not in m_0 to the cluster to which it is closest. We monitor how this allocation changes as the search proceeds. Those units that are firmly clustered stay in the same cluster throughout. Usually only those units about which there is some doubt have an allocation that changes as the search progresses. We ran one such search with the initial subset formed from the central 75% of units yielding our initial clusters of 118 and 115 units, that is the first 75% of this new set of units to enter these clusters in the individual searches shown in Fig. 1. We then obtained a set of units the allocations of which remained constant throughout the search. 75% of this new set of units resulted in an increased value of 204 for m_0. Figure 4 shows a forward plot of the allocation of the seven units that changed allocation during this two-cluster search. The bottom two lines serve as a key. The next band of two lines is for units 118 and 124. The classification of these units in the first cluster was not in doubt in our previous analyses, but they briefly become closer to the second group as the parameter estimates change with the inclusion of new units in the subsets used in fitting. The remaining seven lines, working upward, show the allocation, from $m = 240$, of units 110, 134, 135, 145, 178, 179 and 211. All other units, excluded from the plot, would have a single symbol throughout. As we shall see, these seven units lie between our two groups, so we refer to them as a "bridge". If we repeat the two-group search with the larger value of 268 for m_0 indicated by the results of Fig. 4 we find that the units in the bridge are, indeed, the only ones whose classification changes during the search.

The three panels of Fig. 5 show our proposed classification into two groups, of 145 and 177 units, with seven bridge units. The left-hand panel of the figure shows

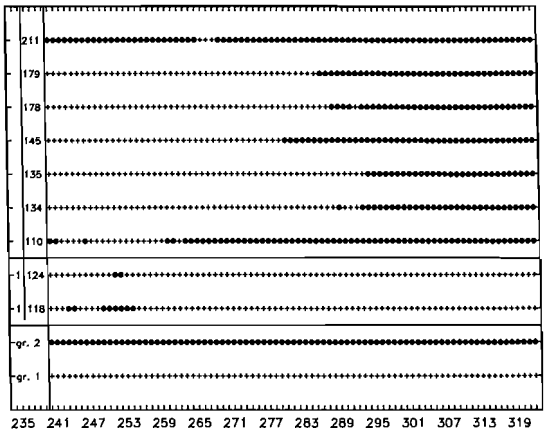


Fig. 4 Logged banking data. Cluster membership during a confirmatory search with two clusters starting with $m_0 = 204$. The bottom two lines serves as a key; the next two lines are units whose classification has never previously been in doubt, whereas the top seven lines give membership for the units that change classification during the search

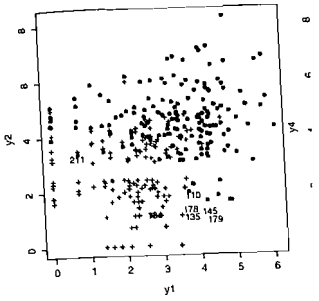


Fig. 5 Logged banking data: scatt
seven units from Fig. 4. Reading ac

the plot of y_2 against y_1, wit
the seven bridge units numbe
in these dimensions, with so
211, seem to lie in Group 1,
y_2. There is a clear division i
seem to cluster in Group 2, a
y_3 again shows the clear se
dispersed.

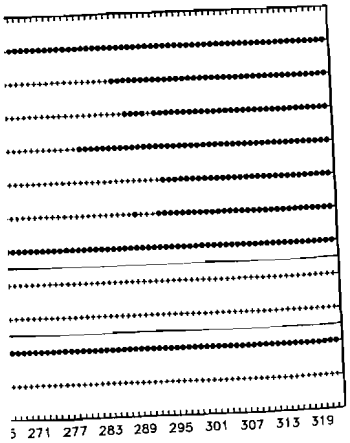
These plots seem to indic
data. But this has been achie
our procedure.

The classification of units
distances calculated using
difficulty, discussed by Atki
two clusters are very differe
are very different. As measu
of a tight cluster may have
from a cluster with a large
a large variance. Due to th
cluster with larger variance
variance become increasin
the search progresses. As a
from the tighter cluster.

A solution to this proble
distances standardised by t
distances behave more like
loose clusters absorbing ob
arise when the variances o
rithms of the data, we have
of our observations and, a

reserve the cluster structure
onsider the properties of the
ts that are not in the initial
ıg values of m_0 that we take
orrectly classified.
m) to allocate each unit not
how this allocation changes
ered stay in the same cluster
is some doubt have an allo-
e such search with the initial
ur initial clusters of 118 and
to enter these clusters in the
a set of units the allocations
5% of this new set of units
shows a forward plot of the
uring this two-cluster search.
of two lines is for units 118
luster was not in doubt in our
e second group as the param-
in the subsets used in fitting.
ıe allocation, from $m = 240$,
ther units, excluded from the
shall see, these seven units lie
ridge". If we repeat the two-
licated by the results of Fig. 4
nly ones whose classification

ssification into two groups, of
hand panel of the figure shows



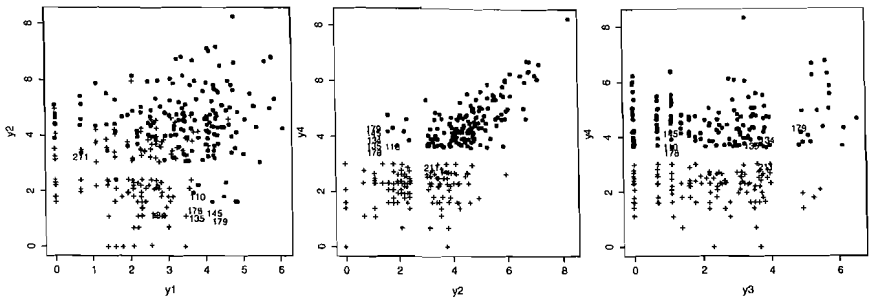


Fig. 5 Logged banking data: scatterplot matrices of the two final clusters with numbering for the seven units from Fig. 4. Reading across: y_2 and y_1, y_4 and y_2 and, *right-hand panel*, y_4 and y_3

the plot of y_2 against y_1, with the two clusters plotted with different symbols and the seven bridge units numbered. The separation of the two groups is not complete in these dimensions, with some interpenetration. Here the bridge units, apart from 211, seem to lie in Group 1, the crosses. The second panel is the plot of y_4 against y_2. There is a clear division into two groups on the values of y_4 and the bridge units seem to cluster in Group 2, again apart from unit 211. The final plot of y_4 against y_3 again shows the clear separation on values of y_4, but now the bridge units are dispersed.

These plots seem to indicate that we have satisfactorily clustered nearly all the data. But this has been achieved without any reference to the statistical properties of our procedure.

The classification of units shown in Fig. 5 is obtained by comparing Mahalanobis distances calculated using parameter estimates from the two groups. A potential difficulty, discussed by Atkinson et al. (2004), [p. 370], arises if the variances of the two clusters are very different. Then Euclidean distances and Mahalanobis distances are very different. As measured by Mahalanobis distance, an observation on the edge of a tight cluster may have a large distance for that cluster, but a smaller distance from a cluster with a larger variance. It will then be assigned to the cluster with a large variance. Due to the inclusion of this unit, the estimate of variance of the cluster with larger variance will increase and other units in the cluster with small variance become increasingly less remote from the cluster with larger variance as the search progresses. As a result the cluster with the looser structure absorbs units from the tighter cluster.

A solution to this problem, suggested by Atkinson et al. (2004), is to use instead distances standardised by the determinant of the estimated covariance matrix. These distances behave more like Euclidean distances and avoid the particular problem of loose clusters absorbing observations from tight clusters. However, these problems arise when the variances of the groups are very different. As a result of taking logarithms of the data, we have broken the relationship between the means and variances of our observations and, as Fig. 3 indicates, have obtained two groups with roughly

equal variances. In fact, here a search with standardised distances yields the same classification as that found using unstandardized distances.

In Fig. 1 we used envelopes derived from the multivariate normal distribution to establish preliminary clusters. We now repeat this procedure to confirm the two clusters that we have found. If we look at the scatterplots of the final clusters in Fig. 5 and compare them with the preliminary clusters in Fig. 3, we see that our final clusters have become appreciably less elliptical in outline and so can be expected to be relatively poorly described by a multivariate normal distribution. This feature is revealed in the confirmatory forward plots of minimum Mahalanobis distance for the two separate groups.

Figure 6 shows the forward plot from the 145 units we finally classified in Group 1, together with 0.1 and 99.9% envelopes. We have taken these broader envelopes as a way of allowing for the very approximate normality of our groups. As the figure shows, the 200 random searches settle down as the search progresses to give a trajectory that lies towards the upper part of the distribution but without any systematic peak and trough of the sort that indicated the presence of clusters in Fig. 1.

The similar Fig. 7 shows the plot for the 170 units of Group 2, together with the 7 units in the "bridge". Here again there is no clear indication of any presence of clusters. The general shape of this plot, lying rather high in the envelope and then gradually decreasing is an indication of slight non-normality; Fig. 11 of Riani and Atkinson (2007) shows a more dramatic example of a plot with a related structure for regression with beta distributed error. The jump in the plot around $m = 120$ corresponds, as we saw in Fig. 1, to the end of the normally distributed central part

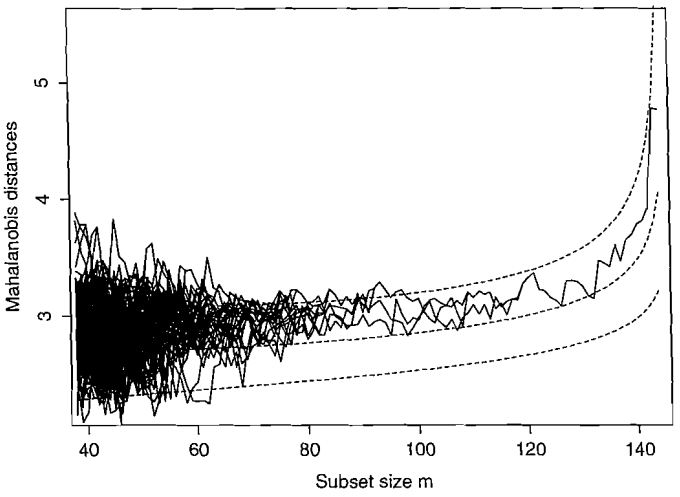


Fig. 6 Logged Banking data. Validation of Group 1. Forward plot of minimum Mahalanobis distances for the 145 units included in Group 1

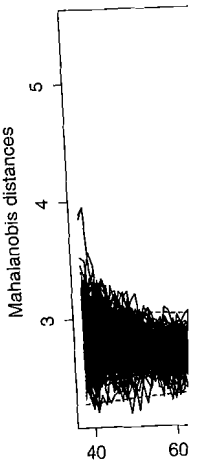


Fig. 7 Logged Banking data
distances for the 170 units incl

of the cluster in the scat
extreme observation, 211

An alternative metho
Raftery (2006) in which a
son and Riani (2007) pr
clusters than our robust
simulated data in which
example of the current
The forward plots of Fi
These forward plots can
do not at all lie within t
satisfactorily homogene
that the data consist mo
have found another ex
clusters.

Acknowledgements This
la valutazione integrata de
del fornitore, soddisfazion
la valutazione della regola

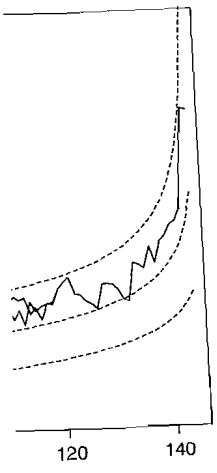
References

Atkinson, A. C., & Riani,
tional Statistics and D
Atkinson, A. C., Riani, M
search. New York: Sp

distances yields the same
s.
ariate normal distribution
cedure to confirm the two
ots of the final clusters in
Fig. 3, we see that our final
ie and so can be expected
il distribution. This feature
i Mahalanobis distance for

ts we finally classified in
have taken these broader
:e normality of our groups.
vn as the search progresses
he distribution but without
l the presence of clusters in

Group 2, together with the
lication of any presence of
gh in the envelope and then
nality; Fig. 11 of Riani and
plot with a related structure
the plot around $m = 120$
ially distributed central part



d plot of minimum Mahalanobis

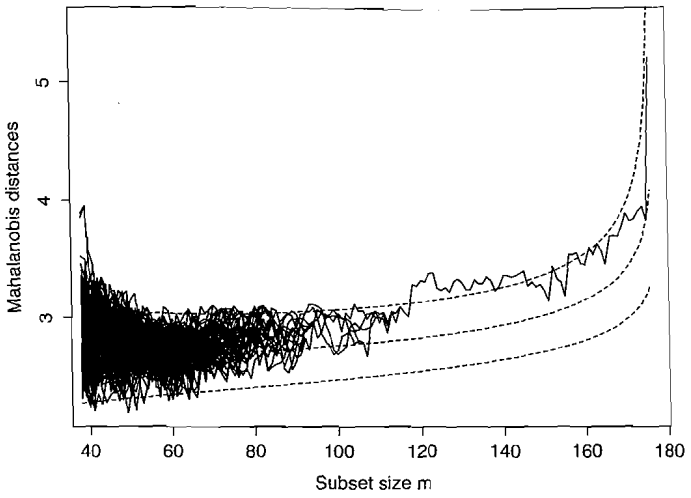


Fig. 7 Logged Banking data. Validation of Group 2. Forward plot of minimum Mahalanobis distances for the 170 units included in Group 2 and the seven"bridge" units

of the cluster in the scatterplots of Fig. 3. At the end of this search there is one extreme observation, 211, that has already been identified as the least well grouped.

An alternative method of clustering is the `mclust` procedure of Fraley and Raftery (2006) in which a mixture of normal distributions is fitted to the data. Atkinson and Riani (2007) provide examples in which `mclust` incorrectly finds more clusters than our robust method. The "incorrectness" is a feature of the analysis of simulated data in which, of course, we know the true number of clusters. In the example of the current paper, the BIC plot from `mclust` indicates five clusters. The forward plots of Figs. 6 and 7 however give no indication of such a structure. These forward plots can also be produced for the five tentative clusters. The searches do not at all lie within the envelopes, indicating that these five clusters are far from satisfactorily homogeneous. There are two conclusions from these analyses. One is that the data consist mostly of two rather non-normal clusters. The other is that we have found another example in which `mclust` indicates an excessive number of clusters.

Acknowledgements This work was supported by the grants "Metodi statistici multivariati per la valutazione integrata della qualità dei servizi di pubblica utilità: efficacia-efficienza, rischio del fornitore, soddisfazione degli utenti" and "Metodologie statistiche per lanalisi di impatto e la valutazione della regolamentazione" of Ministero dell'Università e della Ricerca PRIN 2006.

References

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

Atkinson, A. C., Riani, M., & 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

Atkinson, A. C., Riani, M., & Laurini, F. (2007). Approximate envelopes for finding an unknown number of multivariate outliers in large data sets. In S. Aivazian, P. Filzmoser, & Y. Kharin (eds.), *Proceedings of the Eighth International Conference on Computer Data Analysis and Modeling* (pp. 11–18). Russian Federation : Artia, Minsk

Bini, M., Riani, M., Atkinson, A., & Cerioli, A. (2004). Analisi di efficienza e di efficacia del sistema universitario italiano attraverso nuove metodologie statistiche multivariate robuste. Research report 03, Comitato Nazionale per la Valutazione del Sistema Universitario (CNVSU), MIUR, Ministero dell'Istruzione dell'Universit e della Ricerca. RDR document produced on behalf of CNVSU. http://www.cnvsu.it/_library/downloadfile.asp?id=11265

Cerioli, A., Riani, M., & Atkinson, A. C. (2006). Robust classification with categorical variables. In A. Rizzi & M. Vichi (eds.), *COMPSTAT 2006: Proceedings in Computational Statistics* (pp. 507–519). Heidelberg: Physica

Fraley, C., & Raftery, A. E. (2006). MCLUST version 3: an R package for normal mixture modeling and model-based clustering. Tech. Rep. 504, University of Washington, Department of Statistics, Seattle, WA

Peck, L. (2005). Using cluster analysis in program evaluation. *Evaluation Review*, *29*, 178–196

Riani, M., & Atkinson, A. C. (2007). Fast calibrations of the forward search for testing multiple outliers in regression. *Advances in data analysis and classification*, *1*, 123–141

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., Cerioli, A., Atkinson, A., Perrotta, D., & Torti, F. (2008). Fitting mixtures of regression lines with the forward search. In F. Fogelman-Soulié, D. Perrotta, J. Piskorski, & R. Steinberger (eds.), *Mining massive data sets for security* (pp. 271–286). Amsterdam: IOS

Outliers Detecti Clustering Algo

Balzanella Antonio, Elvir

Abstract In recent years applicative fields. Howeve liers. This paper aims to de by using a Dynamic Curve more robust. The approac rule on the distances dist identify outliers regions. T

1 Introduction

In many applicative fields $\Re$ rather than on a finite *data* Ramsay and Silverm where low fidelity and fre years, strategies to detec functional data, have rec on the concept of impart k-means algorithm Cues Albertos and Fraiman (2 from a random process the center of a cluster or authors change the func an $\alpha \in (0, 1)$ parameter trimmed. In the minimiz order to obtain the globa

A method, also based This clustering strategy

B. Antonio (✉)
Università degli Studi di Na
e-mail: balzanella2@alice.it

F. Palumbo et al. (eds.), *Da*
Studies in Classification, D
DOI 10.1007/978-3-642-03