Robust classification with categorical variables

Andrea Cerioli[1], Marco Riani[1], and Anthony C. Atkinson[2]

[1] Dipartimento di Economia, Sezione di Statistica e Informatica, Università di Parma, Italy andrea.cerioli@unipr.it mriani@unipr.it
[2] Department of Statistics, London School of Economics, London WC2A 2AE, UK A.C.Atkinson@lse.ac.uk

Summary. The forward search provides a powerful and computationally simple approach for the robust analysis of multivariate data. In this paper we suggest a new forward search algorithm for clustering multivariate categorical observations. Classification based on categorical information poses a number of challenging issues that are addressed by our algorithm. These include selection of the number of groups, identification of outliers and stability of the suggested solution. The performance of the algorithm is shown with both simulated and real examples.

Key words: Cluster analysis, forward search, dissimilarity, random starts.

1 Introduction

Clustering objects on the basis of information provided by categorical variables is an important goal in almost all application fields. For instance, in marketing research consumers typically need to be classified according to several categorical attributes describing their purchasing behaviour. This mass of categorical information is essential for uncovering market segments, that can help both to understand the differential consumption patterns across segments and to address them through specific advertising programmes. The evaluation of service quality data is another emerging application field, where it is important to have fast and reliable clustering algorithms suited to categorical variables. An example from this field will be seen in §4, where Italian municipalities are classified according to their degree of activity in e-government.

In spite of its practical relevance, clustering of discrete multivariate observations has received relatively little attention. A commonly used approach is to compute suitable measures of pairwise dissimilarity, such as the simple matching coefficient (e.g. [G99], §2.2), and then to use these measures as input for hierarchical clustering algorithms. Hierarchical agglomeration plays an important role also in the clustering algorithm of [FM04], which can be used with categorical information. The main problem with hierarchical algorithms is that they rapidly become computationally unacceptable and provide results

that are difficult to represent as the number of objects grows. The k-modes algorithm of [H98] and [CGC01] is a notable exception which tries to combine the efficiency of the standard k-means paradigm with the need to take categorical information into account. This is accomplished by running a k-means type algorithm with simple matching dissimilarities instead of Euclidean distances and cluster modes instead of means. However, as with k-means, the results from k-modes can be very sensitive to the choice of the starting solution and even to the order of the observations in the data set. An additional shortcoming is that cluster modes may not be uniquely defined at some steps of the iterative procedure, thus leading to indeterminacy in the clustering solution.

In this paper we take a different view and address the issue of clustering data sets with categorical information through the robust forward search approach. The forward search is a powerful general method for detecting unidentified subsets and multiple masked outliers and for determining their effect on models fitted to the data. The search for multivariate data, including cluster analysis with quantitative variables, is given book length treatment by Atkinson, Riani and Cerioli [ARC04]. It is our purpose to extend their robust clustering technique to cope with non-numeric attributes. This poses a number of novel problems, such as providing a suitable definition for the "centre" of a population along the search and for the "distance" of an individual from that centre. The suggested method is described in §2. It is computationally affordable and provides an assessment of the impact of each observation on the fitted clustering solution. It also helps to shed light on the actual number of clusters in the data, a critical issue with most, if not all, partitioning techniques. The performance of our technique is evaluated in §3 with several simulated datasets under known clustering conditions, including contamination by a small group of outliers. A real dataset is then analysed in §4.

2 Cluster detection through diagnostic monitoring

2.1 Distance from a discrete multivariate population

Let $S = \{u_1, u_2, \ldots, u_n\}$ be a set of n units for which we observe v nominal categorical variables $X_1, X_2, \ldots, X_v$. Unit u_i is represented as $[x_{i1}, x_{i2}, \ldots, x_{iv}]'$, where $x_{ij} \in \mathcal{C}^{(j)}$ is the observed class of variable X_j in unit u_i, and $\mathcal{C}^{(j)}$ is the set of possible classes for X_j. The number of such classes is c_j. For each variable the elements of $\mathcal{C}^{(j)}$ are unordered. We compute the dissimilarity between u_i and u_l through the simple matching coefficient

$$d(u_i, u_l) = \sum_{j=1}^{v} I(x_{ij} \neq x_{lj}), \qquad i, l = 1, \ldots, n, \tag{1}$$

where $I(\cdot)$ is the indicator function.

An alternative representation of simple matching is obtained through dummy coding of the categorical variables $X_1, X_2, \ldots, X_v$. Let $X_j^{(1)}, \ldots, X_j^{(c_j)}$

be dummy variables giving the observed class for X_j, i.e. $x_{ij}^{(c)} = 1$ if $x_{ij} = c$ and $x_{ij}^{(c)} = 0$ otherwise. The dissimilarity between u_i and u_l is measured as

$$d(u_i, u_l) = \sum_{j=1}^{v} \sum_{c \in \mathcal{C}^{(j)}} (x_{ij}^{(c)} - x_{lj}^{(c)})^2, \qquad i, l = 1, \ldots, n. \tag{2}$$

It is easy to see that (1) and (2) are equivalent, since they provide the same ordering of the dissimilarities among pairs of units. However, definition (2) has the advantage of being easily generalized to encompass differential weighting of the categories of each variable. The weighted measure is

$$d(u_i, u_l) = \sum_{j=1}^{v} \sum_{c \in \mathcal{C}^{(j)}} w_j^{(c)} (x_{ij}^{(c)} - x_{lj}^{(c)})^2, \qquad i, l = 1, \ldots, n, \tag{3}$$

where $w_j^{(c)} \geq 0$ is the weight given to category c of variable X_j in the classification process. Popular choices for $w_j^{(c)}$ include equal weighting, i.e. $w_j^{(c)} = 1$ for all $c \in \mathcal{C}^{(j)}$ and $j = 1, \ldots, v$, so that (3) reduces to (2), and

$$w_j^{(c)} = \{\hat{\pi}_j^{(c)}(1 - \hat{\pi}_j^{(c)})\}^{-1}, \qquad 0 < \hat{\pi}_j^{(c)} < 1, \tag{4}$$

where $\hat{\pi}_j^{(c)} = \sum_{i=1}^{n} x_{ij}^{(c)}/n$ is the proportion of units in S for which $X_j = c$. Equation (4) gives the inverse of the variance of $X_j^{(c)}$ in S, a scaling measure adopted for clustering purposes in [FM04] among others. Definition (3) can be further generalized to obtain a Mahalanobis-type simple matching dissimilarity measure, but this extension is not considered here.

The next step is to define a measure of closeness between a unit and a population. Recalling the definition of $x_{ij}^{(c)}$, unit u_i can be represented through $x_i = [x_{i1}^{(1)}, \ldots, x_{i1}^{(c_1)}, \ldots, x_{iv}^{(1)}, \ldots, x_{iv}^{(c_v)}]'$, a vector of dimension $C = \sum_{j=1}^{v} c_j$. We suppose that x_i is a random observation from a population with class probabilities $\pi = [\pi_1^{(1)}, \ldots, \pi_1^{(c_1)}, \ldots, \pi_v^{(1)}, \ldots, \pi_v^{(c_v)}]'$, so that $\mathrm{E}(x_{ij}^{(c)}) = \pi_j^{(c)}$, for $j = 1, \ldots, v$ and $c \in \mathcal{C}^{(j)}$. Following (3), we compute the dissimilarity between u_i and the mean vector π, or, when π is unknown, its sample estimate $\hat{\pi} = [\hat{\pi}_1^{(1)}, \ldots, \hat{\pi}_1^{(c_1)}, \ldots, \hat{\pi}_v^{(1)}, \ldots, \hat{\pi}_v^{(c_v)}]'$, as

$$d_i = d(u_i, \hat{\pi}) = \sum_{j=1}^{v} \sum_{c \in \mathcal{C}^{(j)}} w_j^{(c)} (x_{ij}^{(c)} - \hat{\pi}_j^{(c)})^2.$$

2.2 The forward search and the identification of clusters

The basic idea of the forward search is to start from a small, robustly chosen, subset of the data and to fit subsets of increasing size, in such a way that outliers and subsets of data not following the general structure are clearly revealed by diagnostic monitoring. With multiple groups, searches from more than one starting point are often needed to reveal the clustering structure. For continuous populations, [ARC06] demonstrate the usefulness of starting the search from randomly selected subsets that avoid any preliminary data

analysis. Here we provide evidence for multiple random starts from subsets of discrete multivariate observations.

In the forward search for clustering categorical data the mean estimate $\hat{\pi}$ is repeatedly computed on a subset of m observations, $S(m)$ say, yielding the C-dimensional vector $\hat{\pi}(m) = [\hat{\pi}_j^{(c)}(m)]'$, for $j = 1, \ldots, v$ and $c \in \mathcal{C}^{(j)}$. From this subset we obtain n dissimilarities

$$d_i(m) = d(u_i, \hat{\pi}(m)) = \sum_{j=1}^{v} \sum_{c \in \mathcal{C}^{(j)}} w_j^{(c)} (x_{ij}^{(c)} - \hat{\pi}_j^{(c)}(m))^2 \qquad i = 1, \ldots, n. \quad (5)$$

We start with a randomly selected subset of m_0 observations which grows in size during the search. When subset $S(m)$ is used in fitting, we order the dissimilarities (5) and take the units corresponding to the $m + 1$ smallest as the new subset $S(m+1)$. To detect potential clusters, we look at forward plots of quantities derived from the dissimilarities (5). One of the most useful plots is that of the minimum dissimilarity amongst units not in the subset

$$d_{\min}(m) = \min\, d_i(m) \qquad i \notin S(m). \quad (6)$$

Apart from some initial noise, the searches started in subsets of units with similar features will lead to the same forward plot of $d_{\min}(m)$. We look at bunches of similar trajectories to identify the number of clusters and how they originate. Furthermore, at the step where all the units of a homogeneous group have been included in $S(m)$, the dissimilarity (6) will be large compared to the maximum dissimilarity within $S(m)$. We look at peaks in the forward plots of $d_{\min}(m)$ for precisely identifying such clusters. Cluster membership is obtained by looking at the units in the subsets just before the peaks.

Other valuable tools for detecting important cluster features include the forward plot of individual dissimilarities $d_i(m)$, $i = 1, \ldots, n$, the entry plot showing the composition of $S(m)$ at each step of the search, and the forward plot of sample proportions $\hat{\pi}_j^{(c)}(m)$, $j = 1, \ldots, v$. The last plot can also be a useful aid to the interpretation of clusters along the search.

2.3 Outlier detection

The definition of an outlier as an extreme observation is not suitable for categorical data, since each X_j can take at most c_j distinct labels for which no ordering is available. More generally, we define an outlier as an observation not following the general structure of the data. In the present context, where detection of several groups is of concern, an outlier is then a unit not falling in any of the main groups forming the "clean" part of the data. This broader definition encompasses both "isolated" outliers, i.e. units which are far from all major groups, and "intermediate" outliers, i.e. units which fall within the boundaries of two clusters. Both types of outlier can have strong effects on the results of standard methods for cluster analysis. The identification of intermediate outliers is of particular concern because the absence of sharp cluster boundaries is a common occurrence in applications of clustering methods.

These units might also be difficult to detect using multivariate techniques with a high breakdown point, as is shown in [ARC04] in the case of bivariate continuous populations. Alternative clustering methods that try to reduce the effect of outliers are described by [CAGM97], [FR02] and [H03], but most of them are not easily extended to deal with nominal categorical variables.

2.4 Computational issues

The algorithm suggested in this paper shares the main properties of the forward search approach described in [ARC04], including computational simplicity and effectiveness of its graphical displays. It can be implemented by suitable modification of the software `Rfwdmv`, an R package specifically devised to perform the forward search for the analysis of continuous multivariate observations. The `Rfwdmv` package is available from CRAN, or can be downloaded from the web site `http://www.riani.it/arc/software.html`. It includes many functions and utilities for cluster analysis, such as the possibility of highlighting or removing, by a simple click, a set of trajectories in the forward plot of Mahalanobis distances. See [CK06] for further details.

3 Performance of the method

We evaluate the performance of our forward clustering technique through its ability to recover known clusters. For this purpose, we simulate several datasets under different scenarios and apply the algorithm proposed in §2 to each of them. The main findings are reported in §3.

3.1 Design of the simulation study

The setting of this study broadly mimics the structure of the application of §4. In each simulated dataset there are 240 "uncontaminated" units and 30 independent categorical variables. The datasets differ with respect to:

- the number of groups k (either $k = 3$ or $k = 6$);
- the number of classes c_j for each variable (either $c_j = 2$ or $c_j = 4$, $j = 1, \ldots, v$);
- the amount of noise in the simulation of $x_{ij}^{(c)}$ (either moderate, i.e. $\Pr(x_{ij}^{(c)} \neq \pi_j^{(c)}) = 0.10$, or high, i.e. $\Pr(x_{ij}^{(c)} \neq \pi_j^{(c)}) = 0.20$);
- the amount of contamination (either no contamination or contamination by a cluster of 4 intermediate outliers and 1 isolated outlier).

Group sizes are taken to be $n_1 = 100$, $n_2 = 80$ and $n_3 = 60$ if $k = 3$, and $n_1 = n_2 = 50$, $n_3 = n_4 = 40$ and $n_5 = n_6 = 30$ if $k = 6$. Each group is defined through a set of v/k variables, for which there is a class c such that

$$\pi_j^{(c)} \gg \pi_j^{(c')} \qquad c' \neq c \in \mathcal{C}^{(j)}. \tag{7}$$

Intermediate outliers lying between different groups are obtained by letting condition (7) hold for some variables in each group. For the isolated outlier, (7) holds for every j. In each dataset we start the forward search with the smallest possible subset size $m_0 = 2$ to enhance the possibility of detecting small clusters, although this may slightly increase noise in the initial steps of the algorithm.

3.2 Simulation results

We only describe the results for binary variables and the equal weight dissimilarity (2). Those for polytomous attributes are similar and are not detailed here. The simulated datasets analyzed in this paper are available at the web site `http://www.riani.it/c06`.

We start by considering the dataset with three clusters, moderate noise, four intermediate and one isolated outlier. The left-hand panel of Figure 1 contains the results of 500 forward searches from randomly selected starting subsets of size $m_0 = 2$. For each search we have plotted the dissimilarity $d_{\min}(m)$, defined in (6). The underlying three-group structure dominates the plot. The most striking feature is that, as early as $m = 20$, the searches follow only three different trajectories, regardless of starting point. These trajectories then merge towards the end of the algorithm, as units from different clusters enter into $S(m)$. The effect of contamination is visible both at the end, when the remote isolated outlier is included in $S(m)$, and at the very beginning of the plot, where random inclusion of outliers in the starting subset and their immediate removal by the search lead to an early peak in $d_{\min}(m)$. The peak at $m = 60$ is for searches containing units from the smaller group. At these values of m the observations from the other groups are all remote and have large dissimilarities from $S(m)$. The peaks just after $m = 80$ and $m = 100$ are for searches starting in the other two clusters. They are slightly delayed because the intermediate outliers lie on the border of these groups.

The plot shows the clear information that can be obtained by looking at the data from more than one viewpoint. It also shows how quickly the search settles down; because of the way in which units are included and excluded from the subset, the searches rapidly tend to produce subsets located in one specific cluster. The structure emerging from one of the possible viewpoints is depicted in the right-hand panel, the forward plot of individual dissimilarity measures $d_i(m)$ for searches starting in the largest cluster. The effect of adding units from a different group to a homogeneous subset is evident just after $m = 100$. The isolated outlier and the borderline units are clearly revealed (black solid lines), due to the peculiar shape of their trajectories.

A dissection of the units into three separate clusters can be obtained by looking at the composition of $S(m)$ just before the peaks in the forward plots of $d_{\min}(m)$. The resulting classification allocates 234 observations to their correct cluster, reveals the unique features of the isolated outlier and identifies 10 "borderline" units which could be equally assigned to different groups.

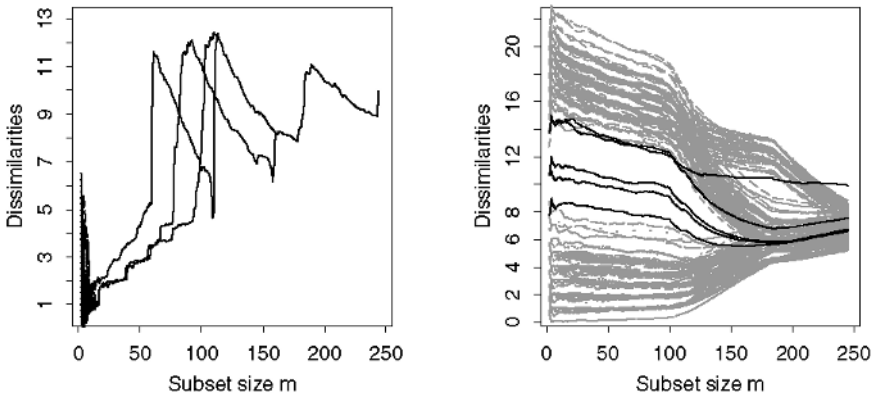

Fig. 1. Simulated categorical data with $c_j = 2$, three clusters, moderate noise, four intermediate and one isolated outlier. Left-hand panel: forward plots of $d_{\min}(m)$ for 500 searches with random starting points. Right-hand panel: forward plot of individual dissimilarities starting from the largest cluster; the trajectories of the five outliers are shown in black.

Among these borderline units, we find the four intermediate outliers, three of which can been seen to form a separate cluster in the right-hand panel of Figure 1. This clustering solution is more satisfactory than the one given by the `S-Plus` version of k-means (12 misclassified units) in the optimistic situation where the true data structure is known and we set $k = 5$. Another disappointing feature of standard partitioning methods is that slightly different algorithmic options can lead to widely different results. For instance, we found on these data that the clustering solution provided by the k-means implementation of `SPSS`, with $k = 5$, was remarkably different from that of `S-Plus`.

Both panels of Figure 1 can indeed be interpreted as revealing the clusters. But we also need to demonstrate that we are not finding structure where none exists. Figure 2 repeats Figure 1 for a sample from a homogeneous population with no outliers and moderate noise. These plots show none of the clustering structure that we have found in our previous example. The left-hand panel however does show again how quickly the search settles down, regardless of starting point.

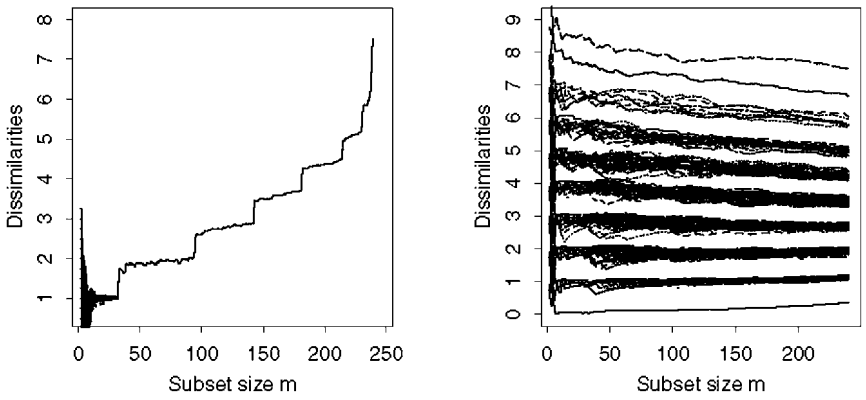

Fig. 2. Simulated categorical data from a homogeneous population with $c_j = 2$, moderate noise and no outliers. Left-hand panel: forward plots of $d_{\min}(m)$ for 500 searches with random starting points. Right-hand panel: forward plot of individual dissimilarities from a generic search. No clustering structure emerges

The simulated dataset with 6 clusters, high noise and 5 outliers provides a challenging task for classification methods. Given the unfavourable v/k ratio and the large amount of random error, we can expect the clusters to overlap considerably in the observed sample.

Both the k-means and the k-modes algorithms have a poor performance in this example even if we know, from external information, that $k = 8$. The left-hand panel of Figure 3 contains the forward plots of $d_{\min}(m)$ for 500 searches with random starting points. The right-hand panel is a zoom taken for $40 \leq m \leq 80$. The search settles down later than before, an evidence of higher noise. Although the structure of the plot is less striking than in Figure 1, there is still some definite evidence of clustering. The right-hand panel shows that, from around $m = 40$, all the searches converge to six stable trajectories.

We produce further forward plots to investigate the clustering structure implied by the six distinct trajectories. For instance, Figure 4 gives the viewpoint from one of these subsets. The trajectory shapes in the upper panel show a bunch of units forming a fairly distinct cluster from the rest of the data. The isolated outlier is also apparent at the top of the plot. The lower panel contains another useful forward plot, that of sample proportions $\hat{\pi}_j^{(1)}(m)$. There is a sudden change in the trajectories for variables $X_{11} - X_{15}$ at around $m = 40$, suggesting the existence of a cluster of that size defined by such variables.

The analysis can be repeated from different viewpoints, corresponding to the other stable trajectories in Figure 3.

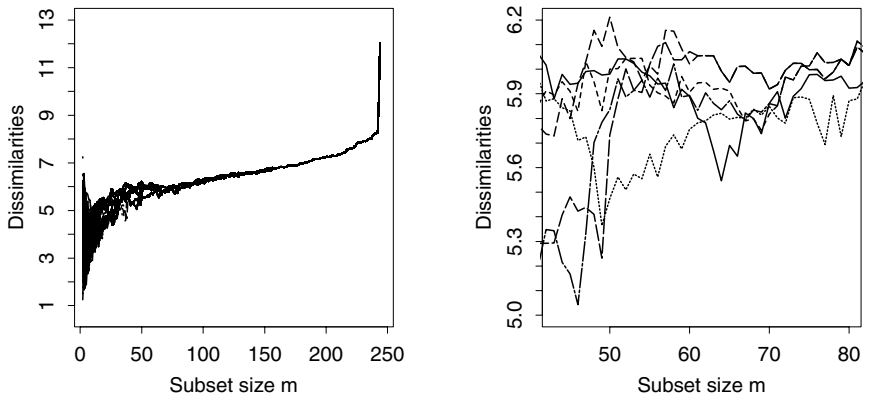

Fig. 3. Simulated categorical data with $c_j = 2$, six clusters, high noise, four intermediate and one isolated outlier: a difficult example for cluster analysis. Left-hand panel: forward plots of $d_{\min}(m)$ for 500 searches with random starting points. Right-hand panel: a zoom taken for $40 \leq m \leq 80$.

Again, looking at the data from different perspectives and inspecting the composition of $S(m)$ just before the major peaks in the forward plots of $d_{\min}(m)$ lead in the end to the identification of the whole group structure. Procedures similar to those described in [ARC04] and plots like those in Figure 4 could be used to confirm this structure and to explore borderline units.

4 E-government data

The data we analyze come from a questionnaire sent to all the municipalities of Emilia-Romagna, a region of northern Italy, in October 2004. The questionnaire aim was to measure the implementation of e-government technologies in a wide range of activities, from labour organization to on-line services for citizens. The data set consists of 229 municipalities, i.e. those filling out the questionnaire and having an active web site. Almost all the information was collected on a categorical scale. Here we focus on 34 binary indicators describing implementation of advanced e-government services in citizen-oriented activities, such as availability of interactive on-line services and e-democracy facilities. The number of implemented services ranges from 22 (in the case of Bologna, the regional capital) to only 1. A more detailed description of the data, together with further references and some preliminary analyses, can be found in [CMMZ05].

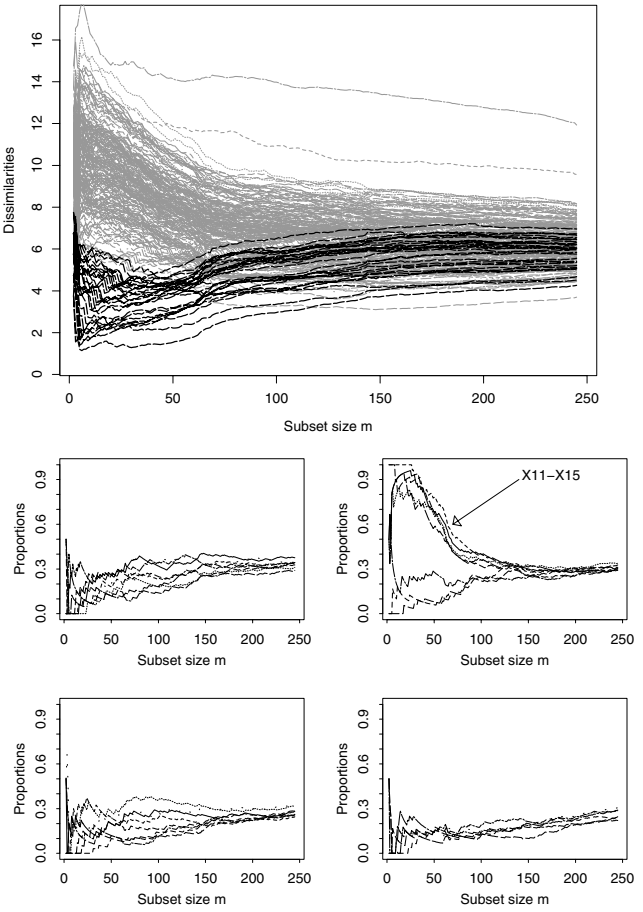

Fig. 4. Simulated categorical data as in Figure 3. Further forward plots starting from one cluster. Upper panel: individual dissimilarities $d_i(m)$; a cluster of units with increasing trajectories is shown in black. Lower panel: sample proportions $\hat{\pi}_j^{(1)}(m)$ (the first plot corresponds to variables X_1–X_8, the second plot to X_9–X_{16}, the third plot to X_{17}–X_{24} and the fourth plot to X_{25}–X_{30}).

The upper panel of Figure 5 shows the results of 1000 forward searches from randomly selected starting subsets. Here we start with $m_0 = 5$, the results being similar but slightly less noisy than with $m_0 = 2$. For each search we provide the forward plot of $d_{\min}(m)$ for the equal weight dissimilarity measure (2). From m around 40 all searches follow the same trajectory, regardless of the starting point. The end of the search shows a few possible outliers, without evidence of masking.

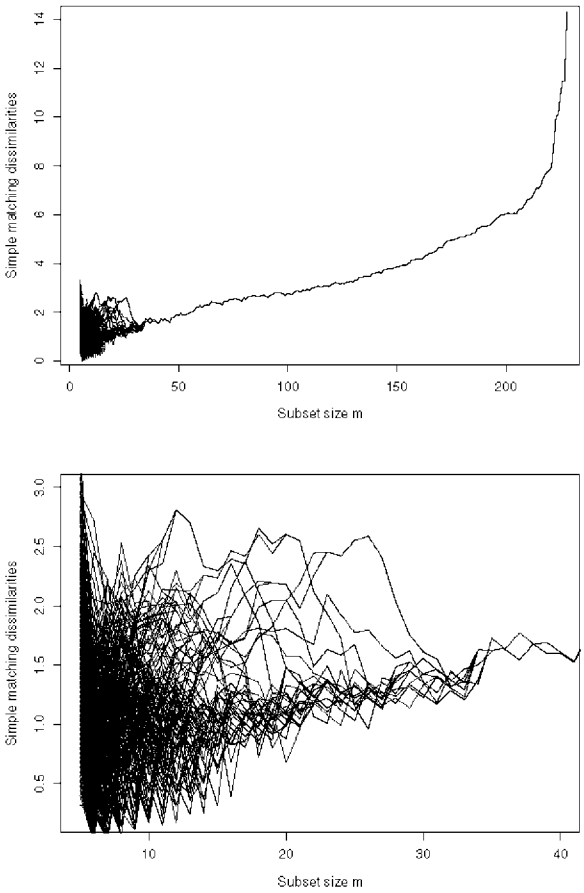

Fig. 5. E-government data. Upper panel: forward plots of $d_{\min}(m)$ for 1000 searches with random starting points. Lower panel: a zoom taken for $m \leq 40$.

Bologna and Modena, the two most remote units, are those with the highest number of e-government services. Other municipalities with good e-government facilities enter towards the end of the search, but without evidence of clustering. This means that these units are fairly well separated in the space spanned by the 34 indicators, although they are far from the majority of the municipalities of Emilia-Romagna.

The plot also leads to identification of a few clusters giving rise to the peaks visible in some trajectories for $m < 40$. The peaks are perhaps best seen in the lower panel of Figure 5. As they contain only few searches, we may anticipate a structure with some small and not very well separated clusters. For instance, consider the peak at $m = 12$. It comes from a set of 22 searches

containing municipalities with a relatively large, but not extreme, number of e-government services. The slow decay of the trajectory originating in this subset confirms a smooth transition from the cluster to the bulk of the data. The searches that do not give rise to such peaks either contain observations from more than one group, or come from an unstructured population. Therefore, they do not provide evidence of clustering.

For a better understanding of the cluster structure, further forward plots could be displayed. As in § 3, we monitor how individual distances (5) and proportions $\hat{\pi}_j^{(1)}(m)$ evolve for selected searches, such as those producing the peak at $m = 12$. These pictures (not shown) reinforce the idea of a large unstructured population to which most municipalities of Emilia-Romagna belong, of a few relatively small and not well separated clusters added to this "noisy" background, and of some "outstanding" but isolated municipalities.

Acknowledgement

We are grateful to the staff at Direzione Generale Organizzazione, Sistemi Informativi e Telematica of Emilia-Romagna for making the e-government data available to us.

References

[ARC04] Atkinson A.C., Riani M., Cerioli A.: Exploring Multivariate Data with the Forward Search. Springer, New York (2004)

[ARC06] Atkinson A.C., Riani M., Cerioli A.: Random start forward searches with envelopes for detecting clusters in multivariate data. In: Zani, S., Cerioli, A., Riani, M., Vichi, M. (eds.) Data Analysis, Classification and the Forward Search. Springer, Berlin (in press)

[CMMZ05] Cerioli A., Milioli M.A., Morlini I., Zani S.: L'ICT nella pubblica amministrazione: un'applicazione ai comuni dell'Emilia-Romagna (in Italian). In: Atti della Riunione Scientifica su Valutazione e Customer Satisfaction per la Qualità dei Servizi. Facoltà di Scienze Statistiche, Università di Roma, 65–68 (2005)

[CGC01] Chaturvedi A., Green P.E., Carroll J.D.: K-modes clustering. Journal of Classification, **18**, 35–55 (2001)

[CK06] Corbellini A., Konis K.: An R package for the robust analysis of multivariate data. In: Zani, S., Cerioli, A., Riani, M., Vichi, M. (eds.) Data Analysis, Classification and the Forward Search. Springer, Berlin (in press)

[CAGM97] Cuesta-Albertos, J.A., Gordaliza A., Matran C.: Trimmed k-means: an attempt to robustify quantizers. Annals of Statistics, **25**, 553–576 (1997)

[FR02] Fraley C., Raftery A.E.: Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association, **97**, 611–631 (2002)

[FM04] Friedman J.H., Meulman J.J.: Clustering objects on subsets of attributes. Journal of the Royal Statistical Society B, **66**, 815–849 (2004)

[G99] Gordon A.D.: Classification, 2nd Ed. Chapman & Hall/CRC, Boca Raton (1999)

[H03] Hennig C.: Clusters, outliers, and regression: fixed point clusters. Journal of Multivariate Analysis, **86**, 183–212 (2003)

[H98] Huang Z.: Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery, **2**, 283–304 (1998)