Model-based clustering using S-PLUS

MORVEN LEESE,^{1,2} SABINE LANDAU²

1 Department of Health Services Research, Institute of Psychiatry, King's College, London, UK

2 Department of Biostatistics and Computing, Institute of Psychiatry, King's College, London, UK

Abstract

Cluster analysis can be used to identify homogenous subgroups in many fields, including psychology and psychiatry. However, most clustering methods implemented in general-purpose statistical packages are heuristic and can be criticized in principle for their lack of an underlying statistical model. Furthermore correlations between variables are generally ignored by standard methods. The question addressed here is whether currently available commercial software (S-PLUS), which provides model-based methods for clustering correlated continuous data, should be used for clustering data derived from questionnaires. Such data may be either continuous or ordinal in nature and typically exhibit correlations. Performance is assessed in this study on simulated data sets containing distinct multivariate normal subpopulations, both before and after mapping the simulated data onto an ordinal scale. A practical example showing how correlated data can be cluster-analysed using these methods is given. The conclusion is that model-based methods are certainly worthwhile for continuous data. However, their benefit, in particular their ability to deal with correlated data, is not marked for ordinal data. Simpler methods such as Ward's method may be almost as effective in this situation. Copyright © 2006 John Wiley & Sons, Ltd.

Key words: cluster analysis, model-based methods, Ward's method, ordinal data, classification likelihood, S-PLUS

Introduction

Model-based clustering of ordinal data

The term *cluster analysis* encompasses a wide-ranging group of techniques for classifying 'objects' (which may be people, products, hospitals or any other units of interest including variables) into subgroups that are internally homogeneous (cohesion) but which differ from each other (isolation). In cluster analysis the group structure is unknown and the aim is to discover it. Many of the widely used techniques are heuristic in that the rules for allocating objects to clusters do not have an underlying theoretical model (probabilistic or deterministic). Rather they optimize an objective function based on cohesion and isolation defined in some way. This may be performed in a stepwise manner, either by successively agglomerating objects on the basis of their similarity (defined in some way) or by starting with the whole group of objects and successively dividing it. In such methods a hierarchical

nested sequence of clusters is formed. Examples are the widely used *single*, *average* or *complete linkage* methods and Ward's method (Ward, 1963). Alternatively partitioning methods may be used, in which a single partition into clusters is formed, again using a heuristic algorithm to optimize the objective function for a given number of clusters, a well-known example being *k-means* (MacQueen, 1967). Reviews of cluster analysis, both hierarchical and partitioning, are given by McLachlan (1992), Arabie et al. (1996), Gordon (1999) and Everitt et al. (2001).

Heuristic methods of clustering can be criticized for their lack of an underlying statistical model. Furthermore, different similarity measures and clustering algorithms often give widely different results and there is little basis on which to choose one method rather than another. In that respect *model-based methods* may be preferable. These usually assume that the target population divides into distinct subpopulations with

differing multivariate distributions (typically multivariate normal distributions for continuous data or multinomial for nominal categorical data, the latter leading to latent class analysis). If the model is realistic, inferences about the data-generating process may be possible. Furthermore, the assumptions on which the clusters have been derived are explicit. Of particular interest is the ability of methods based on the multivariate normal model to detect clusters in highly correlated data, because most heuristic methods do not cater for this. Typically traditional methods, such as the linkage methods, Ward's method and k-means, treat variables as independent within clusters.

The underlying concept in many model-based methods is that a number of overlapping subpopulations have been mixed together. Such an assumption is often plausible in psychiatry but whether the model is reasonable in any particular circumstance needs to be judged on the basis of subject matter. The current debate about taxometric methods that distinguish discrete classes versus methods that assume dimensional data is relevant here (see Meehl, 1992, for example). There are many examples of the mixtures approach but, to give just one example, Bellivier et al. (2003) show evidence for heterogeneity in the age of onset of bipolar affective disorder using the 'admixture' method. This analysis is based on univariate data but extension to multivariate data poses no additional conceptual problems, although it is computationally more difficult because of the multiplicity of parameters that need to be fitted. Banfield and Raftery (1993) provided a solution to these computational problems, using 'classification likelihood' to fit mixtures of multivariate normal distributions as a hierarchical sequence of clusters. The method has been incorporated in the package S-PLUS and so is now accessible for researchers in psychiatry or psychology.

One potential limitation to the method is that it assumes multivariate normality within clusters. Psychiatric data are often skewed so this may need to be addressed, for example by transformation (bearing in mind that it is multivariate normality within clusters that is required and that cluster membership is unknown at the start of the analysis). However, another important limitation is that psychiatric data are often in the form of ordered categorical scales for which the method was not designed. For example, questionnaires often include adjectival or Likert scales (Likert, 1932), the latter typically having end points representing 'agree

strongly' and 'disagree strongly', with a few fixed intermediate points. When the number of points becomes 'large' the data are often analysed as if they were continuous (Streiner and Norman, 2002). It is difficult to give guidelines about how many points are necessary for this to be acceptable because it depends on the context and the degree of approximation that can be tolerated. Bollen and Barb (1981), for example, showed that correlations based on scales of about five to seven points were often acceptable for many practical purposes. In a more complex application, Jorgensen and Hunt (1999) discuss circumstances in which ordinal data might be clustered as if they were continuous. The question addressed in this paper is whether the classification likelihood approach to clustering based on multivariate normal mixtures can be used, at least as an approximation, for ordinal data. We do this by simulating highly correlated, clustered data both before and after mapping onto an ordered categorical scale. First we provide some background on the methods as they have been incorporated in S-PLUS 6.1 and S-PLUS 2000 (Krause and Olsen, 2000).

The classification likelihood approach to clustering implemented in S-PLUS

The model underlying the (model-based) clustering method for continuous data in S-PLUS assumes that the population consists of G subpopulations (clusters) each of which is multivariate normal with respect to an m -dimensional outcome X , but with different means μ_k , $k \in \{1, \dots, G\}$ and possibly different variance-covariance matrices Σ_k . Let $f_k(\mathbf{x}; \theta_k)$ with $\theta_k = \{\mu_k, \Sigma_k\}$ denote the multivariate normal density for the k th subpopulation. Further let $\mathbf{x}_1, \dots, \mathbf{x}_n$ denote a random sample of m -dimensional observations from the population. Then in order to estimate the group membership of the observations Banfield and Raftery (1993) suggest maximizing the *classification likelihood* defined by

$$L_{\text{classification}}(\gamma, \theta; \mathbf{x}_1, \dots, \mathbf{x}_n) = \prod_{i=1}^n f_{\gamma_i}(\mathbf{x}_i; \theta_{\gamma_i}) \quad (1)$$

where $\gamma = (\gamma_1, \dots, \gamma_n)^T$ denote identifying labels with $\gamma_i = k$ if $\mathbf{x}_i$ comes from the k th subpopulation. That is, group membership itself is modelled as a parameter for each observation and group memberships and other model parameters are estimated, simultaneously providing a cluster solution for G subpopulations.

The classification likelihood approach as implemented in S-PLUS is very flexible in that a number of options are provided as to how the variance-covariance matrices Σ_k can vary between the subpopulations. Key to this is the eigenvalue decomposition of the variance-covariance matrix of the multivariate normal distribution for the k th cluster

$$\Sigma_k = \lambda_{1,k} \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T \quad (2)$$

where $\mathbf{D}_k$ is the $m \times m$ -matrix of eigenvectors of Σ_k , $\mathbf{A}_k$ is an $m \times m$ -matrix with (diagonal) elements proportional to the eigenvalues of Σ_k and $\lambda_{1,k}$ is the largest eigenvalue of Σ_k . Then $\mathbf{D}_k$ reflects the orientation of the cluster distribution, $\lambda_{1,k}$ its size and $\mathbf{A}_k$ its shape. (Size here refers to volume in multidimensional space, rather than sample size.) For example, for a bivariate normal distribution with variances $\sigma_{x,k}^2$ and $\sigma_{y,k}^2$ and correlation ρ_k , the two eigenvalues are given by

$$\lambda_{1,k}, \lambda_{2,k} = \frac{1}{2} \left[(\sigma_{x,k}^2 + \sigma_{y,k}^2) \pm \sqrt{4\rho_k^2 \sigma_{x,k}^2 \sigma_{y,k}^2 + (\sigma_{x,k}^2 - \sigma_{y,k}^2)^2} \right]$$

and matrix $\mathbf{A}_k$ becomes $\begin{bmatrix} 1 & 0 \\ 0 & \alpha_k \end{bmatrix}$ with $\alpha_k = \lambda_{2,k}/\lambda_{1,k}$. (If

the variances in the two dimensions are equal the formula reduces to $\alpha_k = (1 - |\rho_k|)/(1 + |\rho_k|)$, showing that the degree of correlation informs the shape parameter.)

This decomposition can be employed to reduce the number of variance-covariance matrix parameters by allowing some features of the distributions (orientation, size or shape) to vary between clusters while constraining others to be the same. The two extreme choices of variance-covariance parameterisations are the fully constrained spherical model – cluster distributions are spherical and of the same size in all clusters – and the completely unconstrained model – all three distributional features (orientation, size and shape) are allowed to vary with the clusters. In the former $\Sigma_k = \sigma^2 \mathbf{I}_m$ and only a single parameter is used. It can be shown that the criterion resulting from maximizing the respective classification likelihood is equivalent to the optimisation criterion employed by the standard Ward's method (also called the 'trace' method, because it minimizes the trace of the within-cluster covariance matrix). Similarly it can be shown that the unconstrained model leads to a cluster criterion previously proposed by Scott and Symons (1971).

Banfield and Raftery (1993) suggest that intermediate models might be useful and derive two new cluster criteria called 'S' and 'S*'. Cluster method S maximizes the classification likelihood of a multivariate normal model where the cluster distributions have the same size and (known) shape but different orientations. Cluster method S* is derived from a model that also allows the sizes to vary and is therefore slightly less restrictive. The classification likelihood approach for different choices of cluster orientation, size and shape (including S, S* and Ward's method) has been programmed in S-PLUS as a form of agglomerative hierarchical clustering, in which clusters are merged according to the increase in the classification likelihood at each stage. As with the more standard hierarchical methods, a particular partition corresponding to a specific number of clusters must be chosen. Banfield and Raftery (1993) suggest an index known as the *approximate weight of evidence* (AWE) for this. Values of AWE that are greater than zero indicate evidence for clustering and the maximum indicates the value of k for which the evidence is greatest, although a range of plausible values should usually be considered.

Objective of this study

Through the S-PLUS software, model-based methods for clustering continuous correlated data have become easily usable by researchers in psychiatry and psychology. The objective of this paper is therefore to examine the performance of the routines mentioned above when applied to data of the type often found in psychiatry and psychology. Some preliminary investigations showed little difference between the performances of S* and S, and this conclusion was also reported by Papageorgiou (2001) in an empirical study. We therefore focus on the performances of the S method and Ward's method (the latter being already well-known and widely used). We assess performance empirically on simulated bivariate data sets. The first consists of data generated from multivariate normal populations with a proportion of the populations with a high degree of correlation (the type of data for which the S method was designed) and a proportion having zero correlation (the type of data for which Ward's method is appropriate). All models assumed equal-sized distributions within population subgroups. The second data set consists of a gridded version of the same data – mapped onto an ordered categorical scale.

Simulation study

Simulation strategy

The overall strategy was to apply each of two cluster methods in three stages. The methods were firstly the commonly used Ward's method (which can be derived from a multivariate normal model with constant size spherical distributions within clusters) and, secondly, the S method (derived from a model with cluster distributions of common size and shape but without assuming a spherical shape). The three stages were as follows:

1. Apply the methods to simulated data of the type used by Banfield and Raftery, with the shape setting of the S method given the value $\alpha^{(s)}$ to 0.2, the value these authors used (and as recommended in the software manual). This was aimed at demonstrating the baseline performance of the classification likelihood method on bivariate normal data with a high degree of correlation, and for which the number of clusters is assumed known (the ideal situation for the S method).
2. Remove the unrealistic assumption that the number of clusters is known, because this is likely to impact on performance, and apply the method to the same simulated continuous data as in stage 1. The number of clusters in each of the 'realistic' runs was estimated on the basis of the maximum AWE, as recommended in the software manual.

3. Apply the methods as in stage 2, but to ordinal data derived from the simulated continuous data, because this is of interest in the fields of psychiatry and psychology.

Simulation parameters used to generate data to be clustered

The parameters chosen for generating the data were those reflecting high correlations between variables – low α_k , for which the method S ought to perform well. The parameter choices also included, as a special case, completely spherical cluster distributions (zero correlations between variables) for which Ward's method ought to perform well.

Two series of 100 samples from each of 25 overall conditions (one for continuous data and one for the corresponding ordinal data) were chosen to produce samples from distributions where the shapes and sizes did not vary between the clusters in a given run. When the population clusters had parameter $\alpha < 1$ this gave data that fulfilled the assumptions of method S. The subset of the simulated data generated with $\alpha = 1$ represented samples from spherical and equal sized cluster distributions and so fulfilled the ideal conditions for Ward's method. The size parameter λ_1 and shape parameter α were held constant for the clusters within a run whereas number of clusters per run, number of points per cluster, bivariate cluster means and cluster orientations were chosen at random within a run.

Table 1. Parameters used in simulations (100×25 runs in each of two series, continuous and ordinal data)

Number of clusters per run	Discrete uniform distribution on {2, 3, 4, 5, 6}				
Number of points per cluster	Discrete uniform distribution on {15, 16, . . . , 25, 26}				
Means per cluster	Uniform distribution on unit square				
Orientations per cluster (radians)	Uniform distribution on [0, π]				
α and λ_1 were held constant across clusters per run;					
PNND (%), a measure of clustering difficulty, is shown in the body of the table (median over 100 simulations of continuous data)					
	Level for α				
Level for λ_1	0.001	0.005	0.025	0.1	1 ^a
0.01	0.00	0.00	0.42	1.19	3.36
0.05	3.03	3.66	3.15	4.52	5.62
0.1	3.13	3.85	3.47	5.78	6.65
0.5	4.73	4.82	6.37	7.23	8.55
1	5.56	5.65	7.21	7.40	8.23

^aSpherical clusters of the same size, fulfilling the assumptions of Ward's method.

Table 1 shows the parameters used in the simulation runs and a measure of the overlap between simulated clusters, the average percentage of nearest neighbours, defined by Euclidean distance, from a different cluster (PNND). This index constitutes a measure of the clustering difficulty of a simulated data set.

The simulated bivariate normal data were then mapped onto a seven-point scale by categorizing values, in each dimension, in the intervals $(-\infty, -0.75]$, $(-0.75, -0.25]$, $(-0.25, 0.25]$, $(0.25, 0.75]$, $(0.75, 1.25]$, $(1.25, 1.75]$, $(1.75, \infty)$ onto a grid with discrete levels 1, . . . , 7. We thus used equal sized intervals for the bulk of the data. This is equivalent to the converse transformation, whereby rating scales are analysed as if they produced interval level with equal sized intervals between points (see Streiner and Norman, 2003 for a discussion of this issue). In the original simulations data lay predominantly in the $[-1, 2] \times [-1, 2]$ square, with values above 2 or below -1 only occurring for large cluster size parameters ($\lambda_1 = 0.5$ or 1). Hence 1.75 and -0.75 were chosen as the limits for the largest and smallest interval respectively. We chose to use a scale of length seven because Miller (1956) suggested that, in a large number of different situations, a scale of seven points (plus or minus two) was about the limit of the human capacity for discriminating between categories. Munshi (1990) empirically assessed categorizations of continuous scales and found that splitting the continuous scale into seven discrete intervals best matched the internal responses of his subjects. Seven is thus a reasonable upper limit for semantic scales, and Streiner and Norman (2002) suggest that the minimum number of categories in semantic scales should be between five and seven.

This gave two simulated datasets to be cluster-analysed, one continuous and one gridded. Column 1 of Figure 1 illustrates some simulated bivariate data sets and column 2 illustrates the resulting gridded data. This figure indicates the level of overlap and correlation to be expected for varying values of the parameters, and also gives the PNND (measure of clustering difficulty) of these values.

Measure of performance

In the 'unknown number of clusters' case, the obvious cluster recovery performance criterion, the percentage of points misclassified, is not available. This is because true clusters and clusters found by the clustering method

do not necessarily have one-to-one correspondence and, indeed, there may be different numbers of clusters in the two partitions being compared. For this reason the *adjusted* (or *corrected*) *Rand index* (Hubert and Arabie, 1985), which depends only on the summed agreement between pairs of cases, was used.

Agreement between assigned and true clusters for any particular pair of cases arises when:

- the pair is assigned to the same cluster if both belong to the same (true) cluster; or
- the pair is assigned to different clusters if both belong to different (true) clusters.

The maximum is 1 for perfect agreement and the value for chance agreement, given the marginal totals, approaches 0.

Performance of model-based methods in identifying the simulated clusters

Stage 1: continuous data, assuming that the number of clusters is known

The comparisons of primary interest are made under 'real-life' conditions, when the number of clusters is unknown (see stages 2 and 3 below). However preliminary runs were first made on continuous data assuming that this number was known (using the values actually used in the simulations, which varied between 2 and 6, with a median of 4). This was aimed at establishing a benchmark level of performance for later more realistic runs. The default value for $\alpha^{(s)}$ (0.2) was used. Table 2 shows the values of the adjusted Rand index (medians over the simulation runs with the indicated parameters) in this situation. Performance of both methods ranged from almost perfect clustering success (Rand indices about 0.9) to almost no success (Rand indices about 0.05) as clustering difficulty increased; see the PNND values in Table 1. As expected, method S was superior to Ward's method in 19/20 conditions in which the data were drawn from subpopulations with non-spherical distributions ($\alpha < 1$). Where $\alpha = 1$, i.e. where Ward's method would be expected to do well, Ward's method was indeed very slightly superior.

Stage 2: continuous data, without assuming that the number of clusters is known

Table 3 shows adjusted Rand indices for the S method, with $\alpha^{(s)}$ set to the default value of 0.2, and Ward's method, both applied to continuous data without

(a) $N = 82$, $\lambda_1 = 0.01$, $\alpha = 0.025$, PNND 0.42 (b) $N = 94$, $\lambda_1 = 0.1$, $\alpha = 0.005$, PNND 3.85
(c) $N = 87$, $\lambda_1 = 0.5$, $\alpha = 0.005$, PNND 4.82 (d) $N = 80$, $\lambda_1 = 1$, $\alpha = 0.025$, PNND 7.21.

Figure 1. Simulated data from bivariate normal distributions with 4–6 clusters (first column) and the corresponding gridded data (second column).

Table 2. Performance of S and Ward’s methods on continuous data, using actual number of clusters as simulated

Adjusted Rand index: median over simulation runs (the median number of clusters is 4 in all cases)										
(a) S method (setting $\alpha^{(S)} = 0.2$)						(b) Ward’s method				
α	0.001	0.005	0.025	0.1	1	0.001	0.005	0.025	0.100	1
λ_1										
0.01	0.961	0.905	0.908	0.878	0.605	0.853	0.797	0.835	0.839	0.654
0.05	0.597	0.579	0.607	0.472	0.266	0.445	0.447	0.482	0.454	0.269
0.1	0.489	0.464	0.454	0.278	0.155	0.340	0.322	0.350	0.265	0.173
0.5	0.250	0.243	0.179	0.105	0.032	0.129	0.155	0.140	0.108	0.032
1	0.177	0.170	0.126	0.081	0.012	0.091	0.101	0.096	0.074	0.013

Table 3. Performance of S and Ward's methods on continuous data, estimating the number of clusters from the AWE

Adjusted Rand index: median over simulation runs (subscript is the number of clusters estimated from the AWE)										
(a) S method (setting $\alpha^{(s)} = 0.2$)						(b) Ward's method				
α	0.001	0.005	0.025	0.1	1	0.001	0.005	0.025	0.100	1
λ_1										
0.01	0.844 _{2.5}	0.785 ₃	0.620 _{2.5}	0.375 ₂	0.292 ₂	0.809 ₄	0.768 ₄	0.820 ₄	0.801 ₃	0.642 ₄
0.05	0.527 ₃	0.530 ₃	0.536 ₃	0.414 ₃	0.190 ₃	0.405 ₇	0.400 ₇	0.420 ₆	0.397 ₆	0.215 ₆
0.1	0.470 ₃	0.438 ₃	0.431 ₃	0.255 ₃	0.153 ₃	0.317 ₈	0.316 ₈	0.316 ₇	0.250 ₆	0.118 ₈
0.5	0.253 ₃	0.241 ₃	0.162 ₃	0.084 ₃	0.032 ₄	0.177 ₁₁	0.181 ₁₂	0.163 ₁₁	0.095 ₉	0.022 _{12.5}
1	0.176 ₃	0.158 ₃	0.110 ₃	0.068 ₃	0.012 ₄	0.146 ₁₆	0.145 _{14.5}	0.125 ₁₂	0.071 ₁₂	0.009 ₁₄

Table 4. Performance of S and Ward's methods on ordinal data, estimating the number of clusters from the AWE

Adjusted Rand index: median over simulation runs (subscript is the number of clusters estimated from the AWE)										
(a) S method (setting $\alpha^{(s)} = 0.2$)						(b) Ward's method				
α	0.001	0.005	0.025	0.1	1	0.001	0.005	0.025	0.1	1
λ_1										
0.01	0.581 ₅	0.548 ₅	0.570 ₅	0.541 ₅	0.451 ₆	0.581 ₅	0.548 ₅	0.560 ₅	0.535 ₅	0.451 ₆
0.05	0.362 ₅	0.338 ₆	0.341 ₄	0.299 ₅	0.192 ₅	0.366 ₇	0.339 ₇	0.340 _{7.5}	0.310 ₈	0.180 ₁₁
0.1	0.297 ₅	0.267 ₅	0.287 ₄	0.220 ₅	0.117 ₅	0.275 ₉	0.261 ₉	0.261 _{9.5}	0.208 ₁₀	0.084 ₁₅
0.5	0.182 ₅	0.173 ₄	0.141 ₅	0.097 ₄	0.029 ₄	0.156 ₁₇	0.144 ₁₈	0.111 ₂₀	0.051 ₂₁	0.007 ₃₂
1	0.127 ₅	0.138 ₅	0.113 ₄	0.062 ₄	0.014 ₄	0.109 ₂₂	0.108 _{23.5}	0.069 ₂₆	0.032 ₂₉	0.002 _{38.5}

assuming anything about the number of clusters. For these runs the AWE-predicted number of clusters was chosen. The median number of clusters, predicted by the AWE, is shown as a subscript (the actual median number of simulated clusters was 4 in all cases).

As might be expected, estimating the number of clusters reduces performance levels for both methods – compare Tables 2 and 3. For well separated continuous data from spherical cluster distributions ($\alpha^{(s)} = 1$, λ_1 low) Ward's method was superior to the S method: see final columns of Tables 3a and 3b. Ward's method was also superior for well-separated clusters, even when there was some degree of correlation. However performance of the S method was superior to Ward's method for moderate to well-separated clusters in very highly correlated continuous data, as in Figures 1b and c. This is shown in columns 1 and 2 of Tables 3a and 3b. For clusters that were neither well separated nor with very

high correlation, as for example in Figure 1d, both methods performed poorly, with little ability to detect clusters above chance. This is shown in the bottom right hand corner of Tables 3a and 3b. The AWE-predicted number of clusters for S was underestimated (typically by 1) compared with the true median of 4: see Table 3a. Underestimation was most marked for well-separated clusters (λ_1 low). The number of clusters predicted by the AWE using Ward's method was generally too high and very inaccurate for other than very well separated clusters: see Table 3b.

Stage 3: ordinal data, without assuming that the number of clusters is known

Table 4 shows adjusted Rand indices for the S method and Ward's method applied to ordinal data without assuming knowledge of the number of clusters. Performance on ordered categorical data was worse than on

the equivalent continuous data, as expected, and there was little to choose between Ward's method and the S method, whatever the values of the α and λ_i parameters. There was a very slight advantage in using the S method when the clusters were not well separated and when the underlying correlation was very high: see lower halves of the first columns of Tables 4a and 4b. The S method overestimated the number of clusters, rather than underestimating it as in the case of the continuous data, typically by 1: see Table 4a. As before, the number of clusters predicted using Ward's method was very inaccurate for other than very well separated clusters, and became worse as the clusters became more spherical and/or less well separated: see Table 4b.

An application to data on high security psychiatric hospital patients

In order to illustrate the practical steps involved, including the necessary S-PLUS commands, and to show how correlations within clusters can be accommodated, a clustering of data from an assessment of the individual and placement needs of high security psychiatric hospital patients in England (Harty et al., 2004) is illustrated. Patients, classified as mentally ill, personality disordered or mentally impaired, were assessed for their service needs using the Security, Dependency and Treatment Needs Scale (SDTP) (Shaw et al., 2001), an 11 (0–10) point scale completed by the primary nurse with 10 representing low needs and 0 high needs. Security needs included items relating to violence risk and absconson risk; dependency needs reflected the degree to which the person could function independently, and treatment needs encapsulated compliance and supervision issues (Shaw et al., 2001).

Patients in high-security hospitals are referred from a number of sources and, while they will all exhibit some degree of need in all three areas, it is feasible to suppose that there are distinct subgroups represented among them. For example one might expect to see a subgroup, weighted towards those classified under the legal definition of personality disorder, with relatively high security needs, moderate treatment needs and low dependency needs. Another subgroup, weighted towards those classified as having with mental impairment, might be characterized by high mean dependency, and there may well be other subgroups of interest. If this is so, the overall sample can be modelled conceptually as a mixture of distributions, which would be expected to overlap and within which the three types of needs

might be correlated in some way. The statistical mixture model therefore seems to be appropriate *a priori*, apart perhaps from the ordinal nature of the data. However the simulations reported in this paper have shown the ordinal nature of the data is not necessarily a barrier to the use of the model-based methods discussed (if the clusters are well-defined). Figure 2 shows the S-PLUS commands necessary for a cluster analysis of the SDTP data using S and Ward's method.

First a random number seed is set; this is necessary to ensure that the same results are obtained on rerunning the program (because a random starting configuration is used for the clustering algorithm). The less clear-cut the clusters are, the more likely such numerical instability is to be a problem. The first 100 records for three variables are read into a matrix, after excluding three missing values. The S method with the default setting of $\alpha^{(s)}$ (0.2) is applied and the AWE printed. This suggests three clusters, so the cases are classified into three groups. Ward's method is then applied and the two three-cluster solutions found by S and Ward's method are compared in a crosstabulation.

There is an obvious correspondence between the clusters found by the S method and Ward's method, although the clusters 1 and 3 from both methods are somewhat overlapping. (Note that the numbering is arbitrary but in this case they are obviously equivalent.) The cluster means (low values = high needs) are shown for the S method but are similar using Ward's method. These suggest that the data can be characterized by a cluster with moderately high needs in all three areas; a cluster with mainly treatment needs, and a third cluster with treatment and security needs but little need in the area of dependency.

Since it is the ability of model-based clustering methods such as S to pick up within-cluster correlation that makes them attractive, this aspect of the clustering is now the focus. The overall correlations were -0.41 for Security and Treatment, 0.31 for Dependency and Security, and -0.36 for Treatment and Dependency. The pooled within-cluster correlations for Security and Dependency dropped to 0.11 after clustering with S (which allows for within-cluster correlation). The small Security-Dependency correlation that remained after clustering is perhaps more credible than the moderately high correlation found before clustering. Dependency is associated with lack of ability or will to abscond or act violently (an ability that would generally lead to security needs) so that a positive


```

Set random number seed; read data set, 3 11-point items from SDTP scale, 1st 100 rows
set.seed(123111)
SDTP <- sdtpt[1:100, c(2:4)]

Remove missing values and assign data to a data matrix
exclude <- na.exclude(SDTP)
overall <- as.matrix(exclude)

Run the S method with the default alpha (0.2; 1st parameter is always 1)
modelS <- mclust(x = overall, method = "S", shape = c(1, rep(0.2, 2)))

Print out the AWE: maximum for these data is at 3 clusters
modelS$awe
[1]      0.000000      268.150757      418.362915      409.566406      390.356232
etc

Classify cases into one of the 3 clusters found by S
classmodelS <- mclass(modelS, 3)

Repeat clustering using Ward's method (use 3-cluster solution)
modelW <- mclust(x = overall, method = "trace")
classmodelW <- mclass(modelW, 3)

Compare the S and Ward's method results
ctable <- table(classmodelS$classification, classmodelW$classification)
      ctable
      1  2  3
1 27  0  2
2  0 42  0
3  8  1 17

Add S method classification to original data
clusterdata <- cbind(overall, classmodelS$classification)

Compute means for each cluster
aggregate( clusterdata[, c(1:3)], by = clusterdata[, 4], FUN = mean)
      sdtpts1  sdtpt1  sdtptd1
1      1 3.931034 4.379310 3.206897
2      2 8.190476 2.428571 8.904762
3      3 2.307692 3.384615 8.653846

Overall & within-cluster correlation (based on residuals from regression on cluster)
within <- lm(formula = clusterdata[, c(1, 2, 3)] ~ factor(clusterdata[,4]))$res

cor(overall)
      sdtpts1      sdtpt1      sdtptd1
sdtpts1  1.0000000 -0.4090060  0.3111045
sdtpt1 -0.4090060  1.0000000 -0.3614394
sdtptd1  0.3111045 -0.3614394  1.0000000

cor(within)
      sdtpts1      sdtpt1      sdtptd1
sdtpts1  1.0000000 -0.3957878  0.1143610
sdtpt1 -0.3957878  1.0000000 -0.2796649
sdtptd1  0.1143610 -0.2796649  1.0000000

```

Figure 2. Typical S-PLUS commands to run S and Ward's method.

association seems somewhat implausible. This particular example illustrates the general point that if indeed the clusters are genuine and represent distinct subgroups of patients post-clustering correlations are likely to be better estimates of the underlying associations, given subgroup membership, than the simple overall estimates.

Discussion

Summary of findings

For continuous data using the correct number of clusters the methods performed as expected, with the S method (using the default setting for $\alpha^{(s)}$) outperforming Ward's method for non-spherical clusters. The converse was true with Ward's method outperforming the S method for spherical clusters, not surprisingly because the model underlying Ward's method is appropriate for this type of data and the S method would be imposing a substantial correlation where none existed.

In the more interesting, realistic scenario, where the number of clusters has to be estimated from the AWE, the S method using the default setting for $\alpha^{(s)}$ was worthwhile for continuous data with substantial within-cluster correlation and low-to-moderate overlap. Figures 1b and c (left hand side) give an idea of what this might look like in a scatter plot. In this scenario the S method outperformed Ward's method. However, for data from clusters with spherical and equal size distributions Ward's method was superior.

For ordinal data the most striking finding was that differences in performance between the methods were minimal. There was, however, still a slight benefit for the S method, which was more marked for correlated clusters with the most overlap. As was the case for continuous data, the AWE-estimated number of clusters was very inaccurate for Ward's method (except for very well separated clusters). Having said this, neither method worked very well for ordinal data, probably due to the attenuation of correlation and consequent increase in overlap when data are mapped from a continuous scale onto a grid. If the mapping process is viewed as introducing a measurement error to an underlying latent variable, the attenuation of correlation is as expected.

Although the simulation results suggested very little difference in performance between S and Ward's method, we concluded from the practical example that the S method was nevertheless useful. Firstly, the AWE derived from the S method estimated the number of clusters reasonably well. Secondly, it allowed

investigation of the within-cluster correlation without imposing a spherical distribution (as Ward's method does). Within-cluster correlation may be a more accurate reflection of underlying associations between variables than correlation over the combined sample, since the latter may be induced by the variation in cluster means.

Limitations of the simulation study

A limitation of this study is the relatively restricted range of α values (low values plus $\alpha = 1$) and the use of two-dimensional data. These are, however, the situations in which model-based methods are expected to perform best and they were based on the original simulations of Banfield and Raftery (1993), which were used as a benchmark. In fact performance dropped off markedly as α increased. This implies that simulation using moderate values for α would not have been informative. A further limitation is that the routines tested are those available in the commercial version of S-PLUS. Extensions to the method have been developed which use the iterative 'EM' (Expectation-Maximization) method rather than the hierarchical method (see Fraley and Raftery, 2002). This however is not a part of the standard S-PLUS package although it can be downloaded from <http://www.stat.washington.edu/mclust/>. This method finds the number of clusters as part of the overall estimation process. It is possible that this might perform better than the method tested here. It is worth noting that this Web page also has links to the R package.

Recommendations

The findings here suggest that model-based methods as implemented in the commercially available S-PLUS package are relatively easy to apply and would be worthwhile considering for continuous data, although less useful for ordinal scales as often used in psychology and psychiatry. For the latter Ward's method may be almost as successful in recovering clusters. The main problem is in the use of the AWE to predict the number of clusters with Ward's method since this is very inaccurate. Furthermore, although the simulation study showed little advantage in terms of cluster recovery performance for the S method, it may have benefits in terms of allowing the investigation of within-cluster correlation. The best approach may be to run both methods in parallel, using the AWE from the S method to suggest the number of clusters, and to compare the within-cluster and overall correlations.

References

- Arabie P, Hubert LJ, De Soete G. Clustering and Classification. Singapore: World Scientific, 1996.
- Banfield JD, Raftery AE. Model-based Gaussian and non-Gaussian clustering. *Biometrics* 1993; 49: 803–22.
- Bellivier F, Golmard J-L, Rietschel M, Schulze TG, Malafosse A, Preisig M, McKeon P, Mynett-Johnson L, Henry C, Leboyer M. Age at onset in bipolar I affective disorder: further evidence for three subgroups. *Am J Psychiatry* 2003; 160: 999–1001.
- Bollen KA, Barb KH. Pearson's R and Coarsely Categorized Measures. *American Sociological Review* 1981; 46: 232–39.
- Everitt BS, Landau S, Leese M. Cluster Analysis. 4 edn. London: Edward Arnold, 2001.
- Fraley C, Raftery E. Model-based clustering, discriminant analysis, and density estimation *JASA* 2002; 97: 611–31.
- Gordon AE. Classification. 2 edn. New York: Chapman & Hall, 1999.
- Harty M-A, Shaw J, Thomas S, Dolan M, Davies L, Thornicroft G, Carlisle J, Moreno M, Leese M, Appleby L, Jones P. The security, clinical and social needs of patients in high security psychiatric hospitals in England. *J Forensic Psychiatr Psychol* 2004; 5: 208–21.
- Hubert LJ, Arabie P. Comparing partitions. *Journal of Classification* 1985; 2: 193–218.
- Jorgensen M, Hunt LA. Mixture model clustering using the MULTIMIX program. *Aust New Zeal J Stat* 1999; 41: 153–71.
- Krause A, Olsen M. The Basics of S and S-PLUS. 2 edn. New York: Springer, 2000.
- Likert R. A technique for the measurement of attitudes. *Arch Psychol* 1932; 140: 5–55.
- McLachlan GJ. Cluster analysis and related techniques in medical research. *Stat Meth Med Res* 1992; 1: 27–48.
- MacQueen JB. Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1967; 1: 281–97.
- Meehl PE. Factors and taxa, traits and types, differences of degree and differences in kind. *J Pers* 1992; 60: 117–74.
- Miller GA. The magic number seven plus or minus two: Some limits on our capacity for processing information. *Psychol Rev* 1956; 63: 81–97.
- Munshi J. A method for constructing Likert scales. Research Report, Sonoma State University, CA 1990: <http://munshi.4t.com/papers/likert.html>.
- Papageorgiou I, Baxter MJ, Cau MA. Model-based cluster analysis of artefact compositional data. *Archaeometry* 2001; 43: 571–88.
- Shaw J, Davies J, Morey H. An assessment of the security, dependency and treatment needs of all patients in secure services in a UK health region. *J Forensic Psychiatr* 2001; 12: 610–37.
- Streiner DL, Norman GR. Health Measurement Scales: A Practical Guide to their Development and Use. 3 edn. Oxford: Oxford Medical Publications, 2002.
- Scott AJ, Symons M. Clustering methods based on likelihood ratio criteria. *Biometrics* 1971; 27: 387–97.
- Ward JH. Hierarchical groupings to optimize an objective function. *J Am Stat Assoc* 1963; 58: 236–44.

Correspondence: Morven Leese, P029, Department of Health Services Research, Institute of Psychiatry, King's College London, De Crespigny Park, London SE5 8AF, UK.

Telephone (+44) (0) 20 7848 0710

Fax (+44) (0) 20 7277 1462

Email: m.leese@iop.kcl.ac.uk