Statistical Methods in Medical Research 2004; **13**: 347–361

Mixture modelling for cluster analysis

GJ McLachlan Department of Mathematics and the Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia and **SU Chang** Department of Mathematics, University of Queensland, Brisbane, Australia

Cluster analysis via a finite mixture model approach is considered. With this approach to clustering, the data can be partitioned into a specified number of clusters g by first fitting a mixture model with g components. An outright clustering of the data is then obtained by assigning an observation to the component to which it has the highest estimated posterior probability of belonging; that is, the ith cluster consists of those observations assigned to the ith component ($i = 1, \ldots, g$). The focus is on the use of mixtures of normal components for the cluster analysis of data that can be regarded as being continuous. But attention is also given to the case of mixed data, where the observations consist of both continuous and discrete variables.

1 Introduction

Finite mixture models are being widely used in medical and other applications to model the distributions of a wide variety of random phenomena and to cluster data sets. Examples may be found in the recent monograph of McLachlan and Peel.[1] Here we focus on applications of mixture models where the clustering of the data at hand is the primary aim of the analysis. In this case, the mixture model is being used purely as a device for exposing any grouping that may underlie the data. McLachlan and Basford[2] have highlighted the usefulness of mixture models as a way of providing an effective clustering of various data sets under a variety of experimental designs.

With a mixture model based approach to clustering, it is assumed that the data to be clustered are from a mixture of an initially specified number g of groups in various proportions. That is, each data point $\boldsymbol{y}_j$ is taken to be a realization of the mixture density

$$f(\boldsymbol{y}_j) = \sum_{i=1}^{g} \pi_i \, f_i(\boldsymbol{y}_j) \tag{1}$$

where the g components correspond to the g groups. In Equation (1), the $f_i(\boldsymbol{y}_j)$ & are densities and the π_i & are non-negative quantities (the mixing proportions) that sum to 1. The $f_i(\boldsymbol{y}_j)$ are called the *component densities* of the mixture.

Address for correspondence: GJ McLachlan, Department of Mathematics, The University of Queensland, Brisbane, Queensland 4072, Australia. E-mail: gjm@maths.uq.edu.au

 10.1191/0962280204sm372ra

On specifying a parametric form $f_i(\boldsymbol{y}_j; \boldsymbol{\theta}_i)$ for each component density, we can fit this parametric mixture model

$$f(\boldsymbol{y}_j; \Psi) = \sum_{i=1}^{g} \pi_i f_i(\boldsymbol{y}_j; \boldsymbol{\theta}_i) \tag{2}$$

by maximum likelihood (ML) via the expectation maximization (EM) algorithm of Dempster *et al.*[3] (refer to McLachlan and Krishnan[4]). Here $\Psi = (\boldsymbol{\xi}^{\mathrm{T}}, \pi_1, \ldots, \pi_{g-1})^{\mathrm{T}}$ is the vector of unknown parameters, where $\boldsymbol{\xi}$ consists of the elements of $\boldsymbol{\theta}_i$ known *a priori* to be distinct. Once the mixture model has been fitted, a probabilistic clustering of the data into g clusters can be obtained in terms of the fitted posterior probabilities of component membership for the data. An outright assignment of the data into g clusters is achieved by assigning each data point to the component to which it has the highest estimated posterior probability of belonging. Although these estimated posterior probabilities may have limited reliability in small samples, they may well give a satisfactory outright assignment of the data. The choice of the number of components g in the mixture model can be considered on the basis of the likelihood, which is discussed in Section 4.

In the earlier case, there is a one to one correspondence between the mixture components and the groups. For multivariate data of a continuous nature, attention has been concentrated on the use of multivariate normal components because of their computational convenience. In those cases where the underlying population consists of g groups in each of which the feature vector is able to be modelled by a single normal distribution, the number of components g in the fitted normal mixture model corresponds to the number of groups. However, when the distribution of a group is unable to be modelled adequately by a single normal distribution but rather needs a normal mixture distribution, the components in the fitted g component normal mixture model and in the consequent clusters will correspond to g subgroups rather than to the smaller number of actual groups represented in the data.

It can be seen that this mixture likelihood based approach to clustering is model based in that the form of each component density of an observation has to be specified in advance. Hawkins *et al.*[5] commented that most writers on cluster analysis 'lay more stress on algorithms and criteria in the belief that intuitively reasonable criteria should produce good results over a wide range of possible (and generally unstated) models'. For example, the trace $\boldsymbol{W}$ criterion, where $\boldsymbol{W}$ is the pooled within-cluster sums of squares and products matrix, is predicated on normal groups with (equal) spherical covariance matrices; but as they pointed out, many users apply this criterion even in the face of evidence of nonspherical clusters or, equivalently, would use Euclidean distance as a metric. They strongly supported the increasing emphasis on a model based approach to clustering. Indeed, as remarked by Aitkin *et al.*[6] in the reply to the discussion of their paper, 'when clustering samples from a population, no cluster method is *a priori* believable without a statistical model'. Concerning the use of mixture models to represent nonhomogeneous populations, they noted in their paper that 'Clustering methods based on such mixture models allow estimation and hypothesis testing within the framework of standard statistical theory'. Previously, Marriott[7] had

noted that the mixture likelihood based approach 'is about the only clustering technique that is entirely satisfactory from the mathematical point of view. It assumes a well defined mathematical model, investigates it by well established statistical techniques and provides a test of significance for the results'.

We shall focus here on normal mixture models for the clustering of continuous data, where the ith component density for the jth observation $\boldsymbol{y}_j$ is specified as

$$f_i(\boldsymbol{y}_j; \boldsymbol{\theta}_i) = \phi(\boldsymbol{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{3}$$

and $\phi(\boldsymbol{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ denotes the p variate normal density function with mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i (i = 1, \ldots, g)$.

A robust version is presented too by using the t distribution in place of the normal in the specification of the component distributions in the mixture model. We also consider the case of high dimensional data, where the number of parameters is large relative to the number of observations, we consider the use of mixtures of factor analysers. This approach enables a normal mixture model to be fitted to a sample of n data points of dimension p, where p is large relative to n. The number of free parameters is controlled through the dimension of the latent factor space. By working in this reduced space, it allows a model for each component covariance matrix with complexity lying between that of the isotropic and full covariance structure models. The extension of the normal mixture model to handle the clustering of data with mixed variables (continuous and mixed features) is covered.

2 ML estimation

The ML estimate of Ψ is obtained as an appropriate root of the likelihood equation

$$\frac{\partial \log L(\Psi)}{\partial \Psi} = 0 \tag{4}$$

where $L(\Psi)$ denotes the likelihood function for Ψ formed from the observed random sample $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$. Solutions of Equation (4) corresponding to local maxima can be found by application of the EM algorithm. The EM algorithm is applied in the framework where an observation $\boldsymbol{y}_j$ is conceptualized to have arisen from one of the components and the indicator vector $\mathbf{z}_j$ denoting its component of origin is taken to be missing, where $z_{ij} = (\mathbf{z}_j)_i$ is defined to be one or zero, according as to whether $\boldsymbol{y}_j$ did or did not arise from the ith component of the mixture ($i = 1, \ldots, g; j = 1, \ldots, n$). The complete data vector is therefore declared to be

$$\boldsymbol{y}_c = (\boldsymbol{y}^T, \mathbf{z}^T)^T \tag{5}$$

where

$$\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T)^T \tag{6}$$

contains the observed data and

$$\mathbf{z} = (\mathbf{z}_1^{\mathrm{T}}, \ldots, \mathbf{z}_n^{\mathrm{T}})^{\mathrm{T}} \tag{7}$$

contains the unobservable component indicator variables.

The complete data log likelihood for Ψ, $\log L_{\mathrm{c}}(\Psi)$, is given by

$$\log L_{\mathrm{c}}(\Psi) = \sum_{i=1}^{g} \sum_{j=1}^{n} z_{ij} \{ \log \pi_i + \log f_i(\boldsymbol{y}_j; \boldsymbol{\theta}_i) \} \tag{8}$$

The E step of the EM algorithm requires averaging the complete data log likelihood $\log L_{\mathrm{c}}(\Psi)$ over the conditional distribution of $\mathbf{z}$ given the observed data vector $\boldsymbol{y}$, using the current fit for the vector of unknown parameters Ψ. As $\log L_{\mathrm{c}}(\Psi)$ is linear in the unobservable data z_{ij}, the E step [on the $(k+1)$th iteration] simply requires the calculation of the current conditional expectation of Z_{ij} given the observed data $\boldsymbol{y}$, where Z_{ij} is the random variable corresponding to z_{ij}. This yields the Q function given by

$$Q(\Psi; \Psi^{(k)}) = \sum_{i=1}^{g} \sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j; \Psi^{(k)}) \{ \log \pi_i + \log f_i(\boldsymbol{y}_j; \boldsymbol{\theta}_i) \} \tag{9}$$

where

$$\begin{aligned} \tau_i(\boldsymbol{y}_j; \Psi^{(k)}) &= E_{\Psi^{(k)}}(Z_{ij} \mid \boldsymbol{y}) \\ &= \mathrm{pr}_{\Psi^{(k)}}\{Z_{ij} = 1 \mid \boldsymbol{y}_j\} \\ &= \frac{\pi_i^{(k)} f_i(\boldsymbol{y}_j; \boldsymbol{\theta}_i^{(k)})}{f(\boldsymbol{y}_j; \Psi^{(k)})} \end{aligned} \tag{10}$$

is the posterior probability that the jth observation $\boldsymbol{y}_j$ belongs to the ith component $(i = 1, \ldots, g; j = 1, \ldots, n)$. In Equation (10), $E_{\Psi^{(k)}}$ and $\mathrm{pr}_{\Psi^{(k)}}$ denote expectation and probability, respectively, using the current value $\Psi^{(k)}$ for Ψ.

The M step on the $(k+1)$th iteration requires the global maximization of $Q(\Psi; \Psi^{(k)})$ with respect to Ψ over the parameter space Ω to give the updated estimate $\Psi^{(k+1)}$. The updated estimate of the ith mixing proportion π_i is given then by

$$\pi_i^{(k+1)} = \sum_{j=1}^{n} \frac{\tau_i(\boldsymbol{y}_j; \Psi^{(k)})}{n}, \quad i = 1, \ldots, g \tag{11}$$

The updated estimate of the vector $\boldsymbol{\xi}$ containing the distinct parameters in the component densities satisfies the equation

$$\sum_{i=1}^{g}\sum_{j=1}^{n}\tau_{ij}(\boldsymbol{y}_j;\,\boldsymbol{\Psi}^{(k)})\frac{\partial \log f_i(\boldsymbol{y}_j;\,\theta_i)}{\partial \boldsymbol{\xi}} = 0 \tag{12}$$

For normal component densities, the estimates of the component means $\boldsymbol{\mu}_i$ and covariance matrices $\boldsymbol{\Sigma}_i$ are given in closed form, namely

$$\boldsymbol{\mu}_i^{(k+1)} = \frac{\sum_{j=1}^{n}\tau_{ij}^{(k)}\boldsymbol{y}_j}{\sum_{j=1}^{n}\tau_{ij}^{(k)}} \tag{13}$$

and

$$\boldsymbol{\Sigma}_i^{(k+1)} = \frac{\sum_{j=1}^{n}\tau_{ij}^{(k)}(\boldsymbol{y}_j-\boldsymbol{\mu}_i^{(k+1)})(\boldsymbol{y}_j-\boldsymbol{\mu}_i^{(k+1)})^T}{\sum_{j=1}^{n}\tau_{ij}^{(k)}} \tag{14}$$

for $i = 1, \ldots, g$, where $\tau_{ij}^{(k)} = \tau_i(\boldsymbol{y}_j;\,\boldsymbol{\Psi}^{(k)})$.

The E and M steps are alternated repeatedly until the difference

$$L(\boldsymbol{\Psi}^{(k+1)}) - L(\boldsymbol{\Psi}^{(k)})$$

changes by an arbitrarily small amount in the case of convergence of the sequence of likelihood values $\{L(\boldsymbol{\Psi}^{(k)})\}$. Dempster *et al.*[3] showed that the (incomplete data) likelihood function $L(\boldsymbol{\Psi})$ is not decreased after an EM iteration, that is,

$$L(\boldsymbol{\Psi}^{(k+1)}) \geq L(\boldsymbol{\Psi}^{(k)}) \tag{15}$$

for $k = 0, 1, 2, \ldots$. Hence, convergence must be obtained with a sequence of likelihood values that are bounded above. Recently, Ng and McLachlan[8] have investigated the speeding up the fitting of normal mixtures by the use of the incremental EM algorithm and variants whereby the available observations are divided into B ($B \leq n$) blocks and the E step is implemented for only a block of observations at a time before the next M step is performed.

As the likelihood equation (4) tends to have multiple roots corresponding to local maxima, the EM algorithm needs to be started from a variety of initial values for the parameter vector $\boldsymbol{\Psi}$ or for a variety of initial partitions of the data into g groups. The latter can be obtained by randomly dividing the data into g groups corresponding to the g components of the mixture model. With random starts, the effect of the central limit theorem tends to have the component parameters initially being similar at least in large samples. One way to reduce this effect is to first select a small random subsample from the data, which is then randomly assigned to the g components. The first M step is then performed on the basis of the subsample. The

subsample has to be sufficiently large to ensure that the first M step is able to produce a nondegenerate estimate of the parameter vector Ψ. Coleman *et al.*[9] have considered using a combinatorial search for a good starting point to apply the EM algorithm. They compared two local searches with a hierarchical agglomerative approach where the objective function to be minimized was taken to be the determinant of the pooled within-cluster covariance matrix $\boldsymbol{W}$.[10]

The choice of root of the likelihood equation in the case of homoscedastic components is straightforward in the sense that the ML estimate exists as the global maximizer of the likelihood function. The situation is less straightforward in the case of heteroscedastic components as the likelihood function is unbounded. In practice, consideration has to be given to the problem of relatively large local maxima that occur as a consequence of a fitted component having a very small (but non-zero) variance for univariate data or generalized variance (the determinant of the covariance matrix) for multivariate data. Such a component corresponds to a cluster containing a few data points either relatively close together or almost lying in a lower dimensional subspace in the case of multivariate data. There is thus a need to monitor the relative size of the fitted mixing proportions and of the component variances for univariate observations, or of the generalized component variances for multivariate data, in an attempt to identify these spurious local maximizers.

The reader is referred to the appendix in McLachlan and Peel[1] for the availability of software for the fitting of normal mixture models, including the EMMIX programme of McLachlan *et al.*[11] The current version of EMMIX is available from http://www.maths.uq.edu.au/~gjm/emmix/emmix.html. Concerning the availability of mixture modelling facilities in general purpose statistical packages, there is the MCLUST software package of Fraley and Raftery,[12] which is interfaced to the S-PLUS commercial software.

3 Classification likelihood approach

Another likelihood based approach to clustering is what is sometimes called the classification likelihood approach, whereby Ψ and $\mathbf{z}$ are chosen to maximize the complete data log likelihood $\log L_c(\Psi)$. That is, $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are treated as unknown parameters to be estimated along with Ψ. This procedure has been considered by several authors, Hartley and Rao,[13] John[14] and Scott and Symons,[15] among others. This maximization can be approached iteratively, using the EM equations[16] in which the current estimate $\tau_i(\boldsymbol{y}_j; \Psi^{(k)})$ of the posterior probability of ith component membership of the mixture model is replaced by $z_{ij}^{(k)}$, which is equal to one if

$$i = \arg\max_h \tau_h(\boldsymbol{y}_j; \Psi^{(k)})$$

and is zero otherwise. Unfortunately, with this procedure, the $\mathbf{z}_j$ increases in number with the number of observations, and under such conditions, the ML estimate of

Ψ need not be consistent.[17] Under the assumption of equal mixing proportions and equal spherical component covariance matrices

$$\boldsymbol{\Sigma}_i = \sigma^2 \mathbf{I}_p, \quad i = 1, \ldots, g \tag{16}$$

where $\mathbf{I}_p$ is the $p \times p$ identity matrix, the classification ML approach leads to clustering via the trace $\boldsymbol{W}$ criterion[18] or, equivalently, k means. If the sphericity assumption (16) is relaxed to equal component covariance matrices of arbitrary shape, then the classification ML approach is equivalent to the det $\boldsymbol{W}$ criterion, as originally suggested by Friedman and Rubin.[19]

4 Choice of the number of components

With a mixture model based approach to clustering, the question of how many clusters there are can be considered in terms of the number of components of the mixture model being used. It is sensible in practice to approach the latter question of the number of components g in a mixture model in terms of an assessment of the smallest number of components compatible with the data. A guide to the final choice of g can be obtained from monitoring the increase in the log likelihood as g is increased from a single component. Unfortunately, it is difficult to carry out formal tests at any stage of this sequential process for the need of an additional component, since, as is well known, regularity conditions fail to hold for the likelihood ratio statistic λ to have its usual asymptotic null distribution of chi square with degrees of freedom equal to the difference between the number of parameters under the null and alternative hypotheses.

A formal test of the null hypothesis H_0: $g = g_0$ versus the alternative H_1: $g = g_1 (g_1 > g_0)$ can be undertaken using a resampling method, as described by McLachlan.[20] Bootstrap samples are generated from the mixture model fitted under the null hypothesis of g_0 components. That is, the bootstrap samples are generated from the g_0 component mixture model with the vector Ψ of unknown parameters replaced by its ML estimate $\hat{\boldsymbol{\Psi}}_{g_0}$ computed by consideration of the log likelihood formed from the original data under H_0. The value of $-2 \log \lambda$ is computed for each bootstrap sample after fitting mixture models for $g = g_0$ and g_1 in turn to it. The process is repeated independently a number of times (B), and the replicated values of $-2 \log \lambda$ formed from the successive bootstrap samples provide an assessment of the bootstrap, and hence of the true, null distribution of $-2 \log \lambda$. It enables an approximation to be made to the achieved level of significance P corresponding to the value of $-2 \log \lambda$ evaluated from the original sample. The rth order statistic of the B bootstrap replications can be used to estimate the quantile of order $r/(B + 1)$. A preferable alternative would be to use the rth order statistic as an estimate of the quantile of order $(3r - 1)/(3B + 1)$.

A commonly used method of testing the above hypotheses is to adopt the BIC criterion of Schwarz,[21] which, when applied in the present context, leads to H_0 being rejected if twice the increase in the log likelihood (i.e., $-2 \log \lambda$) is greater than $d \log n$, where d denotes the difference between the number of parameters under the two hypotheses. Further information based criteria for tests on the number of components

in a mixture model have been considered, including the integrated classification likelihood criterion, as proposed by Biernacki *et al.*[22]

5 Mixtures of *t* distributions

For many applied problems, the tails of the normal distribution are often shorter than appropriate. Also, the estimates of the component means and covariance matrices can be affected by observations that are atypical of the components in the normal mixture model being fitted. McLachlan and Peel[23] and Peel and McLachlan[24] have considered the fitting of mixtures of (multivariate) t distributions. The t distribution provides a longer tailed alternative to the normal distribution. Hence it provides a more robust approach to the fitting of normal mixture models, as observations that are atypical of a normal component are given reduced weight in the calculation of its parameters.

The t density with location parameter $\boldsymbol{\mu}_i$, positive definite matrix $\boldsymbol{\Sigma}_i$ and v_i degrees of freedom is given by

$$\phi_t(\mathbf{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, v_i) = \frac{\Gamma((v_i + p)/2)|\boldsymbol{\Sigma}_i|^{-1/2}}{(\pi v_i)^{(1/2)p}\Gamma(v_i/2)\{1 + \delta(\mathbf{y}_j, \boldsymbol{\mu}_i; \boldsymbol{\Sigma}_i)/v_i\}^{(1/2)(v_i+p)}} \tag{17}$$

where

$$\delta(\mathbf{y}_j, \boldsymbol{\mu}_i; \boldsymbol{\Sigma}_i) = (\mathbf{y}_j - \boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_j - \boldsymbol{\mu}_i) \tag{18}$$

denotes the Mahalanobis squared distance between $\mathbf{y}_j$ and $\boldsymbol{\mu}_i$ (with $\boldsymbol{\Sigma}_i$ as the covariance matrix). If $v_i > 1$, $\boldsymbol{\mu}_i$ is the mean of $\mathbf{Y}_j$, and if $v_i > 2$, $v_i(v_i - 2)^{-1}\boldsymbol{\Sigma}_i$ is its covariance matrix. As v_i tends to infinity, $\mathbf{Y}_j$ becomes marginally multivariate normal with mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$. Hence this parameter v_i may be viewed as a robustness tuning parameter. It can be fixed in advance or it can be inferred from the data for each component, thereby providing an *adaptive* robust procedure.[1]

The t distribution does not have substantially better breakdown behaviour than the normal. The advantage of the t mixture model is that, although the number of outliers needed for breakdown is almost the same as with the normal mixture model, the outliers have to be much larger. This point is made more precise in Hennig[25] who has provided an excellent account of breakdown points for the ML estimation of location scale mixtures with a fixed number of components g.

6 Reduction in dimension of parameter vector

For the parametric mixture model (2), the vector Ψ of unknown parameters consists of the distinct elements in the component parameter vectors $\boldsymbol{\theta}_i$, along with the mixing proportions. In practice, it is important that the dimension of Ψ is not large relative to the sample size n.

The g component normal mixture model with unrestricted component covariance matrices is a highly parameterized model with $(1/2)p(p+1)$ parameters for each component covariance matrix $\mathbf{\Sigma}_i (i = 1, \ldots, g)$. Banfield and Raftery[26] introduced a parameterization of the component covariance matrix $\mathbf{\Sigma}_i$ based on a variant of the standard spectral decomposition of $\mathbf{\Sigma}_i$,

$$\mathbf{\Sigma}_i = \sum_{v=1}^{p} \lambda_{iv} \boldsymbol{a}_{iv} \boldsymbol{a}_{iv}^{\mathrm{T}} \tag{19}$$

where $\boldsymbol{a}_{i1}, \ldots, \boldsymbol{a}_{ip}$ denote the eigenvectors corresponding to the eigenvalues $\lambda_{i1} \geq \lambda_{i2} \geq \cdots \lambda_{ip} > 0$ of $\mathbf{\Sigma}_i (i = 1, \ldots, g)$. They expressed $\mathbf{\Sigma}_i$ further as

$$\mathbf{\Sigma}_i = \lambda_i \boldsymbol{A}_i \Lambda_i \boldsymbol{A}_i^{\mathrm{T}} \tag{20}$$

where $\boldsymbol{A}_i = (\boldsymbol{a}_{i1}, \ldots, \boldsymbol{a}_{ip})$ is the (orthogonal) matrix of the eigenvectors of $\mathbf{\Sigma}_i$. Conventions for normalizing λ_i and Λ_i include taking $\lambda_i = \lambda_{i1}$ (the largest eigenvalue of $\mathbf{\Sigma}_i$) for which then

$$\Lambda_i = \mathrm{diag}\left(1, \frac{\lambda_{i2}}{\lambda_{i1}}, \ldots, \frac{\lambda_{ip}}{\lambda_{i1}}\right) \tag{21}$$

Another requirement is $|\mathbf{\Lambda}_i| = 1$ for which $\lambda_i = |\mathbf{\Sigma}_i|^{1/p}$ and

$$\Lambda_i = \mathrm{diag}\left(\frac{\lambda_{i1}}{\lambda_i}, \ldots, \frac{\lambda_{ip}}{\lambda_i}\right)$$

The parameter λ_i controls the volume of the cluster corresponding to the ith component, Λ_i its shape and $\boldsymbol{A}_i$ its orientation. A reduction in the number of parameters is achieved by imposing various constraints on $\boldsymbol{A}_i$, Λ_i and λ_i. For example, the constraint $\boldsymbol{A}_i = \boldsymbol{A} (i = 1, \ldots, g)$ imposes the same orientation on the g clusters. Applications of mixture models under the model (20) for the component covariance matrices have been considered by Bensmail *et al.*[27] and Fraley and Raftery,[12] among others.

A common approach to reducing the number of dimensions is to perform a principal component analysis (PCA) and then to perform the cluster analysis on the basis of the first few leading principal components. But as is well known, projections of the feature data $\boldsymbol{y}_j$ onto the first few principal axes are not always useful in portraying the group structure [refer to McLachlan and Peel[1] (p. 239) for for an illustrative example of this].

7 Mixtures of factor analysers

One approach for reducing the number of unknown parameters in the forms for the component covariance matrices $\mathbf{\Sigma}_i$ is to adopt the mixtures of factor analysers model, as considered by McLachlan and Peel.[1,28,29] This model was originally proposed

by Ghahramani and Hinton[30] for the purposes of visualizing high dimensional data in a lower dimensional space to explore for group structure; refer to Tipping and Bishop[31] who considered the related model of mixtures of principal component analysers for the same purpose. With the mixture of factor analysers model, the ith component covariance matrix $\mathbf{\Sigma}_i$ has the form

$$\mathbf{\Sigma}_i = \boldsymbol{B}_i \boldsymbol{B}_i^{\mathrm{T}} + \boldsymbol{D}_i, \quad i = 1, \ldots, g \tag{22}$$

where $\boldsymbol{B}_i$ is a $p \times q$ matrix of factor loadings and $\boldsymbol{D}_i$ is a diagonal matrix. It assumes that the component correlations between the observations can be explained by the conditional linear dependence of the latter on q latent or unobservable variables specific to the given component. Unlike the PCA model, the factor analysis model (22) enjoys a powerful invariance property: changes in the scales of the feature variables in $\boldsymbol{y}_j$ appear only as scale changes in the appropriate rows of the matrix $\boldsymbol{B}_i$ of factor loadings.

If the number of factors q is chosen sufficiently smaller than p, the representation (22) imposes some constraints on the component covariance matrix $\mathbf{\Sigma}_i$ and thus reduces the number of free parameters to be estimated. Note that in the case of $q > 1$, there is an infinity of choices for $\boldsymbol{B}_i$, since Equation (22) is still satisfied if $\boldsymbol{B}_i$ is replaced by $\boldsymbol{B}_i \boldsymbol{C}_i$, where $\boldsymbol{C}_i$ is any orthogonal matrix of order q. One (arbitrary) way of uniquely specifying $\boldsymbol{B}_i$ is to choose the orthogonal matrix $\boldsymbol{C}_i$ so that $\boldsymbol{B}_i^{\mathrm{T}} \boldsymbol{D}_i^{-1} \boldsymbol{B}_i$ is diagonal (with its diagonal elements arranged in decreasing order). Assuming that the eigenvalues of $\boldsymbol{B}_i \boldsymbol{B}_i^{\mathrm{T}}$ are positive and distinct, the condition that $\boldsymbol{B}_i^{\mathrm{T}} \boldsymbol{D}_i^{-1} \boldsymbol{B}_i$ is diagonal as mentioned above imposes $(1/2)q(q-1)$ constraints on the parameters. Hence then the number of free parameters for each component covariance matrix is

$$pq + p - (1/2)q(q-1)$$

With the factor analysis model, we avoid to compute the inverses of iterates of the estimated $p \times p$ covariance matrix $\mathbf{\Sigma}_i$ that may be singular for large p relative to n. This is because the inversion of the current value of the $p \times p$ matrix $(\boldsymbol{B}_i \boldsymbol{B}_i^{\mathrm{T}} + \boldsymbol{D}_i)$ on each iteration can be undertaken using the result that

$$(\boldsymbol{B}_i \boldsymbol{B}_i^{\mathrm{T}} + \boldsymbol{D}_i)^{-1} = \boldsymbol{D}_i^{-1} - \boldsymbol{D}_i^{-1} \boldsymbol{B}_i (\mathbf{I}_q + \boldsymbol{B}_i^{\mathrm{T}} \boldsymbol{D}_i^{-1} \boldsymbol{B}_i)^{-1} \boldsymbol{B}_i^{\mathrm{T}} \boldsymbol{D}_i^{-1} \tag{23}$$

where the right hand side of Equation (23) involves only the inverses of $q \times q$ matrices, since $\boldsymbol{D}_i$ is a diagonal matrix. The determinant of $(\boldsymbol{B}_i \boldsymbol{B}_i^{\mathrm{T}} + \boldsymbol{D}_i)$ can then be calculated as

$$|\boldsymbol{B}_i \boldsymbol{B}_i^{\mathrm{T}} + \boldsymbol{D}_i| = \frac{|\boldsymbol{D}_i|}{|\mathbf{I}_q - \boldsymbol{B}_i^{\mathrm{T}} (\boldsymbol{B}_i \boldsymbol{B}_i^{\mathrm{T}} + \boldsymbol{D}_i)^{-1} \boldsymbol{B}_i|}$$

Direct differentiation of the log likelihood function shows that the ML estimate of the diagonal matrix $\boldsymbol{D}_i$ satisfies

$$\hat{\boldsymbol{D}}_i = \mathrm{diag}(\hat{\boldsymbol{V}}_i - \hat{\boldsymbol{B}}_i \hat{\boldsymbol{B}}_i^{\mathrm{T}}) \tag{24}$$

where

$$\hat{\boldsymbol{V}}_i = \frac{\sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j; \hat{\boldsymbol{\Psi}})(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_i)^{\mathrm{T}}}{\sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j; \hat{\boldsymbol{\Psi}})} \tag{25}$$

It can be seen from Equation (24) that some of the estimates of the elements of the diagonal matrix $\boldsymbol{D}_i$ (the uniqueness) will be close to zero if effectively not more than q observations are unequivocally assigned to the ith component of the mixture in terms of the fitted posterior probabilities of component membership. This will lead to spikes or near singularities in the likelihood. One way to avoid this is to impose the condition of a common value $\boldsymbol{D}$ for the $\boldsymbol{D}_i$,

$$\boldsymbol{D}_i = \boldsymbol{D}, \quad i = 1, \ldots, g \tag{26}$$

The mixture of probabilistic component analysers (PCAs) model, as proposed by Tipping and Bishop,[31] has the form (22) with each $\boldsymbol{D}_i$ now having the isotropic structure

$$\boldsymbol{D}_i = \sigma_i^2 \mathbf{I}_p, \quad i = 1, \ldots, g \tag{27}$$

Under this isotropic restriction (27) the iterative updating of $\boldsymbol{B}_i$ and $\boldsymbol{D}_i$ is not necessary since given the component membership of the mixture of PCAs, $\boldsymbol{B}_i^{(k+1)}$ and $\sigma_i^{(k+1)^2}$ are given explicitly by an eigenvalue decomposition of the current value of $\boldsymbol{V}_i$.

The mixtures of factor analysers model can be fitted by using the alternating expectation conditional maximization algorithm.[32] We can make use of the link of factor analysis with the probabilistic PCA model (27) to specify an initial value $\Psi^{(0)}$ for Ψ.[29]

8 Mixed feature data

We consider the case where some of the feature variables are discrete. That is, the observation vector $\boldsymbol{y}_j$ on the jth entity to be clustered consists of m discrete variables in the vector $\boldsymbol{y}_{1j}$ in addition to p continuous variables now represented by the vector $\boldsymbol{y}_{2j} (j = 1, \ldots, n)$. The ith component density of the jth observation

$$\boldsymbol{y}_j = (\boldsymbol{y}_{1j}^{\mathrm{T}}, \boldsymbol{y}_{2j}^{\mathrm{T}})^{\mathrm{T}}$$

can then be written as

$$f_i(y_j) = f_i(\boldsymbol{y}_{1j}) f_i(\boldsymbol{y}_{2j} \mid \boldsymbol{y}_{1j}) \tag{28}$$

The symbol f_i is being used generically here to denote a density where, for discrete random variables, the density is really a probability function.

In discriminant and cluster analyses, it has been found that it is reasonable to proceed by treating the discrete variables as if they are independently distributed within a class or cluster. This is known as the NAIVE assumption.[33,34] Under this assumption, the ith component conditional density of the vector $\boldsymbol{y}_{2j}$ of discrete features is given by

$$f_i(\boldsymbol{y}_{1j}) = \prod_{v=1}^{m} f_{iv}(y_{1vj}) \tag{29}$$

where $f_{iv}(y_{1vj})$ denotes the ith component conditional density of the vth discrete feature variable y_{1vj} in $\boldsymbol{y}_{1j}$.

If y_{1v} denotes one of the distinct values taken on by the discrete variable y_{1vj}, then under Equation (29) the $(k+1)$th update of $f_{iv}(y_{1v})$ is

$$f_{iv}^{(k+1)}(y_{1v}) = \frac{\sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j; \Psi^{(k)})\, \delta[y_{1vj}, y_{1v}] + c_1}{\sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j; \Psi^{(k)}) + c_2} \tag{30}$$

where $\delta[y_{1vj}, y_{1v}] = 1$ if $y_{1vj} = y_{1v}$ and is zero otherwise, and $\Psi^{(k)}$ is the current estimate of the vector of all the unknown parameters that now include the probabilities for the discrete variables. In Equation (30), the constants c_1 and c_2, which are both equal to zero for the ML estimate, can be chosen to limit the effect of zero estimates of $f_{iv}(y_{1v})$ for rare values y_{1v}. One choice is $c_2 = 1$ and $c_1 = 1/d_v$, where d_v is the number of distinct values in the support of y_{1vj}.[33]

We can allow for some dependence between the vector $\boldsymbol{y}_{2j}$ of continuous variables and the discrete data vector $\boldsymbol{y}_{1j}$ by adopting the location model as, for example, in Hunt and Jorgensen.[35] With the location model, $f_i(\boldsymbol{y}_{2j}|\boldsymbol{y}_{1j})$ is taken to be multivariate normal with a mean that is allowed to be different for some or all of the different levels of $\boldsymbol{y}_{1j}$.

As an alternative to the use of the full mixture model, we may proceed conditionally on the realized values of the discrete feature vector $\boldsymbol{y}_{1j}$. This leads to the use of the conditional mixture model for the continuous feature vector $\boldsymbol{y}_{2j}$,

$$f(\boldsymbol{y}_{2j}|\boldsymbol{y}_{1j}) = \sum_{i=1}^{g} \pi_i(\boldsymbol{y}_{1j})\, f_i(\boldsymbol{y}_{2j} \mid \boldsymbol{y}_{1j}) \tag{31}$$

where $\pi_i(\boldsymbol{y}_{1j})$ denotes the conditional probability of ith component membership of the mixture given the discrete data in $\boldsymbol{y}_{1j}$. A common model for $\pi_i(\boldsymbol{y}_{1j})$ is the logistic model under which

$$\pi_i(\boldsymbol{y}_{1j}) = \frac{\exp(\beta_{i0} + \boldsymbol{\beta}_i^{\mathrm{T}}\boldsymbol{y}_{1j})}{1 + \sum_{h=1}^{g-1} \exp(\beta_{h0} + \boldsymbol{\beta}_h^{\mathrm{T}}\boldsymbol{y}_{1j})} \tag{32}$$

Table 1 Clusters and outcomes for treated and untreated patients

Patient group		Outcome			
		Alive	Prostate death	Cardio death	Other death
Untreated patients					
Cluster 1	Stage 3	39	17	37	32
	Stage 4	3	4	3	2
Cluster 2	Stage 3	1	5	2	4
	Stage 4	14	49	18	7
Treated patients					
Cluster 1	Stage 3	50	4	53	20
	Stage 4	4	0	2	6
Cluster 2	Stage 3	1	5	1	9
	Stage 4	25	37	23	11

where $\boldsymbol{\beta}_i = (\beta_{i1}, , \beta_{ip_1})^{\mathrm{T}}$ for $i = 1, \ldots, g-1$, and

$$\pi_g(\boldsymbol{y}_{1j}) = 1 - \sum_{h=1}^{g-1} \pi_h(\boldsymbol{y}_{1j})$$

9 Example

We consider the clustering of patients on the basis of pretrial covariates alone for the prostate cancer clinical trial data of Byar and Green.[36] The data are analysed in the form used by Hunt and Jorgensen.[35] There are $n = 475$ patients on which 12 pretrial variates are measured. Eight are taken to be continuous (age, weight index, systolic and diastolic blood pressure, serum haemoglobin, size of primary tumour, index of tumour stage histolic grade, serum prostatic acid phosphatase) and four to be discrete (performance rating, cardiovascular disease history, electrocardiogram code and bone metastases). The number of levels of these latter four categorical variates as analysed was 3, 2, 7 and 2, respectively.

We fitted a two component mixture model to these mixed feature data to compare the two clusters obtained with the clinical stage (3 or 4) and trial outcomes of the patients, who are stratified according to their treatment status ('treated' or 'untreated'). In Table 1, we present the results for the NAIVE model for the component probability functions for the discrete variates and the normal model (with unrestricted covariance matrices) for the component densities for the continuous variates, which are taken to be independent of the discrete features. We also relaxed the latter assumption by adopting a location model for the continuous variates with respect to the discrete feature variate of bone metastases. But it led to very similar results, as did other variants of the mixture model that involved using the logistic model (32).

It can be seen in Table 1 that cluster 1 membership and clinical Stage 3 status are associated with a better chance of survival. The patterns of outcomes for the 42 patients whose model and clinical classifications conflict suggest that the mixture model based classifications are better indicators of prognosis than the clinical criteria used.

References

1 McLachlan GJ, Peel D. *Finite mixture models*. New York: Wiley, 2000.
2 McLachlan GJ, Basford, KE. *Mixture models: inference and applications to clustering*. New York: Marcel Dekker, 1988.
3 Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B* 1977; **39**: 1–38.
4 McLachlan GJ, Krishnan T. *The EM algorithm and extensions*. New York: Wiley, 1997.
5 Hawkins DM, Muller MW, ten Krooden JA. Cluster analysis. In Hawkins DM, ed. *Topics in applied multivariate analysis*. Cambridge: Cambridge University Press, 1982; 303–56.
6 Aitkin M, Anderson D, Hinde J. Statistical modelling of data on teaching styles (with discussion). *Journal of the Royal Statistical Society B* 1981; **144**: 419–61.
7 Marriott FHC. *The interpretation of multiple observations*. London: Academic Press, 1974.
8 Ng, SK, McLachlan, GJ. On the choice of the number of blocks with the incremental EM algorithm for the fitting of normal mixtures. *Statistics and Computing* 2003; **13**: 45–55.
9 Coleman D, Dong X, Hardin J, Rocke DM, Woodruff DL. Some computational issues in cluster analysis with no a priori metric. *Computational Statistics and Data Analysis* 1999; **31**: 1–11.
10 Biernacki C, Celeux G, Govaert G. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics and Data Analysis* 2003; **41**: 561–75.
11 McLachlan GJ, Peel D, Basford KE, Adams P. The EMMIX software for the fitting of mixtures of normal and *t*-components. *Journal of Statistical Software* 1999; **4** (No. 2).
12 Fraley C, Raftery AE. How many clusters? Which clustering method? Answers via model-based cluster analysis. *Computer Journal* 1998; **41**: 578–88.
13 Hartley HO, Rao JNK. Classification and estimation in analysis of variance problems. *International Statistical Review* 1968; **36**: 141–47.
14 John S. On identifying the population of origin of each observation in a mixture of observations from two normal populations. *Technometrics* 1970; **12**: 553–63.
15 Scott AJ, Symons MJ. Clustering methods based on likelihood ratio criteria. *Biometrics* 1971; **27**: 387–97.
16 McLachlan GJ. The classification and mixture maximum likelihood approaches to cluster analysis. In Krishnaiah PR, Kanal L, eds. *Handbook of statistics*, Volume 2. Amsterdam: North-Holland, 1982; 199–208.
17 Bryant PG. Large-sample results for optimization-based clustering. *Journal of Classification* 1991; **8**: 31–44.
18 Edwards AWF, Cavalli-Sforza LL. A method for cluster analysis. *Biometrics* 1965; **21**: 362–75.
19 Friedman HP, Rubin H. On some invariant criteria for grouping data. *Journal of the American Statistical Association* 1967; **62**: 1159–78.
20 McLachlan GJ. On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics* 1987; **36**: 318–24.
21 Schwarz G. Estimating the dimension of a model. *Annals of Statistics* 1978; **6**: 461–64.
22 Biernacki C, Celeux G, Govaert G. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2000; **22**: 719–25.
23 McLachlan GJ, Peel D. Robust cluster analysis via mixtures of multivariate *t*-distributions. In Amin A, Dori D, Pudil P, Freeman H, eds. *Lecture notes in computer science*, Volume 1451. Berlin: Springer-Verlag, 1998; 658–66.
24 Peel D. and McLachlan. GJ. Robust mixture modelling using the *t*-distribution. *Statistics and Computing* 2000; **10**: 335–44.
25 Hennig C. Breakdown points of maximum likelihood-estimators of location-scale mixtures. *Annals of Statistics* 2002; in press.
26 Banfield JD, Raftery AE. Model-based Gaussian and non-Gaussian clustering. *Biometrics* 1993; **49**: 803–21.
27 Bensmail H, Celeux G, Raftery AE, Robert CP. Inference in model-based cluster analysis. *Statistics and Computing* 1997; **7**: 1–10.
28 McLachlan GJ, Peel, D. Mixtures of factor analyzers, In Langley P, ed. *Proceedings of the Seventeenth International Conference on Machine Learning*. San Francisco: Morgan Kaufmann, 2000; 599–606.

29 McLachlan GJ, Peel D, Bean RW. Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics and Data Analysis* 2003; **41**: 379–88.
30 Ghahramani Z, Hinton GE. The EM algorithm for factor analyzers. Technical Report No. CRG-TR-96-1, Toronto: The University of Toronto, 1997.
31 Tipping ME, Bishop CM. Mixtures of probabilistic principal component analysers. *Neural Computation* 1999; **11**: 443–82.
32 Meng XL, van Dyk D. The EM algorithm – an old folk song sung to a fast new tune (with discussion). *Journal of the Royal Statistical Society B* 1997; **59**: 511–67.
33 Titterington DM, Murray GD, Murray LS, Spiegelhalter DJ, Skene AM, Habbema JDF, Gelpke, GJ. Comparison of discrimination techniques applied to a complex data set of head injured patients (with discussion). *Journal of the Royal Statistical Society A* 1981; **144**: 145–75.
34 Hand DJ, Yi K. Idiot's Bayes – not so stupid after all? *International Statistical Review* 2001; **69**: 385–98.
35 Hunt LA, Jorgensen MA. Mixture model clustering: a brief introduction to the MULTIMIX program. *Australian and New Zealand Journal of Statistics* 1999; **40**: 153–71.
36 Byar DP, Green SB. The choice of treatment for cancer patients based on covariate information: application to prostate cancer. *Bulletin du Cancer (Paris)* 1980; **67**: 477–90.