Model-Based Clustering

G.J. McLachlan [1],

Department of Mathematics and the Institute for Molecular Bioscience, University of Queensland, St. Lucia, Brisbane 4072, Australia

Abstract

Finite mixture models are being commonly used in a wide range of applications in practice concerning density estimation and clustering. An attractive feature of this approach to clustering is that it provides a sound statistical framework in which to assess the important question of how many clusters there are in the data and their validity.

Key words: Finite mixture modelling; Maximum likelihood; EM algorithm; Normal components; Multivariate t-distribution; Factor analyzers; Choice of number of components

1 Introduction

Clustering procedures based on finite mixture models are being increasingly preferred over heuristic methods due to their sound mathematical basis and to the interpretability of their results. Mixture model-based procedures provide a probabilistic clustering that allows for overlapping clusters corresponding to the components of the mixture model. The uncertainties that the observations belong to the clusters are provided in terms of the fitted values for their posterior probabilities of component membership of the mixture. As each component in a finite mixture model corresponds to a cluster, the problem of choosing an appropriate clustering method can be recast as statistical model choice. It also allows the important question of how many clusters there are in the data to be approached through an assessment of how many components are needed in the mixture model. These questions of model choice can be considered in terms of the likelihood function.

Email address: gjm@maths.uq.edu.au (G.J. McLachlan).

[1] Phone: +61 7 3365 2150, Fax +61 7 3365 1477

Scott and Symons[1] were one of the first to adopt a model-based approach to clustering. Assuming that the data were normally distributed within a cluster, they showed that their approach is equivalent to some commonly used clustering criteria with various constraints on the cluster covariance matrices. However, from an estimation point of view, this approach yields inconsistent estimators of the parameters; see, for example, Bryant and Williamson[2] and McLachlan[3].

This inconsistency can be avoided by working with the mixture likelihood formed under the assumption that the observed data are from a mxture of classes corresponding to the clusters to be imposed on the data, as proposed by Wolfe[4] and Day[5]. Finite mixture models have since been increasingly used to model the distributions of a wide variety of random phenomena and to cluster data sets; see, for example, the recent books by Böhning[6], McLachlan and Peel[7] and Frühwirth-Schnatter[8], and the references therein. Earlier references on mixture models may be found in the previous books by Everitt and Hand[9], Titterington et al.[10], McLachlan and Basford[11], and Lindsay[12].

2 Definition of Mixture Models

We let $\boldsymbol{Y}$ denote a random vector consisting of p feature variables associated with the random phenomenon of interest. We let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ denote an observed random sample of size n on $\boldsymbol{Y}$. With the finite mixture model-based approach to density estimation and clustering, the density of $\boldsymbol{Y}$ is modelled as a mixture of a number (g) of component densities $f_i(\boldsymbol{y})$ in some unknown proportions $\pi_1, \ldots, \pi_g$. That is, each data point is taken to be a realization of the mixture probability density function (p.d.f.),

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

where the mixing proportions π_i are nonnegative and sum to one. In density estimation, the number of components g can be taken sufficiently large for (1) to provide an arbitrarily accurate estimate of the underlying density function; see, for example, Li and Barron[13]. For clustering purposes, each component in the mixture model (1) corresponds to a cluster. The posterior probability that an observation with feature vector $\boldsymbol{y}_j$ belongs to the ith component of the mixture is given by

$$\tau_i(\boldsymbol{y}_j) = \pi_i f_i(\boldsymbol{y}_j)/f(\boldsymbol{y}_j) \tag{2}$$

for $i = 1, \ldots, g$. 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 partitioning of the observations into g nonoverlapping clusters $C_1, \ldots, C_g$ is effected by assigning each observation to the component to which it has the highest estimated posterior probability of belonging. Thus the ith cluster C_i contains those observations assigned to group G_i. That is, C_i contains those observations j with $\hat{z}_{ij} = (\hat{\boldsymbol{z}}_j)_i = 1$, where

$$\begin{aligned} \hat{z}_{ij} &= 1, \qquad \text{if } \hat{\tau}_i(\boldsymbol{y}_j) \geq \hat{\tau}_h(\boldsymbol{y}_j), \qquad (h = 1, \ldots, g; h \neq i), \\ &= 0, \qquad \text{otherwise}, \end{aligned} \tag{3}$$

where $\hat{\tau}_i(\boldsymbol{y}_j)$ is an estimate of $\tau_i(\boldsymbol{y}_j)$. As the notation implies, $\hat{z}_{ij}$ can be viewed as an estimate of z_{ij} which, under the assumption that the observations come from a mixture of g groups $G_1, \ldots, G_g$, is defined to be one or zero according as the jth observation does or does not come from G_i $(i = 1, \ldots, g; j = 1, \ldots, n)$.

3 Maximum Likelihood Estimation

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

$$f(\boldsymbol{y}_j; \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i \, f_i(\boldsymbol{y}_j; \boldsymbol{\theta}_i) \tag{4}$$

by maximum likelihood (ML). Here $\boldsymbol{\Psi} = (\boldsymbol{\omega}^T, \pi_1, \ldots, \pi_{g-1})^T$ is the vector of unknown parameters, where $\boldsymbol{\omega}$ consists of the elements of the $\boldsymbol{\theta}_i$ known *a priori* to be distinct. In order to estimate $\boldsymbol{\Psi}$ from the observed data, it must be identifiable. This will be so if the representation (4) is unique up to a permutation of the component labels. The maximum likelihood estimate (MLE) of $\boldsymbol{\Psi}$, $\hat{\boldsymbol{\Psi}}$, is given by an appropriate root of the likelihood equation,

$$\partial \log L(\boldsymbol{\Psi}) / \partial \boldsymbol{\Psi} = \mathbf{0}, \tag{5}$$

where $L(\boldsymbol{\Psi})$ denotes the likelihood function for $\boldsymbol{\Psi}$,

$$L(\boldsymbol{\Psi}) = \prod_{j=1}^{n} f(\boldsymbol{y}_j; \boldsymbol{\Psi}).$$

Solutions of (5) corresponding to local maximizers of $\log L(\mathbf{\Psi})$ can be obtained via the expectation-maximization (EM) algorithm of Dempster, Laird, and Rubin[14]; see also McLachlan and Krishnan[15]. Let $\hat{\mathbf{\Psi}}$ denote the estimate of $\mathbf{\Psi}$ so obtained.

4 Fitting Mixture Models Via the EM Algorithm

We consider now the ML fitting of the mixture model (4) via the EM algorithm. It is straightforward, at least in principle, to find solutions of (5) using the EM algorithm. It is easy to program for this problem and proceeds iteratively in two steps, E (for expectation) and M (for maximization).

For the purpose of the application of the EM algorithm, the observed data are regarded as being incomplete. The complete data are taken to be the observed feature vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, along with their component-indicator vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$, which are unobservable in the framework of the mixture model being fitted. Consistent with the notation introduced in the last section, the ith element z_{ij} of $\boldsymbol{z}_j$ is defined to be one or zero, according as the jth with feature vector $\boldsymbol{y}_j$ does or does not come from the ith component of the mixture, that is, from group G_i $(i = 1, \ldots, g; j = 1, \ldots, n)$. The data are thus conceptualized to have come from g groups $G_1, \ldots, G_g$, irrespective of whether these groups do externally exist.

For this specification, the complete-data log likelihood is

$$\log L_c(\mathbf{\Psi}) = \sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij} \log \pi_i + \sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij} \log f_i(\boldsymbol{y}_j; \boldsymbol{\theta}_i). \tag{6}$$

4.1 E-Step

The addition of the unobservable data to the problem (here the $\boldsymbol{z}_j$) is handled by the E-step, which takes the conditional expectation of the complete-data log likelihood, $\log L_{\mathrm{c}}(\mathbf{\Psi})$, given the observed data

$$\boldsymbol{y}_{\mathrm{obs}} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T)^T,$$

using the current fit for $\mathbf{\Psi}$. Let $\mathbf{\Psi}^{(0)}$ be the value specified initially for $\mathbf{\Psi}$. Then on the first iteration of the EM algorithm, the E-step requires the computation of the conditional expectation of $\log L_c(\mathbf{\Psi})$ given $\boldsymbol{y}$, using $\mathbf{\Psi}^{(0)}$ for $\mathbf{\Psi}$, which can be written as

$$Q(\mathbf{\Psi}; \mathbf{\Psi}^{(0)}) = E_{\mathbf{\Psi}^{(0)}}\{\log L_{\mathrm{c}}(\mathbf{\Psi}) \mid \boldsymbol{y}_{\mathrm{obs}}\}. \tag{7}$$

The expectation operator E has the subscript $\boldsymbol{\Psi}^{(0)}$ to explicitly convey that this expectation is being effected using $\boldsymbol{\Psi}^{(0)}$ for $\boldsymbol{\Psi}$.

It follows that on the $(k+1)$th iteration, the E-step requires the calculation of $Q(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k)})$, where $\boldsymbol{\Psi}^{(k)}$ is the value of $\boldsymbol{\Psi}$ after the kth EM iteration. As the complete-data log likelihood, $\log L_{\rm c}(\boldsymbol{\Psi})$, 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 feature observation $\boldsymbol{y}_j$, where Z_{ij} is the random variable corresponding to z_{ij}. Now

$$\begin{aligned} E_{\boldsymbol{\Psi}^{(k)}}(Z_{ij} \mid \boldsymbol{y}_j) &= \mathrm{pr}_{\boldsymbol{\Psi}^{(k)}}\{Z_{ij} = 1 \mid \boldsymbol{y}_j\} \\ &= \tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k)}), \end{aligned} \tag{8}$$

where, corresponding to (2),

$$\begin{aligned} \tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k)}) &= \pi_i^{(k)} f_i(\boldsymbol{y}_j;\, \boldsymbol{\theta}_i^{(k)}) / f(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k)}) \\ &= \pi_i^{(k)} f_i(\boldsymbol{y}_j;\, \boldsymbol{\theta}_i^{(k)}) / \sum_{h=1}^{g} \pi_h^{(k)} f_h(\boldsymbol{y}_j;\, \boldsymbol{\theta}_h^{(k)}) \end{aligned} \tag{9}$$

for $i = 1, \ldots, g;\ j = 1, \ldots, n$. The quantity $\tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k)})$ is the posterior probability that the jth member of the sample with observed value $\boldsymbol{y}_j$ belongs to the ith component of the mixture. Using (8), we have on taking the conditional expectation of (6) given $\boldsymbol{y}_{\rm obs}$ that

$$Q(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k)}) = \sum_{i=1}^{g} \sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k)}) \{\log \pi_i + \log f_i(\boldsymbol{y}_j;\, \boldsymbol{\theta}_i)\}. \tag{10}$$

4.2 M-Step

The M-step on the $(k+1)$th iteration requires the global maximization of $Q(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k)})$ with respect to $\boldsymbol{\Psi}$ over the parameter space $\boldsymbol{\Omega}$ to give the updated estimate $\boldsymbol{\Psi}^{(k+1)}$. For the finite mixture model, the updated estimates $\pi_i^{(k+1)}$ of the mixing proportions π_i are calculated independently of the updated estimate $\boldsymbol{\omega}^{(k+1)}$ of the parameter vector $\boldsymbol{\omega}$ containing the unknown parameters in the component densities.

If the z_{ij} were observable, then the complete-data MLE of π_i would be given simply by

$$\hat{\pi}_i = \sum_{j=1}^{n} z_{ij}/n \qquad (i = 1, \ldots, g). \tag{11}$$

As the E-step simply involves replacing each z_{ij} with its current conditional expectation $\tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k)})$ in the complete-data log likelihood, the updated estimate of π_i is given by replacing each z_{ij} in (11) by $\tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k)})$ to give

$$\pi_i^{(k+1)} = \sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k)})/n \qquad (i = 1, \ldots, g). \tag{12}$$

Thus in forming the estimate of π_i on the $(k+1)$th iteration, there is a contribution from each observation $\boldsymbol{y}_j$ equal to its (currently assessed) posterior probability of membership of the ith component of the mixture model.

Concerning the updating of $\boldsymbol{\omega}$ on the M-step of the $(k+1)$th iteration, it can be seen from (10) that $\boldsymbol{\omega}^{(k+1)}$ is obtained as an appropriate root of

$$\sum_{i=1}^{g} \sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k)}) \partial \log f_i(\boldsymbol{y}_j;\, \boldsymbol{\theta}_i)/\partial \boldsymbol{\omega} = \mathbf{0}. \tag{13}$$

One nice feature of the EM algorithm is that the solution of (13) often exists in closed form, as is to be demonstrated for the normal mixture model in Section 6.

The E- and M-steps are alternated repeatedly until the difference

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

changes by an arbitrarily small amount in the case of convergence of the sequence of likelihood values $\{L(\boldsymbol{\Psi}^{(k)})\}$. Dempster et al.[14] showed that the (incomplete-data) likelihood function $L(\boldsymbol{\Psi})$ is not decreased after an EM iteration; that is,

$$L(\boldsymbol{\Psi}^{(k+1)}) \geq L(\boldsymbol{\Psi}^{(k)}) \tag{14}$$

for $k = 0, 1, 2, \ldots$. Hence, convergence must be obtained with a sequence of likelihood values $\{L(\boldsymbol{\Psi}^{(k)})\}$ that are bounded above. In almost all cases, the limiting value L^* is a local maximum. In any event, if an EM sequence $\{\boldsymbol{\Psi}^{(k)}\}$ is trapped at some stationary point $\boldsymbol{\Psi}^*$ that is not a local or global maximizer of $L(\boldsymbol{\Psi})$ (for example, a saddle point), a small random perturbation of $\boldsymbol{\Psi}$ away from the saddle point $\boldsymbol{\Psi}^*$ will cause the EM algorithm to diverge from the saddle point. Further details may be found in McLachlan and Krishnan[15] (Chapter 3).

Let $\hat{\boldsymbol{\Psi}}$ be the chosen solution of the likelihood equation. For an observed sample, $\hat{\boldsymbol{\Psi}}$ is usually taken to be the root of (5) corresponding to the largest of the local maxima located. That is, in those cases where $L(\boldsymbol{\Psi})$ has a global

maximum in the interior of the parameter space, $\hat{\mathbf{\Psi}}$ is the global maximizer, assuming that the global maximum has been located.

5 Choice of Starting Values for the EM Algorithm

McLachlan and Peel[7] provide an in-depth account of the fitting of finite mixture models. Briefly, with mixture models the likelihood typically will have multiple maxima; that is, the likelihood equation will have multiple roots. Thus the EM algorithm needs to be started from a variety of initial values for the parameter vector $\mathbf{\Psi}$ 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. Nonrandom partitions of the data can be obtained via some clustering procedure such as k-means. Also, Coleman et al.[16] have proposed some procedures for obtaining nonrandom starting partitions.

The choice of root of the likelihood equation in the case of homoscedastic normal 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 normal components as the likelihood function is unbounded. It is known that as the sample size goes to infinity, there exists a sequence of roots of the likelihood equation that is consistent and asymptotically efficient. With probability tending to one, these roots correspond to local maxima in the interior of the parameter space; see McLachlan and Peel[7]. Usually, the intent is to choose as the ML estimate of the parameter vector $\mathbf{\Psi}$ the local maximizer corresponding to the largest of the local maxima located. But 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 nonzero) 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.

6 Clustering Via Normal Mixtures

Frequently, in practice, the clusters in the data are essentially elliptical, so that it is reasonable to consider fitting mixtures of elliptically symmetric component densities. Within this class of component densities, the multivariate normal density is a convenient choice given its computational tractability.

6.1 Heteroscedastic Components

Under the assumption of multivariate normal components, the ith component-conditional density $f_i(\boldsymbol{y};\, \boldsymbol{\theta}_i)$ is given by

$$f_i(\boldsymbol{y};\, \boldsymbol{\theta}_i) = \phi(\boldsymbol{y};\, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{15}$$

where $\boldsymbol{\theta}_i$ consists of the elements of $\boldsymbol{\mu}_i$ and the $\frac{1}{2}p(p+1)$ distinct elements of $\boldsymbol{\Sigma}_i$ $(i = 1, \ldots, g)$. Here

$$\phi(\boldsymbol{y};\, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}_i|^{-1/2} \exp\{-\tfrac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{y} - \boldsymbol{\mu}_i)\}. \tag{16}$$

It follows that on the M-step of the $(k+1)$th iteration, the updates of the component means $\boldsymbol{\mu}_i$ and component-covariance matrices $\boldsymbol{\Sigma}_i$ are given explicitly by

$$\boldsymbol{\mu}_i^{(k+1)} = \sum_{j=1}^{n} \tau_{ij}^{(k)} \boldsymbol{y}_j \Big/ \sum_{j=1}^{n} \tau_{ij}^{(k)} \tag{17}$$

and

$$\boldsymbol{\Sigma}_i^{(k+1)} = \sum_{j=1}^{n} \tau_{ij}^{(k)} (\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k+1)})(\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k+1)})^T \Big/ \sum_{j=1}^{n} \tau_{ij}^{(k)} \tag{18}$$

for $i = 1, \ldots, g$, where

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

The updated estimate of the ith mixing proportion π_i is given by (12).

One attractive feature of adopting mixture models with elliptically symmetric components such as the normal or t-densities, is that the implied clustering

is invariant under affine transformations of the data; that is, invariant under transformations of the feature vector $\boldsymbol{y}$ of the form,

$$\boldsymbol{y} \rightarrow \boldsymbol{C}\boldsymbol{y} + \boldsymbol{a}, \tag{19}$$

where $\boldsymbol{C}$ is a nonsingular matrix. If the clustering of a procedure is invariant under (19) for only diagonal $\boldsymbol{C}$, then it is invariant under change of measuring units but not rotations. But as commented upon by Hartigan[17], this form of invariance is more compelling than affine invariance.

6.2 Homoscedastic Components

Often in practice, the component-covariance matrices $\boldsymbol{\Sigma}_i$ are restricted to being the same,

$$\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma} \qquad (i = 1, \ldots, g), \tag{20}$$

where $\boldsymbol{\Sigma}$ is unspecified. In this case of homoscedastic normal components, the updated estimate of the common component-covariance matrix $\boldsymbol{\Sigma}$ is given by

$$\boldsymbol{\Sigma}^{(k+1)} = \sum_{i=1}^{g} \pi_i^{(k)} \boldsymbol{\Sigma}_i^{(k+1)} / n, \tag{21}$$

where $\boldsymbol{\Sigma}_i^{(k+1)}$ is given by (18), and the updates of π_i and $\boldsymbol{\mu}_i$ are as above in the heteroscedastic case.

6.3 Spherical Components

A further simplification is to take the component-covariance matrices to have a common spherical form, where the covariance matrix of each component is taken to be a (common) multiple of the $p \times p$ identity matrix $\boldsymbol{I}_p$, namely

$$\boldsymbol{\Sigma}_i = \sigma^2 \boldsymbol{I}_p \qquad (i = 1, \ldots, g). \tag{22}$$

The constraint (22) means that the clusters produced are spherical. If we also take the mixing proportions to be equal, then it is equivalent to a "soft" version of k-means clustering. It is a soft version as with k-means, the observations are assigned outright at each of the iterations.

7 Spectral Representation of Component-Covariances Matrices

It can be seen from (16) that the mixture model with unrestricted group-covariance matrices in its normal component distributions is a highly parameterized one with $\frac{1}{2}p(p+1)$ parameters for each component-covariance matrix $\boldsymbol{\Sigma}_i$ $(i = 1, \ldots, g)$. As an alternative to taking the component-covariance matrices to be the same or diagonal, we can adopt some model for the component-covariance matrices that is intermediate between homoscedasticity and the unrestricted model, as in the approach of Banfield and Raftery[18]; see also Fraley and Raftery[19].

Banfield and Raftery[18] introduced a parameterization of the component-covariance matrix $\boldsymbol{\Sigma}_i$ based on a variant of the standard spectral decomposition of $\boldsymbol{\Sigma}_i$,

$$\boldsymbol{\Sigma}_i = \sum_{v=1}^{p} \lambda_{iv} \boldsymbol{a}_{iv} \boldsymbol{a}_{iv}^T, \tag{23}$$

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 $\boldsymbol{\Sigma}_i$ $(i = 1, \ldots, g)$. They expressed $\boldsymbol{\Sigma}_i$ further as

$$\boldsymbol{\Sigma}_i = \lambda_i \boldsymbol{A}_i \boldsymbol{\Lambda}_i \boldsymbol{A}_i^T, \tag{24}$$

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

$$\boldsymbol{\Lambda}_i = \mathrm{diag}(1, \lambda_{i2}/\lambda_{i1}, \ldots, \lambda_{ip}/\lambda_{i1}). \tag{25}$$

Another requires $\mid \boldsymbol{\Lambda}_i \mid = 1$ for which $\lambda_i = \mid \boldsymbol{\Sigma}_i \mid^{1/p}$ and

$$\boldsymbol{\Lambda}_i = \mathrm{diag}(\lambda_{i1}/\lambda_i, \ldots, \lambda_{ip}/\lambda_i).$$

The parameter λ_i controls the volume of the cluster corresponding to the ith component, $\boldsymbol{\Lambda}_i$ its shape, and $\boldsymbol{A}_i$ its orientation. A reduction in the number of parameters is achieved by imposing various constraints on the $\boldsymbol{A}_i, \boldsymbol{\Lambda}_i$, and the λ_i. For example, the constraint $\boldsymbol{A}_i = \boldsymbol{A}$ $(i = 1, \ldots, g)$ imposes the same orientation on the g clusters.

8 Multivariate t-distribution

The mixture model with normal components (16) is sensitive to outliers since it adopts the multivariate normal family for the distributions of the errors. An obvious way to improve the robustness of this model for data which have longer tails than the normal or atypical observations is to consider using the multivariate t-family of elliptically symmetric distributions. It has an additional parameter called the degrees of freedom that controls the length of the tails of the distribution. Although the number of outliers needed for breakdown is almost the same as with the normal distribution, the outliers have to be much larger. This point is made more precise in Hennig[20] who has provided an excellent account of breakdown points for ML estimation of location-scale mixtures with a fixed number of components g.

The t-distribution for the ith component-conditional distribution of $\boldsymbol{Y}_j$ is obtained by embedding the normal $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ distribution in a wider class of elliptically symmetric distributions with an additional parameter ν_i called the degrees of freedom. This t-distribution can be characterized by letting W_j denote a random variable distributed as

$$W_j \sim \text{ gamma}\,(\tfrac{1}{2}\nu_i,\, \tfrac{1}{2}\nu_i), \tag{26}$$

where the gamma$\,(\alpha,\, \beta)$ density function is equal to

$$f_{\mathrm{G}}(w; \alpha, \beta) = \{\beta^{\alpha} w^{\alpha-1}/\Gamma(\alpha)\} \exp(-\beta w) I_{[0,\infty)}(w) \qquad (\alpha,\, \beta > 0), \tag{27}$$

and $I_A(w)$ denotes the indicator function that is 1 if w belongs to A and is zero otherwise. Then, if the conditional distribution of $\boldsymbol{Y}_j$ given $W_j = w_j$ is specified to be

$$\boldsymbol{Y}_j \mid w_j \sim N_p(\boldsymbol{\mu}_i,\, \boldsymbol{\Sigma}_i/w_j), \tag{28}$$

the unconditional distribution of $\boldsymbol{Y}_j$ has a (multivariate) t-distribution with mean $\boldsymbol{\mu}_i$, scale matrix $\boldsymbol{\Sigma}_i$, and degrees of freedom ν_i. The mean of this t-distribution is $\boldsymbol{\mu}_i$ and its covariance matrix is $\{\nu_i/(\nu_i - 2)\}\boldsymbol{\Sigma}_i$. We write

$$\boldsymbol{Y}_j \sim t_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \nu_i), \tag{29}$$

and we let $f_{\mathrm{t}}(\boldsymbol{y}_j;\, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \nu_i)$ denote the corresponding density. As ν_i tends to infinity, the t-distribution approaches the normal distribution. Hence this parameter ν_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.

9 ML Estimation of Mixtures of t distributions

McLachlan and Peel[7] and Peel and McLachlan[21] have implemented the E- and M-steps of the EM algorithm and its variant, the ECM (expectation–conditional maximization) algorithm for the ML estimation of multivariate t components. The ECM algorithm proposed by Meng and Rubin[22] replaces the M-step of the EM algorithm by a number of computationally simpler conditional maximization (CM) steps.

In the EM framework for this problem, the unobservable variable w_j in the characterization (28) of the t distribution for the ith component of the t mixture model and the component-indicator labels z_{ij} are treated as being the "missing" data.

It can be shown that the conditional expectation of W_j given $\boldsymbol{y}_j$ and $z_{ij} = 1$ can be expressed as

$$E\{W_j \mid \boldsymbol{y}_j, z_{ij} = 1\} = w_i(\boldsymbol{y}_j; \boldsymbol{\Psi}),$$

where

$$w_i(\boldsymbol{y}_j; \boldsymbol{\Psi}) = \frac{\nu_i + p}{\nu_i + \delta(\boldsymbol{y}_j, \boldsymbol{\mu}_i; \boldsymbol{\Sigma}_i)} \tag{30}$$

and where

$$\delta(\boldsymbol{y}_j, \boldsymbol{\mu}_i; \boldsymbol{\Sigma}_i) = (\boldsymbol{y}_j - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{y}_j - \boldsymbol{\mu}_i) \tag{31}$$

denotes the squared Mahalanobis distance between $\boldsymbol{y}_j$ and $\boldsymbol{\mu}_i$ ($i = 1, \ldots g$; $j = 1, \ldots, n$).

On the $(k+1)$th iteration of the EM algorithm, the updated estimates of the mixing proportion, the mean vector $\boldsymbol{\mu}_i$, and the scale matrix $\boldsymbol{\Sigma}_i$ are given by

$$\pi_i^{(k+1)} = \sum_{j=1}^{n} \tau_{ij}^{(k)} / n, \tag{32}$$

$$\boldsymbol{\mu}_i^{(k+1)} = \sum_{j=1}^{n} \tau_{ij}^{(k)} w_{ij}^{(k)} \boldsymbol{y}_j / \sum_{j=1}^{n} \tau_{ij}^{(k)} w_{ij}^{(k)} \tag{33}$$

and

$$\boldsymbol{\Sigma}_i^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_{ij}^{(k)} w_{ij}^{(k)} (\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k+1)})(\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k+1)})^T}{\sum_{j=1}^{n} \tau_{ij}^{(k)}}. \tag{34}$$

In the above,

$$\tau_{ij}^{(k)} = \frac{\pi_i^{(k)} f(\boldsymbol{y}_j;\, \boldsymbol{\mu}_i^{(k)},\, \boldsymbol{\Sigma}_i^{(k)},\, \nu_i^{(k)})}{f(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k)})} \tag{35}$$

is the posterior probability that $\boldsymbol{y}_j$ belongs to the ith component of the mixture, using the current fit $\boldsymbol{\Psi}^{(k)}$ for $\boldsymbol{\Psi}$ $(i = 1, \ldots, g;\, j = 1, \ldots, n)$. Also,

$$w_{ij}^{(k)} = \frac{\nu_i^{(k)} + p}{\nu_i^{(k)} + \delta(\boldsymbol{y}_j,\, \boldsymbol{\mu}_i^{(k)};\, \boldsymbol{\Sigma}_i^{(k)})}, \tag{36}$$

which is the current stimate of the conditional expectation of W_j given $\boldsymbol{y}_j$ and $z_{ij} = 1$.

The updated estimate $\nu_i^{(k+1)}$ of ν_i does not exist in closed form, but is given as a solution of the equation

$$\begin{aligned}\left\{ -\psi(\tfrac{1}{2}\nu_i) + \log(\tfrac{1}{2}\nu_i) + 1 + \frac{1}{n_i^{(k)}} \sum_{j=1}^{n} \tau_{ij}^{(k)} (\log w_{ij}^{(k)} - w_{ij}^{(k)}) \right. \\ \left. + \psi\left(\frac{\nu_i^{(k)} + p}{2}\right) - \log\left(\frac{\nu_i^{(k)} + p}{2}\right) \right\} = 0,\end{aligned} \tag{37}$$

where $n_i^{(k)} = \sum_{j=1}^{n} \tau_{ij}^{(k)}$ $(i = 1, \ldots, g)$ and $\psi(\cdot)$ is the Digamma function.

Following the proposal of Kent, Tyler, and Vardi[23] in the case of a single-component t distribution, we can replace the divisor $\sum_{j=1}^{n} \tau_{ij}^{(k)}$ in (34) by

$$\sum_{j=1}^{n} \tau_{ij}^{(k)} w_{ij}^{(k)},$$

which should improve the speed of convergence. It corresponds to an application of the the parameter-expanded EM (PX-EM) algorithm (Liu, Rubin, and Wu[24]).

These E- and M-steps are alternated until the changes in the estimated parameters or the log likelihood are less than some specified threshold. It can be seen that if the degrees of freedom ν_i is fixed in advance for each component, then the M-step exists in closed form. In this case where ν_i is fixed beforehand, the estimation of the component parameters is a form of M-estimation. However, an attractive feature of the use of the t distribution to model the component distributions is that the degrees of robustness as controlled by ν_i can be inferred from the data by computing its MLE.

10 Choice of the Number of Components in a Mixture Model

With a mixture model-based approach to clustering, the question of how many clusters there are can be considered in terms of the smallest number of components needed for the mixture model to be compatible with the data. The estimation of the order of a mixture model has been considered mainly by consideration of the likelihood, using two main ways. One way is based on a penalized form of the log likelihood. The other main way is based on a resampling approach.

10.1 Bayesian Information Criterion

The main Bayesian-based information criteria use an approximation to the integrated likelihood, as in the original proposal by Schwarz[25] leading to his Bayesian information criterion (BIC). Available general theoretical justifications of this approximation rely on the same regularity conditions that break down for inference on the number of components in a frequentist framework.

In the literature, the information criteria so formed are generally expressed in terms of twice the negative difference between the log likelihood and the penalty term. This difference for the Bayesian information criterion (BIC) is given by

$$-2\log L(\hat{\boldsymbol{\Psi}}) + d\log n \tag{38}$$

where d is the number of parameters in the model. The intent is to minimize the criterion (38) in model selection, including the present situation for the number of components g in a mixture model.

10.2 Resampling Approach

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 in McLachlan[26]. With this approach, 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 $\boldsymbol{\Psi}$ 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 $-2\log\lambda$, where λ is the likelihood ratio statistic, is computed for each bootstrap sample after fitting mixture models for

$g = g_0$ and g_1 to it in turn. The process is repeated independently B times, 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 $-2\log\lambda$. Other resampling approaches include that based on the Gap statistic of Tibshirani et al.[27] and the Clest method of Dudoit and Fridlyand[28].

11 Advantages of Mixture Model-Based Clustering

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, Muller, and ten Krooden[29] 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, Anderson, and Hinde[30] 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[31] 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." In the context of the analysis of gene expression data, Yeung et al.[32] commented that "in the absence of a well-grounded statistical model, it seems difficult to define what is meant by a 'good' clustering algorithm or the 'right' number of clusters."

This mixture model-based approach also provides a framework for assessing the number of clusters and for clustering data in the presence of outliers, as discussed above.

12 Factor Analysis Model for Dimension Reduction

As remarked earlier, the g-component normal mixture model with unrestricted component-covariance matrices is a highly parameterized model with $\frac{1}{2}p(p+1)$ parameters for each component-covariance matrix $\mathbf{\Sigma}_i\,(i = 1, \ldots, g)$. As discussed in Section 7, Banfield and Raftery[18] introduced a parameterization of the component-covariance matrix $\mathbf{\Sigma}_i$ based on a variant of the standard spectral decomposition of $\mathbf{\Sigma}_i\,(i = 1, \ldots, g)$. However, if p is large relative to the sample size n, it may not be possible to use this decomposition to infer an appropriate model for the component-covariance matrices. Even if it is possible, the results may not be reliable due to potential problems with near-singular estimates of the component-covariance matrices when p is large relative to n.

A common approach to reducing the number of dimensions is to perform a principal component analysis (PCA). 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; see McLachlan and Peel[7]. Another approach for reducing the number of unknown parameters in the forms for the component-covariance matrices is to adopt the mixture of factor analyzers model, as considered in McLachlan and Peel[7,33] and McLachlan, Peel, and Bean[34]. This model was originally proposed by Ghahramani and Hinton[35] and Hinton, Dayan, and Revow[36] for the purposes of visualizing high dimensional data in a lower dimensional space to explore for group structure; see also Tipping and Bishop[37]. who considered the related model of mixtures of principal component analyzers for the same purpose.

With this approach, 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 without any restrictions on the covariance matrices.

13 Mixtures of Normal Factor Analyzers

13.1 Formulation of factor analysis submodels

A global nonlinear approach to dimension reduction can be obtained by postulating a finite mixture of linear submodels for the distribution of the full observation vector $\boldsymbol{Y}_j$ given the (unobservable) factors $\boldsymbol{u}_j$. That is, we can provide a local dimensionality reduction method by assuming that the distri-

bution of the observation $\boldsymbol{Y}_j$ can be modelled as

$$\boldsymbol{Y}_j = \boldsymbol{\mu}_i + \boldsymbol{B}_i \boldsymbol{U}_{ij} + \boldsymbol{e}_{ij} \qquad \text{with prob. } \pi_i \quad (i = 1, \ldots, g) \tag{39}$$

for $j = 1, \ldots, n$, where the factors $\boldsymbol{U}_{i1}, \ldots, \boldsymbol{U}_{in}$ are distributed independently $N(\boldsymbol{0}, \boldsymbol{I}_q)$, independently of the $\boldsymbol{e}_{ij}$, which are distributed independently $N(\boldsymbol{0}, \boldsymbol{D}_i)$, where $\boldsymbol{D}_i$ is a diagonal matrix $(i = 1, \ldots, g)$.

Thus the mixture of factor analyzers model is given by

$$f(\boldsymbol{y}_j; \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i \phi(\boldsymbol{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{40}$$

where the ith component-covariance matrix $\boldsymbol{\Sigma}_i$ has the form

$$\boldsymbol{\Sigma}_i = \boldsymbol{B}_i \boldsymbol{B}_i^T + \boldsymbol{D}_i \quad (i = 1, \ldots, g) \tag{41}$$

and where $\boldsymbol{B}_i$ is a $p \times q$ matrix of factor loadings and $\boldsymbol{D}_i$ is a diagonal matrix $(i = 1, \ldots, g)$. The parameter vector $\boldsymbol{\Psi}$ now consists of the mixing proportions π_i and the elements of the $\boldsymbol{\mu}_i$, the $\boldsymbol{B}_i$, and the $\boldsymbol{D}_i$.

We can think of the use of this mixture of factor analyzers model as being purely a method of regularization, but in several applications it is possible to make a case for it being a reasonable model for the correlation structure between the variables within a cluster.

The mixture of factor analyzers model can be fitted by using the alternating expectation–conditional maximization (AECM) algorithm (Meng and van Dyk[38]). The AECM algorithm is an extension of the ECM algorithm, where the specification of the complete data is allowed to be different on each CM-step. Meng and van Dyk[38] established that monotone convergence of the sequence of likelihood values is retained with the AECM algorithm.

13.2 An AECM algorithm for mixture of factor analyzers models

To apply the AECM algorithm to the fitting of the mixture of factor analyzers model, we partition the vector of unknown parameters $\boldsymbol{\Psi}$ as $(\boldsymbol{\Psi}_1^T, \boldsymbol{\Psi}_2^T)^T$, where $\boldsymbol{\Psi}_1$ contains the mixing proportions $\pi_i \, (i = 1, \ldots, g-1)$ and the elements of the component means $\boldsymbol{\mu}_i \, (i = 1, \ldots, g)$. The subvector $\boldsymbol{\Psi}_2$ contains the elements of the $\boldsymbol{B}_i$ and the $\boldsymbol{D}_i \, (i = 1, \ldots, g)$.

We let $\boldsymbol{\Psi}^{(k)} = (\boldsymbol{\Psi}_1^{(k)^T}, \boldsymbol{\Psi}_2^{(k)^T})^T$ be the value of $\boldsymbol{\Psi}$ after the kth iteration of the AECM algorithm. For this application of the AECM algorithm, one iteration

consists of two cycles, and there is one E-step and one CM-step for each cycle. The two CM-steps correspond to the partition of $\boldsymbol{\Psi}$ into the two subvectors $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$. For the first cycle of the AECM algorithm, we specify the missing data to be just the component-indicator vectors, $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$.

13.3 E-step

In order to carry out the E-step, we need to be able to compute the conditional expectation of the sufficient statistics. To carry out this step, we need to be able to calculate the conditional expectations,

$$\boldsymbol{C}_{yui} = E\{Z_{ij}\boldsymbol{y}_j\boldsymbol{U}_{ij}^T \mid \boldsymbol{y}_j\} \tag{42}$$

and

$$\boldsymbol{C}_{uui} = E\{Z_{ij}\boldsymbol{U}_{ij}\boldsymbol{U}_{ij}^T \mid \boldsymbol{y}_j\}. \tag{43}$$

To do this, we need the result that the random vector $(\boldsymbol{Y}_j^T, \boldsymbol{U}_{ij}^T)^T$ given its membership of the ith component of the mixture (that is, $z_{ij} = 1$) has a multivariate normal distribution,

$$\begin{pmatrix} \boldsymbol{Y}_j \\ \boldsymbol{U}_{ij} \end{pmatrix} \mid z_{ij} = 1 \sim N_{p+q}(\boldsymbol{\mu}_i^*, \boldsymbol{\xi}_i) \quad (i = 1, \ldots, g), \tag{44}$$

where

$$\boldsymbol{\mu}_i^* = (\boldsymbol{\mu}_i^T, \boldsymbol{0}^T)^T \tag{45}$$

and the covariance matrix $\boldsymbol{\xi}_i$ is given by

$$\boldsymbol{\xi}_i = \begin{pmatrix} \boldsymbol{B}_i\boldsymbol{B}_i^T + \boldsymbol{D}_i & \boldsymbol{B}_i \\ \boldsymbol{B}_i^T & \boldsymbol{I}_q \end{pmatrix}. \tag{46}$$

It follows that the conditional distribution of $\boldsymbol{U}_{ij}$ given $\boldsymbol{y}_j$ and $z_{ij} = 1$ is given by

$$\boldsymbol{U}_j \mid \boldsymbol{y}_j, z_{ij} = 1 \sim N(\boldsymbol{\gamma}_i^T(\boldsymbol{y}_j - \boldsymbol{\mu}_i), \boldsymbol{\Omega}_i) \tag{47}$$

for $i = 1, \ldots, g$; $j = 1, \ldots, n$, where

$$\boldsymbol{\gamma}_i = (\boldsymbol{B}_i\boldsymbol{B}_i^T + \boldsymbol{D}_i)^{-1}\boldsymbol{B}_i. \tag{48}$$

and where

$$\boldsymbol{\Omega}_i = \boldsymbol{I}_q - \boldsymbol{\gamma}_i^T \boldsymbol{B}_i. \tag{49}$$

Using (47),

$$\boldsymbol{C}_{yui} = \tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi})\boldsymbol{\gamma}_i^T \boldsymbol{y}_j \tag{50}$$

and

$$\boldsymbol{C}_{uui} = \tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi})\{\boldsymbol{\gamma}_i^T(\boldsymbol{y}_j - \boldsymbol{\mu}_i)(\boldsymbol{y}_j - \boldsymbol{\mu}_i)^T\boldsymbol{\gamma}_i + \boldsymbol{\Omega}_i\}. \tag{51}$$

13.4 CM-steps

The first conditional CM-step leads to $\pi_i^{(k)}$ and $\boldsymbol{\mu}_i^{(k)}$ being updated to

$$\pi_i^{(k+1)} = \sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k)})/n \tag{52}$$

and

$$\boldsymbol{\mu}_i^{(k+1)} = \sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k)})\, \boldsymbol{y}_j / \sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k)}) \tag{53}$$

for $i = 1, \ldots, g$, where $\tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi})$ is the ith component-posterior probability of $\boldsymbol{y}_j$.

For the second cycle for the updating of $\boldsymbol{\Psi}_2$, we specify the missing data to be the factors $\boldsymbol{u}_{i1}, \ldots \boldsymbol{u}_{in}$, as well as the component-indicator vectors, $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$. On setting $\boldsymbol{\Psi}^{(k+1/2)}$ equal to $(\boldsymbol{\Psi}_1^{(k+1)^T}, \boldsymbol{\Psi}_2^{(k)^T})^T$, an E-step is performed to calculate $Q(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k+1/2)})$, which is the conditional expectation of the complete-data log likelihood given the observed data, using $\boldsymbol{\Psi} = \boldsymbol{\Psi}^{(k+1/2)}$. The CM-step on this second cycle is implemented by the maximization of $Q(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k+1/2)})$ over $\boldsymbol{\Psi}$ with $\boldsymbol{\Psi}_1$ set equal to $\boldsymbol{\Psi}_1^{(k+1)}$. This yields the updated estimates $\boldsymbol{B}_i^{(k+1)}$ and $\boldsymbol{D}_i^{(k+1)}$. The former is given by

$$\boldsymbol{B}_i^{(k+1)} = \boldsymbol{V}_i^{(k+1/2)}\boldsymbol{\gamma}_i^{(k)}(\boldsymbol{\gamma}_i^{(k)^T}\boldsymbol{V}_i^{(k+1/2)}\boldsymbol{\gamma}_i^{(k)} + \boldsymbol{\Omega}_i^{(k)})^{-1}, \tag{54}$$

where

$$\boldsymbol{V}_i^{(k+1/2)} = \frac{\sum_{j=1}^n \tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k+1/2)})\,(\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k+1)})(\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k+1)})^T}{\sum_{j=1}^n \tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k+1/2)})}, \tag{55}$$

$$\boldsymbol{\gamma}_i^{(k)} = (\boldsymbol{B}_i^{(k)} \boldsymbol{B}_i^{(k)^T} + \boldsymbol{D}_i^{(k)})^{-1} \boldsymbol{B}_i^{(k)}, \tag{56}$$

and

$$\boldsymbol{\Omega}_i^{(k)} = \boldsymbol{I}_q - \boldsymbol{\gamma}_i^{(k)^T} \boldsymbol{B}_i^{(k)} \tag{57}$$

for $i = 1, \ldots, g$. The updated estimate $\boldsymbol{D}_i^{(k+1)}$ is given by

$$\begin{aligned} \boldsymbol{D}_i^{(k+1)} &= \mathrm{diag}\{\boldsymbol{V}_i^{(k+1/2)} - \boldsymbol{B}_i^{(k+1)} \boldsymbol{H}_i^{(k+1/2)} \boldsymbol{B}_i^{(k+1)^T}\} \\ &= \mathrm{diag}\{\boldsymbol{V}_i^{(k+1/2)} - \boldsymbol{V}_i^{(k+1/2)} \boldsymbol{\gamma}_i^{(k)} \boldsymbol{B}_i^{(k+1)^T}\}, \end{aligned} \tag{58}$$

where

$$\begin{aligned} \boldsymbol{H}_i^{(k+1/2)} &= \frac{\sum_{j=1}^n \tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k+1/2)})\, E_i^{(k+1/2)}(\boldsymbol{U}_j \boldsymbol{U}_j^T \mid \boldsymbol{y}_j)}{\sum_{j=1}^n \tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k+1/2)})} \\ &= \boldsymbol{\gamma}_i^{(k)^T} \boldsymbol{V}_i^{(k+1/2)} \boldsymbol{\gamma}_i^{(k)} + \boldsymbol{\Omega}_i^{(k)} \end{aligned} \tag{59}$$

and $E_i^{(k+1/2)}$ denotes conditional expectation given membership of the ith component, using $\boldsymbol{\Psi}^{(k+1/2)}$ for $\boldsymbol{\Psi}$.

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

$$(\boldsymbol{B}_i \boldsymbol{B}_i^T + \boldsymbol{D}_i)^{-1} = \boldsymbol{D}_i^{-1} - \boldsymbol{D}_i^{-1} \boldsymbol{B}_i (\boldsymbol{I}_q + \boldsymbol{B}_i^T \boldsymbol{D}_i^{-1} \boldsymbol{B}_i)^{-1} \boldsymbol{B}_i^T \boldsymbol{D}_i^{-1}, \tag{60}$$

where the right-hand side of (60) involves only the inverses of $q \times q$ matrices, since $\boldsymbol{D}_i$ is a diagonal matrix. The determinant of $(\boldsymbol{B}_i \boldsymbol{B}_i^T + \boldsymbol{D}_i)$ can then be calculated as

$$\mid \boldsymbol{B}_i \boldsymbol{B}_i^T + \boldsymbol{D}_i \mid = \mid \boldsymbol{D}_i \mid \, / \, \mid \boldsymbol{I}_q - \boldsymbol{B}_i^T (\boldsymbol{B}_i \boldsymbol{B}_i^T + \boldsymbol{D}_i)^{-1} \boldsymbol{B}_i \mid .$$

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

$$\hat{\boldsymbol{D}}_i = \mathrm{diag}(\hat{\boldsymbol{V}}_i - \hat{\boldsymbol{B}}_i \hat{\boldsymbol{B}}_i^T), \tag{61}$$

where

$$\hat{\boldsymbol{V}}_i = \sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j;\, \hat{\boldsymbol{\Psi}})\, (\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_i)^T / \sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j;\, \hat{\boldsymbol{\Psi}}). \tag{62}$$

As remarked by Lawley and Maxwell[39](Page 30) in the context of direct computation of the ML estimate for a single-component factor analysis model, the equation (61) looks temptingly simple to use to solve for $\hat{\boldsymbol{D}}_i$, but was not recommended due to convergence problems.

On comparing (61) with (58), it can be seen that with the calculation of the ML estimate of $\boldsymbol{D}_i$ directly from the (incomplete-data) log likelihood function, the unconditional expectation of $\boldsymbol{U}_j \boldsymbol{U}_j^T$, which is the identity matrix, is used in place of the conditional expectation in (59) on the E-step of the AECM algorithm. Unlike the direct approach of calculating the ML estimate, the EM algorithm and its variants such as the AECM version have good convergence properties in that they ensure the likelihood is not decreased after each iteration regardless of the choice of starting point.

It can be seen from (61) that some of the estimates of the elements of the diagonal matrix $\boldsymbol{D}_i$ (the uniquenesses) 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 $\boldsymbol{D}_i$,

$$\boldsymbol{D}_i = \boldsymbol{D} \quad (i = 1, \ldots, g). \tag{63}$$

An alternative way of proceeding is to adopt some prior distribution for the $\boldsymbol{D}_i$ as in the Bayesian approach of Fokoué and Titterington[40].

The mixture of probabilistic component analyzers (PCAs) model, as proposed by Tipping and Bishop[37], has the form (41) for each $\boldsymbol{\Sigma}_i$ with each $\boldsymbol{D}_i$ now having the isotropic structure

$$\boldsymbol{D}_i = \sigma_i^2 \boldsymbol{I}_p \quad (i = 1, \ldots, g). \tag{64}$$

Under this isotropic restriction (64) 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 $\boldsymbol{V}_i$.

13.5 Initialization of AECM algorithm

We can make use of the link of factor analysis with the probabilistic PCA model (64) to specify an initial value $\boldsymbol{\Psi}^{(0)}$ for $\boldsymbol{\Psi}$ in the ML fitting of the mixture of factor analyzers via the AECM algorithm. On noting that the transformed data $\boldsymbol{D}_i^{-1/2}\boldsymbol{Y}_j$ satisfies the probabilitistic PCA model (64) with $\sigma_i^2 = 1$, it follows that for a given $\boldsymbol{D}_i^{(0)}$ and $\boldsymbol{\Sigma}_i^{(0)}$, we can specify $\boldsymbol{B}_i^{(0)}$ as

$$\boldsymbol{B}_i^{(0)} = \boldsymbol{D}_i^{(0)^{1/2}} \boldsymbol{A}_i (\boldsymbol{\Lambda}_i - \tilde{\sigma}_i^2 \boldsymbol{I}_q)^{1/2} \quad (i = 1, \ldots, g), \tag{65}$$

where

$$\tilde{\sigma}_i^2 = \sum_{h=q+1}^{p} \lambda_{ih}/(p-q).$$

The q columns of the matrix $\boldsymbol{A}_i$ are the eigenvectors corresponding to the eigenvalues $\lambda_{i1} \geq \lambda_{i2} \geq \cdots \geq \lambda_{iq}$ of

$$\boldsymbol{D}_i^{(0)^{-1/2}} \boldsymbol{\Sigma}_i^{(0)} \boldsymbol{D}_i^{(0)^{-1/2}}, \tag{66}$$

and $\boldsymbol{\Lambda}_i = \mathrm{diag}(\lambda_{i1}, \ldots, \lambda_{iq})$. The use of $\tilde{\sigma}_i^2$ instead of unity is proposed in (65), because it avoids the possibility of negative values for $(\boldsymbol{\Lambda}_i - \boldsymbol{I}_q)$, which can occur since estimates are being used for the unknown values of $\boldsymbol{D}_i$ and $\boldsymbol{\Sigma}_i$ in (66).

To specify $\boldsymbol{\Sigma}_i^{(0)}$ for use in (66), we can randomly assign the data into g groups and take $\boldsymbol{\Sigma}_i^{(0)}$ to be the sample covariance matrix of the ith group ($i = 1, \ldots, g$). Concerning the choice of $\boldsymbol{D}_i^{(0)}$, we can take $\boldsymbol{D}_i^{(0)}$ to be the diagonal matrix formed from the diagonal elements of $\boldsymbol{\Sigma}_i^{(0)}$ ($i = 1, \ldots, g$). In this case, the matrix (66) has the form of a correlation matrix.

The eigenvalues and eigenvectors for use in (66) can be found by a singular value decomposition of each $p \times p$ sample component-covariance matrix $\boldsymbol{\Sigma}_i^{(0)}$. But if the number of dimensions p is appreciably greater than the sample size n, then it is much quicker to find them by a singular value decomposition of the $n_i \times n_i$ matrix $\tilde{\boldsymbol{\Sigma}}_i^{(0)}$, the sample matrix formed by taking the observations to be the rows rather than the columns of the $p \times n_i$ data matrix whose n_i columns are the p-dimensional observations assigned initially to the ith

component $(i = 1, \ldots, g)$. The eigenvalues of this latter matrix are equal to those of $\boldsymbol{\Sigma}_i^{(0)}$ apart from a common multiplier due to the different divisors in their formation.

A formal test for the number of factors can be undertaken using the likelihood ratio λ, as regularity conditions hold for this test conducted at a given value for the number of components g. For the null hypothesis that $H_0 : q = q_0$ versus the alternative $H_1 : q = q_0 + 1$, the statistic $-2\log\lambda$ is asymptotically chi-squared with $d = g(p - q_0)$ degrees of freedom. However, in situations where n is not large relative to the number of unknown parameters, we prefer the use of the BIC criterion of Schwarz[25]. Applied in this context, it means that twice the increase in the log likelihood $(-2\log\lambda)$ has to be greater than $d\log n$ for the null hypothesis to be rejected.

14 Mixtures of t Factor Analyzers

The mixture of factor analyzers model is sensitive to outliers since it uses normal errors and factors. Recently, McLachlan, Bean, and Ben-Tovim Jones[41] have considered the use of mixtures of t analyzers in an attempt to make the model less sensitive to outliers. Zhao and Jiang[42] have independently considered this problem in the special case of spherical $\boldsymbol{D}_i$.

14.1 Formulation of the mixture of t-factor analyers model

Following McLachlan et al.[41], we now formulate our mixture of t analyzers model by replacing the multivariate normal distribution in (16) for the ith component-conditional distribution of $\boldsymbol{Y}_j$ by the multivariate t distribution with mean vector vector $\boldsymbol{\mu}_i$, scale matrix $\boldsymbol{\Sigma}_i$, and ν_i degrees of freedom with the factor analytic restriction (41) on the component-scale matrices $\boldsymbol{\Sigma}_i$. Thus our postulated mixture model of t factor analyzers assumes that $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ is an observed random sample from the t mixture density

$$f(\boldsymbol{y}_j; \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i f_t(\boldsymbol{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \nu_i), \tag{67}$$

where

$$\boldsymbol{\Sigma}_i = \boldsymbol{B}_i\boldsymbol{B}_i^T + \boldsymbol{D}_i \quad (i = 1, \ldots, g) \tag{68}$$

and where now the vector of unknown parameters $\boldsymbol{\Psi}$ consists of the degrees of freedom ν_i in addition to the mixing proportions π_i and the elements of

the $\boldsymbol{\mu}_i, \boldsymbol{B}_i$, and the $\boldsymbol{D}_i$ $(i = 1, \ldots, g)$. As in the mixture of factor analyzers model, $\boldsymbol{B}_i$ is a $p \times q$ matrix and $\boldsymbol{D}_i$ is a diagonal matrix.

In order to fit this model (67) with the restriction (68), it is computationally convenient to exploit its link with factor analysis. Accordingly, corresponding to (39), we assume that

$$\boldsymbol{Y}_j = \boldsymbol{\mu}_i + \boldsymbol{B}_i \boldsymbol{U}_{ij} + \boldsymbol{e}_{ij} \qquad \text{with prob. } \pi_i \quad (i = 1, \ldots, g) \tag{69}$$

for $j = 1, \ldots, n$, where the joint distribution of the factor $\boldsymbol{U}_{ij}$ and of the error $\boldsymbol{e}_{ij}$ needs to be specified so that it is consistent with the t mixture formulation (67) for the marginal distribution of $\boldsymbol{Y}_j$.

For the normal factor analysis model, we have that conditional on membership of the ith component of the mixture the joint distribution of $\boldsymbol{Y}_j$ and its associated vector of factors $\boldsymbol{U}_{ij}$ is multivariate normal,

$$\begin{pmatrix} \boldsymbol{Y}_j \\ \boldsymbol{U}_{ij} \end{pmatrix} \mid z_{ij} = 1 \ \sim N_{p+q}(\boldsymbol{\mu}_i^*, \boldsymbol{\xi}_i) \quad (i = 1, \ldots, g). \tag{70}$$

where the mean $\boldsymbol{\mu}_i^*$ and the covariance matrix $\boldsymbol{\xi}_i$ are given by

$$\boldsymbol{\mu}_i^* = (\boldsymbol{\mu}_i^T, \boldsymbol{0}^T)^T \tag{71}$$

and

$$\boldsymbol{\xi}_i = \begin{pmatrix} \boldsymbol{B}_i \boldsymbol{B}_i^T + \boldsymbol{D}_i & \boldsymbol{B}_i \\ \boldsymbol{B}_i^T & \boldsymbol{I}_q \end{pmatrix}. \tag{72}$$

We now replace the normal distribution by the t distribution in (70) to postulate that

$$\begin{pmatrix} \boldsymbol{Y}_j \\ \boldsymbol{U}_{ij} \end{pmatrix} \mid z_{ij} = 1 \ \sim t_{p+q}(\boldsymbol{\mu}_i^*, \boldsymbol{\xi}_i, \nu_i) \quad (i = 1, \ldots, g). \tag{73}$$

This specification of the joint distribution of $\boldsymbol{Y}_j$ and its associated factors in (69) will imply the t mixture model (67) for the marginal distribution of $\boldsymbol{Y}_j$ with the restriction (68) on its component-scale matrices.

Using the characterization of the t distribution discussed earlier, it follows that we can express (73) alternatively as

$$\begin{pmatrix} \boldsymbol{Y}_j \\ \boldsymbol{U}_{ij} \end{pmatrix} \mid w_j, z_{ij} = 1 \ \sim N_{p+q}(\boldsymbol{\mu}_i^*, \boldsymbol{\xi}_i / w_j), \tag{74}$$

where w_{ij} is a value of the weight variable W_j taken to have the gamma distribution (27).

It can be established from (74) that

$$\boldsymbol{U}_{ij} \mid w_j, z_{ij} = 1 \ \sim N_q(\mathbf{0},\ \boldsymbol{I}_q/w_j) \tag{75}$$

and

$$\boldsymbol{e}_{ij} \mid z_{ij} = 1 \ \sim N_p(\mathbf{0},\ \boldsymbol{D}_i/w_j), \tag{76}$$

and hence that

$$\boldsymbol{U}_{ij} \mid z_{ij} = 1 \ \sim t_q(\mathbf{0}, \boldsymbol{I}_q, \nu_i) \tag{77}$$

and

$$\boldsymbol{e}_{ij} \mid z_{ij} = 1 \ \sim t_p(\mathbf{0}, \boldsymbol{D}_i, \nu_i). \tag{78}$$

Thus with this formulation, the error terms $\boldsymbol{e}_{ij}$ and the factors $\boldsymbol{U}_{ij}$ are distributed according to the t distribution with the same degrees of freedom. However, the factors and error terms are no longer independently distributed as in the normal-based model for factor analysis, but they are uncorrelated. To see this, we have from (74) that conditional on w_j, $\boldsymbol{U}_{ij}$ and $\boldsymbol{e}_{ij}$ are uncorrelated, and hence, unconditionally uncorrelated.

14.2 An AECM algorithm for mixtures of t-factor analyzers

We can fit the mixture of t factor analyzers model specified by (67) and (68) using the AECM algorithm (Meng and van Dyk[38]), as described in McLachlan et al.[41]. More specifically, we declare the missing data to be the component-indicators z_{ij}, the factors $\boldsymbol{u}_{ij}$ in (69), and the weights w_j in the characterization (74) of the t-distribution for the ith component distribution of $\boldsymbol{Y}_j$ and $\boldsymbol{U}_{ij}$. We have from (74) that

$$\boldsymbol{Y}_j \mid \boldsymbol{u}_{ij}, w_j, z_{ij} = 1 \ \sim N_p(\boldsymbol{\mu}_i + \boldsymbol{B}_i\boldsymbol{u}_{ij},\ \boldsymbol{D}_i/w_j) \tag{79}$$

for $i = 1, \ldots, g$.

Thus in the EM framework for this problem, the complete data consist, in addition to the observed data $\boldsymbol{y}_j$, of the component-indicators z_{ij}, the unobservable weights w_j, and the latent factors $\boldsymbol{u}_{ij}$.

Using the results (77) to (79), it is straightforward to show that an AECM algorithm can be formulated as in Section 13.2 to iteratively fit the mixture of t-factor analyzers model as specified by (67) and (68).

We use two CM steps in the AECM algorithm, which correspond to the partition of $\boldsymbol{\Psi}$ into the two subvectors $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$, where $\boldsymbol{\Psi}_1$ contains the mixing proportions, the elements of the $\boldsymbol{\mu}_i$, and the degrees of freedom $\nu_i\,(i = 1, \ldots, g)$. The subvector $\boldsymbol{\Psi}_2$ contains the elements of the matrix $\boldsymbol{B}_i$ of factor loadings and of the diagonal matrix $\boldsymbol{D}_i$.

On the first cycle, we specify the missing data to be the component-indicator variables Z_{ij} and the weights w_j in the characterization (74) of the t-distribution for the component distribution of $\boldsymbol{y}_j$. On the $(k+1)$th iteration of the algorithm, we update the estimates of the mixing proportions using (52). The updated estimate of the ith component mean $\boldsymbol{\mu}_i$ is given by

$$\boldsymbol{\mu}_i^{(k+1)} = \sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j;\ \boldsymbol{\Psi}^{(k)}) w_{ij}^{(k)} \boldsymbol{y}_j / \sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j;\ \boldsymbol{\Psi}^{(k)}) w_{ij}^{(k)}, \tag{80}$$

where the current weight $w_{ij}^{(k)}$ is formed using the current value $\boldsymbol{\Psi}^{(k)}$ for $\boldsymbol{\Psi}$ in (30).

In the case where the degrees of freedom ν_i in the component t-distributions are not specified but are to be estimated from the data, we have to update the estimate of ν_i on the second cycle. The updated estimate $\nu_i^{(k+1)}$ of ν_i does not exist in closed form, but is given as a solution of the equation

$$\left\{ -\psi(\tfrac{1}{2}\nu_i) + \log(\tfrac{1}{2}\nu_i) + 1 + \frac{1}{n_i^{(k)}} \sum_{j=1}^{n} \tau_{ij}^{(k)} (\log w_{ij}^{(k)} - w_{ij}^{(k)}) \right.$$
$$\left. + \psi\left(\frac{\nu_i^{(k)} + p}{2}\right) - \log\left(\frac{\nu_i^{(k)} + p}{2}\right) \right\} = 0, \tag{81}$$

where $\tau_{ij}^{(k)} = \tau_i(\boldsymbol{y}_j;\ \boldsymbol{\Psi}^{(k)}), n_i^{(k)} = \sum_{j=1}^{n} \tau_{ij}^{(k)}\ (i = 1, \ldots, g)$, and $\psi(\cdot)$ is the Digamma function.

The estimate of $\boldsymbol{\Psi}$ is updated so that its current value after the first cycle is given by

$$\boldsymbol{\Psi}^{(k+1/2)} = (\boldsymbol{\Psi}_1^{(k+1)^T},\ \boldsymbol{\Psi}_2^{(k)^T})^T. \tag{82}$$

On the second cycle of this iteration, the complete data are expanded to include the unobservable factors $\boldsymbol{U}_{ij}$ associated with the $\boldsymbol{y}_j$. The estimates of

the matrix of factor loadings $\boldsymbol{B}_i$ and the diagonal matrix $\boldsymbol{D}_i$ can be updated using (54) and (55), but where the ith component sample covariance matrix is calculated as

$$\boldsymbol{V}_i^{(k+1/2)} = \frac{\sum_{j=1}^n \tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k+1/2)})\, w_{ij}^{(k+1/2)} (\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k+1)})(\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k+1)})^T}{\sum_{j=1}^n \tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k+1/2)})}, \quad (83)$$

where $w_{ij}^{(k+1/2)}$ is updated partially by using $\boldsymbol{\Psi}^{(k+1/2)}$ for $\boldsymbol{\Psi}$ in (30).

15 Available Software

The reader is referred to the appendix in McLachlan and Peel[7] for the availability of software for the fitting of normal mixture models, including the EMMIX program of McLachlan et al.[43] The current version of EMMIX is available from the World Wide Web address

`http://www.maths.uq.edu.au/~gjm/emmix/emmix.html`

Concerning the availability of mixture modeling facilities in general-purpose statistical packages, there is the MCLUST software package of Fraley and Raftery[19], which is interfaced to the S-PLUS commercial software and the open source R code.

16 Example

We now illustrate the use of normal mixture models by applying them to two data sets, the so-called Vietnam and Thyroid data sets as considered in Smyth et al.[44] The Vietnam data set consists of the log transformed and z-standardized concentrations of $p = 17$ chemical elements to which synthetic noise variables were added by Smyth et al.[44] to study methods for clustering high dimensional data. The concentrations were measured in hair samples from six classes of Vietnamese. These classes differed in their age and exposure to coal. There were $n = 224$ subjects and the classes were: (1) Control adults: n_1=31 males with low exposure to coal; (2) Control children: n_2=children with low exposure to coal; (3) Miner adults: n_3=56 males employed at a coal mine; (4) Miner children: n_4=47 children of male coal workers; (5) Burner adults: n_5=18 females using coal for cooking; (6) Burner children: n_6=41 children with exposure to coal through its use in cooking.

The Thyroid data set consists of the z-standardized concentrations of $p = 5$

hormones to which 400 synthetic noise variables similar to those in the Vietnam data set were added by Smyth et al.[44] They were measured in $n = 215$ patients. The patients were divided into three classes according to Thyroid function. The classes were: (1) Normal thyroid function ($n_1 = 150$); (2) Hyperthyroid function ($n_2 = 35$); (3) Hypothyroid function ($n_3 = 30$).

A plot of the first two canonical variates of these two data sets is given in Figures 1 and 2, respectively. Note that since there are only three classes for the Thyroid data set, the data can be fully represented in the space of the first two canonical variates.

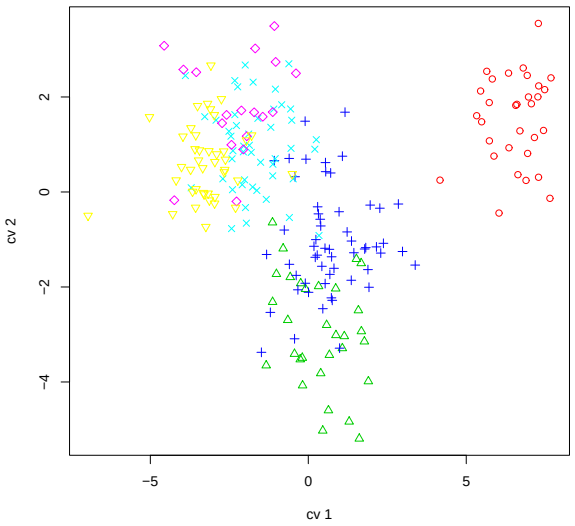

Fig. 1. Plot of the first two canonical variates for the $g = 6$ classes in the Vietnam data.

The 400 additional noise variables added by Smyth et al.[44] to these two data sets consisted of 100 of each of four different types of noise variable. The noise variables were all z-standardized. They examined the effects on clustering methods of adding p, 50 and 100 of each of the noise variable types, where p was equal to the dimension of the original feature vector.

For these two data sets, we looked at the effect on normal mixture models and mixtures of factor analyzers of adding 50 of each of two types of their noise variables, which were normally distributed noise and uniformly distributed noise.

If we had tried to fit normal mixture mdoels with unrestricted component-covariance matrices to the Vietnam data set, then we would have had problems with singularities or near-singular estimates of the component-covariances matrices. This is because there are $p = 17$ variables, but only 18 observations in one of the classes. Similarly, with the Thyroid data set, because the two

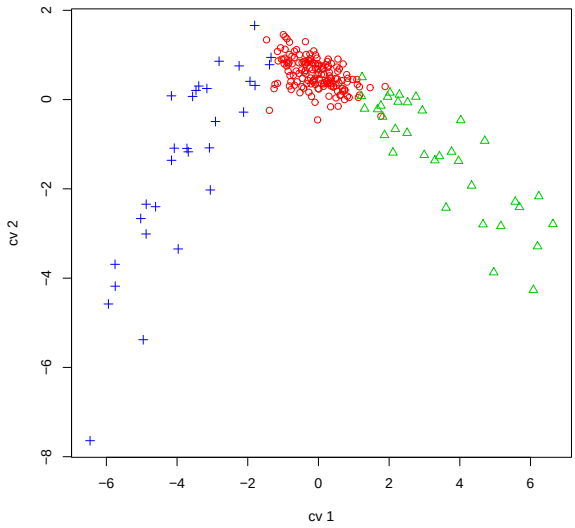

Fig. 2. Plot of the first two canonical variates for the $g = 3$ classes in the thyroid data.

smaller classes had only 30 and 35 patients in them, use of normal mixture models with unrestricted component-covariance matrices would have led to singularities with 50 noise variables added. We therefore fitted mixtures of factor analyzers both with equal and unequal component-covariance matrices. In the latter case, the diagonal matrices $\boldsymbol{D}_i$ in the factor analysis representation (41) of the component-covariance matrices were constrained to be equal; that is, we took the uniquenesses to be common across the components. For the mixture of factor analyzers models we examined models with $q = 2, 3, 4$, and 5 factors per component, but have only presented the results for q=2 factors per cluster here. The models with $q = 2$ factor analyzers per cluster were the ones which most accurately reproduced the classes described above.

We fitted the normal mixture models and mixtures of factor analyzers using 50 random and 50 k-means starts and then choosing the one with highest likelihood. We then used the EM-algorithm to obtain our final solution using this as our initial classification. In doing this we specified the number of components g to be the same as the number of *a priori* specified number of classes; that is, $g = 6$ for the Vietnam data and $g = 3$ for the Thyroid data.

For each clustering obtained, we permuted the cluster labels so that the error rate of misallocation was minimized when the clusters were identified with the true classes. We also calculated the adjusted Rand index[45]. This is a measure of the agreement of two partitions of a data set. For two given partitions of a data set (not necessarily having the same number of clusters or classes) the

Rand index[46] is

$$R = (a+b)/\binom{n}{2},$$

where a is the number of pairs of elements which are in the same cluster in the first partition and in the same cluster in the second partition, b is the number of pairs of elements which are in different clusters in the first partition and in different clusters in the second partition and n is the number of elements. The adjusted Rand index is modification of the Rand index so that its expected value when comparing two random partitions is zero.

The results of the clustering are given in Tables 1 and 2, corresponding to the Vietnam and Thyroid data sets, respectively. In these tables, we have listed the adjusted Rand index with the error of misallocation below in parentheses for the normal mixture model (NMM) and the mixture of factor analyzers model (MFA) with q =2 factors and for no noise, uniform noise and normal noise. The "equal" and "unequal" in these tables refers to whether the component-covariance matrices were set to be equal or were unequal in the sense that the component-factor analyzers were allowed to have different loadings (that is, different $\boldsymbol{B}_i$), but common uniquenesses (that is, common $\boldsymbol{D}_i$).

Table 1
Vietnam data Adjusted Rand Index and misclassification rate

Start	Noise	NMM	MFA	MFA
		equal	equal	unequal
True	None	0.961	0.958	0.939
		(0.031)	(0.036)	(0.049)
	Uniform	0.988	0.949	0.992
		(0.009)	(0.054)	(0.009)
	Normal	0.974	0.922	0.981
		(0.022)	(0.067)	(0.018)
50r/50k	None	0.762	0.766	0.907
		(0.201)	(0.196)	(0.098)
	Uniform	0.778	0.825	0.867
		(0.214)	(0.143)	(0.143)
	Normal	0.818	0.783	0.867
		(0.143)	(0.210)	(0.143)

We found that if there were no noise, the clusters created by the mixture of factor analyzers model (with unrestricted factor loadings) agreed very well

Table 2
Thyroid data Adjusted Rand Index and misclassification rate

Start	Noise	NMM equal	NMM unequal	MFA equal	MFA unequal
True	None	0.380	0.931	0.380	0.923
		(0.191)	(0.042)	(0.191)	(0.037)
	Uniform	0.819		0.528	0.884
		(0.061)		(0.149)	(0.061)
50r/50k	None	0.147	0.931	0.312	0.906
		(0.279)	(0.042)	(0.219)	(0.047)
	Uniform	0.104		0.254	0.807
		(0.488)		(0.488)	(0.102)

with the true classes. Also, in this case, the full normal mixture model performed well in those instances where it was able to be fitted. The imposition of equal component-covariance matrices led to more misallocations. For the Vietnam data set, the agreement between the output clusters and the true classes was still acceptable, but for the Thyroid data set it was quite poor. The Thyroid data set appears to have a class configuration that cannot be adequately represented by a model with equal component-covariance matrices. It has little separation between classes and large differences in their variance structure. This can be seen by examination of the class structure in Figure 2.

Concerning the introduction of the noise variables to the originally measured variables, we concluded that if the component-covariance matrices were allowed to be unequal, added noise did not greatly reduce the accuracy of the clustering we obtained from the mixtures of factor analyzers models. We were not able to evaluate the effect of noise on normal mixture models with unrestricted component-covariance matrices since we were unable to fit these models to these data sets with noise variables added. The effect of noise on the full normal mixture model with equal component-covariance matrices was erratic. It was found that it could lead to either increases or to decreases in the accuracy of the clusterings obtained.

The adjusted Rand indices for our model-based clusterings of these two data sets were higher than those calculated in Smyth et al.[44] For the case with 50 normal noise variables and q=2 factor analyzers per cluster, they calculated an adjusted Rand index of 0.415 for the Vietnam data and 0.735 for the Thyroid data. Our calculated values are 0.867 and 0.807, respectively. The reason for this large difference on the Vietnam data set is unknown. Their

preferred classification method was multivariate regression trees using global factor scores. Using this method the adjusted Rand indices they obtained for these data sets were 0.808 and 0.766.

Up to now for each of these two data sets, we have specified the number of components in our mixture models to be the same as the number of classes. We also looked at the case where the number of component g in the mixture model were selected via consideration of the log likelihood. We used the resampling approach of McLachlan[26] as described in Section 10.2. Starting with a single factor analyzer ($g = 1$), we proceeded to fit an additional component analyzer (with the same uniquenesses as the existing component analyzers) provided the likelihood ratio test for an additional factor analyzer was found to be significant. It led to the choice of $g = 6$ and $g = 3$ for the Vietnam and Thyroid data sets, respectively, coinciding with the specified number of classes in these two sets.

17 Some recent extensions for high-dimensional data

The EMMIX-GENE program of McLachlan, Peel, and Bean[47] has been designed for the normal mixture model-based clustering of a limited number of observations that may be of extremely high-dimensions. It was called EMIX-GENE as it was designed specifically for problems in bioinformatics that require the clustering of a relatively small number of tissue samples containing the expression levels of possibly thousands of genes. But it is applicable to clustering problems outside the field of bioinformatics involving high-dimensional data. In situations where the sample size n is very large relative to the dimension p, it might not be practical to fit mixtures of factor analyzers, as it would involve a considerable amount of computation time. Thus initially some of the variables may have to be removed. Indeed, the simultaneous use of too many variables in the cluster analysis may serve only to create noise that masks the effect of a smaller number of variables. Also, the intent of the cluster analysis may not be to produce a clustering of the observations on the basis of all the available genes, but rather to discover and study different clusterings of the observations corresponding to different subsets of the variables.

Therefore, the EMMIX-GENE procedure has two optional steps before the final step of clustering the observations. The first step considers the selection of a subset of relevant variables from the available set of variables by screening the variables on an individual basis to eliminate those which are of little use in clustering the observations. The usefulness of a given variable to the clustering process can be assessed formally by a test of the null hypothesis that it has a single component normal distribution over the observations. A faster but *ad hoc* way is to make this decision on the basis of the interquartile range.

Even after this step has been completed, there may still remain too many variables. Thus there is a second step in EMMIX-GENE in which the retained variables are clustered (after standardization) into a number of groups on the basis of Euclidean distance so that variables with similar profiles are put into the same group. In general, care has to be taken with the scaling of variables before clustering of the observations, as the nature of the variables can be intrinsically different. Also, as noted above, the clustering of the observations via normal mixture models is invariant under changes in scale and location. The clustering of the observations can be carried out on the basis of the groups considered individually using some or all of the variables within a group or collectively. For the latter, we can replace each group by a representative (a metavariable) such as the sample mean as in the EMMIX-GENE procedure.

(a) there are no replications on any particular entity specifically identified as such;
(b) all the observations on the entities are independent of one another.

These assumptions should hold for the clustering of, say, tissue samples as discussed above, although the tissue samples have been known to be correlated for different tissues due to flawed experimental conditions. However, condition (b) will not hold for the clustering of gene profiles, since not all the genes are independently distributed, and condition (a) will generally not hold either as the gene profiles may be measured over time or on technical replicates. While this correlated structure can be incorporated into the normal mixture model (15) by appropriate specification of the component-covariance matrices Σ_i, it is difficult to fit the model under such specifications. For example, the M-step may not exist in closed form.

Accordingly, Ng et al.[48] have developed the procedure called EMMIX-WIRE (**EM**-based **MIX**ture analysis **W**ith **R**andom **E**ffects) to handle the clustering of correlated data that may be replicated. They adopted conditionally a mixture of linear mixed models to specify the correlation structure between the variables and to allow for correlations among the observations. It also enables covariate information to be incorporated into the clustering process.

18 Mixed feature data

We consider now 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 p_1 discrete variables, represented by the subvector $\boldsymbol{y}_{1j}$, in addition to p_2 continuous variables represented by the subvector $\boldsymbol{y}_{2j}$ $(j = 1, \ldots, n)$. The ith

component density of the jth observation

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

can then be written as

$$f_i(y_j) = f_i(\boldsymbol{y}_{1j}) f_i(\boldsymbol{y}_{2j} \mid \boldsymbol{y}_{1j}), \tag{84}$$

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 (Titterington et al.[49], Hand and Yi[50]). Under this assumption, the ith component-conditional density of the vector $\boldsymbol{y}_{1j}$ of discrete features is given by

$$f_i(\boldsymbol{y}_{1j}) = \prod_{v=1}^{p_1} f_{iv}(y_{1vj}), \tag{85}$$

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 (85) the $(k+1)$th update of $f_{iv}(y_{1v})$ is

$$f_{iv}^{(k+1)}(y_{1v}) = \frac{\sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j;\ \boldsymbol{\Psi}^{(k)})\ \delta[y_{1vj},\ y_{1v}] + c_1}{\sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j;\ \boldsymbol{\Psi}^{(k)}) + c_2}, \tag{86}$$

where $\delta[y_{1vj},\ y_{1v}] = 1$ if $y_{1vj} = y_{1v}$ and is zero otherwise, and $\boldsymbol{\Psi}^{(k)}$ is the current estimate of the vector of all the unknown parameters that now include the probabilities for the discrete variables. In (86), the constants c_1 and c_2, which are both equal to zero for the maximum likelihood 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} (Titterington et al.[49]).

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[51]. With the location model, $f_i(\boldsymbol{y}_{2j} \mid \boldsymbol{y}_{1j})$ 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 $\boldsymbol{y}_{1j}$, as in McLachlan and Chang[52]. This leads to the use of the conditional mixture model for the continuous feature vector $\boldsymbol{y}_{2j}$,

$$f(\boldsymbol{y}_{2j} \mid \boldsymbol{y}_{1j}) = \sum_{i=1}^{g} \pi_i(\boldsymbol{y}_{1j})\, f_i(\boldsymbol{y}_{2j} \mid \boldsymbol{y}_{1j}), \tag{87}$$

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

$$\pi_i(\boldsymbol{y}_{1j}) = \frac{\exp(\beta_{i0} + \boldsymbol{\beta}_i^T \boldsymbol{y}_{1j})}{1 + \sum_{h=1}^{g-1} \exp(\beta_{h0} + \boldsymbol{\beta}_h^T \boldsymbol{y}_{1j})}, \tag{88}$$

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

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

19 Acknowledgement

The author would like to thank Lloyd Flack for his assistance with the analyses undertaken in presenting the illustrative example.

20 References

1. Scott, A.J.; Symons, M.J. Clustering methods based on likelihood ratio criteria. *Biometrics* **1971**, 27, 387–397.

2. Bryant, P.G.; Williamson, J.A. Asymptotic behaviour of classification maximum likelihood estimates. *Biometrika* **1978**, 65, 273–281.

3. McLachlan, G.J. The classification and mixture maximum likelihood approaches to cluster analysis. In *Handbook of Statistics*, Vol. 2, Krishnaiah, P.I., Kanal, L., Eds.; North–Holland: Amsterdam, 1982, pp 199–208.

4. Wolfe, J.H. A computer program for the computation of maximum likelihood analysis of types. *Research Memo. SRM 65-12*. San Diego: U.S. Naval Personnel Research Activity, 1965.

5. Day, N.E. Estimating the components of a mixture of two normal distributions. *Biometrika* **1969**, 56, 463–474.

6. Böhning, D. *Computer-Assisted Analysis of Mixtures and Applications: Meta-Analysis, Disease Mapping and Others*; *Chapman & Hall/CRC*: New York, 1999.

7. McLachlan, G.J.; Peel, D. *Finite Mixture Models*; Wiley: New York, 2000.

8. Frühwirth-Schnatter, S. *Finite Mixture and Markov Switching Models*; Springer: New York, 2006.

9. Everitt, B.S.; Hand, D.J. *Finite Mixture Distributions*; Chapman & Hall: London, 1981.

10. Titterington, D.M.; Smith, A.F.M.; Makov, U.E.; *Statistical Analysis of Finite Mixture Distributions*; Wiley: New York, 1985.

11. McLachlan, G.J.; Basford, K.E. *Mixture Models: Inference and Applications to Clustering*; Marcel Dekker: New York, 1988.

12. Lindsay, B.G. *Mixture Models: Theory, Geometry and Applications*. NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 5; Institute of Mathematical Statistics and the American Statistical Association, Alexandria, VA., 1995.

13. Li, J.Q.; Barron, A.R. Mixture density estimation. Technical Report, Department of Statistics, Yale University, New Haven, Connecticut, 2000.

14. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. B* **1977**, 39, 1–38.

15. McLachlan, G.J.; Krishnan, T. *The EM Algorithm and Extensions*; Wiley: New York, 1997.

16. Coleman, D.; Dong, X.; Hardin, J.; Rocke, D.M.; Woodruff, D.L. Some computational issues in cluster analysis with no a priori metric. *Comput.*

Statist. Data Anal. **1999**, 31, 1–11.

17. Hartigan, J.A. *Clustering Algorithms*; Wiley: New York, 1975.

18. Banfield, J.D.; Raftery, A.E. Model-based Gaussian and non-Gaussian clustering. *Biometrics* **1993**, 49, 803–821.

19. Fraley, C.; Raftery, A.E. Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Statist. Assoc.* **2002**, 97, 611–631.

20. Hennig, C. Breakdown points for maximum likelihood-estimators of location-scale mixtures. *Ann. Statist.* **2004**, 32, 1313–1340.

21. Peel, D.; McLachlan. G.J. Robust mixture modelling using the t distribution. *Statist. Comput.* **2000**, 10, 335–344.

22. Meng, X.L.; Rubin, D.B. Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* **1993**, **80**, 267–278.

23. Kent, J.T.; Tyler, D.E.; Vardi, Y. A curious likelihood identity for the multivariate t-distribution. *Comm. Statist. Simul. Comput.* **1994**, 23, 441–453.

24. Liu, C.; Rubin, D.B.; Wu, Y.N. Parameter expansion to accelerate EM: the PX-EM algorithm. *Biometrika* **1998**, 85, 755–770.

25. Schwarz, G. Estimating the dimension of a model. *Ann. Statist.* **1978**, 6, 461–464.

26. McLachlan, G.J. On bootstrapping thelikelihood ratio test statistic for the number of components in a normal mixture. *Appl. Statist.* **1987**, 36, 318–324.

27. Tibshirani, R.; Walther, G.; Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *J. Roy. Statist. Soc. B* **2001**, 63, 411–423.

28. Dudoit, S.; Fridlyand, J. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biol.* **2002**, 3, research0036.1–0036.21 .

29. Hawkins, D.M.; Muller, M.W.; ten Krooden, J.A. Cluster analysis. In *Topics in Applied Multivariate Analysis*, D.M. Hawkins (Ed.); Cambridge,

Cambridge University Press, 1982, pp. 303–356.

30. Aitkin, M.; Anderson, D.; Hinde, J. Statistical modelling of data on teaching styles (with discussion). *J. Roy. Statist. Soc. A* **1981**, 144, 414–461.

31. Marriott, F.H.C. *The Interpretation of Multiple Observations*; Academic Press: London, 1974.

32. Yeung, K.Y.; Fraley, C.; Murua, A.; Raftery, A.E.; Ruzzo, W.L. Model-based clustering and data transformations for gene expression data. *Bioinformatics* **2001**, 17, 977–987.

33. McLachlan, G.J.; Peel, D. Mixtures of factor analyzers. Proceedings of the Seventeenth International Conference on Machine Learning, 2000, Morgan Kaufmann, San Francisco,, 599–606.

34. McLachlan, G.J.; Peel, D.; Bean, R.W. Modelling high-dimensional data by mixtures of factor analyzers. *Comput. Statist. Data Anal.* **2003**, 41, 379–388.

35. Ghahramani, Z.; Hinton, G.E. The EM algorithm for factor analyzers. Technical Report No. CRG-TR-96-1, The University of Toronto, Toronto, 1997.

36. Hinton, G.E.; Dayan, P.; Revow, M. Modeling the manifolds of images of handwritten digits. *IEEE Trans. Neural Networks* **1997**, 8, 65–73.

37. Tipping, M.E.; Bishop, C.M. Mixtures of probabilistic principal component analysers. Technical Report No. NCRG/97/003, Neural Computing Research Group, Aston University, Birmingham, 1997.

38. Meng, X.L.; van Dyk, D. The EM algorithm—an old folk song sung to a fast new tune (with discussion). *J. Roy. Statist. Soc. B* **1997**, 59, 511–567.

39. Lawley, D.N.; Maxwell, A.E. *Factor Analysis as a Statistical Method*; 2nd ed.; Butterworths: London, 1971.

40. Fokoué, E.; Titterington, D.M. Mixtures of factor analyzers. Bayesian estimation and inference by stochastic simulation. *Machine Learning* **2002**, 50, 73–94.

41. McLachlan, G.J.; Bean, R.W.; Peel, D. Extension of the mixture of fac-

tor analyzers model to incorporate the multivariate t distribution. *Comput. Statist. Data Anal.* **2007**. To appear.

42. Zhao, J.; Jiang, Q. Probabilistic PCA for t distributions. *Neurocomputing* **2006**, 69, 2217–2226.

43. McLachlan, G.J.; Peel, D.; Basford, K.E.; Adams, P. The EMMIX software for the fitting of mixtures of normal and t-components. *J. Statist. Software* **1999**, 4, No. 2.

44. Smyth, C.; Coomans, D.; Everingham, Y. Clustering noisy data in a reduced dimension space via multivariate regression trees. *Pattern Recognition* **2006**, 39, 424–431.

45. Hubert,L.; Arabie, P. Comparing Partitions. *J. Classification* **1985**, 2, 193–218.

46. Rand, W.M. Objective criteria for the evaluation of clustering methods. *J. Amer. Statist. Assoc.* **1971**, 66, 846–850.

47. McLachlan, G.J.; Bean, R.W.; Peel, D. A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics* **2002**, 18, 413–422.

48. Ng, S. K.; McLachlan G. J.; Wang, K.; Ben-Tovim Jones L.; Ng S.-W. A mixture model with random-effects components for clustering correlated gene-expression profiles. *Bioinformatics* **2006**, 22, 1745–1752.

49. Titterington, D.M.; Murray, G.D., Murray, L.S.; Spiegelhalter, D.J.; Skene, A.M.; Habbema, J.D.F.; Gelpke, G.J. Comparison of discrimination techniques applied to a complex data set of head injured patients (with discussion). *J. Roy. Statist. Soc. A* **1981**, 144, 145–175.

50. Hand, D.J.; Yi, K. Idiot's Bayes – not so stupid after all? *Inter. Statist. Rev.* **2001**, 69, 385–398.

51. Hunt, L.A.; Jorgensen, M.A. Mixture model clustering: a brief introduction to the MULTIMIX program. *Austral. New Zeal. J. Statist.* **1999**, 40, 153–171.

52. McLachlan, G.J.; Chang, S.U. Mixture modelling for cluster analysis. *Statist. Meth. Med. Res.* **2004**, 13, 347–361.