Extension of the Mixture of Factor Analyzers Model to Incorporate the Multivariate t-Distribution

G.J. McLachlan *,1, R.W. Bean, L. Ben-Tovim Jones

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

Abstract

Mixtures of factor analyzers enable model-based density estimation to be undertaken for high-dimensional data, where the number of observations n is very large relative to their dimension p. However, this approach is sensitive to outliers as it is based on a mixture model in which the multivariate normal family of distributions is assumed for the component error and factor distributions. An extension to mixtures of t-factor analyzers is considered, whereby the multivariate t-family is adopted for the component error and factor distributions. An EM-based algorithm is developed for the fitting of mixtures of t-factor analyzers. Its application is demonstrated in the clustering of some microarray gene-expression data.

Key words: Mixture modeling; Factor analyzers; Multivariate t-distribution; EM algorithm

1 Introduction

Finite mixture models are being 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 (1999) and McLachlan and Peel (2000a) and the references therein. Earlier references on mixture models may be found in the previous books by Everitt and Hand (1981), Titterington et al. (1985), McLachlan and Basford (1988), and Lindsay (1995).

* Corresponding author.
Email address: gjm@maths.uq.edu.au (G.J. McLachlan).
[1] Phone: +61 7 3365 2150, Fax +61 7 3365 1477

Let

$$\boldsymbol{Y} = (Y_1, \ldots, Y_p)^T \tag{1}$$

be a p-dimensional vector of feature variables. For continuous features Y_j, the density of $\boldsymbol{Y}$ can be modelled by a mixture of a sufficiently large enough number g of multivariate normal component distributions,

$$f(\boldsymbol{y};\, \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i \, \phi(\boldsymbol{y};\, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{2}$$

where $\phi(\boldsymbol{y};\, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the p-variate normal density function with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Here the vector $\boldsymbol{\Psi}$ of unknown parameters consists of the mixing proportions π_i, the elements of the component means $\boldsymbol{\mu}_i$, and the distinct elements of the component-covariance matrix $\boldsymbol{\Sigma}_i$.

The parameter vector $\boldsymbol{\Psi}$ can be estimated by maximum likelihood. For an observed random sample, $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, the log likelihood function for $\boldsymbol{\Psi}$ is given by

$$\log L(\boldsymbol{\Psi}) = \sum_{j=1}^{n} \log f(\boldsymbol{y}_j;\, \boldsymbol{\Psi}). \tag{3}$$

The maximum likelihood estimate (MLE) of $\boldsymbol{\Psi}$, $\hat{\boldsymbol{\Psi}}$, is given by an appropriate root of the likelihood equation,

$$\partial \log L(\boldsymbol{\Psi})/\partial \boldsymbol{\Psi} = \boldsymbol{0}. \tag{4}$$

Solutions of (4) corresponding to local maximizers of $\log L(\boldsymbol{\Psi})$ can be obtained via the expectation-maximization (EM) algorithm of Dempster, Laird, and Rubin (1977); see also McLachlan and Krishnan (1997).

Besides providing an estimate of the density function of $\boldsymbol{Y}$, the normal mixture model (2) provides a probabilistic clustering of the observed data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ into g clusters in terms of their estimated posterior probabilities of component membership of the mixture. The posterior probability $\tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi})$ that the jth feature vector with observed value $\boldsymbol{y}_j$ belongs to the ith component of the mixture can be expressed by Bayes' theorem as

$$\tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}) = \frac{\pi_i \phi(\boldsymbol{y}_j;\, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{h=1}^{g} \pi_h \phi(\boldsymbol{y}_j;\, \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)}, \tag{5}$$

where π_i is the prior probability that an observation belongs to the ith component $(i = 1, \ldots, g; j = 1, \ldots, n)$. An outright assignment of the data is obtained by assigning each data point $\boldsymbol{y}_j$ to the component to which it has the highest estimated posterior probability of belonging.

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, under operations relating to changes in location, scale, and rotation of the data); see, for example, Coleman et al. (1999). However, the g-component normal mixture model (2) with unrestricted component-covariance matrices is a highly parameterized model with $d = \frac{1}{2}p(p+1)$ parameters for each component-covariance matrix $\boldsymbol{\Sigma}_i\,(i = 1, \ldots, g)$. Banfield and Raftery (1993) introduced a parameterization of the component-covariance matrix $\boldsymbol{\Sigma}_i$ based on a variant of the standard spectral decomposition of $\boldsymbol{\Sigma}_i\,(i = 1, \ldots, g)$. But 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 simple way of proceeding in the clustering of high-dimensional data would be to take the component-covariances matrices $\boldsymbol{\Sigma}_i$ to be diagonal. But this leads to clusters whose axes are aligned with those of the feature space, whereas in practice the clusters are of arbitrary orientation. For instance, taking the $\boldsymbol{\Sigma}_i$ to be a common multiple of the identity matrix leads to a soft-version of k-means which produces spherical clusters. Another way commonly used in practice for 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.

In this paper, we focus on the use of mixtures of factor analyzers as a means of fitting normal mixture models in situations where p is sufficiently large relative to the sample size n to cause potential problems with singular or near-singular estimates of the component-covariance matrices. 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. This approach has been studied in a series of articles by McLachlan (2000a, 2000b) and McLachlan et al. (2003).

However, the mixture of factor analyzers model is sensitive to outliers since it uses normal errors and factors. Here we consider the use of mixtures of t-factor analyzers in an attempt to make the model less sensitive to outliers. An EM-type algorithm is described for the fitting of factor mixture models

where the family of multivariate t-distributions is adopted for the distributions of the latent factors and error terms. We demonstrate the implementation of this algorithm in its application to a real data set involving the clustering of gene expression levels on the basis of tissue samples from some microarray experiments.

2 Mixtures of Factor Analyzers

Factor analysis is commonly used for explaining data, in particular, correlations between variables in multivariate observations. It can be used also for dimensionality reduction. However, a single-factor analysis model like a principal component analysis, provides only a global linear model for the representation of the data in a lower-dimensional subspace. Thus it has limited scope in revealing group structure in a data set.

A global nonlinear approach 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 distribution of the observation $\boldsymbol{Y}_j$ can be modelled as

$$\boldsymbol{Y}_j = \boldsymbol{\mu}_i + \boldsymbol{B}_i \boldsymbol{U}_{ij} + \boldsymbol{e}_{ij} \qquad \text{with prob. } \pi_i \quad (i = 1, \ldots, g) \tag{6}$$

for $j = 1, \ldots, n$, where $\boldsymbol{U}_{ij}$ is a q-dimensional ($q < p$) vector of latent or unobservable variables called factors and $\boldsymbol{B}$ is a $p \times q$ matrix of factor loadings (parameters). The factor (vector) $\boldsymbol{U}_{ij}$ is distributed $N_q(\boldsymbol{0}, \boldsymbol{I}_q)$, independently of $\boldsymbol{e}_{ij}$, which is distributed $N_p(\boldsymbol{0}, \boldsymbol{D}_i)$, where $\boldsymbol{D}_i$ is a diagonal matrix ($i = 1, \ldots, g$) and where $\boldsymbol{I}_q$ denotes the $q \times q$ identity matrix.

Thus the mixture of factor analyzers model is given by (2), where the ith component-covariance matrix $\boldsymbol{\Sigma}_i$ has the form

$$\boldsymbol{\Sigma}_i = \boldsymbol{B}_i \boldsymbol{B}_i^T + \boldsymbol{D}_i \quad (i = 1, \ldots, g), \tag{7}$$

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 elements of the $\boldsymbol{\mu}_i$, the $\boldsymbol{B}_i$, and the $\boldsymbol{D}_i$, along with the mixing proportions π_i ($i = 1, \ldots, g-1$), on putting $\pi_g = 1 - \sum_{i=1}^{g-1} \pi_i$. Unlike the principal component analysis model, the mixture of factor analyzers model (6) 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 (in conjunction with scale changes in the elements of the vectors of means and errors.)

If q is chosen sufficiently smaller than p, the representation (7) imposes some constraints on the component-covariance matrices $\boldsymbol{\Sigma}_i$ 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 (7) is still satisfied if $\boldsymbol{B}_i$ is post multiplied by any orthogonal matrix of order q. As $\frac{1}{2}q(q-1)$ constraints are needed for $\boldsymbol{B}$ to be uniquely defined, the number of free parameters is $pq + p - \frac{1}{2}q(q-1)$. With a principal component analysis, there are $\frac{1}{2}p_1(p_1+1)$ parameters where p_1 is the number of principal components chosen.

3 An AECM algorithm for mixture of factor analyzers models

The mixture of factor analyzers model can be fitted by maximum likelihood via the alternating expectation–conditional maximization (AECM) algorithm (Meng and van Dyk, 1997), as described in McLachlan et al. (2003). We give again the equations here for the updating of the parameters, as they are needed to describe the extension of the algorithm to handle the fitting of mixtures of t-factor analyzers.

The expectation–conditional maximization (ECM) algorithm proposed by Meng and Rubin (1993) replaces the M-step of the EM algorithm by a number of computationally simpler conditional maximization (CM) steps. 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 (1997) established that monotone convergence of the sequence of likelihood values is retained with the AECM algorithm and that under standard regularity conditions, the sequence converges to a stationary value of the likelihood function (which in practice is usually a local maximum).

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 two CM-steps 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, $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$, where $z_{ij} = (\boldsymbol{z}_j)_i$ is one or zero, according to whether $\boldsymbol{y}_j$ arose or did not arise from the ith component

$(i = 1, \ldots, g;\ j = 1, \ldots, n)$. In this conceptualization of the mixture model, it is valid to assume that the observation $\boldsymbol{y}_j$ has arisen from one of the g components.

3.1 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,

$$\boldsymbol{C}_{yui} = E\{Z_{ij}\boldsymbol{y}_j\boldsymbol{U}_{ij}^T \mid \boldsymbol{y}_j\} \tag{8}$$

and

$$\boldsymbol{C}_{uui} = E\{Z_{ij}\boldsymbol{U}_{ij}\boldsymbol{U}_{ij}^T \mid \boldsymbol{y}_j\}. \tag{9}$$

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,

$$\begin{pmatrix} \boldsymbol{Y}_j \\ \boldsymbol{U}_{ij} \end{pmatrix} \mid z_{ij} = 1 \ \sim N_{p+q}(\boldsymbol{\mu}_i^*, \boldsymbol{\xi}_i) \quad (i = 1, \ldots, g), \tag{10}$$

where

$$\boldsymbol{\mu}_i^* = (\boldsymbol{\mu}_i^T, \boldsymbol{0}^T)^T \tag{11}$$

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

$$\boldsymbol{\xi}_i = \begin{pmatrix} \boldsymbol{B}_i\boldsymbol{B}_i^T + \boldsymbol{D}_i & \boldsymbol{B}_i \\ \boldsymbol{B}_i^T & \boldsymbol{I}_q \end{pmatrix}. \tag{12}$$

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

$$\boldsymbol{U}_j \mid \boldsymbol{y}_j,\ z_{ij} = 1 \sim N(\boldsymbol{\gamma}_i^T(\boldsymbol{y}_j - \boldsymbol{\mu}_i),\ \boldsymbol{\Omega}_i) \tag{13}$$

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

$$\boldsymbol{\gamma}_i = (\boldsymbol{B}_i\boldsymbol{B}_i^T + \boldsymbol{D}_i)^{-1}\boldsymbol{B}_i. \tag{14}$$

and where

$$\boldsymbol{\Omega}_i = \boldsymbol{I}_q - \boldsymbol{\gamma}_i^T \boldsymbol{B}_i. \tag{15}$$

Using (13),

$$\boldsymbol{C}_{yui} = \tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi})\boldsymbol{\gamma}_i^T \boldsymbol{y}_j \tag{16}$$

and

$$\boldsymbol{C}_{uui} = \tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi})\{\boldsymbol{\gamma}_i^T(\boldsymbol{y}_j - \boldsymbol{\mu}_i)(\boldsymbol{y}_j - \boldsymbol{\mu}_i)^T\boldsymbol{\gamma}_i + \boldsymbol{\Omega}_i\}; \tag{17}$$

see McLachlan et al. (2003) for further details.

3.2 CM-steps

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

$$\pi_i^{(k+1)} = \sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k)})/n \tag{18}$$

and

$$\boldsymbol{\mu}_i^{(k+1)} = \sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k)})\, \boldsymbol{y}_j / \sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k)}) \tag{19}$$

for $i = 1, \ldots, g$, where $\tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi})$ is the ith component-posterior probability of $\boldsymbol{y}_j$ defined by (5).

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

$$\boldsymbol{B}_i^{(k+1)} = \boldsymbol{V}_i^{(k+1/2)}\boldsymbol{\gamma}_i^{(k)}(\boldsymbol{\gamma}_i^{(k)^T}\boldsymbol{V}_i^{(k+1/2)}\boldsymbol{\gamma}_i^{(k)} + \boldsymbol{\Omega}_i^{(k)})^{-1}, \tag{20}$$

where

$$\boldsymbol{V}_i^{(k+1/2)} = \frac{\sum_{j=1}^n \tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k+1/2)})\,(\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k+1)})(\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k+1)})^T}{\sum_{j=1}^n \tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k+1/2)})}, \tag{21}$$

$$\boldsymbol{\gamma}_i^{(k)} = (\boldsymbol{B}_i^{(k)}\boldsymbol{B}_i^{(k)^T} + \boldsymbol{D}_i^{(k)})^{-1}\boldsymbol{B}_i^{(k)}, \tag{22}$$

and

$$\boldsymbol{\Omega}_i^{(k)} = \boldsymbol{I}_q - \boldsymbol{\gamma}_i^{(k)^T}\boldsymbol{B}_i^{(k)} \tag{23}$$

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

$$\begin{aligned}\boldsymbol{D}_i^{(k+1)} &= \mathrm{diag}\{\boldsymbol{V}_i^{(k+1/2)} - \boldsymbol{B}_i^{(k+1)}\boldsymbol{H}_i^{(k+1/2)}\boldsymbol{B}_i^{(k+1)^T}\} \\ &= \mathrm{diag}\{\boldsymbol{V}_i^{(k+1/2)} - \boldsymbol{V}_i^{(k+1/2)}\boldsymbol{\gamma}_i^{(k)}\boldsymbol{B}_i^{(k+1)^T}\},\end{aligned} \tag{24}$$

where

$$\begin{aligned}\boldsymbol{H}_i^{(k+1/2)} &= \frac{\sum_{j=1}^n \tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k+1/2)})\, E_i^{(k+1/2)}(\boldsymbol{U}_j\boldsymbol{U}_j^T \mid \boldsymbol{y}_j)}{\sum_{j=1}^n \tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k+1/2)})} \\ &= \boldsymbol{\gamma}_i^{(k)^T}\boldsymbol{V}_i^{(k+1/2)}\boldsymbol{\gamma}_i^{(k)} + \boldsymbol{\Omega}_i^{(k)}\end{aligned} \tag{25}$$

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

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 $\boldsymbol{D}_i$,

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

Another way is to impose constraints on the ratios of the diagonal elements of each $\boldsymbol{D}_i$; see Hathaway (1985) and Ingrassia (2004). Alternatively, one can adopt a Bayesian approach as, for example, in Fokoué and Titterington (2002) and Svensén and Bishop (2005).

Under the mixture of probabilistic component analyzers (PCAs) model as proposed by Tipping and Bishop (1997), the ith component-covariance matrix

$\boldsymbol{\Sigma}_i$ has the form (7) with each $\boldsymbol{D}_i$ now having the isotropic structure

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

Under this isotropic restriction (27), $\boldsymbol{B}_i^{(k+1)}$ and $\sigma_i^{(k+1)^2}$ are given explicitly by an eigenvalue decomposition of the current value of $\boldsymbol{V}_i$ without the need to introduce the latent factors $\boldsymbol{u}_{ij}$ as "missing" data.

We can make use of the link of factor analysis with the probabilistic PCA algorithm to specify an initial starting value for $\boldsymbol{\Psi}$; see McLachlan et al. (2003).

4 Multivariate t-distribution

The mixture of factor analyzers model is sensitive to outliers since it adopts the multivariate normal family for the distributions of the errors and the latent factors. 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 (Hennig, 2004).

Before we proceed to consider a mixture model that adopts the t-family for modelling the distribution of the component errors and also the latent factors, we give a brief account of the multivariate t-distribution. 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

$$W_j \sim \text{ gamma}\,(\tfrac{1}{2}\nu_i, \tfrac{1}{2}\nu_i), \tag{28}$$

where the gamma (α, β) density function is equal to

$$f_{\mathrm{G}}(w; \alpha, \beta) = \{\beta^\alpha w^{\alpha-1}/\Gamma(\alpha)\} \exp(-\beta w) I_{[0,\infty)}(w) \qquad (\alpha, \beta > 0), \tag{29}$$

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

$$\boldsymbol{Y}_j \mid w_j \sim N_p(\boldsymbol{\mu}_i,\, \boldsymbol{\Sigma}_i/w_j), \tag{30}$$

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

$$\boldsymbol{Y}_j \sim t_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \nu_i), \tag{31}$$

and we let $f_{\rm t}(\boldsymbol{y}_j;\, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \nu_i)$ denote the corresponding density; see, for example, McLachlan and Peel (2000a, Chapter 7) and Kotz and Nadarajah (2004). 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.

McLachlan and Peel (2000a, Chapter 7) and Peel and McLachlan (2000) have considered the fitting of mixtures of t-components via the ECM algorithm. In this framework, the unobservable random variables W_j defined by (29) are introduced as "missing data." Thus on the E-step, their conditional expectation given the observed data has to be computed, using the current estimate for $\boldsymbol{\Psi}$. 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

$$w_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}) = \frac{\nu_i + p}{\nu_i + \delta(\boldsymbol{y}_j,\, \boldsymbol{\mu}_i;\, \boldsymbol{\Sigma}_i)} \tag{32}$$

and where

$$\delta(\boldsymbol{y}_j,\, \boldsymbol{\mu}_i;\, \boldsymbol{\Sigma}_i) = (\boldsymbol{y}_j - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{y}_j - \boldsymbol{\mu}_i) \tag{33}$$

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

5 Formulation of mixture of t-factor analyzers model

McLachlan and Peel (1998, 2000a) replaced the multivariate normal component distributions in normal mixture models by multivariate t-distribution

components in an attempt to make the model more robust to outliers. We now follow their approach in the present context with factor analytic components with arbitrary component-diagonal matrices $\boldsymbol{D}_i$. Zhao and Jiang (2006) have independently considered this problem in the special case of spherical $\boldsymbol{D}_i$.

We now formulate our mixture of t-analyzers model by replacing the multivariate normal distribution in (10) for the ith component-conditional distribution of $\boldsymbol{Y}_j$ by the multivariate t-distribution with mean vector vector $\boldsymbol{\mu}_i$, scale matrix $\boldsymbol{\xi}_i$, and ν_i degrees with the factor analytic restriction (7) on the 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

$$f(\boldsymbol{y}_j; \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i f_t(\boldsymbol{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \nu_i), \tag{34}$$

where

$$\boldsymbol{\Sigma}_i = \boldsymbol{B}_i \boldsymbol{B}_i^T + \boldsymbol{D}_i \quad (i = 1, \ldots, g) \tag{35}$$

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 (34) with the restriction (35), it is computationally convenient to exploit its link with factor analysis. Accordingly, corresponding to (6), we assume that

$$\boldsymbol{Y}_j = \boldsymbol{\mu}_i + \boldsymbol{B}_i \boldsymbol{U}_{ij} + \boldsymbol{e}_{ij} \qquad \text{with prob. } \pi_i \quad (i = 1, \ldots, g) \tag{36}$$

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 (34) for the marginal distribution of $\boldsymbol{Y}_j$.

From (10), we have for the usual factor analysis model that conditional on membership of the ith component of the mixture the joint distribution of $\boldsymbol{Y}_j$ and its associated factor (vector) $\boldsymbol{U}_{ij}$ is multivariate normal,

$$\begin{pmatrix} \boldsymbol{Y}_j \\ \boldsymbol{U}_{ij} \end{pmatrix} \mid z_{ij} = 1 \ \sim N_{p+q}(\boldsymbol{\mu}_i^*, \boldsymbol{\xi}_i) \quad (i = 1, \ldots, g). \tag{37}$$

where the mean $\boldsymbol{\mu}_i^*$ and the covariance matrix $\boldsymbol{\xi}_i$ are defined by (11) and

(12). We now replace the normal distribution by the t-distribution in (37) to postulate that

$$\begin{pmatrix} \boldsymbol{Y}_j \\ \boldsymbol{U}_{ij} \end{pmatrix} \mid z_{ij} = 1 \ \sim t_{p+q}(\boldsymbol{\mu}_i^*, \boldsymbol{\xi}_i, \nu_i) \quad (i = 1, \ldots, g). \tag{38}$$

This specification of the joint distribution of $\boldsymbol{Y}_j$ and its associated factors in (36) will imply the t-mixture model (34) for the marginal distribution of $\boldsymbol{Y}_j$ with the restriction (35) on its component-scale matrices.

Using the characterization of the t-distribution discussed in Section 4, it follows that we can express (37) alternatively as

$$\begin{pmatrix} \boldsymbol{Y}_j \\ \boldsymbol{U}_{ij} \end{pmatrix} \mid w_j, z_{ij} = 1 \ \sim N_{p+q}(\boldsymbol{\mu}_i^*, \boldsymbol{\xi}_i/w_j), \tag{39}$$

where w_{ij} is a value of the weight variable W_j taken to have the gamma distribution (29). It can be established from (39) that

$$\boldsymbol{U}_{ij} \mid w_j, z_{ij} = 1 \ \sim N_q(\boldsymbol{0}, \boldsymbol{I}_q/w_j) \tag{40}$$

and

$$\boldsymbol{e}_{ij} \mid w_j, z_{ij} = 1 \ \sim N_p(\boldsymbol{0}, \boldsymbol{D}_i/w_j), \tag{41}$$

and hence that

$$\boldsymbol{U}_{ij} \mid z_{ij} = 1 \ \sim t_q(\boldsymbol{0}, \boldsymbol{I}_q, \nu_i) \tag{42}$$

and

$$\boldsymbol{e}_{ij} \mid z_{ij} = 1 \ \sim t_p(\boldsymbol{0}, \boldsymbol{D}_i, \nu_i). \tag{43}$$

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 (39) that conditional on w_j, $\boldsymbol{U}_{ij}$ and $\boldsymbol{e}_{ij}$ are uncorrelated, and hence, unconditionally uncorrelated.

6 An AECM algorithm for mixtures of t-factor analyzers

We can use maximum likelihood to provide an estimator of the vector of unknown parameters in the mixture of t-factor analyzers model specified by (34) and (35). We use the AECM algorithm as outlined in Section 3 for mixtures of factor analyzers. The results as outlined in McLachlan and Peel (2000, Section 3.8) on the consistency of the ML estimator in the case of normal mixture components should carry over here if the adopted factor analysis model holds true for the component distributions.

More specifically, we declare the missing data to be the component-indicators z_{ij}, the factors $\boldsymbol{u}_{ij}$ in (36), and the weights w_j in the characterization (39) of the t-distribution for the ith component distribution of $\boldsymbol{Y}_j$ and $\boldsymbol{U}_{ij}$. We have from (39) that

$$\boldsymbol{Y}_{ij} \mid \boldsymbol{u}_{ij}, w_j, z_{ij} = 1 \ \sim N_p(\boldsymbol{\mu}_i + \boldsymbol{B}_i \boldsymbol{u}_{ij},\ \boldsymbol{D}_i/w_j) \tag{44}$$

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 $\boldsymbol{u}_{ij}$. The complete-data log likelihood for $\boldsymbol{\Psi}$ formed on the basis of the complete data is given by

$$\log L_{\rm c}(\boldsymbol{\Psi}) = \sum_{i=1}^{g} \sum_{j=1}^{n} z_{ij} \log a_{ij} \tag{45}$$

where

$$a_{ij} = \pi_i f_{\rm G}(w_j; \tfrac{1}{2}\nu_i, \tfrac{1}{2}\nu_i) \phi(\boldsymbol{u}_{ij}; \boldsymbol{0}, \boldsymbol{I}_q/w_j) \phi(\boldsymbol{y}_j;\ \boldsymbol{\mu}_i + \boldsymbol{B}_i \boldsymbol{u}_{ij},\ \boldsymbol{D}_i/w_j). \tag{46}$$

From (46), $\log a_{ij}$ can be expressed as

$$\log a_{ij} = \sum_{h=1}^{4} a_{hij}, \tag{47}$$

where

$$a_{1ij} = \log \pi_i, \tag{48}$$

$$a_{2ij} = -\log \Gamma(\tfrac{1}{2}\nu_i) + \tfrac{1}{2}\nu_i \log(\tfrac{1}{2}\nu_i) + \tfrac{1}{2}\nu_i(\log w_j - w_j) - \log w_j, \tag{49}$$

$$a_{3ij} = -\tfrac{1}{2}q\log(2\pi) + \tfrac{1}{2}q\log w_j + \tfrac{1}{2}\boldsymbol{u}_{ij}^T\boldsymbol{u}_{ij}/w_j, \tag{50}$$

and

$$\begin{aligned} a_{4ij} = &-\tfrac{1}{2}p\log(2\pi) - +\tfrac{1}{2}p\log w_j - \tfrac{1}{2}\log|\boldsymbol{D}_i| \\ &-\tfrac{1}{2}w_j(\boldsymbol{y}_j - \boldsymbol{\mu}_i - \boldsymbol{B}_i\boldsymbol{u}_{ij})^T\boldsymbol{D}_i^{-1}(\boldsymbol{y}_j - \boldsymbol{\mu}_i - \boldsymbol{B}_i\boldsymbol{u}_{ij}). \end{aligned} \tag{51}$$

6.1 E-step

It can be seen from (51) that in order to carry out the E-step, we need to be able to calculate the conditional expectation of terms like

$$Z_{ij}W_j\boldsymbol{U}_{ij} \tag{52}$$

and

$$Z_{ij}W_j\boldsymbol{U}_{ij}\boldsymbol{U}_{ij}^T. \tag{53}$$

From (39), we have that conditional on $\boldsymbol{y}_j$ and w_j, the ith component-conditional distribution of $\boldsymbol{U}_{ij}$ is multivariate normal with mean

$$\boldsymbol{\gamma}_i^T(\boldsymbol{y}_j - \boldsymbol{\mu}_i) \tag{54}$$

and covariance matrix $\boldsymbol{\Omega}_i/w_j$, where $\boldsymbol{\gamma}_i$ and $\boldsymbol{\Omega}_i$ are defined by (14) and (15).

We can now use (39) and (54) to compute the required conditional expectations (52) and (53). It follows that conditional on w_j and $z_{ij} = 1$

$$E\{Z_{ij}W_j\boldsymbol{U}_{ij} \mid \boldsymbol{y}_j, w_j\} = w_j\boldsymbol{\gamma}_i^T(\boldsymbol{y}_j - \boldsymbol{\mu}_i) \tag{55}$$

and

$$E\{Z_{ij}W_j\boldsymbol{U}_{ij}\boldsymbol{U}_{ij}^T \mid \boldsymbol{y}_j, w_j\} = \boldsymbol{\Omega}_i + w_j\boldsymbol{\gamma}_i^T(\boldsymbol{y}_j - \boldsymbol{\mu}_i)(\boldsymbol{y}_j - \boldsymbol{\mu}_i)^T\boldsymbol{\gamma}_i \tag{56}$$

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

The conditional expectation of W_j given $\boldsymbol{y}_j$ and $z_{ij} = 1$ is given by (32), and the conditional expectation of Z_{ij} given $\boldsymbol{y}_j$ is given by the posterior probability that $\boldsymbol{y}_j$ belongs to the ith component of the mixture. This posterior

probability can be expressed as in (5) on replacing the multivariate normal density $\phi(\boldsymbol{y}_j;\, \boldsymbol{\mu}_i, \boldsymbol{\Psi})$ by the multivariate t-density $f(\boldsymbol{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \nu_i)$. That is,

$$\tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}) = \frac{\pi_i f_{\rm t}(\boldsymbol{y}_j;\, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \nu_i)}{\sum_{h=1}^{g} \pi_h f_{\rm t}(\boldsymbol{y}_j;\, \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h, \nu_h)} \quad (i = 1, \ldots, g; j = 1, \ldots, n), \tag{57}$$

where now the vector of parameters $\boldsymbol{\Psi}$ consists of the mixing proportions, the elements of $\boldsymbol{\mu}_i, \boldsymbol{B}_i$, and of $\boldsymbol{D}_i$, and the degrees of freedom $\nu_i\, (i = 1, \ldots, g)$.

Using the results (56) and (57), it is straightforward to show that an AECM algorithm can be formulated as in Section 3 to iteratively fit the mixture of t-factor analyzers model as specified by (35) and (37).

6.2 CM-steps

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 (39) 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 (18), where now the posterior probabilities are calculated using (57). The updated estimate of the ith component mean $\boldsymbol{\mu}_i$ is given by

$$\boldsymbol{\mu}_i^{(k+1)} = \sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k)}) w_{ij}^{(k)} \boldsymbol{y}_j / \sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k)}) w_{ij}^{(k)}, \tag{58}$$

where the current weight $w_{ij}^{(k)}$ is formed using the current value $\boldsymbol{\Psi}^{(k)}$ for $\boldsymbol{\Psi}$ in (32).

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

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

$$+\psi\left(\frac{\nu_i^{(k)}+p}{2}\right)-\log\left(\frac{\nu_i^{(k)}+p}{2}\right)\Bigg\}=0, \tag{59}$$

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

$$\boldsymbol{\Psi}^{(k+1/2)} = (\boldsymbol{\Psi}_1^{(k+1)^T},\, \boldsymbol{\Psi}_2^{(k)^T})^T. \tag{60}$$

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 (20) to (25), but where the ith component sample covariance matrix is calculated as

$$\boldsymbol{V}_i^{(k+1/2)} = \frac{\sum_{j=1}^n \tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k+1/2)})\, w_{ij}^{(k+1/2)}(\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k+1)})(\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k+1)})^T}{\sum_{j=1}^n \tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k+1/2)})}, \tag{61}$$

where $w_{ij}^{(k+1/2)}$ is updated partially by using $\boldsymbol{\Psi}^{(k+1/2)}$ for $\boldsymbol{\Psi}$ in (32).

7 Clustering of microarray gene-expression data

As an example of the use of mixtures of t-factor analyzers to cluster high-dimensional data, we consider the clustering of 234 tissue samples on the basis of 70 genes. van't Veer et al. (2002) developed a set of 70 marker genes to predict the clinical outcome of breast-cancer patients. van de Vijver (2002) used these genes to study a set of 295 patients; 234 of these patients were not examined in the van't Veer et al study. Weigelt et al. (2005), referring to Perou et al. (2000), reported that there were at least four molecular subtypes associated with distinct patient outcomes from breast cancer. These four types were luminal, normal-like, basal-like, and HER2+, having different metastatic properties and outcomes. We wish to investigate the claims of Weigelt at al. (2005) and Perou et al. (2000) about the number of molecular subtypes with respect to these 234 tissue samples. But firstly, we note some results for the supervised classification of these 234 tissue samples.

In the study of van de Vijver et al. (2002), these 234 tissues were assigned to a good-prognosis class G_1 and a poor-prognosis class G_2, using a classifier

Table 1
Confusion matrix for van de Vijver classification versus our classification using an SVM

	van de Vijver	
	G_1	G_2
SVM G_1	75	7
SVM G_2	18	134

based on 61 lymph-node negative tissues from an earlier study of van't Veer et al. (2002); 93 tissues were assigned to G_1 and 141 tissues to G_2. The rule was essentially equivalent to the nearest-centroid classifier, as it was based on the correlation between the feature vector (the gene-signature vector) and the mean of the good-prognosis class G_1. We also classified these 234 tissues, using a support vector machine (SVM) based on the same 61 tissue samples from van't Veer et al. (2002). With its application, 18 of the 93 tissues assigned to the class G_1 were put in the class G_2, while 7 of the 141 tissues assigned to G_2 were put in G_1. These results are shown in the confusion matrix in Table 1. We examined the 18 tissues classified by our SVM as being in the bad-prognosis class, but in the good-prognosis class by van de Vijver et al. (2002). Of the 18 tissues, four have an outcome of distant metastasis during the follow up, two of which had distant metastases within five years and these two died. The other two were still alive at the time of follow up (about nine years later). Next, we examined the seven tissues classified by our SVM as in the good-prognosis class but by van de Vijver et al. (2002) as in the bad-prognosis class. Of these seven, five of these had no distant metastases. Of the other two with distant metastases, both were still alive at the last follow up (one at more than eight years, the other at only two years).

We now consider the clustering of these 234 tissues without use of the 61 tissues used by van de Vijver et al. (2002). We first worked with all $p = 70$ genes and fitted a g-component normal mixture model, with g ranging from 2 to 4, with the restriction that the component-covariance matrices Σ_i are equal. To assess the number of components g to be used in this normal mixture model, we used BIC (the Bayesian information criterion) of Schwarz (1978). With this criterion, we compare twice the increase in the log likelihood, that is, $-2\log\lambda$ where λ is the likelihood ratio statistic, to $d\log n$, where d is the increase in the number of parameters fitted in proceeding from g to $g+1$ components. Using BIC, we concluded that there was not sufficient support in the data to reject $g = 2$ components in favour of $g = 3$; likewise for rejecting $g = 3$ in favour of $g = 4$. For example, for $g = 2$ versus $g = 3$, $-2\log\lambda$ is 224.9, which is less than $d\log n = 71 \times 5.46 = 388$. The $g = 2$ clustering puts 121 tissues in one cluster and 113 in the other. Compared with the (supervised) classification of these tissues from the support vector machine, the larger cluster contained 72

Table 2
Confusion matrix for our classification using mixture of normals with equal covariance matrices versus our classification using an SVM

	mixture of normals	
	G_1	G_2
SVM G_1	72	10
SVM G_2	49	103

from the good-prognosis class and the smaller cluster contained 103 from the bad-prognosis class. These results are given in a confusion matrix in Table 2.

Although we can fit mixtures of normal components with no numerical difficulties due to the imposition of the restriction of equal component-covariance matrices $\Sigma_i = \Sigma (i = 1, \ldots, g)$, the number of distinct parameters in Σ is still very large relative to the sample size of $n = 234$. Also, there is no reason why the homoscedasticity should hold. Hence we decided to explore these data further for group structure without the restriction of homoscedasticity. However, we are unable to fit a mixture of normal distributions with unrestricted component-covariance matrices for $p = 70$ variables (genes) as there will be problems with singular or near-singular estimates of the Σ_i. Accordingly, we considered the fitting of mixtures of g factor analyzers. We adopted $q = 6$ factors for this purpose of exploring the data for group structure.

As before, we used BIC to assess the number g of components (factor analyzers) to be used in the mixture model. We also used this criterion to assess the number q of factors to be used for a given choice of g. Regularity conditions hold for this latter testing problem provided g is fixed (McLachlan, Chapter 8, 2000). Application of this criterion here for $g = 2, 3$, and 4 suggests that it is reasonable to use $q = 6$ factors, although its use interpreted rigidly would lead to values of q up to $q = 10$. But we settled on $q = 6$ factors, bearing in mind that the sample size is limited relative to the number of parameters that would be needed in a model with $q = 10$.

On fitting mixtures of g factor analyzers with $q = 6$ factors, we proceeded to see if we could find evidence in the data to support (according to BIC) using $g = 3$ or 4 over $g = 2$ components. The component-covariance matrices were unconstrained apart from the assumptions that the diagonal matrices $\boldsymbol{D}_i$ were equal; that is, $\boldsymbol{D}_i = \boldsymbol{D}$. As in the case of equal component-covariance matrices we could find no evidence in support of $g = 3$ or 4 components over $g = 2$.

The clustering according to a mixture of $g = 2$ factor analyzers with $q = 6$ puts 181 in the larger cluster and 53 in the smaller cluster. The cardinalities of the

clusters are found to be quite different from that obtained using mixtures of normals with equal component-covariance matrices. When we identified these two clusters with the classification into good- and bad-prognosis classes using SVM, we found that the larger cluster contains 81 of the 82 tissues assigned to the good-prognosis class. We also fitted a mixture of $g = 2$ t-factor analyzers, obtaining one cluster with 156 tissues and the other with 78. The former contained the 82 tissues classified to the good-prognosis class by the SVM. The difference between the clusterings obtained with the use of t- instead of ordinary factor analyzers indicates that there are tissues with atypical gene expressions. These tissues can be identified by those that have small weights with respect to both components in the mixture model. The biological significance of these atypical gene expressions is still under investigation.

8 Acknowledgements

The authors would like to thank the editors, associate editor, and the reviewers for their very helpful suggestions for the revision of the paper.

9 References

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

Bishop, C.M., 1998. Latent variable models. In: Jordan, M.I. (Ed.), Learning in Graphical Models. Kluwer, Dordrecht, 371-403.

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

Coleman, D., Dong, X., Hardin, J., Rocke, D.M., and Woodruff, D.L., 1999. Some computational issues in cluster analysis with no a priori metric. Computational Statistics & Data Analysis, 31, 1-11.

Dempster, A.P., Laird, N.M., and Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistics Society B, 39, 1-38.

Everitt, B.S. and Hand, D.J., 1981. Finite Mixture Distributions. Chapman & Hall, London.

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

Hathaway, R.J., 1985. A constrained formulation of maximum-likelihood estimation for normal mixture distributions, Annals of Statistics, 13, 795-800.

Hennig, C., 2004. Breakdown points for maximum likelihood-estimators of location-scale mixtures, Annals of Statistics, 32, 1313-1340.

Ingrassia, S., 2004. A likelihood-based constrained algorithm for multivariate normal mixture models. Statistical Methods and Applications, 13, 151-166.

Kotz, S. and Nadarajah S., 2004. Multivariate t-distributions and their applications. Cambridge University Press, New York.

Lindsay, B.G., 1995. 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.

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

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

McLachlan, G.J. and Peel, D., 1998. Robust cluster analysis via mixtures of multivariate t-distributions. In: A. Amin, D. Dori, P. Pudil, and H. Freeman (Eds.), Lecture Notes in Computer Science, Vol. 1451, Springer-Verlag, Berlin, 658-666.

McLachlan, G.J. and Peel, D., 2000a. Finite Mixture Models. Wiley, New York.

McLachlan, G.J. and Peel, D., 2000b. Mixtures of factor analyzers. In: Langley, P. (Ed.), Proceedings of the Seventeenth International Conference on Machine Learning. Morgan Kaufmann, San Francisco, 599-606.

McLachlan, G.J., Bean, R.W., and Peel, D., 2002. A mixture model-based approach to the clustering of microarray expression data. Bioinformatics, 18(3) 413-422.

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

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

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

Peel, D. and McLachlan, G.J., 2000. Robust mixture modelling using the t distribution. Statistics and Computing, 10, 335-344.

Schwarz, G., 1978. Estimating the dimension of a model. Annals of Statistics, 6, 461-464.

Svensén, M. and Bishop, C.M., 2005. Robust Bayesian mixture modelling, Neurocomputing, 64, 234-252.

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

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

van de Vijver, M.J., He, Y.D., van't Veer, L.J., et al., 2002. A gene-expression signature as a predictor of survival in breast cancer. New England Journal of Medicine 347, 1999-2009.

van't Veer, L.J., Dai, H., van de Vijver, M.J., et al. 2002. Gene expression profiling predicts clinical outcome of breast cancer. Nature 415, 530-536.

Zhao, J. and Jiang, Q., 2006. Probabilistic PCA for t distributions. Neurocomputing. To appear.