# Mixtures of Factor Analyzers with Common Factor Loadings for the Clustering and Visualisation of High-Dimensional Data

Jangsun Baek[1] and Geoffrey J. McLachlan[2]

[1]Department of Statistics, Chonnam National University, Gwangju 500-757, South Korea

and

[2]Department of Mathematics and Institute for Molecular Bioscience, University of Queensland, Brisbane QLD 4072, Australia

**Abstract** – Mixtures of factor analyzers enable model-based density estimation and clustering to be undertaken for high-dimensional data, where the number of observations $n$ is very large relative to their dimension $p$. In practice, there is often the need to reduce further the number of parameters in the specification of the component-covariance matrices. To this end, we propose the use of common component-factor loadings, which considerably reduces further the number of parameters. Moreover, it allows the data to be displayed in low-dimensional plots.

**Keywords** – Normal mixture models, mixtures of factor analyzers, common factor loadings, model-based clustering .

# 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, [1]. 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 $\pi_i$, the elements of the component means $\boldsymbol{\mu}_i$, and the distinct elements of the component-covariance matrices $\boldsymbol{\Sigma}_i\, (i = 1, \ldots, g)$.

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 [2]; see also [3].

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 $j$th feature vector with observed value $\boldsymbol{y}_j$ belongs to the $i$th component of the mixture can be expressed by Bayes' theorem as

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

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.

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 [4] 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 $n$.

In this paper, we focus on the use of mixtures of factor analyzers to reduce the number of parameters in the specification of the component-covariance matrices, as discussed in [1, 5, 6]; see also [7]. With the factor-analytic representation of the component-covariance matrices, we have that

$$\boldsymbol{\Sigma}_i = \boldsymbol{B}_i \boldsymbol{B}_i^T + \boldsymbol{D}_i \quad (i = 1, \ldots, g), \tag{6}$$

where $\boldsymbol{B}_i$ is a $p \times q$ matrix and $\boldsymbol{D}_i$ is a diagonal matrix. As $\frac{1}{2}q(q-1)$ constraints are needed for $\boldsymbol{B}_i$ to be uniquely defined, the number of free parameters in (6) is

$$pq + p - \tfrac{1}{2}q(q-1). \tag{7}$$

Thus with this representation (6), the reduction in the number of parameters for $\boldsymbol{\Sigma}_i$ is

$$
\begin{aligned}
r &= \tfrac{1}{2}p(p+1) - pq - p + \tfrac{1}{2}q(q-1) \\
&= \tfrac{1}{2}\{(p-q)^2 - (p+q)\},
\end{aligned} \tag{8}
$$

assuming that $q$ is chosen sufficiently smaller than $p$ so that this difference is positive. The total number of parameters is

$$d_1 = (g-1) + 2gp + g\{pq - \tfrac{1}{2}q(q-1)\}. \tag{9}$$

We shall refer to this approach as MFA (mixtures of factor analyzers).

Even with this MFA approach, the number of parameters still might not be manageable, particularly if the number of dimensions $p$ is large and/or the number of components (clusters) $g$ is not small.

In this paper, we therefore consider how this factor-analytic approach can be modified to provide a greater reduction in the number of parameters. We extend the model of [8, 9] to propose the normal mixture model (2) with the restrictions

$$\boldsymbol{\mu}_i = \boldsymbol{A}\boldsymbol{\xi}_i \quad (i = 1, \ldots, g) \tag{10}$$

and

$$\boldsymbol{\Sigma}_i = \boldsymbol{A}\boldsymbol{\Omega}_i \boldsymbol{A}^T + \boldsymbol{D} \quad (i = 1, \ldots, g), \tag{11}$$

where $\boldsymbol{A}$ is a $p \times q$ matrix, $\boldsymbol{\xi}_i$ is a $q$-dimensional vector, $\boldsymbol{\Omega}_i$ is a $q \times q$ positive definite symmetric matrix, and $\boldsymbol{D}$ is a diagonal $p \times p$ matrix. As to be made more precise in the next section, $\boldsymbol{A}$ is a matrix of loadings on $q$ unobservable factors and its $p$ columns are taken to be orthonormal; that is,

$$\boldsymbol{A}^T \boldsymbol{A} = \boldsymbol{I}_q, \tag{12}$$

where $\boldsymbol{I}_q$ is the $q \times q$ identity matrix. With the restrictions (10) and 11) on the component mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$, respectively, the total number of parameters is reduced to

$$d_2 = (g-1) + p + q(p+g) + \tfrac{1}{2}(g-1)q(q+1). \tag{13}$$

3

We shall refer to this approach as MCFA (mixtures of common factor analyzers) since it is formulated via the adoption of a common matrix ($\boldsymbol{A}$) for the component-factor loadings. We shall show for this approach how the EM algorithm can be implemented to fit this normal mixture model under the constraints (10) and (11). We shall also illustrate how it can be used to provide lower-dimensional plots of the data $\boldsymbol{y}_j$ ($j = 1, \ldots, n$). It provides an alternative to canonical variates which are calculated from the clusters under the assumption of equal component-covariance matrices.

# 2  Mixtures of Common Factor Analyzers (MCFA)

In this section, we examine the motivation underlying the MCFA approach with its constraints (10) and (11) on the $g$ component means and covariance matrices $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ ($i = 1, \ldots, g$). We shall show that it can be viewed as a special case of the MFA approach.

To see this we first note that the MFA approach with the factor-analytic representation (6) on $\boldsymbol{\Sigma}_i$ is equivalent to assuming that the distribution of the difference $\boldsymbol{Y}_j - \boldsymbol{\mu}_i$ can be modelled as

$$\boldsymbol{Y}_j - \boldsymbol{\mu}_i = \boldsymbol{B}_i \boldsymbol{U}_{ij} + \boldsymbol{e}_{ij} \qquad \text{with prob. } \pi_i \quad (i = 1, \ldots, g) \tag{14}$$

for $j = 1, \ldots, n$, where the (unobservable) 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 ($i = 1, \ldots, g$).

As noted in the introductory section, this model may not lead to a sufficiently large enough reduction in the number of parameters, particularly if $g$ is not small. Hence if this is the case, we propose the MCFA approach whereby the distribution of $\boldsymbol{Y}_j$ is modelled as

$$\boldsymbol{Y}_j = \boldsymbol{A} \boldsymbol{U}_{ij} + \boldsymbol{e}_{ij} \qquad \text{with prob. } \pi_i \quad (i = 1, \ldots, g) \tag{15}$$

for $j = 1, \ldots, n$, where the (unobservable) factors $\boldsymbol{U}_{i1}, \ldots, \boldsymbol{U}_{in}$ are distributed independently $N(\boldsymbol{\xi}, \boldsymbol{\Omega}_i)$, independently of the $\boldsymbol{e}_{ij}$, which are distributed independently $N(\boldsymbol{0}, \boldsymbol{D})$, where $\boldsymbol{D}$ is a diagonal matrix ($i = 1, \ldots, g$). Here $\boldsymbol{A}$ is a $p \times q$ matrix of factor loadings, which we take to satisfy the relationship (12); that is, $\boldsymbol{A}^T \boldsymbol{A} = \boldsymbol{I}_q$.

To see that the MCFA model as specified by (15) is a special case of the MFA approach as specified by (14), we note that we can rewrite (15) as

$$\begin{aligned}
\boldsymbol{Y}_j &= \boldsymbol{A} \boldsymbol{U}_{ij} + \boldsymbol{e}_{ij} \\
&= \boldsymbol{A} \boldsymbol{\xi}_i + \boldsymbol{A}(\boldsymbol{U}_{ij} - \boldsymbol{\xi}_i) + \boldsymbol{e}_{ij} \\
&= \boldsymbol{\mu}_i + \boldsymbol{A} \boldsymbol{K}_i \boldsymbol{K}_i^{-1}(\boldsymbol{U}_{ij} - \boldsymbol{\xi}_i) + \boldsymbol{e}_{ij} \\
&= \boldsymbol{\mu}_i + \boldsymbol{B}_i \boldsymbol{U}_{ij}^* + \boldsymbol{e}_{ij},
\end{aligned} \tag{16}$$

where

$$\begin{aligned}
\boldsymbol{\mu}_i &= \boldsymbol{A} \boldsymbol{\xi}_i, \tag{17} \\
\boldsymbol{B}_i &= \boldsymbol{A} \boldsymbol{K}_i, \tag{18} \\
\boldsymbol{U}_{ij}^* &= \boldsymbol{K}_i^{-1}(\boldsymbol{U}_{ij} - \boldsymbol{\xi}_i), \tag{19}
\end{aligned}$$

and where the $\boldsymbol{U}_{ij}^*$ are distributed independently $N(\boldsymbol{0}, \boldsymbol{I}_q)$. The covariance matrix of $\boldsymbol{U}_{ij}^*$ is equal to $\boldsymbol{I}_q$, since $\boldsymbol{K}_i$ can be been chosen so that

$$\boldsymbol{K}_i^{-1}\boldsymbol{\Omega}_i\boldsymbol{K}_i^{-1^T} = \boldsymbol{I}_q \quad (i = 1, \ldots, g). \tag{20}$$

On comparing (16) with (14), it can be seen that the MCFA model is a special case of the MFA model with the additional restrictions that

$$\boldsymbol{\mu}_i = \boldsymbol{A}\boldsymbol{\xi}_i \quad (i = 1, \ldots, g), \tag{21}$$

$$\boldsymbol{B}_i = \boldsymbol{A}\boldsymbol{K}_i \quad (i = 1, \ldots, g), \tag{22}$$

and

$$\boldsymbol{D}_i = \boldsymbol{D} \quad (i = 1, \ldots, g). \tag{23}$$

The latter restriction of equal diagonal covariance matrices for the component-specific error terms $(\boldsymbol{D}_i = \boldsymbol{D})$ is sometimes imposed with applications of the MFA approach to avoid potential singularities with small clusters (see [5]).

Concerning the restriction (22) that the matrix of factor of loadings is equal to $\boldsymbol{A}\boldsymbol{K}_i$ for each component, it can be viewed as adopting common factor loadings before the use of the transformation $\boldsymbol{K}_i$ to transform the factors so that they have unit variances and zero covariances. Hence this is why we call this approach mixtures of common factor analyzers. It is also different to the MFA approach in that it considers the factor-analytic representation of the observations $\boldsymbol{Y}_j$ directly, rather than the error terms $\boldsymbol{Y}_j - \boldsymbol{\mu}_i$.

As the MFA approach allows a more general representation of the component-covariance matrices and places no restrictions on the component means it is in this sense preferable to the MCFA approach if its application is feasible given the values of $p$ and $g$. If the dimension $p$ and/or the number of components $g$ is too large, then the MCFA provides a more feasible approach at the expense of more distributional restrictions on the data. In empirical results some of which are to be reported in the sequel we have found the performance of the MCFA approach is usually at least comparable to the MFA approach for data sets to which the latter is practically feasible. The MCFA approach also has the advantage in that the latent factors in its formulation are allowed to have different means and covariance matrices and are not white noise as with the formulation of the MFA approach. Thus the (estimated) posterior means of the factors corresponding to the observed data can be used to portray the latter in low-dimensional spaces.

# 3   Some Related Approaches

The MCFA approach is similar in form to the approach proposed by Yoshida et al. [8, 9], who also imposed the additional restrictions that the common diagonal covariance matrix $\boldsymbol{D}$ of the error terms is spherical,

$$\boldsymbol{D} = \sigma^2 \boldsymbol{I}_p, \tag{24}$$

Table 1: The number of parameters for three models

|  | $p$ | $g$ | $q$ | Number of parameters |
|---|---|---|---|---|
| MFA | 1000 | 2 | 2 | 7999 |
|  | 1000 | 4 | 2 | 15999 |
|  | 5000 | 2 | 2 | 39999 |
|  | 5000 | 4 | 2 | 79999 |
| MCFA | 1000 | 2 | 2 | 3008 |
|  | 1000 | 4 | 2 | 3020 |
|  | 5000 | 2 | 2 | 15008 |
|  | 5000 | 4 | 2 | 15020 |
| MCUFSA | 1000 | 2 | 2 | 2007 |
|  | 1000 | 4 | 2 | 2017 |
|  | 5000 | 2 | 2 | 10007 |
|  | 5000 | 4 | 2 | 10017 |

and that the component-covariance matrices of the factors are diagonal. We shall call this approach MCUFSA (mixtures of common uncorrelated factor spherical-error analyzers). The total number of parameters with this approach is

$$d_3 = (g - 1) + pq + 1 + 2gq - \tfrac{1}{2}q(q + 1). \tag{25}$$

In our experience, we have found that these restrictions of sphericity of the errors and of diagonal covariance matrices in the component distributions of the factors can have an adverse effect on the clustering of high-dimensional data sets. The relaxation of these restrictions does considerably increase the complexity of the problem of fitting the model. But we shall show how it can be effected via the EM algorithm with the E- and M-steps being able to be carried out in closed form.

In Table 1, we have listed the number of parameters to be estimated for the MFA, MCFA, and MCUFSA models when $p = 1000, 5000$; $q = 2$; and $g = 2, 4$. For example, when we cluster $p = 1000$ dimensional gene expression data into $g = 2$ groups using $q = 2$ dimensional factors, the MFA model requires 7999 parameters to be estimated, while the MCUFSA needs only 2007 parameters. Moreover, as the number of clusters grows from 2 to 4 the number of parameters for the MFA model grows twice as large as before, but that for MCUFSA remains almost the same. As MCUFSA needs less parameters to characterize the structure of the clusters, it does not always provide a good fit. It may fail to fit the data unless the directions of the cluster-error distributions are parallel to the axes of the original feature space due the condition of sphericity on the cluster errors. Also, it is assuming that the component-covariance matrices of the factors are diagonal.

Recently, Sanguinetti [10] has considered a method of dimensionality reduction in a cluster analysis context. However, its underlying model assumes sphericity in the specification of the variances/covariances of the factors in each cluster. Our proposed method allows for oblique factors, which provides the extra flexibility needed to cluster more effectively high-dimensional data sets in practice.

# 4   Fitting of Factor-Analytic Models

The fitting of mixtures of factor analyzers as with the MFA approach has been considered in [5], using a variant of the EM algorithm known as the alternating expectation-conditional maximization algorithm (AECM). With the MCFA approach, we have fit to the same mixture model of factor analyzers but with the additional restrictions (10)and (11) on the component means $\boldsymbol{\mu}_i$ and covariance matrices $\boldsymbol{\Sigma}_i$. We also have to impose the restriction (23) of common diagonal covariance matrices $\boldsymbol{D}$. The implementation of the EM algorithm for this model is described in the Appendix. In the EM framework, the component label $\boldsymbol{z}_j$ associated with the observation $\boldsymbol{y}_j$ is introduced as missing data, where $z_{ij} = (\boldsymbol{z}_j)_i$ is one or zero according as $\boldsymbol{y}_j$ belongs or does not belong to the $i$th component of the mixture $(i = 1, \ldots, g; j = 1, \ldots, n)$. The unobservable factors $\boldsymbol{u}_{ij}$ are also introduced as missing data in the EM framework.

As part of the E-step, we require the conditional expectation of the component labels $z_{ij}$ $(i = 1, \ldots, g)$ given the observed data point $\boldsymbol{y}_j$ $(j = 1, \ldots, n)$ . It follows that

$$
\begin{aligned}
E_{\boldsymbol{\Psi}}\{Z_{ij} \mid \boldsymbol{y}_j\} &= \mathrm{pr}_{\boldsymbol{\Psi}}\{Z_{ij} = 1 \mid \boldsymbol{y}_j\} \\
&= \tau_i(\boldsymbol{y}_j; \boldsymbol{\Psi}) \qquad (i = 1, \ldots, g; j = 1, \ldots, n),
\end{aligned} \tag{26}
$$

where $\tau_i(\boldsymbol{y}_j; \boldsymbol{\Psi})$ is the posterior probability that $\boldsymbol{y}_j$ belongs to the $i$th component of the mixture. From (5), it can be expressed under the MCFA model as

$$
\tau_i(\boldsymbol{y}_j; \boldsymbol{\Psi}) = \frac{\pi_i \phi(\boldsymbol{y}_j; \boldsymbol{A}\boldsymbol{\xi}_i, \boldsymbol{A}\boldsymbol{\Omega}_i\boldsymbol{A}^T + \boldsymbol{D})}{\sum_{h=1}^{g} \pi_h \phi(\boldsymbol{y}_j; \boldsymbol{A}\boldsymbol{\xi}_h, \boldsymbol{A}\boldsymbol{\Omega}_h\boldsymbol{A}^T + \boldsymbol{D})} \tag{27}
$$

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

We also require the conditional distribution of the unobservable (latent) factors $\boldsymbol{U}_{ij}$ given the observed data $\boldsymbol{y}_j$ $(j = 1, \ldots, n)$. The conditional distribution of $\boldsymbol{U}_{ij}$ given $\boldsymbol{y}_j$ and its membership of the $i$th component of the mixture (that is, $z_{ij} = 1$) is multivariate normal,

$$
\boldsymbol{U}_{ij} \mid \boldsymbol{y}_j, z_{ij} = 1 \ \sim N(\boldsymbol{\xi}_{ij}, \boldsymbol{\Omega}_{iy}), \tag{28}
$$

where

$$
\boldsymbol{\xi}_{ij} = \boldsymbol{\xi}_i + \boldsymbol{\gamma}_i^T(\boldsymbol{y}_j - \boldsymbol{A}\boldsymbol{\xi}_i) \tag{29}
$$

and

$$
\boldsymbol{\Omega}_{iy} = (\boldsymbol{I}_q - \boldsymbol{\gamma}_i^T\boldsymbol{A})\boldsymbol{\Omega}_i, \tag{30}
$$

and where

$$
\boldsymbol{\gamma}_i = (\boldsymbol{A}\boldsymbol{\Omega}_i\boldsymbol{A}^T + \boldsymbol{D})^{-1}\boldsymbol{A}\boldsymbol{\Omega}_i. \tag{31}
$$

We can portray the observed data $\boldsymbol{y}_j$ in $q$-dimensional space by plotting the corresponding values of the $\hat{\boldsymbol{u}}_{ij}$, which are estimated conditional expectations of the factors $\boldsymbol{U}_{ij}$, corresponding to the observed data points $\boldsymbol{y}_j$. From (28) and (29),

$$
\begin{aligned}
E(\boldsymbol{U}_{ij} \mid \boldsymbol{y}_j, z_{ij} = 1) &= \boldsymbol{\xi}_{ij} \\
&= \boldsymbol{\xi}_i + \boldsymbol{\gamma}_i^T(\boldsymbol{y}_j - \boldsymbol{A}\boldsymbol{\xi}_i). 
\end{aligned} \tag{32}
$$

We let $\hat{\boldsymbol{u}}_{ij}$ denote the value of the right-hand side of (32) evaluated at the maximum likelihood estimates of $\boldsymbol{\xi}_i, \boldsymbol{\gamma}_i$, and $\boldsymbol{A}$. We can define the estimated value $\hat{\boldsymbol{u}}_j$ of the $j$th factor corresponding to $\boldsymbol{y}_j$ as

$$\hat{\boldsymbol{u}}_j = \sum_{i=1}^{g} \tau_i(\boldsymbol{y}_j; \hat{\boldsymbol{\Psi}})\, \hat{\boldsymbol{u}}_{ij} \quad (j = 1, \ldots, n) \tag{33}$$

where, from (27), $\tau_i(\boldsymbol{y}_j; \hat{\boldsymbol{\Psi}})$ is the estimated posterior probability that $\boldsymbol{y}_j$ belongs to the $i$th component. An alternative estimate of the posterior expectation of the factor corresponding to the $j$th observation $\boldsymbol{y}_j$ is defined by replacing $\tau_i(\boldsymbol{y}_j; \hat{\boldsymbol{\Psi}})$ by $\hat{z}_{ij}$ in (33), where

$$\hat{z}_{ij} = \arg\max_{h} \tau_h(\boldsymbol{y}_j; \hat{\boldsymbol{\Psi}}). \tag{34}$$

# 5 Accuracy of Factor-Analytic Approximations

To illustrate the accuracy of the three factor-analytic approximations as defined above, we performed a small simulation experiment. We generated 100 random vectors from each of $g = 2$ different three-dimensional multivariate normal distributions. The first distribution had the mean vector $\boldsymbol{\mu}_1 = (0, 0, 0)^T$ and covariance matrix

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 4 & -1.8 & -1 \\ -1.8 & 2 & 0.9 \\ -1 & 0.9 & 2 \end{pmatrix},$$

while the second distribution had mean vector $\boldsymbol{\mu}_2 = (2, 2, 6)^T$ and covariance matrix

$$\boldsymbol{\Sigma}_2 = \begin{pmatrix} 4 & 1.8 & 0.8 \\ 1.8 & 2 & 0.5 \\ 0.8 & 0.5 & 2 \end{pmatrix}.$$

We applied the MFA, MCFA, and the MCUFSA approaches with $q = 2$ to cluster the data into two groups. We used the ArrayCluster `http://www.ism.ac.jp/~higuchi/arraycluster.htm`, which was developed by Yoshida *et al.* [9] to implement the MCUFSA approach. There were 2 misclassifications for MFA, 4 for MCFA, and 8 for MCUFSA. As we obtained the parameter estimates for each model we can also predict each observation based on the estimated factor scores and the parameter estimates. In Figures 1, 2, and 3, we have plotted the predicted observations $\hat{\boldsymbol{y}}_j$ along with the actual observations $\boldsymbol{y}_j$ by the MFA, MCFA, and the MCUFSA approaches. For the MFA approach, the predicted observation is obtained as

$$\hat{\boldsymbol{y}}_j = \sum_{i=1}^{g} \tau_i(\boldsymbol{y}_j; \hat{\boldsymbol{\Psi}})(\hat{\boldsymbol{\mu}}_i + \hat{\boldsymbol{B}}_i \hat{\boldsymbol{u}}_{ij}), \tag{35}$$

where

$$\hat{\boldsymbol{u}}_{ij} = \hat{\boldsymbol{\alpha}}_i^T (\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_i) \tag{36}$$

and

$$\hat{\boldsymbol{\alpha}}_i = (\hat{\boldsymbol{B}}_i \hat{\boldsymbol{B}}_i^T + \hat{\boldsymbol{D}}_i)^{-1} \hat{\boldsymbol{B}}_i. \tag{37}$$

8

Figure 1: Original observations and the predicted observations by MFA: □ Group 1; o Group 2; * predicted for Group 1; + predicted for Group 2
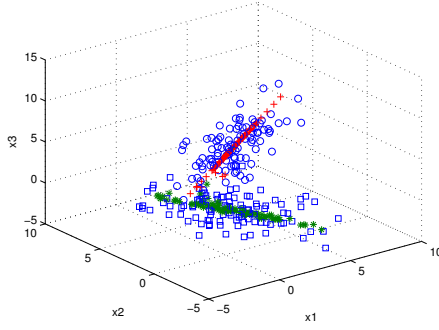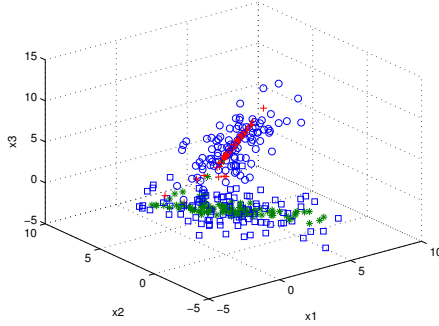


Figure 2: Original observations and the predicted observations by MCFA: □ Group 1; o Group 2; * predicted for Group 1; + predicted for Group 2

For the MCFA approach, the predicted observation is

$$\hat{\boldsymbol{y}}_j = \hat{\boldsymbol{A}}\hat{\boldsymbol{u}}_j, \tag{38}$$

where $\hat{\boldsymbol{A}}$ is the estimated projection matrix $\boldsymbol{A}$ and where $\hat{\boldsymbol{u}}_j$ is the estimated factor score for the $j$th observation, as defined by (33); similarly, for the MCUFSA approach.

The figures show that the original distribution structure of two groups is recovered by the estimated factor scores for MFA and MCFA approaches. Their assumed models are sufficiently flexible to fit the data where the directions of the two cluster-error distributions are not parallel to the axes of the original feature sapce. On the other hand the predicted observations for the MCUFSA approach are not fitted well to the actual distribution of two groups as shown in Figure 3. With this approach, the predicted observations tend to be higher than the actual observations from the first group and lower for those from the second group. This lack of fit is due to the strict assumption of a spherical covariance matrix for each component-error distribution. We measured the difference between the predicted and observed observations by the mean squared error (MSE),where MSE=$\sum_{j=1}^{200}(\boldsymbol{y}_j - \hat{\boldsymbol{y}}_j)^T(\boldsymbol{y}_j - \hat{\boldsymbol{y}}_j)/200$. The value of the MSE for the simulated data is 2.30, 3.80, 17.34 for MFA, MCFA, and MCUFSA, respectively. As to be expected, the MSE increases in going from MFA to MCFA and then markedly to MCUFSA.
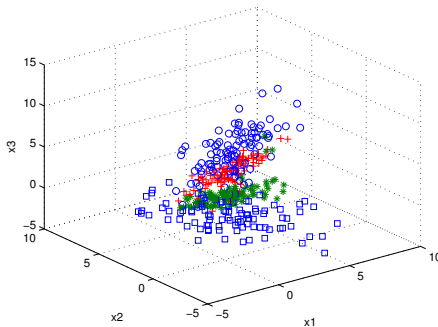
9

Figure 3: Original observations and the predicted observations by MCUFSA: □ Group 1; o Group 2; * predicted for Group 1; + predicted for Group 2

# 6 Applications of MCFA Approach to Clustering

We implemented the MFA, MCFA, and MCUFSA approaches to cluster tissues in two different gene expression data sets and individuals in one chemical measurement data set. We compared the consistency between the implied clustering obtained with each approach with the true group membership of each data set. We adopted the clustering corresponding to the local maximizer that gave the largest value of the likelihood as obtained by implementing the EM algorithm for 50 trial starts, comprising 25 $k$-means starts and 25 random starts. To measure the agreement between a clustering of the data and their true group membership, we used the consistency measures of Jaccard Index [11] and the Adjusted Rand Index (ARI)[12]. Both indices take the value 1 when there is perfect validation between the clustering and the true grouping. The Jaccard Index takes any value between 0 and 1, but the ARI can have negative values.

## 6.1 Example 1: Leukaemia Gene-Expression Data

The first data set concerns the leukaemia tissue samples of Golub et al. [13], in which there are two types of acute leukaemia: Acute Lymphoblastic Leukaemia (ALL) and Acute Myeloid Leukaemia (AML). The data contain the expression levels of 7129 genes on 72 tissues comprising 47 cases of ALL and 25 cases of AML. We preprocessed the data by the following steps: (i) thresholding with a floor of 100 and a ceiling of 16000; (ii) filtering: exclusion of genes with $\max / \min \leq 5$ and $(\max - \min) \leq 500$, where max and min refer to the maximum and minimum expression levels, respectively, of a particular gene across a tissue sample; (iii) taking logs of the expression levels and then standardizing them across genes to have mean zero and unit standard deviation. Finally, standardizing them across samples to have mean zero and unit standard deviation. This preprocessing resulted in thousand genes being retained. We reduced this set further to 100 genes by selecting genes according to the simple two-sample $t$-test as used in [14].

As observed in previous studies, the data on the ALL and AML cases are well separated, confirming the biological distinction between ALL and AML subtypes. Hence all three approaches were able to cluster the into two clusters that almost corresponded

10

Table 2: Agreement between the clustering result and the true membership of leukaemia data

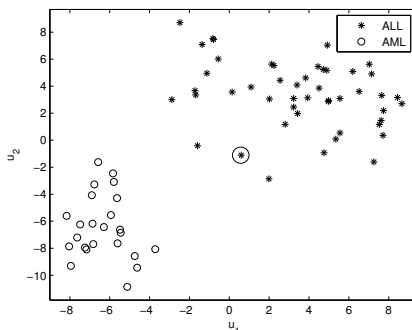| Similarity Indices | Model | $q$ 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| Jaccard | MFA | 0.951 | 0.951 | 0.951 | 0.951 |
| | MCFA | 0.951 | 0.951 | 0.951 | 0.951 |
| | MCUFSA | 0.951 | 0.951 | 0.951 | 0.951 |
| ARI | MFA | 0.945 | 0.945 | 0.945 | 0.945 |
| | MCFA | 0.945 | 0.945 | 0.945 | 0.945 |
| | MCUFSA | 0.945 | 0.945 | 0.945 | 0.945 |



Figure 4: Plot of the factor scores with the MCFA approach

perfectly with the ALL and AML subtypes. There was only one misallocation with the three approaches using either $q = 1, 2, 3,$ or 4 factors. Hence in Table 2 the values of the Jaccard Index are the same for all three approaches, as are the ARI values.

We plotted the estimated factor scores $\hat{\boldsymbol{u}}_j = (\hat{u}_{1j}, \hat{u}_{2j})^T$ of the tissues in two-dimensional space according to their clustered membership determined by the MCFA with $q = 2$ factors (Figure 4). It can be seen from Figure 4 that the two classes ALL and AMl are well separated. The one AML tissue that is misallocated as ALL is circled in Figure 4.

## 6.2 Example 2: Paediatric Leukaemia Gene-Expression Data

In the second example, we considered the clustering of the Paediatric Acute Lymphoblastic Leukaemia (ALL) data of Yeoh et al. [15], who analyzed some gene expressions on 7 subtypes of 327 ALL tissues, consisting of BCR-ABL, E2A-PBXI, Hyperdiploid ($> 50$), MLL, Novel, T-ALL, and TEL-AML1 subtypes. They performed an hierarchical cluster analysis of the 327 diagnostic cases using genes selected by several different methods including chi-squared and the $t$-statistic. We chose for our analysis the 20 genes with highest chi-squared statistics in each subtype. (This resulted in a total 132 unique genes as some were chosen for more than one subtype. Given that the

Table 3: Agreement between the clustering result and the true membership of paediatric ALL data

| Similarity Indices | Model | 1 | 2 | $q$<br>3 | 4 | 5 |
|---|---|---|---|---|---|---|
| | MFA | 0.632 | 0.642 | 0.648 | 0.555 | 0.540 |
| Jaccard | MCFA | NA | 0.355 | 0.597 | 0.610 | 0.650 |
| | MCUFSA | - | - | 0.593 | - | 0.584 |
| | MFA | 0.720 | 0.735 | 0.741 | 0.650 | 0.632 |
| ARI | MCFA | NA | 0.392 | 0.685 | 0.700 | 0.740 |
| | MCUFSA | - | - | 0.687 | - | 0.677 |

number of components (subtypes) here is not small with $g = 7$, we imposed the constraint (23) of common diagonal matrices $\boldsymbol{D}_i$ in the formulation of the MFA approach. This constraint is always imposed with the MCFA approach.

The values of the Jaccard Index and the ARI are listed in Table 3 for the three approaches for each of five levels of the number of factors $q$ ($q = 1, 2, 3, 4, 5$). In the case of a single factor ($q = 1$), the MFCA approach using $g = 7$ components clustered the 327 tissues into only 5 clusters, and so the indices were unable to be calculated. This is noted by NA (not available) for $q = 1$ in Table 3. It can be seen that the performance of the MCFA approach for $q \geq 3$ factors is comparable to the MFA approach, although their best results for the Jaccard and Adjusted Rand Indices occur for different values of $q$. The indices for the latter are greatest for $q = 3$ factors, whereas they are greatest for the MCFA approach for $q = 5$ factors. The ArrayCluster program for the implementation of the MCUFSA approach gave results for only $q = 3$ and 5 factors, which are not as high as those for the MCFA approach.

## 6.3   Example 3: Chemical Data with Additional Noise Added

The third example considers the so-called Vietnam data which was considered in Smyth *et al.* [16]. The Vietnam data set consists of the log transformed and standardized concentrations of 17 chemical elements to which four types of synthetic noise variables were added in [16] to study methods for clustering high-dimensional data. We used these data consisting of a total of 67 variables ($p = 67$; 17 chemical concentration variables plus 50 normal noise variables). The concentrations were measured in hair samples from six classes ($g = 6$) of Vietnamese, and the total number of subjects were $n = 224$.

The values of the indices for the clustering results for this set are presented in Table 4. It can be seen that the highest value (0.815) of the ARI for the MFCA approach was obtained for $q = 4$ factors, compared to a highest value of 0.611 for this index at $q = 3$ factors with the MFA approach. A similar comparison can be made on the basis of the Jaccard Index. Again, it was not possible to obtain results for the MCUFSA approach for all values of $q$.

Table 4: Agreement between the clustering result and the true membership of Vietnamese data

| Similarity Indices | Model | q 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Jaccard | MFA | 0.416 | 0.507 | 0.526 | 0.483 | 0.425 |
| | MCFA | 0.316 | 0.605 | 0.590 | 0.738 | 0.691 |
| | MCUFSA | - | 0.374 | 0.583 | 0.577 | 0.576 |
| ARI | MFA | 0.491 | 0.590 | 0.611 | 0.565 | 0.505 |
| | MCFA | 0.331 | 0.700 | 0.676 | 0.815 | 0.778 |
| | MCUFSA | - | 0.440 | 0.681 | 0.671 | 0.675 |

# 7  Low-Dimensional Plots via MCFA Approach

To illustrate the usefulness of the MFCA approach for portraying the results of a clustering in low-dimensional space, we have plotted in Figure 5 the estimated mean posterior values of the factors $q_{ij}$ as defined by (33). It can be seen that the clusters are represented in this plot with very little overlap. This is not the case in Figure 6, where the first two canonical variates are plotted. They were calculated using the implied clustering labels. It can be seen from Figure 6 that one cluster is essentially on top of another. The canonical variates are calculated on the basis of the assumption of equal cluster-covariance matrices, which does not apply here. The MCFA approach is not predicated on this assumption and so has more flexibility in representing the data in reduced dimensions.
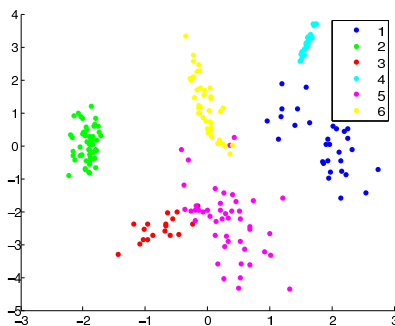


Figure 5: Plot of the (estimated) posterior mean factor scores

# 8  Discussion and Conclusions

In practice, much attention is being given to the use of normal mixture models in density estimation and clustering. However, for high-dimensional data sets, the component-covariance matrices are highly parameterized and some form of reduction in the number of parameters is needed, particularly when the number of observations $n$ is not large
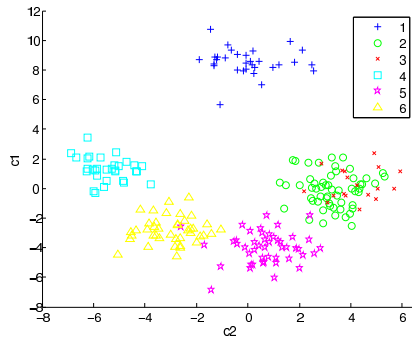
Figure 6: Plot of the first two canonical variates based on the implied clustering via MCFA approach

relative to the number of dimensions $p$. One way of proceeding is to work with mixtures of factor analyzers (MFA) as studied in [1, Chapter 8]. This approach achieves a reduction in the number of parameters through its factor-analytic representation of the component-covariance matrices. But it may not provide a sufficient reduction in the number of parameters, particularly when the the number $g$ of clusters (components)to be imposed on the data is not small. In this paper, we show how in such instances the number of parameters can be reduced appreciably by using a factor-analytic representation of the component-covariance matrices with common factor loadings. The approach is called mixtures of common factor analyzers (MCFA). This sharing of the factor loadings enables the model to be used to cluster high-dimensional into many clusters and to provide low-dimensional plots of the clusters so obtained. The latter plots are given in terms of the (estimated) posterior means of the factors corresponding to the observed data. These projections are not useful with the MFA approach as in its formulation the factors are taken to be white noise with no cluster-specific discriminatory features for the factors.

The MFA approach does allow a more general representation of the component variances/covariances and places no restrictions on the component means. Thus it is more flexible in its modelling of the data. But in this paper we demonstrate that MCFA provides a comparable approach that can be applied in situations where the dimension $p$ and the number of clusters $g$ can be quite large. We have presented analyses of both simulated and real data sets to demonstrate the usefulness of the MCFA approach.

In practice, we can use the Bayesian Information Criterion (BIC) of Schwartz [18] to provide a guide to the choice of the number of factors $q$ and the number of number of components $g$ to be used. On the latter choice it is well known that regularity conditions do not hold for the usual chi-squared approximation to the asymptotic null distribution of the likelihood ratio test statistic to be valid. However, they do hold for tests on the number of factors at a given level of $g$, and so we can also use the likelihood ratio test statistic to choose $q$; see [1, Chapter 8].

# 9   Appendix

The model (15) underlying the MCFA approach can be fitted via the EM algorithm to estimate the vector $\boldsymbol{\Psi}$ of unknown parameters. It consists of the mixing proportions $\pi_i$, the factor component-mean vectors $\boldsymbol{\xi}_i$, the distinct elements of the factor component-covariance matrices $\boldsymbol{\Omega}_i$, the projection matrix $\boldsymbol{A}$ based on sharing of factor loadings, and the common diagonal matrix $\boldsymbol{D}$ of the residuals given the factor scores within a component of the mixture. In order to apply the EM algorithm to this problem, we introduce the component-indicator labels $z_{ij}$, where $z_{ij}$ is one or zero according to whether $\boldsymbol{y}_j$ belongs or does not belong to the $i$th component of the model. We let $\boldsymbol{z}_j$ be the component-label vector, $\boldsymbol{z}_j = (z_{1j}, \ldots, z_{gj})^T$. The $\boldsymbol{z}_j$ are treated as missing data, along with the (unobservable) latent factors $\boldsymbol{u}_{ij}$ within this EM framework. The complete-data log likelihood is then given by

$$\log L_c(\boldsymbol{\Psi}) = \sum_{i=1}^{g} \sum_{j=1}^{n} z_{ij} \{ \log \pi_i + \log \phi(\boldsymbol{y}_j; \boldsymbol{A}\boldsymbol{u}_{ij}, \boldsymbol{D}) + \log \phi(\boldsymbol{u}_{ij}; \boldsymbol{\xi}_i, \boldsymbol{\Omega}_i) \}. \tag{39}$$

## • E-step

On the E-step, we require the conditional expectation of the complete-data log likelihood, $\log L_c(\boldsymbol{\Psi})$, given the observed data $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T)^T$, using the current fit for $\boldsymbol{\Psi}$. Let $\boldsymbol{\Psi}^{(k)}$ be the value of $\boldsymbol{\Psi}$ after the $k$th iteration of the EM algorithm. Then more specifically, on the $(k+1)$th iteration the E-step requires the computation of the conditional expectation of $\log L_c(\boldsymbol{\Psi})$ given $\boldsymbol{y}$, using $\boldsymbol{\Psi}^{(k)}$ for $\boldsymbol{\Psi}$, which is denoted by $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})$.

We let
$$\tau_{ij}^{(k)} = \tau_i(\boldsymbol{y}_j; \boldsymbol{\Psi}^{(k)}), \tag{40}$$
where $\tau_i(\boldsymbol{y}_j; \boldsymbol{\Psi})$ is defined by (27). Also, we let $E_{\boldsymbol{\Psi}^{(k)}}$ refer to the expectation operator, using $\boldsymbol{\Psi}^{(k)}$ for $\boldsymbol{\Psi}$. Then the so-called $Q$-function, $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})$, can be written as

$$Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)}) = \sum_{i=1}^{g} \sum_{j=1}^{n} \tau_{ij}^{(k)} \{ \log \pi_i + w_{1ij}^{(k)} + w_{1ij}^{(k)} \}, \tag{41}$$

where
$$w_{1ij}^{(k)} = E_{\boldsymbol{\Psi}^{(k)}} \{ \log \phi(\boldsymbol{y}_j; \boldsymbol{A}\boldsymbol{u}_{ij}, \boldsymbol{D}) \mid \boldsymbol{y}_j, z_{ij} = 1 \} \tag{42}$$
and
$$w_{2ij}^{(k)} = E_{\boldsymbol{\Psi}^{(k)}} \{ \log \phi(\boldsymbol{u}_{ij}; \boldsymbol{\xi}_i, \boldsymbol{\Omega}_i) \mid \boldsymbol{y}_j, z_{ij} = 1 \}. \tag{43}$$

## • M-step

On the $(k+1)$th iteration of the EM algorithm, the M-step consists of calculating the updated estimates $\pi_i^{(k+1)}, \boldsymbol{\xi}_i^{(k+1)}, \boldsymbol{\Omega}_i^{(k+1)}, \boldsymbol{A}^{(k+1)}$, and $\boldsymbol{D}^{(k+1)}$ by solving the equation

$$\partial Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})/\partial \boldsymbol{\Psi} = \boldsymbol{0}. \tag{44}$$

The updated estimates of the mixing proportions $\pi_i$ are given as in the case of the normal mixture model by

$$\pi_i^{(k+1)} = \sum_{j=1}^{n} \tau_{ij}^{(k)}/n \quad (i = 1, \ldots, g). \tag{45}$$

Concerning the other parameters, it can be shown using vector and matrix differentiation that

$$\partial Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})/\partial \boldsymbol{\xi}_i = \boldsymbol{\Omega}_i^{-1} \sum_{j=1}^{n} \tau_{ij}^{(k)} E_{\boldsymbol{\Psi}^{(k)}}\{(\boldsymbol{u}_{ij} - \boldsymbol{\xi}_i) \mid \boldsymbol{y}_j\}, \tag{46}$$

$$\partial Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})/\partial \boldsymbol{\Omega}_i^{-1} = \sum_{j=1}^{n} \tau_{ij}^{(k)} \tfrac{1}{2}[\boldsymbol{\Omega}_i - E_{\boldsymbol{\Psi}^{(k)}}\{(\boldsymbol{u}_{ij} - \boldsymbol{\xi}_i)(\boldsymbol{u}_{ij} - \boldsymbol{\xi}_i)^T \mid \boldsymbol{y}_j\}], \tag{47}$$

$$\partial Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})/\partial \boldsymbol{D}^{-1} = \sum_{i=1}^{g} \sum_{j=1}^{n} \tau_{ij}^{(k)} \tfrac{1}{2}[\boldsymbol{D} - E_{\boldsymbol{\Psi}^{(k)}}\{(\boldsymbol{y}_j - \boldsymbol{A}\boldsymbol{u}_{ij})(\boldsymbol{y}_j - \boldsymbol{u}_{ij})^T \mid \boldsymbol{y}_j\}], \tag{48}$$

$$\partial Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})/\partial \boldsymbol{A} = \sum_{i=1}^{g} \sum_{j=1}^{n} \tau_{ij}^{(k)}[\boldsymbol{D}^{-1}\{\boldsymbol{y}_j E_{\boldsymbol{\Psi}^{(k)}}(\boldsymbol{u}_{ij}^T \mid \boldsymbol{y}_j)$$
$$-\boldsymbol{A} E_{\boldsymbol{\Psi}^{(k)}}(\boldsymbol{u}_{ij}\boldsymbol{u}_{ij}^T \mid \boldsymbol{y}_j)\}]. \tag{49}$$

On equating (46) to the zero vector, it follows that $\boldsymbol{\xi}_i^{(k+1)}$ can be expressed as

$$\boldsymbol{\xi}_i^{(k+1)} = \boldsymbol{\xi}_i^{(k)} + \frac{\sum_{j=1}^{n} \tau_{ij}^{(k)} \boldsymbol{\gamma}_i^{(k)T}(\boldsymbol{y}_j - \boldsymbol{A}^{(k)}\boldsymbol{\xi}_i^{(k)})}{\sum_{j=1}^{n} \tau_{ij}^{(k)}}, \tag{50}$$

where
$$\boldsymbol{\gamma}_i^{(k)} = (\boldsymbol{A}^{(k)}\boldsymbol{\Omega}_i^{(k)}\boldsymbol{A}^{(k)T} + \boldsymbol{D}^{(k)})^{-1}\boldsymbol{A}^{(k)}\boldsymbol{\Omega}_i^{(k)}. \tag{51}$$

On equating (47) to the null matrix, it follows that

$$\boldsymbol{\Omega}_i^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_{ij}^{(k)} \boldsymbol{\gamma}_i^{(k)T}(\boldsymbol{y}_j - \boldsymbol{A}^{(k)}\boldsymbol{\xi}_i^{(k)})(\boldsymbol{y}_j - \boldsymbol{A}^{(k)}\boldsymbol{\xi}_i^{(k)})^T\boldsymbol{\gamma}^{(k)}}{\sum_{j=1}^{n} \tau_{ij}^{(k)}}$$
$$+(\boldsymbol{I}_q - \boldsymbol{\gamma}_i^{(k)T}\boldsymbol{A}^{(k)})\boldsymbol{\Omega}_i^{(k)} \tag{52}$$

On equating (48) to the zero vector, we obtain

$$\boldsymbol{D}^{(k+1)} = \mathrm{diag}(\boldsymbol{D}_1^{(k)} + \boldsymbol{D}_2^{(k)}), \tag{53}$$

where
$$\boldsymbol{D}_1^{(k)} = \frac{\sum_{i=1}^{g} \sum_{j=1}^{n} \tau_{ij}^{(k)} \boldsymbol{D}^{(k)}(\boldsymbol{I}_p - \boldsymbol{\beta}_i^{(k)})}{\sum_{i=1}^{g} \sum_{j=1}^{n} \tau_{ij}^{(k)}} \tag{54}$$

16

and

$$\boldsymbol{D}_2^{(k)} = \frac{\sum_{i=1}^{g} \sum_{j=1}^{n} \tau_{ij}^{(k)} \boldsymbol{\beta}_i^{(k)^T} (\boldsymbol{y}_j - \boldsymbol{A}^{(k)} \boldsymbol{\xi}_i^{(k)})(\boldsymbol{y}_j - \boldsymbol{A}^{(k)} \boldsymbol{\xi}_i^{(k)})^T \boldsymbol{\beta}_i^{(k)}}{\sum_{i=1}^{g} \sum_{j=1}^{n} \tau_{ij}^{(k)}}, \tag{55}$$

and where

$$\boldsymbol{\beta}_i^{(k)} = (\boldsymbol{A}^{(k)} \boldsymbol{\Omega}_i^{(k)} \boldsymbol{A}^{(k)^T} + \boldsymbol{D}^{(k)})^{-1} \boldsymbol{D}^{(k)}. \tag{56}$$

On equating (49) to the null matrix, we obtain

$$\boldsymbol{A}^{(k+1)} = (\sum_{i=1}^{g} \boldsymbol{A}_{1i}^{(k)})(\sum_{i=1}^{g} \boldsymbol{A}_{2i}^{(k)})^{-1}, \tag{57}$$

where

$$\boldsymbol{A}_{1i}^{(k+1)} = \sum_{j=1}^{n} \tau_{ij}^{(k)} \{ \boldsymbol{y}_j \boldsymbol{\xi}_i^{(k)^T} + (\boldsymbol{y}_j - \boldsymbol{A}^{(k)} \boldsymbol{\xi}_i^{(k)})^T \boldsymbol{\gamma}_i^{(k)} \}, \tag{58}$$

$$\boldsymbol{A}_{2i}^{(k+1)} = \sum_{j=1}^{n} \tau_{ij}^{(k)} \{ (\boldsymbol{I}_q - \boldsymbol{\gamma}_i^{(k)^T} \boldsymbol{A}^{(k)}) \boldsymbol{\Omega}_i^{(k)} + \boldsymbol{r}_i^{(k)} \boldsymbol{r}_i^{(k)^T} \}, \tag{59}$$

and

$$\boldsymbol{r}_i^{(k)} = \boldsymbol{\xi}_i^{(k)} + \boldsymbol{\gamma}_i^{(k)^T} (\boldsymbol{y}_j - \boldsymbol{A}^{(k)} \boldsymbol{\xi}_i^{(k)}) \tag{60}$$

The factor loading matrix $\boldsymbol{A}^{(k+1)}$ must be orthogonal in columns, that is,

$$\boldsymbol{A}^{(k+1)^T} \boldsymbol{A}^{(k+1)} = \boldsymbol{I}_q. \tag{61}$$

We can use the Cholesky decomposition to find the upper triangular matrix $\boldsymbol{C}^{(k+1)}$ of order $q$ so that

$$\boldsymbol{A}^{(k+1)^T} \boldsymbol{A}^{(k+1)} = \boldsymbol{C}^{(k+1)^T} \boldsymbol{C}^{(k+1)}. \tag{62}$$

Then it follows that if we replace $\boldsymbol{A}^{(k+1)}$ by

$$\boldsymbol{A}^{(k+1)} \boldsymbol{C}^{(k+1)^{-1}}, \tag{63}$$

then it will satisfy the requirement (61). With the adoption of the estimate (63) for $\boldsymbol{A}$, we need to adjust the updated estimates $\boldsymbol{\xi}_i^{(k+1)}$ and $\boldsymbol{\Omega}_i^{(k+1)}$ to be

$$\boldsymbol{C}^{(k+1)} \boldsymbol{\xi}_i^{(k+1)} \tag{64}$$

and

$$\boldsymbol{C}^{(k+1)} \boldsymbol{\Omega}_i^{(k+1)} \boldsymbol{C}^{(k)^T}, \tag{65}$$

where $\boldsymbol{\xi}_i^{(k+1)}$ and $\boldsymbol{\Omega}_i^{(k+1)}$ are given by (50) and (52), respectively.

We have to specify an initial value for the vector $\boldsymbol{\Psi}$ of unknown parameters in the application of the EM algorithm. A random start is obtained by first randomly assigning the data into $g$ groups. Let $n_i$, $\bar{\boldsymbol{y}}_i$, and $\boldsymbol{S}_i$ be the number of observations, the sample mean, and the sample covariance matrix, respectively, of the $i$th group of the data so obtained $(i = 1, \ldots, g)$. We then proceed as follows:

- Set $\pi_i^{(0)} = n_i/n$.

- Generate random numbers from the standard normal distribution $N(0,1)$ to obtain values for the $(j,k)$th element of $\boldsymbol{A}^*$ $(j = 1, \ldots, p; k = 1, \ldots, q)$.

- Implement the Cholesky decomposition so that $\boldsymbol{A}^{*T}\boldsymbol{A}^* = \boldsymbol{C}^T\boldsymbol{C}$, where $\boldsymbol{C}$ is the upper triangle matrix of order $q$, and define $\boldsymbol{A}^{(0)}$ by $\boldsymbol{A}^*\boldsymbol{C}^{-1}$.

- On noting that the transformed data $\boldsymbol{D}^{-1/2}\boldsymbol{Y}_j$ satisfies the probabilistic PCA model of Tipping and Bishop [17] with $\sigma_i^2 = 1$, it follows that for a given $\boldsymbol{D}^{(0)}$ and $\boldsymbol{A}^{(0)}$, we can specify $\boldsymbol{\Sigma}_i^{(0)}$ as

$$\boldsymbol{\Omega}_i^{(0)} = \boldsymbol{A}^{(0)^T}\boldsymbol{D}^{(0)^{1/2}}\boldsymbol{H}_i(\boldsymbol{\Lambda}_i - \tilde{\sigma}_i^2 I_q)\boldsymbol{H}_i^T\boldsymbol{D}^{(0)^{1/2}}\boldsymbol{A}^{(0)},$$

where $\tilde{\sigma}_i^2 = \sum_{h=q+1}^p \lambda_{ih}/(p-q)$. The $q$ columns of the matrix $\boldsymbol{H}_i$ are the eigenvectors corresponding to the eigenvalues $\lambda_{i1} \geq \lambda_{i2} \geq \cdots \geq \lambda_{iq}$ of

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

where $\boldsymbol{\Omega}_i^{(0)}$ is the covariance matrix of the $\boldsymbol{y}_j$ in the $i$th group, and $\boldsymbol{\Lambda}_i$ is the diagonal matrix with diagonal elements equal to $\lambda_{i1}, \ldots, \lambda_{iq}$. Concerning the choice of $\boldsymbol{D}^{(0)}$, we can take $\boldsymbol{D}^{(0)}$ to be the diagonal matrix formed from the diagonal elements of the (pooled) within-cluster sample covariance matrix of the $\boldsymbol{y}_j$. The initial value for $\boldsymbol{\xi}_i$ is $\boldsymbol{\xi}_i^{(0)} = \boldsymbol{A}^{(0)^T}\bar{\boldsymbol{y}}_i$.

Some clustering procedure such as $k$-means can be used to provide non-random partitions of the data, which can be used to obtain another set of initial values for the parameters. In our analyses we used both initialization methods.

# Acknowledgement

## REFERENCES

[1] G.J. McLachlan and D. Peel, *Finite Mixture Models.* New York: Wiley, 2000.

[2] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm (with discussion)," *Journal of the Royal Statistical Society: Series B*, vol. 39, pp. 1–38, 1977.

[3] G.J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions.* Second Edition. New York: Wiley, 2008.

[4] J.D. Banfield and A.E. Raftery, "Model-based Gaussian and non-Gaussian clustering," *Biometrics*, Vol. 49, pp. 803–821, 1993.

[5] G.J. McLachlan, D. Peel, and R.W. Bean, "Modelling high-dimensional data by mixtures of factor analyzers," *Computational Statistics & Data Analysis*, vol. 41, pp. 379–388, 2003.

[6] G.J. McLachlan, R.W. Bean, and L. Ben-Tovim Jones, "Extension of the mixture of factor analyzers model to incorporate the multivariate $t$ distribution," *Computational Statistics & Data Analysis*, Vol. 51, 5327–5338, 2007.

[7] G.E. Hinton, P. Dayan, and M. Revow, M. "Modeling the manifolds of images of handwritten digits," *IEEE Transactions on Neural Networks*, vol. 8, pp. 65–73, 1997.

[8] R. Yoshida, T. Higuchi, and S. Imoto, "A mixed factors model for dimension reduction and extraction of a group structure in gene expression data," *Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference*, pp. 161–172, 2004.

[9] R. Yoshida, T. Higuchi, S. Imoto, and S. Miyano, "ArrayCluster: an analytic tool for clustering, data visualization and model finder on gene expression profiles," *Bioinformatics*, vol. 22, pp. 1538–1539, 2006.

[10] G. Sanguinetti, "Dimensionality rediction of clustered data sets," *IEEE Transactions Pattern Analysis and Machine Intelligence*, Vol. 30, pp. 535–540.

[11] P. Jaccard, "Distribution de la florine alpine dans la Bassin de Dranses et dans quelques regiones voisines," *Bulletin de la Société Vaudoise des Sciences Naturel les*, Vol. 37, pp. 241–272, 1901.

[12] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, Vol. 2, 193–218, 1985.

[13] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeck, P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, *et al.*, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, 286, 531-537, 1999.

[14] D. Nguyen, and D. Rocke, "Tumor classification by partial least squares using microarray gene expression data," *Bioinformatics*, Vol. 18, pp. 39–50, 2002.

[15] E. Yeoh, and Ross, M.E., *et al.*, "Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling," *Cancer Cell*, Vol. 1, pp. 133–143, 2002.

[16] C. Smyth, D. Coomans, and Y. Everingham, "Clustering noisy data in a reduced dimension space via multivariate regression trees," *Pattern Recognition*, Vol. 39, pp. 424–431, 2006.

[17] M.E. Tipping, and C.M. Bishop, "Mixtures of probabilistic principal component analysers," *Neural Computation*, Vol. 11, pp. 443–482, 1999.

[18] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, Vol. 6, pp. 461–464, 1978.