Mixtures of factor analyzers with common factor loadings: applications to the clustering and visualisation of high-dimensional data

Jangsun Baek, Geoffrey J. McLachlan and Lloyd K. Flack

Abstract—Mixtures of factor analyzers enable model-based density estimation to be undertaken for high-dimensional data, where the number of observations n is not 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.

Index Terms—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 π_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

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

- *J. Baek is with the Department of Statistics, Chonnam National University, Gwangju 500-757, South Korea.*
- *G.J. McLachlan and L.K. Flack are with the Department of Mathematics and Institute for Molecular Bioscience, University of Queensland, Brisbane QLD 4072, Australia. E-mail:gjm@maths.uq.edu.au*

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

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

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. As considered initially in [8], we extend the model of [9, 10] 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. The representation (10) and (11) is not unique, as it still holds if $\boldsymbol{A}$ were to be postmultiplied by any nonsingular matrix. Hence the number of free parameters in $\boldsymbol{A}$ is

$$pq - q^2. \tag{12}$$

Thus 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

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

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

In our implementation of this procedure, we postmultiply the solution $\hat{\boldsymbol{A}}$ for $\boldsymbol{A}$ by the nonsingular matrix as defined in the Appendix that achieves the result

$$\hat{\boldsymbol{A}}^T\hat{\boldsymbol{A}} = \boldsymbol{I}_q, \tag{14}$$

where $\boldsymbol{I}_q$ denotes the $q \times q$ identity matrix. That is, the p columns of $\hat{\boldsymbol{A}}$ are taken to be orthonormal. This solution is unique up to postmultiplication by an orthogonal matrix.

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

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

for $j = 1, \ldots, n$, where the (unobservable) factors $\boldsymbol{U}_{i1}, \ldots, \boldsymbol{U}_{in}$ are distributed independently $N(\boldsymbol{0}, \boldsymbol{I}_q)$, 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

$$\boldsymbol{Y}_j = \boldsymbol{A}\boldsymbol{U}_{ij} + \boldsymbol{e}_{ij} \qquad \text{with prob. } \pi_i \quad (i = 1, \ldots, g) \tag{16}$$

for $j = 1, \ldots, n$, where the (unobservable) factors $\boldsymbol{U}_{i1}, \ldots, \boldsymbol{U}_{in}$ are distributed independently $N(\boldsymbol{\xi}_i, \boldsymbol{\Omega}_i)$, 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 (14).

To see that the MCFA model as specificed by (16) is a special case of the MFA approach as specified by (15),

we note that we can rewrite (16) as

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

where

$$\boldsymbol{\mu}_i = \boldsymbol{A}\boldsymbol{\xi}_i, \tag{18}$$

$$\boldsymbol{B}_i = \boldsymbol{A}\boldsymbol{K}_i, \tag{19}$$

$$\boldsymbol{U}_{ij}^* = \boldsymbol{K}_i^{-1}(\boldsymbol{U}_{ij} - \boldsymbol{\xi}_i), \tag{20}$$

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 chosen so that

$$\boldsymbol{K}_i^{-1}\boldsymbol{\Omega}_i\boldsymbol{K}_i^{-1^T} = \boldsymbol{I}_q \quad (i = 1, \ldots, g). \tag{21}$$

On comparing (17) with (15), it can be seen that the MCFA model is a special case of the MFA model with the additional restrictions that

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

$$\boldsymbol{B}_i = \boldsymbol{A}\boldsymbol{K}_i \quad (i = 1, \ldots, g), \tag{23}$$

and

$$\boldsymbol{D}_i = \boldsymbol{D} \quad (i = 1, \ldots, g), \tag{24}$$

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 (23) 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. [9, 10], who also imposed the additional restriction that the common diagonal covariance matrix $\boldsymbol{D}$ of the error terms is spherical,

$$\boldsymbol{D} = \sigma^2\boldsymbol{I}_p, \tag{25}$$

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

$$d_3 = (g-1) + pq + 1 + 2gq - \tfrac{1}{2}q(q+1). \tag{26}$$

In our experience, we have found that this restriction 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. 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 models with the MFA, MCFA, and MCUFSA approaches when $p = 50, 100$; $q = 2$; and $g = 4, 8$. For example, when we cluster $p = 50$ dimensional gene expression data into $g = 4$ groups using $q = 2$ dimensional factors, the MFA model requires 799 parameters to be estimated, while the MCUFSA needs only 117 parameters. Moreover, as the number of clusters grows from 4 to 8 the number of parameters for the MFA model grows almost twice as large as before, but that for MCUFSA remains almost the same (137 parameters). However, 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 adequately as it is assuming that the component-covariance matrices of the factors are diagonal and that the cluster-error distributions conditional on the factors are spherical. The MCFA model has 170 and 194 parameters for $g = 4$ and 8, respectively, with $q = 2$ factors. It will be seen that the MCFA approach provides a good parsimonious compromise between the MFA and MCUFSA approaches.

In the context of the analysis of speech recognition data, Rosti and Gales [11] considered a factor-analytic approach in which separate mixture distributions are adopted independently for the factors and for the error terms conditional on the factors. It thus contains our model (16) as a special case where the error distribution conditional on the factors is specified by a single normal distribution. However, they developed their procedure

TABLE 1
The number of parameters in models for three factor-analytic approaches

	p	g	q	Number of parameters
MFA	50	4	2	799
	50	8	2	1599
	100	4	2	1599
	100	8	2	3199
MCFA	50	4	2	169
	50	8	2	193
	100	4	2	319
	100	8	2	343
MCUFSA	50	4	2	117
	50	8	2	137
	100	4	2	217
	100	8	2	237

for only diagonal component-covariance matrices for the factors, whereas in our MCFA model these factor covariance matrices have no restrictions imposed on them. Some related work in this context of speech recognition includes [12-16].

In a related approach with common factor loadings adopted recently by Galimberti et al. [17] for the data after mean centring, the factors in (16) are taken to have a mixture distribution with the constraints that its mean is the null vector and its covariance matrix is the identity matrix. As their program applies only to mean-centred data, it cannot be used to visualize the original data.

In other recent work, Sanguinetti [18] 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 (24) 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 ith 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

$$\begin{aligned} E_{\boldsymbol{\Psi}}\{Z_{ij} \mid \boldsymbol{y}_j\} &= \mathrm{pr}_{\boldsymbol{\Psi}}\{Z_{ij} = 1 \mid \boldsymbol{y}_j\} \\ &= \tau_i(\boldsymbol{y}_j; \boldsymbol{\Psi}) \\ &\quad (i = 1, \ldots, g; j = 1, \ldots, n), \end{aligned} \tag{27}$$

where $\tau_i(\boldsymbol{y}_j; \boldsymbol{\Psi})$ is the posterior probability that $\boldsymbol{y}_j$ belongs to the ith component of the mixture. From (16), it can be expressed under the MCFA model as

$$\tau_i(\boldsymbol{y}_j; \boldsymbol{\Psi}) = \frac{\pi_i \phi(\boldsymbol{y}_j; \boldsymbol{A}\boldsymbol{\xi}_i, \boldsymbol{A}\boldsymbol{\Omega}_i\boldsymbol{A}^T + \boldsymbol{D})}{\sum_{h=1}^{g} \pi_h \phi(\boldsymbol{y}_j; \boldsymbol{A}\boldsymbol{\xi}_h, \boldsymbol{A}\boldsymbol{\Omega}_h\boldsymbol{A}^T + \boldsymbol{D})} \tag{28}$$

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 ith component of the mixture (that is, $z_{ij} = 1$) is multivariate normal,

$$\boldsymbol{U}_{ij} \mid \boldsymbol{y}_j, z_{ij} = 1 \sim N(\boldsymbol{\xi}_{ij}, \boldsymbol{\Omega}_i^*), \tag{29}$$

where

$$\boldsymbol{\xi}_{ij} = \boldsymbol{\xi}_i + \boldsymbol{\gamma}_i^T(\boldsymbol{y}_j - \boldsymbol{A}\boldsymbol{\xi}_i) \tag{30}$$

and

$$\boldsymbol{\Omega}_i^* = (\boldsymbol{I}_q - \boldsymbol{\gamma}_i^T\boldsymbol{A})\boldsymbol{\Omega}_i, \tag{31}$$

and where

$$\boldsymbol{\gamma}_i = (\boldsymbol{A}\boldsymbol{\Omega}_i\boldsymbol{A}^T + \boldsymbol{D})^{-1}\boldsymbol{A}\boldsymbol{\Omega}_i. \tag{32}$$

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 (29) and (30),

$$\begin{aligned} E(\boldsymbol{U}_{ij} \mid \boldsymbol{y}_j, z_{ij} = 1) &= \boldsymbol{\xi}_{ij} \\ &= \boldsymbol{\xi}_i + \boldsymbol{\gamma}_i^T(\boldsymbol{y}_j - \boldsymbol{A}\boldsymbol{\xi}_i). \end{aligned} \tag{33}$$

We let $\hat{\boldsymbol{u}}_{ij}$ denote the value of the right-hand side of (33) 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 jth factor corresponding to $\boldsymbol{y}_j$ as

$$\hat{\boldsymbol{u}}_j = \sum_{i=1}^{g} \tau_i(\boldsymbol{y}_j; \hat{\boldsymbol{\Psi}})\,\hat{\boldsymbol{u}}_{ij} \quad (j = 1, \ldots, n) \tag{34}$$

where, from (28), $\tau_i(\boldsymbol{y}_j; \hat{\boldsymbol{\Psi}})$ is the estimated posterior probability that $\boldsymbol{y}_j$ belongs to the ith component. An alternative estimate of the posterior expectation of the factor corresponding to the jth observation $\boldsymbol{y}_j$ is defined by replacing $\tau_i(\boldsymbol{y}_j; \hat{\boldsymbol{\Psi}})$ by $\hat{z}_{ij}$ in (34), where

$$\begin{aligned} \hat{z}_{ij} &= 1, \quad \text{if } \hat{\tau}_i(\boldsymbol{y}_j; \hat{\boldsymbol{\Psi}}) \geq \hat{\tau}_h(\boldsymbol{y}_j; \hat{\boldsymbol{\Psi}}), \\ &\qquad\qquad (h = 1, \ldots, g; h \neq i), \\ &= 0, \quad \text{otherwise.} \end{aligned} \tag{35}$$

5 Simulation Experiment 1: 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 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. 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

$$\hat{\boldsymbol{y}}_j = \sum_{i=1}^{g} \tau_i(\boldsymbol{y}_j; \hat{\boldsymbol{\Psi}})(\hat{\boldsymbol{\mu}}_i + \hat{\boldsymbol{B}}_i \hat{\boldsymbol{u}}_{ij}) \tag{36}$$

where

$$\hat{\boldsymbol{u}}_{ij} = \hat{\boldsymbol{\alpha}}_i^T(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_i), \tag{37}$$

where

$$\hat{\boldsymbol{\alpha}}_i = (\hat{\boldsymbol{B}}_i \hat{\boldsymbol{B}}_i^T + \hat{\boldsymbol{D}}_i)^{-1} \hat{\boldsymbol{B}}_i. \tag{38}$$

For the MCFA approach, the predicted observation is

$$\hat{\boldsymbol{y}}_j = \hat{\boldsymbol{A}} \hat{\boldsymbol{u}}_j, \tag{39}$$

where $\hat{\boldsymbol{A}}$ is the estimated projection matrix $\boldsymbol{A}$ and where $\hat{\boldsymbol{u}}_j$ is the estimated factor score for the jth observation, as defined by (34); 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 space. 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 and diagonal component-covariance matrices for the factors. 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.

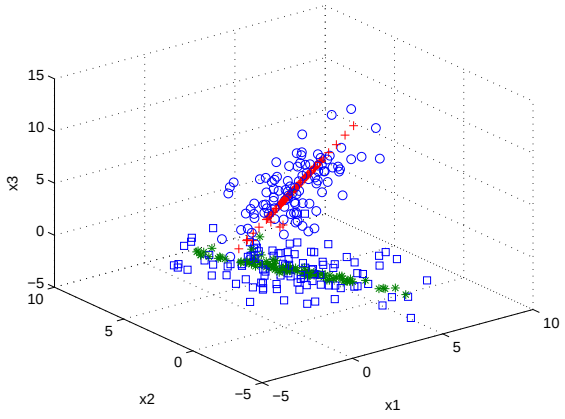


Fig. 1. Original observations and the predicted observations by MFA: □ Group 1; o Group 2; * predicted for Group 1; + predicted for Group 2

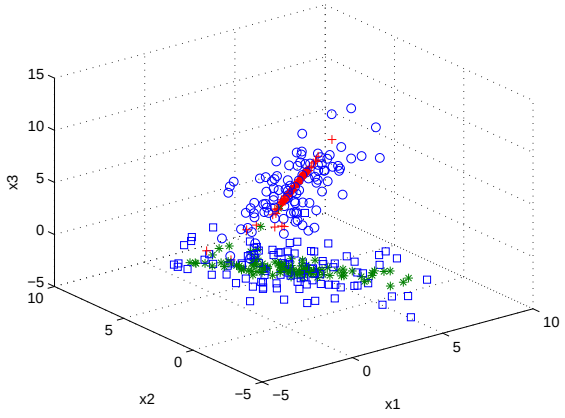


Fig. 2. Original observations and the predicted observations by MCFA: □ Group 1; o Group 2; * predicted for Group 1; + predicted for Group 2

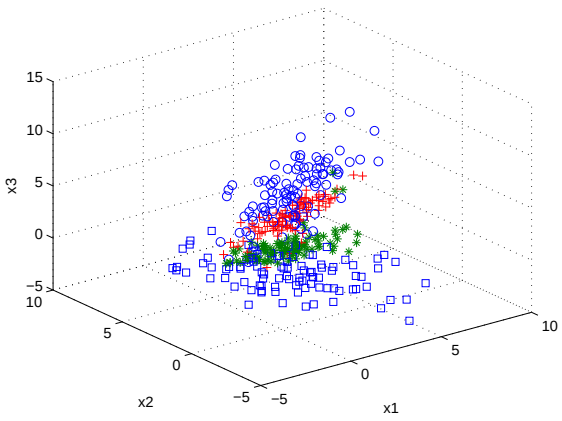


Fig. 3. Original observations and the predicted observations by MCUFSA: □ Group 1; o Group 2; * predicted for Group 1; + predicted for Group 2

6 Simulation Experiment 2

To illustrate further the application of the MCFA approach, we report now the results of a second simulation experiment performed in situations in which the MCFA model is valid and in which we know the true underlying group structure of the data. A sample of $n = 200$ p-dimensional observations

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

was generated from a $g = 5$ component mixture model of bivariate normal factor $(q = 2)$. Here $\boldsymbol{y}_{1j}$ is a $p_1 = 10$ dimensional subvector containing the signal, while $\boldsymbol{y}_{2j}$ is a p_2-dimensional subvector of noise variables with $p_2 = p - p_1$, where there are five levels of $p_2\,(p_2 = 0, 10, 20, 30, 40)$.

The generated data can be viewed as coming from the factor-analytic model (16),

$$\boldsymbol{Y}_j = (\boldsymbol{A}_1^T, \boldsymbol{A}_2^T)^T \boldsymbol{U}_{ij} + \boldsymbol{e}_{ij} \quad \text{with prob. } \pi_i\,(i = 1, \ldots, 5), \tag{40}$$

where $\pi_1 = 0.15, \pi_2 = 0.2, \pi_3 = 0.15, \pi_4 = 0.2$, and $\pi_5 = 0.3$. The $p_1 \times q$ submatrix $\boldsymbol{A}_1$ of factor loadings was specified to be

$$\boldsymbol{A}_1^T = \begin{pmatrix} 0.5 & -0.9 & 0.3 & 0.6 & 0.2 & -0.7 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.8 & -0.7 & 0.5 & 0.6 & -0.4 & 0.3 & -0.5 \end{pmatrix}.$$

The $p_2 \times q$ submatrix $\boldsymbol{A}_2$ corresponding to the p_2 noise variables has all elements zero. The mean vectors of the factors $\boldsymbol{U}_{ij}$ were specified as $\boldsymbol{\xi}_1 = (0\ \ 2.5)^T$, $\boldsymbol{\xi}_2 = (-2.5\ \ 0)^T$, $\boldsymbol{\xi}_3 = (2.5\ \ 0)^T$, $\boldsymbol{\xi}_4 = (0\ \ -2.5)^T$, $\boldsymbol{\xi}_5 = (0\ \ 0)^T$, while their covariance matrices were taken to be

$$\boldsymbol{\Omega}_1 = \begin{pmatrix} 0.10 & 0 \\ 0 & 0.45 \end{pmatrix}, \boldsymbol{\Omega}_2 = \begin{pmatrix} 0.45 & 0 \\ 0 & 0.10 \end{pmatrix},$$

$$\boldsymbol{\Omega}_3 = \begin{pmatrix} 0.45 & 0 \\ 0 & 0.10 \end{pmatrix}, \boldsymbol{\Omega}_4 = \begin{pmatrix} 0.10 & 0 \\ 0 & 0.45 \end{pmatrix},$$

$$\boldsymbol{\Omega}_5 = \begin{pmatrix} 1 & 0.90 \\ 0.90 & 1 \end{pmatrix}.$$

The error terms $\boldsymbol{e}_{ij}\,(j = 1, \ldots, n)$ were taken to be distributed independently $N(\boldsymbol{0}, \boldsymbol{D})$, where $\boldsymbol{D}$ is a diagonal matrix $(i = 1, \ldots, 5)$. Its first p_1 elements were randomly generated from a uniform distribution on the interval $[0.1, 0.3]$, while its remaining p_2 elements were randomly generated from a uniform distribution on the interval $[0.3, 0.8]$.

The generated bivariate factor scores $\boldsymbol{u}_{ij}$ are displayed in Figure 4 for each component $i\,(i = 1, \ldots, 5)$.

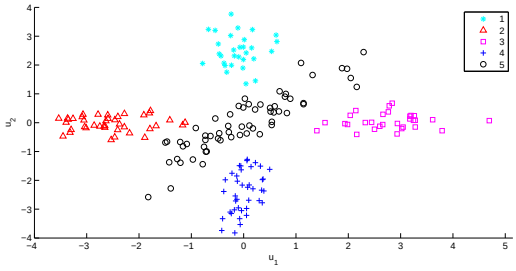


Fig. 4. Scatter plot of generated bivariate factors for each of $g = 5$ components

We implemented the MFA and MCFA approaches on the five data sets corresponding to the five levels of the number p_2 of noise variables for each of 24 combinations of the number of components g and the number of factors $q\,(g = 2, \ldots, 7; q = 2, \ldots, 5)$. To measure the agreement between a clustering of the data and their true group membership, we computed the error rate, where the latter corresponds to the lowest proportion of errors with respect to the true grouping over all permutations of the cluster labels.

In practice, we can use the Bayesian Information Criterion (BIC) of Schwarz [19] to provide a guide to the choice of the number of factors q and the 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]. In this paper, we used the BIC criterion to choose $g\,(q)$ for given $q\,(g)$ both in this simulation example and in the real data sets to be discussed later. With BIC, $d \log n$ is added to twice the negative of the log likelihood at the solution, where d denotes the number of (free) parameters in the model. The intent is to choose $g\,(q)$ to minimize the negative of this penalized form of the log likelihood.

With the MFCA approach, the correct combination of g and q was selected for all 5 data sets. On the other hand, the correct combination was not selected for any of the data sets with the MFA and MCUFSA approaches.

In Table 2, we give the values of the error rate (ERR)

and the adjusted Rand index (ARI) for the MFA, MCFA, and MCUFSA approaches for the correct combination of g and $q\,(g = 5, q = 2)$ for each of the five levels of the number of noise variables p_2. The adjusted Rand index (ARI)[20] takes the value 1 when there is perfect agreement between the clustering and the true grouping, and it can be negative.

It can be seen that in the case of no noise variables ($p_2 = 0$), the error rate for the MFA approach is nearly seven times as large as for the MCFA (46 misallocations versus 7). As the number of noise variables p_2 increases, the performance of the error rate of the MCFA approach only slightly increases (resulting in at most 2 more misallocations), while the ARI only falls slightly. In contrast, as p_2 increases from $p_2 = 0$, the performance of the MFA approach relative to MCFA becomes even poorer in terms of the error rate and the ARI. Comparing MCFA with the MCUFSA approach, it can be seen that the error rate of the latter is around three times that of the former over the five levels of the number p_2 of noise variables.

In summary, the MCFA approach has been essentially unaffected by the presence of the additional noise variables in this simulated data set. Of course it has been generated under a model in which the MCFA holds exactly. The MFA model also holds here as we have seen in Section 2 that it is more general than the MCFA model. But as it has to fit more parameters, its performance here falls away as the signal in the $p_1 = 10$ variables is degraded by the presence of a larger number of noise variables.

TABLE 2
Values of the error rate (ERR) and the adjusted Rand index (ARI) for the MFA, MCFA, and MCUFSA approaches in the case of $g = 5$ and $q = 2$

Model		Number p_2 of noise variables 0	10	20	30	40
MFA	ERR	0.2300	0.2900	0.2700	0.3250	0.2750
	ARI	0.5353	0.4660	0.4830	0.4568	0.4760
MCFA	ERR	0.0350	0.0350	0.0450	0.0450	0.0400
	ARI	0.9017	0.9017	0.8760	0.8760	0.8883
MCUFSA	ERR	0.1100	0.0900	0.1100	0.1050	0.1100
	ARI	0.7273	0.7626	0.7162	0.7279	0.7200

7 Applications of MCFA Approach to Clustering of Real Data Sets

We now report on the application of the MFA, MCFA, and MCUFSA approaches to cluster tissues in one gene expression data set and individuals in one chemical measurement data set. We compared the agreement between the implied clustering obtained with each approach with the true group membership of each data set. In both examples there are multiple groups ($g = 6$) for which the MCFA model is specifically designed to handle through its use of shared factor loadings for the groups.

7.1 Example 1: Paediatric Leukaemia Gene-Expression Data

The first real data set concerns the clustering of the Paediatric Acute Lymphoblastic Leukaemia (ALL) data of Yeoh et al. [21]. This data set had $n = 327$ samples from 9 subtypes of ALL and one normal group. The classes were BCR-ABL, E2A-PBX1, Hyperdip (> 50), MLL, T-ALL, TEL-AML1, Hyperdip47-50, Hypodip, Normal, and Pseudodip. We used here only the $n = 248$ samples in the first $g = 6$ classes. The final four classes were either too small to be reliably classified or there was probably a group structure present that was different from the given classes. There is some evidence for the last mentioned possibility as Yeoh *et al.* [21] claim to have found a group within the last four classes that did not correspond to the given classes.

The files selected contained both the value calculated by MAS 4.0 and an indicator of whether MAS evaluated the gene as being expressed, absent, or whether data were missing. The samples were standardized to all have the same mean of 2500, but the variance was not standardized. The values of each gene were then stanardized so that the minimum value for samples with that gene identified as being expressed was set to one. Values of a gene where MAS identified it as being unexpressed or where data were missing were then set to one. Genes, which had a range of values of less than 500 or which had less than 30 samples where that gene was identified as being expressed, were deleted. This reduced the number of genes from 12,625 to 6,350. The genes were then logarithmically transformed and then standardized to have mean zero and unit variance.

The remaining 6,350 genes were further reduced by running the select-genes step of the EMMIX-GENE procedure [22], whereby a gene is retained if twice the increase in the log likelihood is greater than a specificed threshold (here 8) in testing for a single t-component versus a mixture of two t-components. Also, the minimum size of the two clusters had to be greater than an imposed threshold of 8.

This reduced the data set to 5,483 genes. The top 2,000 genes in terms of the aforementioned likelihood ratio statistic were selected and clustered into 50 clusters using the cluster-genes step of the EMMIX-GENE procedure, which uses essentially a soft-version of k-means to cluster the genes with the intent that highly correlated genes (genes close in Euclidean distance) are put together in the same cluster. We then took the means of these 50 clusters (metagenes) to be our p variables to which the MFA, MCFA, and MCUFSA approaches were applied. Given that the number of components (subtypes) here is not small with $g = 6$, we imposed the constraint (24) of common diagonal matrices $\boldsymbol{D}_i$ in the formulation of the MFA approach. This constraint is always imposed with the MCFA approach.

We implemented the MFA, MCFA, and MCUFSA approaches with $g = 6$ components for the number of

factors q ranging from 1 to 9. To measure the agreement between a clustering of the data and their true group membership, we calculated the error rate and the ARI plus one other similarity measure, namely the Jaccard index [23].

The results are presented in Table 3. In the case of a single factor ($q = 1$), the MFUCSA approach put the 248 tissues into only 5 clusters, and so the error rate was unable to be calculated. This is noted by NA (not available) for $q = 1$ in Table 3. We have also listed in this table under the heading of BIC, twice the negative of the log likelihood augmented by $\log n$ times the number of (free) parameters d.

TABLE 3
Comparison of MFA, MCFA, and MCUFSA approaches for implied clustering versus the true membership of paediatric ALL data

Model	Factors	BIC	ARI	Jaccard	Error rate
MFA	**1**	**4684**	**0.8006**	**0.7264**	**0.1331**
	2	5261	0.4561	0.3910	0.3911
	3	6092	0.2679	0.2649	0.5323
	4	6933	0.2300	0.2360	0.5726
	5	7855	0.3162	0.2915	0.4839
	6	8868	0.2932	0.2767	0.5040
	7	9797	0.2080	0.2232	0.5524
	8	10719	0.1879	0.2086	0.5806
	9	10205	0.2750	0.2653	0.5202
MCFA	1	8300	-0.0040	0.1064	0.7540
	2	6425	0.3380	0.3129	0.4435
	3	5091	0.6159	0.5285	0.3065
	4	3664	0.7435	0.6604	0.2258
	5	3274	**0.8447**	**0.7816**	**0.1169**
	6	**3069**	0.7147	0.6305	0.2419
	7	3119	0.6763	0.5878	0.2298
	8	3159	0.8327	0.7672	0.1411
	9	3251	0.831	0.7649	0.1492
MCUFSA	1	9570	−0.0064	0.1760	NA
	2	6918	0.2901	0.2865	0.5161
	3	5749	0.6210	0.5412	0.2903
	4	4368	**0.8301**	**0.7638**	**0.1411**
	5	4077	0.6573	0.5701	0.2218
	6	**3756**	0.8101	0.7403	0.1452
	7	3808	0.6822	0.5978	0.2621
	8	3862	0.5274	0.4472	0.3185
	9	3954	0.3738	0.3462	0.3831

It can be seen that the MCFA leads to good values for the indices and error rate for $q \geq 4$ factors, achieving its lowest error rate (and highest ARI and Jaccard index) for $q = 5$ factors. Apart from its performance for a single factor ($q = 1$), the MFA approach is not as good as MCFA. However, the use of BIC to choose q would lead to this choice of q, whereas with the MCFA and MCUFSA approaches, it does not lead to the choice of q with the smallest (largest) error rate (AR/Jaccard indices).

7.2 Example 2: Vietnam Chemical Data with Additional Noise Added

The second example considers the so-called Vietnam data which was considered in Smyth *et al.* [24]. 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 [24] 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 uniform 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 noise variables were generated from the uniform distribution on the interval $[-2, 2]$.

We implemented the MFA, MCFA, and MCUFSA approaches with $g = 6$ components for the number of factors q ranging from 1 to 5. Again with the MFA model, we imposed the assumption of equal diagonal matrices D_i for the error terms. For each value of q, we computed the ARI, Jaccard index, and the error rate. They are displayed in Table 4. It was not possible to obtain results for the MCUFSA approach for all values of q less than 5.

It can be seen that the lowest error rate and highest values of the ARI and Jaccard index are obtained by using $q = 3$ factors with the MCFA model, which coincides with the choice on the basis of BIC. The best result with the MFA model is obtained for $q = 2$ factors (BIC suggests using $q = 1$). It can be seen that the error rate, ARI, and Jaccard index for MFA and MCUFSA are not nearly as good as for MCFA.

TABLE 4
Comparison of MFA, MCFA, and MCUFSA approaches for implied clustering versus the true membership of Vietnam data

Model	Factors	BIC	ARI	Jaccard	Error rate
MFA	**1**	**46758**	0.5925	0.4974	0.2277
	2	48212	**0.6585**	**0.5600**	**0.1696**
	3	49743	0.6322	0.5342	0.1830
	4	51351	0.5392	0.4510	0.2589
	5	52846	0.5700	0.4767	0.2679
MCFA	1	45171	0.3444	0.3447	0.4777
	2	44950	0.7288	0.6380	0.1384
	3	**44825**	**0.8063**	**0.7248**	**0.0893**
	4	44984	0.7081	0.6241	0.2277
	5	45151	0.6259	0.5385	0.2634
MCUFSA	1	NA	NA	NA	NA
	2	NA	NA	NA	NA
	3	NA	NA	NA	NA
	4	NA	NA	NA	NA
	5	46523	0.5479	0.4572	0.2768

8 Low-Dimensional Plots via MCFA Approach

To illustrate the usefulness of the MCFA approach for portraying the results of a clustering in low-dimensional space, we have plotted in Figure 5 for the Vietnam data the estimated posterior means of the factors $\hat{u}_j$ as defined by (34) with the implied cluster labels shown. In this

plot, we have chosen the second and third factors in the MCFA model with $q = 3$ factors. 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.

We have also given in Figure 7 the plot correspsonding to that in Figure 5 with the true cluster labels shown. It can be seen there is good agreement between the two plots. This is to be expected since the error rate of the MCFA model fitted with $q = 3$ factors is quite low (0.0893).

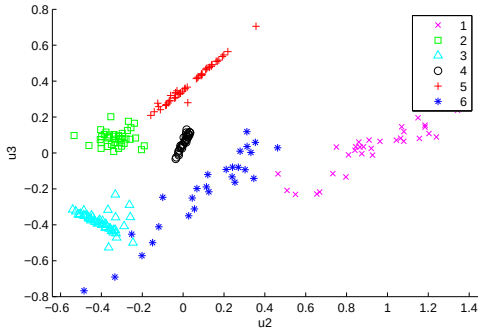


Fig. 5. Plot of the (estimated) posterior mean factor scores via the MCFA approach with the six cluster labels shown for the Vietnam data

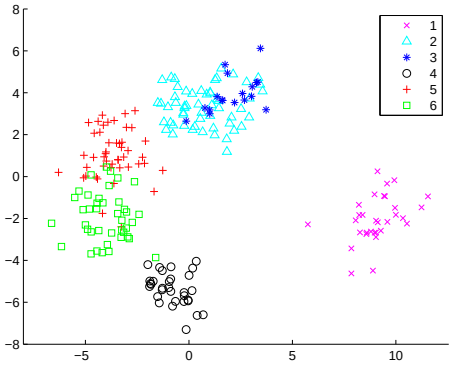


Fig. 6. Plot of the first two canonical variates based on the implied clustering via MCFA approach with the six cluster labels shown for the Vietnam data

9 Discussion and Conclusions

In practice, much attention is being given to the use of normal mixture models in density estimation and

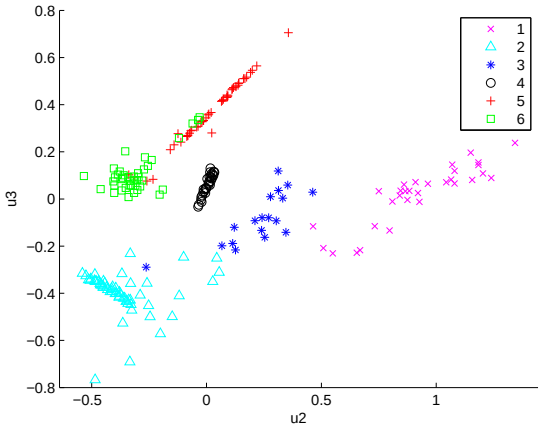


Fig. 7. Plot of the (estimated) posterior mean factor scores via the MCFA approach with the true labels shown for the six classes in the Vietnam data

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 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 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 data 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 Cri-

terion (BIC) of Schwartz [19] to provide a guide to the choice of the number of factors q and the 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]. In our examples and simulation experiments preseneted here, we used BIC, although it did not always lead to the correct choice of the number of factors q as, for example, in the paediatric ALL data example. In future work, we wish to investigate the use of BIC and other criteria on choosing the number of factors q for a given number of components g.

10 Appendix

The model (16) 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 π_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 ith 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

$$\log L_c(\boldsymbol{\Psi}) = \sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij}\{\log \pi_i + \log \phi(\boldsymbol{y}_j; \boldsymbol{A}\boldsymbol{u}_{ij}, \boldsymbol{D}) + \log \phi(\boldsymbol{u}_{ij}; \boldsymbol{\xi}_i, \boldsymbol{\Omega}_i)\}. \quad (41)$$

- **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 kth 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

$$\tau_{ij}^{(k)} = \tau_i(\boldsymbol{y}_j; \boldsymbol{\Psi}^{(k)}), \quad (42)$$

where $\tau_i(\boldsymbol{y}_j; \boldsymbol{\Psi})$ is defined by (28). 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

$$Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)}) = \sum_{i=1}^{g}\sum_{j=1}^{n} \tau_{ij}^{(k)}\{\log \pi_i + w_{1ij}^{(k)} + w_{1ij}^{(k)}\}, \quad (43)$$

where

$$w_{1ij}^{(k)} = E_{\boldsymbol{\Psi}^{(k)}}\{\log \phi(\boldsymbol{y}_j; \boldsymbol{A}\boldsymbol{u}_{ij}, \boldsymbol{D}) \mid \boldsymbol{y}_j, z_{ij} = 1\} \quad (44)$$

and

$$w_{2ij}^{(k)} = E_{\boldsymbol{\Psi}^{(k)}}\{\log \phi(\boldsymbol{u}_{ij}; \boldsymbol{\xi}_i, \boldsymbol{\Omega}_i) \mid \boldsymbol{y}_j, z_{ij} = 1\}. \quad (45)$$

- **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

$$\partial Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})/\partial \boldsymbol{\Psi} = \boldsymbol{0}. \quad (46)$$

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

$$\pi_i^{(k+1)} = \sum_{j=1}^{n} \tau_{ij}^{(k)}/n \quad (i = 1, \ldots, g). \quad (47)$$

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

$$\partial Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})/\partial \boldsymbol{\xi}_i = \boldsymbol{\Omega}_i^{-1}\sum_{j=1}^{n} \tau_{ij}^{(k)} E_{\boldsymbol{\Psi}^{(k)}}\{(\boldsymbol{u}_{ij} - \boldsymbol{\xi}_i) \mid \boldsymbol{y}_j\}, \quad (48)$$

$$\partial Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})/\partial \boldsymbol{\Omega}_i^{-1} = \sum_{j=1}^{n} \tau_{ij}^{(k)} \tfrac{1}{2}[\boldsymbol{\Omega}_i - E_{\boldsymbol{\Psi}^{(k)}}\{(\boldsymbol{u}_{ij} - \boldsymbol{\xi}_i)(\boldsymbol{u}_{ij} - \boldsymbol{\xi}_i)^T \mid \boldsymbol{y}_j\}], \quad (49)$$

$$\partial Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})/\partial \boldsymbol{D}^{-1} = \sum_{i=1}^{g}\sum_{j=1}^{n} \tau_{ij}^{(k)} \tfrac{1}{2}[\boldsymbol{D} - E_{\boldsymbol{\Psi}^{(k)}}\{(\boldsymbol{y}_j - \boldsymbol{A}\boldsymbol{u}_{ij})(\boldsymbol{y}_j - \boldsymbol{u}_{ij})^T \mid \boldsymbol{y}_j\}], \quad (50)$$

$$\partial Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})/\partial \boldsymbol{A} = \sum_{i=1}^{g}\sum_{j=1}^{n} \tau_{ij}^{(k)}[\boldsymbol{D}^{-1}\{\boldsymbol{y}_j E_{\boldsymbol{\Psi}^{(k)}}(\boldsymbol{u}_{ij}^T \mid \boldsymbol{y}_j) - \boldsymbol{A}E_{\boldsymbol{\Psi}^{(k)}}(\boldsymbol{u}_{ij}\boldsymbol{u}_{ij}^T \mid \boldsymbol{y}_j)\}]. \quad (51)$$

On equating (48) to the zero vector, it follows that $\boldsymbol{\xi}_i^{(k+1)}$ can be expressed as

$$\boldsymbol{\xi}_i^{(k+1)} = \boldsymbol{\xi}_i^{(k)} + \frac{\sum_{j=1}^{n} \tau_{ij}^{(k)} \boldsymbol{\gamma}_i^{(k)T} \boldsymbol{y}_{ij}^{(k)}}{\sum_{j=1}^{n} \tau_{ij}^{(k)}}, \quad (52)$$

where

$$\boldsymbol{y}_{ij}^{(k)} = \boldsymbol{y}_j - \boldsymbol{A}^{(k)}\boldsymbol{\xi}_i^{(k)} \quad (53)$$

and

$$\boldsymbol{\gamma}_i^{(k)} = (\boldsymbol{A}^{(k)}\boldsymbol{\Omega}_i^{(k)}\boldsymbol{A}^{(k)^T} + \boldsymbol{D}^{(k)})^{-1}\boldsymbol{A}^{(k)}\boldsymbol{\Omega}_i^{(k)}. \quad (54)$$

On equating (49) to the null matrix, it follows that

$$\boldsymbol{\Omega}_i^{(k+1)} = \frac{\sum_{j=1}^n \tau_{ij}^{(k)} \boldsymbol{\gamma}_i^{(k)^T} \boldsymbol{y}_{ij}^{(k)} \boldsymbol{y}_{ij}^{(k)^T} \boldsymbol{\gamma}^{(k)}}{\sum_{j=1}^n \tau_{ij}^{(k)}} + (\boldsymbol{I}_q - \boldsymbol{\gamma}_i^{(k)^T} \boldsymbol{A}^{(k)}) \boldsymbol{\Omega}_i^{(k)} \quad (55)$$

On equating (50) to the zero vector, we obtain

$$\boldsymbol{D}^{(k+1)} = \mathrm{diag}(\boldsymbol{D}_1^{(k)} + \boldsymbol{D}_2^{(k)}), \quad (56)$$

where

$$\boldsymbol{D}_1^{(k)} = \frac{\sum_{i=1}^g \sum_{j=1}^n \tau_{ij}^{(k)} \boldsymbol{D}^{(k)} (\boldsymbol{I}_p - \boldsymbol{\beta}_i^{(k)})}{\sum_{i=1}^g \sum_{j=1}^n \tau_{ij}^{(k)}} \quad (57)$$

and

$$\boldsymbol{D}_2^{(k)} = \frac{\sum_{i=1}^g \sum_{j=1}^n \tau_{ij}^{(k)} \boldsymbol{\beta}_i^{(k)^T} \boldsymbol{y}_{ij}^{(k)} \boldsymbol{y}_{ij}^{(k)^T} \boldsymbol{\beta}_i^{(k)}}{\sum_{i=1}^g \sum_{j=1}^n \tau_{ij}^{(k)}}, \quad (58)$$

and where

$$\boldsymbol{\beta}_i^{(k)} = (\boldsymbol{A}^{(k)} \boldsymbol{\Omega}_i^{(k)} \boldsymbol{A}^{(k)^T} + \boldsymbol{D}^{(k)})^{-1} \boldsymbol{D}^{(k)}. \quad (59)$$

On equating (51) to the null matrix, we obtain

$$\boldsymbol{A}^{(k+1)} = (\sum_{i=1}^g \boldsymbol{A}_{1i}^{(k)})(\sum_{i=1}^g \boldsymbol{A}_{2i}^{(k)})^{-1}, \quad (60)$$

where

$$\boldsymbol{A}_{1i}^{(k+1)} = \sum_{j=1}^n \tau_{ij}^{(k)} \{\boldsymbol{y}_j \boldsymbol{\xi}_i^{(k)^T} + \boldsymbol{y}_{ij}^{(k)^T} \boldsymbol{\gamma}_i^{(k)}\}, \quad (61)$$

$$\boldsymbol{A}_{2i}^{(k+1)} = \sum_{j=1}^n \tau_{ij}^{(k)} \{(\boldsymbol{I}_q - \boldsymbol{\gamma}_i^{(k)^T} \boldsymbol{A}^{(k)}) \boldsymbol{\Omega}_i^{(k)} + \boldsymbol{r}_i^{(k)} \boldsymbol{r}_i^{(k)^T}\}, \quad (62)$$

and

$$\boldsymbol{r}_i^{(k)} = \boldsymbol{\xi}_i^{(k)} + \boldsymbol{\gamma}_i^{(k)^T} \boldsymbol{y}_{ij}^{(k)}. \quad (63)$$

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 ith 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 distrubution $N(0, 1)$ to obtain values for the (j, k)th element of $\boldsymbol{A}^*$ $(j = 1, \ldots, p; k = 1, \ldots, q)$.
- Define $\boldsymbol{A}^{(0)}$ by $\boldsymbol{A}^*$.
- On noting that the transformed data $\boldsymbol{D}^{-1/2}\boldsymbol{Y}_j$ satisfies the probabilistic PCA model of Tipping and Bishop [25] with $\sigma_i^2 = 1$, it follows that for a given $\boldsymbol{D}^{(0)}$ and $\boldsymbol{A}^{(0)}$, we can specify $\boldsymbol{\Omega}_i^{(0)}$ 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

$$\boldsymbol{D}^{(0)^{-1/2}} \boldsymbol{S}_i \boldsymbol{D}^{(0)^{-1/2}}, \quad (64)$$

where $\boldsymbol{S}_i$ is the covariance matrix of the $\boldsymbol{y}_j$ in the ith 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 partititons 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.

As noted previously, the solution $\hat{\boldsymbol{A}}$ for the matrix of factor loadings is unique only up to postmultiplication by a nonsingular matrix. We chose to postmultiply by the nonsingular matrix for which the solution is orthonormal; that is,

$$\hat{\boldsymbol{A}}^T \hat{\boldsymbol{A}} = \boldsymbol{I}_q. \quad (65)$$

To achieve this with $\hat{\boldsymbol{A}}$ computed as above, we note that we can use the Cholesky decomposition to find the upper triangular matrix $\boldsymbol{C}$ of order q so that

$$\hat{\boldsymbol{A}}^T \hat{\boldsymbol{A}} = \boldsymbol{C}^T \boldsymbol{C}. \quad (66)$$

Then it follows that if we replace $\hat{\boldsymbol{A}}$ by

$$\hat{\boldsymbol{A}} \boldsymbol{C}^{-1}, \quad (67)$$

then it will satisfy the requirement (65). With the adoption of the estimate (67) for $\hat{\boldsymbol{A}}$, we need to adjust the updated estimates $\hat{\boldsymbol{\xi}}_i$ and $\hat{\boldsymbol{\Omega}}_i$ to be

$$\boldsymbol{C}\hat{\boldsymbol{\xi}}_i \quad (68)$$

and

$$\boldsymbol{C}\hat{\boldsymbol{\Omega}}_i \boldsymbol{C}^T, \quad (69)$$

where $\hat{\boldsymbol{\xi}}_i$ and $\hat{\boldsymbol{\Omega}}_i$ are given by the limiting values of (52) and (55), respectively.

An R version of our program is available at http://www.maths.uq.edu.au/~gjm/

ACKNOWLEDGEMENT

The authors would like to thank the Associate Editor and the anonymous reviewers for their helpful comments in improving the manuscript. The work of J. Baek was supported by a grant from the Korean Research Foundation funded by the Korean Government (MOEHRD, Basic Research Promotion Fund, KRF-2007-521-C00048). The work of G. McLachlan was supported by the Australian Research Council. Also, he wishes to thank the Isaac Newton Institute for Mathematical Sciences for support to participate in its research program on Statistical Theory and Methods for Complex, High-Dimensional Data.

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] J. Baek and G.J. McLachlan, "Mixtures of factor analyzers with common factor loadings for the clustering and visualisation of high-dimensional data," Preprint Series of the Isaac Newton Institute for Mathematical Sciences, Cambridge, Technical Report NI08018-SCH, 2008.

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

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

[11] A.-V.I. Rosti and M.J.F. Gales, "Factor analysis hidden Markov models for speech recognition," *Computer Speech and Language*, vol. 18, pp. 181–200, 2004.

[12] N. Kumar, "Investigation of silicon-auditory models and generalization of linear discriminant analysis for improved speech recognition", doctoral dissertation, Johns Hopkins University, Maryland, USA 1997.

[13] R. Gopinath, B. Ramabhadran, and S. Dharanipragada, "Factor analysis invariant to linear transformations of data", in *Proceedings International Conference on Speech and Language Processing* pp. 397–400, 1998.

[14] M. Gales, "Semi-tied covariance matrices for hidden Markov models" *IEEE Transactions on Speech and Audio Processing*, vol.7, pp.272–281, 1999.

[15] S. Axelrod, R. Gopinath, and P. Olsen, "Modeling with a subspace constraint on inverse covariance matrices", in *Proc. ICSLP*, 2002.

[16] P. Olsen, and R. Gopinath "Modeling inverse covariance matrices by basis expansion", in *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, vol.1, pp. 945-948, 2002.

[17] G. Galimberti, A. Montanari and C. Viroli, "Latent classes of objects and variable selection," in *Proceedings of COMPSTAT 2008, P. Brito (Ed.)*, Heidelberg: Springer, pp. 373–383, 2008.

[18] G. Sanguinetti, "Dimensionality rediction of clustered data sets," *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol. 30, pp. 535–540, 2008.

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

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

[21] 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.

[22] G.J. McLachlan, R.W. Bean, and D. Peel, "Mixture Model-Based Approach to the Clustering of Microarray Expression Data," *Bioinformatics*, vol. 18, pp. 413–422, 2002.

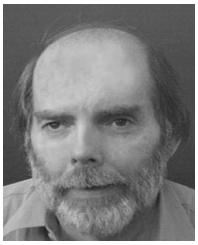
[23] P. Jaccard, "Distribution de la florine alpine dans la Bassin de Dranses et dans quelques regiones voisines," *Bulletin de la Société Vaudoise des Sciences Naturel les*, vol. 37, pp. 241–272, 1901.

[24] 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.

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

Jangsun Baek received the B.S. and M.S. degrees in Applied Statistics at Yonsei University, South Korea, in 1981 and 1984, respectively, and the Ph.D. degree in Statistics at Texas A&M University, U.S.A., in 1991. From 1991 to 1993 he was a postdoctoral fellow at the Department of Statistical Science, Southern Methodist University, U.S.A. Since 1993 he has been a faculty member of the Department of Statistics, Chonnam National University, South Korea.

Professor Baek is a member of the American Statistical Association and the Korean Statistical Society. His research interests include pattern recognition, image segmentation, multivariate statistics. Recently he is interested in developing clustering and classification methods for high-dimensional data in bioinformatics.



Geoff McLachlan received the B.Sc. (Hons.) and Ph.D. degrees from the University of Queensland in 1969 and 1973, respectively. Since 1975 he has been a faculty member of the Department of Mathematics of the University of Queensland. In 1994, he was awarded a D.Sc. degree by the University of Queensland on the basis of his publications in the scientific literature. Since 2002, he has had a joint appointment with the Institute for Molecular Bioscience and he is a chief investigator of the Australian Research Council Centre of Excellence in Biomathematics. In 2007, he was awarded an Australian Professorial Fellowship.

Professor McLachlan is a fellow of the American Statistical Association, the Royal Statistical Society, and the Australian Mathematical Society. His research interests have been concentrated in the related fields of classification, cluster and discriminant analyses, image analysis, machine learning, neural networks, pattern recognition, and data mining, and in the field of statistical inference. More recently, he has become actively involved in the field of bioinformatics with the focus on the statistical analysis of microarray gene-expression data. In these fields, he has published over 190 research articles, including six monographs. The last five monographs, which are volumes in the Wiley Series in Probability and Statistics, are on the topics of discriminant analysis, the EM algorithm (including a second edition), finite mixture models, and the analysis of microarray data.

Professor McLachlan is on the editorial board of several international journals and has served on the program committee for many international conferences. He is a member of the College of Experts of the Australian Research Council and is President-elect of the International Federation of Classification Societies.

Lloyd Flack received the B.Sc. degree from the University of Sydney in 1973 and the M.Stats degree from the University of New South Wales in 1995. From 1985 to 1987, hes was a professional officer at the University of Sydney and then at the University of New South Wales. From 1987 to 1998, he was a Biometrican for NSW Agriculture, Sydney Water, NSW Department of Natural Resources, and the NSW Environment Protection Authority. In 1999 he was a biostatistician for Polartechnics. From 2000 to 2003 he was a methodologist for the Australian Bureau of Statistics. In 2004 he was a data analyst for the CSIRO. Since 2007, he has been a research assistant in the Department of Mathematics and the Institute for Molecular Bioscience at the University of Queensland.

His research interests are mostly in the fields of classification and clustering methods and in smoothers. He is especially interested in the application of statistics to biological and environmental problems.