Clustering of High-Dimensional Data via Finite Mixture Models

G.J. McLachlan and Jangsun Baek

[1] Department of Mathematics and Institute for Molecular Bioscience, University of Queensland Brisbane, QLD 4072, Australia gjm@maths.uq.edu.au
[2] Department of Statistics, Chonnam National University, Gwangju 500-757, South Korea jbaek@chonnam.ac.kr

Summary. Finite mixture models are being commonly used in a wide range of applications in practice concerning density estimation and clustering. An attractive feature of this approach to clustering is that it provides a sound statistical framework in which to assess the important question of how many clusters there are in the data and their validity. We review the application of normal mixture models to high-dimensional data of a continuous nature. One way to handle the fitting of normal mixture models is to adopt mixtures of factor analyzers. They enable model-based density estimation and clustering 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. We focus here on a new modified approach that uses common component-factor loadings, which considerably reduces further the number of parameters. Moreover, it allows the data to be displayed in low-dimensional plots.

Key words: Model-based clustering, Normal mixture densities, Mixtures of factor analyzers, Common factor analyzers

1 Introduction

Clustering procedures based on finite mixture models are being increasingly preferred over heuristic methods due to their sound mathematical basis and to the interpretability of their results. Mixture model-based procedures provide a probabilistic clustering that allows for overlapping clusters corresponding to the components of the mixture model. The uncertainties that the observations belong to the clusters are provided in terms of the fitted values for their posterior probabilities of component membership of the mixture. As each component in a finite mixture model corresponds to a cluster, it allows the important question of how many clus-

ters there are in the data to be approached through an assessment of how many components are needed in the mixture model. These questions of model choice can be considered in terms of the likelihood function; see, for example, McLachlan (1982) and McLachlan and Peel (2000).

2 Definition of Mixture Models

We let $\boldsymbol{Y}$ denote a random vector consisting of p feature variables associated with the random phenomenon of interest. We let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ denote an observed random sample of size n on $\boldsymbol{Y}$. With the finite mixture model-based approach to density estimation and clustering, the density of $\boldsymbol{Y}$ is modelled as a mixture of a number (g) of component densities $f_i(\boldsymbol{y}; \boldsymbol{\theta}_i)$ in some unknown proportions $\pi_1, \ldots, \pi_g$, where $f_i(\boldsymbol{y}; \boldsymbol{\theta}_i)$ is specified up to an unknown parameter vector $\boldsymbol{\theta}_i$ ($i = 1, \ldots, g$). That is, each data point is taken to be a realization of the mixture probability density function (p.d.f.),

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

where the mixing proportions π_i are nonnegative and sum to one. In density estimation, the number of components g can be taken sufficiently large for (1) to provide an arbitrarily accurate estimate of the underlying density function.

The vector of all unknown parameters is given by $\boldsymbol{\Psi} = (\boldsymbol{\omega}^T, \pi_1, \ldots, \pi_{g-1})^T$, where $\boldsymbol{\omega}$ consists of the elements of the $\boldsymbol{\theta}_i$ known *a priori* to be distinct. For an observed random sample, $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, the log likelihood function for $\boldsymbol{\Psi}$ is given by

$$\log L(\boldsymbol{\Psi}) = \sum_{j=1}^{n} \log f(\boldsymbol{y}_j; \boldsymbol{\Psi}). \tag{2}$$

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

$$\partial \log L(\boldsymbol{\Psi})/\partial \boldsymbol{\Psi} = \boldsymbol{0}. \tag{3}$$

Solutions of (3) corresponding to local maximizers of $\log L(\boldsymbol{\Psi})$ can be obtained via the expectation-maximization (EM) algorithm (Dempster et al., 1977). In the event that the EM sequence is trapped at some stationary point that is not a local or global maximizer of $\log L(\boldsymbol{\Psi})$ (for example, a saddle point), a small random perturbation of $\boldsymbol{\Psi}$ away from the saddle point will cause the EM algorithm to diverge from the saddle point; see McLachlan and Krishnan (2008).

For clustering purposes, each component in the mixture model (1) corresponds to a cluster. The posterior probability that an observation with feature vector $\boldsymbol{y}_j$ belongs to the ith component of the mixture can be expressed by Bayes' theorem as

$$\tau_i(\boldsymbol{y}_j; \boldsymbol{\Psi}) = \frac{\pi_i f_i(\boldsymbol{y}_j; \boldsymbol{\theta}_i)}{\sum_{h=1}^{g} \pi_h f_h(\boldsymbol{y}_j; \boldsymbol{\theta}_h)} \quad (i = 1, \ldots, g; j = 1, \ldots, n). \tag{4}$$

for $i = 1, \ldots, g$. A probabilistic clustering of the data into g clusters can be obtained in terms of the fitted posterior probabilities of component membership for the data.

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

$$\begin{aligned} \hat{z}_{ij} &= 1, \qquad \text{if } \hat{\tau}_i(\boldsymbol{y}_j; \hat{\boldsymbol{\Psi}}) \geq \hat{\tau}_h(\boldsymbol{y}_j; \hat{\boldsymbol{\Psi}}), \qquad (h = 1, \ldots, g; h \neq i), \\ &= 0, \qquad \text{otherwise.} \end{aligned} \tag{5}$$

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

3 Choice of Starting Values for the EM Algorithm

McLachlan and Peel (2000) provide an in-depth account of the fitting of finite mixture models. Briefly, with mixture models the likelihood typically will have multiple maxima; that is, the likelihood equation will have multiple roots. Thus the EM algorithm needs to be started from a variety of initial values for the parameter vector $\boldsymbol{\Psi}$ or for a variety of initial partitions of the data into g groups. The latter can be obtained by randomly dividing the data into g groups corresponding to the g components of the mixture model. With random starts, the effect of the central limit theorem tends to have the component parameters initially being similar at least in large samples. Nonrandom partitions of the data can be obtained via some clustering procedure such as k-means. Also, Coleman et al. (1999) have proposed some procedures for obtaining nonrandom starting partitions.

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

4 Clustering via Normal Mixtures

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

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

$$f_i(\boldsymbol{y};\, \boldsymbol{\theta}_i) = \phi(\boldsymbol{y};\, \boldsymbol{\mu}_i,\, \boldsymbol{\Sigma}_i), \tag{6}$$

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

$$\phi(\boldsymbol{y};\, \boldsymbol{\mu}_i,\, \boldsymbol{\Sigma}_i) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}_i|^{-1/2} \exp\{-\tfrac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{y} - \boldsymbol{\mu}_i)\}. \tag{7}$$

One attractive feature of adopting mixture models with elliptically symmetric components such as the normal or t-densities, is that the implied clustering is invariant under affine transformations of the data; that is, invariant under transformations of the feature vector $\boldsymbol{y}$ of the form,

$$\boldsymbol{y} \rightarrow \boldsymbol{C}\boldsymbol{y} + \boldsymbol{a}, \tag{8}$$

where $\boldsymbol{C}$ is a nonsingular matrix. If the clustering of a procedure is invariant under (8) for only diagonal $\boldsymbol{C}$, then it is invariant under change of measuring units but not rotations. But as commented upon by Hartigan (1975), this form of invariance is more compelling than affine invariance.

It can be seen from (7) that the mixture model with unrestricted component-covariance matrices in its normal component distributions is a highly parameterized one with $\frac{1}{2}p(p+1)$ parameters for each component-covariance matrix $\boldsymbol{\Sigma}_i$ $(i = 1, \ldots, g)$. As an alternative to taking the component-covariance matrices to be the same or diagonal, we can adopt some model for the component-covariance matrices that is intermediate between homoscedasticity and the unrestricted model, as in the approach of Banfield and Raftery (1993). They introduced a parameterization of the component-covariance matrix $\boldsymbol{\Sigma}_i$ based on a variant of the standard spectral decomposition of $\boldsymbol{\Sigma}_i$.

The mixture model with normal components (7) is sensitive to outliers since it adopts the multivariate normal family for the distributions of the errors. An obvious way to improve the robustness of this model for data which have longer tails than the normal or atypical observations is to consider using the multivariate t-family of elliptically symmetric distributions; see McLachlan and Peel (1998) and McLachlan and Peel (2000, Chapter 7). It has an additional parameter called the degrees of freedom that controls the length of the tails of the distribution. Although the number of outliers needed for breakdown is almost the same as with the normal distribution, the outliers have to be much larger; see Hennig (2003, 2004).

5 Some Recent Extensions for High-Dimensional Data

The EMMIX-GENE program of McLachlan et al. (1999) is an extension of the EMMIX program of McLachlan et al. (2002) for the normal mixture model-based clustering of a limited number of observations that may be of extremely high-dimensions. It was called EMMIX-GENE as it was designed specifically for problems in bioinformatics that require the clustering of a relatively small number of tissue samples containing the expression levels of possibly thousands of genes. But it is applicable to clustering problems outside the field of bioinformatics involving high-dimensional data. In situations where the number of variables p is large, it might not be practical to fit mixtures of factor analyzers to data on all the variables, as it would involve a considerable amount of computation time. Thus initially some of the variables may have to be removed. Indeed, the simultaneous use of too many variables in the cluster analysis may serve only to create noise that masks the effect of a smaller number of variables. Also, the intent of the cluster analysis may not be to produce a clustering of the observations on the basis of all the available variables, but rather to discover and study different clusterings of the observations corresponding to different subsets of the variables.

Therefore, the EMMIX-GENE procedure has two optional steps before the final step of clustering the observations. The first step considers the selection of a subset of relevant variables from the available set of variables by screening the variables on an individual basis to eliminate those which are of little use in clustering the observations. The usefulness of a given variable to the clustering process can be assessed formally by a test of the null hypothesis that it has a single component normal distribution over the observations (McLachlan et al., 2002). A faster but *ad hoc* way is to make this decision on the basis, say, of the sample interquartile range; if a variable has a distribution that is a mixture of normals, then its interquartile range will be greater than that for a single normal population. Even after this step has been completed, there may still remain too many variables. Thus there is a second step in EMMIX-GENE in which the retained variables are clustered (after standardization) into a number of groups on the basis of Euclidean distance so that variables with similar profiles are put into the same group. In general, care has to be taken with the scaling of variables before clustering of the observations, as the nature of the variables can be intrinsically different. Also, as noted above, the clustering of the observations via normal mixture models is invariant under changes in scale and location. The clustering of the observations can be carried out on the basis of the groups considered individually using some or all of the variables within a group or collectively. For the latter, we can replace each group by a representative (a metavariable) such as the sample mean as in the EMMIX-GENE procedure.

6 Factor Analysis Model for Dimension Reduction

As remarked earlier, the g-component normal mixture model with unrestricted component-covariance matrices is a highly parameterized model with $\frac{1}{2}p(p+1)$ parameters for each component-covariance matrix $\boldsymbol{\Sigma}_i$ $(i = 1, \ldots, g)$. As discussed

above, Banfield and Raferty (1993) introduced a parameterization of the component-covariance matrix $\boldsymbol{\Sigma}_i$ based on a variant of the standard spectral decomposition of $\boldsymbol{\Sigma}_i$ $(i = 1, \ldots, g)$. However, if p is large relative to the sample size n, it may not be possible to use this decomposition to infer an appropriate model for the component-covariance matrices. Even if it is possible, the results may not be reliable due to potential problems with near-singular estimates of the component-covariance matrices when p is large relative to n.

A common approach to reducing the number of dimensions is to perform a principal component analysis (PCA). But as is well known, projections of the feature data $\boldsymbol{y}_j$ onto the first few principal axes are not always useful in portraying the group structure; see the example in McLachlan and Peel (2000, Section 8.2). A global nonlinear approach can be obtained by postulating a factor-analytic model for each component-covariance matrix of the full feature vector $\boldsymbol{Y}_j$; see Hinton et al. (1997), McLachlan and Peel (2000), and McLachlan et al. (2003). This leads to the mixture of factor analyzers (MFA) model given by

$$f(\boldsymbol{y}_j; \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i \phi(\boldsymbol{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{9}$$

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

$$\boldsymbol{\Sigma}_i = \boldsymbol{B}_i \boldsymbol{B}_i^T + \boldsymbol{D}_i \quad (i = 1, \ldots, g) \tag{10}$$

and where $\boldsymbol{B}_i$ is a $p \times q$ matrix of factor loadings and $\boldsymbol{D}_i$ is a diagonal matrix $(i = 1, \ldots, g)$.

This MFA approach with the factor-analytic representation (10) on $\boldsymbol{\Sigma}_i$ is equivalent to assuming that the distribution of the difference $\boldsymbol{Y}_j - \boldsymbol{\mu}_i$ can be modelled as

$$\boldsymbol{Y}_j - \boldsymbol{\mu}_i = \boldsymbol{B}_i \boldsymbol{U}_{ij} + \boldsymbol{e}_{ij} \qquad \text{with prob. } \pi_i \quad (i = 1, \ldots, g) \tag{11}$$

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 $(i = 1, \ldots, g)$.

The parameter vector $\boldsymbol{\Psi}$ now consists of the mixing proportions π_i and the elements of the $\boldsymbol{\mu}_i$, the $\boldsymbol{B}_i$, and the $\boldsymbol{D}_i$. With this approach, the number of free parameters is controlled through the dimension of the latent factor space. By working in this reduced space, it allows a model for each component-covariance matrix with complexity lying between that of the isotropic and full covariance structure models without any restrictions on the covariance matrices. The mixture of factor analyzers model can be fitted by using the alternating expectation–conditional maximization (AECM) algorithm of Meng and van Dyk (1997).

A formal test for the number of factors can be undertaken using the likelihood ratio λ, as regularity conditions (Rao, 1973) hold for this test conducted at a given value for the number of components g. For the null hypothesis that $H_0 : q = q_0$ versus the alternative $H_1 : q = q_0 + 1$, the statistic $-2 \log \lambda$ is asymptotically chi-squared with $d = g(p - q_0)$ degrees of freedom. However, in situations where n is not large relative to the number of unknown parameters, we prefer the use of the BIC

criterion (Schwarz, 1978). Applied in this context, it means that twice the increase in the log likelihood ($-2\log\lambda$) has to be greater than $d\log n$ for the null hypothesis to be rejected.

The mixture of factor analyzers model is sensitive to outliers since it uses normal errors and factors. Recently, McLachlan et al. (2007) have considered the use of mixtures of t analyzers in an attempt to make the model less sensitive to outliers. In some other recent work, Montanari and Viroli (2007) have considered the use of mixtures of factor analyzers with covariates.

As $\frac{1}{2}q(q-1)$ constraints are needed for $\boldsymbol{B}_i$ to be uniquely defined, the number of free parameters in (10) is

$$pq + p - \tfrac{1}{2}q(q-1). \tag{12}$$

Thus with this representation (10), the reduction in the number of parameters for $\boldsymbol{\Sigma}_i$ is

$$\begin{aligned} r &= \tfrac{1}{2}p(p+1) - pq - p + \tfrac{1}{2}q(q-1) \\ &= \tfrac{1}{2}\{(p-q)^2 - (p+q)\}, \end{aligned} \tag{13}$$

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

$$d_1 = (g-1) + 2gp + g\{pq - \tfrac{1}{2}q(q-1)\}. \tag{14}$$

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 the sequel, we focus on how the MFA approach can be modified to provide a greater reduction in the number of parameters.

7 Mixtures of Common Factor Analyzers (MCFA)

Baek and McLachlan (2008) have proposed the Mixtures of Common Factor Analyzers (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 $N(\boldsymbol{\xi}_i, \boldsymbol{\Omega}_i)$, 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. The representation (15) is not unique, as it still has the same form if $\boldsymbol{A}$ were to be postmultiplied by any nonsingular matrix. Hence the number of free parameters in $\boldsymbol{A}$ is

$$pq - q^2. \tag{16}$$

To see that the MCFA model as specified by (15) is a special case of the MFA approach as specified by (11), we note that we can rewrite (15) as

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

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 been chosen so that

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

On comparing (17) with (11), 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 McLachlan et al., 2003). It follows from (23) that the ith component-covariance matrix $\boldsymbol{\Sigma}_i$ has the form

$$\boldsymbol{\Sigma}_i = \boldsymbol{B}_i\boldsymbol{B}_i^T + \boldsymbol{D} \quad (i = 1, \ldots, g). \tag{25}$$

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 Baek and McLachlan (2008) called 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 $\boldsymbol{Y}_j - \boldsymbol{\mu}_i$.

With the the restrictions (22) and (25) on the component mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$, respectively, the total number of free parameters is

$$d_2 = (g-1) + p + q(p+g) + \tfrac{1}{2}gq(q+1) - q^2. \tag{26}$$

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.

The MCFA approach is similar in form to the approach proposed by Yoshida et al. (2004, 2006) who also imposed the additional restrictions that the common diagonal covariance matrix $\boldsymbol{D}$ of the error terms is spherical,

$$\boldsymbol{D} = \sigma^2 \boldsymbol{I}_p, \tag{27}$$

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

$$d_3 = (g-1) + pq + 1 + 2gq - \tfrac{1}{2}q(q+1). \tag{28}$$

8 Fitting of Factor-Analytic Models

The fitting of mixtures of factor analyzers as with the MFA approach has been considered in McLachlan et al. (2003), using a variant of the EM algorithm known as the alternating expectation-conditional maximization algorithm (AECM). With the MCFA approach, we have to fit the same mixture model of factor analyzers but with the additional restrictions (23) and (25) on the component means $\boldsymbol{\mu}_i$ and covariance matrices $\boldsymbol{\Sigma}_i$. The implementation of the EM algorithm for this model was developed in Baek and McLachlan (2008). 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

$$\begin{aligned} E_{\boldsymbol{\Psi}}\{Z_{ij} \mid \boldsymbol{y}_j\} &= \mathrm{pr}_{\boldsymbol{\Psi}}\{Z_{ij} = 1 \mid \boldsymbol{y}_j\} \\ &= \tau_i(\boldsymbol{y}_j; \boldsymbol{\Psi}) \qquad (i = 1, \ldots, g; j = 1, \ldots, n), \end{aligned} \tag{29}$$

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 (4), it can be expressed under the MCFA model as

$$\tau_i(\boldsymbol{y}_j; \boldsymbol{\Psi}) = \frac{\pi_i \phi(\boldsymbol{y}_j; \boldsymbol{A}\boldsymbol{\xi}_i, \boldsymbol{A}\boldsymbol{\Omega}_i\boldsymbol{A}^T + \boldsymbol{D})}{\sum_{h=1}^{g} \pi_h \phi(\boldsymbol{y}_j; \boldsymbol{A}\boldsymbol{\xi}_h, \boldsymbol{A}\boldsymbol{\Omega}_h\boldsymbol{A}^T + \boldsymbol{D})} \tag{30}$$

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,

$$\boldsymbol{U}_{ij} \mid \boldsymbol{y}_j, z_{ij} = 1 \sim N(\boldsymbol{\xi}_{ij}, \boldsymbol{\Omega}_{iy}), \tag{31}$$

where

$$\boldsymbol{\xi}_{ij} = \boldsymbol{\xi}_i + \boldsymbol{\gamma}_i^T(\boldsymbol{y}_j - \boldsymbol{A}\boldsymbol{\xi}_i) \tag{32}$$

and

$$\boldsymbol{\Omega}_{iy} = (\boldsymbol{I}_q - \boldsymbol{\gamma}_i^T\boldsymbol{A})\boldsymbol{\Omega}_i, \tag{33}$$

and where

$$\boldsymbol{\gamma}_i = (\boldsymbol{A}\boldsymbol{\Omega}_i\boldsymbol{A}^T + \boldsymbol{D})^{-1}\boldsymbol{A}\boldsymbol{\Omega}_i. \tag{34}$$

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 (31) and (32),

$$\begin{aligned} E(\boldsymbol{U}_{ij} \mid \boldsymbol{y}_j, z_{ij} = 1) &= \boldsymbol{\xi}_{ij} \\ &= \boldsymbol{\xi}_i + \boldsymbol{\gamma}_i^T(\boldsymbol{y}_j - \boldsymbol{A}\boldsymbol{\xi}_i). \end{aligned} \tag{35}$$

We let $\hat{\boldsymbol{u}}_{ij}$ denote the value of the right-hand side of (35) 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

$$\hat{\boldsymbol{u}}_j = \sum_{i=1}^{g} \tau_i(\boldsymbol{y}_j; \hat{\boldsymbol{\Psi}})\, \hat{\boldsymbol{u}}_{ij} \quad (j = 1, \ldots, n). \tag{36}$$

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 (36).

Acknowledgement

The work of J. Baek was supported by the Korea Research Foundation Grant 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.

REFERENCES

[1.] J. Baek and G.J. McLachlan. Mixtures of factor analyzers with common factor loadings for the clustering and visualization of high-dimensional data. Technical Report NI08020-HOP, Preprint Series of the Isaac Newton Institute for Mathematical Sciences, Cambridge, 2008.

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

[3.] D. Coleman, X. Dong, J. Hardin, D. Rocke, and D. Woodruff. Some computational issues in cluster analysis with no a priori metric. *Computational Statistics & Data Analysis*, 31:1–11, 1999.

[4.] 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*, 39:1–38, 1977.

[5.] J. Hartigan. *Clustering Algorithms*. Wiley, New York, 1975.

[6.] C. Hennig. Clusters, outliers and regression: fixed point clusters. *Journal of Multivariate Analysis*, 86:183–212, 2003.

[7.] C. Hennig. Breakdown points for maximum likelihood-estimators of location-scale mixtures. *Annals of Statistics*, 32:1313–1340., 2004.

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

[10.] G.J. McLachlan. The classification and mixture maximum likelihood approaches to cluster analysis. In P.R. Krishnaiah and L. Kanal, editors, *Handbook of Statistics Vol. 2*, pages 199–208. North-Holland, Amsterdam, 1982.

[11.] 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*, 51:5327–5338, 2007.

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

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

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

[15.] G.J. McLachlan and D. Peel. Robust cluster analysis via mixtures of multivariate t-distributions. *Lecture Notes in Computer Science*, 1451:658–666, 1998.

[16.] G.J. McLachlan, D. Peel, K.E. Basford, and P. Adams. The EMMIX software for the fitting of mixtures of normal and t-components. *Journal of Statistical Software*, 4:, No. 2, 1999.

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

[18.] X. Meng and D. van Dyk. The EM algorithm—an old folk song sung to a fast new tune (with discussion). *Journal of the Royal Statistical Society B*, 59:511–567, 1997.

[19.] A. Montanari and C. Viroli. Two layer latent regression. Technical Report, International Statistical Institute, Voorburg, Netherlands. 2007.

[20.] C.R. Rao (1973). *Linear Statistical Inference and its Applications.* Wiley, New York, 1973.

[21.] G. Schwarz. Estimating the dimension of a model. *Annals of Statisics*, 6:461–464, 1978.

[22.] 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*, 161–172, 2004.

[23.] 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*, 22:1538–1539, 2006.