# Mixtures of Factor Analyzers

**Geoffrey McLachlan** GJM@MATHS.UQ.EDU.AU
**David Peel** DP@MATHS.UQ.EDU.AU
Department of Mathematics, University of Queensland, St Lucia, 4072, Australia

## Abstract

In this paper, we shall focus here on mixtures of factor analyzers from the perspective of a method for model-based density estimation from high-dimensional data, and hence for the clustering of such data. This model enables a normal mixture model to be fitted to high-dimensional data. The number of free parameters is controlled through the dimension of the latent factor space. By working in this reduced space it allows an interpolation in model complexities from isotropic to full covariance structures without any restrictions. We shall show how the mixtures of factor analyzers model can be fitted efficiently using an extension of the EM algorithm, called the alternating expectation conditional maximization (AECM) algorithm. We shall also discuss the close link of mixtures of factor analyzers with mixtures of probabilistic principal component analyzers.

## 1. Introduction

Factor analysis is commonly used for explaining data, in particular, correlations between variables in multivariate observations. It can be used also for dimensionality reduction, although the method of principal component analysis is more widely used in this role. However, the effectiveness of these two statistical techniques is limited by their global linearity. The applicability of these two methods can be widened by combining local models of them in the form of a finite mixture, along the lines proposed by Ghahramani and Hinton (1997) and Tipping and Bishop (1997).

In this paper, we shall focus here on mixtures of factor analyzers from the perspective of a method for model-based density estimation from high-dimensional data, and hence for the clustering of such data. We shall also discuss the close link of mixtures of factor analyzers with mixtures of probabilistic principal component analyzers. The mixtures of factor analyzers model enables a normal mixture model to be fitted to high-dimensional data. The number of free parameters is controlled through the dimension $q$ of the latent factor space $q$.

## 2. Single Factor Analysis Model

We let $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ denote a random sample of size $n$ on a $p$-dimensional random vector. In a typical factor analysis model, each observation $\boldsymbol{Y}_j$ is modelled as

$$\boldsymbol{Y}_j = \boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{U}_j + \boldsymbol{e}_j \quad (j = 1, \ldots, n), \tag{1}$$

where $\boldsymbol{U}_j$ is a $q$-dimensional $(q < p)$ vector of latent or unobservable variables called factors and $\boldsymbol{B}$ is a $p \times q$ matrix of factor loadings (parameters). It is assumed that

$$(\boldsymbol{Y}_1^T, \boldsymbol{U}_1^T)^T, \ldots, (\boldsymbol{Y}_n^T, \boldsymbol{U}_n^T)^T$$

are independently and identically distributed. The $\boldsymbol{U}_j$ are assumed to be independently and identically distributed as $N(\boldsymbol{0}, \boldsymbol{I}_q)$, independently of the errors $\boldsymbol{e}_j$, which are assumed to be independently and identically distributed as $N(\boldsymbol{0}, \boldsymbol{D})$, where $\boldsymbol{D}$ is a diagonal matrix and where $\boldsymbol{I}_q$ denotes the $q \times q$ identity matrix. The $\sigma_i^2$ are called the uniquenesses. Thus, conditional on the $u_j$, the $Y_j$ are independently distributed as $N(\boldsymbol{\mu} + \boldsymbol{B}\,\boldsymbol{u}_j,\, \boldsymbol{D})$. Unconditionally, the $\boldsymbol{Y}_j$ are independently and identically distributed according to a normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix

$$\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}^T + \boldsymbol{D}. \tag{2}$$

Under the model (1), the variables in $\boldsymbol{Y}_j$ are conditionally independent given $\boldsymbol{u}_j$. Thus the factors in $\boldsymbol{u}_j$ are intended to explain the correlations between the variables in $\boldsymbol{Y}_j$, while the error terms $\boldsymbol{e}_j$ represent the unexplained noise unique to a particular $\boldsymbol{y}_j \,(j = 1, \ldots, n)$. Note that in the case of $q > 1$, there is an infinity of choices for $\boldsymbol{B}$, since this model is still satisfied if we replace $u_j$ by $\boldsymbol{H}\boldsymbol{u}_j$ and $\boldsymbol{B}$ by $\boldsymbol{B}\boldsymbol{H}^T$, where $\boldsymbol{H}$ is any orthogonal matrix of order

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

$$pq + p - \tfrac{1}{2}q(q-1); \tag{3}$$

see Lawley and Maxwell (1971, Chapter 1).

## 3. EM Algorithm for a Single Factor Analyzer

The factor analysis model (1) can be fitted by maximum likelihood, although the solution has to be computed iteratively as no closed form expressions exist for the ML estimates of $\boldsymbol{B}$ and $\boldsymbol{D}$. They can be computed iteratively the via the EM algorithm as considered in Dempster et al. (1997); see Rubin and Thayer (1982) and McLachlan and Krishnan (1997, Chapter 5) for further details. The ML estimate of the mean $\boldsymbol{\mu}$ is obviously the sample mean $\overline{\boldsymbol{y}}$ of the $n$ observed values $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ corresponding to the random sample $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$. Hence in the sequel, $\boldsymbol{\mu}$ is set equal to $\overline{\boldsymbol{y}}$ without loss of generality. Accordingly, we let the parameter vector $\boldsymbol{\Psi}$ of unknown parameters consist of the elements of $\boldsymbol{B}$ and the diagonal elements of $\boldsymbol{D}$. The (incomplete-data) log likelihood for $\boldsymbol{\Psi}$ that can be formed from the observed data $\boldsymbol{x}_{\mathrm{o}} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T)^T$ is, apart from an additive constant,

$$\log L(\boldsymbol{\Psi}) = -\tfrac{1}{2}n\{\log \mid \boldsymbol{\Sigma} \mid + \sum_{j=1}^{m}(\boldsymbol{y}_j - \overline{\boldsymbol{y}})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_j - \overline{\boldsymbol{y}})\}. \tag{4}$$

In order to apply the EM and its variants to this problem, we follow Dempster et al. (1977) and formulate $\boldsymbol{x}_{\mathrm{c}} = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_m^T)^T$ as the complete-data vector, where

$$\boldsymbol{x}_j = (\boldsymbol{y}_j^T, \boldsymbol{u}_j^T)^T \quad (j = 1, \ldots, n);$$

that is, $\boldsymbol{x}_{\mathrm{miss}} = (\boldsymbol{u}_1^T, \ldots, \boldsymbol{u}_n^T)^T$ is the missing-data vector, where $\boldsymbol{u}_j$ corresponds to $\boldsymbol{U}_j$. The complete-data log likelihood is, but for an additive constant,

$$\log L_c(\boldsymbol{\Psi}) = -\tfrac{1}{2}n \log \mid \boldsymbol{D} \mid - \tfrac{1}{2}\sum_{j=1}^{n}\{\boldsymbol{r}_j^T\boldsymbol{D}^{-1}\boldsymbol{r}_j + \boldsymbol{u}_j^T\boldsymbol{u}_j\}. \tag{5}$$

where $\boldsymbol{r}_j = \boldsymbol{y}_j - \overline{\boldsymbol{y}} - \boldsymbol{B}\boldsymbol{u}_j$ $(j = 1, \ldots, n)$.

It follows (McLachlan and Krishnan (1997, Chapter 5) that after the $(k+1)$ th iteration of the EM algorithm, the updated estimates of $\boldsymbol{B}$ and $\boldsymbol{D}$ are given by

$$\boldsymbol{B}^{(k+1)} = \boldsymbol{C}_{yy}\boldsymbol{\gamma}^{(k)}(\boldsymbol{\gamma}^{(k)^T}\boldsymbol{C}_{yy}\boldsymbol{\gamma}^{(k)} + n\boldsymbol{\omega}^{(k)})^{-1}$$

and

$$\boldsymbol{D}^{(k+1)} = n^{-1}\,\mathrm{diag}\{\boldsymbol{C}_{yy} - \boldsymbol{C}_{yy}\boldsymbol{\gamma}^{(k)}\boldsymbol{B}^{(k+1)^T}\},$$

where

$$\begin{aligned}\boldsymbol{C}_{yy} &= \sum_{j=1}^{n}(\boldsymbol{y}_j - \overline{\boldsymbol{y}})(\boldsymbol{y}_j - \overline{\boldsymbol{y}})^T; \\ \boldsymbol{C}_{yu} &= \sum_{j=1}^{n}(\boldsymbol{y}_j - \overline{\boldsymbol{y}})\boldsymbol{u}_j^T,\end{aligned}$$

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

and

$$\boldsymbol{\omega}^{(k)} = \boldsymbol{I}_q - \boldsymbol{\gamma}^{(k)^T}\boldsymbol{B}^{(k)}.$$

The inversion of the current value of the $p \times p$ matrix $(\boldsymbol{B}\boldsymbol{B}^T + \boldsymbol{D})$ on each iteration can be undertaken using the result that $(\boldsymbol{B}\boldsymbol{B}^T + \boldsymbol{D})^{-1}$ can be expressed as

$$\boldsymbol{D}^{-1} - \boldsymbol{D}^{-1}\boldsymbol{B}(\boldsymbol{I}_q + \boldsymbol{B}^T\boldsymbol{D}^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{D}^{-1}, \tag{6}$$

where the right-hand side of (6) involves only the inverses of $q \times q$ matrices, since $\boldsymbol{D}$ is a diagonal matrix. The determinant of $(\boldsymbol{B}\boldsymbol{B}^T + \boldsymbol{D})$ can then be calculated as

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

## 4. Mixtures of Factor Analyzers

As the single factor analysis model (1) provides only a global linear model for the representation of the data in a lower dimensional subspace, the scope of its application is limited. A global nonlinear approach can be obtained by postulating a finite mixture of linear submodels for the distribution of the full observation vector $\boldsymbol{Y}_j$ given the (unobservable) factors $\boldsymbol{u}_j$. That is, we can provide a local dimensionality reduction method by assuming that the distribution of the observation $\boldsymbol{Y}_j$ can be modelled as

$$\boldsymbol{Y}_j = \boldsymbol{\mu}_i + \boldsymbol{B}_i\boldsymbol{U}_{ij} + \boldsymbol{e}_{ij} \tag{7}$$

with prob. $\pi_i$ $(i = 1, \ldots, g)$ for $j = 1, \ldots, n$, where the factors $\boldsymbol{U}_{i1}, \ldots, \boldsymbol{U}_{in}$ are distributed independently $N(\boldsymbol{0}, \boldsymbol{I}_{\mathbf{q}})$, independently of the the $\boldsymbol{e}_{ij}$, which are distributed independently $N(\boldsymbol{0}, \boldsymbol{D}_{\mathbf{i}})$, where $\boldsymbol{D}_i$ is a diagonal matrix $(i = 1, \ldots, g)$. The so-called mixing proportions $\pi_i$ are nonnegative and sum to one.

Thus, unconditionally, the density of each observation $\boldsymbol{Y}_j$ is a mixture of g normal densities in proportions $\pi_1, \ldots, \pi_g$; that is,

$$f(\boldsymbol{y}_j; \boldsymbol{\Psi}) = \sum_{i=1}^{g}\pi_i\,\phi(\boldsymbol{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{8}$$

where

$$\boldsymbol{\Sigma}_i = \boldsymbol{B}_i \boldsymbol{B}_i^T + \boldsymbol{D}_i \quad (i = 1, \ldots, g). \qquad (9)$$

The parameter vector $\boldsymbol{\Psi}$ now consists of the elements of the $\boldsymbol{\mu}_i$, the $\boldsymbol{B}_i$, and of the $\boldsymbol{D}_i$, along with the mixing proportions $\pi_i \, (i = 1, \ldots, g)$. In Ghahramani and Hinton (1997), $\boldsymbol{D}_i$ is taken to be the same for each component $i$; see also Yung (1997).

The mixtures of factor analyzers model (7) is also useful in the modeling of high-dimensional data by mixtures of normal components, particularly in situations where there are limited data. With the fitting of a mixture of normal components with unrestricted covariance matrices $\boldsymbol{\Sigma}_i$, there are $\frac{1}{2}p(p+1)$ parameters for each $\boldsymbol{\Sigma}_i \, (i = 1, \ldots, g)$. This means that as the number of components $g$ in the mixture model grows, the total number of parameters can quickly become very large relative to the sample size $n$, leading to over fitting. The mixture of factor analyzers provides a way of controlling the number of parameters through the reduced model (9) for the component-covariance matrices. It provides a model intermediate between the independence and unrestricted models. The adequacy of the fit of a mixture of factor analyzers with $q$ factors can be tested using the likelihood ratio statistic. It will have its usual asymptotic null distribution of chi-squared for this problem, although of course it does not for the problem of tests on the number of components.

# 5. AECM Algorithm for Fitting Mixtures of Factor Analyzers

## 5.1 AECM Framework

The log likelihood for $\boldsymbol{\Psi}$ that can be formed from the observed data $\boldsymbol{x}_{\rm o} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T)^T$ is

$$\log L(\boldsymbol{\Psi}) = \sum_{j=1}^{n} \log\{\sum_{i=1}^{g} \pi_i \phi(\boldsymbol{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\}. \qquad (10)$$

We can use the alternating expectation conditional-maximization (AECM) algorithm as developed by Meng and van Dyk (1997) to fit the mixture of factor analyzers model (8) by maximum likelihood. The expectation conditional-maximization (ECM) algorithm proposed by Meng and Rubin (1993) replaces the M-step of the EM algorithm by a number of computationally simpler conditional maximization (CM) steps. The AECM algorithm is an extension of the ECM algorithm, where the specification of the complete-data is allowed to be different on each CM-step.

To apply the AECM algorithm to the fitting of the mixture model (8), we partition the vector of unknown parameters $\boldsymbol{\Psi}$ as $(\boldsymbol{\Psi}_1^T, \boldsymbol{\Psi}_2^T)^T$, where $\boldsymbol{\Psi}_1$ contains the mixing proportions $\pi_i$ and the elements of the component means $\boldsymbol{\mu}_i \, (i = 1, \ldots, g)$. The subvector $\boldsymbol{\Psi}_2$ contains the elements of the $\boldsymbol{B}_i$ and the $\boldsymbol{D}_i \, (i = 1, \ldots, g)$. Concerning the specification of the incomplete data, it is useful to conceptualize each observation $\boldsymbol{y}_j$ as having arisen from one of the components of the mixture and then to declare the component-indicator vector $\boldsymbol{z}_j$ so associated with $\boldsymbol{y}_j$ as missing data. In this framework, $z_{ij} = (\boldsymbol{z}_j)_i$ is one or zero according as $\boldsymbol{y}_j$ arose or did not arise from the $i$th component $(i = 1, \ldots, g;\, j = 1, \ldots, n)$. The conditional expectation of $Z_{ij}$ given $\boldsymbol{y}_j$ is the posterior probability that the $j$th observation comes from the $i$th component, given by

$$\begin{aligned} \tau_i(\boldsymbol{y}_j; \boldsymbol{\Psi}) &= \mathrm{pr}\{Z_{ij} = 1 \mid \boldsymbol{y}_j\} \\ &= \frac{\pi_i \, \phi(\boldsymbol{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{h=1}^{g} \pi_h \phi(\boldsymbol{y}_j; \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)}, \end{aligned} \qquad (11)$$

where $\boldsymbol{\Sigma}_i$ has the form (9) $(i = 1, \ldots, g;\, j = 1, \ldots, n)$.

We let $\boldsymbol{\Psi}^{(k)} = (\boldsymbol{\Psi}_i^{(k)^T}, \boldsymbol{\Psi}_2^{(k)^T})^T$ be the value of $\boldsymbol{\Psi}$ after the $k$th iteration of the AECM algorithm. For this application of the AECM algorithm, one iteration consists of two cycles, and there is one E-step and one CM-step for each cycle. The two CM-steps correspond to the partition of $\boldsymbol{\Psi}$ into the two subvectors $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$.

## 5.2 First Cycle

For the first cycle of the AECM algorithm, we specify the missing data to be just the component-indicator vectors, $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$. The complete-data log likelihood is then given by

$$\log L_c(\boldsymbol{\Psi}) = \sum_{i=1}^{g} \sum_{j=1}^{n} z_{ij} \log\{\pi_i \phi(\boldsymbol{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\}. \qquad (12)$$

Hence the E-step on the first cycle on the $(k+1)$th iteration requires the calculation of $Q_1(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})$, where

$$Q_1(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)}) = E_{\boldsymbol{\Psi}^{(k)}}\{\log L_c(\boldsymbol{\Psi}) \mid \boldsymbol{x}_{\rm o}\}$$

is the conditional expectation of the complete-data log likelihood (12), using $\boldsymbol{\Psi}^{(k)}$ for $\boldsymbol{\Psi}$. This E-step is achieved simply by replacing each $z_{ij}$ in (12) by its current conditional expectation given the observed data (effectively $\boldsymbol{y}_j$); that is, we replace $z_{ij}$ by $\tau_i(\boldsymbol{y}_j; \boldsymbol{\Psi}^{(k)})$. The first CM-step is implemented by maximizing $Q_1(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})$ over $\boldsymbol{\Psi}$ with $\boldsymbol{\Psi}_2$ held fixed at $\boldsymbol{\Psi}_2^{(k)}$. The updated estimate $\boldsymbol{\Psi}_1^{(k+1)}$ of $\boldsymbol{\Psi}_1$ so obtained contains the new estimates of the $\pi_i$ and $\boldsymbol{\mu}_i$ given by

$$\pi_i^{(k+1)} = \sum_{i=1}^{g} \tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k)})/n \tag{13}$$

and

$$\boldsymbol{\mu}_i^{(k+1)} = \sum_{i=1}^{g} \tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k)})\, \boldsymbol{y}_j / \sum_{j=1}^{g} \tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k)}) \tag{14}$$

for $i = 1, \ldots, g$. We now set $\boldsymbol{\Psi}^{(k+1/2)}$ equal to $(\boldsymbol{\Psi}_1^{(k+1)^T},\, \boldsymbol{\Psi}_2^{(k)^T})^T$.

### 5.3 Second Cycle

For the second cycle for the updating of $\boldsymbol{\Psi}_2$, which contains the elements of the $\boldsymbol{B}_i$ and the $\boldsymbol{D}_i$, we specify the missing data to be the factors $\boldsymbol{u}_1, \ldots \boldsymbol{u}_n$, as well as the component-indicator vectors, $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$. The E-step on the second cycle on the $(k+1)$th requires the calculation of $Q_2(\boldsymbol{\Psi};\, \boldsymbol{\Psi}_2^{(k+1/2)})$, which denotes the conditional expectation of (12) given the observed data $\boldsymbol{x}_{\mathrm{o}}$, using $\boldsymbol{\Psi}^{(k+1/2)}$ for $\boldsymbol{\Psi}$. In addition to updating the posterior probabilities of component membership to $\tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k+1/2)})$, it requires calculating the conditional expectation of

$$E_{\boldsymbol{\Psi}^{(k+1/2)}}\{Z_{ij}(\boldsymbol{U}_{ij} - \boldsymbol{\mu}_i) \mid \boldsymbol{y}_j\} \tag{15}$$

and

$$E_{\boldsymbol{\Psi}^{(k+1/2)}}\{Z_{ij}(\boldsymbol{U}_{ij} - \boldsymbol{\mu}_i)(\boldsymbol{U}_{ij} - \boldsymbol{\mu}_i)^T \mid \boldsymbol{y}_j\}, \tag{16}$$

which are given by

$$\tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k+1/2)})\, \boldsymbol{\gamma}_i^{(k)^T}(\boldsymbol{y}_j - \boldsymbol{\mu}_i) \tag{17}$$

and

$$\begin{aligned}\tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k+1/2)})\{&\boldsymbol{\gamma}_i^{(k)^T}(\boldsymbol{y}_j - \boldsymbol{\mu}_i)\\ &(\boldsymbol{y}_j - \boldsymbol{\mu}_i)^T\boldsymbol{\gamma}_i^{(k)} + \boldsymbol{\omega}_i^{(k)}\},\end{aligned} \tag{18}$$

respectively, where

$$\boldsymbol{\gamma}_i^{(k)} = (\boldsymbol{B}_i^{(k)}\boldsymbol{B}_i^{(k)^T} + \boldsymbol{D}_i^{(k)})^{-1}\boldsymbol{B}_i^{(k)} \tag{19}$$

and

$$\boldsymbol{\omega}_i^{(k)} = \boldsymbol{I}_q - \boldsymbol{\gamma}_i^{(k)^T}\boldsymbol{B}_i^{(k)} \tag{20}$$

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

The CM-step on this second cycle is implemented by the maximization of $Q_2(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k+1/2)})$ over $\boldsymbol{\Psi}$ with $\boldsymbol{\Psi}_1$ set equal to $\boldsymbol{\Psi}_1^{(k+1)}$. This yields the updated estimates $\boldsymbol{B}_i^{(k+1)}$ and $\boldsymbol{D}_i^{(k+1)}$ for $\boldsymbol{B}_i$ and $\boldsymbol{D}_i$, given by

$$\boldsymbol{V}_i^{(k+1/2)}\boldsymbol{\gamma}_i^{(k)}(\boldsymbol{\gamma}_i^{(k)^T}\boldsymbol{V}_i^{(k+1/2)}\boldsymbol{\gamma}_i^{(k)} + \boldsymbol{\omega}_i^{(k+1/2)})^{-1} \tag{21}$$

and

$$\mathrm{diag}\{\boldsymbol{V}_i^{(k+1/2)}\boldsymbol{V}_i^{(k+1/2)}\boldsymbol{\gamma}_i^{(k)}\boldsymbol{B}_i^{(k+1)^T}\}, \tag{22}$$

respectively, where

$$\boldsymbol{V}_i^{(k+1/2)} = \frac{\sum_{j=1}^{n}\tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k+1/2)})\, \boldsymbol{d}_{ij}^{(k)}}{\sum_{j=1}^{n}\tau_i(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k+1/2)})} \tag{23}$$

where $\boldsymbol{d}_{ij}^{(k)} = (\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k+1)})(\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k+1)})^T$.

By construction of this AECM algorithm,

$$Q_1(\boldsymbol{\Psi}^{(k+1/2)};\, \boldsymbol{\Psi}^{(k)}) \geq Q_1(\boldsymbol{\Psi}^{(k)};\, \boldsymbol{\Psi}^{(k)}),$$

and

$$Q_2(\boldsymbol{\Psi}^{(k+1)};\, \boldsymbol{\Psi}^{(k+1/2)}) \geq Q_2(\boldsymbol{\Psi}^{(k+1/2)};\, \boldsymbol{\Psi}^{(k+1/2)}),$$

which ensures that

$$L(\boldsymbol{\Psi}^{(k+1/2)}) \geq L(\boldsymbol{\Psi}^{(k)})$$

and

$$L(\boldsymbol{\Psi}^{(k+1)}) \geq L(\boldsymbol{\Psi}^{(k+1/2)}),$$

respectively; see Meng and Rubin (1993) and Meng van Dyk (1997). Thus the (incomplete-data) likelihood $L(\boldsymbol{\Psi})$ is not decreased after each cycle, and hence after each iteration, of the AECM algorithm.

### 5.4 Representation of Original Data

With this $g$-component mixture model of factor analyzers, the dimension of the data is being reduced locally by fitting $g$ normal component distributions with covariance matrices of the form (9), where the contribution of a point to the process is weighted by its posterior probabilities of component membership.

To consider the representation of an original data point $\boldsymbol{y}_j$ in $q$-dimensions, we have that the estimated conditional expectation of the unobservable factor $\boldsymbol{U}_{ij}$ given $\boldsymbol{y}_j$ and also its membership of the $i$th component is from (15) equal to

$$\begin{aligned}\hat{\boldsymbol{u}}_{ij} &= E_{\hat{\boldsymbol{\Psi}}}\{\boldsymbol{U}_{ij} \mid \boldsymbol{y}_j,\, z_{ij} = 1\}\\ &= \hat{\gamma}_i^T(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_i).\end{aligned} \tag{24}$$

If we average these (estimated) conditional expected component scores $\hat{\boldsymbol{u}}_{ij}$ over the estimated posterior distribution of the component membership of $\boldsymbol{y}_j$, we obtain

$$\hat{\boldsymbol{u}}_j = \sum_{i=1}^{g}\tau_i(\boldsymbol{y}_j;\, \hat{\boldsymbol{\Psi}})\, \hat{\gamma}_i^T(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_i) \tag{25}$$

as the estimated factor score corresponding to $\boldsymbol{y}_j$ ($j = 1, \ldots, n$).

We can thus display $\boldsymbol{y}_j$ in two-dimensional space by mapping it either onto the corresponding posterior mean $\hat{\boldsymbol{u}}_j$ for $q = 2$ factors or onto $\hat{\boldsymbol{u}}_{ij}$ for that value of $i$ for which $\hat{z}_{ij} = 1$. These plots of the estimated posterior factor means may be useful for exploring for group structure in a data set, as demonstrated in the series of papers by Tipping and Bishop (1997, 1998) and Bishop and Tipping (1997, 1998) on the use of the related model of mixtures of probabilistic PCAs.

## 6. Link of Factor Analysis with Probabilistic PCA

Principal component analysis (PCA) is a popular technique for reducing the dimension of data, in particular for visualizing the data in two or three dimensions. However, its effectiveness is limited by its global linearity. One nonlinear extension of this technique can be obtained by attempting to combine separate principal component analyses in the form of a finite mixture model. But conventional PCA does not correspond to an underlying probability density function for the data, and so there has been no useful way to combine the separate analyses.

Bishop and Tipping (1998) and Tipping and Bishop (1997, 1998) formulated a mixture of probabilistic principal component analyzers by adopting a model that is closely related to the factor analysis model (7). To discuss now this link between factor analysis and PCA, we firstly consider the case of a single factor analyzer and a single principal component analyzer.

In factor analysis, the subspace defined by the columns of $\boldsymbol{B}$ will generally not correspond to the principal subspace. However, as noted by Tipping and Bishop (1998, 1999), it had been observed in the literature that the factor loadings in $\boldsymbol{B}$ in a single factor analyzer (1) are quite similar to the principal component axes in situations where the estimates of the diagonal matrix $\boldsymbol{D}$ are approximately equal; see, for example, Rao (1955), Anderson (1963), and Basilevsky (1994).

In their model for a probabilistic PCA Tipping and Bishop (1997, 1998) proposed taking the diagonal covariance matrix $\boldsymbol{D}$ for the error terms to have the isotropic structure,

$$\boldsymbol{D} = \sigma^2 \boldsymbol{I}_p. \tag{26}$$

In this isotropic case, they showed that the ML estimates $\hat{\boldsymbol{B}}$ and $\hat{\sigma}^2$ are related to a PCA, since $\hat{\boldsymbol{B}}$ is the matrix whose columns are scaled eigenvectors of the sample covariance matrix $\boldsymbol{V}$ and $\hat{\sigma}^2$ is the average of the variance in the discarded dimensions; see also Tipping and Bishop (1999). More specifically, the likelihood function (10) with $\boldsymbol{D}$ specified by (26), is maximized by

$$\hat{\boldsymbol{B}} = \boldsymbol{A}(\boldsymbol{\Lambda} - \hat{\sigma}^2 \boldsymbol{I}_q)^{1/2}\boldsymbol{R}, \tag{27}$$

where

$$\begin{aligned} \boldsymbol{A} &= (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_q), \\ \boldsymbol{\Lambda} &= \operatorname{diag}(\lambda_1, \ldots, \lambda_q), \\ \hat{\sigma}^2 &= \sum_{h=q+1}^{p} \lambda_h/(p-q), \end{aligned}$$

and $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_q$ are the (unit) eigenvectors, corresponding to the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_q$ of $\boldsymbol{V}$, and $\boldsymbol{R}$ is an arbitrary $q \times q$ orthogonal matrix.

From (24) and (27), we have for the probabilistic PCA model that the estimated conditional expectation of $\boldsymbol{U}_j$ corresponding to $\boldsymbol{y}_j, \hat{\boldsymbol{u}}_j$, is given by

$$\begin{aligned} \hat{\boldsymbol{u}}_j &= \hat{\boldsymbol{B}}^T(\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^T + \hat{\sigma}^2\boldsymbol{I}_p)^{-1}(\boldsymbol{y}_j - \overline{\boldsymbol{y}}) \qquad (28) \\ &= (\hat{\boldsymbol{B}}^T\hat{\boldsymbol{B}} + \hat{\sigma}^2\boldsymbol{I}_q)^{-1}\hat{\boldsymbol{B}}^T(\boldsymbol{y}_j - \overline{\boldsymbol{y}}), \qquad (29) \end{aligned}$$

where (29) follows from (28) on noting that

$$\hat{\boldsymbol{B}}^T(\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^T + \hat{\sigma}^2\boldsymbol{I}_p)^{-1} = (\hat{\boldsymbol{B}}^T\hat{\boldsymbol{B}} + \hat{\sigma}^2\boldsymbol{I}_q)^{-1}\hat{\boldsymbol{B}}^T.$$

Although the columns of $\hat{\boldsymbol{B}}$ in a probabilistic PCA will span the principal subspace, they need not be an orthogonal projection in the factor space, since

$$\hat{\boldsymbol{B}}^T\hat{\boldsymbol{B}} = \boldsymbol{R}^T(\boldsymbol{\Lambda} - \hat{\sigma}^2\boldsymbol{I}_q)\boldsymbol{R},$$

which is not diagonal for $\boldsymbol{R} \neq \boldsymbol{I}_q$ (Tipping and Bishop, 1999). As $\hat{\sigma}^2 \to 0$, we see from (29) that the probabilistic PCA becomes equivalent in the limit to standard PCA. On further distinctions between factor analysis and probabilistic PCA, neither of the factors found by, say, a two-factor model is necessarily the same as that found by a single-factor model. However, with probabilistic PCA, we see that the principal axes can be found incrementally. On the effect of transformations of the original data, $\boldsymbol{H}\boldsymbol{y}_j$, the form of the factor analysis solution is preserved under diagonal $\boldsymbol{H}$; that is, under scaling of each variable in the observation vector $\boldsymbol{y}_j$. On the other hand, $\boldsymbol{H}$ needs to be orthogonal (corresponding to a rotation of the axes) for the probabilistic PCA solution to be preserved. This is where factor analysis and standard PCA fundamentally differ. With the factor analysis model (7), the latent variables are intended to explain the correlations between the variables in the observation vector $\boldsymbol{Y}_j$, while the error term $\boldsymbol{e}_j$ represents variability unique to a particular $\boldsymbol{Y}_j$. With standard PCA, covariance and variance terms are treated identically.

## 7. Mixtures of Probabilistic PCAs

The mixtures of PCAs as proposed by Tipping and Bishop (1997) has the form (8) and (9) with each $\boldsymbol{D}_i$ now having the isotropic structure

$$\boldsymbol{D}_i = \sigma_i^2 \boldsymbol{I}_p \quad (i = 1, \ldots, g). \tag{30}$$

From the previous section, it follows that under this isotropic restriction assumption (30) the CM-step for the updating of $\boldsymbol{B}_i$ and $\boldsymbol{D}_i$ is not necessary since, given the component membership of the mixture of PCAs, $\boldsymbol{B}_i^{(k+1)}$ and $\sigma_i^{(k+1)^2}$ are given explicitly by an eigenvalue decomposition of the $p \times p$ component-weighted covariance matrix, $\boldsymbol{V}_i^{(k+1/2)}$. That is, application of the EM algorithm will suffice with just the component-indicator vectors as the declared missing data. However, we may still wish to use the AECM algorithm as described in Section 5, since it necessitates only the inversion of $q \times q$ matrices and is of computational complexity $O(q^3)$. Under the additional isotropic restriction (30), the common diagonal element $\sigma_i^2$ of $\boldsymbol{D}_i$ is estimated by the trace of the right-hand side of (23) divided by $p$.

Hinton, Dayan, and Revow (1997) have given an excellent discussion of the use of mixtures of factor and probabilistic analyzers in the context of the clustering of handwritten digits. More recently, Bishop, Svensén, and Williams (1999) have considered a nonlinear form of factor analysis by replacing the factor vector $\boldsymbol{U}_{ij}$ in (7) by a nonlinear function so as to give a nonlinear mapping from the observed data space to the latent space. They call their latent model the generative topographic mapping; see also Bishop (1998).

We can make use of the link of factor analysis with the probabilistic PCA model (30) to specify an initial value $\boldsymbol{\Psi}^{(0)}$ for $\boldsymbol{\Psi}$ in the ML fitting of the mixture of factor analyzers via the AECM algorithm, as described in Section 5; see also Ghahramani and Hinton (1997).

## 8. Example 5.1: Wine Data

To illustrate the use of the mixture of factor analyzers model in a cluster analysis, we examine the so-called wine data set, as available at the machine learning repository of the University of California, Irvine (Merz, Murphy, and Aha, 1997). These data give the results of a chemical analysis of wines grown in the same region in Italy, but derived from three different cultivars. The analysis determined the quantities of $p = 13$ constituents found in each of $n = 178$ wines.

We now consider the usefulness of the mixture of factor analyzers as a model for density estimation for providing a model-based approach to clustering. To cluster this data set, we can model the density of each $p = 13$ dimensional observation $\boldsymbol{y}_j$ by a $g = 3$ component normal mixture model. A soft or probabilistic clustering of the data can be obtained in terms of the estimated posterior probabilities of component membership, $\tau_i(\boldsymbol{y}_j; \hat{\boldsymbol{\Psi}})$; a hard or outright clustering is obtained by assigning each observation to the component to which it has the highest estimated posterior probability of belonging. As $p = 13$ in this application, the covariance matrix $\boldsymbol{\Sigma}_i$ in the $i$th component normal distribution has 91 parameters for each $i$ $(i = 1, 2, 3)$, which means that the total number of parameters is very large relative to the sample size of $n = 178$. Hence we consider fitting a mixture of factor analyzers for small values of $q$ as a way of reducing the number of parameters to be fitted.

*Table 1.* Overall Error Rates for Values of $q$.

| $q$ | $\log L$ | Err | %Err | $-2\log\lambda$ |
|---|---|---|---|---|
| 1 | -3102.254 | 2 | 1.12 | - |
| 2 | -2995.334 | 1 | 0.56 | 213.840 |
| 3 | -2913.122 | 1 | 0.56 | 164.424 |
| 4 | -2871.655 | 3 | 1.69 | 82.934 |
| 5 | -2831.860 | 4 | 2.25 | 79.590 |
| 6 | -2811.290 | 4 | 2.25 | 41.140 |
| 7 | -2799.204 | 4 | 2.25 | 24.172 |
| 8 | -2788.542 | 4 | 2.25 | 21.324 |
| Full | -2813.153 | 7 | 3.93 | – |
| True | -2781.244 | 1 | 0.56 | – |

In Table 1, we report the overall error rate of the normal mixture model-based approach using restricted forms of the $\boldsymbol{\Sigma}_i$ through the adoption of factor analyzer components for various levels of the number of factors $q$ from $q = 1$ to $q = 8$, and for the full model (no restrictions on the $\boldsymbol{\Sigma}_i$). In the latter case, we have also reported the results obtained when the AECM algorithm is started from the true classification of these data. We have also reported the value of minus twice the likelihood ratio statistic $\lambda$ (that is, twice the increase in the log likelihood) as we proceed from fitting a mixture of $q$ factor analyzers to one with $q + 1$ component factors.

The error rates reported in Table 1 are for the case where the normal mixture models were fitted using ten random starting points with each application of the EM algorithm for unrestricted $\boldsymbol{\Sigma}_i$ and the AECM algorithm for factor analyzer components. this was done so as to reflect the situation in practice where the group structure of the data is unknown in a cluster analysis. We did also fit these mixture models using the known classification of the data to start the EM and AECM algorithms, and they gave an error rate of 0.56% for each model. It can be seen from Table 1, that the error rate of the clustering based on the mixture model with unrestricted component-covariance matrices (that is, the full model) is increased when the known group-membership of the data is not used to start the EM algorithm. This lack of knowledge affects also the clustering for all but $q = 2$ and 3 factors, as we can achieve an error rate of one for the other values of $q$ when this knowledge is used to start the AECM algorithm.

The log likelihood reported in Table 1 in the case where the full model is fitted without restrictions on the component-covariance matrices is smaller than the log likelihood for $q = 6, 7$, and 8. Indeed, when we start the EM algorithm in the latter case using the known classification of the data, we do find a larger local maximum that is greater than the values reported in Table 1 for $q = 6$ to 8.

In practice in a typical cluster analysis, the true classification of the data would not be available. It can be seen from Table 1 that without this knowledge to start the AECM algorithm, the error rate of the outright clustering tends to increase as the number of factors $q$ increases. This apparent error rate is smallest for $q = 2$ and 3. Concerning the use of the likelihood ratio test to decide on the number of factors $q$, the test of $q = q_0 = 6$ versus $q = q_0 + 1 = 7$ is not significant at a conventional level of significance ($P = 0.28$), on taking $-2 \log \lambda$ to be chi-squared with $d = g(p - q_0) = 21$ degrees of freedom under the null hypothesis that $q = q_0 = 6$.

In Table 2, we list the eigenvalues $\lambda_{ih}$ $(h = 1, \ldots, 8)$ of our initial estimates of the component-correlation matrices. It can been seen that for each $i$, the $\lambda_{ih}$ start to fall away after $h = 5$ or 6, which is consistent with the above result based on the likelihood ratio test that $q = 6$ factors would appear to be adequate. It suggests that the relative sizes of the $\lambda_{ih}$ $(h = 1, \ldots, p)$ for each $i$ $(i = 1, \ldots, g)$ can be used as guide to an initial choice for the number of factors $q$. A more informed choice if desired can be made then on the basis of the log likelihood ratio statistic applied for nearby values of $q$.

## 9. Some Other Examples

In this section, we briefly report the results of using the mixtures of factor analyzers model to cluster two further data sets. The first concerns a simulated data set, consisting of $n = 150$ observations generated from a mixture of $g = 4$ groups of $p = 39$ dimensions. The second the so-called sonar data set available from the UCI machine learning repository (Merz et al., 1997). It consists of $n = 104$ observations of $p = 60$ dimensions from $g = 2$ groups.

*Table 2.* Eigenvalues of the Initial Estimates of the Component-Correlation Matrices.

| $h$ | $\lambda_{1h}$ | $\lambda_{2h}$ | $\lambda_{3h}$ |
|---|---|---|---|
| 1 | 13.604 | 7.149 | 17.060 |
| 2 | 7.411 | 4.985 | 8.176 |
| 3 | 4.184 | 4.185 | 7.497 |
| 4 | 3.215 | 2.126 | 4.547 |
| 5 | 2.392 | 1.849 | 2.936 |
| 6 | 2.098 | 1.417 | 2.195 |
| 7 | 1.687 | 0.952 | 2.012 |
| 8 | 1.178 | 0.670 | 0.866 |

For each of these data sets, we report in Table 3 the error rate when it is clustered by fitting a mixtures of factor analyzers model. Also reported in this table is the error rate of the clustering procedure that replaces the $p$ original variables by the first $q$ principal components of the set where, for comparative purposes, $q$ was set equal to the number of factors in the final form of the mixtures of analyzers model applied to the data set. This latter procedure represents perhaps the most commonly used way of clustering data sets where the number of variables is large relative to the number of observations.

It can be seen for each data set that the use of the mixtures of factor analyzers model gives a more effective clustering of the data at hand.

*Table 3.* Overall Error Rates for Values of $q$ for Two Data Sets.

| | Sonar | | Simulated | |
|---|---|---|---|---|
| $q$ | MFA | PC | MFA | PC |
| 1 | 42.3% | 40.4% | 24.7% | 69.3% |
| 2 | 34.6% | 42.3% | 19.3% | 62.0% |
| 3 | 22.1% | 43.3% | 16.7% | 65.3% |
| 4 | 32.7% | 42.3% | 18.7% | 63.3% |
| 5 | 32.7% | 42.3% | 17.3% | 64.0% |
| 6 | 31.7% | 42.3% | 13.3% | 59.3% |
| 7 | 26.0% | 38.5% | 10.7% | 61.3% |
| 8 | 19.2% | 36.5% | 10.0% | 50.7% |
| 9 | 18.3% | 36.5% | 4.7% | 51.3% |
| 10 | 16.4% | 35.6% | 4.7% | 51.3% |
| 20 | 11.5% | 26.9% | 1.3% | 10.7% |

## References

Anderson, T.W. (1963). Asymptotic theory for principal component analysis. *Annals of Mathematical Statistics*, *34*, 122–148.

Basilevsky, A. (1994). *Statistical factor analysis and related methods*. New York: Wiley.

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

Ghahramani, Z., & Hinton, G.E. (1997). *The EM algorithm for factor analyzers* (Technical Report No. CRG-TR-96-1). Toronto: The University of Toronto.

Lawley, D.N., & Maxwell, A.E. (1971). *Factor analysis as a statistical method*. Second Edition. London: Butterworths.

McLachlan, G.J., & Krishnan, T. (1997). *The EM algorithm and extensions*. New York: Wiley.

Merz, C., Murphy, P., & Aha, D.W. (1997). *UCI repository of machine learning databases*. Irvine, CA: University of California, Department of Information and Computer Science.

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

Meng, X.L., & van Dyk, D. (1997). 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.

Rao, C.R. (1955). Estimation and tests of significance in factor analysis. *Psychometrika*, *47*, 69–76.

Rubin, D.B., & Thayer, D.T. (1982). EM algorithms for ML factor analysis. *Psychometrika*, *47*, 69–76.

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

Tipping, M.E., & Bishop. C.M. (1998). Mixtures of probabilistic principal component analysers. *Neural Computation*, *10*, 443–482.

Tipping, M.E., & Bishop, C.M. (1999). Probabilistic principal components. *Journal of the Royal Statistical Society B*, *61*, 611–622.

Yung, Y.-F. (1997). Finite mixtures in confirmatory factor-analysis models. *Psychometrika*, *62*, 297–330.