ELSEVIER Computational Statistics & Data Analysis 41 (2003) 379–388

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Modelling high-dimensional data by mixtures of factor analyzers

G.J. McLachlan*, D. Peel, R.W. Bean

Department of Mathematics, University of Queensland, St. Lucia, Brisbane 4072, Australia

Received 1 March 2002

Abstract

We focus 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 approach enables a normal mixture model to be fitted to a sample of n data points of dimension p, where p is large relative to n. 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. We shall illustrate the use of mixtures of factor analyzers in a practical example that considers the clustering of cell lines on the basis of gene expressions from microarray experiments.

Keywords: Mixture modelling; Factor analyzers; EM algorithm

1. Introduction

Finite mixtures of distributions have provided a mathematical-based approach to the statistical modelling of a wide variety of random phenomena; see, for example, McLachlan and Peel (2000a). For multivariate data of a continuous nature, attention has focussed on the use of multivariate normal components because of their computational convenience. With the normal mixture model-based approach to density estimation and clustering, the density of the (p-dimensional) random variable $\boldsymbol{Y}$ of interest is modelled as a mixture of a number (g) of multivariate normal densities in some

* Corresponding author. Tel.: +61-7-3365-2150; fax: +61-7-3365-1477.
E-mail address: gjm@maths.uq.edu.au (G.J. McLachlan).

0167-9473/03/$ - see front matter
PII: S0167-9473(02)00183-4

unknown proportions $\pi_1,\ldots,\pi_g$. That is, each data point is taken to be a realization of the mixture probability density function (p.d.f.),

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

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

The normal mixture model (1) can be fitted iteratively to an observed random sample $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n$ by maximum likelihood (ML) via the expectation-maximization (EM) algorithm of Dempster et al. (1977); see also McLachlan and Krishnan (1997). The number of components g can be taken sufficiently large to provide an arbitrarily accurate estimate of the underlying density function; see, for example, Li and Barron (2000). For clustering purposes, 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 assignment of the data into g clusters is achieved by assigning each data point to the component to which it has the highest estimated posterior probability of belonging.

The g-component normal mixture model (1) with unrestricted component-covariance matrices is a highly parameterized model with $\frac{1}{2}p(p+1)$ parameters for each component-covariance matrix $\boldsymbol{\Sigma}_i$ $(i=1,\ldots,g)$. Banfield and Raftery (1993) introduced a parameterization of the component-covariance matrix $\boldsymbol{\Sigma}_i$ based on a variant of the standard spectral decomposition of $\boldsymbol{\Sigma}_i$ $(i=1,\ldots,g)$. 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 McLachlan and Peel (2000a, p. 239). This point was also stressed by Chang (1983), who showed in the case of two groups that the principal component of the feature vector that provides the best separation between groups in terms of Mahalanobis distance is not necessarily the first component.

Another approach for reducing the number of unknown parameters in the forms for the component-covariance matrices is to adopt the mixture of factor analyzers model, as considered in McLachlan and Peel (2000a, 2000b). This model was originally proposed by Ghahramani and Hinton (1997) and Hinton et al. (1997) for the purposes of visualizing high dimensional data in a lower dimensional space to explore for group structure; see also Tipping and Bishop (1997, 1999) and Bishop (1998) who considered the related model of mixtures of principal component analyzers for the same purpose. Further references may be found in McLachlan and Peel (2000a, Chapter 8).

In this paper, we investigate further the modelling of high-dimensional data through the use of mixtures of factor analyzers, focussing on computational issues not addressed in McLachlan and Peel (2000a, Chapter 8). We shall also demonstrate the usefulness of the methodology in its application to the clustering of microarray expression data, which is a very important but nonstandard problem in cluster analysis. Initial attempts on this problem used hierarchical clustering, but there is no reason why the clusters

should be hierarchical for this problem. Also, a mixture model-based approach enables the clustering of microarray data to be approached on a sound mathematical basis. Indeed, as remarked by Aitkin et al. (1981), "when clustering samples from a population, no cluster analysis method is a priori believable without a statistical model". For microarray data, the number of tissues n is usually very small relative to the number of genes (the dimension p), and so the use of factor models to represent the component-covariance matrices allows the mixture model to be fitted by working in the lower dimensional space implied by the factors.

2. Single-factor analysis model

Factor analysis is commonly used for explaining data, in particular, correlations between variables in multivariate observations. It can be used also for dimensionality reduction. In a typical factor analysis model, each observation $\boldsymbol{Y}_j$ is modelled as

$$\boldsymbol{Y}_j = \boldsymbol{\mu} + \boldsymbol{B}U_j + \boldsymbol{e}_j \quad (j = 1, \ldots, n), \tag{2}$$

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). The $\boldsymbol{U}_j$ are assumed to be i.i.d. as $N(\boldsymbol{0}, \boldsymbol{I}_q)$, independently of the errors $\boldsymbol{e}_j$, which are assumed to be i.i.d. as $N(\boldsymbol{0}, \boldsymbol{D})$, where $\boldsymbol{D}$ is a diagonal matrix,

$$\boldsymbol{D} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$$

and where $\boldsymbol{I}_q$ denotes the $q \times q$ identity matrix. Thus, conditional on $\boldsymbol{U}_j = \boldsymbol{u}_j$, the $\boldsymbol{Y}_j$ are independently distributed as $N(\boldsymbol{\mu} + \boldsymbol{B}u_j, \boldsymbol{D})$. Unconditionally, the $\boldsymbol{Y}_j$ are i.i.d. according to a normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix

$$\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}^{\mathrm{T}} + \boldsymbol{D}. \tag{3}$$

If q is chosen sufficiently smaller than p, representation (3) imposes some constraints on the component-covariance matrix $\boldsymbol{\Sigma}$ and thus reduces the number of free parameters to be estimated. Note that in the case of $q > 1$, there is an infinity of choices for $\boldsymbol{B}$, since (3) is still satisfied if $\boldsymbol{B}$ is replaced by $\boldsymbol{B}C$, where $\boldsymbol{C}$ is any orthogonal matrix of order q. One (arbitrary) way of uniquely specifying $\boldsymbol{B}$ is to choose the orthogonal matrix $\boldsymbol{C}$ so that $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{B}$ is diagonal (with its diagonal elements arranged in decreasing order); see Lawley and Maxwell (1971, Chapter 1). Assuming that the eigenvalues of $\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}$ are positive and distinct, the condition that $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{B}$ is diagonal as above imposes $\frac{1}{2}q(q-1)$ constraints on the parameters. Hence then the number of free parameters is $pq + p - \frac{1}{2}q(q-1)$.

The factor analysis model (2) can be fitted by the EM algorithm and its variants as to be discussed in the subsequent section for the more general case of mixtures of such models. Note that with the factor analysis model, we avoid having to compute the inverses of iterates of the estimated $p \times p$ covariance matrix $\boldsymbol{\Sigma}$ that may be singular for large p relative to n. This is because the inversion of the current value of the $p \times p$ matrix $(\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}} + \boldsymbol{D})$ on each iteration can be undertaken using the result that

$$(\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}} + \boldsymbol{D})^{-1} = \boldsymbol{D}^{-1} - \boldsymbol{D}^{-1}\boldsymbol{B}(\boldsymbol{I}_q + \boldsymbol{B}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{D}^{-1}, \tag{4}$$

where the right-hand side of (4) involves only the inverses of $q \times q$ matrices, since $\boldsymbol{D}$ is a diagonal matrix. The determinant of $(\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}} + \boldsymbol{D})$ can then be calculated as

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

Unlike the PCA model, the factor analysis model (2) enjoys a powerful invariance property: changes in the scales of the feature variables in $\boldsymbol{y}_j$, appear only as scale changes in the appropriate rows of the matrix $\boldsymbol{B}$ of factor loadings.

3. Mixtures of factor analyzers

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

$$\boldsymbol{Y}_j = \boldsymbol{\mu}_i + \boldsymbol{B}_i\boldsymbol{U}_{ij} + \boldsymbol{e}_{ij} \quad \text{with prob. } \pi_i \ (i = 1,\ldots,g) \tag{5}$$

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

Thus the mixture of factor analyzers model is given by (1), where the ith component-covariance matrix $\boldsymbol{\Sigma}_i$ has the form

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

where $\boldsymbol{B}_i$ is a $p \times q$ matrix of factor loadings and $\boldsymbol{D}_i$ is a diagonal matrix $(i = 1,\ldots,g)$. The parameter vector $\boldsymbol{\Psi}$ now consists of the elements of the $\boldsymbol{\mu}_i$, the $\boldsymbol{B}_i$, and the $\boldsymbol{D}_i$, along with the mixing proportions π_i $(i = 1,\ldots,g-1)$, on putting $\pi_g = 1 - \sum_{i=1}^{g-1}\pi_i$.

4. Maximum likelihood estimation of mixture of factor analyzers models

The mixture of factor analyzers model can be fitted by using the alternating expectation–conditional maximization (AECM) algorithm (Meng and van Dyk, 1997). 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 of factor analyzers model, we partition the vector of unknown parameters $\boldsymbol{\Psi}$ as $(\boldsymbol{\Psi}_1^{\mathrm{T}}, \boldsymbol{\Psi}_2^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\Psi}_1$ contains the mixing proportions π_i $(i = 1,\ldots,g-1)$ and the elements of the component means $\boldsymbol{\mu}_i$ $(i = 1,\ldots,g)$. The subvector $\boldsymbol{\Psi}_2$ contains the elements of the $\boldsymbol{B}_i$ and the $\boldsymbol{D}_i$ $(i = 1,\ldots,g)$.

We let $\boldsymbol{\Psi}^{(k)} = (\boldsymbol{\Psi}_1^{(k)^{\mathrm{T}}}, \boldsymbol{\Psi}_2^{(k)^{\mathrm{T}}})^{\mathrm{T}}$ be the value of $\boldsymbol{\Psi}$ after the kth iteration of the AECM algorithm. For this application of the AECM algorithm, one iteration consists of two

cycles, and there is one E-step and 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 $\boldsymbol{\Psi}_2$.

For the first cycle of the AECM algorithm, we specify the missing data to be just the component-indicator vectors, $\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n$, where $z_{ij}=(\boldsymbol{z}_j)_i$ is one or zero, according to whether $\boldsymbol{y}_j$ arose or did not arise from the ith component ($i=1,\ldots,g; j=1,\ldots,n$). The first conditional CM-step leads to $\pi_i^{(k)}$ and $\boldsymbol{\mu}_i^{(k)}$ being updated to

$$\pi_i^{(k+1)}=\sum_{j=1}^{n}\tau_i(\boldsymbol{y}_j;\boldsymbol{\Psi}^{(k)})/n \tag{7}$$

and

$$\boldsymbol{\mu}_i^{(k+1)}=\sum_{j=1}^{n}\tau_i(\boldsymbol{y}_j;\boldsymbol{\Psi}^{(k)})\boldsymbol{y}_j \Bigg/ \sum_{j=1}^{n}\tau_i(\boldsymbol{y}_j;\boldsymbol{\Psi}^{(k)}) \tag{8}$$

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

$$\tau_i(\boldsymbol{y}_j;\boldsymbol{\Psi})=\pi_i\phi(\boldsymbol{y}_j;\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i) \Bigg/ \sum_{h=1}^{g}\pi_h\phi(\boldsymbol{y}_j;\boldsymbol{\mu}_h,\boldsymbol{\Sigma}_h) \tag{9}$$

is the ith component posterior probability of $\boldsymbol{y}_j$.

For the second cycle for the updating of $\boldsymbol{\Psi}_2$, we specify the missing data to be the factors $\boldsymbol{u}_1,\ldots,\boldsymbol{u}_n$, as well as the component-indicator vectors, $\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n$. On setting $\boldsymbol{\Psi}^{(k+1/2)}$ equal to $(\boldsymbol{\Psi}_1^{(k+1)^{\mathrm{T}}},\boldsymbol{\Psi}_2^{(k)^{\mathrm{T}}})^{\mathrm{T}}$, an E-step is performed to calculate $Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k+1/2)})$, which is the conditional expectation of the complete-data log likelihood given the observed data, using $\boldsymbol{\Psi}=\boldsymbol{\Psi}^{(k+1/2)}$. The CM-step on this second cycle is implemented by the maximization of $Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k+1/2)})$ over $\boldsymbol{\Psi}$ with $\boldsymbol{\Psi}_1$ set equal to $\boldsymbol{\Psi}_1^{(k+1)}$. This yields the updated estimates $\boldsymbol{B}_i^{(k+1)}$ and $\boldsymbol{D}_i^{(k+1)}$. The former is given by

$$\boldsymbol{B}_i^{(k+1)}=\boldsymbol{V}_i^{(k+1/2)}\boldsymbol{\gamma}_i^{(k)}(\boldsymbol{\gamma}_i^{(k)^{\mathrm{T}}}\boldsymbol{V}_i^{(k+1/2)}\boldsymbol{\gamma}_i^{(k)}+\boldsymbol{\omega}_i^{(k)})^{-1}, \tag{10}$$

where

$$\boldsymbol{V}_i^{(k+1/2)}=\frac{\sum_{j=1}^{n}\tau_i(\boldsymbol{y}_j;\boldsymbol{\Psi}^{(k+1/2)})(\boldsymbol{y}_j-\boldsymbol{\mu}_i^{(k+1)})(\boldsymbol{y}_j-\boldsymbol{\mu}_i^{(k+1)})^{\mathrm{T}}}{\sum_{j=1}^{n}\tau_i(\boldsymbol{y}_j;\boldsymbol{\Psi}^{(k+1/2)})}, \tag{11}$$

$$\boldsymbol{\gamma}_i^{(k)}=(\boldsymbol{B}_i^{(k)}\boldsymbol{B}_i^{(k)^{\mathrm{T}}}+\boldsymbol{D}_i^{(k)})^{-1}\boldsymbol{B}_i^{(k)} \tag{12}$$

and

$$\boldsymbol{\omega}_i^{(k)}=\boldsymbol{I}_q-\boldsymbol{\gamma}_i^{(k)^{\mathrm{T}}}\boldsymbol{B}_i^{(k)} \tag{13}$$

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

$$\begin{aligned}\boldsymbol{D}_i^{(k+1)}&=\mathrm{diag}\{\boldsymbol{V}_i^{(k+1/2)}-\boldsymbol{B}_i^{(k+1)}\boldsymbol{H}_i^{(k+1/2)}\boldsymbol{B}_i^{(k+1)^{\mathrm{T}}}\}\\&=\mathrm{diag}\{\boldsymbol{V}_i^{(k+1/2)}-\boldsymbol{V}_i^{(k+1/2)}\boldsymbol{\gamma}_i^{(k)}\boldsymbol{B}_i^{(k+1)^{\mathrm{T}}}\},\end{aligned} \tag{14}$$

where

$$\boldsymbol{H}_i^{(k+1/2)} = \frac{\sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j; \boldsymbol{\Psi}^{(k+1/2)}) E_i^{(k+1/2)}(\boldsymbol{U}_j \boldsymbol{U}_j^{\mathrm{T}} \mid \boldsymbol{y}_j)}{\sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j; \boldsymbol{\Psi}^{(k+1/2)})},$$

$$= \boldsymbol{\gamma}_i^{(k)^{\mathrm{T}}} \boldsymbol{V}_i^{(k+1/2)} \boldsymbol{\gamma}_i^{(k)} + \boldsymbol{\omega}_i^{(k)} \tag{15}$$

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

Direct differentiation of the log-likelihood function shows that the ML estimate of the diagonal matrix $\boldsymbol{D}_i$ satisfies

$$\hat{\boldsymbol{D}}_i = \mathrm{diag}(\hat{\boldsymbol{V}}_i - \hat{\boldsymbol{B}}_i \hat{\boldsymbol{B}}_i^{\mathrm{T}}), \tag{16}$$

where

$$\hat{\boldsymbol{V}}_i = \sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j; \hat{\boldsymbol{\Psi}})(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_i)^{\mathrm{T}} \Big/ \sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j; \hat{\boldsymbol{\Psi}}). \tag{17}$$

As remarked by Lawley and Maxwell (1971, p. 30) in the context of direct computation of the ML estimate for a single-component factor analysis model, Eq. (16) looks temptingly simple to use to solve for $\hat{\boldsymbol{D}}_i$, but was not recommended due to convergence problems.

On comparing (16) with (14), it can be seen that with the calculation of the ML estimate of $\boldsymbol{D}_i$ directly from the (incomplete-data) log-likelihood function, the unconditional expectation of $\boldsymbol{U}_j \boldsymbol{U}_j^{\mathrm{T}}$, which is the identity matrix, is used in place of the conditional expectation in (15) on the E-step of the AECM algorithm. Unlike the direct approach of calculating the ML estimate, the EM algorithm and its variants such as the AECM version have good convergence properties in that they ensure the likelihood is not decreased after each iteration regardless of the choice of starting point.

It can be seen from (16) that some of the estimates of the elements of the diagonal matrix $\boldsymbol{D}_i$ (the uniquenesses) will be close to zero if effectively not more than q observations are unequivocally assigned to the ith component of the mixture in terms of the fitted posterior probabilities of component membership. This will lead to spikes or near singularities in the likelihood. One way to avoid this is to impose the condition of a common value $\boldsymbol{D}$ for the $\boldsymbol{D}_i$,

$$\boldsymbol{D}_i = \boldsymbol{D} \quad (i = 1, \ldots, g). \tag{18}$$

An alternative way of proceeding is to adopt some prior distribution for the $\boldsymbol{D}_i$ as in the Bayesian approaches of Fokoué and Titterington (2000), Ghahramani and Beal (2000) and Utsugi and Kumagai (2001).

The mixture of probabilistic component analyzers (PCAs) model, as proposed by Tipping and Bishop (1997), has form (6) with each $\boldsymbol{D}_i$ now having the isotropic structure

$$\boldsymbol{D}_i = \sigma_i^2 \boldsymbol{I}_p \quad (i = 1, \ldots, g). \tag{19}$$

Under this isotropic restriction (19) the iterative 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 current value of $\boldsymbol{V}_i$.

5. Initialization of AECM algorithm

We can make use of the link of factor analysis with the probabilistic PCA model (19) 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. On noting that the transformed data $\boldsymbol{D}_i^{-1/2}\boldsymbol{Y}_j$ satisfies the probabilitistic PCA model (19) with $\sigma_i^2 = 1$, it follows that for a given $\boldsymbol{D}_i^{(0)}$ and $\boldsymbol{\Sigma}_i^{(0)}$, we can specify $\boldsymbol{B}_i^{(0)}$ as

$$\boldsymbol{B}_i^{(0)} = \boldsymbol{D}_i^{(0)^{1/2}}\boldsymbol{A}_i(\boldsymbol{\Lambda}_i - \tilde{\sigma}_i^2\boldsymbol{I}_q)^{1/2} \quad (i = 1,\ldots,g), \tag{20}$$

where

$$\tilde{\sigma}_i^2 = \sum_{h=q+1}^{p} \lambda_{ih}/(p-q).$$

The q columns of the matrix $\boldsymbol{A}_i$ are the eigenvectors corresponding to the eigenvalues $\lambda_{i1} \geqslant \lambda_{i2} \geqslant \cdots \geqslant \lambda_{iq}$ of

$$\boldsymbol{D}_i^{(0)^{-1/2}}\boldsymbol{\Sigma}_i^{(0)}\boldsymbol{D}_i^{(0)^{-1/2}} \tag{21}$$

and $\boldsymbol{\Lambda}_i = \mathrm{diag}(\lambda_{i1},\ldots,\lambda_{iq})$. The use of $\tilde{\sigma}_i^2$ instead of unity is proposed in (20), because it avoids the possibility of negative values for $(\boldsymbol{\Lambda}_i - \boldsymbol{I}_q)$, which can occur since estimates are being used for the unknown values of $\boldsymbol{D}_i$ and $\boldsymbol{\Sigma}_i$ in (21).

To specify $\boldsymbol{\Sigma}_i^{(0)}$ for use in (21), we can randomly assign the data into g groups and take $\boldsymbol{\Sigma}_i^{(0)}$ to be the sample covariance matrix of the ith group $(i = 1,\ldots,g)$. Concerning the choice of $\boldsymbol{D}_i^{(0)}$, we can take $\boldsymbol{D}_i^{(0)}$ to be the diagonal matrix formed from the diagonal elements of $\boldsymbol{\Sigma}_i^{(0)}$ $(i = 1,\ldots,g)$. In this case, the matrix (21) has the form of a correlation matrix.

The eigenvalues and eigenvectors for use in (21) can be found by a singular value decomposition of each $p \times p$ sample component-covariance matrix $\boldsymbol{\Sigma}_i^{(0)}$. But if the number of dimensions p is appreciably greater than the sample size n, then it is much quicker to find them by a singular value decomposition of the $n_i \times n_i$ matrix $\tilde{\boldsymbol{\Sigma}}_i^{(0)}$, the sample matrix formed by taking the observations to be the rows rather than the columns of the $p \times n_i$ data matrix whose n_i columns are the p-dimensional observations assigned initially to the ith component $(i = 1,\ldots,g)$. The eigenvalues of this latter matrix are equal to those of $\boldsymbol{\Sigma}_i^{(0)}$ apart from a common multiplier due to the different divisors in their formation.

A formal test for the number of factors can be undertaken using the likelihood ratio λ, as regularity conditions 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 of 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.

6. Example: colon data

In this example, we consider the clustering of tissue samples on the basis of two thousand genes for the colon data of Alon et al. (1999). They used Affymetrix oligonucleotide arrays to monitor absolute measurements on expressions of over 6500 human gene expressions in 40 tumour and 22 normal colon tissue samples, These samples were taken from 40 different patients so that 22 patients supplied both a tumour and normal tissue sample. Alon et al. (1999) focussed on the 2000 genes with highest minimal intensity across the samples, and it is these 2000 genes that comprised our data set. The matrix $\boldsymbol{A}$ of microarray data for this data set thus has $p = 2000$ rows and $n = 62$ columns. Before we considered the clustering of this set, we processed the data by taking the (natural) logarithm of each expression level in the matrix $\boldsymbol{A}$. Then each column of this matrix was standardized to have mean zero and unit standard deviation. Finally, each row of the consequent matrix was standardized to have mean zero and unit standard deviation.

We are unable to proceed directly with the fitting of a normal mixture model to these data in this form. But even if we were able to do so, it is not perhaps the ideal way of proceeding because with such a large number p of feature variables, there will be a lot of noise introduced into the problem and this noise is unable to be modelled adequately because of the very small number ($n = 62$) of observations available relative to the dimension $p = 2000$ of each observation. We therefore applied the screening procedure in the software EMMIX-GENE of McLachlan et al. (2001). With this screening procedure, the genes are ranked in decreasing size of $-2\log\lambda$, where λ is essentially the likelihood ratio statistic for the test of $g = 1$ versus $g = 2$ component t distributions fitted to the 62 tissues with each gene considered individually. If the value of $-2\log\lambda$ were greater than some threshold (here taken to be 8) but the minimum size of the implied clusters was less than some threshold (here taken to be 8 also), this value of λ was replaced by its value for the test of $g = 2$ versus 3 components. This screening of the genes here resulted in 446 genes being retained.

We first clustered the $n = 62$ tissues on the basis of the retained set of 446 genes. We fitted mixtures of factor analyzers for various levels of the number q of factors ranging from $q = 2$ to 8. Using 50 random and 50 k-means-based starts, the clustering corresponding to the largest of the local maxima obtained gave the following clustering for $q = 6$ factors,

$$C_1 = \{1 - 12, 20, 25, 41 - 52\} \cup \{13 - 39, 21 - 24, 26 - 40, 53 - 62\}. \qquad (22)$$

Getz et al. (2000) and Getz (2001) reported that there was a change in the protocol during the conduct of the microarray experiments. The 11 tumour tissue samples

(labelled 1–11 here) and 11 normal tissue samples (41–51) were taken from the first 11 patients using a poly detector, while the 29 tumour tissue samples (12–40) and normal tissue samples (52–62) were taken from the remaining 29 patients using total extraction of RNA. It can be seen from (22) that this clustering C_1 almost corresponds to the dichotomy between tissues obtained under the "old" and "new" protocols. A more detailed account of mixture model-based clustering of this colon data set may be found in McLachlan et al. (2001).

References

Aitkin, M., Anderson, D., Hinde, J., 1981. Statistical modelling of data on teaching styles (with discussion) J. Roy. Statist. Soc. Ser. B 144, 419–461.

Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J., 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Nat. Acad. Sci. 96, 6745–6750.

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

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

Chang, W.C., 1983. On using principal components before separating a mixture of two multivariate normal distributions. Appl. Statist. 32, 267–275.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion) J. Roy. Statist. Soc. Ser. B 39, 1–38.

Fokoué, E., Titterington, D.M., 2000. Bayesian sampling for mixtures of factor analysers. Technical Report, Department of Statistics, University of Glasgow, Glasgow.

Getz, G., 2001. Private communication.

Getz, G., Levine, E., Domany, E., 2000. Coupled two-way clustering analysis of gene microarray data. Cell Biol. 97, 12079–12084.

Ghahramani, Z., Beal, M.J., 2000. Variational inference for Bayesian mixtures of factor analyzers. In: Solla, S.A., Leen, T.K., Miller, K.-R. (Eds.), Neural Information Processing Systems 12. MIT Press, MA, pp. 449–455.

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

Hinton, G.E., Dayan, P., Revow, M., 1997. Modeling the manifolds of images of handwritten digits. IEEE Trans. Neural Networks 8, 65–73.

Lawley, D.N., Maxwell, A.E., 1971. Factor Analysis as a Statistical Method, 2nd Edition. Butterworths, London.

Li, J.Q., Barron, A.R., 2000. Mixture density estimation. Technical Report, Department of Statistics, Yale University, New Haven, Connecticut.

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

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

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

McLachlan, G.J., Bean, R.W., Peel, D., 2001. EMMIX-GENE: a mixture model-based program for the clustering of microarray expression data. Technical Report, Centre for Statistics, University of Queensland.

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) J. Roy. Statist. Soc. Ser. B 59, 511–567.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6, 461–464.

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

Tipping, M.E., Bishop, C.M., 1999. Mixtures of probabilistic principal component analysers. Neural Comput. 11, 443–482.

Utsugi, A., Kumagai, T., 2001. Bayesian analysis of mixtures of factor analyzers. Neural Comput. 13, 993–1002.