BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 9 2011, pages 1269–1276
doi:10.1093/bioinformatics/btr112

Gene expression Advance Access publication March 3, 2011

Mixtures of common *t*-factor analyzers for clustering high-dimensional microarray data

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

[1]Department of Statistics, Chonnam National University, Gwangju 500-757, South Korea and [2]Department of Mathematics and Institute for Molecular Bioscience, University of Queensland, Brisbane QLD 4072, Australia

Associate Editor: Trey Ideker

ABSTRACT

Motivation: Mixtures of factor analyzers enable model-based clustering to be undertaken for high-dimensional microarray data, where the number of observations n is small relative to the number of genes p. Moreover, when the number of clusters is not small, for example, where there are several different types of cancer, there may be the need to reduce further the number of parameters in the specification of the component-covariance matrices. A further reduction can be achieved by using mixtures of factor analyzers with common component-factor loadings (MCFA), which is a more parsimonious model. However, this approach is sensitive to both non-normality and outliers, which are commonly observed in microarray experiments. This sensitivity of the MCFA approach is due to its being based on a mixture model in which the multivariate normal family of distributions is assumed for the component-error and factor distributions.

Results: An extension to mixtures of *t*-factor analyzers with common component-factor loadings is considered, whereby the multivariate *t*-family is adopted for the component-error and factor distributions. An EM algorithm is developed for the fitting of mixtures of common *t*-factor analyzers. The model can handle data with tails longer than that of the normal distribution, is robust against outliers and allows the data to be displayed in low-dimensional plots. It is applied here to both synthetic data and some microarray gene expression data for clustering and shows its better performance over several existing methods.

Availability: The algorithms were implemented in Matlab. The Matlab code is available at http://blog.naver.com/aggie100.

Contact: jbaek@jnu.ac.kr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 20, 2010; revised on February 10, 2011; accepted on February 27, 2011

1 INTRODUCTION

Model-based methods have been widely used for both clustering and classifying high-dimensional microarray data (McLachlan *et al.*, 2002; Yeung *et al.*, 2001). Thalamuthu *et al.* (2006) compared various clustering techniques and showed that model-based method performed well for microarray gene clustering. The finite normal mixture model with unrestricted component-covariance matrices is a highly parameterized model (McLachlan and Basford, 1988; McLachlan and Peel, 2000a). Banfield and Raftery (1993) introduced a parameterization of the component-covariance matrix based on a variant of the standard spectral decomposition, and its program MCLUST (Fraley and Raftery, 2003) has been often used. But if the number of genes 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 were possible, the results may not be reliable due to potential problems with near-singular estimates of the component-covariance matrices when p is large relative to n. In this case, mixtures of factor analyzers (MFA) is a useful model to reduce the number of parameters by allowing factor-analytic representation of the component-covariance matrices. Hinton *et al.* (1997) proposed the MFA adopting a finite mixture of factor analysis models, which was considered for the purposes of clustering by McLachlan and Peel (2000a, 2000b) and McLachlan *et al.* (2003, 2007). McLachlan *et al.* (2002, 2003) applied MFA to tissue samples with microarray gene expression data for clustering. Martella (2006) used MFA to classify microarray data successfully. Recently, Xie *et al.* (2010) proposed a penalized MFA to allow both selection of effective genes and clustering of high-dimensional data simultaneously. Zhou *et al.* (2009) has proposed another penalized model-based clustering method with unconstrained covariance matrices.

In practice, for example, where there are several different types of cancer, there is often the need to reduce further the number of parameters in the specification of the component-covariance matrices by factor-analytic representations. McNicholas and Murphy (2008) introduced some parsimonious MFA models, which include various MFA models with fewer parameters. Galimberti *et al.* (2009) proposed another parsimonious factor mixture model to allow both dimension reduction and variable selection. Baek and McLachlan (2008) and Baek *et al.* (2010) proposed the use of mixtures of factor analyzers with common component-factor loadings (MCFA) and applied it to a microarray dataset for clustering. The method considerably reduces further the number of parameters, and allows the data to be displayed in low-dimensional plots in a straightforward manner in contrast to MFA. Several analyses of many real datasets, however, have suggested that the empirical distribution of gene expression levels is approximately log-normal or sometimes with a slightly heavier tailed *t*-distribution depending on the biological samples under investigation (Li, 2002). In particular, Giles and Kipling (2003) applied the Shapiro–Wilks test to Affymetrix microarray expression data and concluded that

*To whom correspondence should be addressed.

non-normal distributions are common (up to 46% of probe sets). Lönnstedt and Speed (2002) also noted that outliers occur frequently in microarray experiments. Therefore, the above approaches are sensitive to both non-normality of the data and extreme expression levels as they are based on a mixture model in which the multivariate normal family of distributions is assumed for the component-error and factor distributions. McLachlan *et al.* (2007) extended MFA to incorporate the multivariate t-distribution for the component-error and factor distributions. In this article, we propose an extension of MCFA to incorporate the multivariate t-distribution to handle the data with tails longer than that of the normal distribution.

In the next section, we review briefly the MCFA approach as proposed by Baek and McLachlan (2008) and considered further in Baek *et al.* (2010). We then describe the mixtures of t-factor analyzers model with common factor loadings (MCtFA) and develop its implementation via the expectation-maximization (EM) algorithm. In Section 3, its application is demonstrated in the clustering of two microarray gene expression datasets. The results so obtained illustrate the improved performance of MCtFA over MCLUST, MFA and MCFA for these two datasets. A short discussion is given in Section 4.

2 METHODS

2.1 Mixtures of common t-factor analyzers and its EM algorithm

Finite mixture models are being increasingly used to model the distributions of a wide variety of random phenomena and to cluster datasets; see, for example, McLachlan and Peel (2000a). Let $\boldsymbol{Y}=(Y_1,\ldots,Y_p)^T$ be a p-dimensional vector of feature variables. For continuous features Y_j, the density of $\boldsymbol{Y}$ can be modelled by a mixture of a sufficiently large enough number g of multivariate normal component distributions,

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

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 matrices $\boldsymbol{\Sigma}_i (i=1,\ldots,g)$.

We focus on the use of mixtures of factor analyzers to reduce the number of parameters in the specification of the component-covariance matrices, as discussed in Hinton *et al.* (1997), McLachlan and Peel (2000a) and McLachlan *et al.* (2003). With the factor-analytic representation of the component-covariance matrices, we have that

$$\boldsymbol{\Sigma}_i=\boldsymbol{A}_i\boldsymbol{A}_i^T+\boldsymbol{D}_i \quad (i=1,\ldots,g), \tag{2}$$

where $\boldsymbol{A}_i$ is a $p\times q$ matrix and $\boldsymbol{D}_i$ is a diagonal matrix. To see this, we first note that the MFA approach with the factor-analytic representation (2) on $\boldsymbol{\Sigma}_i$ is equivalent to assuming that the distribution of the difference $\boldsymbol{Y}_j-\boldsymbol{\mu}_i$ can be modelled as

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

for $j=1,\ldots,n$, where the (unobservable) factors $\boldsymbol{U}_{i1},\ldots,\boldsymbol{U}_{in}$ are distributed independently $N(\boldsymbol{0},\boldsymbol{I}_q)$, independent of the $\boldsymbol{e}_{ij}$, which are distributed independently $N(\boldsymbol{0},\boldsymbol{D}_i)$, where $\boldsymbol{D}_i$ is a diagonal matrix $(i=1,\ldots,g)$.

However, this model may not lead to a sufficiently large enough reduction in the number of parameters, particularly if g is not small. Hence for this case, Baek and McLachlan (2008) and Baek *et al.* (2010) proposed the MCFA approach whereby the distribution of $\boldsymbol{Y}_j$ is modelled as

$$\boldsymbol{Y}_j=\boldsymbol{A}\boldsymbol{U}_{ij}+\boldsymbol{e}_{ij} \qquad \text{with prob. } \pi_i \quad (i=1,\ldots,g) \tag{3}$$

for $j=1,\ldots,n$, where the (unobservable) factors $\boldsymbol{U}_{i1},\ldots,\boldsymbol{U}_{in}$ are distributed independently $N(\boldsymbol{\xi}_i,\boldsymbol{\Omega}_i)$, independent of the $\boldsymbol{e}_{ij}$, which are distributed independently $N(\boldsymbol{0},\boldsymbol{D})$, where $\boldsymbol{D}$ is a diagonal matrix $(i=1,\ldots,g)$. Here $\boldsymbol{A}$ is a $p\times q$ matrix of factor loadings, which satisfies $\boldsymbol{A}^T\boldsymbol{A}=\boldsymbol{I}_q$. Then MCFA is considered as the normal mixture model (1) with the restrictions

$$\boldsymbol{\mu}_i=\boldsymbol{A}\boldsymbol{\xi}_i \quad (i=1,\ldots,g) \tag{4}$$

and

$$\boldsymbol{\Sigma}_i=\boldsymbol{A}\boldsymbol{\Omega}_i\boldsymbol{A}^T+\boldsymbol{D} \quad (i=1,\ldots,g), \tag{5}$$

where $\boldsymbol{A}$ is a $p\times q$ matrix, $\boldsymbol{\xi}_i$ is a q-dimensional vector, $\boldsymbol{\Omega}_i$ is a $q\times q$ positive definite symmetric matrix, and $\boldsymbol{D}$ is a diagonal $p\times p$ matrix. With the restrictions (4) and (5) on the component mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$, respectively, the MCFA approach provides a greater reduction in the number of parameters compared with MFA (see Table 1 in Baek *et al.* 2010). The implementation of the EM algorithm to fit this model is described in the Appendix of Baek *et al.* (2010). Another useful feature of the MCFA approach is its ability to portray the results of a clustering in low-dimensional space. We can plot the estimated posterior means $\hat{u}_j$ of the factors [as defined by Equation (34) of Baek *et al.* (2010)] in 2D or 3D space, using the implied cluster labels. In contrast, MFA does not have the ability to project the high-dimensional objects in low-dimensional space since the mean vector of the factor is assumed to be $\boldsymbol{0}$ for each cluster. This is illustrated in McLachlan and Peel (2000a, Chapter 8).

Now we assume that $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n$ are an observed random sample from the t-mixture density

$$f(\boldsymbol{y};\boldsymbol{\Psi})=\sum_{i=1}^{g}\pi_i f_t(\boldsymbol{y};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i,\nu_i), \tag{6}$$

where $\boldsymbol{\mu}_i=\boldsymbol{A}\boldsymbol{\xi}_i$ and

$$\boldsymbol{\Sigma}_i=\boldsymbol{A}\boldsymbol{\Omega}_i\boldsymbol{A}^T+\boldsymbol{D} \ (i=1,\ldots,g), \tag{7}$$

and where the multivariate t-probability density function $f_t(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)$ is defined as

$$f_t(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)=\frac{\Gamma((\nu+p)/2)|\boldsymbol{\Sigma}|^{-1/2}}{(\pi\nu)^{p/2}\Gamma(\nu/2)\{1+\delta(\boldsymbol{y},\boldsymbol{\mu};\boldsymbol{\Sigma})/\nu\}^{(\nu+p)/2}},$$

where $\delta(\boldsymbol{y},\boldsymbol{\mu};\boldsymbol{\Sigma})=(\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})$, and the vector of unknown parameters $\boldsymbol{\Psi}$ consists of the degrees of freedom ν_i in addition to the mixing proportions π_i and the elements of the $\boldsymbol{\xi}_i,\boldsymbol{\Omega}_i,\boldsymbol{A}$ and $\boldsymbol{D}$ $(i=1,\ldots,g)$. As in the mixture of common factor analyzers model, $\boldsymbol{A}$ is a $p\times q$ matrix and $\boldsymbol{D}$ is a diagonal matrix.

In order to fit the model (6) with the restriction (7), it is computationally convenient to exploit its link with factor analysis. Therefore, we assume that the distribution of $\boldsymbol{Y}_j$ of MCtFA is modelled as (3), where the joint distribution of the factor $\boldsymbol{U}_{ij}$ and the error $\boldsymbol{e}_{ij}$ needs to be specified so that it is consistent with the t-mixture formulation (6) for the marginal distribution of $\boldsymbol{Y}_j$. In the EM framework, the component label $\boldsymbol{z}_j$ associated with the observation $\boldsymbol{y}_j$ is introduced as missing data, where $z_{ij}=(\boldsymbol{z}_j)_i$ is one or zero according as $\boldsymbol{y}_j$ belongs or does not belong to the i-th component of the mixture $(i=1,\ldots,g;j=1,\ldots,n)$. The unobservable factors $\boldsymbol{u}_{ij}$ are also introduced as missing data in the EM framework.

Now we postulate that conditional on membership of the i-th component of the mixture the joint distribution of $\boldsymbol{Y}_j$ and its associated factor (vector) $\boldsymbol{U}_{ij}$ is multivariate t-distribution. That is,

$$\begin{pmatrix}\boldsymbol{Y}_j\\ \boldsymbol{U}_{ij}\end{pmatrix}|z_{ij}=1\sim t_{p+q}(\boldsymbol{\mu}_i^*,\boldsymbol{K}_i,\nu_i) \qquad (i=1,\ldots,g), \tag{8}$$

where $\boldsymbol{\mu}_i^*=(\boldsymbol{\mu}_i,\boldsymbol{\xi}_i)=(\boldsymbol{A}^T,\boldsymbol{I}_q)^T\boldsymbol{\xi}_i$ and

$$\boldsymbol{K}_i=\begin{pmatrix}\boldsymbol{A}\boldsymbol{\Omega}_i\boldsymbol{A}^T+\boldsymbol{D} & \boldsymbol{A}\boldsymbol{\Omega}_i\\ \boldsymbol{\Omega}_i\boldsymbol{A}^T & \boldsymbol{\Omega}_i\end{pmatrix}.$$

This specification of the joint distribution of $\boldsymbol{Y}_j$ and its associated factors in (3) will imply the t-mixture model (6) for the marginal distribution of $\boldsymbol{Y}_j$ with the restriction (7). Using the characterization of the t-distribution related to the normal distribution, it follows that we can express (8) alternatively as

$$\begin{pmatrix}\boldsymbol{Y}_j\\ \boldsymbol{U}_{ij}\end{pmatrix}|w_j,z_{ij}=1\sim N_{p+q}(\boldsymbol{\mu}_i^*,\boldsymbol{K}_i/w_j), \tag{9}$$

where w_j is a value of the weight variable W_j taken to have the gamma distribution $f_G(w_j;\nu_i/2,\nu_i/2)$, where

$f_G(w;\alpha,\beta)=\{\beta^\alpha w^{\alpha-1}/\Gamma(\alpha)\}\exp(-\beta w)I_{[0,\infty)}(w)(\alpha,\beta>0)$. Therefore, it can be established from (9) that

$$\boldsymbol{U}_{ij}|w_j,z_{ij}=1\sim N_q(\boldsymbol{\xi}_i,\boldsymbol{\Omega}_i/w_j)$$

and

$$\boldsymbol{e}_{ij}|w_j,z_{ij}=1\sim N_p(\boldsymbol{0},\boldsymbol{D}/w_j)$$

and hence that

$$\boldsymbol{U}_{ij}|z_{ij}=1\sim t_q(\boldsymbol{\xi}_i,\boldsymbol{\Omega}_i,\nu_i)$$

and

$$\boldsymbol{e}_{ij}|z_{ij}=1\sim t_p(\boldsymbol{0},\boldsymbol{D},\nu_i).$$

Thus, with this formulation, the error terms $\boldsymbol{e}_{ij}$ and the factors $\boldsymbol{U}_{ij}$ are distributed according to the t-distribution with the same degrees of freedom. However, the factors and error terms are no longer independently distributed as in the normal-based model for common factor analysis, but they are uncorrelated. To see this, we have from (9) that conditional on w_j, $\boldsymbol{U}_{ij}$ and $\boldsymbol{e}_{ij}$ are uncorrelated, and hence, unconditionally uncorrelated.

By adopting a common factor loading matrix and the t-distribution for the factors and error terms, the MCtF model has fewer parameters and is more robust against extreme observations, thus providing a better fit to data with skewed heavy tails.

We can obtain the maximum likelihood estimator of the vector of unknown parameters in the mixture of common t-factor analyzers model specified by (6) and (7) as follows. We use a modified version of the AECM algorithm outlined in McLachlan *et al.* (2007) for mixtures of t-factor analyzers. We assume that the component-indicators z_{ij}, the factors $\boldsymbol{U}_{ij}$ in (3) and the weights w_j in the characterization (9) of the t-distribution for the i-th component distribution of $\boldsymbol{Y}_j$ and $\boldsymbol{U}_{ij}$ are all missing. We have from (9) that

$$\boldsymbol{Y}_j|\boldsymbol{u}_{ij},w_j,z_{ij}=1\sim N_p(\boldsymbol{A}\boldsymbol{u}_{ij},\boldsymbol{D}/w_j)\quad(i=1,\ldots,g).$$

Thus in the EM framework for this problem, the complete data consist, in addition to the observed data $\boldsymbol{y}_j$, of the component-indicators z_{ij}, the unobservable weights w_j, and the latent factors $\boldsymbol{u}_{ij}$. The complete-data log likelihood for $\boldsymbol{\Psi}$ formed on the basis of the complete data is given by

$$\log L_c(\boldsymbol{\Psi})=\sum_{i=1}^{g}\sum_{j=1}^{n}z_{ij}\log a_{ij},$$

where

$$a_{ij}=\pi_i f_G(w_j;\nu_i/2,\nu_i/2)\phi(\boldsymbol{u}_{ij};\boldsymbol{\xi}_i,\boldsymbol{\Omega}_i/w_j)\phi(\boldsymbol{y}_j;\boldsymbol{A}\boldsymbol{u}_{ij},\boldsymbol{D}/w_j).$$

2.1.1 E-step In order to carry out the E-step, we need to be able to calculate the conditional expectation of the terms $Z_{ij}W_j(\boldsymbol{U}_{ij}-\boldsymbol{\xi}_i)$ and $Z_{ij}W_j(\boldsymbol{U}_{ij}-\boldsymbol{\xi}_i)(\boldsymbol{U}_{ij}-\boldsymbol{\xi}_i)^T$. From (9), we have that conditional on $\boldsymbol{y}_j$ and w_j, the i-th component-conditional distribution of $\boldsymbol{U}_{ij}-\boldsymbol{\xi}_i$ is multivariate normal with mean $\boldsymbol{\gamma}_i^T(\boldsymbol{y}_j-\boldsymbol{A}\boldsymbol{\xi}_i)$ and covariance matrix $(\boldsymbol{I}_q-\boldsymbol{\gamma}_i^T\boldsymbol{A})\boldsymbol{\Omega}_i/w_j$, where $\boldsymbol{\gamma}_i=(\boldsymbol{A}\boldsymbol{\Omega}_i\boldsymbol{A}^T+\boldsymbol{D})^{-1}\boldsymbol{A}\boldsymbol{\Omega}_i$. Thus, $E\{\boldsymbol{U}_{ij}-\boldsymbol{\xi}_i|(\boldsymbol{y}_j,w_j,z_{ij}=1)\}=\boldsymbol{\gamma}_i^T(\boldsymbol{y}_j-\boldsymbol{A}\boldsymbol{\xi}_i)$, and $E\{(\boldsymbol{U}_{ij}-\boldsymbol{\xi}_i)(\boldsymbol{U}_{ij}-\boldsymbol{\xi}_i)^T|(\boldsymbol{y}_j,w_j,z_{ij}=1)\}=\boldsymbol{\gamma}_i^T(\boldsymbol{y}_j-\boldsymbol{A}\boldsymbol{\xi}_i)(\boldsymbol{y}_j-\boldsymbol{A}\boldsymbol{\xi}_i)^T\boldsymbol{\gamma}_i+(\boldsymbol{I}_q-\boldsymbol{\gamma}_i^T\boldsymbol{A})\boldsymbol{\Omega}_i/w_j$. The conditional expectation of W_j given $\boldsymbol{y}_j$ and $z_{ij}=1$ is given by

$$w_i(\boldsymbol{y}_j;\boldsymbol{\Psi})=\frac{\nu_i+p}{\nu_i+\delta(\boldsymbol{y}_j,\boldsymbol{A}\boldsymbol{\xi}_i;\boldsymbol{A}\boldsymbol{\Omega}_i\boldsymbol{A}^T+\boldsymbol{D})},\tag{10}$$

where $\delta(\boldsymbol{y}_j,\boldsymbol{A}\boldsymbol{\xi}_i;\boldsymbol{A}\boldsymbol{\Omega}_i\boldsymbol{A}^T+\boldsymbol{D})=(\boldsymbol{y}_j-\boldsymbol{A}\boldsymbol{\xi}_i)^T(\boldsymbol{A}\boldsymbol{\Omega}_i\boldsymbol{A}^T+\boldsymbol{D})^{-1}(\boldsymbol{y}_j-\boldsymbol{A}\boldsymbol{\xi}_i)$. The conditional expectation of Z_{ij} given $\boldsymbol{y}_j$ is given by the posterior probability $\tau_i(\boldsymbol{y}_j;\boldsymbol{\Psi})$ that $\boldsymbol{y}_j$ belongs to the i-th component of the mixture;

$$\tau_i(\boldsymbol{y}_j;\boldsymbol{\Psi})=\frac{\pi_i f_t(\boldsymbol{y}_j;\boldsymbol{A}\boldsymbol{\xi}_i,\boldsymbol{A}\boldsymbol{\Omega}_i\boldsymbol{A}^T+\boldsymbol{D},\nu_i)}{\sum_{h=1}^{g}\pi_h f_t(\boldsymbol{y}_j;\boldsymbol{A}\boldsymbol{\xi}_h,\boldsymbol{A}\boldsymbol{\Omega}_h\boldsymbol{A}^T+\boldsymbol{D},\nu_h)}\tag{11}$$

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

2.1.2 CM-step We use two CM steps in the AECM algorithm, which correspond to the partition of $\boldsymbol{\Psi}$ into the two subvectors $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$, where $\boldsymbol{\Psi}_1$ consists of the mixing proportions, the elements of $\boldsymbol{\xi}_i$ and the degrees of freedom $\nu_i(i=1,\ldots,g)$. The subvector $\boldsymbol{\Psi}_2$ consists of the elements of the common factor loadings matrix $\boldsymbol{A}$, the $\boldsymbol{\Omega}_i$ and the diagonal matrix $\boldsymbol{D}$.

On the first cycle, we specify the missing data to be the component-indicator variables Z_{ij} and the weights w_j in the characterization (9) of the t-distribution for the component distribution of $\boldsymbol{y}_j$. On the $(k+1)$-th iteration of the algorithm, we update the estimators of the mixing proportions using

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

where the posterior probabilities are calculated using (11). The updated estimate of the i-th component factor mean is given by

$$\boldsymbol{\xi}_i^{(k+1)}=\sum_{j=1}^{n}\tau_i(\boldsymbol{y}_j;\boldsymbol{\Psi}^{(k)})w_{ij}^{(k)}\boldsymbol{A}^{(k)^T}\boldsymbol{y}_j/\sum_{j=1}^{n}\tau_i(\boldsymbol{y}_j;\boldsymbol{\Psi}^{(k)})w_{ij}^{(k)},$$

where the current weight $w_{ij}^{(k)}=w_i(\boldsymbol{y}_j;\boldsymbol{\Psi})$ is formed using the current value $\boldsymbol{\Psi}^{(k)}$ for $\boldsymbol{\Psi}$ in (10).

In the case where the degrees of freedom ν_i in the component t-distributions are not specified but are to be estimated from the data, we have to update the estimate of ν_i on the first cycle. The updated estimate $\nu_i^{(k+1)}$ of ν_i does not exist in closed form, but is given as a solution of the equation

$$-\psi(\frac{\nu_i}{2})+\log(\frac{\nu_i}{2})+1+\frac{1}{n_i^{(k)}}\sum_{j=1}^{n}\tau_{ij}^{(k)}(\log w_{ij}^{(k)}-w_{ij}^{(k)})$$

$$+\psi(\frac{\nu_i^{(k)}+p}{2})-\log(\frac{\nu_i^{(k)}+p}{2})=0,$$

where $\tau_{ij}^{(k)}=\tau_i(\boldsymbol{y}_j;\boldsymbol{\Psi}^{(k)}),n_i^{(k)}=\sum_{j=1}^{n}\tau_{ij}^{(k)}$ $(i=1,\ldots,g)$, and $\psi(\cdot)$ is the digamma function.

The estimate of $\boldsymbol{\Psi}$ is updated so that its current value after the first cycle is given by $\boldsymbol{\Psi}^{(k+1/2)}=(\boldsymbol{\Psi}_1^{(k+1)^T},\boldsymbol{\Psi}_2^{(k)^T})^T$.

On the second cycle of this iteration, the complete data are expanded to include the unobservable factors $\boldsymbol{U}_{ij}$ associated with the $\boldsymbol{y}_j$. 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 $\boldsymbol{\Psi}=\boldsymbol{\Psi}^{(k+1/2)}$. Then the new posterior probability, $\tau_i(\boldsymbol{y}_j;\boldsymbol{\Psi}^{(k+1/2)})$, is estimated by

$$\frac{\pi_i^{(k+1)}f_t(\boldsymbol{y}_j;\boldsymbol{A}^{(k)}\boldsymbol{\xi}_i^{(k+1)},\boldsymbol{A}^{(k)}\boldsymbol{\Omega}_i^{(k)}\boldsymbol{A}^{(k)^T}+\boldsymbol{D}^{(k)},\nu_i^{(k+1)})}{\sum_{h=1}^{g}\pi_h^{(k+1)}f_t(\boldsymbol{y}_j;\boldsymbol{A}^{(k)}\boldsymbol{\xi}_h^{(k+1)},\boldsymbol{A}^{(k)}\boldsymbol{\Omega}_h^{(k)}\boldsymbol{A}^{(k)^T}+\boldsymbol{D}^{(k)},\nu_h^{(k+1)})}.$$

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 $\boldsymbol{A}^{(k+1)}$, $\boldsymbol{\Omega}_i^{(k+1)}$ and $\boldsymbol{D}^{(k+1)}$. Set

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

$$w_{ij}^{(k+1/2)}=\frac{\nu_i^{(k+1)}+p}{\nu_i^{(k+1)}+\delta(\boldsymbol{y}_j,\boldsymbol{A}^{(k)}\boldsymbol{\xi}_i^{(k+1)};\boldsymbol{A}^{(k)}\boldsymbol{\Omega}_i^{(k)}\boldsymbol{A}^{(k)^T}+\boldsymbol{D}^{(k)})},$$

$$n_i^{(k+1/2)}=\sum_{j=1}^{n}\tau_i(\boldsymbol{y}_j;\boldsymbol{\Psi}^{(k+1/2)}),$$

$$\boldsymbol{S}_i^{(k+1/2)}=\frac{\sum_{j=1}^{n}\tau_i(\boldsymbol{y}_j;\boldsymbol{\Psi}^{(k+1/2)})w_{ij}^{(k+1/2)}SE_i^{(k+1/2)}}{\sum_{j=1}^{n}\tau_i(\boldsymbol{y}_j;\boldsymbol{\Psi}^{(k+1/2)})},$$

where $\mathrm{SE}_i^{(k+1/2)}=(\boldsymbol{y}_j-\boldsymbol{A}^{(k)}\boldsymbol{\xi}_i^{(k+1)})(\boldsymbol{y}_j-\boldsymbol{A}^{(k)}\boldsymbol{\xi}_i^{(k+1)})^T$. Then $\boldsymbol{\Omega}_i^{(k+1)}$ is given by

$$\boldsymbol{\Omega}_i^{(k+1)}=\boldsymbol{\gamma}_i^{(k+1/2)^T}\boldsymbol{S}_i^{(k+1/2)}\boldsymbol{\gamma}_i^{(k+1/2)}+\boldsymbol{\Omega}_i^{(k)}(\boldsymbol{I}_q-\boldsymbol{A}^{(k)^T}\boldsymbol{\gamma}_i^{(k+1/2)}),$$

and the updated estimate $\boldsymbol{D}^{(k+1)}$ is given by

$$\boldsymbol{D}^{(k+1)} = \frac{1}{\sum_{i=1}^{g} n_i^{(k+1/2)}} \sum_{i=1}^{g} n_i^{(k+1/2)} \{(\boldsymbol{A}^{(k)} \boldsymbol{\gamma}_i^{(k+1/2)^T} - \boldsymbol{I}_p) \boldsymbol{S}_i^{(k+1/2)}$$

$$\cdot (\boldsymbol{A}^{(k)} \boldsymbol{\gamma}_i^{(k+1/2)^T} - \boldsymbol{I}_p)^T + \boldsymbol{A}^{(k)} \boldsymbol{\Omega}_i^{(k)} (\boldsymbol{I}_q - \boldsymbol{A}^{(k)^T} \boldsymbol{\gamma}_i^{(k+1/2)}) \boldsymbol{A}^{(k)^T} \}.$$

Let $\boldsymbol{\eta}_{ij}^{(k+1/2)} = \boldsymbol{\gamma}_i^{(k+1/2)^T} (\boldsymbol{y}_j - \boldsymbol{A}^{(k)} \boldsymbol{\xi}_i^{(k+1)}) + \boldsymbol{\xi}_i^{(k+1)}$. Then the updated estimate $\boldsymbol{A}^{(k+1)}$ is obtained by

$$\boldsymbol{A}^{(k+1)} = \left(\sum_{i=1}^{g} \sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j; \boldsymbol{\Psi}^{(k+1/2)}) w_{ij}^{(k+1/2)} \boldsymbol{y}_j \boldsymbol{\eta}_{ij}^{(k+1/2)^T} \right)$$

$$\cdot \{ \sum_{i=1}^{g} (\sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j; \boldsymbol{\Psi}^{(k+1/2)}) w_{ij}^{(k+1/2)} \boldsymbol{\eta}_{ij}^{(k+1/2)} \boldsymbol{\eta}_{ij}^{(k+1/2)^T}$$

$$+ n_i^{(k+1/2)} (\boldsymbol{I}_q - \boldsymbol{\gamma}_i^{(k+1/2)^T} \boldsymbol{A}^{(k)}) \boldsymbol{\Omega}_i^{(k)}) \}^{-1}.$$

We have to specify an initial value for the vector $\boldsymbol{\Psi}$ of unknown parameters in the application of the EM algorithm. A random start is obtained by first randomly assigning the data into g groups. Using the sample mean and sample covariance matrix of each randomly partitioned data, the initial parameter estimates are obtained as described in the Appendix of Baek *et al.* (2010).

We can portray the observed data $\boldsymbol{y}_j$ in q-dimensional space by plotting the corresponding value of the $\hat{\boldsymbol{u}}_{ij}$, which are the estimated conditional expectations of the factors $\boldsymbol{U}_{ij}$, corresponding to the observed data points $\boldsymbol{y}_j$. Note that

$$E(\boldsymbol{U}_{ij} | \boldsymbol{y}_j, z_{ij} = 1) = E\{E(\boldsymbol{U}_{ij} | \boldsymbol{y}_j, w_j, z_{ij} = 1)\}$$

$$= \boldsymbol{\xi}_i + \boldsymbol{\gamma}_i^T (\boldsymbol{y}_j - \boldsymbol{A}\boldsymbol{\xi}_i) \qquad (12)$$

We let $\hat{\boldsymbol{u}}_{ij}$ denote the value of the right-hand side of (12) evaluated at the maximum likelihood estimates of $\boldsymbol{\xi}_i$, $\boldsymbol{\gamma}_i$ and $\boldsymbol{A}$. We can define the estimated value $\hat{\boldsymbol{u}}_j$ of the j-th factor corresponding to $\boldsymbol{y}_j$ as

$$\hat{\boldsymbol{u}}_j = \sum_{i=1}^{g} \tau_i(\boldsymbol{y}_j; \hat{\boldsymbol{\Psi}}) \hat{\boldsymbol{u}}_{ij} \qquad (j = 1, \ldots, n). \qquad (13)$$

An alternative estimate of the posterior expectation of the factor corresponding to the j-th observation $\boldsymbol{y}_j$ is defined by replacing $\tau_i(\boldsymbol{y}_j; \hat{\boldsymbol{\Psi}})$ by $\hat{z}_{ij}$ in (13).

3 RESULTS

Souto *et al.* (2008) compared different clustering methods for the analysis of 35 cancer gene expression datasets. For our experiment, we considered 2 of these 35 datasets. We applied MC*t*FA to cluster each of these two datasets, and compared its performance with other methods. We compare our method with MCLUST, MFA and MCFA. The performance is measured by the Adjusted Rand Index (ARI; Hubert and Arabie, 1985) and the error rate since the true membership of each observation is known.

MCLUST is a software package that implements Gaussian mixture models via EM algorithm and the Bayesian Information Criterion (BIC, Schwarz, 1978) for model-based clustering (Fraley and Raftery, 2003). In MCLUST, the component-covariance matrix $\boldsymbol{\Sigma}_i$ is parameterized by eigenvalue decomposition in the form $\boldsymbol{\Sigma}_i = \rho_i \boldsymbol{E}_i \boldsymbol{\Lambda}_i \boldsymbol{E}_i^T$, where ρ_i is a constant, $\boldsymbol{E}_i$ is the matrix of eigenvectors, $\boldsymbol{\Lambda}_i$ is a diagonal matrix with elements proportional to the eigenvalues of $\boldsymbol{\Sigma}_i$. Different conditions on $\rho_i, \boldsymbol{\Lambda}_i$ and $\boldsymbol{E}_i$ characterize the volume, shape and orientation of each component distribution in MCLUST. We deal with the 10 submodels of MCLUST: EII, VII, EEI, VEI, EVI, VVI, EEE, EEV, VEV, VVV (Fraley and Raftery, 2003, Table 1). For MFA, we assumed the covariance matrix of errors is equal for each component. We took advantage of the mclust software for R (Team RDC 2004) and the EMMIX program (McLachlan *et al.,* 1999) for MFA, and developed programs for the MCFA and the MC*t*FA approaches, using the MATLAB language.

Table 1. Comparison of MCLUST, MFA, MCFA and MC*t*FA models for implied clustering versus the true membership of Chowdary's 104 cancer tissues

Model	Factors	BIC	AWE	ARI	Error rate
MCLUST	VVI	**242 872**		**0.0657**	**0.3462**
MFA	1	250 841	257 805	0.0296	0.3750
	2	241 124	250 855	0.0296	0.3750
	3	234 888	**247 371**	**0.5858**	**0.1154**
	4	232 917	248 137	0.0386	0.3750
	5	230 485	248 426	0.0657	0.3462
	6	230 326	250 974	0.4726	0.1538
	7	230 461	253 800	0.2790	0.2308
	8	229 825	255 839	0.1431	0.2981
	9	**229 433**	258 107	0.1696	0.2885
MCFA	1	261 337	264 164	**0.6800**	**0.0865**
	2	253 301	257 532	0.0464	0.3654
	3	246 291	251 934	0.1282	0.3077
	4	243 084	250 139	0.0657	0.3462
	5	240 297	248 767	0.1587	0.2885
	6	236 654	246 539	0.0657	0.3462
	7	233 076	**244 374**	0.0657	0.3462
	8	232 083	244 795	0.2128	0.2596
	9	**231 226**	245 354	0.1240	0.3077
MC*t*FA	1	234 450	237 278	0	0.4038
	2	229 677	233 920	0.8505	0.0385
	3	228 106	233 764	0.8505	0.0385
	4	226 891	233 976	0.0675	0.3558
	5	225 387	233 876	0.5564	0.1250
	6	223 518	**233 419**	**0.8867**	**0.0288**
	7	222 455	233 772	0.6808	0.0865
	8	222 156	234 884	0.7465	0.0673
	9	**221 134**	235 276	0.4742	0.1538

The bold numbers are the optimal values of BIC, AWE, ARI and Error rate for each model.

The first set concerns both breast and colon cancer data (Chowdary *et al.*, 2006), which consists of 104 gene expressions for 52 matched tissue pairs of two different cancer types (32 pairs of breast tumour and 20 pairs of colon tumour). There are 22 283 genes in the original data, but Souto *et al.* (2008) selected 182 genes by filtering uninformative genes. It has been reported in many analyses of real datasets that the empirical distribution of gene expression levels is approximately log-normal or sometimes (on the log scale) with a slightly heavier tailed t-distribution depending on the biological samples under investigation (Li, 2002). Thus, these data may also have many extreme expressions for each gene. Figure 1 shows boxplots of the expression levels for the first 10 genes. The distribution of each gene is skewed and has a very long tail. The rest of the genes also have similar shaped distributions. In particular, gene 6 has very high expression levels for six particular tissues.

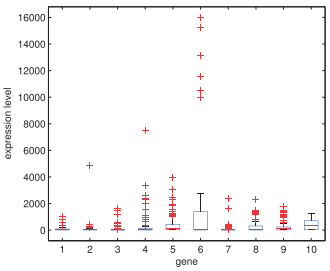

Fig. 1. The boxplots for the first 10 genes in the cancer data of Chowdary *et al.* (2006).

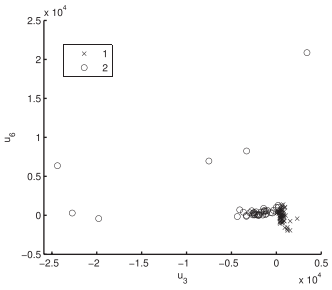

Fig. 2. Plot of the (estimated) posterior mean factor scores via the MC*t*FA approach based on the implied clustering for the cancer data of Chowdary *et al.* (2006).

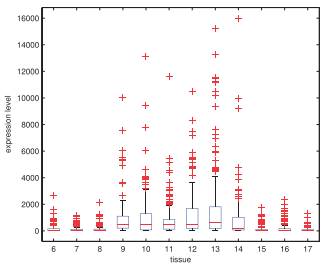

Fig. 3. The boxplots of expression levels of all genes for the 6th–17th colon tumor tissues.

We implemented the MCLUST, MFA, MCFA and MC*t*FA procedures with $g=2$ components with the number of factors q ranging from 1 to 9, using 50 starting values for the parameters. For each value of q, we computed the ARI and the associated error rate. The results are presented in Table 1. We have also listed in this table the values of the BIC and the Approximate Weight of Evidence (AWE: Banfield and Raftery, 1993) for each model with different q. AWE is a model selection criterion based on an approximation to the classification log-likelihood. AWE is defined as

$$-2\log L(\boldsymbol{\Psi})+2EN(\boldsymbol{\tau})+2m(3/2+\log(n)), \qquad (14)$$

where $EN(\boldsymbol{\tau})=-\sum_{j=1}^{n}\sum_{i=1}^{g}\tau_i(\boldsymbol{y}_j;\boldsymbol{\Psi})\log(\tau_i(\boldsymbol{y}_j;\boldsymbol{\Psi}))$ is the entropy of the classification matrix with (i,j)-th element equal to $\tau_i(\boldsymbol{y}_j;\boldsymbol{\Psi})$ and m is the number of (free) parameters. It penalizes complex models more severely than BIC, and thus selects more parsimonious models than BIC. It would appear that BIC works well at a practical level; see, for example, Fraley and Raftery (1998). Further, Keribin (2000) proved that BIC provides a consistent estimator of g (Celeux, 2007). But BIC can lead to too few or too many clusters in practice, depending on the problem at hand. For the present problem of choosing the number of factors q, it would appear from Table 1 that it leads to too many factors being fitted in the mixtures of factor analyzers model. An apparent explanation for this is that for the present dataset (and the other one to follow), the data are not well represented by the true models of clusters and/or the true clusters are not well separated. As a consequence, BIC leads to too many factors in the mixture model being fitted to the data. The AWE criterion is preferable to BIC here as it leads to fewer factors since it places a higher penalty on more complex model due to the presence of twice of the entropy and the extra constant term $(2EN(\boldsymbol{\tau})+3m+m\log(n))$ in (14). We also considered the ICL criterion which chooses q to minimize $-2\log L(\boldsymbol{\Psi})+2EN(\boldsymbol{\tau})+m\log(n)$, which is similar to the AWE criterion. They are the same apart from the additional penalty of $3m+m\log(n)$ imposed by AWE, which for our present problem leads to a better choice of q.

It can be seen that the lowest error rate (0.0288) and highest value (0.8867) of the ARI is obtained by using $q=6$ factors with the MC*t*FA model, which coincides with the choice on the basis of AWE. The lowest error rate (0.1154) of the MFA model is obtained for $q=3$ factors. The best result of MCFA model is obtained with its lowest error rate 0.0865 for $q=1$ factor (AWE suggests using $q=7$). MCLUST chose VVI model as its best and its error rate is 0.3462. It can be seen that the error rate and ARI for MC*t*FA are better than those for MCLUST, MFA and MCFA. We have also calculated BIC for all models. It can be seen that it failed to select the best model for each method. BIC reached its minimum for largest $q(q=9)$ of each method, so it selected a more complex model than the one with highest ARI and lowest error rate. In the case where the distribution from which the data arose is not in the collection of considered mixture models, BIC criterion tends to overestimate the correct size regardless of the separation of the clusters (Celeux, 2007).

To illustrate the usefulness of the MC*t*FA approach for portraying the results of a clustering in low-dimensional space, we have plotted in Figure 2 the estimated factor scores $\hat{\boldsymbol{u}}_j$ as defined by (13) with the implied cluster labels shown. In this plot, we used the third and sixth factors in the fit of the MC*t*FA model with $q=6$ factors. It can be seen that the clusters are represented in this plot with very little overlap. The estimated factor scores were plotted according to the implied clustering labels. The degrees of freedom of the factor t-distributions for both groups were estimated as 1.0 and 1.0, which means their tails of the distributions are very thick and long. We can easily detect 6 distinct extreme tissues as shown in Figure 2, which are known to be the 9th–14th colon tumor tissue. Figure 3 shows the expression levels of all genes for the 6th–17th colon tumor tissues. In Figure 3, we observe that all of these 6 outliers are very different from others since they have extremely large expression levels not only of the 6th gene shown in Figure 1, but also of other genes.

The second dataset to which we applied our method is a lung cancer data (Bhattacherjee *et al.*, 2001), for which the number of classes is not small ($g=5$). It consists of 203 gene expressions partitioned into five subpopulations: four lung cancer types and normal tissues. Souto *et al.* (2008) selected 1543 informative genes from the original 12 600 genes. There are big differences among the class sizes of the data. The number of tissues for each class is 139, 17, 6, 21 and 20, respectively. Figure 4 shows the

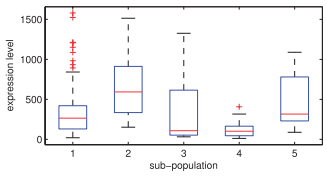

Fig. 4. The boxplots for the expressions of a gene by subpopulation: the lung cancer data of Bhattacherjee *et al.* (2001).

boxplots of the expressions of a gene plotted for each subpopulation. It can be seen that there exist skewed distributions mixed with symmetric distribution with or without extreme observations for five components in the plot. Since the selected (1543) genes are still too many for the mixture model, we grouped the genes into 50 clusters and selected the centroid from each cluster of genes. That is, we applied the k-means algorithm to the 1543 gene expressions and clustered them into 50 groups of similar characteristics. Then we extracted the centroid from each group to make 50 new features for the mixture models.

We implemented the MCLUST, MFA, MCFA, and MCtFA approaches with $g=5$ components for the number of factors q ranging from 1 to 10, using 50 starting values for the parameters. For each value of q, we computed the ARI and the error rate. The results are presented in Table 2. We have also listed in this table the values of BIC and AWE for each model with different levels of q. MCtFA attains its largest ARI (0.7322) and lowest error rate (0.1133) for $q=6$, although AWE suggested the model with $q=7$. We notice that there is little difference between the AWE values for $q=6$ and for $q=7$. The lowest error rate for MCFA is 0.2611 for $q=2$ and the largest ARI is 0.4570 for $q=9$. The error rate (NA) of MCFA for $q=1$ was not able to be calculated since the estimated number of clusters was less than the true value 5. Neither BIC nor AWE indicated the best model for MCFA. MFA reached its best ARI (0.3487) and error rate (0.3498) for $q=7$. The minimum BIC and AWE were obtained at $q=6$, and at $q=1$, respectively. The best model for MCLUST showed similar performance (ARI: 0.3021, error rate: 0.3350) to MFA. MCtFA again performed better than the other methods for this dataset.

We display the data using the estimated factor scores of our model in 3D space (Figure 5). In the latter, we used the second, the fourth and the fifth factors in the fit of the MCtFA model with $q=6$ factors. The estimated factor scores were plotted according to the implied clustering labels. It can be seen that the five clusters are represented in this plot with very little overlap. The degrees of freedom of the factor t-distributions for the components were estimated as 1.1, 1.3, 7.8, 4.0 and 4.1. There are two distributions with long tails [$\nu_1 = 1.1$(*triangle*), $\nu_2 = 1.3$(*circle*)] in the plot. We have also given in Figure 6 the plot corresponding to that in Figure 5 with the true cluster labels shown. There are 23 misallocated tissues which can be seen in other's clusters, but as a whole there is a good agreement between the two plots.

Table 2. Comparison of MCLUST, MFA, MCFA and MCtFA models for implied clustering versus the true membership of Bhattacharjee's 203 lung cancer tissues

Model	Factors	BIC	AWE	ARI	Error rate
MCLUST	VVI	**135 023**		**0.3021**	**0.3350**
MFA	1	140 710	**145 324**	0.3100	0.4286
	2	139 479	146 125	0.3219	0.3941
	3	139 313	147 955	0.3156	0.4089
	4	139 016	149 611	0.3013	0.4039
	5	139 068	151 573	0.3368	0.3645
	6	**138 870**	153 244	0.2445	0.4236
	7	139 358	155 561	**0.3487**	**0.3498**
	8	139 747	157 738	0.2616	0.4335
	9	140 207	159 943	0.2571	0.4433
	10	140 414	161 854	0.2368	0.4680
MCFA	1	148 255	149 674	0.0721	NA
	2	144 994	146 585	0.3348	**0.2611**
	3	142 978	145 042	0.3090	0.3300
	4	141 826	144 453	0.4376	0.2759
	5	140 943	144 139	0.3703	0.3153
	6	140 123	143 943	0.3269	0.3251
	7	139 362	**143 800**	0.3692	0.3153
	8	138 921	144 015	0.3775	0.3153
	9	138 420	144 194	**0.4570**	0.2709
	10	**138 134**	144 633	0.2418	0.4335
MCtFA	1	144 424	145 607	−0.1269	0.4384
	2	142 708	144 294	0.3619	0.2413
	3	141 155	143 236	0.4735	0.2266
	4	139 683	142 334	0.6179	0.1527
	5	138 937	142 154	0.6657	0.1379
	6	138 194	142 025	**0.7322**	**0.1133**
	7	137 538	**142 009**	0.5875	0.1675
	8	136 973	142 105	0.6417	0.1773
	9	136 660	142 474	0.4408	0.2365
	10	**136 431**	142 967	0.3215	0.3153

The bold numbers are the optimal values of BIC, AWE, ARI and Error rate for each model. NA means Not Available.

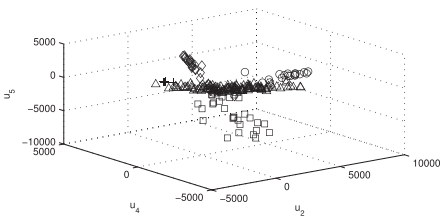

Fig. 5. Plot of the (estimated) posterior mean factor scores via the MCtFA approach based on the implied clustering for the lung cancer data of Bhattacherjee *et al.* (2001).

4 DISCUSSION

For clustering high-dimensional data such as microarray gene expressions, MFA is a useful technique since it can reduce the number of parameters through its factor-analytic representation of the component-covariance matrices. However, this approach is sensitive to outliers as it is based on a mixture model in which the multivariate normal family of distributions is assumed for the component factor and error distributions. McLachlan *et al.* (2007) extended MFA to incorporate t-distributions for the component factor and error in the mixture model for dealing with unusual extreme observations (MtFA). These methods, however, may

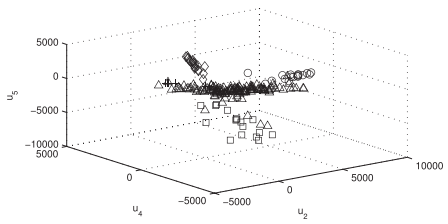

Fig. 6. Plot of the (estimated) posterior mean factor scores via the MC*t*FA approach with the true labels shown for the lung cancer data of Bhattacherjee *et al*. (2001).

not provide a sufficient reduction in the number of parameters, particularly when the number of clusters (subpopulations) is not small. In this article, we proposed a new mixture model which can reduce the number of parameters further in such instances and cluster the data containing outliers simultaneously by introducing a mixture of *t*-distributions with both component-mean and component covariance represented by common factor loadings. We call this approach mixtures of common *t*-factor analyzers (MC*t*FA). We describe the implementation of an EM algorithm for fitting the MC*t*FA. This approach also has the ability to portray the results of a clustering in low-dimensional space. We can plot the estimated posterior means of the factors $\hat{u}_j$ as defined by (13) with the implied cluster labels. On the other hand, the approaches MCLUST, MFA and M*t*FA cannot project high-dimensional objects in low-dimensional space. The applications of MC*t*FA to two cancer microarray datasets have demonstrated the usefulness and its relative superiority in clustering performance over MCLUST, MFA and MCFA. It has shown that our method works well for clustering data containing outliers. Moreover, it provides information on the distribution structure of each subpopulation by displaying the estimated factor scores in low-dimensional space. We observed also that the proposed approach fitted the experimental datasets better than the other approaches, and the performance difference between MC*t*FA and the others becomes even greater when the number of clusters is not small, such as in the case of second dataset (Section 1.1 of Supplementary Material).

Often BIC is used to provide a guide to the choice of the number of factors q and the number of components g to be used. However, it did not always lead to the correct choice of the best model. That is, BIC can lead to too simple or too complex model in practice, depending on the problem at hand. Simulation studies reported in Biernacki and Govaert (1997), Biernacki *et al.* (2000) and McLachlan and Peel (2000a) show that BIC will overrate the number of clusters under misspecification of the component density, whereas several alternative criteria such as the AWE and ICL criterion are able to identify the correct number of clusters even when the component densities are misspecified (Frühwirth-Schnatter and Pyne, 2010). In both of our real data applications, we observed that BIC did not choose the best q. An apparent explanation for this is that BIC tries to choose more complex model since some of the subpopulations of the datasets have skewed distributions and have several extreme outliers. On the other hand, AWE leads to the best or almost the best model with smallest error rate since it is more robust against misspecification of the component densities for the experimental datasets. Recently, Frühwirth-Schnatter and Pyne (2010) reported that AWE picked the correct model for both skew-*t* and skew-normal mixture distributions. Also a small simulation study confirms the better performance of the AWE over the BIC when the distribution of the data has skewed heavy tails due to some extreme observations (Section 1.2 of Supplementary Material). In future work, we wish to investigate the use of various model selection criteria on choosing the number of factors q and the number of components g in mixtures of t or skewed distributions.

Funding: Korea Research Foundation Grant funded by the Korean Government (MOEHRD, Basic Research Promotion Fund, KRF-2007-521-C00048 to J.B.). Australian Research Council (to G.J.M.).

Conflict of Interest: none declared.

REFERENCES

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

Baek,J. *et al.* (2010) Mixtures of factor analyzers with common factor loadings: applications to the clustering and visualisation of high-dimensional data. *IEEE Trans. Pattern Anal. Mach. Intel.*, **32**, 1298–1309.

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

Bhattacherjee,A. *et al.* (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl Acad. Sci. USA*, **98**, 13790–13795.

Biernacki,C. and Govaert,G. (1997) Using the classification likelihood to choose the number of clusters. *Comput. Sci. Stat.*, **29**, 451–457.

Biernacki,C. *et al.* (2000) Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Mach. Intel.*, **22**, 719–725.

Celeux,G. (2007) Mixture models for classification. In Decker,R. *et al.* (ed.) *Advances in Data Analysis*. Springer, Berlin.

Chowdary,D. *et al.* (2006) Prognostic gene expression signatures can be measured in tissues collected in RNAlater preservative. *J. Mol. Diagn.*, **8**, 31–39.

Fraley,C. and Raftery,A.E. (1998) How many clusters? Which clustering methods? Answers via model-based cluster analysis. *Comput.J.*, **41**, 578–588.

Fraley,C. and Raftery,A.E. (2003) Enhanced model-based clustering, density estimation, and discriminant analysis software: MCLUST. *J. Classific.*, **20**, 263–286.

Frühwirth-Schnatter,S. and Pyne,S. (2010) Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-*t* distributions. *Biostatistics*, **11**, 317–336.

Galimberti,G. *et al.* (2009) Penalized factor mixture analysis for variable selection in Clustered Data, *Comput. Stat. Data Anal.*, **53**, 4301–4310.

Giles,P.J. and Kipling,D. (2003) Normality of oligonucleotide microarray data and implications for parametric statistical analyses. *Bioinformatics*, **19**, 2254–2262.

Hinton,G.E. *et al.* (1997) Modeling the manifolds of images of handwritten digits. *IEEE Trans. Neural Netw.*, **8**, 65–73.

Hubert,L. and Arabie,P. (1985) Comparing partitions. *J. Classific.*, **2**, 193–218.

Keribin,C. (2000) Consistent estimation of the order of mixture models. *Sankhya Ser. A*, **62**, 49–66.

Li,K.C. (2002) Genome-wide coexpression dynamics: theory and application. *Proc. Natl Acad. Sci. USA*, **99**, 16875–16880.

Lönnstedt,I. and Speed,T. (2002) Replicated microarray data. *Stat. Sinica*, **12**, 31–46.

Martella,F. (2006) Classification of microarray data with factor mixture models. *Bioinformatics*, **22**, 202–208.

McLachlan,G.J. and Basford,K.E. (1988) *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker Inc, New York.

McLachlan,G.J. and Peel,D. (2000a) *Finite Mixture Models*. Wiley, New York.

McLachlan,G.J. and 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. *et al.* (1999) The EMMIX software for the fitting of mixtures of normal and t-components. *J. Stat. Softw.*, **4**, 2.

McLachlan,G.J. *et al.* (2002) Mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, **18**, 413–422.

McLachlan,G.J. *et al.*(2003) Modelling high-dimensional data by mixtures of factor analyzers. *Comput. Stat. Data Anal.*, **41**, 379–388.

McLachlan,G.J. *et al.* (2007) Extension of the mixture of factor analyzers model to incorporate the multivariate *t* distribution. *Comput. Stat. Data Anal.*, **51**, 5327–5338.

McNicholas,P.D. and Murphy,T.B. (2008) Parsimonious Gaussian mixture models. *Stat. Comput.*, **18**, 285–296.

Schwarz,G. (1978) Estimating the dimension of a model. *Ann. Stat.*, **6**, 461–464.

Souto,M. *et al.* (2008) Clustering cancer gene expression data: a comparative study. *BMC Bioinformatics*, **9**, 497.

Team RDC (2004) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Thalamuthu,A. *et al.* (2006) Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*, **22**, 2405–2412.

Xie,B. *et al.* (2010) Penalized mixtures of factor analyzers with application to clustering high dimensional microarray data. *Bioinformatics*, **26**, 501–508.

Yeung,K.Y. *et al.* (2001) Model-based clustering and data transformations for gene expression data. *Bioinformatics*, **17**, 977–987.

Zhou,H. *et al.* (2009) Penalized model-based clustering with unconstrained covariance matrices. *Electron. J. Stat.*, **3**, 1473–1496.