Issues of Robustness and High Dimensionality in Cluster Analysis

Kaye Basford[1], Geoff McLachlan[2], and Richard Bean[3]

[1] School of Land and Food Sciences
University of Queensland
Australia
k.e.basford@uq.edu.au
[2] Department of Mathematics & Institute for Molecular Bioscience
University of Queensland
Australia
gjm@maths.uq.edu.au
[3] Institute for Molecular Bioscience
University of Queensland
Australia
rbean@maths.uq.edu.au

Summary. Finite mixture models are being increasingly used to model the distributions of a wide variety of random phenomena. While normal mixture models are often used to cluster data sets of continuous multivariate data, a more robust clustering can be obtained by considering the t mixture model-based approach. Mixtures of factor analyzers enable model-based density estimation to be undertaken for high-dimensional data where the number of observations n is very large relative to their dimension p. As the approach using the multivariate normal family of distributions is sensitive to outliers, it is more robust to adopt the multivariate t family for the component error and factor distributions. The computational aspects associated with robustness and high dimensionality in these approaches to cluster analysis are discussed and illustrated;

Key words: finite mixture models, normal components, mixtures of factor analyzers, t distributions, EM algorithm

1 Introduction

Finite mixture models are being increasingly used to model the distributions of a wide variety of random phenomena. As in [MNB06], consider their application in the context of cluster analysis. Let the p-dimensional vector $\boldsymbol{x} = (x_1, \ldots, x_p)^T$ contain the values of p variables measured on each of n (independent) entities to be clustered, and let $\boldsymbol{x}_j$ denote the value of $\boldsymbol{x}$ corresponding to the jth entity ($j = 1, \ldots, n$). With the mixture approach to clustering, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are assumed to be an observed random sample from mixture of a finite number, say g, of groups in some unknown proportions $\pi_1, \ldots, \pi_g$. The mixture density of $\boldsymbol{x}_j$ is expressed as

$$f(\boldsymbol{x}_j; \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i f_i(\boldsymbol{x}_j; \boldsymbol{\theta}_i) \qquad (j = 1, \ldots, n), \tag{1}$$

where the mixing proportions $\pi_1, \ldots, \pi_g$ sum to one and the group-conditional density $f_i(\boldsymbol{x}_j; \boldsymbol{\theta}_i)$ is specified up to a vector $\boldsymbol{\theta}_i$ of unknown parameters $(i = 1, \ldots, g)$. The vector of all unknown parameters is given by $\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_{g-1}, \boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_g^T)^T$, where the superscript T denotes vector transpose. Using an estimate of $\boldsymbol{\Psi}$, this approach gives a probabilistic clustering of the data into g clusters in terms of estimates of the posterior probabilities of component membership,

$$\tau_i(\boldsymbol{x}_j) = \frac{\pi_i f_i(\boldsymbol{x}_j; \boldsymbol{\theta}_i)}{f(\boldsymbol{x}_j; \boldsymbol{\Psi})}, \tag{2}$$

where $\tau_i(\boldsymbol{x}_j)$ is the posterior probability that $\boldsymbol{x}_j$ (really the entity with observation $\boldsymbol{x}_j$) belongs to the ith component of the mixture$(i = 1, \ldots, g;\ j = 1, \ldots, n)$.

The parameter vector $\boldsymbol{\Psi}$ can be estimated by maximum likelihood. The maximum likelihood estimate (MLE) of $\boldsymbol{\Psi}, \hat{\boldsymbol{\Psi}}$, is given by an appropriate root of the likelihood equation,

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

where

$$\log L(\boldsymbol{\Psi}) = \sum_{j=1}^{n} \log f_i(\boldsymbol{x}_j; \boldsymbol{\theta}_i) \tag{4}$$

is the log likelihood function for $\boldsymbol{\Psi}$. Solutions of (3) corresponding to local maximizers of $\log L(\boldsymbol{\Psi})$ can be obtained via the expectation-maximization (EM) algorithm of [DLR77].

For the modelling of continuous data, the group-conditional densities are usually taken to belong to the same parametric family, for example, the normal. In this case,

$$f_i(\boldsymbol{x}_j; \boldsymbol{\theta}_i) = \phi(\boldsymbol{x}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{5}$$

where $\phi(\boldsymbol{x}_j; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the p-dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

One attractive feature of adopting mixture models with elliptically symmetric components such as the normal or t densities, is that the implied clustering is invariant under affine transformations of the data (that is, under operations relating to changes in location, scale, and rotation of the data); see, for example, [CDH99]. Thus the clustering process does not depend on irrelevant factors such as the units of measurement or the orientation of the clusters in space. Concerning the desirability of the latter, [Har75] has commented that affine invariance is less compelling that than invariance under the change of measuring units of each variable.

Unfortunately, as with many other applications of ML estimation for normal-based models, the ML fitting of normal mixture models is not robust against gross outliers, at least if the number of components g is not fixed. The problem of providing protection against outliers in multivariate data is a very difficult problem and increases in difficulty with the dimension of the data. There is now a vast literature on robust modelling methods some of which focus on outlier identification, while others are more for outlier accommodation ([Hub81]). In a series of papers, [ROC96], [RW96], [RW97], [WR93], and [WR94] have considered robust estimation of multivariate location and shape, and the consequent identification of outliers and leverage points. More recently, [DG05] have discussed the concept of breakdown points ([Ham71]; [DH83]). In the context of mixtures, [Hen04] has given an account of robustness issues with ML estimation of univariate normal mixture models.

One-way to broaden the normal mixture parametric family for potential outliers or data with longer-than-normal tails is to adopt mixtures of t distributions, as proposed by [MP98] and [MP00b], and [PM00]. Mixtures of a fixed number of t components, are not robust against outliers. The advantage of the t mixture model is that, although the number of outliers needed for breakdown is almost the same as with the normal mixture model, the outliers have to be much larger. This point is made more precise in [Hen04].

Robust estimation in the context of mixture models has been considered in the past by [Cam84] and [MB88], among others, using M-estimates of the means and covariance matrices of the normal components of the mixture model. [Mar00] has provided a formal approach to robust mixture estimation by applying weighted likelihood methodology ([MBL98] in the context of mixture models. [MN04] and [NFD04] have considered the trimmed likelihood methodology ([HL97]; [VN98]) in the fitting of mixtures of normals and generalized linear models. Also, [TK99] have proposed the technique of otstrap "bumping," which can be used for resistant fitting.

We give a brief review of the fitting of mixtures of t components and the use of mixture models for the clustering of high-dimensional data. With mixtures of normal or t component distributions, there may be problems with potential singularities in the estimates of the component scale matrices. One way to avoiding such singularities for mixture of normal components is to fit mixtures of factor analyzers. Thus we will then discuss how this latter model can be made less sensitive to outliers by considering the implementation of mixtures of t factor analyzers whereby the multivariate t family is adopted for the component error and factor distributions.

2 Multivariate t Distribution

For mixtures of normal components, the ith component-conditional distribution of the jth observation vector $\boldsymbol{X}_j$ is given by

$$\boldsymbol{X}_j \sim N(\boldsymbol{\mu}_i, \, \boldsymbol{\Sigma}_i),$$

denoting the multivariate normal distribution with mean vector μ_i and covariance matrix $\boldsymbol{\Sigma}_i$.

With the t mixture model, the normal distribution for the ith component is embedded in a wider class of elliptically symmetric distributions with an additional parameter ν_i called the degrees of freedom. Then the ith-conditional distribution of $\boldsymbol{X}_j$ is given by

$$\boldsymbol{X}_j \sim t(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \nu_i), \tag{6}$$

where $t(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \nu_i)$ denotes the multivariate t distribution with mean $\boldsymbol{\mu}_i$, scale matrix $\boldsymbol{\Sigma}_i$, and ν_i degrees of freedom. The mean of this t distribution is $\boldsymbol{\mu}_i$ and its covariance matrix is $\{\nu_i/(\nu_i - 2)\}\boldsymbol{\Sigma}_i$.

The density corresponding to (6) is given by

$$f(\boldsymbol{x}_j;\, \boldsymbol{\mu}_i,\, \boldsymbol{\Sigma}_i,\, \nu_i) = \frac{\Gamma(\frac{\nu_i+p}{2})\, |\boldsymbol{\Sigma}_i|^{-1/2}}{(\pi\nu_i)^{\frac{1}{2}p}\Gamma(\frac{\nu_i}{2})\{1+\delta(\boldsymbol{x}_j,\, \boldsymbol{\mu}_i;\, \boldsymbol{\Sigma}_i)/\nu_i\}^{\frac{1}{2}(\nu_i+p)}}, \tag{7}$$

where

$$\delta(\boldsymbol{x}_j,\, \boldsymbol{\mu}_i;\, \boldsymbol{\Sigma}_i) = (\boldsymbol{x}_j - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu}_i) \tag{8}$$

denotes the squared Mahalanobis distance between $\boldsymbol{x}_j$ and $\boldsymbol{\mu}_i$ (with $\boldsymbol{\Sigma}_i$ as the covariance matrix).

The t distribution (6) can be characterized by letting W_j denote a random variable distributed as

$$W_j \sim \text{ gamma}\,(\tfrac{1}{2}\nu_i,\, \tfrac{1}{2}\nu_i), \tag{9}$$

where the gamma $(\alpha,\, \beta)$ density function is equal to

$$\{\beta^\alpha w^{\alpha-1}/\Gamma(\alpha)\}\exp(-\beta w) I_{[0,\infty)}(w) \qquad (\alpha,\, \beta > 0), \tag{10}$$

and $I_A(w)$ denotes the indicator function that is 1 if w belongs to A and is zero otherwise.

If the conditional distribution of $\boldsymbol{X}_j$ given $W_j = w_j$ is specified by

$$\boldsymbol{X}_j \mid w_j \sim N(\boldsymbol{\mu}_i,\, \boldsymbol{\Sigma}_i/w_j), \tag{11}$$

then the unconditional distribution of $\boldsymbol{X}_j$ is given by the t distribution (6); see, for example, the monograph of [KN04] on the t distribution. As ν_i tends to infinity, the t distribution approaches the normal distribution. Hence this parameter ν_i may be viewed as a robustness tuning parameter. It can be fixed in advance or it can be inferred from the data for each component.

For ML estimation in the case of a single t distribution, the reader is referred to [Rub83], [LR87], [LR94], [LR95], [Liu97], and [LRW98]. A brief history of the development of ML estimation of a single-component t distribution is given in [LR95].

3 ML Estimation of Mixtures of t Components

[MP00a] have implemented the E- and M-steps of the EM algorithm and its variant, the ECM (expectation–conditional maximization) algorithm for the ML estimation of multivariate t components. The ECM algorithm proposed by [MR93] replaces the M-step of the EM algorithm by a number of computationally simpler conditional maximization (CM) steps.

In the EM framework for this problem, the unobservable variable w_j in the characterization (11) of the t distribution for the ith component of the t mixture model and the component-indicator labels z_{ij} are treated as being the "missing" data, where z_{ij} is defined to be one or zero according as $\boldsymbol{x}_j$ belongs or does not belong to the ith component of the mixture $(i = 1, \ldots, g;\; j = 1, \ldots, n)$. On the $(k+1)$th iteration of the EM algorithm, the updated estimates of the mixing proportion, the mean vector $\boldsymbol{\mu}_i$, and the scale matrix $\boldsymbol{\Sigma}_i$ are given by

$$\pi_i^{(k+1)} = \sum_{j=1}^{n} \tau_{ij}^{(k)}/n, \tag{12}$$

$$\boldsymbol{\mu}_i^{(k+1)} = \sum_{j=1}^{n} \tau_{ij}^{(k)} w_{ij}^{(k)} \boldsymbol{x}_j / \sum_{j=1}^{n} \tau_{ij}^{(k)} w_{ij}^{(k)} \tag{13}$$

and

$$\boldsymbol{\Sigma}_i^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_{ij}^{(k)} w_{ij}^{(k)} (\boldsymbol{x}_j - \boldsymbol{\mu}_i^{(k+1)})(\boldsymbol{x}_j - \boldsymbol{\mu}_i^{(k+1)})^T}{\sum_{j=1}^{n} \tau_{ij}^{(k)}}. \tag{14}$$

In the above,

$$\tau_{ij}^{(k)} = \frac{\pi_i^{(k)} f(\boldsymbol{x}_j;\, \boldsymbol{\mu}_i^{(k)},\, \boldsymbol{\Sigma}_i^{(k)},\, \nu_i^{(k)})}{f(\boldsymbol{x}_j;\, \boldsymbol{\Psi}^{(k)})} \tag{15}$$

is the posterior probability that $\boldsymbol{x}_j$ belongs to the ith component of the mixture, using the current fit $\boldsymbol{\Psi}^{(k)}$ for $\boldsymbol{\Psi}$ $(i = 1, \ldots, g;\, j = 1, \ldots, n)$. Also,

$$w_{ij}^{(k)} = \frac{\nu_i^{(k)} + p}{\nu_i^{(k)} + \delta(\boldsymbol{x}_j,\, \boldsymbol{\mu}_i^{(k)};\, \boldsymbol{\Sigma}_i^{(k)})}, \tag{16}$$

which is the current stimate of the conditional expectation of U_j given $\boldsymbol{x}_j$ and $z_{ij} = 1$.

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

$$\left\{ -\psi(\tfrac{1}{2}\nu_i) + \log(\tfrac{1}{2}\nu_i) + 1 + \frac{1}{n_i^{(k)}} \sum_{j=1}^{n} \tau_{ij}^{(k)} (\log w_{ij}^{(k)} - w_{ij}^{(k)}) \right.$$
$$\left. + \, \psi\left(\frac{\nu_i^{(k)} + p}{2}\right) - \log\left(\frac{\nu_i^{(k)} + p}{2}\right) \right\} = 0, \tag{17}$$

where $n_i^{(k)} = \sum_{j=1}^{n} \tau_{ij}^{(k)}$ $(i = 1, \ldots, g)$ and $\psi(\cdot)$ is the Digamma function.

Following the proposal of [KTV94] in the case of a single-component t distribution, we can replace the divisor $\sum_{j=1}^{n} \tau_{ij}^{(k)}$ in (14) by

$$\sum_{j=1}^{n} \tau_{ij}^{(k)} w_{ij}^{(k)},$$

which should improve the speed of convergence; see also [Liu97] and [LRW98].

These E- and M-steps are alternated until the changes in the estimated parameters or the log likelihood are less than some specified threshold. It can be seen that if the degrees of freedom ν_i is fixed in advance for each component, then the M-step exists in closed form. In this case where ν_i is fixed beforehand, the estimation of the component parameters is a form of M-estimation. However, an attractive feature of the use of the t distribution to model the component distributions is that the degrees of robustness as controlled by ν_i can be inferred from the data by computing its MLE.

4 Factor Analysis Model for Dimension Reduction

The g-component normal mixture model with unrestricted component-covariance matrices is a highly parameterized model with $d = \frac{1}{2}p(p+1)$ parameters for each component-covariance matrix $\boldsymbol{\Sigma}_i \, (i = 1, \ldots, g)$. [BR93] introduced a parameterization of the component-covariance matrix $\boldsymbol{\Sigma}_i$ based on a variant of the standard spectral decomposition of $\boldsymbol{\Sigma}_i \, (i = 1, \ldots, g)$. However, if p is large relative to the sample size n, it may not be possible to use this decomposition to infer an appropriate model for the component-covariance matrices. Even if it is possible, the results may not be reliable due to potential problems with near-singular estimates of the component-covariance matrices when p is large relative to n.

A common approach to reducing the number of dimensions is to perform a principal component analysis (PCA). But as is well known, projections of the feature data $\boldsymbol{x}_j$ onto the first few principal axes are not always useful in portraying the group structure; see [MP00a] and [Cha83]. 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 [MP00b]. This model was originally proposed by [GH97] and [HDR97] for the purposes of visualizing high dimensional data in a lower dimensional space to explore for group structure; see also [TB97] who considered the related model of mixtures of principal component analyzers for the same purpose. Further references may be found in [MP00a].

In the next section, 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 without any restrictions on the covariance matrices.

5 Mixtures of Normal 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 $\boldsymbol{X}_j$ 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 $\boldsymbol{X}_j$ can be modelled as

$$\boldsymbol{X}_j = \boldsymbol{\mu}_i + \boldsymbol{B}_i \boldsymbol{U}_{ij} + \boldsymbol{e}_{ij} \qquad \text{with prob. } \pi_i \quad (i = 1, \ldots, g) \tag{18}$$

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

$$f(\boldsymbol{x}_j; \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i \phi(\boldsymbol{x}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{19}$$

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

$$\boldsymbol{\Sigma}_i = \boldsymbol{B}_i \boldsymbol{B}_i^T + \boldsymbol{D}_i \quad (i = 1, \ldots, g) \tag{20}$$

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

The mixture of factor analyzers model can be fitted by using the alternating expectation–conditional maximization (AECM) algorithm ([MV97]). 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 $(\boldsymbol{\Psi}_1^T, \boldsymbol{\Psi}_2^T)^T$, 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 $\boldsymbol{\Psi}^{(k)} = (\boldsymbol{\Psi}_1^{(k)^T}, \boldsymbol{\Psi}_2^{(k)^T})^T$ 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 $\boldsymbol{\Psi}_2$.

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

$$\hat{\boldsymbol{D}}_i = \operatorname{diag}(\hat{\boldsymbol{V}}_i - \hat{\boldsymbol{B}}_i \hat{\boldsymbol{B}}_i^T), \tag{21}$$

where

$$\hat{\boldsymbol{V}}_i = \sum_{j=1}^{n} \tau_i(\boldsymbol{x}_j; \hat{\boldsymbol{\Psi}})\,(\boldsymbol{x}_j - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{x}_j - \hat{\boldsymbol{\mu}}_i)^T / \sum_{j=1}^{n} \tau_i(\boldsymbol{x}_j; \hat{\boldsymbol{\Psi}}). \tag{22}$$

As remarked by [LM71] in the context of direct computation of the ML estimate for a single-component factor analysis model, the equation (21) looks temptingly simple to use to solve for $\hat{\boldsymbol{D}}_i$, but was not recommended due to convergence problems.

On comparing (21) with (16), 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 $\boldsymbol{U}_j \boldsymbol{U}_j^T$, which is the identity matrix, is used in place of the conditional expectation in 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; see [MPB03] for further discussion.

It can be seen from (21) 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 $\boldsymbol{D}_i$,

$$\boldsymbol{D}_i = \boldsymbol{D} \quad (i = 1, \ldots, g). \tag{23}$$

An alternative way of proceeding is to adopt some prior distribution for the $\boldsymbol{D}_i$ as, for example, in the Bayesian approach of [FT02].

The mixture of probabilistic component analyzers (PCAs) model, as proposed by [TB97], has the form (20) with each $\boldsymbol{D}_i$ now having the isotropic structure

$$\boldsymbol{D}_i = \sigma_i^2 \boldsymbol{I}_p \quad (i = 1, \ldots, g). \tag{24}$$

Under this isotropic restriction (24) 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 $\boldsymbol{V}_i$.

6 Mixtures of *t* Factor Analyzers

The mixture of factor analyzers model is sensitive to outliers since it uses normal errors and factors. Recently, [MBB06] have considered the use of mixtures of t analyzers in an attempt to make the model less sensitive to outliers. With mixtures of t factor analyzers, the error terms $\boldsymbol{e}_{ij}$ and the factors $\boldsymbol{U}_{ij}$ are assumed to be distributed according to the t distribution with the same degrees of freedom. Under this model, the factors and error terms are no longer independently distributed but they are uncorrelated.

Following [MBB06], we now formulate our mixture of t analyzers model by replacing the multivariate normal distribution in (19) for the ith component-conditional distribution of $\boldsymbol{X}_j$ by the multivariate t distribution with mean vector vector $\boldsymbol{\mu}_i$, scale matrix $\boldsymbol{\Sigma}_i$, and ν_i degrees of freedom with the factor analytic restriction (20) on the component-scale matrices $\boldsymbol{\Sigma}_i$. Thus our postulated mixture model of t factor analyzers assumes that $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ is an observed random sample from the t mixture density

$$f(\boldsymbol{x}_j; \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i f_{\mathrm{t}}(\boldsymbol{x}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \nu_i), \tag{25}$$

where

$$\boldsymbol{\Sigma}_i = \boldsymbol{B}_i \boldsymbol{B}_i^T + \boldsymbol{D}_i \quad (i = 1, \ldots, g) \tag{26}$$

and where now 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 $\boldsymbol{\mu}_i$, $\boldsymbol{B}_i$, and the $\boldsymbol{D}_i$ $(i = 1, \ldots, g)$. As in the mixture of factor analyzers model, $\boldsymbol{B}_i$ is a $p \times q$ matrix and $\boldsymbol{D}_i$ is a diagonal matrix.

In order to fit this model (25) with the restriction (26), it is computationally convenient to exploit its link with factor analysis. Accordingly, corresponding to (18), we assume that

$$\boldsymbol{X}_j = \boldsymbol{\mu}_i + \boldsymbol{B}_i \boldsymbol{U}_{ij} + \boldsymbol{e}_{ij} \qquad \text{with prob. } \pi_i \quad (i = 1, \ldots, g) \tag{27}$$

for $j = 1, \ldots, n$, where the joint distribution of the factor $\boldsymbol{U}_{ij}$ and of the error $\boldsymbol{e}_{ij}$ needs to be specified so that it is consistent with the t mixture formulation (25) for the marginal distribution of $\boldsymbol{X}_j$.

For the normal factor analysis model, we have that conditional on membership of the ith component of the mixture the joint distribution of $\boldsymbol{X}_j$ and its associated factor (vector) $\boldsymbol{U}_{ij}$ is multivariate normal,

$$\begin{pmatrix} \boldsymbol{X}_j \\ \boldsymbol{U}_{ij} \end{pmatrix} \mid z_{ij} = 1 \ \sim N_{p+q}(\boldsymbol{\mu}_i^*, \boldsymbol{\xi}_i) \quad (i = 1, \ldots, g). \tag{28}$$

where the mean $\boldsymbol{\mu}_i^*$ and the covariance matrix $\boldsymbol{\xi}_i$ are defined by

$$\boldsymbol{\mu}_i^* = (\boldsymbol{\mu}_i^T,\, \mathbf{0}^T)^T \tag{29}$$

and the covariance matrix $\boldsymbol{\xi}_i$ is given by

$$\boldsymbol{\xi}_i = \begin{pmatrix} \boldsymbol{B}_i\boldsymbol{B}_i^T + \boldsymbol{D}_i & \boldsymbol{B}_i \\ \boldsymbol{B}_i^T & \boldsymbol{I}_q \end{pmatrix}. \tag{30}$$

We now replace the normal distrubution by the t distribution in (28) to postulate that

$$\begin{pmatrix} \boldsymbol{Y}_j \\ \boldsymbol{U}_{ij} \end{pmatrix} \mid z_{ij} = 1 \;\sim t_{p+q}(\boldsymbol{\mu}_i^*, \boldsymbol{\xi}_i, \nu_i) \quad (i = 1, \ldots, g). \tag{31}$$

This specification of the joint distribution of $\boldsymbol{X}_j$ and its associated factors in (27) will imply the t mixture model (25) for the marginal distribution of $\boldsymbol{X}_j$ with the restriction (26) on its component-scale matrices.

Using the characterization of the t distribution discussed earlier, it follows that we can express (26) alternatively as

$$\begin{pmatrix} \boldsymbol{Y}_j \\ \boldsymbol{U}_{ij} \end{pmatrix} \mid w_j, z_{ij} = 1 \;\sim N_{p+q}(\boldsymbol{\mu}_i^*,\, \boldsymbol{\xi}_i/w_j), \tag{32}$$

where w_{ij} is a value of the weight variable W_j taken to have the gamma distribution (10).

It can be established from (32) that

$$\boldsymbol{U}_{ij} \mid w_j, z_{ij} = 1 \;\sim N_q(\mathbf{0},\, \boldsymbol{I}_q/w_j) \tag{33}$$

and

$$\boldsymbol{e}_{ij} \mid z_{ij} = 1 \;\sim N_p(\mathbf{0},\, \boldsymbol{D}_i/w_j), \tag{34}$$

and hence that

$$\boldsymbol{U}_{ij} \mid z_{ij} = 1 \;\sim t_q(\mathbf{0}, \boldsymbol{I}_q, \nu_i) \tag{35}$$

and

$$\boldsymbol{e}_{ij} \mid z_{ij} = 1 \;\sim t_p(\mathbf{0}, \boldsymbol{D}_i, \nu_i). \tag{36}$$

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 factor analysis, but they are uncorrelated. To see this, we have from (32) that conditional on w_j, $\boldsymbol{U}_{ij}$ and $\boldsymbol{e}_{ij}$ are uncorrelated, and hence, unconditionally uncorrelated.

We fit the mixture of t factor analyzers model specified by (25) and (26) using the AECM algorithm ([MV97]), as described in [MBB06].

7 Discussion

We have considered the use of mixtures of multivariate t distributions instead of normal components as a more robust approach to the clustering of multivariate continuous data which have longer tails that the normal or atypical observations.

As pointed out by [Hen04], although the number of outliers needed for breakdown with the t mixture model is almost the same as with the normal version, the outliers have to be much larger.

In considering the robustness of mixture models, it is usual to consider the number of components as fixed. This is because the existence of outliers in a data set can be handled by the addition of further components in the mixture model if the number of components is not fixed. Breakdown can still occur if the contaminating points lie between the clusters of the main body of points and fill in the feature space to the extent that a fewer number of components is needed in the mixture model than the actual number of clusters ([Hen04]). But obviously the situation is fairly straightforward if the number of clusters are known *a priori*. However, this is usually not the case in clustering applications.

We consider also the case of clustering high-dimensional feature data via normal mixture models. These models can be fitted by adopting the factor analysis model to represent the component-covariance matrices. It is shown how the resulting model known as mixtures of factor analyzers can be made more robust by using the multivariate t distribution for the component distributions of the factors and errors.

Examples will be presented in the oral presentation and computational aspects associated with these approaches further discussed and illustrated.

References

[BR93] Banfield, J.D., Raftery, A.E.: Model-based Gaussian and non-Gaussian clustering. Biometrics, **49**, 803–821 (1993)

[Cam84] Campbell, N.A.: Mixture models and atypical values. Math. Geol., **16**, 465–477 (1984)

[Cha83] Chang, W.C.: On using principal components before separating a mixture of two multivariate normal distributions. Appl. Stat., **32**, 267–275 (1983)

[CDH99] Coleman, D., Dong, X., Hardin, J., Rocke, D.M., Woodruff, D.L.: Some computational issues in cluster analysis with no a priori metric. Comp. Stat. Data Anal., **31**, 1–11 (1999)

[DG05] Davies, P.L., Gather, U.: Breakdown and groups (with discussion). Ann. Stat., **33**, 977–1035 (2005)

[DLR77] Dempster, A.P, Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. R. Stat. Soc. B, **39**, 1–38 (1977)

[DH83] Donoho, D.L., Huber, J.: The notion of breakdown point. In: Bickel, P.J., Doksum, K.A., Hodges, J.L. (eds) A Festschrift for Erich L. Lehmann. Wadsworth, Belmont, CA (1983)

[FT02] Fokoué, E., Titterington, D.M.: Mixtures of factor analyzers. Bayesian estimation and inference by stochastic simulation. Mach. Learn., **50**, 73–94 (2002)

[GH97] Ghahramani, Z., Hinton, G.E.: The EM algorithm for mixtures of factor analyzers. Techncial Report, University of Toronto (1997)

[HL97] Hadi, A.S., Luccño, A.: Maximum trimmed likelihood estimators: a unified approach, examples, and algorithms. Comp. Stat. Data Anal., **25**, 251–272 (1997)

[Ham71] Hampel, F.R. A general qualitative definition of robustness. Ann. Math. Stat., **42**, 1887–1896 (1971)

[Har75] Hartigan, J.A.: Statistical theory in clustering. J. Classif., **2**, 63–76 (1975)

[Hen04] Hennig, C.: Breakdown points for maximum likelihood estimators of location-scale mixtures. Ann. Stat., **32**, 1313–1340 (2004)

[HDR97] Hinton, G.E., Dayan, P., Revov, M.: Modeling the manifolds of images of handwritten digits. IEEE Trans. Neur. Networks, **8**, 65–73

[Hub81] Huber, P.J.: Robust Statistics. Wiley, New York (1981)

[KTV94] Kent, J.T., Tyler, D.E., Vardi, Y.: A curious likelihood identity for the multivariate t-distribution. Comm. Stat. Sim Comp., **23**, 441–453 (1994)

[KN04] Kotz, S. Nadarajah, S.: Multivariate t distributions and their applications. Cambridge University Press, New York (2004)

[LM71] Lawley, D.N., Maxwell, A.E.: Factor Analysis as a Statistical Method. Butterworths, London (1971)

[LR87] Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data. Wiley, New York (1987)

[Liu97] Liu, C.: ML estimation of the multivariate t distribution and the EM algorithm. J. Multiv. Anal., **63**, 296–312 (1997)

[LR94] Liu, C., Rubin, D.B.: The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. Biometrika, **81**, 633–648 (1994)

[LR95] Liu, C., Rubin, D.B.: ML estimation of the t distribution using EM and its extensions, ECM and ECME. Statistica Sinica, 5:19–39 (1995)

[LRW98] Liu, C., Rubin, D.B., Wu, Y.N.: Parameter expansion to accelerate EM: the PX-EM algorithm. Biometrika, **85**, 755–770 (1998)

[Mar00] Markatou, M.: Mixture models, robustness and the weighted likelihood methodology. Biom., **56**, 483–486 (2000)

[MBL98] Markatou, M., Basu, A., Lindsay, B.G.: Weighted likelihood equations with bootstrap root search. J. Amer. Stat. Assoc., **93**, 740–750 (1998)

[MB88] McLachlan, G.J., Basford, K.E.: Mixture Models: Inference and Applications to Clustering. Marcel Dekker, New York (1988)

[MP98] McLachlan, G.J., Peel, D.: Robust cluster analysis via mixtures of multivariate t distributions. Lec. Notes Comput. Sci., **1451**, 658–666 (1998)

[MP00a] McLachlan, G.J., Peel, D.: Finite Mixture Models. Wiley, New York (2000)

[MP00b] McLachlan, G.J., Peel, D.: Mixtures of factor analyzers. In: Langley, P. (ed) Proceedings of the Seventeenth International Conference on Machine Learning. Morgan Kaufmann, San Francisco (2000)

[MBB06] McLachlan, G.J., Bean, R.W., Ben-Tovim Jones, L.: Extension of mixture of factor analyzers model to incorporate the multivariate t distribution. To appear in Comp. Stat. Data Anal. (2006)

[MNB06] McLachlan, G.J., Ng, S.-K., Bean, R.W.: Robust cluster analysis via mixture models. To appear in Aust. J. Stat. (2006)

[MPB03] McLachlan, G.J., Peel, D., Bean, R.: Modelling high-dimensional data by mixtures of factor analyzers. Comp. Stat. Data Anal., **41**, 379–388 (2003)

[MV97] Meng, X.L., van Dyk, D.: The EM algorithm—an old folk song sung to a fast new tune (with discussion). J. R. Stat. Soc. B, **59**, 511–567 (1997)

[MR93] Meng, X.L., Rubin, D.B.: Maximum likelihood estimation via the ECM algorithm: a general framework. Biometrika, **80**, 267–278 (1993)

[MN04] Müller, C.H., Neykov, N.: Breakdown points of trimmed likelihood estimators and related estimators in generalized linear models. J. Stat. Plann. Infer., **116**, 503–519 (2004)

[NFD04] Neykov, N., Filzmoser, P., Dimova, R., Neytchev, P.: Compstat 2004, Proceedings Computational Statistics. Physica-Verlag, Vienna (2004)

[PM00] Peel, D., McLachlan, G.J.: Robust mixture modelling using the *t* distribution. Stat. Comput., **10**, 335–344 (2000)

[ROC96] Rocke, D.M.: Robustness properties of S-estimators of multivariate location and shape in high dimension. Ann. Stat., **24**, 1327–1345 (1996)

[RW96] Rocke, D.M., Woodruff, D.L.: Identification of outliers in multivariate data. J. Amer. Stat. Assoc., **91**, 1047-1061 (1996)

[RW97] Rocke, D.M., Woodruff, D.L.: Robust estimation of multivariate location and shape. J. Stat. Plann. Infer., **57**, 245–255 (1997)

[Rub83] Rubin, D.B.: Iteratively reweighted least squares. In: Kotz, S., Johnson, N.L., and Read, C.B. (eds) Encyclopedia of Statistical Sciences, Vol. 4. Wiley, New York (1983)

[TK99] Tibshirani, R., Knight, K.: Model search by bootstrap "bumping". J. Comp. Graph. Stat., **8**, 671–686 (1999)

[TB97] Tipping, M.E., Bishop, C.M.: Mixtures of probabilistic principal component analysers. Technical Report, Neural Computing Research Group, Aston University (1997)

[VN98] Vandev, D.L., Neykov, N.: About regression estimators with high breakdown point. Ann. Stat., **32**, 111–129 (1998)

[WR93] Woodruff, D.L., Rocke, D.M.: Heuristic search algorithms for the minimum volume ellipsoid. J. Comp. Graph. Stat., **2**, 69–95 (1993)

[WR94] Woodruff, D.L., Rocke, D.M.: Computable robust estimation of multivariate location and shape using compound estimators. J. Amer. Stat. Assoc., **89**, 888–896 (1994)