ON COMPUTATIONAL ASPECTS OF CLUSTERING VIA MIXTURES OF NORMAL AND t-COMPONENTS

G.J. McLachlan and D. Peel, University of Queensland
G.J. McLachlan, Department of Mathematics,
University of Queensland, Brisbane, Queensland 4072, AUSTRALIA

Key Words: Finite mixture models, Normal components, Multivariate t components, Maximum likelihood, EM algorithm, Cluster analysis

ABSTRACT Normal mixture models are being increasingly used to model the distributions of a wide variety of random phenomena and to cluster sets of continuous multivariate data. However, for a set of data containing a group or groups of observations with longer than normal tails or atypical observations, the use of normal components may unduly affect the fit of the mixture model. We consider here a robust approach by modeling the data by a mixture of multivariate t distributions. Examples of its use are given in the context of clustering multivariate data in the presence of atypical observations, which may be in the form of background noise. In presenting these examples, we also consider some of the computational issues in clustering via normal and t mixture models.

1. INTRODUCTION

Finite mixtures of distributions have provided a mathematical-based approach to the statistical modeling of a wide variety of random phenomena; see, for example, McLachlan and Basford (1988) and Fraley and Raftery (1999). Because of their usefulness as an extremely flexible method of modeling, finite mixture models have continued to receive increasing attention over the years, both from a practical and theoretical point of view. For multivariate data of a continuous nature, attention has focused on the use of multivariate normal components because of their computational convenience. They can be easily fitted iteratively by maximum likelihood (ML) via the expectation-maximization (EM) algorithm of Dempster, Laird, and Rubin (1977); see also McLachlan and Krishnan (1997).

With a normal mixture model-based approach to clustering, it is assumed that the data to be clustered are from a mixture of an initially specified number g of multivariate normal densities in some 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.),

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

where $\phi(\boldsymbol{y};\, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ denotes the p-variate normal density probability function with mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$. 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 matrics $\boldsymbol{\Sigma}_i$. Once the mixture model has been fitted, 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.

However, for many applied problems, the tails of the normal distribution are often shorter than required. Also, the estimates of the component means and covariance matrices can be affected by observations that are atypical of the components in the normal mixture model being fitted. 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 (Rocke and Woodruff, 1997).

In this paper, we consider the fitting of mixtures of (multivariate) t distributions. The t distribution provides a longer tailed alternative to the normal distribution. Hence it provides a more robust approach to the fitting of normal mixture models, as observations that are atypical of a component are given reduced weight in the calculation of its parameters. Also, the use of t components gives less extreme estimates of the posterior probabilities of component membership of the mixture model, as demonstrated in McLachlan and Peel (1998); see also Peel and

McLachlan (2000) and McLachlan and Peel (2000a, Chapter 7).

Robust estimation in the context of mixture models has been considered in the past by Campbell (1984), McLachlan and Basford (1988, Chapter 3), and De Veaux and Kreiger (1990), among others, using M-estimates of the means and covariance matrices of the normal components of the mixture model. Recently, Markatou (1998) has provided a formal approach to robust mixture estimation by applying weighted likelihood methodology in the context of mixture models.

2. FITTING OF NORMAL MIXTURES

Before we proceed to consider the fitting of normal mixture models, we consider some of the computational issues associated with the fitting of mixtures of normal components. The maximum likelihood estimate of $\boldsymbol{\Psi}$ is obtained as an appropriate root of the likelihood equation

$$\partial \log L(\boldsymbol{\Psi})/\partial \boldsymbol{\Psi} = \mathbf{0}, \tag{2}$$

where $L(\boldsymbol{\Psi})$ denotes the likelihood function for $\boldsymbol{\Psi}$ formed from the observed data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$. Solutions of (2) corresponding to local maxima can be found by application of the EM algorithm. The latter is applied in the framework where an observation $\boldsymbol{y}_j$ is conceptualized to have arisen from one of the components and the indicator variable denoting its component of origin is taken to be missing. The E-step of the EM algorithm thus involves replacing these unobservable indicator variables by their conditional expectations given the observed data, since the complete-data log likelihood is linear in them; see, for example, McLachlan Peel (2000a, Chapter 3). It follows that on the $(k+1)$th iteration of the EM algorithm, the estimates of $\pi_i, \boldsymbol{\mu}_i$, and $\boldsymbol{\Sigma}_i$ are given by

$$\pi_i^{(k+1)} = \sum_{j=1}^{n} \tau_{ij}^{(k)}/n, \tag{3}$$

$$\boldsymbol{\mu}_i^{(k+1)} = \sum_{j=1}^{n} \tau_{ij}^{(k)} \boldsymbol{y}_j / \sum_{j=1}^{n} \tau_{ij}^{(k)}, \tag{4}$$

and

$$\boldsymbol{\Sigma}_i^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_{ij}^{(k)} (\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k+1)})(\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k+1)})^T}{\sum_{j=1}^{n} \tau_{ij}^{(k)}}, \tag{5}$$

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

$$\tau_{ij}^{(k)} = \frac{\pi_i^{(k)} \phi(\boldsymbol{y}_j;\, \boldsymbol{\mu}_i^{(k)}, \boldsymbol{\Sigma}_i^{(k)})}{f(\boldsymbol{y}_j;\, \boldsymbol{\Psi}^{(k)})} \tag{6}$$

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

We note in passing that with the development of Markov chain Monte Carlo methods, there has been increasing attention given to the implementation of the Bayesian approach for mixture models; see, for example, Lavine and West (1992) and Diebolt and Robert (1994) and the references in McLachlan and Peel (2000a, Chapter 4).

3. SOME COMPUTATIONAL ISSUES

As the likelihood equation (2) tends to have multiple roots corresponding to local maxima, the EM algorithm needs to be started from a variety of initial values for the parameter vector $\boldsymbol{\Psi}$, or equivalently, for a variety of initial partitions of the data into g groups. The latter can be obtained by randomly dividing the data into g groups corresponding to the g components of the mixture model. With random starts, the effect of the central limit theorem tends to have the component parameters initially being similar at least in large samples. One way to reduce this effect is to first select a small random subsample from the data, which is then randomly assigned to the g components. The first M-step is then performed on the basis of the subsample. The subsample has to be sufficiently large to ensure that the first M-step is able to produce a nondegenerate estimate of the parameter vector $\boldsymbol{\Psi}$.

An alternative method of specifying a random start in the context of g normal components with means $\boldsymbol{\mu}_i$ and covariance matrices $\boldsymbol{\Sigma}_i$, is to randomly generate the $\boldsymbol{\mu}_i^{(0)}$ independently as

$$\boldsymbol{\mu}_1^{(0)}, \ldots, \boldsymbol{\mu}_g^{(0)} \overset{\text{i.i.d.}}{\sim} N(\overline{\boldsymbol{y}}, \boldsymbol{V}), \tag{7}$$

where $\overline{\boldsymbol{y}}$ is the sample mean and

$$\boldsymbol{V} = \sum_{j=1}^{n} (\boldsymbol{y}_j - \overline{\boldsymbol{y}})(\boldsymbol{y}_j - \overline{\boldsymbol{y}})^T/n \tag{8}$$

is the sample covariance matrix of the observed data. With this method, there is more variation between the initial values $\boldsymbol{\mu}_i^{(0)}$ for the component means $\boldsymbol{\mu}_i$ than with a random partition of the data into g groups, and it is also computationally less demanding.

The component-covariance matrices $\boldsymbol{\Sigma}_i$ and the mixing proportions π_i can be specified as

$$\boldsymbol{\Sigma}_i^{(0)} = \boldsymbol{V} \text{ and } \pi_i^{(0)} = 1/g \qquad (i = 1, \ldots, g). \quad (9)$$

As the EM algorithm updates the posterior probabilities of component membership for each observation before the next M-step is performed, it can take some time to implement for large data sets. Hence variants of the EM algorithm have been considered for reducing the computation time. With the incremental EM (IEM) algorithms as proposed by Neal and Hinton (1998), the n available observations are divided into $B\,(B \leq n)$ blocks and the E-step is implemented for only a block of observations at a time before the next M-step is performed. The choice of the number of blocks B so as to optimize the convergence time of the IEM algorithm is an interesting problem. The initial work of McLachlan and Ng (2000) suggests using $B \approx n^{2/5}$ as a simple guide. The optimal choice will depend on the number of unknown parameters, and McLachlan and Ng (2000) suggest modifying this guide to $B \approx n^{1/3}$ in the case of component-covariance matrices specified to be diagonal. Neal and Hinton (1998) also proposed a sparse version of the EM algorithm for which only those component-posterior probabilities that are above a specified threshold are updated. McLachlan and Ng (2000) have considered a sparse version of the IEM algorithm that reduced the time to convergence of the standrad EM by up to a factor of 56% in some examples. Thiesson et al. (2000) suggest a search method to choose the number of blocks.

Other approaches for speeding up the EM algorithm for mixtures have been considered by Moore (1999) and Bradley, Fayyad, and Reina (1999). The former has made use of multiresolution kd-trees to speed up the fitting process of the EM algorithm. Impressive results were obtained for the reduction in the time to convergence for small values of the dimension p of the observation vector. However, as p increases, the time taken to fit the kd-tree will increase and so the reduction in computation time will fall away; see Thiesson et al. (2000). Bradley et al. (1999) developed a scalable version of the EM algorithm for handling very large databases with a limited memory buffer. It is based on identifying regions that are compressible and regions that must be maintained in memory.

4. MODELS FOR THE COVARIANCE MATRICES

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

$$\boldsymbol{\Sigma}_i = \sum_{v=1}^{p} \lambda_{iv} \boldsymbol{a}_{iv} \boldsymbol{a}_{iv}^T, \quad (10)$$

where $\boldsymbol{a}_{i1}, \ldots, \boldsymbol{a}_{ip}$ denote the eigenvectors corresponding to the eigenvalues $\lambda_{i1} \geq \lambda_{i2} \geq \cdots \lambda_{ip} > 0$ of $\boldsymbol{\Sigma}_i\,(i = 1, \ldots, g)$. They expressed $\boldsymbol{\Sigma}_i$ further as

$$\boldsymbol{\Sigma}_i = \lambda_i \boldsymbol{A}_i \boldsymbol{\Lambda}_i \boldsymbol{A}_i^T, \quad (11)$$

where $\boldsymbol{A}_i = (\boldsymbol{a}_{i1}, \ldots, \boldsymbol{a}_{ip})$ is the (orthogonal) matrix of the eigenvectors of $\boldsymbol{\Sigma}_i$. Conventions for normalizing λ_i and $\boldsymbol{\Lambda}_i$ include taking $\lambda_i = \lambda_{i1}$ (the largest eigenvalue of $\boldsymbol{\Sigma}_i$) for which then

$$\boldsymbol{\Lambda}_i = \operatorname{diag}(1, \lambda_{i2}/\lambda_{i1}, \ldots, \lambda_{ip}/\lambda_{i1}). \quad (12)$$

Another requires $\mid \boldsymbol{\Lambda}_i \mid = 1$ for which $\lambda_i = \mid \boldsymbol{\Sigma}_i \mid^{1/p}$ and

$$\boldsymbol{\Lambda}_i = \operatorname{diag}(\lambda_{i1}/\lambda_i, \ldots, \lambda_{ip}/\lambda_i).$$

The parameter λ_i controls the volume of the cluster corresponding to the ith component, $\boldsymbol{\Lambda}_i$ its shape, and $\boldsymbol{A}_i$ its orientation. A reduction in the number of parameters is achieved by imposing various constraints on the $\boldsymbol{A}_i, \boldsymbol{\Lambda}_i$, and the λ_i. For example, the constraint $\boldsymbol{A}_i = \boldsymbol{A}\,(i = 1, \ldots, g)$ imposes the same orientation on the g clusters.

Another approach for reducing the number of unknown parameters in the forms for the component-covariance matrices is to adopt the mixtures of factor analyzers model, as considered in McLachlan and Peel (2000a, 2000b). This model was originally proposed by Ghahramani and Hinton (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) who considered the related model of mixtures of principal component analyzers for the same purpose. With the mixture of factor analyzers model, the ith component-covariance matrix $\boldsymbol{\Sigma}_i$ has the form

$$\boldsymbol{\Sigma}_i = \boldsymbol{B}_i \boldsymbol{B}_i^T + \boldsymbol{D}_i \quad (i = 1, \ldots, g), \quad (13)$$

where $\boldsymbol{B}_i$ is a $p \times q$ matrix of factor loadings and $\boldsymbol{D}_i$ is a diagonal matrix.

If q is chosen sufficiently smaller than p, the representation (13) imposes some constraints on the component-covariance matrix $\boldsymbol{\Sigma}_i$ 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 $\boldsymbol{B}_i$, since (13) is still satisfied if $\boldsymbol{B}_i$ is replaced by $\boldsymbol{B}_i\boldsymbol{C}_i$, where $\boldsymbol{C}_i$ is any orthogonal matrix of order q. One (arbitrary) way of uniquely specifying $\boldsymbol{B}_i$ is to choose the orthogonal matrix $\boldsymbol{C}_i$ so that $\boldsymbol{B}_i^T\boldsymbol{D}_i^{-1}\boldsymbol{B}_i$ is diagonal (with its diagonal elements arranged in decreasing order); see Lawley and Maxwell (1971, Chapter 1). Assuming that the eigenvalues of $\boldsymbol{B}_i\boldsymbol{B}_i^T$ are positive and distinct, the condition that $\boldsymbol{B}_i^T\boldsymbol{D}_i^{-1}\boldsymbol{B}_i$ is diagonal as above imposes $\frac{1}{2}q(q-1)$ constraints on the parameters. Hence then the number of free parameters is

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

5. MULTIVARIATE t DISTRIBUTION

In the above, the component densities have been modeled by the normal parametric family. One way to broaden this family for potential outliers or data with longer than normal tails is to adopt the normal scale mixture model, which can be written as

$$\int \phi(\boldsymbol{y}; \boldsymbol{\mu}; \boldsymbol{\Sigma}/u)dH(u), \qquad (14)$$

where H is the probability distribution that places mass $(1-\epsilon)$ at the point $u = 1$ and mass ϵ at the point $u = 1/c$. Suppose we now replace H by the p.d.f. of a chi-squared random variable on its degrees of freedom ν. We then obtain the t density with location parameter $\boldsymbol{\mu}$, positive definite inner product matrix $\boldsymbol{\Sigma}$, and ν degrees of freedom,

$$f(\boldsymbol{y};\, \boldsymbol{\mu},\, \boldsymbol{\Sigma},\, \nu) =$$

$$\frac{\Gamma(\frac{\nu+p}{2})\,|\boldsymbol{\Sigma}|^{-1/2}}{(\pi\nu)^{\frac{1}{2}p}\Gamma(\frac{\nu}{2})\{1+\delta(\boldsymbol{y},\,\boldsymbol{\mu};\,\boldsymbol{\Sigma})/\nu\}^{\frac{1}{2}(\nu+p)}}, \qquad (15)$$

where

$$\delta(\boldsymbol{y},\, \boldsymbol{\mu};\, \boldsymbol{\Sigma}) = (\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}) \qquad (16)$$

denotes the Mahalanobis squared distance between $\boldsymbol{y}$ and $\boldsymbol{\mu}$ (with $\boldsymbol{\Sigma}$ as the covariance matrix). If $\nu > 1$, $\boldsymbol{\mu}$ is the mean of $\boldsymbol{Y}$, and if $\nu > 2$, $\nu(\nu-2)^{-1}\boldsymbol{\Sigma}$ is its covariance matrix. As ν tends to infinity, U converges to one with probability one, and so $\boldsymbol{Y}$ becomes marginally multivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The family of t distributions thus provides a heavy-tailed alternative to the normal family with mean $\boldsymbol{\mu}$ and covariance matrix that is equal to a scalar multiple of $\boldsymbol{\Sigma}$ (if $\nu > 2$).

6. ML ESTIMATION OF t DISTRIBUTION

A brief history of the development of ML estimation of a single component t distribution is given in Liu and Rubin (1995). An account of more recent work is given in Liu (1997). Liu and Rubin (1994, 1995) have shown that the ML estimates can be found much more efficiently by using an extension of the EM algorithm called the expectation-conditional maximization either (ECME) algorithm. Meng and van Dyk (1997) demonstrated that the more promising versions of the ECME algorithm for the t distribution can be obtained using alternative data augmentation schemes. They called this algorithm the alternating expectation-conditional maximization (AECM) algorithm. Following Meng and van Dyk (1997), Liu (1997) considered a class of data augmentation schemes even more general than the class of Meng and van Dyk (1997). This led to new versions of the ECME algorithm for ML estimation of the t distribution with possible missing values, corresponding to applications of the parameter-expanded EM (PX-EM) algorithm (Liu, Wu, and Rubin, 1998).

7. ML ESTIMATION OF MIXTURE OF t DISTRIBUTIONS

We consider now ML estimation for a g-component mixture of t distributions, given by

$$f(\boldsymbol{y};\, \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i f(\boldsymbol{y};\, \boldsymbol{\mu}_i,\, \boldsymbol{\Sigma}_i,\, \nu_i), \qquad (17)$$

where

$$\boldsymbol{\Psi} = (\pi_1,\, \ldots\,,\, \pi_{g-1},\, \boldsymbol{\theta}^T,\, \boldsymbol{\nu}^T)^T,$$

$\boldsymbol{\nu} = (\nu_1,\, \ldots\,,\, \nu_g)^T$, and $\boldsymbol{\xi} = (\boldsymbol{\theta}_1^T,\, \ldots\,,\, \boldsymbol{\theta}_g^T)^T$, and where $\boldsymbol{\theta}_i$ contains the elements of $\boldsymbol{\mu}_i$ and the distinct elements of $\boldsymbol{\Sigma}_i$ $(i = 1,\, \ldots\,,\, g)$. The application of the EM algorithm for ML estimation in the case of a single component t distribution has been described in McLachlan and Krishnan (1997, Sections 2.6 and 5.8). The results there can be extended to cover the present case of a g-component mixture of multivariate t distributions.

In the EM-framework, the complete-data vector is given by

$$\boldsymbol{x}_{\mathrm{c}} = (\boldsymbol{x}_{\mathrm{o}}^T,\, \boldsymbol{z}_1^T,\, \ldots\,,\, \boldsymbol{z}_n^T,\, u_1,\, \ldots\,,\, u_n)^T \qquad (18)$$

where $\boldsymbol{x}_{\mathrm{o}} = (\boldsymbol{y}_1^T,\, \ldots\,,\, \boldsymbol{y}_n^T)^T$ denotes the observed-data vector, $\boldsymbol{z}_1,\, \ldots\,,\, \boldsymbol{z}_n$ are the component-label vectors defining the component of origin of $\boldsymbol{y}_1,\, \ldots\,,$

$\boldsymbol{y}_n$, respectively, and $z_{ij} = (\boldsymbol{z}_j)_i$ is 1 or zero, according as to whether $\boldsymbol{y}_j$ belongs or does not belong to the ith component. In the light of the above characterization of the t distribution, it is convenient to view the observed data augmented by the $\boldsymbol{z}_j$ as still being incomplete and introduce into the complete-data vector the additional missing data, $u_1, \ldots, u_n$, which are defined so that given $z_{ij} = 1$,

$$\boldsymbol{Y}_j \mid u_j, z_{ij} = 1 \;\sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i/u_j), \tag{19}$$

independently for $j = 1, \ldots, n$, and

$$U_j \mid z_{ij} = 1 \sim \text{gamma}\,(\tfrac{1}{2}\nu_i,\, \tfrac{1}{2}\nu_i). \tag{20}$$

Given $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$, the $U_1, \ldots, U_n$ are independently distributed according to (20).

The complete-data likelihood $L_c(\boldsymbol{\Psi})$ can be factored into the product of the marginal densities of the $\boldsymbol{Z}_j$, the conditional densities of the U_j given the $\boldsymbol{z}_j$, and the conditional densities of the $\boldsymbol{Y}_j$ given the u_j and the $\boldsymbol{z}_j$.

8. E-STEP

The E-step on the $(k+1)$th iteration of the EM algorithm requires the calculation of $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})$, the current conditional expectation of the complete-data log likelihood function $\log L_c(\boldsymbol{\Psi})$. This E-step can be effected by first taking the expectation of $\log L_c(\boldsymbol{\Psi})$ conditional also on $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$, as well as $\boldsymbol{x}_o$, and then finally over the $\boldsymbol{z}_j$ given $\boldsymbol{x}_o$. It follows (Peel and McLachlan, 2000) that the conditional expectation of the complete-data log likelihood, $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})$, is given by

$$\begin{aligned} Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)}) &= Q_1(\boldsymbol{\pi}; \boldsymbol{\Psi}^{(k)}) + Q_2(\boldsymbol{\nu}; \boldsymbol{\Psi}^{(k)}) \\ &\quad + Q_3(\boldsymbol{\xi}; \boldsymbol{\Psi}^{(k)}), \end{aligned} \tag{21}$$

where

$$Q_1(\boldsymbol{\pi}; \boldsymbol{\Psi}^{(k)}) = \sum_{i=1}^{g}\sum_{j=1}^{n} \hat{\tau}_{ij}^{(k)} \log \pi_i, \tag{22}$$

$$Q_2(\boldsymbol{\nu}; \boldsymbol{\Psi}^{(k)}) = \sum_{i=1}^{g}\sum_{j=1}^{n} \hat{\tau}_{ij}^{(k)} Q_{2j}(\nu_i; \boldsymbol{\Psi}^{(k)}), \tag{23}$$

and

$$Q_3(\boldsymbol{\xi}; \boldsymbol{\Psi}^{(k)}) = \sum_{i=1}^{g}\sum_{j=1}^{n} \hat{\tau}_{ij}^{(k)} Q_{3j}(\boldsymbol{\theta}_i; \boldsymbol{\Psi}^{(k)}), \tag{24}$$

and where, on ignoring terms not involving the ν_i,

$$\begin{aligned} Q_{2j}(\nu_i; \boldsymbol{\Psi}^{(k)}) &= -\log\Gamma(\tfrac{1}{2}\nu_i) + \tfrac{1}{2}\nu_i \log(\tfrac{1}{2}\nu_i) \\ &\quad + \tfrac{1}{2}\nu_i\{\sum_{j=1}^{n}(\log u_{ij}^{(k)} - u_{ij}^{(k)}) \\ &\quad + \psi(\frac{\nu_i^{(k)}+p}{2}) - \log(\frac{\nu_i^{(k)}+p}{2})\} \end{aligned} \tag{25}$$

and

$$\begin{aligned} Q_{3j}(\boldsymbol{\theta}_i; \boldsymbol{\Psi}^{(k)}) &= \{-\tfrac{1}{2}p\log(2\pi) - \tfrac{1}{2}\log|\boldsymbol{\Sigma}_i| \\ &\quad + \tfrac{1}{2}p\log u_{ij}^{(k)} \\ &\quad - \tfrac{1}{2}u_{ij}(\boldsymbol{y}_j - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_j - \boldsymbol{\mu}_i)\}. \end{aligned} \tag{26}$$

In the above,

$$\tau_{ij}^{(k)} = \frac{\pi_i^{(k)} f(\boldsymbol{y}_j; \boldsymbol{\mu}_i^{(k)}, \boldsymbol{\Sigma}_i^{(k)}, \nu_i^{(k)})}{f(\boldsymbol{y}_j; \boldsymbol{\Psi}^{(k)})} \tag{27}$$

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

$$u_{ij}^{(k)} = \frac{\nu_i^{(k)} + p}{\nu_i^{(k)} + \delta(\boldsymbol{y}_j, \boldsymbol{\mu}_i^{(k)}; \boldsymbol{\Sigma}_i^{(k)})}. \tag{28}$$

Also,

$$\psi(s) = \{\partial\Gamma(s)/\partial s\}/\Gamma(s)$$

is the Digamma function.

9. M-STEP

On the M-step at the $(k+1)$th iteration of the EM algorithm, it follows from (21) that $\boldsymbol{\pi}^{(k+1)}$, $\boldsymbol{\xi}^{(k+1)}$, and $\boldsymbol{\nu}^{(k+1)}$ can be computed independently of each other, by separate consideration of (22), (23), and (24), respectively. The solutions for $\boldsymbol{\pi}_i^{(k+1)}$ and $\boldsymbol{\theta}_i^{(k+1)}$ exist in closed form. Only the updates $\nu_i^{(k+1)}$ for the degrees of freedom ν_i need to be computed iteratively.

The mixing proportions are updated by consideration of the first term $Q_1(\boldsymbol{\pi}; \boldsymbol{\Psi}^{(k)})$ on the right-hand side of (21). This leads to $\pi_i^{(k+1)}$ being given by the average of the posterior probabilities of component membership of the mixture. That is,

$$\pi_i^{(k+1)} = \sum_{j=1}^{n} \tau_{ij}^{(k)}/n \quad (i = 1, \ldots, g). \tag{29}$$

To update the estimates of $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ $(i = 1, \ldots, g)$, we need to consider $Q_3(\boldsymbol{\theta}_i; \boldsymbol{\Psi}^{(k)})$. This is easily undertaken on noting that it corresponds to the log likelihood function formed from n independent observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ with common mean $\boldsymbol{\mu}_i$ and covariance matrices $\boldsymbol{\Sigma}_i/u_{i1}, \ldots, \boldsymbol{\Sigma}_i/u_{in}$, respectively. It is thus equivalent to computing the weighted sample mean and sample covariance matrix of $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ with weights $u_{i1}^{(k)}, \ldots, u_{in}^{(k)}$. Hence

$$\boldsymbol{\mu}_i^{(k+1)} = \sum_{j=1}^{n} \tau_{ij}^{(k)} u_{ij}^{(k)} \boldsymbol{y}_j / \sum_{j=1}^{n} \tau_{ij}^{(k)} u_{ij}^{(k)} \tag{30}$$

and

$$\boldsymbol{\Sigma}_i^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_{ij}^{(k)} u_{ij}^{(k)} (\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k+1)})(\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k+1)})^T}{\sum_{j=1}^{n} \tau_{ij}^{(k)}}. \tag{31}$$

It can be seen that it effectively chooses $\boldsymbol{\mu}_i^{(k+1)}$ and $\boldsymbol{\Sigma}_i^{(k+1)}$ by iteratively reweighted least-squares estimation. The E-step updates the weights $u_{ij}^{(k)}$, while the M-step chooses $\boldsymbol{\mu}_i^{(k+1)}$ and $\boldsymbol{\Sigma}_i^{(k+1)}$ by weighted least-squares estimation.

Following the proposal of Kent, Tyler, and Vardi (1994) in the case of a single component t distribution, we can replace the divisor $\sum_{j=1}^{n} \tau_{ij}^{(k)}$ in (31) by

$$\sum_{j=1}^{n} \tau_{ij}^{(k)} u_{ij}^{(k)}.$$

This modified algorithm, however, converges faster than the conventional EM algorithm, as reported by Kent et al. (1994) and Meng and van Dyk (1997) in the case of a single component t distribution ($g = 1$). In the latter situation, Meng and van Dyk (1997) showed that this modified EM algorithm is optimal among EM algorithms generated from a class of data augmentation schemes. More recently, in the case $g = 1$, Liu (1997) and Liu et al. (1998) have derived this modified EM algorithm using the PX-EM algorithm.

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 ML estimate. In this case, we have to compute also on the M-step the updated estimate $\nu_i^{(k+1)}$ of ν_i. On calculating the left-hand side of the equation

$$\sum_{j=1}^{n} \partial Q_{2j}(\nu_i;\, \boldsymbol{\Psi}^{(k)})/\partial \nu_i = 0,$$

it follows that $\nu_i^{(k+1)}$ is a solution of the equation

$$\begin{aligned}
&\{-\psi(\tfrac{1}{2}\nu_i) + \log(\tfrac{1}{2}\nu_i) + 1 \\
&+ \frac{1}{n_i^{(k)}} \sum_{j=1}^{n} \tau_{ij}^{(k)} (\log u_{ij}^{(k)} - u_j^{(k)}) \\
&\quad + \psi(\frac{\nu_i^{(k)} + p}{2}) - \log(\frac{\nu_i^{(k)} + p}{2})\} = 0,
\end{aligned} \tag{32}$$

where $n_i^{(k)} = \sum_{j=1}^{n} \tau_{ij}^{(k)}$.

10. APPLICATION OF ECM ALGORITHM

For ML estimation of a single t component, Liu and Rubin (1995) noted that the convergence of the EM algorithm is slow for unknown ν and the one-dimensional search for the computation of $\nu^{(k+1)}$ is time consuming. Consequently, they considered extensions of the EM algorithm in the form of the ECM and ECME algorithms; see McLachlan and Krishnan (1997, Section 5.8) and Liu (1997).

We consider now the ECM algorithm for the present problem of more than one t component, where $\boldsymbol{\Psi}$ is partitioned as $(\boldsymbol{\Psi}_1^T, \boldsymbol{\Psi}_2^T)^T$, with $\boldsymbol{\Psi}_1 = (\pi_1, \ldots, \pi_{g-1}, \boldsymbol{\xi}^T)^T$ and with $\boldsymbol{\Psi}_2$ equal to $\boldsymbol{\nu}$. On the $(k+1)$th iteration of the ECM algorithm, the E-step is the same as given above for the EM algorithm, but the M-step of the latter is replaced by two CM-steps, as follows.

CM-Step 1. Calculate $\boldsymbol{\Psi}_1^{(k+1)}$ by maximizing $Q(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k)})$ with $\boldsymbol{\Psi}_2$ fixed at $\boldsymbol{\Psi}_2^{(k)}$; that is, $\boldsymbol{\nu}$ fixed at $\boldsymbol{\nu}^{(k)}$.

CM-Step 2. Calculate $\boldsymbol{\Psi}_2^{(k+1)}$ by maximizing $Q(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k)})$ with $\boldsymbol{\Psi}_1$ fixed at $\boldsymbol{\Psi}_1^{(k+1)}$.

But as seen above, $\boldsymbol{\Psi}_1^{(k+1)}$ and $\boldsymbol{\Psi}_2^{(k+1)}$ are calculated independently of each other on the M-step, and so these two CM-steps of the ECM algorithm are equivalent to the M-step of the EM algorithm. Hence there is no difference between this ECM and the EM algorithms here. But Liu and Rubin (1995) used the ECM algorithm to give two modifications that are different from the EM algorithm. These two modifications are a multicycle version of the ECM algorithm and an ECME extension. The multicycle version of the ECM algorithm has an additional E-step between the two CM-steps. That is, after the first CM-step, the E-step is taken with

$$\boldsymbol{\Psi} = (\boldsymbol{\Psi}_1^{(k+1)^T},\, \boldsymbol{\Psi}_2^{(k)^T})^T,$$

instead of with $\boldsymbol{\Psi} = (\boldsymbol{\Psi}_1^{(k)^T},\, \boldsymbol{\Psi}_2^{(k)^T})^T$ as on the commencement of the $(k+1)$th iteration of the ECM algorithm.

11. EMMIX PROGRAM

The EMMIX program of McLachlan et al. (1999) can be used to fit normal mixtures. It also has the option for the fitting fitting of mixtures of multivariate t components either with or without the specification of the component degrees of freedom. The EMMIX program automatically provides a selection of starting values for the fitting of these mixtures if the user does not provide any. It automatically provides starting values for the application of the EM algorithm by considering a selection obtained from three sources: (a) random starts, (b) hierarchical clustering-based starts, and (c) k-means clustering-based starts. Concerning (a) for randomly selected starts, there is an additional option whereby the user can first subsample the data before using a random start based on the subsample each time. This is to limit the effect of the central limit theorem which would have the randomly selected starts being similar for each component in large samples. Concerning (b), the user has the option of using in either standardized or unstandardized form, the results from seven hierarchical methods (nearest neighbor, farthest neighbor, group average, median, centroid, flexible sorting, and Ward's method). There are several algorithm parameters that the user can optionally specify; alternatively, default values are used. The program fits the normal or t mixture model for each of the initial grouping specified from the three sources (a) to (c). All these computations are automatically carried out by the program. The user only has to provide the data set, the restrictions on the component-covariance matrices (equal, unequal, or diagonal), the extent of the selection of the initial groupings to be used to determine starting values, and the number of components that are to be fitted.

12. EXAMPLE 1

One way in which the presence of atypical observations or background noise in the data has been handled when fitting mixtures of normal components has been to include an additional component having a uniform distribution. The support of the latter component is generally specified by the upper and lower extremities of each dimension defining the rectangular region that contains all the data points. Typically, the mixing proportion for this uniform component is left unspecified to be estimated from the data, as in the analysis of the minefield data set as analyzed by Dasgupta and Raftery (1998). Also, Schroeter et al. (1998) fitted a mixture of three normal components and a uniform distribution to segment magnetic resonance images of the human brain into three regions (gray matter, white matter and cerebro-spinal fluid) in the presence of background noise arising from instrument irregularities and tissue abnormalities.

Here we consider a sample consisting initially of 100 simulated points from a two-component bivariate normal mixture model, to which 50 noise points were added from a uniform distribution over the range -10 to 10 on each variate. The parameters of the mixture model were,

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 0 & 3 \end{pmatrix}^T \boldsymbol{\mu}_2 = \begin{pmatrix} 3 & 0 \end{pmatrix}^T \boldsymbol{\mu}_3 = \begin{pmatrix} -3 & 0 \end{pmatrix}^T$$

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 2 & 0.5 \\ 0.5 & .5 \end{pmatrix} \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & 0 \\ 0 & .1 \end{pmatrix}$$

$$\boldsymbol{\Sigma}_3 = \begin{pmatrix} 2 & -0.5 \\ -0.5 & .5 \end{pmatrix}$$

with mixing proportions $\pi_1 = \pi_2 = \pi_3 = \frac{1}{3}$. The true grouping is shown in Figure **??**. We now consider the clustering obtained by fitting a mixture of three t components with unequal scale matrices but equal degrees of freedom ($\nu_1 = \nu_2 = \nu_3 = \nu$). The values of the weights $u_{ij}^{(k)}$ at convergence, $\hat{u}_{ij}$, were examined. The noise points (points 101-150) generally produced much lower $\hat{u}_{ij}$ values. In this application, an observation $\boldsymbol{y}_j$ is treated as an outlier (background noise) if $\sum_{i=1}^{g} \hat{z}_{ij}\hat{u}_{ij}$ is sufficiently small, or equivalently,

$$\sum_{i=1}^{g} \hat{z}_{ij}\delta(\boldsymbol{y}_j,\, \hat{\boldsymbol{\mu}}_i;\, \hat{\boldsymbol{\Sigma}}_i) \qquad (33)$$

is sufficiently large, where

$$\hat{z}_{ij} = \arg\max_{h} \hat{\tau}_{hj} \quad (i = 1, \ldots, g;\, j = 1, \ldots, n),$$

and $\hat{\tau}_{ij}$ denotes the estimated posterior probability that $\boldsymbol{y}_j$ belongs to the ith component of the mixture.

To decide on how large the statistic (33) must be in order for $\boldsymbol{y}_j$ to be classified as noise, we compared it to the 95th percentile of the chi-squared distribution with p degrees of freedom, where the

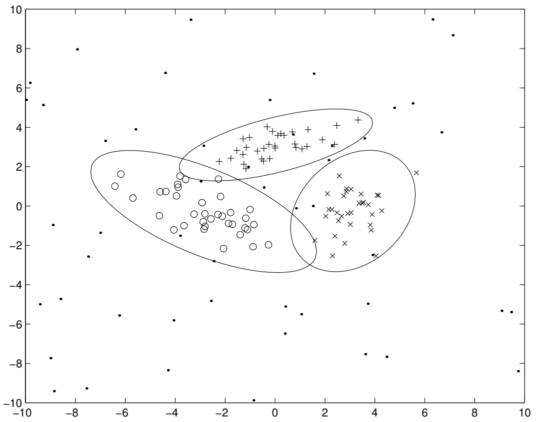

Figure 1: Plot of the true grouping of the simulated noisy data set

latter is used to approximate the true distribution of $\delta(\boldsymbol{Y}_j, \hat{\boldsymbol{\mu}}_i; \hat{\boldsymbol{\Sigma}}_i)$.

The clustering so obtained is displayed in Figure 2. It compares well with the true grouping in Figure 3 and the clustering in Figure 3 obtained by fitting a mixture of three normal components and an additional uniform component. In this particular example the model of three normal components with an additional uniform component to model the noise works well since it is the same model used to generate the data in the first instance. However, this model, unlike the t mixture model, cannot be expected to work as well in situations when the noise is not uniform or is unable to be modeled adequately by the uniform distribution.

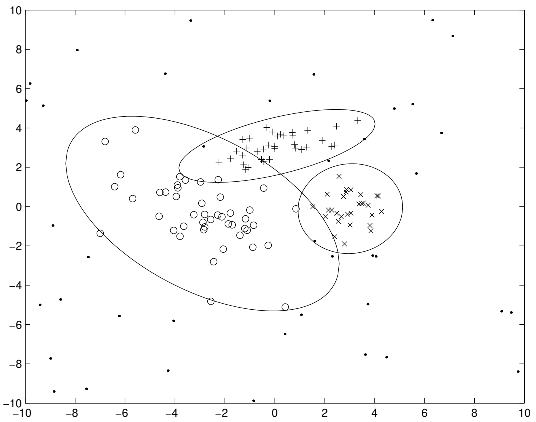

Figure 2: Plot of the result of fitting a mixture of t distributions (with a classification of noise at a significance level of 5%) to the simulated noisy data

It is of interest to note that if the number of groups is treated as unknown and a normal mixture fitted, then the number of groups selected via AIC, BIC, and bootstrapping of $-2 \log \lambda$ is four for each of these criteria. The result of fitting a mixture of four normal components is displayed in Figure 4. Obviously, the additional fourth component is attempting to model the background noise. However, it can be seen from Figure 4 that this normal mixture-based clustering is still affected by the noise.

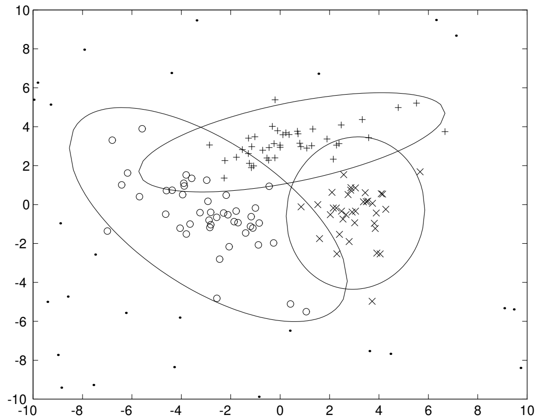

Figure 3: Plot of the result of fitting a three-component normal mixture plus a uniform component model to the simulated noisy data

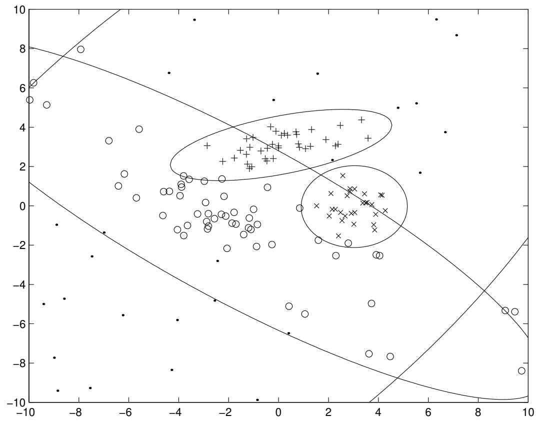

Figure 4: Plot of the result of fitting a four-component normal mixture model to the simulated noisy data

13. EXAMPLE 2

As another demonstration of the use of the t mixture approach to a robust cluster analysis, we now consider the Old Faithful geyser data set consisting of 272 univariate measurements of eruption lengths of the Old Faithful geyser. The data set is originally from Silverman (1986). However in this example, we consider a bivariate version analyzed in García-

Escudero and Gordaliza (1999), where the data have been lagged.

Figure 5 shows the clustering obtained by fitting a mixture of three t components with unequal scale matrices, but equal degrees of freedom ($\nu_1 = \nu_2 = \nu_3 = \nu$). The values of the weights $u_{ij}^{(k)}$ at convergence, $\hat{u}_{ij}$, were examined in a similar manner to that in the previous section to classify points as outliers. The outliers determined at the 1% level are indicated by dots.

Figure 6 gives the corresponding results when the restriction of equal component-scale matrices is applied. These last results are in general agreement with those obtained by García-Escudero and Gordaliza (1999) who provided a robust clustering of these data using a trimmed k-means procedure. The latter assumes also that the equal component-covariance matrices are spherical. However, this would appear to be a reasonable assumption for this data set. The solutions differ on the conclusion of whether two of the observations (labeled A and B in Figure 6) are atypical; the latter are considered atypical on the basis of the t mixture model.

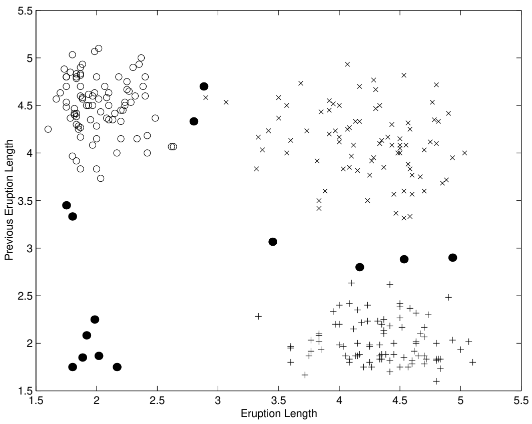

Figure 5: Result of fitting a three component t mixture model with equal ν_i and unequal scale matrices to the geyser data set; dots denote those a observations assessed as atypical at the 1% level

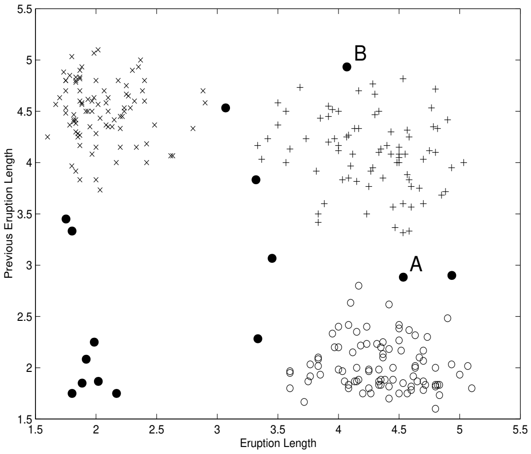

Figure 6: Result of fitting a three component t mixture model with equal ν_i and equal scale matrices to the geyser data set; dots denote those observations assessed as atypical at the 1% level

REFERENCES

Banfield, J. and Raftery, A.E. (1991). Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49**, 803–821.

Bradley, P.S., Fayyad, U.M., and Reina, C.A. (1999). Scaling EM (expectation-maximization) clustering to large databases. *Technical Report No. MSR-TR-98-35*. Seattle: Microsoft Research.

Campbell, N.A. (1984). Mixture models and atypical values. *Mathematical Geology* **16**, 465–477.

Dasgupta, A. and Raftery, A.E. (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association* **93**, 294–302.

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

De Veaux, R.D. and Kreiger, A.M. (1990). Robust estimation of a normal mixture. *Statistics & Probability Letters* **10**, 1–7.

Diebolt, J. and Robert, C.P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society B* **56**, 363–375.

Fraley, C. and Raftery, A.E. (1999). MCLUST: Software for model-based cluster analysis. *Journal of Classification* **16**, 297-306.

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

Kent, J.T., Tyler, D.E., and Vardi, Y. (1994). A curious likelihood identity for the multivariate t-distribution. *Communications in Statistics - Simulation and Computation* **23**, 441–453.

Lange, K., Little, R.J.A., and Taylor, J.M.G. (1989). Robust statistical modeling using the t distribution. *Journal of the American Statistical Association* **84**, 881–896.

Lavine, M. and West, M. (1992). A Bayesian method of classification and discrimination. *Canadian Journal of Statistics* **20**, 451–461.

Lindsay, B.G. (1995). *Mixture Models: Theory, Geometry and Applications, NSF-CBMS Regional Conference Series in Probability and Statistics*, Vol. 5. Alexandria, VA: Institute of Mathematical Statistics and the American Statistical Association.

Liu, C. (1997). ML estimation of the multivariate t distribution and the EM algorithm. *Journal of Multivariate Analysis* **63**, 296–312.

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

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

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

Markatou, M. (1998). Mixture models, robustness and the weighted likelihood methodology. *Technical Report No. 1998-9*. Stanford: Department of Statistics, Stanford University.

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

McLachlan, G.J. and Ng, S.K. (2000). A sparse version of the incremental EM algorithm for large databases. *Technical Report*. Brisbane: Department of Mathematics, University of Queensland.

McLachlan, G.J. and Peel, D. (1998). Robust cluster analysis via mixtures of multivariate t distributions. In *Lecture Notes in Computer Science* Vol. 1451, A. Amin, D. Dori, P. Pudil, and H. Freeman (Eds.). Berlin: Springer-Verlag, pp. 658–666.

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

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

McLachlan, G.J., Peel, D., Basford, K.E., and Adams, P. (1999). Fitting of mixtures of normal and t components. *Journal of Statistical Software* **4**,*No.* 2.

Meng, X.L. and van Dyk, D. (1995). The EM algorithm — an old folk song sung to a fast new tune (with discussion). *Journal of the Royal Statistical Society B* **59**, 511–567.

Moore, A.W. (1999). Very fast EM-based mixture model clustering using multiresolution kd-trees. In *Advances in Neural Information Processing Systems 11*, M.S. Kearns, S.A. Solla, and D.A. Cohn (Eds.). Cambridge, Massachusetts: MIT Press, pp. 543–549.

Neal, R.M. and Hinton, G.E. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*, M.I. Jordan (Ed.). Dordrecht: Kluwer, pp. 355–368.

Peel, D. and McLachlan. G.J. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing* (to appear).

Rocke, D.M. and Woodruff, D.L. (1997). Robust estimation of multivariate location and shape. *Journal of Statistical Planning and Inference* **57**, 245–255.

Schroeter, P., Vesin, J.-M., Langenberger, T., and Meuli, R. (1998). Robust parameter estimation of intensity distributions for brain magnetic resonance images. *IEEE Transactions on Medical Imaging* **17** 172-186.

Thiesson, B., Meek, C., and Heckerman, D. (2000). Accelerating EM for large databases. *Technical Report No. MSR-TR-99-31* (revised version). Seattle: Microsoft Research.

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