Statistics and Computing (2000) **10,** 339–348

Robust mixture modelling using the t distribution

D. PEEL and G. J. McLACHLAN

Department of Mathematics, University of Queensland, St. Lucia, Queensland 4072, Australia
gjm@maths.uq.edu.au

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. In this paper, we consider a more robust approach by modelling the data by a mixture of t distributions. The use of the ECM algorithm to fit this t mixture model is described and examples of its use are given in the context of clustering multivariate data in the presence of atypical observations in the form of background noise.

Keywords: finite mixture models, normal components, multivariate t components, maximum likelihood, EM algorithm, cluster analysis

1. Introduction

Finite mixtures of distributions have provided a mathematical-based approach to the statistical modelling of a wide variety of random phenomena; see, for example, Everitt and Hand (1981), Titterington, Smith and Makov (1985), McLachlan and Basford (1988), Lindsay (1995), and Böhning (1999). Because of their usefulness as an extremely flexible method of modelling, 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 focussed 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).

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 with the difficulty of the dimension of the data (Rocke and Woodruff (1997) and Kosinski (1999)).

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). In their conference paper, they reported briefly on robust clustering via mixtures of t components, but did not include details of the implementation of the EM algorithm nor the examples to be given here.

With this t mixture model-based approach, the normal distribution for each component in the mixture is embedded in a wider class of elliptically symmetric distributions with an additional parameter called the degrees of freedom ν. As ν tends to infinity, the t distribution approaches the normal distribution. Hence this parameter ν 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 thereby providing an *adaptive* robust procedure, as explained in Lange, Little and Taylor (1989), who considered the use of a single component t distribution in linear and nonlinear regression problems; see also Rubin (1983) and Sutradhar and Ali (1986).

In the past, there have been many attempts at modifying existing methods of cluster analysis to provide robust clustering procedures. Some of these have been of a rather *ad hoc*

nature. The use of a mixture model of t distributions provides a sound mathematical basis for a robust method of mixture estimation and hence clustering. We shall illustrate its usefulness in the latter context by a cluster analysis of a simulated data set with background noise added and of an actual data set.

2. Previous work

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. This line of approach is to be discussed in Section 9.

Recently, Markatou (1998) has provided a formal approach to robust mixture estimation by applying weighted likelihood methodology in the context of mixture models. With this methodology, an estimate of the vector of unknown parameters is obtained as a solution of the equation

$$\sum_{j=1}^{n} w(\boldsymbol{y}_j)\partial \log f(\boldsymbol{y}_j;\ \boldsymbol{\Psi})/\partial \boldsymbol{\Psi} = \mathbf{0}, \tag{1}$$

where $f(\boldsymbol{y};\boldsymbol{\Psi})$ denotes the specified parametric form for the probability density function (p.d.f.) of the random vector $\boldsymbol{Y}$ on which $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ have been observed independently. The weight function $w(\boldsymbol{y})$ is defined in terms of the Pearson residuals; see Markatou, Basu and Lindsay (1998) and the previous work of Green (1984). The weighted likelihood methodology provides robust and first-order efficient estimators, and Markatou (1998) has established these results in the context of univariate mixture models.

One useful application of normal mixture models has been in the important field of cluster analysis. Besides having a sound mathematical basis, this approach is not confined to the production of spherical clusters, such as with k-means-type algorithms that use Euclidean distance rather than the Mahalanobis distance metric which allows for within-cluster correlations between the variables in the feature vector $\boldsymbol{Y}$. Moreover, unlike clustering methods defined solely in terms of the Mahalanobis distance, the normal mixture-based clustering takes into account the normalizing term $|\boldsymbol{\Sigma}_i|^{-1/2}$ in the estimate of the multivariate normal density adopted for the component distribution of $\boldsymbol{Y}$ corresponding to the ith cluster. This term can make an important contribution in the case of disparate group-covariance matrices (McLachlan 1992, Chapter 2).

Although even a crude estimate of the within-cluster covariance matrix $\boldsymbol{\Sigma}_i$ often suffices for clustering purposes (Gnanadesikan, Harvey and Kettenring 1993), it can be severely affected by outliers. Hence it is highly desirable for methods of cluster analysis to provide robust clustering procedures. The problem of making clustering algorithms more robust has received much attention recently as, for example, in Smith, Bailey and Munford (1995), Davé and Krishnapuram (1996), Frigui and Krishnapuram (1996), Jolion, Meer and Bataouche (1996), Kharin (1996), Rousseeuw, Kaufman and Trauwaert (1996) and Zhuang *et al.* (1996).

3. Multivariate *t* distribution

We let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ denote an observed p-dimensional random sample of size n. With a normal mixture model-based approach to drawing inferences from these data, each data point is assumed to be a realization of the random p-dimensional vector $\boldsymbol{Y}$ with the g-component normal mixture probability density function (p.d.f.),

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

where the mixing proportions π_i are nonnegative and sum to one and where

$$\phi(\boldsymbol{y};\ \boldsymbol{\mu}_i,\ \boldsymbol{\Sigma}_i) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}_i|^{-\frac{1}{2}} \times \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_i)\right\} \tag{2}$$

denotes the p-variate multivariate normal p.d.f. with mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$ $(i = 1, \ldots, g)$. Here $\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_{g-1}, \boldsymbol{\theta}^T)^T$, where $\boldsymbol{\theta}$ consists of the elements of the $\boldsymbol{\mu}_i$ and the distinct elements of the $\boldsymbol{\Sigma}_i$ $(i = 1, \ldots, g)$. In the above and sequel, we are using f as a generic symbol for a p.d.f.

One way to broaden this parametric family for potential outliers or data with longer than normal tails is to adopt the two-component normal mixture p.d.f.

$$(1-\epsilon)\phi(\boldsymbol{y};\ \boldsymbol{\mu},\ \boldsymbol{\Sigma}) + \epsilon\phi(\boldsymbol{y};\ \boldsymbol{\mu},\ c\boldsymbol{\Sigma}), \tag{3}$$

where c is large and ϵ is small, representing the small proportion of observations that have a relatively large variance. Huber (1964) subsequently considered more general forms of contamination of the normal distribution in the development of his robust M-estimators of a location parameter, as discussed further in Section 9.

The normal scale mixture model (3) can be written as

$$\int \phi(\boldsymbol{y};\boldsymbol{\mu};\boldsymbol{\Sigma}/u)\, dH(u), \tag{4}$$

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; that is, by the random variable U distributed as

$$U \sim \text{gamma}\left(\frac{1}{2}\nu, \frac{1}{2}\nu\right),$$

where the gamma $(\alpha,\ \beta)$ density function $f(u;\ \alpha,\ \beta)$ is given by

$$f(u;\alpha,\ \beta) = \{\beta^\alpha u^{\alpha-1}/\,\Gamma(\alpha)\}\ \exp(-\beta u) I_{(0,\infty)}(u); \quad (\alpha, \beta > 0),$$

and the indicator function $I_{(0,\infty)}(u) = 1$ for $u > 0$ and is zero elsewhere. We then obtain the t distribution with location

parameter $\boldsymbol{\mu}$, positive definite inner product matrix $\boldsymbol{\Sigma}$, and ν degrees of freedom,

$$f(\boldsymbol{y};\ \boldsymbol{\mu},\ \boldsymbol{\Sigma},\ \nu) = \frac{\Gamma\left(\frac{\nu+p}{2}\right)|\boldsymbol{\Sigma}|^{-1/2}}{(\pi\nu)^{\frac{1}{2}p}\Gamma\left(\frac{\nu}{2}\right)\{1+\delta(\boldsymbol{y},\ \boldsymbol{\mu};\ \boldsymbol{\Sigma})/\nu\}^{\frac{1}{2}(\nu+p)}}, \tag{5}$$

where

$$\delta(\boldsymbol{y},\ \boldsymbol{\mu};\ \boldsymbol{\Sigma}) = (\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}) \tag{6}$$

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 $\nu > 2$).

4. 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).

5. ML estimation of mixture of *t* distributions

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

$$f(\boldsymbol{y};\ \boldsymbol{\Psi}) = \sum_{i=1}^{g}\pi_i f(\boldsymbol{y};\ \boldsymbol{\mu}_i,\ \boldsymbol{\Sigma}_i,\ \nu_i), \tag{7}$$

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 $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T,\ldots,\boldsymbol{\theta}_g^T)^T$, 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

$$\boldsymbol{x}_{\mathrm{c}} = \left(\boldsymbol{x}_{\mathrm{o}}^T,\ \boldsymbol{z}_1^T,\ldots,\boldsymbol{z}_n^T,\ u_1,\ldots,u_n\right)^T \tag{8}$$

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 $z_{ij} = 1$,

$$\boldsymbol{Y}_j \mid u_j,\ z_{ij} = 1 \sim N(\boldsymbol{\mu}_i,\ \boldsymbol{\Sigma}_i/u_j), \tag{9}$$

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

$$U_j \mid z_{ij} = 1 \sim \text{gamma}\left(\frac{1}{2}\nu_i,\ \frac{1}{2}\nu_i\right). \tag{10}$$

Given $\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n$, the $U_1,\ldots,U_n$ are independently distributed according to (10).

The complete-data likelihood $L_{\mathrm{c}}(\boldsymbol{\Psi})$ 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 $\boldsymbol{z}_j$. Accordingly, the complete-data log likelihood can be written as

$$\log L_{\mathrm{c}}(\boldsymbol{\Psi}) = \log L_{1\mathrm{c}}(\boldsymbol{\pi}) + \log L_{2\mathrm{c}}(\boldsymbol{\nu}) + \log L_{3\mathrm{c}}(\boldsymbol{\theta}), \tag{11}$$

where

$$\log L_{1\mathrm{c}}(\boldsymbol{\pi}) = \sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij}\ \log\pi_i, \tag{12}$$

$$\log L_{2\mathrm{c}}(\boldsymbol{\nu}) = \sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij}\left\{-\log\Gamma\left(\frac{1}{2}\nu_i\right) + \frac{1}{2}\nu_i\ \log\left(\frac{1}{2}\nu_i\right) + \frac{1}{2}\nu_i(\log u_j - u_j) - \log u_j\right\}, \tag{13}$$

and

$$\log L_{3\mathrm{c}}(\boldsymbol{\theta}) = \sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij}\left\{-\frac{1}{2}p\ \log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}u_j(\boldsymbol{y}_j-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_j-\boldsymbol{\mu}_i)\right\}. \tag{14}$$

In (11), $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_g)^T$.

6. E-step

Now 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 can be seen from (12) to (14) that in order to do this, we need to calculate

$$E_{\boldsymbol{\Psi}^{(k)}}(Z_{ij} \mid \boldsymbol{y}_j),$$
$$E_{\boldsymbol{\Psi}^{(k)}}(U_j \mid \boldsymbol{y}_j,\ \boldsymbol{z}_j),$$

and

$$E_{\boldsymbol{\Psi}^{(k)}}(\log U_j \mid \boldsymbol{y}_j,\ \boldsymbol{z}_j) \tag{15}$$

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

It follows that

$$E_{\boldsymbol{\Psi}^{(k)}}(Z_{ij} \mid \boldsymbol{y}_j) = \tau_{ij}^{(k)}$$

where

$$\tau_{ij}^{(k)} = \frac{\pi_i^{(k)} f\big(\boldsymbol{y}_j;\ \boldsymbol{\mu}_i^{(k+1)},\ \boldsymbol{\Sigma}_i^{(k+1)},\ \nu_i^{(k+1)}\big)}{f\big(\boldsymbol{y}_j;\ \boldsymbol{\Psi}^{(k+1)}\big)}, \tag{16}$$

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

Since the gamma distribution is the conjugate prior distribution for U_j, it is not difficult to show that the conditional distribution of U_j given $\boldsymbol{Y}_j = \boldsymbol{y}_j$ and $Z_{ij} = 1$ is

$$U \mid \boldsymbol{y}_j,\ z_{ij} = 1 \sim \text{ gamma}\,(m_{1i},\ m_{2i}), \tag{17}$$

where

$$m_{1i} = \frac{1}{2}(\nu_i + p)$$

and

$$m_{2i} = \frac{1}{2}\{\nu_i + \delta(\boldsymbol{y}_j,\ \boldsymbol{\mu}_i;\ \boldsymbol{\Sigma}_i)\}. \tag{18}$$

From (17), we have that

$$E(U_j \mid \boldsymbol{y}_j,\ z_{ij} = 1) = \frac{\nu_i + p}{\nu_i + \delta(\boldsymbol{y}_j,\ \boldsymbol{\mu}_i;\ \boldsymbol{\Sigma}_i)}. \tag{19}$$

Thus from (19),

$$E_{\boldsymbol{\Psi}^{(k)}}(U_j \mid \boldsymbol{y}_j, z_{ij} = 1) = u_{ij}^{(k)},$$

where

$$u_{ij}^{(k)} = \frac{\nu_i^k + p}{\nu_i^k + \delta\big(\boldsymbol{y}_j,\ \boldsymbol{\mu}_i^{(k)};\ \boldsymbol{\Sigma}_i^{(k)}\big)}. \tag{20}$$

To calculate the conditional expectation (15), we need the result that if a random variable R has a gamma (α, β) distribution, then

$$E(\log R) = \psi(\alpha) - \log\beta, \tag{21}$$

where

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

is the Digamma function. Applying the result (21) to the conditional density of U_j given $\boldsymbol{y}_j$ and $z_{ij} = 1$, as specified by (10), it follows that

$$\begin{aligned} &E_{\boldsymbol{\Psi}^{(k)}}(\log U_j \mid \boldsymbol{y}_j,\ z_{ij} = 1) \\ &\quad = \psi\left(\frac{\nu_i^{(k)} + p}{2}\right) - \log\left[\frac{1}{2}\left\{\nu_i^{(k)} + \delta\big(\boldsymbol{y}_j,\ \boldsymbol{\mu}_i^{(k)};\ \boldsymbol{\Sigma}_i^{(k)}\big)\right\}\right] \\ &\quad = \log u_{ij}^{(k)} + \left\{\psi\left(\frac{\nu_i^{(k)} + p}{2}\right) - \log\left(\frac{\nu_i^{(k)} + p}{2}\right)\right\} \end{aligned} \tag{22}$$

for $j = 1, \ldots, n$. The last term on the right-hand side of (22),

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

can be interpreted as the correction for just imputing the conditional mean value $u_{ij}^{(k)}$ for u_j in $\log u_j$.

On using the results (16), (19) and (22) to calculate the conditional expectation of the complete-data log likelihood from (11), we have that $Q(\boldsymbol{\Psi};\ \boldsymbol{\Psi}^{(k)})$ is given by

$$Q\big(\boldsymbol{\Psi};\ \boldsymbol{\Psi}^{(k)}\big) = Q_1\big(\boldsymbol{\pi};\ \boldsymbol{\Psi}^{(k)}\big) + Q_2\big(\boldsymbol{\nu};\ \boldsymbol{\Psi}^{(k)}\big) + + Q_3\big(\boldsymbol{\theta};\ \boldsymbol{\Psi}^{(k)}\big), \tag{23}$$

where

$$Q_1\big(\boldsymbol{\pi};\ \boldsymbol{\Psi}^{(k)}\big) = \sum_{i=1}^{g}\sum_{j=1}^{n} \hat{\tau}_{ij}^{(k)}\ \log\pi_i, \tag{24}$$

$$Q_2\big(\boldsymbol{\nu};\ \boldsymbol{\Psi}^{(k)}\big) = \sum_{i=1}^{g}\sum_{j=1}^{n} \hat{\tau}_{ij}^{(k)} Q_{2j}\big(\nu_i;\ \boldsymbol{\Psi}^{(k)}\big), \tag{25}$$

and

$$Q_3\big(\boldsymbol{\theta};\ \boldsymbol{\Psi}^{(k)}\big) = \sum_{i=1}^{g}\sum_{j=1}^{n} \hat{\tau}_{ij}^{(k)} Q_{3j}\big(\boldsymbol{\theta}_i;\ \boldsymbol{\Psi}^{(k)}\big), \tag{26}$$

and where, on ignoring terms not involving the ν_i,

$$\begin{aligned} Q_{2j}\big(\nu_i;\ \boldsymbol{\Psi}^{(k)}\big) = &-\log\Gamma\left(\frac{1}{2}\nu_i\right) + \frac{1}{2}\nu_i\ \log\left(\frac{1}{2}\nu_i\right) \\ &+ \frac{1}{2}\nu_i\left\{\sum_{j=1}^{n}\left(\log u_{ij}^{(k)} - u_{ij}^{(k)}\right)\right. \\ &\left. + \psi\left(\frac{\nu_i^{(k)} + p}{2}\right) - \log\left(\frac{\nu_i^{(k)} + p}{2}\right)\right\}, \end{aligned} \tag{27}$$

and

$$Q_{3j}(\boldsymbol{\theta}_i;\, \boldsymbol{\Psi}^{(k)}) = \Big\{ -\frac{1}{2}p\ \log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_i| + \frac{1}{2}p\ \log u_{ij}^{(k)} - \frac{1}{2}u_{ij}(\boldsymbol{y}_j - \boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_j - \boldsymbol{\mu}_i)\Big\}. \quad (28)$$

7. M-step

On the M-step at the $(k+1)$th iteration of the EM algorithm, it follows from (23) that $\boldsymbol{\pi}^{(k+1)}$, $\boldsymbol{\theta}^{(k+1)}$, and $\boldsymbol{\nu}^{(k+1)}$ can be computed independently of each other, by separate consideration of (24), (25), and (26), respectively. The solutions for $\boldsymbol{\pi}_i^{(k+1)}$ and $\boldsymbol{\theta}^{(k+1)}$ 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 (23). This leads to $\pi_i^{(k+1)}$ being given by the average of the posterior probabilities of component membership of the mixture. That is,

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

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 $\boldsymbol{\Sigma}_i/u_1^k, \ldots, \boldsymbol{\Sigma}_i/u_n^k$, 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 $u_1^{(k)}, \ldots, u_n^{(k)}$. Hence

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

and

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

It can be seen that it effectively chooses $\boldsymbol{\mu}_i^{(k+1)}$ and $\boldsymbol{\Sigma}_i^{(k+1)}$ by weighted least-squares estimation. The E-step updates the weights $u_{ij}^{(k)}$, while the M-step effectively chooses $\boldsymbol{\mu}_i^{(k+1)}$ and $\boldsymbol{\Sigma}_i^{(k+1)}$ by weighted least-squares estimation. It can be seen from the form of the equation (30) derived for the MLE of $\boldsymbol{\mu}_i$ that, as $\nu_i^{(k)}$ decreases, the degree of downweighting of an outlier increases. For finite $\nu_i^{(k)}$ as $\|\boldsymbol{y}_j\| \to \infty$, the effect on the ith component location parameter estimate goes to zero, whereas the effect on the ith component scale estimate remains bounded but does not vanish.

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, Tyler and Vardi (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, Rubin and Wu (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; see Lange, Little and Taylor (1989, Page 882). 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

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

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

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

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

8. 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 the ECM algorithm for this problem, where $\boldsymbol{\Psi}$ is partitioned as $(\boldsymbol{\Psi}_1^T,\ \boldsymbol{\Psi}_2)^T$, with $\boldsymbol{\Psi}_1 = (\pi_1, \ldots, \pi_{g-1},\ \boldsymbol{\theta}^T)^T$ 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

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

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. The EMMIX algorithm of McLachlan *et al.* (1999) has an option for the fitting of mixtures of multivariate t components either with or without the specification of the component degrees of freedom. It is available from the software archive StatLib or from the first author's homepage with website address `http://www.maths.uq.edu.au/~gjm/`.

For a single component t distribution, Liu and Rubin (1994, 1995), Kowalski *et al.* (1997), Liu (1997), Meng and van Dyk (1997), and Liu, Rubin and Wu (1998) have considered further extensions of the ECM algorithm corresponding to various versions of the ECME algorithm. However, the implementation of the ECME algorithm for mixtures of t distributions is not as straightforward, and so it is not applied here.

9. Previous work on M-estimation of mixture components

A common way in which robust fitting of normal mixture models has been undertaken, is by using M-estimates to update the component estimates on the M-step of the EM algorithm, as in Campbell (1984) and McLachlan and Basford (1988). In this case, the updated component means $\boldsymbol{\mu}_i^{(k+1)}$ are given by (30), but where now the weights $u_{ij}^{(k)}$ are defined as

$$u_{ij}^{(k)} = \psi\left(d_{ij}^{(k)}\right) \Big/ d_{ij}^{(k)}, \tag{33}$$

where

$$d_{ij}^{(k)} = \left\{\left(\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k)}\right)^T \boldsymbol{\Sigma}_i^{(k)^{-1}} \left(\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k)}\right)\right\}^{1/2}$$

and $\psi(s) = -\psi(-s)$ is Huber's (1964) ψ-function defined as

$$\begin{aligned} \psi(s) &= s, && |s| \le a, \\ &= \text{sign}(s)a, && |s| > a, \end{aligned} \tag{34}$$

for an appropriate choice of the tuning constant a. The ith component-covariance matrix $\boldsymbol{\Sigma}_i^{(k+1)}$ can be updated as (31), where $u_{ij}^{(k)}$ is replaced by $\{\psi(d_{ij}^{(k)})/d_{ij}^{(k)}\}^2$. An alternative to Huber's ψ-function is a redescending ψ-function, for example, Hampel's (1973) piecewise linear function. However, there can be problems in forming the posterior probabilities of component membership, as there is the question as to which parametric family to use for the component p.d.f.'s (McLachlan and Basford 1988, Section 2.8). One possibility is to use the form of the p.d.f. corresponding to the ψ-function adopted. However, in the case of any redescending ψ-function with finite rejection points, there is no corresponding p.d.f. In Campbell (1984), the normal p.d.f. was used, while in the related univariate work in De Veaux and Kreiger (1990), the t density with three degrees of freedom was used, with the location and scale component parameters estimated by the (weighted) median and mean absolute deviation, respectively.

It can be therefore seen that the use of mixtures of t distributions provides a sound statistical basis for formalizing and implementing the somewhat *ad hoc* approaches that have been proposed in the past. It also provides a framework for assessing the degree of robustness to be incorporated into the fitting of the mixture model through the specification or estimation of the degrees of freedom ν_i in the t component p.d.f.'s.

As noted in the introduction, the use of t components in place of the normal components will generally give less extreme estimates of the posterior probabilities of component membership of the mixture model. The use of the t distribution in place of the normal distribution leading to less extreme posterior probabilities of group membership was noted in a discriminant analysis context, where the group-conditional densities correspond to the component densities of the mixture model (Aitchison and Dunsmore 1975, Chapter 2). If a Bayesian approach is adopted and the conventional improper or vague prior specified for the mean and the inverse of the covariance matrix in the normal distribution for each group-conditional density, it leads to the so-called predictive density estimate, which has the form of the t distribution; see McLachlan (1992, Section 3.5).

10. Example 1: Simulated noisy data set

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. For example, 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 cerbro-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,

$$\boldsymbol{\mu}_1 = (0 \quad 3)^T \quad \boldsymbol{\mu}_2 = (3 \quad 0)^T \quad \boldsymbol{\mu}_3 = (-3 \quad 0)^T$$

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 2 & 0.5 \\ 0.5 & .5 \end{pmatrix} \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & 0 \\ 0 & .1 \end{pmatrix} \quad \boldsymbol{\Sigma}_3 = \begin{pmatrix} 2 & -0.5 \\ -0.5 & .5 \end{pmatrix}$$

with mixing proportions $\pi_1 = \pi_2 = \pi_3 = \frac{1}{3}$. The true grouping is shown in Fig. 1. 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,

$$\sum_{i=1}^{g} \hat{z}_{ij}\delta(\boldsymbol{y}_j,\ \hat{\boldsymbol{\mu}}_i;\ \hat{\boldsymbol{\Sigma}}_i) \tag{35}$$

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.

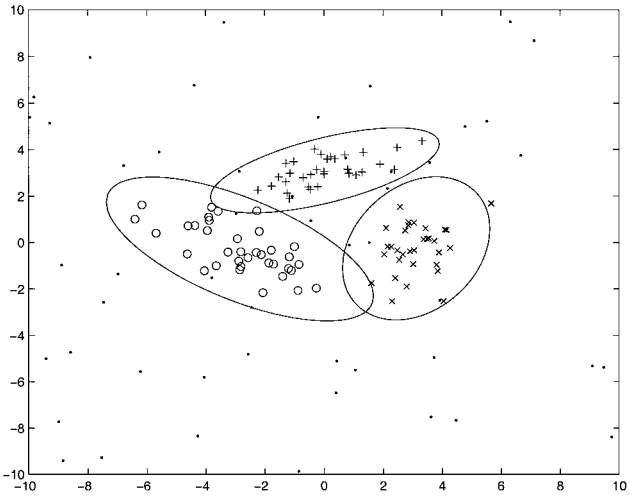

Fig. 1. *Plot the true grouping of the simulated noisy data set*

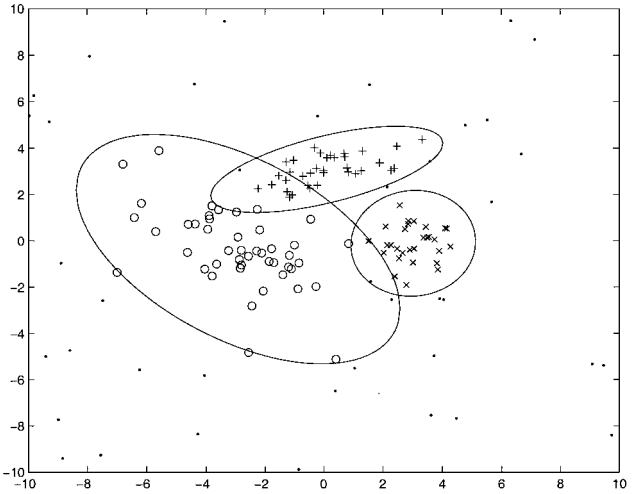

Fig. 2. *Plot of the result of fitting a mixture of t distributions with a classification of noise at a significance level of 5% to the simulated noisy data*

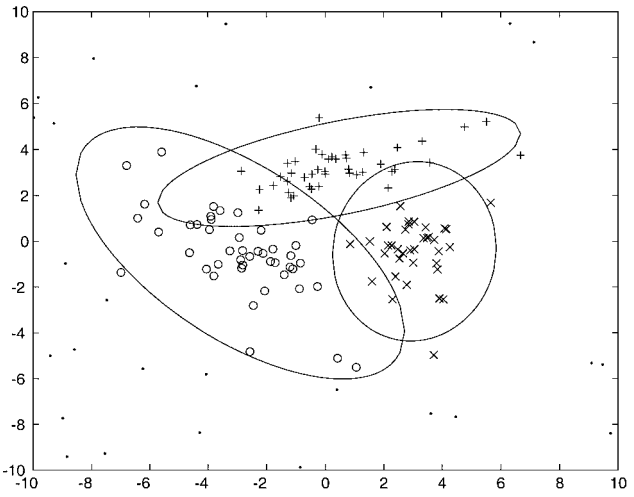

Fig. 3. *Plot of the result of fitting a three component normal mixture plus a uniform component model to the simulated noisy data*

To decide on how large the statistic (35) 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 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 Fig. 2. It compares well with the true grouping in Fig. 1 and the clustering in Fig. 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 modelled adequately by the uniform distribution.

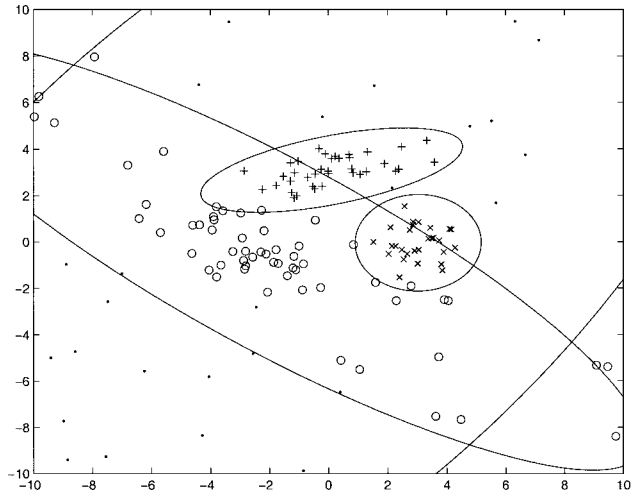

Fig. 4. *Plot of the result of fitting a four component normal mixture model to the simulated noisy data*

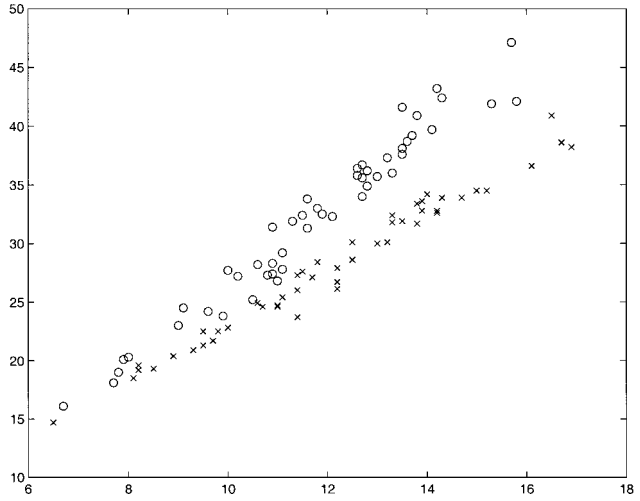

Fig. 5. *Scatter plot of the second and third variates of the Blue Crab data set with their true group of origin noted*

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 Fig. 4. Obviously, the additional fourth component is attempting to model the background noise. However, it can be seen from Fig. 4 that this normal mixture-based clustering is still affected by the noise.

11. Example 2: Blue crab data set

To further illustrate the t mixture model-based approach to clustering, we consider now the crab data set of Campbell and Mahon (1974) on the genus *Leptograpsus*. Attention is focussed on the sample of $n = 100$ blue crabs, there being $n_1 = 50$ males and $n_2 = 50$ females, which we shall refer to as groups G_1 and G_2 respectively. Each specimen has measurements on the width of the frontal lip *FL*, the rear width *RW*, and length along the midline *CL* and the maximum width *CW* of the carapace, and the body depth *BD* in mm. In Fig. 5, we give the scatter plot of the second and third variates with their group of origin noted. Hawkins' (1981) simultaneous test for multivariate normality and equal covariance matrices (homoscedasticity) suggests it is reasonable to assume that the group-conditional distributions are normal with a common covariance matrix. Consistent with this, fitting a mixture of two t components (with equal scale matrices and equal degrees of freedoms) gives only a slightly improved outright clustering over that obtained using a mixture of two normal homoscedastic components. The t mixture model-based clustering results in one cluster containing 32 observations from G_1 and another containing all 50 observations from G_2, along with the remaining 18 observations from G_1; the normal mixture model leads to one additional member of G_1 being assigned to the cluster corresponding to G_2. We note in passing that, although the groups are homoscedastic, a much improved clustering is obtained without restrictions on the scale matrices, with the t and normal mixture model-based clusterings both resulting in 17 fewer misallocations.

In this example, where the normal model for the components appears to be a reasonable assumption, the estimated degrees of freedom for the t components should be large, which they are. The estimate of their common value ν in the case of equal scale matrices and equal degrees of freedom ($\nu_1 = \nu_2 = \nu$), is $\hat{\nu} = 22.5$; the estimates of ν_1 and ν_2 in the case of unequal scale matrices and unequal degrees of freedom are $\hat{\nu}_1 = 23.0$ and $\hat{\nu}_2 = 120.3$.

The likelihood function can be fairly flat near the maximum likelihood estimates of the degrees of freedom of the t components. To illustrate this, we have plotted in Fig. 6 the profile likelihood function in the case of equal scale matrices and equal

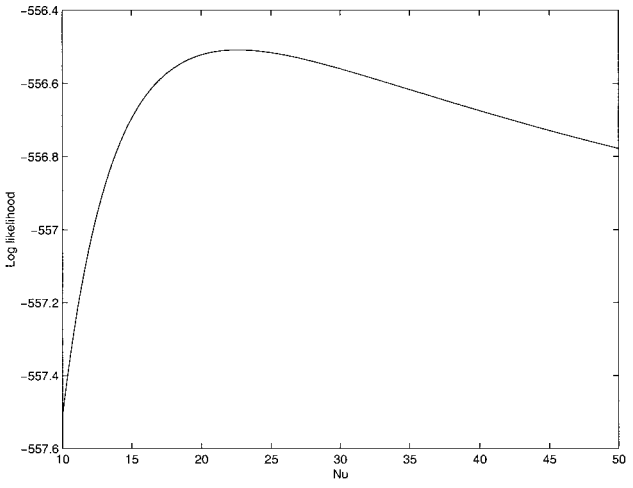

Fig. 6. *Plot of the profile log likelihood for various values of ν for the Blue Crab data set with equal scale matrices and equal degrees of freedom ($\nu_1 = \nu_2 = \nu$)*

Table 1. *Summary of comparison of error rates when fitting normal and t-distributions with the modification of a single point*

Constant	Normal error rate	t component error rate	$\hat{\nu}$	$\hat{u}_{1,25}$	$\hat{u}_{2,25}$
−15	49	19	5.76	.0154	.0118
−10	49	19	6.65	.0395	.0265
−5	21	20	13.11	.1721	.3640
0	19	18	23.05	.8298	1.1394
5	21	20	13.11	.1721	.3640
10	50	20	7.04	.0734	.0512
15	47	20	5.95	.0138	.0183
20	49	20	5.45	.0092	.0074

degrees of freedom ($\nu_1 = \nu_2 = \nu$), while in Fig. 7, we have plotted the profile likelihood function in the case of unequal scale matrices and unequal degrees of freedom.

Finally, we now compare the t and normal mixture based-clusterings after some outliers are introduced into the original data set. This was done by adding various values to the second variate of the 25th point. In Table 1, we report the overall misallocation rate of the normal and t mixture-based clusterings for each perturbed version of the original data set. It can be seen that the t mixture-based clustering is robust to those perturbations, unlike the normal mixture-based clustering. It should be noted in the cases where the constant equals −15, −10, and 20, that fitting a normal mixture model results in an outright classification of the outlier into one cluster. The remaining points are allocated to the second cluster giving an error rate of 49. In practice the user should identify this situation when interpreting the results and hence remove the outlier giving an error rate of 20%. However when the fitting is part of an automatic procedure this would not be the case.

Concerning the effect of outliers by working with the logs of these data under the normal mixture model (as raised by a referee), it was found that the consequent clustering is still very sensitive to atypical observations when introduced as above. However, the assumption of normality is slightly more tenable for the logged data as assessed by Hawkins' (1981) test, and the clustering of the logged data via either the normal or t mixture models does result in fewer misallocations.

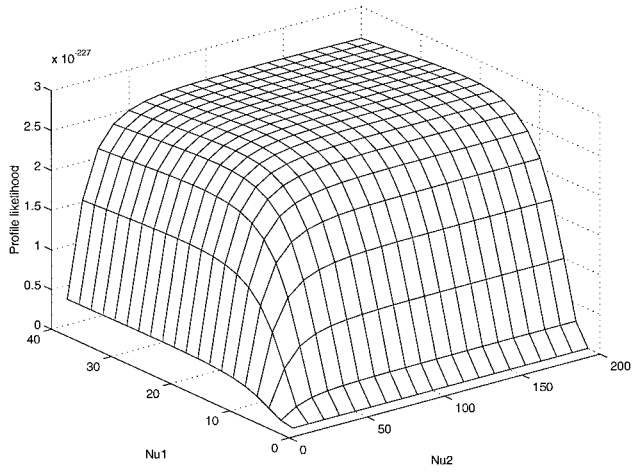

Fig. 7. *Plot of the profile likelihood for ν_1 and ν_2 for the Blue Crab data set with unequal scale matrices and unequal degrees of freedom ν_1 and ν_2*

References

Aitchison J. and Dunsmore I.R. 1975. Statistical Predication Analysis. Cambridge University Press, Cambridge.

Böhning D. 1999. Computer-Assisted Analysis of Mixtures and Applications: Meta-Analysis, Disease Mapping and Others. Chapman & Hall/CRC, New York.

Campbell N.A. 1984. Mixture models and atypical values. Mathematical Geology 16: 465–477.

Campbell N.A. and Mahon R.J. 1974. A multivariate study of variation in two species of rock crab of genus *Leptograpsus*. Australian Journal of Zoology 22: 417–425.

Davé R.N. and Krishnapuram R. 1995. Robust clustering methods: A unified view. IEEE Transactions on Fuzzy Systems 5: 270–293.

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.

Everitt B.S. and Hand D.J. 1981. Finite Mixture Distributions. Chapman & Hall, London.

Frigui H. and Krishnapuram R. 1996. A robust algorithm for automatic extraction of an unknown number of clusters from noisy data. Pattern Recognition Letters 17: 1223–1232.

Gnanadesikan R., Harvey J.W., and Kettenring J.R. 1993. Sankhyā A 55: 494–505.

Green P.J. 1984. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. Journal of the Royal Statistical Society B 46: 149–192.

Hampel F.R. 1973. Robust estimation: A condensed partial survey. Z. Wahrscheinlickeitstheorie verw. Gebiete 27: 87–104.

Hawkins D.M. and McLachlan G.J. 1997. High-breakdown linear discriminant analysis. Journal of the American Statistical Association 92: 136–143.

Huber P.J. 1964. Robust estimation of a location parameter. Annals of Mmathematical Statistics 35: 73–101.

Jolion J.-M., Meer P., and Bataouche S. 1995. Robust clustering with applications in computer vision. IEEE Transactions on Pattern Analysis and Machine Intelligence 13: 791–802.

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.

Kharin Y. 1996. Robustness in Statistical Pattern Recognition. Kluwer, Dordrecht.

Kosinski A. 1999. A procedure for the detection of multivariate outliers. Computational Statistics and Data Analysis 29: 145–161.

Kowalski J., Tu X.M., Day R.S., and Mendoza-Blanco J.R. 1997. On the rate of convergence of the ECME algorithm for multiple regression models with t-distributed errors. Biometrika 84: 269–281.

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.

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

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. Department of Statistics, Stanford University, Stanford.

Markatou M., Basu A., and Lindsay B.G. 1998. Weighted likelihood equations with bootstrap root search. Journal of the American Statistical Association 93: 740–750.

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

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

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(2). (http://www.stat.ucla.edu/journals/jss/).

Meng X.L. and van Dyk D. 1995. The EM algorithm – an old folk song sung to a fast new tune (with discussion).

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

Rousseeuw P.J., Kaufman L., and Trauwaert E. 1996. Fuzzy clustering using scatter matrices. Computational Statistics and Data Analysis 23: 135–151.

Rubin D.B. 1983. Iteratively reweighted least squares. In: Kotz S., Johnson N.L., and Read C.B. (Eds.), Encyclopedia of Statistical Sciences Vol. 4. Wiley, New York, pp. 272–275.

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.

Smith D.J., Bailey T.C., and Munford G. 1993. Robust classification of high-dimensional data using artificial neural networks. Statistics and Computing 3: 71–81.

Sutradhar B. and Ali M.M. 1986. Estimation of parameters of a regression model with a multivariate t error variable. Communications in Statistic – Theory and Methods 15: 429–450.

Titterington D.M., Smith A.F.M., and Makov U.E. 1985. Statistical Analysis of Finite Mixture Distributions. Wiley, New York.

Zhuang X., Huang Y., Palaniappan K., and Zhao Y. 1996. Gaussian density mixture modeling, decomposition and applications. IEEE Transactions on Image Processing 5: 1293–1302.