

Robust Cluster Analysis via Mixture Models

Geoffrey J. McLachlan, Shu-Kay Ng, and Richard Bean
University of Queensland, Australia

Abstract: Finite mixture models are being increasingly used to model the distributions of a wide variety of random phenomena and to cluster data sets. In this paper, we focus on the use of normal mixture models to cluster data sets of continuous multivariate data. As normality based methods of estimation are not robust, we review the use of t component distributions. With the t mixture model-based approach, the normal distribution for each component in the mixture model is embedded in a wider class of elliptically symmetric distributions with an additional parameter called the degrees of freedom. The advantage of the t mixture model is that, although the number of outliers needed for breakdown is almost the same as with the normal mixture model, the outliers have to be much larger. We also consider the use of the t distribution for the robust clustering of high-dimensional data via mixtures of factor analyzers. The latter enable a mixture model to be fitted to data which have high dimension relative to the number of data points to be clustered.

Keywords: Finite Mixture Models, Normal Components, Mixtures of Factor Analyzers, t Distributions, EM Algorithm.

1 Introduction

Finite mixture models are being increasingly used to model the distributions of a wide variety of random phenomena and to cluster data sets. Here we consider their application in the context of cluster analysis. We let the p -dimensional vector $\mathbf{x} = (x_1, \dots, x_p)^T$ contain the values of p variables measured on each of n (independent) entities to be clustered, and we let \mathbf{x}_j denote the value of \mathbf{x} corresponding to the j th entity ($j = 1, \dots, n$). With the mixture approach to clustering, $\mathbf{x}_1, \dots, \mathbf{x}_n$ are assumed to be an observed random sample from mixture of a finite number, say g , of groups in some unknown proportions π_1, \dots, π_g . The mixture density of \mathbf{x}_j is expressed as

$$f(\mathbf{x}_j; \Psi) = \sum_{i=1}^g \pi_i f_i(\mathbf{x}_j; \theta_i), \quad j = 1, \dots, n, \quad (1)$$

where the mixing proportions π_1, \dots, π_g sum to one and the group-conditional density $f_i(\mathbf{x}_j; \theta_i)$ is specified up to a vector θ_i of unknown parameters ($i = 1, \dots, g$). The vector of all the unknown parameters is given by $\Psi = (\pi_1, \dots, \pi_{g-1}, \theta_1^T, \dots, \theta_g^T)^T$, where the superscript T denotes vector transpose. Using an estimate of Ψ , this approach gives a probabilistic clustering of the data into g clusters in terms of estimates of the posterior probabilities of component membership,

$$\tau_i(\mathbf{x}_j) = \frac{\pi_i f_i(\mathbf{x}_j; \theta_i)}{f(\mathbf{x}_j; \Psi)}, \quad (2)$$

where $\tau_i(x_j)$ is the posterior probability that x_j (really the entity with observation x_j) belongs to the i th component of the mixture ($i = 1, \dots, g; j = 1, \dots, n$).

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

$$\partial \log L(\Psi) / \partial \Psi = \mathbf{0} \quad (3)$$

where

$$\log L(\Psi) = \sum_{j=1}^n \log f(x_j; \Psi) \quad (4)$$

is the log likelihood function for Ψ . Solutions of (4) corresponding to local maximizers of $\log L(\Psi)$ can be obtained via the expectation-maximization (EM) algorithm of Dempster et al. (1977).

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

$$f_i(x_j; \theta_i) = \phi(x_j; \mu_i, \Sigma_i), \quad (5)$$

where $\phi(x_j; \mu, \Sigma)$ denotes the p -dimensional multivariate normal distribution with mean vector μ and covariance matrix Σ .

One attractive feature of adopting mixture models with elliptically symmetric components such as the normal or t densities, is that the implied clustering is invariant under affine transformations of the data (that is, under operations relating to changes in location, scale, and rotation of the data); see, for example, Coleman et al. (1999). Thus the clustering process does not depend on irrelevant factors such as the units of measurement or the orientation of the clusters in space. Concerning the desirability of the latter, Hartigan (1975) has commented that affine invariance is less compelling than invariance under the change of measuring units of each variable.

Unfortunately, as with many other applications of ML estimation for normal-based models, the ML fitting of normal mixture models is not robust against gross outliers, at least if the number of components g is not fixed. The problem of providing protection against outliers in multivariate data is a very difficult problem and increases in difficulty with the dimension of the data. There is now a vast literature on robust modelling methods some of which focus on outlier identification, while others are more for outlier accommodation (Huber, 1981). In a series of papers, Rocke (1996), Rocke and Woodruff (1996), Rocke and Woodruff (1997), and Woodruff and Rocke (1993), Woodruff and Rocke (1994) have considered robust estimation of multivariate location and shape, and the consequent identification of outliers and leverage points. More recently, Davies and Gather (2005) have discussed the concept of breakdown points (Hampel, 1971; Donoho and Huber, 1983). In the context of mixtures, Hennig (2004) has given an account of robustness issues with ML estimation of univariate normal mixture models.

One-way to broaden the normal mixture parametric family for potential outliers or data with longer-than-normal tails is to adopt mixtures of t distributions, as proposed by McLachlan and Peel (1998) and McLachlan and Peel (2000b) and Peel and McLachlan (2000). Mixtures of a fixed number of t components are not robust against outliers. The advantage of the t mixture model is that, although the number of outliers needed for

breakdown is almost the same as with the normal mixture model, the outliers have to be much larger. This point is made more precise in Hennig (2004).

In this paper, we give a brief review of the fitting of mixtures of t components. Some illustrative examples are discussed. We also consider the use of mixture models for the clustering of high-dimensional data. With mixtures of normal or t component distributions, there may be problems with potential singularities in the estimates of the component scale matrices. One way to avoiding such singularities for mixture of normal components is to fit mixtures of factor analyzers. We discuss how this latter model can be made less sensitive to outliers by considering the implementation of mixtures of factor analyzers with t distributed errors in the component factor models.

Before we proceed to define the t mixture model, we give a brief account of some other methods for improving the robustness of mixture models.

2 Some Alternatives to Improving the Robustness of Mixture Models

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 (Banfield and Raftery, 1993). As shown by Hennig (2004), the noise component can be affected too by outliers. This can be prevented if the density constant for the noise component is chosen as fixed beforehand, which leads to ML estimation of a mixture model with an improper distribution added to catch the noise.

Robust estimation in the context of mixture models has been considered in the past by Campbell (1984) and McLachlan and Basford (1988, Chapter 3), among others, using M-estimates of the means and covariance matrices of the normal components of the mixture model. Markatou (2000) has provided a formal approach to robust mixture estimation by applying weighted likelihood methodology (Markatou et al., 1998) in the context of mixture models. Müller and Neykov (2003) and Neykov et al. (2004) have considered the trimmed likelihood methodology (Hadi and Luceño, 1997; Vandev and Neykov, 1998) in the fitting of mixtures of normals and generalized linear models. Also, Tibshirani and Knight (1999) have proposed the technique of bootstrap “bumping”, which can be used for resistant fitting.

It is outside the scope of this paper to review procedures besides mixture model-based ones for robust cluster analysis, such as k -medoids. There are also those methods that optimize a target function for only part of the data, such as trimmed k -means (Garcia-Escudero and Gordaliza, 1999) and minimum covariance determinant or minimum volume ellipsoid procedures (Rocke and Woodruff, 2000; Hawkins, 2003; Hawkins, 2004; and Hardin and Rocke, 2004). In the example to be considered here, we shall make use of trimmed k -means to provide a starting partition for the fitting of the t mixture model in clustering the main body of a data set in the presence of local contamination. Although

the k -means procedure corresponds to the assumption of clusters with a common spherical covariance matrix, it is often used to provide a starting partition for the application of the EM algorithm and its variants. But other clustering procedures could be used, for example, the classification method; see Coleman et al. (1999).

3 Multivariate t Distribution

For mixtures of normal components, the i th component-conditional distribution of the j th observation vector \mathbf{X}_j is given by

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

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

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

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

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

The density corresponding to (6) is given by

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

where

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

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

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

$$W_j \sim \text{gamma}(\frac{1}{2}\nu_i, \frac{1}{2}\nu_i), \quad (9)$$

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

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

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

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

$$\mathbf{X}_j | w_j \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i/w_j), \quad (11)$$

then the unconditional distribution of \mathbf{X}_j is given by the t distribution (7); see, for example, the monograph of Kotz and Nadarajah (2004) on the t distribution. As ν_i tends

to infinity, the t distribution approaches the normal distribution. Hence this parameter ν_i may be viewed as a robustness tuning parameter. It can be fixed in advance or it can be inferred from the data for each component.

For ML estimation in the case of a single t distribution, the reader is referred to Rubin (1983), Little and Rubin (1987) Liu and Rubin (1994), Liu and Rubin (1995), Liu (1997), and Liu et al. (1998), A brief history of the development of ML estimation of a single-component t distribution is given in Liu and Rubin (1995).

4 ML Estimation of Mixtures of t Components

McLachlan and Peel (2000a, Chapter 7) have implemented the E- and M-steps of the EM algorithm and its variant, the ECM (expectation–conditional maximization) algorithm for the ML estimation of multivariate t components. The ECM algorithm proposed by Meng and Rubin (1993) replaces the M-step of the EM algorithm by a number of computationally simpler conditional maximization (CM) steps.

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

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

$$\mu_i^{(k+1)} = \sum_{j=1}^n \tau_{ij}^{(k)} w_{ij}^{(k)} \mathbf{x}_j / \sum_{j=1}^n \tau_{ij}^{(k)} w_{ij}^{(k)} \quad (13)$$

and

$$\Sigma_i^{(k+1)} = \frac{\sum_{j=1}^n \tau_{ij}^{(k)} w_{ij}^{(k)} (\mathbf{x}_j - \mu_i^{(k+1)}) (\mathbf{x}_j - \mu_i^{(k+1)})^T}{\sum_{j=1}^n \tau_{ij}^{(k)}}. \quad (14)$$

In the above,

$$\tau_{ij}^{(k)} = \frac{\pi_i^{(k)} f(\mathbf{x}_j; \mu_i^{(k)}, \Sigma_i^{(k)}, \nu_i^{(k)})}{f(\mathbf{x}_j; \Psi^{(k)})} \quad (15)$$

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

$$w_{ij}^{(k)} = \frac{\nu_i^{(k)} + p}{\nu_i^{(k)} + \delta(\mathbf{x}_j, \mu_i^{(k)}; \Sigma_i^{(k)})}, \quad (16)$$

which is the current estimate of the conditional expectation of W_j given \mathbf{x}_j and $z_{ij} = 1$.

The updated estimate $\nu_i^{(k+1)}$ of ν_i does not exist in closed form, but is given as a solution of the equation

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

$$+ \psi \left(\frac{\nu_i^{(k)} + p}{2} \right) - \log \left(\frac{\nu_i^{(k)} + p}{2} \right) \Big\} = 0, \quad (17)$$

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

Following the proposal of 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 (33) by

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

which should improve the speed of convergence; see also Liu (1997) and Liu et al. (1998).

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

5 Example of Clustering via Mixtures of Normal and t Components

We consider some data simulated from a mixture in equal proportions of $g = 3$ groups with configuration similar to that considered in Ueda and Nakano (1998), McLachlan and Peel (2000a, Chapter 2), and, more recently, Cuesta-Albertos et al. (2005). The component means are given by

$$\boldsymbol{\mu}_1^T = (-2, 0), \quad \boldsymbol{\mu}_2^T = (0, 0), \quad \boldsymbol{\mu}_3^T = (2, 0),$$

while the common component-covariance matrix is diagonal,

$$\boldsymbol{\Sigma}_i = \begin{pmatrix} 0.2 & 0 \\ 0 & 2 \end{pmatrix}.$$

Cuesta-Albertos et al. (2005) added some 20 data points in two different schemes to 600 points of which 200 were simulated from each of the three normal distributions as specified above. In the first case, they added 20 points from the uniform distribution on the set,

$$\{(x_1, x_2) \in [-5, 5] \times [-8, 8] : x_1 < -4 \text{ or } > 4, \text{ or } x_2 < -5 \text{ or } > 5\}.$$

Cuesta-Albertos et al. (2005) used this example to show that if the k -means solution is used to start the EM algorithm, then fitting a mixture of $g = 3$ normal components will not lead to the desired solution, as exhibited in Figure 1. But we note here that if we fit a mixture of $g = 3$ t components from the k -means solution, then it will converge

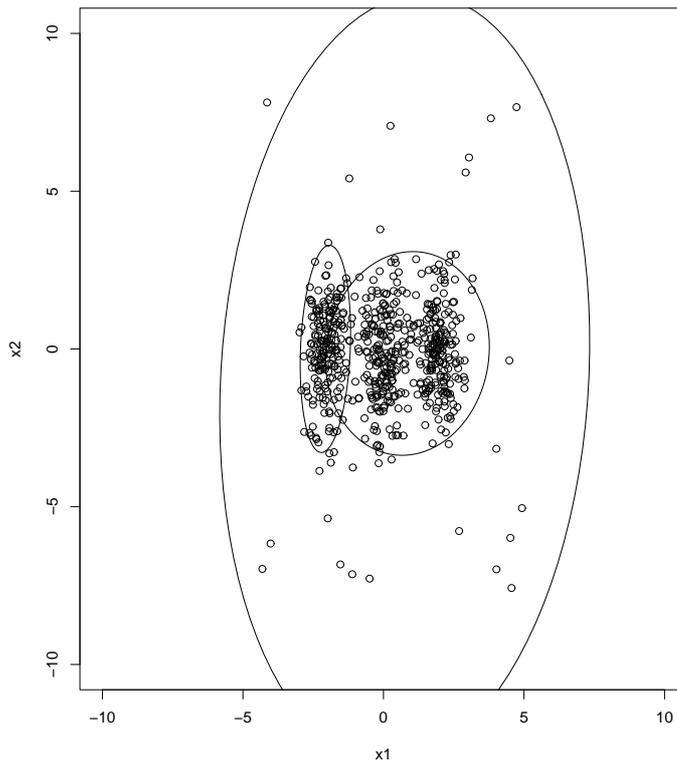


Figure 1: (Asymptotic) ellipsoids for the three clusters obtained by fitting a mixture of $g = 3$ normal components to three normal groups plus uniformly distributed noise.

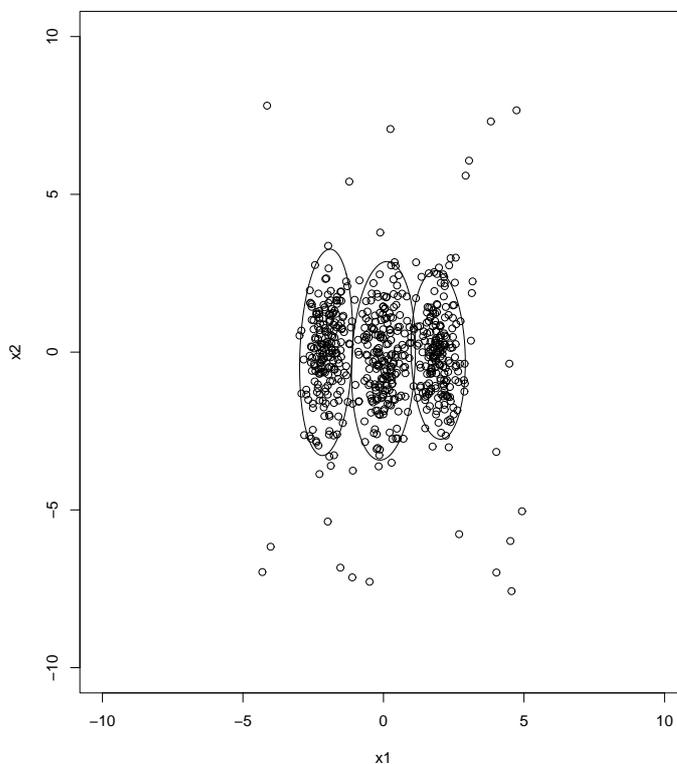


Figure 2: (Asymptotic) ellipsoids for the three clusters obtained by fitting a mixture of $g = 3t$ components to three normal groups plus uniformly distributed noise.

to the desired solution; see Figure 2. The estimated degrees of freedom in the three t components are 190.0, 3.4, and 2.6, respectively, for the components in order of increasing mean of the first variable.

In the second case, Cuesta-Albertos et al. (2005) added 20 points from the uniform distribution on the square

$$[0.5, 1.5] \times [-8, -7],$$

as illustrative of a situation with local contamination.

In this case, Cuesta-Albertos et al. (2005) noted that one gets essentially the same clustering of this contaminated data set of 620 points into $g = 3$ clusters regardless of whether one uses mixtures of normals or t components if the fitting algorithm (EM) is started from the k -means solution; see clustering displayed in Figure 3 from fitting a mixture of $g = 3$ normal components. This is obviously a situation where it helps to know what is the desired number of clusters.

One would expect that with any sensible clustering procedure that uses all of the data, that the 20 locally concentrated data points would be put into a separate cluster of their own. Thus, if the main body of the data is to be clustered into three clusters, then clearly we need to look at clustering the data into $g = 4$ clusters with one cluster for the cell of locally contaminated points or using a procedure that focuses on the main body of data and ignores the local contamination. If we adopt the former approach with mixtures of normals or t components, we will get a four-cluster solution corresponding to the three normal groups and the cell of contaminated data; see Figure 4. The fits produced by the four-component normal and t mixture models are very similar and so only the fit for the normal mixture model has been displayed in Figure 4. In fitting these two mixture models, the EM algorithm was started from the k -means solution using all the data. We also tried several random starts but the solution corresponding to the largest of the local maxima found led to the same clustering obtained using the k -means start.

Concerning the latter approach, we display in Figures 5 and 6 the three clusters obtained by fitting a mixture of $g = 3$ normal and t components, respectively, to all the 620 points, but with the EM algorithm started from the 50% trimmed k -means solution. The estimated degrees of freedom in the three t components are 126.6, 3.2, and 51.6, respectively, for the components in order of increasing mean of the first variable. It can be seen that the normal mixture fit is not robust to the local contamination even when started from a robust solution (50% trimmed k -means), but that the t mixture model is robust to the contamination.

6 Factor Analysis Model for Dimension Reduction

The g -component normal mixture model with unrestricted component-covariance matrices is a highly parameterized model with $d = p(p+1)/2$ parameters for each component-covariance matrix Σ_i ($i = 1, \dots, g$). Banfield and Raftery (1993) introduced a parameterization of the component-covariance matrix Σ_i based on a variant of the standard spectral decomposition of Σ_i . However, if p is large relative to the sample size n , it may not be possible to use this decomposition to infer an appropriate model for the component-covariance matrices. Even if it is possible, the results may not be reliable due to potential

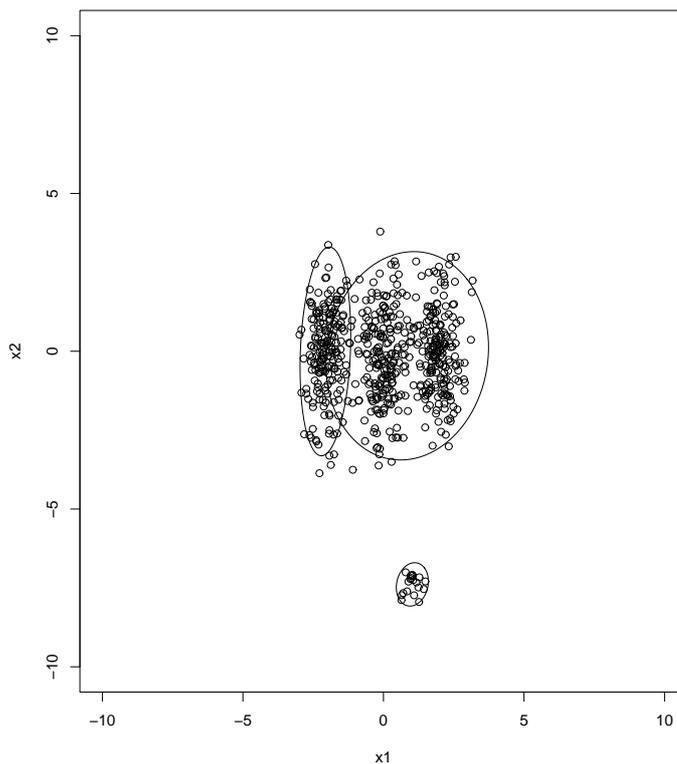


Figure 3: (Asymptotic) ellipsoids for the three clusters obtained by fitting a mixture of $g = 3$ normal components to 3 normal groups with local contamination; EM started from k -means solution.

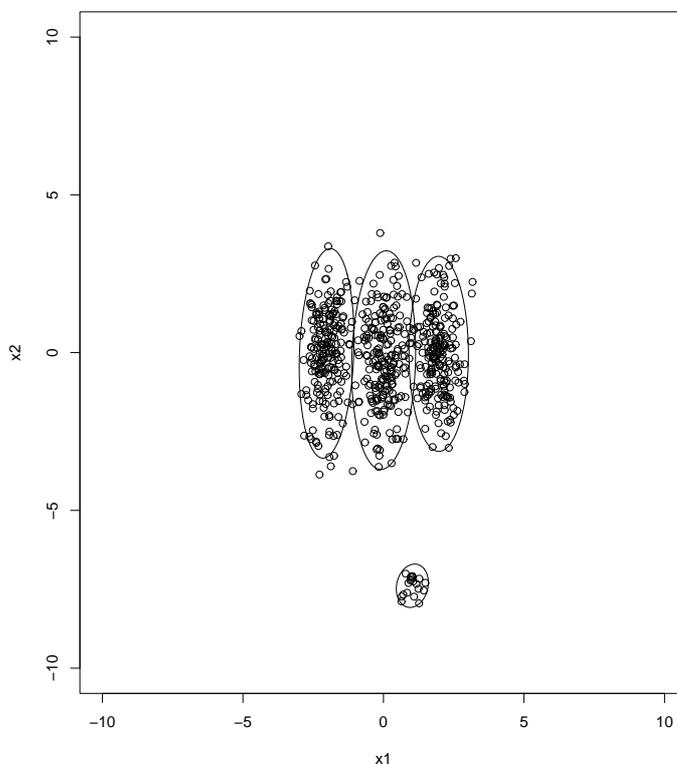


Figure 4: (Asymptotic) ellipsoids for the 4 clusters obtained by fitting a mixture of $g = 4$ normal components to 3 normal groups with local contamination; EM started from k -means solution.

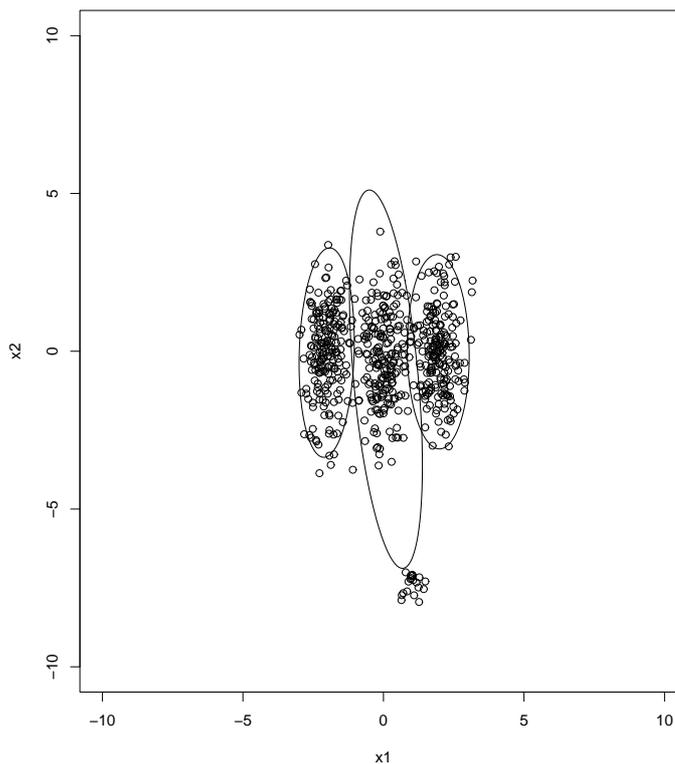


Figure 5: (Asymptotic) ellipsoids for the 3 clusters obtained by fitting a mixture of $g = 3$ normal components to 3 normal groups with local contamination; EM algorithm started from trimmed k -means solution.

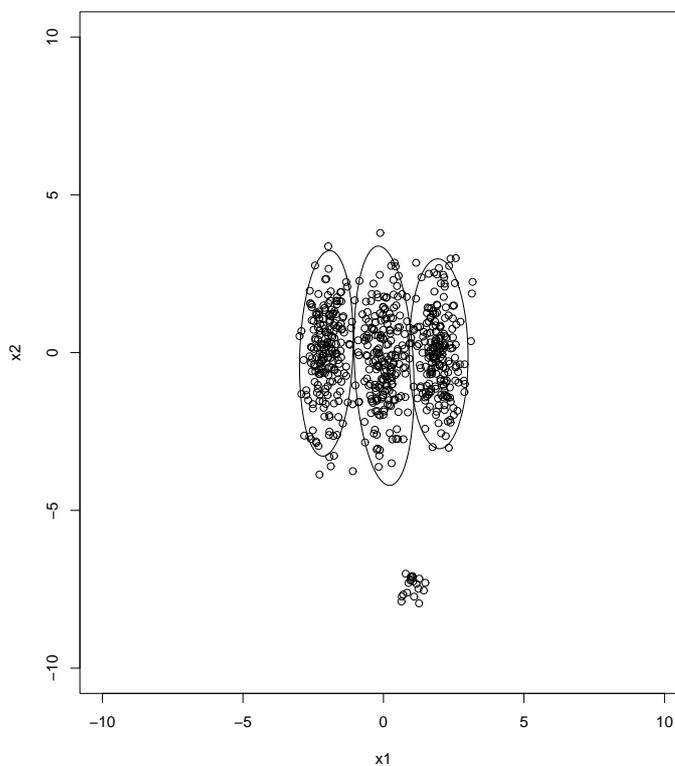


Figure 6: (Asymptotic) ellipsoids for the 3 clusters obtained by fitting a mixture of $g = 3$ t components to 3 normal groups with local contamination; EM algorithm started from trimmed k -means solution.

problems with near-singular estimates of the component-covariance matrices when p is large relative to n .

A common approach to reducing the the number of dimensions is to perform a principal component analysis (PCA). But as is well known, projections of the feature data x_j onto the first few principal axes are not always useful in portraying the group structure; see McLachlan and Peel (2000a, page 239), and Chang (1983). Another approach for reducing the number of unknown parameters in the forms for the component-covariance matrices is to adopt the mixture of factor analyzers model, as considered in McLachlan and Peel (2000a), McLachlan and Peel (2000b). This model was originally proposed by Ghahramani and Hinton (1997) and Hinton, Dayan, and Revow (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. Further references may be found in McLachlan and Peel (2000a, Chapter 8).

In the sequel, we focus on mixtures of factor analyzers from the perspective of a method for model-based density estimation from high-dimensional data, and hence for the clustering of such data. This approach enables a normal mixture model to be fitted to a sample of n data points of dimension p , where p is large relative to n . The number of free parameters is controlled through the dimension of the latent factor space. By working in this reduced space, it allows a model for each component-covariance matrix with complexity lying between that of the isotropic and full covariance structure models without any restrictions on the covariance matrices.

7 Mixtures of Normal Factor Analyzers

A global nonlinear approach can be obtained by postulating a finite mixture of linear submodels for the distribution of the full observation vector X_j given the (unobservable) factors u_j . That is, we can provide a local dimensionality reduction method by assuming that the distribution of the observation X_j can be modelled as

$$X_j = \mu_i + B_i U_{ij} + e_{ij} \quad \text{with prob. } \pi_i, \quad i = 1, \dots, g \quad (18)$$

for $j = 1, \dots, n$, where the factors U_{i1}, \dots, U_{in} are distributed independently $N(\mathbf{0}, I_q)$, independently of the e_{ij} , which are distributed independently $N(\mathbf{0}, D_i)$, where D_i is a diagonal matrix ($i = 1, \dots, g$).

Thus the mixture of factor analyzers model is given by

$$f(x_j; \Psi) = \sum_{i=1}^g \pi_i \phi(x_j; \mu_i, \Sigma_i), \quad (19)$$

where the i th component-covariance matrix Σ_i has the form

$$\Sigma_i = B_i B_i^T + D_i, \quad i = 1, \dots, g \quad (20)$$

and where B_i is a $p \times q$ matrix of factor loadings and D_i is a diagonal matrix ($i = 1, \dots, g$). The parameter vector Ψ now consists of the mixing proportions π_i and the elements of the μ_i , the B_i , and the D_i .

The mixture of factor analyzers model can be fitted by using the alternating expectation–conditional maximization (AECM) algorithm (Meng and van Dyk, 1997). The AECM algorithm is an extension of the ECM algorithm, where the specification of the complete data is allowed to be different on each CM-step.

To apply the AECM algorithm to the fitting of the mixture of factor analyzers model, we partition the vector of unknown parameters Ψ as $(\Psi_1^T, \Psi_2^T)^T$, where Ψ_1 contains the mixing proportions π_i ($i = 1, \dots, g - 1$) and the elements of the component means μ_i ($i = 1, \dots, g$). The subvector Ψ_2 contains the elements of the B_i and the D_i ($i = 1, \dots, g$).

We let $\Psi^{(k)} = (\Psi_1^{(k)T}, \Psi_2^{(k)T})^T$ be the value of Ψ after the k th iteration of the AECM algorithm. For this application of the AECM algorithm, one iteration consists of two cycles, and there is one E-step and one CM-step for each cycle. The two CM-steps correspond to the partition of Ψ into the two subvectors Ψ_1 and Ψ_2 .

For the first cycle of the AECM algorithm, we specify the missing data to be just the component-indicator labels z_{ij} , which are defined as above. The first conditional CM-step leads to $\pi_i^{(k)}$ and $\mu_i^{(k)}$ being updated to

$$\pi_i^{(k+1)} = \frac{1}{n} \sum_{j=1}^n \tau_i(\mathbf{x}_j; \Psi^{(k)}) \quad (21)$$

and

$$\mu_i^{(k+1)} = \frac{\sum_{j=1}^n \tau_i(\mathbf{x}_j; \Psi^{(k)}) \mathbf{x}_j}{\sum_{j=1}^n \tau_i(\mathbf{x}_j; \Psi^{(k)})} \quad (22)$$

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

$$\tau_i(\mathbf{x}_j; \Psi) = \frac{\pi_i \phi(\mathbf{x}_j; \mu_i, \Sigma_i)}{\sum_{h=1}^g \pi_h \phi(\mathbf{x}_j; \mu_h, \Sigma_h)} \quad (23)$$

is the i th component posterior probability of \mathbf{x}_j .

For the second cycle for the updating of Ψ_2 , we specify the missing data to be the factors $\mathbf{u}_1, \dots, \mathbf{u}_n$, as well as the component-indicator labels z_{ij} . On setting $\Psi^{(k+1/2)}$ equal to $(\Psi_1^{(k+1)T}, \Psi_2^{(k)T})^T$, an E-step is performed to calculate $Q(\Psi; \Psi^{(k+1/2)})$, which is the conditional expectation of the complete-data log likelihood given the observed data, using $\Psi = \Psi^{(k+1/2)}$. The CM-step on this second cycle is implemented by the maximization of $Q(\Psi; \Psi^{(k+1/2)})$ over Ψ with Ψ_1 set equal to $\Psi_1^{(k+1)}$. This yields the updated estimates $B_i^{(k+1)}$ and $D_i^{(k+1)}$. The former is given by

$$B_i^{(k+1)} = \mathbf{V}_i^{(k+1/2)} \gamma_i^{(k)} \left(\gamma_i^{(k)T} \mathbf{V}_i^{(k+1/2)} \gamma_i^{(k)} + \omega_i^{(k)} \right)^{-1}, \quad (24)$$

where

$$\mathbf{V}_i^{(k+1/2)} = \frac{\sum_{j=1}^n \tau_i(\mathbf{x}_j; \Psi^{(k+1/2)}) (\mathbf{x}_j - \mu_i^{(k+1)}) (\mathbf{x}_j - \mu_i^{(k+1)})^T}{\sum_{j=1}^n \tau_i(\mathbf{x}_j; \Psi^{(k+1/2)})}, \quad (25)$$

$$\gamma_i^{(k)} = (\mathbf{B}_i^{(k)} \mathbf{B}_i^{(k)T} + \mathbf{D}_i^{(k)})^{-1} \mathbf{B}_i^{(k)}, \quad (26)$$

and

$$\omega_i^{(k)} = \mathbf{I}_q - \gamma_i^{(k)T} \mathbf{B}_i^{(k)} \quad (27)$$

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

$$\begin{aligned} \mathbf{D}_i^{(k+1)} &= \text{diag}\{\mathbf{V}_i^{(k+1/2)} - \mathbf{B}_i^{(k+1)} \mathbf{H}_i^{(k+1/2)} \mathbf{B}_i^{(k+1)T}\} \\ &= \text{diag}\{\mathbf{V}_i^{(k+1/2)} - \mathbf{V}_i^{(k+1/2)} \gamma_i^{(k)} \mathbf{B}_i^{(k+1)T}\}, \end{aligned} \quad (28)$$

where

$$\begin{aligned} \mathbf{H}_i^{(k+1/2)} &= \frac{\sum_{j=1}^n \tau_i(\mathbf{x}_j; \Psi^{(k+1/2)}) \mathbf{E}_i^{(k+1/2)}(\mathbf{U}_j \mathbf{U}_j^T | \mathbf{x}_j)}{\sum_{j=1}^n \tau_i(\mathbf{x}_j; \Psi^{(k+1/2)})} \\ &= \gamma_i^{(k)T} \mathbf{V}_i^{(k+1/2)} \gamma_i^{(k)} + \omega_i^{(k)} \end{aligned} \quad (29)$$

and $\mathbf{E}_i^{(k+1/2)}$ denotes conditional expectation given membership of the i th component, using $\Psi^{(k+1/2)}$ for Ψ .

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

$$\hat{\mathbf{D}}_i = \text{diag}(\hat{\mathbf{V}}_i - \hat{\mathbf{B}}_i \hat{\mathbf{B}}_i^T), \quad (30)$$

where

$$\hat{\mathbf{V}}_i = \frac{\sum_{j=1}^n \tau_i(\mathbf{x}_j; \hat{\Psi})(\mathbf{x}_j - \hat{\boldsymbol{\mu}}_i)(\mathbf{x}_j - \hat{\boldsymbol{\mu}}_i)^T}{\sum_{j=1}^n \tau_i(\mathbf{x}_j; \hat{\Psi})}. \quad (31)$$

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

On comparing (30) with (16), it can be seen that with the calculation of the ML estimate of \mathbf{D}_i directly from the (incomplete-data) log likelihood function, the unconditional expectation of $\mathbf{U}_j \mathbf{U}_j^T$, which is the identity matrix, is used in place of the conditional expectation in (29) on the E-step of the AECM algorithm. Unlike the direct approach of calculating the ML estimate, the EM algorithm and its variants such as the AECM version have good convergence properties in that they ensure the likelihood is not decreased after each iteration regardless of the choice of starting point; see McLachlan et al. (2003) for further discussion.

It can be seen from (30) that some of the estimates of the elements of the diagonal matrix \mathbf{D}_i (the uniquenesses) will be close to zero if effectively not more than q observations are unequivocally assigned to the i th component of the mixture in terms of the fitted

posterior probabilities of component membership. This will lead to spikes or near singularities in the likelihood. One way to avoid this is to impose the condition of a common value D for the D_i ,

$$D_i = D, \quad i = 1, \dots, g. \quad (32)$$

An alternative way of proceeding is to adopt some prior distribution for the D_i as, for example, in the Bayesian approach of Fokoué and Titterington (2002).

The mixture of probabilistic component analyzers (PCAs) model, as proposed by Tipping and Bishop (1997) has the form (20) with each D_i now having the isotropic structure

$$D_i = \sigma_i^2 I_p \quad i = 1, \dots, g. \quad (33)$$

Under this isotropic restriction (33) the iterative updating of B_i and D_i is not necessary since, given the component membership of the mixture of PCAs, $B_i^{(k+1)}$ and $\sigma_i^{(k+1)^2}$ are given explicitly by an eigenvalue decomposition of the current value of V_i .

8 Mixtures of t Factor Analyzers

The mixture of factor analyzers model is sensitive to outliers since it uses normal errors and factors. Recently, McLachlan and Bean (2005) have considered the use of mixtures of t analyzers in an attempt to make the model less sensitive to outliers. With mixtures of t factor analyzers, the error terms e_{ij} and the factors U_{ij} are assumed to be distributed according to the t distribution with the same degrees of freedom. Under this model, the factors and error terms are no longer independently distributed but they are uncorrelated.

It follows that mixtures of t factor analyzers can be fitted essentially as in the previous section for normal factors and errors with minor modification. In equations (22) and (25), the weights $w_{ij}^{(k+1)}$ should be used, and of course the t density $f(\mathbf{x}_j; \boldsymbol{\mu}_i^{(k+1)}, \boldsymbol{\Sigma}_i^{(k+1)}, \nu_i^{(k+1)})$ should be used in forming the current estimates of the posterior probabilities of component membership $\tau_i(\mathbf{x}_j; \boldsymbol{\Psi}^{(k+1)})$. Further details are provided in McLachlan and Bean (2005).

9 Discussion

In this paper, we have considered the use of mixtures of multivariate t distributions instead of normal components as a more robust approach to the clustering of multivariate continuous data which have longer tails than the normal or atypical observations. As pointed out by Hennig (2004), although the number of outliers needed for breakdown with the t mixture model is almost the same as with the normal version, the outliers have to be much larger.

In considering the robustness of mixture models, it is usual to consider the number of components as fixed. This is because the existence of outliers in a data set can be handled by the addition of further components in the mixture model if the number of components is not fixed. Breakdown can still occur if the contaminating points lie between the clusters of the main body of points and fill in the feature space to the extent that a fewer number of components is needed in the mixture model than the actual number of clusters (Hennig,

2004). But obviously the situation is fairly straightforward if the number of clusters are known *a priori*. However, this is usually not the case in clustering applications.

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

References

- Banfield, J. D., and Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics*, 49, 803-821.
- Campbell, N. A. (1984). Mixture models and atypical values. *Mathematical Geology*, 16, 465-477.
- Chang, W. C. (1983). On using principal components before separating a mixture of two multivariate normal distributions. *Applied Statistics*, 32, 267-275.
- Coleman, D., Dong, X., Hardin, J., Rocke, D. M., and Woodruff, D. L. (1999). Some computational issues in cluster analysis with no a priori metric. *Computational Statistics & Data Analysis*, 31, 1-11.
- Cuesta-Albertos, J. A., Matrán, C., and Mayo-Isacar, A. (2005). Estimators based in adaptively trimming cells in the mixture model. <http://personales.unican.es/cuestaj/stemcell.pdf>.
- Davies, P. L., and Gather, U. (2005). Breakdown and groups (with discussion). *Annals of Statistics*, 33, 977-1035.
- 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.
- Donoho, D. L., and Huber, J. (1983). The notion of breakdown point. In J. L. H. P. Bickel K. Doksum (Ed.), *A festschrift for erich l. lehmann* (p. 157-184). Wadworth: Belmont.
- Fokoué, E., and Titterington, D. M. (2002). Mixtures of factor analyzers. bayesian estimation and inference by stochastic simulation. *Machine Learning*, 50, 73-94.
- Garcia-Escudero, L. A., and Gordaliza, A. (1999). Robustness properties of k means and trimmed k means. *Journal of the American Statistical Association*, 956-969.
- Ghahramani, Z., and Hinton, G. E. (1997). *The EM Algorithm for Mixtures of Factor Analyzers* (Vol. CRG-TR-96-1). University of Toronto: Technical Report.
- Hadi, A. S., and Luceño, A. (1997). Maximum trimmed likelihood estimators: a unified approach, examples, and algorithms. *Computational Statistics and Data Analysis*, 25, 251-272.
- Hampel, F. R. (1971). A general qualitative definition of robustness. *Annals of Mathematical Statistics*, 42, 1887-1896.
- Hardin, J., and Rocke, D. M. (2004). Outlier detection in the multiple cluster setting using

- the minimum covariance determinant estimator. *Computational Statistics and Data Analysis*, 44, 625-638.
- Hartigan, J. A. (1975). Statistical theory in clustering. *Journal of Classification*, 2, 63-76.
- Hawkins, D. M. (2003). A feasible solution algorithm for the minimum volume ellipsoid estimator in multivariate data. *Computational Statistics*, 9, 95-107.
- Hawkins, D. M. (2004). The feasible solution algorithm for the minimum covariance determinant estimator in multivariate data. *Computational Statistics and Data Analysis*, 17, 197-210.
- Hennig, C. (2004). Breakdown points for maximum likelihood estimators of location-scale mixtures. *Annals of Statistics*, 32, 1313-1340.
- Hinton, G. E., Dayan, P., and Revow, M. (1997). Modeling the manifolds of images of handwritten digits. *IEEE Transactions in neural Networks*, 8, 65-73.
- Huber, P. J. (1981). *Robust statistics*. New York: J. Wiley.
- Kotz, S., and Nadarajah, S. (2004). *Multivariate t distributions and their applications*. New York: Cambridge University Press.
- Lawley, D. N., and Maxwell, A. E. (1971). *Factor analysis as a statistical method*. London: Butterworths.
- Little, R. J. A., and Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: J. Wiley.
- 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. (2000). Mixture models, robustness and the weighted likelihood methodology. *Biometrics*, 56, 483-486.
- 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. (1988). *Mixture models: Inference and applications to clustering*. New York: Marcel Dekker.
- McLachlan, G. J., and Bean, R. W. (2005). *Maximum likelihood estimation of mixtures of t factor analyzers*. Brisbane, Queensland, Australia: Technical Report, University of Queensland.
- McLachlan, G. J., and Peel, D. (1998). Robust cluster analysis via mixtures of multivariate t distributions. In A. Amin, D. Dori, P. Pudil, and H. Freeman (Eds.), *Lecture notes in computer science* (Vol. 1451, p. 658-666). Berlin: Springer-Verlag.
- McLachlan, G. J., and Peel, D. (2000a). *Finite mixture models*. New York: J. Wiley.
- McLachlan, G. J., and Peel, D. (2000b). Mixtures of factor analyzers. In P. Langley (Ed.), *Proceedings of the seventeenth international conference on machine learning* (p. 599-606). San Francisco: Morgan Kaufmann.
- McLachlan, G. J., Peel, D., and Bean, R. (2003). Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics and Data Analysis*, 41, 379-

388.

- Meng, X. L., and Rubin, D. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, 80, 267-278.
- Meng, X. L., and van Dyk, D. (1997). 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.
- Müller, C. H., and Neykov, N. (2003). Breakdown points of trimmed likelihood estimators and related estimators in generalized linear models. *Journal of Statistical Planning and Inference*, 116, 503-519.
- Neykov, N., Filzmoser, P., Dimova, R., and Neytchev, P. (2004). In *Compstat 2004, proceedings computational statistics* (p. 1585-1592). Vienna: Physica-Verlag.
- Peel, D., and McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. *Statistical Computing*, 10, 335-344.
- Roche, D. M. (1996). Robustness properties of S-estimators of multivariate location and shape in high dimension. *Annals of Statistics*, 24, 1327-1345.
- Roche, D. M., and Woodruff, D. (1996). Identification of outliers in multivariate data. *Journal of the American Statistical Association*, 91, 1047-1061.
- Roche, D. M., and Woodruff, D. (1997). Robust estimation of multivariate location and shape. *Journal of Statistical Planning and Inference*, 57, 245-255.
- Roche, D. M., and Woodruff, D. (2000). A synthesis of outlier detection and cluster identification. *Unpublished manuscript*.
- Rubin, D. B. (1983). Iteratively reweighted least squares. In *Encyclopedia of statistical sciences* (Vol. 4, p. 272-275). New York: J. Wiley.
- Tibshirani, R., and Knight, K. (1999). Model search by bootstrap "bumping". *Journal of Computational and Graphical Statistics*, 8, 671-686.
- Tipping, M. E., and Bishop, C. M. (1997). Mixtures of probabilistic principal component analysers. In *Technical report no. NCRG/97/003*. Birmingham, Aston University: Neural Computing Research Group.
- Tyler, J. T. K. D. E., and Vardi, Y. (1994). A curious likelihood identity for the multivariate t -distribution. *Communications in Statistics - Simulation and Computation*, 23, 441-453.
- Ueda, N., and Nakano, R. (1998). Deterministic annealing EM algorithm. *Neural Networks*, 11, 271-282.
- Vandev, D. L., and Neykov, N. (1998). About regression estimators with high breakdown point. *Statistics*, 32, 111-129.
- Woodruff, D. L., and Roche, D. M. (1993). Heuristic search algorithms for the minimum volume ellipsoid. *Journal of Computational and Graphical Statistics*, 2, 69-95.
- Woodruff, D. L., and Roche, D. M. (1994). Computable robust estimation of multivariate location and shape using compound estimators. *Journal of the American Statistical Association*, 89, 888-896.

Authors' addresses:

Professor Geoffrey J. McLachlan
Department of Mathematics and
the Institute for Molecular Bioscience
The University of Queensland
Brisbane Q4072
Australia
Tel. +61 7 33652150
Fax +61 7 33651477
E-Mail: gjm@maths.uq.edu.au
<http://www.maths.uq.edu.au/gjm/>

Dr. Shu-Kay Ng
Department of Mathematics
The University of Queensland
Brisbane Q4072
Australia
Tel. +61 7 33656139
Fax +61 7 33651477
E-Mail: skn@maths.uq.edu.au
<http://www.maths.uq.edu.au/skn/>

Dr. Richard Bean
The Institute for Molecular Bioscience
The University of Queensland
Brisbane Q4072
Australia
Tel. +61 7 33462627
Fax +61 7 33651477
E-Mail: rbean@maths.uq.edu.au
<http://www.maths.uq.edu.au/rbean/>