Robust Mixture Modeling

G. J. McLachlan, S. K. Ng, and R. W. Bean
Department of Mathematics &
Institute for Molecular Bioscience
University of Queensland, 4072
Brisbane, Australia

Sponsor: Section on Physical and Engineering Sciences

Keywords: finite mixture models; EM algorithm; multiresolution kd-trees; t distributions; mixtures of factor analyzers

Abstract: Finite mixture models are being increasingly used to model the distributions of a wide variety of random phenomena and to cluster datasets. We shall focus on the use of normal mixture models to cluster datasets of continuous multivariate data. We shall consider a robust approach to clustering by modeling the data by a mixture of t distributions. With this 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. Finally, we consider the robust fitting of normal mixtures using multiresolution kd-trees.

1 Introduction

Finite mixtures of distributions have provided a mathematical-based approach to the statistical modeling of a wide variety of random phenomena. Because of their usefulness as an extremely flexible method of modeling, finite mixture models have continued to receive increasing attention over the years, from both a practical and theoretical point of view; see, for example, McLachlan and Basford (1988) and McLachlan and Peel (2000). Mixture distributions have been applied to data with two main purposes in mind: (i) to provide an appealing semiparametric framework in which to model unknown distributional shapes, as an alternative to, say, the kernel density method; (ii) to use the mixture model to provide a model-based clustering. In both situations, there is the question of how many components to include in the mixture.

Frequently, in practice, the clusters in the data are essentially elliptical, so that it is reasonable to consider fitting mixtures of elliptically symmetric component densities. Within this class of component densities, the multivariate normal density is a convenient choice given its computational tractability. Also, any continuous distribution can be approximated arbitrarily well by a finite mixture of normal densities with common variance (or covariance matrix in the multivariate case).

For many applied problems, the tails of the normal distribution are often shorter than required. Also, the estimates of the component means and covariance matrices can be affected by observations that are atypical of the components in the normal mixture model being fitted. The problem of providing protection against outliers in multivariate data is a very difficult problem and increases in difficulty with the dimension of the data (Rocke and Woodruff, 1997; Kosinski, 1999). In this paper, we consider the use of mixtures of t distributions as a more robust approach to the fitting of mixture models.

Recently, Hennig (2004) has provided an in-depth study of breakdown points (including their definitions) for maximum likelihood estimators of g-component location-scale mixtures for both fixed and unfixed g. As he notes, the addition of gross outliers is almost harmless from the theoretical point of view in mixture estimation with g unfixed, because outliers can be accom-

modated by including more components in the mixture model. He goes on to point out that breakdown can occur if the added points lie inside the range of the original data, as it may lead to a solution with a smaller number of clusters than the original number of components g. In this study, we consider the robustness of normal mixture models in the case of a fixed number of components g.

2 Model-Based Clustering

We firstly consider the use of normal mixture models as a device for the clustering of multivariate data.

2.1 Clustering via Mixture Models

In recent times much attention has been given in the statistical literature to the use of finite mixture models as a device for clustering; see, for example, McLachlan and Peel (2000). With this approach, the observed data $\boldsymbol{y}_1, \ldots \boldsymbol{y}_n$, are assumed to have come from a mixture of a finite number, say g, of groups $G_1, \ldots, G_g$ in some unknown proportions $\pi_1, \ldots, \pi_g$. The mixing proportions π_i lie between zero and one, and sum to one. The feature vector $\boldsymbol{Y}$ is taken to have the density $f_i(\boldsymbol{y})$ in group G_i $(i = 1, \ldots, g)$. Thus unconditionally with respect to its group of origin, the feature vector $\boldsymbol{Y}$ has the mixture density

$$f(\boldsymbol{y}) = \sum_{i=1}^{g} \pi_i f_i(\boldsymbol{y}). \tag{1}$$

In this mixture framework, the posterior probability that an observation with feature vector $\boldsymbol{y}_j$ belongs to the ith component of the mixture is given by

$$\tau_i(\boldsymbol{y}_j) = \pi_i f_i(\boldsymbol{y}_j)/f(\boldsymbol{y}_j) \tag{2}$$

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

On specifying a parametric form $f_i(\boldsymbol{y}_j; \boldsymbol{\theta}_i)$ for each component density, we can fit this parametric mixture model

$$f(\boldsymbol{y}_j; \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i \, f_i(\boldsymbol{y}_j; \boldsymbol{\theta}_i) \tag{3}$$

by maximum likelihood via the expectation-maximization (EM) algorithm of Dempster, Laird, and Rubin (1977); see also McLachlan and Krishnan (1997). Here $\boldsymbol{\Psi} = (\boldsymbol{\omega}^T, \pi_1, \ldots, \pi_{g-1})^T$ is the vector of unknown parameters, where $\boldsymbol{\omega}$ consists of the elements of the $\boldsymbol{\theta}_i$ known *a priori* to be distinct. In order to estimate $\boldsymbol{\Psi}$ from the observed data, it must be identifiable. This will be so if the representation (1) is unique up to a permutation of the component labels.

The actual fitting of finite mixture models by maximum likelihood via the EM algorithm. Let $\hat{\boldsymbol{\Psi}}$ denote the estimate of $\boldsymbol{\Psi}$ so obtained. Then

$$\tau_i(\boldsymbol{y}_j; \hat{\boldsymbol{\Psi}}) = \hat{\pi}_i \, f_i(\boldsymbol{y}_j; \hat{\boldsymbol{\theta}}_i) / \sum_{h=1}^{g} \hat{\pi}_h \, f_h(\boldsymbol{y}_j; \hat{\boldsymbol{\theta}}_h) \tag{4}$$

is the estimated posterior probability that the jth observation with feature vector $\boldsymbol{y}_j$ belongs to the ith component of the mixture $(i = 1, \ldots, g; j = 1, \ldots, n)$. The mixture approach gives a probabilistic clustering in terms of these estimated posterior probabilities of component membership. An outright partitioning of the observations into g nonoverlapping clusters $C_1, \ldots, C_g$ is effected by assigning each observation to the component to which it has the highest estimated posterior probability of belonging. Thus the ith cluster C_i contains those observations assigned to group G_i.

2.2 Cluster Analysis with No *A Priori* Metric

Many clustering methods assume that the similarity measure or metric is known *a priori*. Often the Euclidean metric is used as with k-means clustering. However, it is more appropriate to use a distance function (metric) that depends on the shape of the clusters. For example, if a cluster is multivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, the appropriate distance between a point $\boldsymbol{y}$ and the center $\boldsymbol{\mu}$ of the cluster is the squared Mahalanobis distance

$$\delta(\boldsymbol{y}, \boldsymbol{\mu}; \boldsymbol{\Sigma}) = (\boldsymbol{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) \tag{5}$$

between $\boldsymbol{y}$ and $\boldsymbol{\mu}$. The difficulty is that the shape of the clusters is not known until the clusters have been identified, and the clusters cannot be effectively identified unless the shapes are known. Indeed, as noted by Hansen and Tukey (1992), "The shakiest part of any clustering procedure is the choice of the metric."

To avoid reliance on any *a priori* metric, Coleman et al. (1999) advocate the use of affine invariant clustering algorithms. This means that the clustering produced on the transformed data $\boldsymbol{C}\boldsymbol{y} + \boldsymbol{a}$ is the same

as on the untransformed data $\boldsymbol{y}$. Here $\boldsymbol{C}$ is a nonsingular matrix. It means that the clustering is invariant under location (translations of the data), scale (stretchings of the data), and rotation (orientations of the data). Thus affine-invariant metrics are particularly appropriate for use in clustering, since the results do not depend on irrelevant factors such as the units of measurement or the orientation of the clusters in space. Hartigan (1975, Page 63) has commented that "Invariance under this general class of linear transformations seems less compelling than invariance under the change of measuring units of each of the variables."

Essentially, affine invariance of clustering is equivalent to assuming that the metric is quadratic but otherwise unspecified; that is, the distance between any two points $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ is given by

$$d(\boldsymbol{y}_1, \boldsymbol{y}_2) = (\boldsymbol{y}_1 - \boldsymbol{y}_2)^T \boldsymbol{B}^{-1} (\boldsymbol{y}_1 - \boldsymbol{y}_2) \qquad (6)$$

with $\boldsymbol{B}$ a positive-definite symmetric matrix. Quadratic metrics can arise naturally in a number of ways, such as with mixture models with component distributions such as the multivariate normal or other elliptically symmetric distributions (the t distribution). Note that Euclidean distance corresponds to the use of (6) with $\boldsymbol{B}$ equal to the $p \times p$ identity matrix.

2.3 Normal Mixture Models

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). Thus the clustering process does not depend on irrelevant factors such as the units of measurement or the orientation of the clusters in space.

2.4 Advantages of Model-Based Clustering

The mixture likelihood-based approach to clustering is model based in that the form of each component density of an observation has to be specified in advance. Hawkins, Muller, and ten Krooden (1982) commented that most writers on cluster analysis "lay more stress on algorithms and criteria in the belief that intuitively reasonable criteria should produce good results over a wide range of possible (and generally unstated) models." For example, the trace $\boldsymbol{W}$ criterion, where $\boldsymbol{W}$ is the pooled within-cluster sums of squares and products matrix, is predicated on normal groups with (equal) spherical covariance matrices; but as they pointed out, many users apply this criterion even in the face of evidence of nonspherical clusters or, equivalently, would use Euclidean distance as a metric. They strongly supported the increasing emphasis on a model-based approach to clustering. Indeed, as remarked by Aitkin, Anderson, and Hinde (1981) in the reply to the discussion of their paper, "when clustering samples from a population, no cluster method is, *a priori* believable without a statistical model." Concerning the use of mixture models to represent nonhomogeneous populations, they noted in their paper that "Clustering methods based on such mixture models allow estimation and hypothesis testing within the framework of standard statistical theory." Previously, Marriott (1974) had noted that the mixture likelihood-based approach "is about the only clustering technique that is entirely satisfactory from the mathematical point of view. It assumes a well-defined mathematical model, investigates it by well-established statistical techniques, and provides a test of significance for the results." More recently, a model-based approach to clustering has been advocated by Banfield and Raftery (1993) and Fraley and Raftery (1998, 2004).

3 Mixtures of t Distributions

3.1 Extension of Normal Family

One way to broaden the normal parametric family for potential outliers or data with longer-than-normal tails is to adopt the two-component normal mixture density

$$(1-\epsilon)\phi(\boldsymbol{y}_j;\, \boldsymbol{\mu},\, \boldsymbol{\Sigma}) + \epsilon\phi(\boldsymbol{y}_j;\, \boldsymbol{\mu},\, k\boldsymbol{\Sigma}), \qquad (7)$$

where k 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 to be discussed further in Section 4.

The normal scale mixture model (7) can be written as

$$\int \phi(\boldsymbol{y}_j;\, \boldsymbol{\mu},\, \boldsymbol{\Sigma}/u)\, dH(u), \qquad (8)$$

where H is the probability distribution that places mass $(1-\epsilon)$ at the point $u = 1$ and mass ϵ at the

point $u = 1/k$. Suppose we now replace H by the distribution of a chi-squared random variable on its degrees of freedom ν; that is, by the random variable U distributed as

$$U \sim \text{ gamma } (\tfrac{1}{2}\nu, \tfrac{1}{2}\nu), \tag{9}$$

where the gamma (α, β) density function is given by

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

We then obtain the t distribution with location parameter $\boldsymbol{\mu}$, positive definite inner product matrix $\boldsymbol{\Sigma}$, and ν degrees of freedom,

$$\begin{aligned} & f(\boldsymbol{y}_j; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) \\ = \; & \frac{\Gamma(\frac{\nu+p}{2}) \, |\boldsymbol{\Sigma}|^{-1/2}}{(\pi\nu)^{\frac{1}{2}p} \Gamma(\frac{\nu}{2}) \{1 + \delta(\boldsymbol{y}_j, \boldsymbol{\mu}; \boldsymbol{\Sigma})/\nu\}^{\frac{1}{2}(\nu+p)}}, \end{aligned} \tag{11}$$

where

$$\delta(\boldsymbol{y}_j, \boldsymbol{\mu}; \boldsymbol{\Sigma}) = (\boldsymbol{y}_j - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_j - \boldsymbol{\mu}) \tag{12}$$

denotes the Mahalanobis squared distance between $\boldsymbol{y}_j$ and $\boldsymbol{\mu}$ (with $\boldsymbol{\Sigma}$ as the covariance matrix). If $\nu > 1$, $\boldsymbol{\mu}$ is the mean of $\boldsymbol{Y}_j$, 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 $\boldsymbol{Y}_j$ becomes marginally multivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. 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 (McLachlan and Peel, 1998; McLachlan and Peel, 2000, Chapter 7). More recently, Kotz and Nadarajah (2004) have written a book devoted to the t distribution.

The t distribution does not have substantially better breakdown behavior than the normal. 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 by Hennig (2004) who has provided an excellent account of breakdown points for ML estimation of location-scale mixtures with a fixed number of components g. Also, as noted by Lange et al. (1989), the use of the t distribution is not a panacea for all forms of robustness. Data with shorter-than-normal tails, asymmetric distributions, varying degrees of long-tailedness among the feature variables, or with extreme outliers will not be able to be modeled adequately by a mixture of t distributions.

3.2 Maximum Likelihood Estimation

The mixture of t distributions can be fitted by maximum likelihood (ML) via the EM algorithm, as described in McLachlan and Peel (2000, Chapter 7). A history of the development of ML estimation of a single-component t distribution may be found in Liu and Rubin (1994, 1995), Liu (1997), and Meng and van Dyk (1997).

On the M-step of the $(k+1)$th iteration of the EM algorithm, the mixing proportions are updated as

$$\pi_i^{(k+1)} = \sum_{j=1}^{n} \tau_{ij}^{(k)}/n \quad (i = 1, \ldots, g), \tag{13}$$

where $\tau_{ij}^{(k)} = \tau_i(\boldsymbol{y}_j; \hat{\boldsymbol{\Psi}}^{(k)})$ is the current estimate of the posterior probability that observation $\boldsymbol{y}_j$ belongs to the ith component. The updated estimates of $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are given by

$$\boldsymbol{\mu}_i^{(k+1)} = \sum_{j=1}^{n} \tau_{ij}^{(k)} u_{ij}^{(k)} \boldsymbol{y}_j / \sum_{j=1}^{n} \tau_{ij}^{(k)} u_{ij}^{(k)} \tag{14}$$

and

$$\boldsymbol{\Sigma}_i^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_{ij}^{(k)} u_{ij}^{(k)} (\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k+1)})(\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k+1)})^T}{\sum_{j=1}^{n} \tau_{ij}^{(k)}}, \tag{15}$$

where

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

It can be seen that the EM process effectively chooses $\boldsymbol{\mu}_i^{(k+1)}$ and $\boldsymbol{\Sigma}_i^{(k+1)}$ by IRLS. The E-step updates the weights $u_{ij}^{(k)}$, while the M-step chooses $\boldsymbol{\mu}_i^{(k+1)}$ and $\boldsymbol{\Sigma}_i^{(k+1)}$ by weighted least-squares estimation. From the form of the equation (14) derived for the MLE of $\boldsymbol{\mu}_i$, we have that, as $\nu_i^{(k)}$ decreases, the degree of down-weighting 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 (15) by

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

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

It can be seen that if the degrees of freedom ν_i is fixed in advance for each component, then the M-step exists in closed form. In this case where ν_i is fixed beforehand, the estimation of the component parameters is a form of M-estimation; see Lange et al. (1989, p. 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 MLE. In this case, it can be shown that $\nu_i^{(k+1)}$ is a solution of

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

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

With mixture models, the likelihood equation will usually have multiple roots corresponding to local maxima, and so the EM algorithm (or its variants) should be applied from a wide choice of starting values in any search for all local maxima. For example, one can use random starts or partitions obtained via some other clustering procedure such as k-means; see also Biernacki (2004), Biernacki, Celeux, and Govaert (2003), and Coleman and Woodruff (2003). In the absence of the observed value of any known consistent estimator or any other information, an obvious choice for the root of the likelihood equation is the one corresponding to the largest of the local maxima located (excluding so-called spurious local maximizers).

4 Some 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.

With M-estimation, the updated component means $\boldsymbol{\mu}_i^{(k+1)}$ are given by (14), but where now the weights $u_{ij}^{(k)}$ are defined as

$$u_{ij}^{(k)} = \psi(d_{ij}^{(k)})/d_{ij}^{(k)} \tag{18}$$

and where

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

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

$$\begin{aligned} \psi(s) &= s, && \mid s \mid \leq a, \\ &= \text{sign}(s)\, a, && \mid s \mid > a, \end{aligned} \tag{19}$$

for an appropriate choice of the tuning constant a. The ith component-covariance matrix $\boldsymbol{\Sigma}_i^{(k+1)}$ can be updated as (15), 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 densities (McLachlan and Basford, 1988; Section 2.8). One possibility is to use the form of the density corresponding to the ψ-function adopted. However, in the case of any redescending ψ-function with finite rejection points, there is no corresponding density. In Campbell (1984), the normal density 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 deviations, 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 densities.

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).

In other work, 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

$$\sum_{j=1}^{n} w(\boldsymbol{y}_j)\partial \log f(\boldsymbol{y}_j; \boldsymbol{\Psi})/\partial \boldsymbol{\Psi} = \mathbf{0}, \tag{20}$$

where $f(\boldsymbol{y}_j; \boldsymbol{\Psi})$ denotes the specified parametric form for the density of $\boldsymbol{Y}_j$. The weight function $w(\boldsymbol{y}_j)$ is defined in terms of the Pearson residuals; see Markatou, Basu, and Lindsay (1998). The weighted likelihood methodology provides robust and first-order efficient estimators in general, and Markatou (1998) has established these results in the context of univariate mixture models. Also, Tibshirani and Knight (1999) have proposed the technique of bootstrap "bumping," which can be used for resistant fitting.

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, 1994). As the noise (uniform) component defined in this way can be severely affected by outliers, Hennig (2004) has suggested a modified uniform approach whereby the density constant for the noise component is fixed beforehand using an improper prior component.

There are other approaches to robust cluster analysis that are implemented by optimizing a target function for only part of the data. For example, (i) trimmed k-means (Garcia-Escudero and Gordaliza, 1999) and (ii) minimum covariance determinant procedures (Hawkins, 2000; Rocke and Woodruff, 2000).

5 Mixtures of Factor Analyzers

A normal mixture model without restrictions on the component-covariance matrices may be viewed as too general for many situations in practice, particularly with high dimensional data. In exploring high-dimensional data sets for group structure, it is typical to rely on principal component analysis. However, the latter is a global linear method, and may not always be appropriate for finding group structure in a space spanned by the leading principal components. One approach for reducing the number of parameters is to work in a lower dimensional space by adopting mixtures of factor analyzers (McLachlan and Peel, 2000, Chapter 8). The mixture of factor models as given below provides a global nonlinear approach to dimension reduction as it postulates a finite mixture of linear submodels (factor models) for the distribution of the full observation vector given the (unobservable) factors. Thus, it is a local dimensionality reduction method.

The mixture of factor analyzers model is given by

$$f(\boldsymbol{y}_j; \Psi) = \sum_{i=1}^{g} \pi_i \phi(\boldsymbol{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{21}$$

where

$$\boldsymbol{\Sigma}_i = \boldsymbol{B}_i \boldsymbol{B}_i^T + \boldsymbol{D}_i (i = 1, \cdots, g) \tag{22}$$

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

This model can be fitted by an alternating expectation–conditional maximization (AECM) algorithm.

On the first cycle, the missing data are declared to be the component-label vectors in order to update the estimates of π_i and $\boldsymbol{\mu}_i$ as follows

$$\pi_i^{(k+1)} = \sum_{j=1}^{n} \tau_{ij}^{(k+1)}/n \tag{23}$$

$$\boldsymbol{\mu}_i^{(k+1)} = \sum_{j=1}^{n} \tau_{ij}^{(k)} \boldsymbol{y}_j / \sum_{j=1}^{n} \tau_{ij}^{(k)} \tag{24}$$

for $i = 1, \cdots, g$.

On the second cycle, the missing data are declared also to be the unobservable factors in order to update the estimates of $\boldsymbol{B}_i$ and $\boldsymbol{D}_i$ as follows

$$B_i^{(k+1)} = V_i^{(k+1/2)} \gamma_i^{(k)} (\gamma_i^{(k)^T} V_i^{(k+1/2)} \gamma_i^{(k)} + \omega_i^{(k+1/2)})^{-1} \tag{25}$$

$$D_i^{(k+1)} = \mathrm{diag}\{V_i^{(k+1/2)} - V_i^{(k+1/2)} \gamma_i^{(k)} B_i^{(k+1)^T}\} \tag{26}$$

where

$$\gamma_i^{(k)} = (B_i^{(k)} B_i^{(k)^T} + D_i^{(k)})^{-1} B_i^{(k)}, \tag{27}$$

$$\omega_i^{(k)} = I_q - \gamma_i^{(k)^T} B_i, \tag{28}$$

and $V_i^{(k+1/2)}$ is given by

$$\frac{\sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j; \boldsymbol{\Psi}^{(k+1/2)})(\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k+1)})(\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k+1)})^T}{\sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j; \boldsymbol{\Psi}^{(k+1/2)})}. \tag{29}$$

An alternative way of proceeding is to adopt some prior distribution for the $\boldsymbol{D}_i$ as, for example, in the Bayesian approach of Fokoué and Titterington (2003).

We can make this model less sensitive to outliers by introducing in (24) and (29) weights similar to the $u_{ij}^{(k)}$ in the fitting of mixtures of t components. This can be viewed as an *ad hoc* approach to the fitting of the mixture of factor analyzers model in which the errors in the factor submodels are taken to have a t distribution located at the origin.

6 Stem-Cell Approach

Cuesta-Albertos, Matran, and Mayo-Iscar (2004) have proposed the use of so-called stem-cell estimators to improve the robustness of estimators in the mixture model. Their method is applicable in the case where the number of clusters is known *a priori*. This latter knowledge makes the task of robust cluster analysis much easier than in the general case where there is no knowledge on the number of clusters in the data. We shall see this in the examples below. Their algorithm is started from the 50%-trimmed k-means solution.

To demonstrate their method, Cuesta-Albertos et al. (2004) considered a core data set on which they compared their method with the results obtained by fitting mixtures of t and normal mixtures started from the nontrimmed k-means. The core data set consisted of 600 data points generated from a mixture of $g = 3$ normal groups as considered in Ueda and Nakano (1998). The parameters of this model are:

$\mu_1 = (0\ 3)^T, \mu_2 = (3\ 0)^T, \mu_3 = (-3\ 0)^T,$

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 2 & 0.5 \\ 0.5 & 0.5 \end{pmatrix},$$

$$\boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 0.1 \end{pmatrix},$$

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

In their first example for which the 600 core data points were considered uncontaminated, they found that their method gave similar results to the t and normal mixture models started from the nontrimmed k-means solution, as summarized in Figure 1. However, they found that their method gave different results to the t and normal mixture models in their second and third examples in which contamination was introduced into the core data points, as illustrated in Figures 2 and 3. In the first case of contamination (Example 2), the 600 points were contaminated by adding 20 points obtained from the uniform distribution on the distribution uniform on the square $[-5, 5] \times [-8, 8]$. In their second contamination case (Example 3), the 600 points were contaminated by adding 20 points from distribution uniform on the square $[0.5, 1.5] \times [-8, -7]$. However, when we fitted the t mixture model from the 50% k-means solution, it gave similar results to the stem-cell method of Cuesta-Albertos et al. (2004).

These examples demonstrate that knowledge about the true number of clusters can be crucial in robust

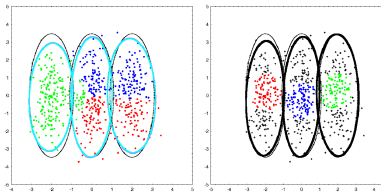

Figure 1: Simulation of 600 points of a mixture of three 2-dimensional Gaussian distributions as described in Example 1. Curves represent the 95% level ellypses of the true distribution (thin ones) and the estimated distributions (cyan and black-thick ones). Estimations on the left hand side graphic (that practically coincide) were made with the t and normal mixture models, while the one on the right hand side was made with the stem-cell procedure. Different colors denote the initial three-clusters as obtained with the non-trimmed 3-means (left-hand side) and with the 50% trimmed 3-means (right-hand side).

cluster analysis. If we know beforehand that the main body of points constitute the true clusters, then it would be wise to consider fitting the mixture model of t components from a clustering obtained by some robust method such as trimmed k-means. But without this knowledge, it may not be appropriate to employ such a clustering procedure initially, particularly in the case where there may be interest in finding break-away clusters. For example, if we fit a mixture of $g = 4$ t components to the contaminated data in Example 3 above, starting from the nontrimmed k-means solution, we obtain four clusters of which three correspond to the main body of points and the fourth to the smaller number of break-away points.

7 Robust Estimation via Multiresolution kd-trees

In this section, we consider a robust implementation of normal mixture models based on multiresolution kd-trees (mrkd-trees). Here kd stands for k-dimensional where, in our notation, $k = p$, the dimension of an observation $\boldsymbol{y}_j$.

7.1 kd-Trees for Mixture Models

Multiresolution kd-tree-based approaches have been adopted to speed up the EM algorithm (Moore, 1999;

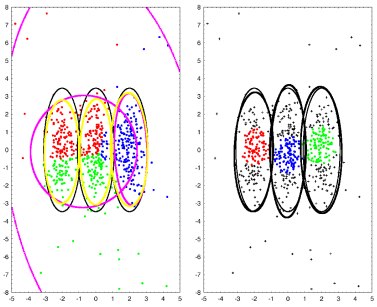

Figure 2: Simulation of 600 points of a mixture of three 2-dimensional Gaussian distributions plus 20 contaminated observations generated as described in Example 2, first contaminated case. Curves represent the 95% level ellypses of the true distribution (thin ones) and the estimated distributions. Estimation on the left-hand side graphic represented by the yellow (resp. violet) colour was made with the t and normal mixture models, while the one on the right hand side was made with the stem-cell procedure. Different colours on the points denote the initial three-clusters as obtained with the non-trimmed 3-means (left-hand side) and with the 50% trimmed 3-means (right-hand side).

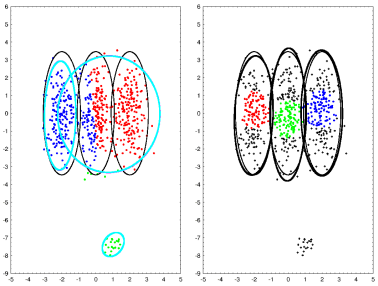

Figure 3: Simulation of 600 points of a mixture of three 2-dimensional Gaussian distributions plus 20 contaminated observations generated as described in Example 3, second contaminated case. Curves represent the 95% level ellypses of the true distribution (thin ones) and the estimated distributions. Estimations on the left-hand side graphic represented by the cyan colour were made with the t and normal mixture models (they practically coincide). The one on the right-hand side was made with the stem-cell procedure. Different symbols denote the initial three-clusters as obtained with the non-trimmed 3-means (left-hand side) and with the 50% trimmed 3-means (right-hand side).

Ng and McLachlan, 2004). Basically, this approach builds a multiresolution data structure (partition) to summarize the data at all resolutions of interest simultaneously. With the mrkd-tree approach, "close-by" observations are grouped into tree-nodes and the conditional expectations of the sufficient statistics are simplified by treating all the data points in a node to have the same posterior probabilities $\tau_i(\bar{\boldsymbol{y}}; \boldsymbol{\Psi}^{(k)})$ calculated at the mean $\bar{\boldsymbol{y}}$ (Ng and McLachlan, 2004). The method thus speeds up the EM algorithm roughly a factor of n/n_L, where n_L is the number of leaf-nodes (the smallest possible partitions this mrkd-tree offers). In practice, the leaf nodes should be very small in order that the approximation using $\tau_i(\bar{\boldsymbol{y}}; \boldsymbol{\Psi}^{(k)})$ be applicable. However, in this situation, n_L will be close to n, and hence there is very little computational gain over the standard EM algorithm. Thus, a further (pruning) step is proposed by Moore (1999) to identify those nodes in which the difference between the minimum and maximum values of the posterior probabilities is small. Such nodes are then treated as if they are "pseudo" leaf nodes and hence their descendants need not be searched at this iteration and time is saved (Ng and McLachlan, 2004).

7.2 Robust Estimation via a Sparse and Incremental mrkd-Tree Algorithm

Recently, a sparse and incremental (SPIEM) mrkd-tree algorithm has been proposed to further increase the speedup factor, while without the compromise on the "quality" of the clustering result (Ng and McLachlan, 2004). The latter is justified by the experimental results showing that mrkd-tree-based algorithms can converge to essentially the same maximum log likelihood value as the EM algorithm. With the SPIEM mrkd-tree algorithm, the nodes at a predetermined level, say L, of the mrkd-tree are divided into B blocks and a "partial" E-step is implemented by searching down from only a block of nodes at level L at a time before the next M-step is performed. Here the number of blocks B is chosen based on the simple rule proposed by Ng and McLachlan (2003a). The argument for improved convergence rate is that the algorithm exploits new information more quickly, rather than waiting for a complete scan of all nodes before parameters are updated by an M-step. Moreover, component-posterior probabilities that are below a specified threshold are held fixed while those for the remaining components in the mixture are updated. Thus, instead of considering all g components, it is possible to "freeze" those $\tau_i(\bar{\boldsymbol{y}}; \boldsymbol{\Psi}^{(k)})$ that are close to zero (say, less than 0.005) and save time; see Ng and McLachlan (2004).

Robust fitting of normal mixtures has been considered by Campbell (1984) using Huber's (1964) M-estimators, where reduced weights are given to observations that are atypical of a component on the M-step of the EM algorithm. With the mrkd-trees structure, it is proposed to perform a robust estimation for normal mixtures by identifying tree-nodes as three different types. Different weights are then given on them in the calculation of parameters (Ng and McLachlan, 2003b). Let n_{PL} denote the number of pseudo leaf nodes. The $\boldsymbol{\mu}_i^{(k+1)}$ and the $\boldsymbol{\Sigma}_i^{(k+1)}$ are updated as follows:

$$\boldsymbol{\mu}_i^{(k+1)} = \frac{\sum_{m=1}^{n_{PL}} \bar{\tau}_{im}^{(k)} n_m u_{im}^{(k)} \bar{\boldsymbol{y}}_m}{\sum_{m=1}^{n_{PL}} \bar{\tau}_{im}^{(k)} n_m u_{im}^{(k)}}, \tag{30}$$

and $\boldsymbol{\Sigma}_i^{(k+1)}$ is given by

$$\frac{\sum_{m=1}^{n_{PL}} \bar{\tau}_{im}^{(k)} n_m u_{im}^{2\ (k)} (\bar{\boldsymbol{y}}_m - \boldsymbol{\mu}_i^{(k+1)})(\bar{\boldsymbol{y}}_m - \boldsymbol{\mu}_i^{(k+1)})^T}{\sum_{m=1}^{n_{PL}} \bar{\tau}_{im}^{(k)} n_m u_{im}^{2\ (k)}}. \tag{31}$$

where $\bar{\tau}_{im}^{(k)}$ denotes $\tau_i(\bar{\boldsymbol{y}}_m; \boldsymbol{\Psi}^{(k)})$, n_m is the number of data points in the mth pseudo-leaf node, and $u_{im}^{(k)}$ is obtained from (18), but where now $d_{ij}^{(k)}$ is replaced by $d_{im}^{(k)}$ which uses $\bar{\boldsymbol{y}}_m$ instead of $\boldsymbol{y}_j$.

It is noted that the categorization of tree-nodes does not induce an extra burden of computation over the original SPIEM mrkd-tree algorithm, as the computations involved can be readily obtained by using only the mrkd-tree code of the original algorithm. The type of each tree node is determined at the pruning process of the mrkd-trees and is based on the "denseness" of the node and the squared distance d_{im}^2 between $\bar{\boldsymbol{y}}_m$ and the current estimated component mean $\boldsymbol{\mu}_i$ $(i = 1, \ldots, g;\ m = 1, \ldots, n_{PL})$.

7.3 Categorization of Tree-Nodes

The first type is that the node is close to at least one of g components. Let λ_i and λ_i' denote the smallest and the largest eigenvalues of $\boldsymbol{\Sigma}_i$ $(i = 1, \ldots, g)$, which are, respectively, the minimum and the maximum values of the Mahalanobis squared distance for all points on unit sphere. For the mth node, if the squared distance

$$d_{hm}^2 < \lambda_h \qquad \text{for some } h \in \{1, \ldots, g\},$$

then data points in this node are considered to come from the main body (inlier) of the normal mixture. Full weight $u_{im} = 1$ is given to this node for all $i = 1, \ldots, g$.

The second type is that the node is far away from all the component centers and is not dense. The former condition is determined if

$$d_{im}^2 > 4\lambda_i' \qquad \text{for all } i \ (i = 1, \ldots, g).$$

The latter is determined if (1) the number of data points in the mth node is smaller than a threshold, say ten, and (2) the maximum diagonal element of the sample covariance matrix $\boldsymbol{S}_m$ of data points in the node, say in the vth dimension $v \in \{1, \ldots, p\}$, satisfies

$$(\boldsymbol{S}_m)_{vv} > 0.1(\boldsymbol{S})_{vv},$$

where $\boldsymbol{S}$ is the global sample covariance of the whole data set. Data points in this node are then considered to be come from the noise (outlier of the normal mixture) and reduced weight $u_{im} = 1/d_{im}$ is given for all $i = 1, \ldots, g$. A dense node is not considered as an outlier automatically, because a moderate size cluster of data points may not arise simply by chance (noise), and could be an interesting feature of the data requiring further investigation.

All nodes that are not identified as one of the above two types form the third category. The weight u_{im} given to these nodes is based on Huber's ψ-function (19) with $a^2 = \chi^2_{p,0.95}$ (McLachlan and Basford, 1988, Section 2.8). Thus, nodes that are atypical of a component are being given reduced weight in the calculation of parameters (equations (30) and (31)).

7.4 Simulated Example

The simulated data consists initially of 50000 data points generated from a eight-component bivariate normal mixture, to which 5000 noise points were added from a uniform distribution over the range -10 to 10 on each variate. The parameters of the mixture model were presented in Table 1. Here, we assume equal mixing proportions $\pi_i = 1/8$ $(i = 1, \ldots, 8)$. The true grouping of the eight-component normal mixture is shown in Fig. 4(a). We now consider the clustering obtained by the robust estimation using the SPIEM mrkd-tree algorithm. The clustering so obtained is given in Fig. 4(b). It compares well with the true grouping in Fig. 4(a). The result of fitting normal mixture of eight components is given in Fig. 4(c) for comparison. It can be seen that the eight-component mixture fails to identify correctly some covariance matrices.

Table 1: Component-means and covariance matrices of the bivariate normal mixture

i	Mean ($\boldsymbol{\mu}_i$)	Covariance matrix ($\boldsymbol{\Sigma}_i$)
1	$\begin{pmatrix} 3 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 0.1 \end{pmatrix}$
2	$\begin{pmatrix} 3 \\ -6 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 0.1 \end{pmatrix}$
3	$\begin{pmatrix} -6 \\ 5 \end{pmatrix}$	$\begin{pmatrix} 1 & 0.1 \\ 0.1 & 0.1 \end{pmatrix}$
4	$\begin{pmatrix} 5 \\ 7 \end{pmatrix}$	$\begin{pmatrix} 1 & -0.1 \\ -0.1 & 0.1 \end{pmatrix}$
5	$\begin{pmatrix} -4 \\ 6 \end{pmatrix}$	$\begin{pmatrix} 1 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}$
6	$\begin{pmatrix} -1 \\ 7 \end{pmatrix}$	$\begin{pmatrix} 2 & 0 \\ 0 & 0.5 \end{pmatrix}$
7	$\begin{pmatrix} 0 \\ 3 \end{pmatrix}$	$\begin{pmatrix} 2 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}$
8	$\begin{pmatrix} -3 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 2 & -0.5 \\ -0.5 & 0.5 \end{pmatrix}$

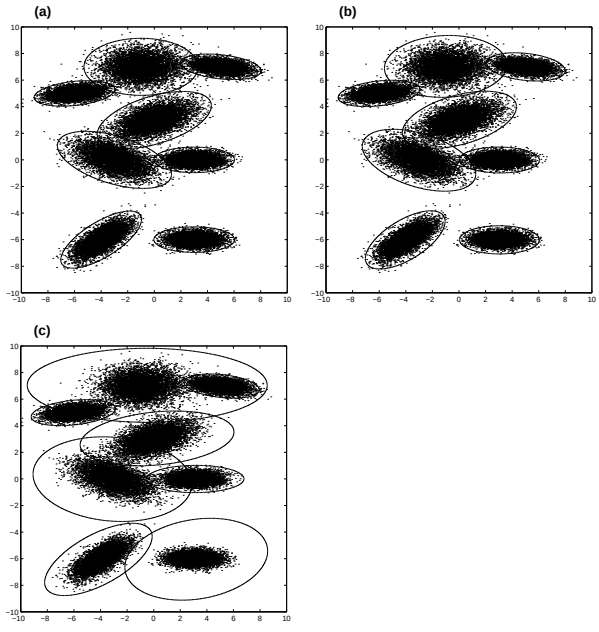

Figure 4: Results for simulated normal mixture with noise

Table 2: Computational performances for simulated normal mixture with noise

Method	No. of iterations	CPU time (seconds)
SPIEM–*mrk*d-tree	12	5
EM (8 components)	44	77
EM (11 components)	134	322

A more complex mixture model may be adopted to model the additional background noise. If the number of components is treated as unknown and a normal mixture is fitted, then the number of components can be selected via BIC (McLachlan and Peel, 2000, Sections 6.8–6.9). The additional three components are attempting to model the background noise. However, estimation of some covariance matrices is still affected by the noise. In comparing the computational performance of these algorithms, the same initialization procedure was used in this simulation study. Ten trials of k-means with two iterations were performed for each model to initialize the EM-based algorithms (McLachlan and Peel, 2000, p. 98). The number of iterations and the CPU time (in seconds) required for various models are presented in Table 2. The results presented in Figure 4 and Table 2 indicate that the SPIEM *mrk*d-tree algorithm is able to speed up the implementation of the EM algorithm and at the same time provide robust estimation without much extra computational burden compared to the fitting of normal mixture models.

8 References

Aitchison, J. and Dunsmore, I.R. (1975). *Statistical Predication Analysis*. Cambridge: Cambridge University Press.

Aitkin, M., Anderson, D., and Hinde, J. (1981). Statistical modelling of data on teaching styles (with discussion). *Journal of the Royal Statistical Society A* **144**, 414–461.

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

Biernacki, C. (2004). initializing EM using the properties of its trajectories in Gaussian mixtures. **14**, 267–279.

Biernacki, C., Celeux, G., and Govaert, G. (2002). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics and Data Analysis* **41**, 561–575.

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

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 and Data Analysis* **31**, 1–11.

Coleman, D.A. and Woodruff, D.L. (2000). Cluster analysis for large datasets: an effective algorithm for maximizing the mixture likelihood. *Journal of Computational and Graphical Statistics* **9**, 672–688.

Cuesta-Albertos, J.A., Matrán, C., and Mayo-Iscar, A. (2004). Stem-cell based estimators in the mixture model. Unpublished technical report. Department of Mathematics, Universidad de Cantabria.

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

Fokoué, E. and Titterington, D.M. (2003). Mixtures of Factor Analysers. Bayesian Estimation and Inference by Stochastic Simulation, *Machine Learning* **50**, 73–94.

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

Fraley, C. and Raftery, A.E. (2004). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* **97**, 611–631.

Fokoué, E. and Titterington, D.M. (2003). Mixtures of factor analysers. Bayesian estimation and inference by stochastic simulation. *Machine Learning* **50**, 73–94.

García-Escudero, L.A. and Gordaliza, A. (1999). Robustness properties of k means and trimmed means. *Journal of the American Statistical Association* **94**, 956–969.

Hampel, F.R. (1973). Robust estimation: a condensed partial survey. *Z. Wahrscheinlickeitstheorie verw. Gebiete* **27**, 87–104

Hansen, K.M. and J.W. Tukey (1992). Tuning a Major Part of a Clustering Algorithm, *International Statistical Review* **60**, 21–44.

Hartigan, J.A. (1975). Statistical theory in clustering, *Journal of Classification* **2**, 63–76.

Hawkins, D.M. (1999). Improved feasible solution algorithms for high breakdown estimation. *Computational Statistics and Data Analysis* **30**, 1–11.

Hawkins, D.M., Muller, M.W., and ten Krooden, J.A. (1982). Cluster analysis. In *Topics in Applied Multivariate Analysis*, D.M. Hawkins (Ed.). Cambridge: Cambridge University Press, pp. 303–356.

Hennig, C. (2004). Breakdown points for maximum-likelihood estimators of location-scale mixtures. *Annals of Statistics* **32**, 1313–1340.

Huber, P.J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics* **35**, 73–101.

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.

Kosinski, A. S., (1999). A procedure for the detection of multivariate outliers. *Computational Statistics and Data Analysis* **29**, 145–161.

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

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.

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. and Rubin, D.B. (1998). Maximum likelihood estimation of factor analysis using the ECME algorithm with complete and incomplete data. *Statistica Sinica* **8**, 729–747.

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

Marriott, F.H.C. (1974). *The Interpretation of Multiple Observations*. London: Academic Press.

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

McLachlan, G.J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. New York: Wiley.

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

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

McLachlan, G.J., Peel, D., and Bean, R.W. (2003). Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics & Data Analysis* **41**, 379–388.

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.

Ng, S.K. and McLachlan, G.J. (2003a). On the choice of the number of blocks with the incremental EM algorithm for the fitting of normal mixtures. **13**, 45–55.

Ng, S.K. and McLachlan, G.J. (2003b). Robust estimation in Gaussian mixtures using multiresolution kd-trees. In *Proceedings of DICTA 2003, 7th Conference of Digital Image Computing: Techniques and Applications* Vol. 1, C. Sun, H. Talbot, S. Ourselin, and T. Adriaansen (Eds.). Sydney: Australian Pattern Recognition Society, pp. 145-154.

Ng, S.K. and McLachlan, G.J. (2004). Speeding up the EM algorithm for mixture model-based segmentation of magnetic resonance images *Pattern Recognition* **37**, 1573–1589.

Peel, D. and McLachlan, G.J. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing* **10**, 335–344.

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

Tibshirani, R. and Knight, K. (1999). Model search by bootstrap "bumping." *Journal of Computational and Graphical Statistics* **8**, 671–686.

Ueda, N. and Nakano, R. (1998). Deterministic annealing EM algorithm. *Neural Networks* **11**, 271–282.