ANNUAL REVIEWS

Annual Review of Statistics and Its Application

Finite Mixture Models

Geoffrey J. McLachlan, Sharon X. Lee, and Suren I. Rathnayake

School of Mathematics and Physics, University of Queensland, St. Lucia, Queensland 4072, Australia; email: g.mclachlan@uq.edu.au

Annu. Rev. Stat. Appl. 2019. 6:355–78

First published as a Review in Advance on September 13, 2018

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-031017-100325

Copyright © 2019 by Annual Reviews. All rights reserved

ANNUAL REVIEWS CONNECT

www.annualreviews.org

- Download figures
- Navigate cited references
- Keyword search
- Explore related articles
- Share via email or social media

Keywords

mixture proportions, EM algorithm, normal and t-mixture distributions, model-based clustering, mixtures of factor analyzers

Abstract

The important role of finite mixture models in the statistical analysis of data is underscored by the ever-increasing rate at which articles on mixture applications appear in the statistical and general scientific literature. The aim of this article is to provide an up-to-date account of the theory and methodological developments underlying the applications of finite mixture models. Because of their flexibility, mixture models are being increasingly exploited as a convenient, semiparametric way in which to model unknown distributional shapes. This is in addition to their obvious applications where there is group-structure in the data or where the aim is to explore the data for such structure, as in a cluster analysis. It has now been three decades since the publication of the monograph by McLachlan & Basford (1988) with an emphasis on the potential usefulness of mixture models for inference and clustering. Since then, mixture models have attracted the interest of many researchers and have found many new and interesting fields of application. Thus, the literature on mixture models has expanded enormously, and as a consequence, the bibliography here can only provide selected coverage.

1. INTRODUCTION

1.1. Flexible Method of Modeling

The importance of finite mixture models in the statistical analysis of data is evident in the ever-increasing rate at which articles on theoretical and practical aspects of mixture models appear in the statistical and general scientific literature. This is because finite mixtures of distributions are being widely used to provide computationally convenient representations for modeling complex distributions of data on random phenomena. Fields in which mixture models have been successfully applied include agriculture, astronomy, bioinformatics, biology, economics, engineering, genetics, imaging, marketing, medicine, neuroscience, psychiatry, and psychology, among many other fields in the biological, physical, and social sciences. In these applications, finite mixture models underpin a variety of techniques in major areas of statistics, including cluster and latent class analyses, discriminant analysis, image analysis, and survival analysis, in addition to their more direct role in data analysis and inference of providing descriptive models for distributions where a single component distribution is apparently inadequate.

In the statistical literature, there are the books on mixture models by Everitt & Hand (1981), Titterington et al. (1985), McLachlan & Basford (1988), Lindsay (1995), Böhning (1999), McLachlan & Peel (2000a), Frühwirth-Schnatter (2006), Mengersen et al. (2011), and McNicholas (2017). In addition, mixture models are addressed in several books involving classification, machine learning, and other fields in multivariate analysis. The reader is referred to the references in these aforementioned books on mixture models and papers cited in this article for further coverage of the topic.

1.2. A Brief History

One of the first major analyses involving the use of mixture models was undertaken nearly 125 years ago by the famous biometrician Karl Pearson. In his now-classic paper, the famous biometrician, statistician, and eugenicist Pearson (1894) fitted a mixture of two normal probability density functions with different means μ_1 and μ_2 and different variances σ_1^2 and σ_2^2 to some crab data provided by his colleague, the evolutionary biologist Weldon (1892, 1893). The possibility of resolving a normal mixture into its constituent components was, of course, implicit in Quetelet's (1846, 1852) work and was mentioned explicitly by Galton (1869); Stigler (1986, chapter 10) provides an absorbing account of this early work on mixtures. Another early reference on mixtures is Holmes (1892), who brought in the concept of mixtures of populations in his suggestion that an average alone was inadequate in consideration of wealth disparity. In another paper predating Pearson's early attempt on mixtures, Newcomb (1886) suggested an iterative reweighting scheme that can be viewed as an application of the EM algorithm of Dempster et al. (1977) to compute the common mean of a mixture in known proportions of a finite number of univariate normal distributions with known variances. The reader is referred to McLachlan & Basford (1988, section 1.2) and McLachlan & Peel (2000a, section 1.1.2) for more discussion and references on the history of mixture models.

But apart from some contributions by Jeffreys (1932) and Rao (1948), the use of maximum likelihood (ML) for fitting mixture models received little attention until the 1960s. Major papers around this time on an iterative scheme for the ML approach to the fitting of mixture distributions were produced by Day (1969) and Wolfe (1970), who also wrote a number of technical reports. However, it was not until Dempster et al. (1977) formalized this iterative scheme in a general context through their expectation–maximization (EM) algorithm that the convergence properties

of the ML solution for the mixture problem were established on a theoretical basis. The EM algorithm proved to be a timely catalyst for further research into the applications of finite mixture models. This can be witnessed by the subsequent stream of papers on finite mixtures in the literature, commencing with, for example, Ganesalingam & McLachlan (1978) and O'Neill (1978).

2. FORMULATION OF MIXTURE DISTRIBUTION

2.1. Basic Definition

The probability density function, or probability mass function in the discrete case of a finite mixture distribution of a p-dimensional random vector $\boldsymbol{Y}$, takes the form

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

where the mixing proportions π_i are nonnegative and sum to one and where the $f_i(\boldsymbol{y})$ are the component densities. We refer to the $f_i(\boldsymbol{y}; \boldsymbol{\theta}_i)$ as densities, since even if the vector $\boldsymbol{Y}$ is discrete, we can still view the $f_i(\boldsymbol{y})$ as densities by the adoption of counting measure. Typically, the component densities are taken to be known up to a vector $\boldsymbol{\theta}_i$ of parameters. In this case, we can write the mixture density as

$$f(\boldsymbol{y}; \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i \, f_i(\boldsymbol{y}; \boldsymbol{\theta}_i), \qquad 2.$$

where $\boldsymbol{\Psi} = (\boldsymbol{\xi}^\top, \pi_1, \ldots, \pi_{g-1})^\top$ denotes the vector of unknown parameters and where $\boldsymbol{\xi}_i$ consists of the elements of the $\boldsymbol{\theta}_i$ known a priori to be distinct. Here, the superscript $\top$ denotes transposition. In many applications, the component densities $f_i(\boldsymbol{y}; \boldsymbol{\theta}_i)$ are taken to belong to the same parametric family, for example, the multivariate normal.

In some applications, the component densities are taken to be different. A particular case of this, called a nonstandard mixture, is that for which $g = 2$, with one of the component distributions being degenerate, concentrated on a single value. The report of the Panel on Nonstandard Mixtures of Distributions (1989) explores nonstandard mixtures in detail.

In the case of common component densities, $f(\boldsymbol{y}; \boldsymbol{\theta})$, the finite mixture model given by Equation 2 can be generalized to the more general form in which

$$f(\boldsymbol{y}) = f(\boldsymbol{y}; H) = \int f(\boldsymbol{y}; \boldsymbol{\theta}) \, \mathrm{d}H(\boldsymbol{\theta}), \qquad 3.$$

where $H(\cdot)$ is a probability measure on the parameter space.

2.2. Identifiability of Mixture Distributions

A parametric family of densities $f(\boldsymbol{y}; \boldsymbol{\Psi})$ is identifiable if distinct values of the parameter $\boldsymbol{\Psi}$ determine distinct members of the family of densities $\{f(\boldsymbol{y}; \boldsymbol{\Psi}) : \boldsymbol{\Psi} \in \boldsymbol{\Omega}\}$, where $\boldsymbol{\Omega}$ is the specified parameter space. Identifiability for mixture distributions is defined slightly differently. For example, if all the g component densities in Equation 2 belong to the same parametric family, then $f(\boldsymbol{y}; \boldsymbol{\Psi})$ is invariant under the $g!$ permutations of the component labels in $\boldsymbol{\Psi}$.

Let $f(\boldsymbol{y}; \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i f_i(\boldsymbol{y}; \boldsymbol{\theta}_i)$ and $f(\boldsymbol{y}; \boldsymbol{\Psi}^*) = \sum_{i=1}^{g^*} \pi_i^* f_i(\boldsymbol{y}; \boldsymbol{\theta}_i^*)$ be any two members of a parametric family of mixture densities. This class of finite mixtures is said to be identifiable for $\boldsymbol{\Psi} \in \boldsymbol{\Omega}$ if $f(\boldsymbol{y}; \boldsymbol{\Psi}) \equiv f(\boldsymbol{y}; \boldsymbol{\Psi}^*)$ if and only if $g = g^*$ and we can permute the component labels so that $\pi_i = \pi_i^*$ and $f_i(\boldsymbol{y}; \boldsymbol{\theta}_i) = f_i(\boldsymbol{y}; \boldsymbol{\theta}_i^*)$ $(i = 1, \ldots, g)$. Here, $\equiv$ implies equality of the densities for almost all $\boldsymbol{y}_j$ relative to the underlying measure on $\mathbb{R}^p$ for $f(\boldsymbol{y}_j; \boldsymbol{\Psi})$.

The lack of identifiability of $\boldsymbol{\Psi}$ due to the interchanging of component labels is of no concern in practice, as it can be easily overcome by the imposition of an appropriate constraint on $\boldsymbol{\Psi}$. However, it can be a major problem in a Bayesian framework where posterior simulation is used to make inferences from the mixture model. In this context, it is known as the label-switching problem.

Another approach to the identifiability problem is to use an identifying function (Kadane 1974). This is essentially the same as Redner's (1981) approach of using the quotient topological space $\tilde{\boldsymbol{\Omega}}$ obtained by mapping equivalent values of $\boldsymbol{\Psi}$ into a single point. Redner (1981) extended Wald's (1949) results on the consistency of the ML estimator of $\boldsymbol{\Psi}$ by using the quotient topological space $\tilde{\boldsymbol{\Omega}}$.

Titterington et al. (1985, section 3.1) gave a lucid account of the concept of identifiability for mixtures. They pointed out that most finite mixtures of continuous densities are identifiable; an exception is a mixture of uniform densities. Teicher (1960) showed that a finite mixture of Poisson distributions is identifiable, whereas mixtures of binomial distributions are not identifiable if $N < 2g - 1$, where N is the common number of trials in the component binomial distributions. Yakowitz & Spragins (1968) showed that finite mixtures of negative binomial component distributions are identifiable.

3. INTERPRETATION OF MIXTURE MODELS

3.1. Conceptualization

We let $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ denote a random sample of size n, where $\boldsymbol{Y}_j$ is a p-dimensional random vector with probability density function given by Equation 2. The vector of the observed values $\boldsymbol{y}_j$ on the $\boldsymbol{Y}_j$ is denoted by $\boldsymbol{y}_{\text{obs}} = (\boldsymbol{y}_1^\top, \ldots, \boldsymbol{y}_n^\top)^\top$.

An obvious way of generating a random vector $\boldsymbol{Y}_j$ with the g-component mixture density $f(\boldsymbol{y}_j)$ given by Equation 2 is as follows. Let Z_j be a categorical random variable taking on the values $1, \ldots, g$, with probabilities $\pi_1, \ldots, \pi_g$, respectively, and suppose that the conditional density of $\boldsymbol{Y}_j$ given $Z = i$ is $f_i(\boldsymbol{y}_j; \boldsymbol{\theta})$ $(i = 1, \ldots, g)$. Then the marginal density of $\boldsymbol{Y}_j$ is given by $f(\boldsymbol{y}; \boldsymbol{\Psi})$. In this context, the variable Z_j can be thought of as the component label of the vector $\boldsymbol{Y}_j$. It is convenient to work with a g-dimensional label vector $\boldsymbol{Z}_j$ in place of the single categorical variable Z_j, where the ith element of $\boldsymbol{Z}_j$, $Z_{ij} = (\boldsymbol{Z}_j)_i$ is defined to be one or zero, according to whether the component of origin of $\boldsymbol{Y}_j$ in the mixture is equal to i or not $(i = 1, \ldots, g)$. Thus, $\boldsymbol{Z}_j$ is distributed according to a multinomial distribution consisting of one draw on g categories with probabilities $\pi_1, \ldots, \pi_g$; that is,

$$\text{pr}\{\boldsymbol{Z}_j = \boldsymbol{z}_j\} = \pi_1^{z_{1j}} \pi_2^{z_{2j}} \ldots \pi_g^{z_{gj}}, \quad \sum_{i=1}^{g} z_{ij} = 1. \tag{4.}$$

We write $\boldsymbol{Z}_j \sim \text{Mult}_g(1, \boldsymbol{\pi})$, where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)^\top$.

The posterior probability that $\boldsymbol{Y}_j$ has arisen from the ith component of the mixture given $\boldsymbol{Y}_j = \boldsymbol{y}_j$ can, by Bayes' Theorem, be expressed as

$$\tau_i(\boldsymbol{y}_j; \boldsymbol{\Psi}) = \text{pr}\{Z_{ij} = 1 \mid \boldsymbol{Y}_j = \boldsymbol{y}_j\} = \pi_i f_i(\boldsymbol{y}_j; \boldsymbol{\theta}_i)/f(\boldsymbol{y}_j; \boldsymbol{\Psi}) \quad (i = 1, \ldots, g;\ j = 1, \ldots, n). \tag{5.}$$

In the interpretation above of a mixture model, an obvious situation where the g-component mixture model (Equation 1) is directly applicable is where each $\boldsymbol{Y}_j$ is drawn from a population G which consists of g groups, $G_1, \ldots, G_g$, in proportions $\pi_1, \ldots, \pi_g$. If the density of $\boldsymbol{Y}_j$ in group G_i is given by $f_i(\boldsymbol{y}_j; \boldsymbol{\theta}_i)$ for $i = 1, \ldots, g$, then the density of $\boldsymbol{Y}_j$ has the g-component mixture form,

as in Equation 1. In this situation, the g components of the mixture can be physically identified with the g externally existing groups, $G_1, \ldots, G_g$.

However, there are also many examples involving the use of mixture models in which the components cannot be identified with externally existing groups as above. In some instances, the components are introduced into the mixture model to allow for greater flexibility in modeling a heterogeneous population that is apparently unable to be modeled by a single component distribution. At the extreme end of this exercise, we obtain the nonparametric kernel estimate of a density if we fit a mixture of $g = n$ components in equal proportions $1/n$, where n is the size of the observed sample.

Thus, it can be seen that mixture models occupy an interesting niche between parametric and nonparametric approaches to statistical estimation. Mixture model–based approaches are parametric in the sense that parametric forms $f_i(\boldsymbol{y}_j; \boldsymbol{\theta}_i)$ may be specified for the component densities, but they can also be regarded as nonparametric by allowing the number of components g to grow.

3.2. Density Estimation

The normal mixture model given by Equation 2 with normal components can be used to estimate an unknown density function. This is because the set of all normal mixture densities is dense in the set of all density functions under the L1 metric (see Li & Barron 1999). Roeder & Wasserman (1997) showed that when a normal mixture model is used to estimate a density nonparametrically, the density estimate that uses the Bayesian information criterion (BIC) of Schwarz (1978) to select the number of components in the mixture is consistent (see also Leroux 1992). The criterion of BIC is defined to be twice the negative of the log likelihood penalized by the addition of $\log n$ times the number of unknown parameters.

In the case where the number of components g is known along with the component labels specified by $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$, this statistical learning problem would be termed supervised classification in the machine learning literature. In the present context, we have the unsupervised classification problem where the indicator variables $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ are unknown and where the number of components g may also be unknown.

3.3. Role of Mixture Models in Clustering of Independent and Identically Distributed Data

The mixture model given by Equation 2 can be used to provide a model-based approach to clustering of the observed data in $\boldsymbol{y}_{\text{obs}}$ by conceptualizing that $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ come from a mixture in proportions $\pi_1, \ldots, \pi_g$ of g groups $G_1, \ldots, G_g$, in which $\boldsymbol{Y}_j$ has density $f_1(\boldsymbol{y}_j; \boldsymbol{\theta}_1), \ldots, f_g(\boldsymbol{y}_j; \boldsymbol{\theta}_g)$, respectively. This is irrespective of whether these groups do externally exist.

For clustering purposes, each component in the mixture model given by Equation 2 is usually taken to correspond to a cluster. The posterior probability that the jth observation $\boldsymbol{y}_j$ arose from group G_i is given by the fitted posterior probability,

$$\tau_i(\boldsymbol{y}_j; \hat{\boldsymbol{\Psi}}) = \hat{\pi}_i f_i(\boldsymbol{y}_j; \hat{\boldsymbol{\theta}}_i)/f(\boldsymbol{y}_j; \hat{\boldsymbol{\Psi}}) \quad (i = 1, \ldots, g;\ j = 1, \ldots, n), \qquad 6.$$

where $\hat{\boldsymbol{\Psi}}$ denotes an estimate of $\boldsymbol{\Psi}$.

A probabilistic clustering of the data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ into g clusters can be obtained in terms of the fitted posterior probabilities of component membership $\tau_i(\boldsymbol{y}_j; \hat{\boldsymbol{\Psi}})\,(i = 1, \ldots, g)$.

An outright partitioning of the observations into g nonoverlapping clusters $C_1, \ldots, C_g$ is effected by assigning each $\boldsymbol{y}_j$ to the group G_i to which it has the highest estimated posterior

probability of belonging. That is, the ith cluster C_i contains those observations $\boldsymbol{y}_j$ with $\hat{z}_{ij} = (\hat{\boldsymbol{z}}_j)_i = 1$, where

$$\begin{aligned} \hat{z}_{ij} &= 1, \qquad \text{if } i = \arg\max_h \hat{\tau}_h(\boldsymbol{y}_j;\, \hat{\boldsymbol{\Psi}}), \\ &= 0, \qquad \text{otherwise.} \end{aligned} \tag{7.}$$

As the notation implies, $\hat{z}_{ij}$ can be viewed as an estimate of z_{ij} which, under the assumption that the observations come from a mixture of g groups $G_1, \ldots, G_g$, is defined to be one or zero according to whether $\boldsymbol{y}_j$ did or did not arise from G_i ($i = 1, \ldots, g$).

The above rule for assigning the unclassified data points $\boldsymbol{y}_j$ to the g groups corresponds to the Bayes rule of allocation in the supervised classified case with known parameter vector $\boldsymbol{\Psi}$ (McLachlan 1992, section 1.3).

4. ESTIMATION OF MIXTURE DISTRIBUTIONS

4.1. Method of Moments

As noted in Section 1.2, one of the first major analyses involving mixture models was undertaken by Pearson (1894), who used the method of moments to fit a mixture of two normal distributions with different means and different variances. It required the solving of a nonic polynomial. In spite of this, the method of moments remained popular until the advent of the EM algorithm that facilitated the computation of ML estimates. However, work by Lindsay and Furman revived interest in moment estimates in certain contexts, particularly in the case of mixtures of normal densities with equal variances (see Furman & Lindsay 1994a,b, and the references therein).

4.2. Maximum Likelihood Estimation

As remarked above, since the advent of the EM algorithm, ML has been by far the most commonly used approach to the fitting of mixture distributions. The application of the EM algorithm for the computation of the ML estimates for parametric mixture models is considered in the next section.

Increasing attention is being given to the use of the minorization–maximization (MM) algorithm (Hunter & Lange 2004, Lange 2013) for the computation of ML estimates for mixtures in situations where the E-step is not straightforward, as in Nguyen & McLachlan (2016a,b).

The so-called nonparametric ML estimation of the mixing distribution H in Equation 3 has also attracted attention, as reviewed in McLachlan (2016) (see also Hall & Zhou 2003, Chen 2017). In spite of the potential generality of the mixing distribution H, there is a likelihood-maximizing measure that is concentrated on a support of cardinality that is at most that of the set of distinct data points. In other words, a finite mixture maximizes the likelihood (see Lindsay 1995 and the references therein).

5. APPLICATION OF EXPECTATION–MAXIMIZATION ALGORITHM TO FINITE MIXTURES

Here, we describe the implementation of the EM algorithm for ML estimation of mixture distributions. As remarked earlier, the EM algorithm greatly stimulated interest in the use of finite mixture distributions to model heterogeneous data. This is because the fitting of mixture models by ML is a classic example of a problem that is simplified considerably by the EM's conceptual unification of ML estimation from data that can be viewed as being incomplete. Indeed, almost

all the post-1977 applications of mixture modeling reported in the books on mixtures use the EM algorithm.

The ML estimate of $\boldsymbol{\Psi}$, $\hat{\boldsymbol{\Psi}}$ is given by an appropriate root of the likelihood equation,

$$\partial \log L(\boldsymbol{\Psi})/\partial \boldsymbol{\Psi} = \mathbf{0}, \qquad 8.$$

where $L(\boldsymbol{\Psi})$ denotes the likelihood function for $\boldsymbol{\Psi}$ formed from the observed data $\boldsymbol{y}_{\text{obs}}$,

$$\log L(\boldsymbol{\Psi}) = \sum_{j=1}^{n} \log \left\{ \sum_{i=1}^{g} \pi_i f_i(\boldsymbol{y}_j; \boldsymbol{\theta}_i) \right\}. \qquad 9.$$

Solutions of Equation 8 corresponding to local maximizers of $\log L(\boldsymbol{\Psi})$ can be obtained via the EM algorithm of Dempster et al. (1977) (see also McLachlan & Krishnan 2008).

5.1. Specification of Expectation–Maximization Framework

It is straightforward, at least in principle, to find solutions of Equation 8 using the EM algorithm. It is easy to program and proceeds iteratively in two steps, E (for expectation) and M (for maximization).

In the EM framework for this mixture problem, the observed data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are regarded as being incomplete. Each $\boldsymbol{y}_j$ is conceptualized to have arisen from one of the component distributions of the mixture model to be fitted with $z_{ij} = (\boldsymbol{z}_j)_i$ equal to one or zero according to whether $\boldsymbol{y}_j$ has arisen or not from the ith component distribution. This is irrespective of how the observed data $\boldsymbol{y}_{\text{obs}}$ were generated. The distribution of the random vector $\boldsymbol{Z}_j$ corresponding to z_j is specified by Equation 4.

The complete-data $\boldsymbol{y}_{\text{comp}}$ vector is taken to be

$$\boldsymbol{y}_{\text{comp}} = (\boldsymbol{y}_{\text{obs}}^{\top}, \boldsymbol{z}^{\top})^{\top}, \qquad 10.$$

where $\boldsymbol{z} = (\boldsymbol{z}_i^{\top}, \ldots, \boldsymbol{z}_n^{\top})^{\top}$. For independent data, it is appropriate to assume that the random indicator-vectors $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ corresponding to $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$, are distributed according to the multinomial distribution

$$\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n \overset{\text{i.i.d.}}{\sim} \text{Mult}_g(1, \boldsymbol{\pi}). \qquad 11.$$

For the specification of the complete-data vector given by Equation 10, the complete-data log likelihood for $\boldsymbol{\Psi}$, $\log L_{\text{c}}(\boldsymbol{\Psi})$, is given by

$$\log L_{\text{c}}(\boldsymbol{\Psi}) = \sum_{i=1}^{g} \sum_{j=1}^{n} z_{ij} \{\log \pi_i + \log f_i(\boldsymbol{y}_j; \boldsymbol{\theta}_i)\}. \qquad 12.$$

5.2. E-step

The addition of the unobservable data of the indicator variables in $\boldsymbol{z}$ to the problem is handled by the E-step, which takes the conditional expectation of the complete-data log likelihood, $\log L_{\text{c}}(\boldsymbol{\Psi})$, given the observed data $\boldsymbol{y}_{\text{obs}} = (\boldsymbol{y}_1^{\top}, \ldots, \boldsymbol{y}_n^{\top})^{\top}$, using the current fit for $\boldsymbol{\Psi}$. Let $\boldsymbol{\Psi}^{(0)}$ be the value specified initially for $\boldsymbol{\Psi}$. Then, on the first iteration of the EM algorithm, the E-step requires the computation of the conditional expectation of $\log L_{\text{c}}(\boldsymbol{\Psi})$ given $\boldsymbol{y}_{\text{obs}}$, using $\boldsymbol{\Psi}^{(0)}$ for $\boldsymbol{\Psi}$, which can be written as

$$Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(0)}) = \text{E}_{\boldsymbol{\Psi}^{(0)}}\{\log L_{\text{c}}(\boldsymbol{\Psi}) \mid \boldsymbol{y}_{\text{obs}}\}. \qquad 13.$$

The expectation operator E has the subscript $\boldsymbol{\Psi}^{(0)}$ to explicitly convey that this expectation is being effected using $\boldsymbol{\Psi}^{(0)}$ for $\boldsymbol{\Psi}$.

It follows that on the $(k+1)$th iteration, the E-step requires the calculation of $Q(\boldsymbol{\Psi};\ \boldsymbol{\Psi}^{(k)})$, where $\boldsymbol{\Psi}^{(k)}$ is the value of $\boldsymbol{\Psi}$ after the kth EM iteration. As the complete-data log likelihood, $\log L_c(\boldsymbol{\Psi})$, is linear in the unobservable data z_{ij}, the E-step [on the $(k+1)$th iteration] simply requires the calculation of the current conditional expectation of Z_{ij} given the observation $\boldsymbol{y}_j$, where Z_{ij} is the random variable corresponding to z_{ij}. Now

$$\mathrm{E}_{\boldsymbol{\Psi}^{(k)}}(Z_{ij} \mid \boldsymbol{y}_j) = \mathrm{pr}_{\boldsymbol{\Psi}^{(k)}}\{Z_{ij} = 1 \mid \boldsymbol{y}_j\} = \tau_i(\boldsymbol{y}_j;\ \boldsymbol{\Psi}^{(k)}), \tag{14.}$$

where, corresponding to Equation 6,

$$\tau_i(\boldsymbol{y}_j;\ \boldsymbol{\Psi}^{(k)}) = \pi_i^{(k)} f_i(\boldsymbol{y}_j;\ \boldsymbol{\theta}_i^{(k)})/f(\boldsymbol{y}_j;\ \boldsymbol{\Psi}^{(k)}) \tag{15.}$$

for $i = 1, \ldots, g;\ j = 1, \ldots, n$. Using Equation 14, we have, after taking the conditional expectation with $\boldsymbol{\Psi} = \boldsymbol{\Psi}^{(k)}$ of Equation 12 given $\boldsymbol{y}_{\mathrm{obs}}$, that

$$Q(\boldsymbol{\Psi};\ \boldsymbol{\Psi}^{(k)}) = \sum_{i=1}^{g}\sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j;\ \boldsymbol{\Psi}^{(k)})\{\log \pi_i + \log f_i(\boldsymbol{y}_j;\ \boldsymbol{\theta}_i)\}. \tag{16.}$$

By using soft allocation through the use of the fractional values of the posterior probabilities $\tau_i(\boldsymbol{y}_j;\ \boldsymbol{\Psi}^{(k)})$, one avoids the biases that are incurred with a hard allocation where the observations $\boldsymbol{y}_j$ are assigned outright to the components on each iteration. This latter procedure corresponds to the so-called classification ML approach (see, for example, McLachlan 1975, 1982).

5.3. M-step

The M-step on the $(k+1)$th iteration requires the global maximization of $Q(\boldsymbol{\Psi};\ \boldsymbol{\Psi}^{(k)})$ with respect to $\boldsymbol{\Psi}$ over the parameter space $\boldsymbol{\Omega}$ to give the updated estimate $\boldsymbol{\Psi}^{(k+1)}$. For the finite mixture model, the updated estimates $\pi_i^{(k+1)}$ of the mixing proportions π_i are calculated independently of the updated estimate $\boldsymbol{\xi}^{(k+1)}$ of the parameter vector $\boldsymbol{\xi}$ containing the unknown parameters in the component densities.

If the z_{ij} were observable, then the complete-data ML estimate of π_i would be given simply by

$$\hat{\pi}_i = \sum_{j=1}^{n} z_{ij}/n \qquad (i = 1, \ldots, g). \tag{17.}$$

As the E-step simply involves replacing each z_{ij} with its current conditional expectation $\tau_i(\boldsymbol{y}_j;\ \boldsymbol{\Psi}^{(k)})$ in the complete-data log likelihood, the updated estimate of π_i is given by replacing each z_{ij} in Equation 6 by $\tau_i(\boldsymbol{y}_j;\ \boldsymbol{\Psi}^{(k)})$ to give

$$\pi_i^{(k+1)} = \sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j;\ \boldsymbol{\Psi}^{(k)})/n \qquad (i = 1, \ldots, g). \tag{18.}$$

Thus, in forming the estimate of π_i on the $(k+1)$th iteration, there is a contribution from each observation $\boldsymbol{y}_j$ equal to its (currently assessed) posterior probability of membership of the ith component of the mixture model.

Concerning the updating of $\boldsymbol{\xi}$ on the M-step of the $(k+1)$th iteration, it can be seen from Equation 16 that $\boldsymbol{\xi}^{(k+1)}$ is obtained as an appropriate root of

$$\sum_{i=1}^{g}\sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j;\ \boldsymbol{\Psi}^{(k)})\partial \log f_i(\boldsymbol{y}_j;\ \boldsymbol{\theta}_i)/\partial \boldsymbol{\xi} = \mathbf{0}. \tag{19.}$$

One nice feature of the EM algorithm is that the solution of Equation 19 often exists in closed form as with the normal mixture model.

The E- and M-steps are alternated repeatedly until the difference $\log L(\boldsymbol{\Psi}^{(k+1)}) - \log L(\boldsymbol{\Psi}^{(k)})$ changes by an arbitrarily small amount in the case of convergence of the sequence of likelihood values $\{L(\boldsymbol{\Psi}^{(k)})\}$. Dempster et al. (1977) showed that the likelihood function $L(\boldsymbol{\Psi})$ is not decreased after an EM iteration; that is,

$$L(\boldsymbol{\Psi}^{(k+1)}) \geq L(\boldsymbol{\Psi}^{(k)}) \qquad 20.$$

for $k = 0, 1, 2, \ldots$. Hence, convergence must be obtained with a sequence of likelihood values $\{L(\boldsymbol{\Psi}^{(k)})\}$ that are bounded above. In almost all cases, the limiting value L^* is a local maximum. A detailed account of the convergence properties of the EM algorithm in a general setting has been given by Wu (1983), who addressed, in particular, the problem that the convergence of $L(\boldsymbol{\Psi}^{(k)})$ to L^* does not automatically imply the convergence of $\boldsymbol{\Psi}^{(k)}$ to a point $\boldsymbol{\Psi}^*$; McLachlan & Krishnan (2008) provide further details.

Let $\hat{\boldsymbol{\Psi}}$ be the chosen solution of the likelihood equation. For an observed sample, $\hat{\boldsymbol{\Psi}}$ is usually taken to be the root of Equation 8 corresponding to the local maximizer at which the likelihood is largest. That is, in those cases where $L(\boldsymbol{\Psi})$ has a global maximum in the interior of the parameter space, $\hat{\boldsymbol{\Psi}}$ is the global maximizer, assuming that the global maximum has been located.

The EM algorithm needs to be started from a variety of initial values for the parameter vector $\boldsymbol{\Psi}$ or for a variety of initial partitions of the data into g groups. The latter can be obtained by randomly dividing the data into g groups corresponding to the g components of the mixture model and estimating the component parameters as if these g groups provide a perfect segmentation of the data with respect to the g components. With random starts performed in this manner, the effect of the central limit theorem tends to have these initial estimates of the component parameters being similar, at least in large samples. Nonrandom partitions of the data can be obtained via some clustering procedure such as k-means. Also, Coleman et al. (1999) proposed some procedures for obtaining nonrandom starting partitions.

6. BAYESIAN ANALYSIS

We consider here the case of a proper prior density $p(\boldsymbol{\Psi})$ for the parameter vector $\boldsymbol{\Psi}$. In the following, we shall use $p(\cdot)$ as a generic notation for a density function. We can write the posterior density of $\boldsymbol{\Psi}$ as

$$p(\boldsymbol{\Psi} \mid \boldsymbol{y}) = C^{-1} L(\boldsymbol{\Psi}) p(\boldsymbol{\Psi}) = C^{-1} \sum_{\boldsymbol{z}} L_{\mathrm{c}}(\boldsymbol{\Psi}) p(\boldsymbol{z} \mid \boldsymbol{\Psi}) p(\boldsymbol{\Psi}), \qquad 21.$$

where $p(\boldsymbol{z} \mid \boldsymbol{\Psi})$ denotes the conditional density of $\boldsymbol{Z}$ given $\boldsymbol{\Psi}$. The normalizing constant C in Equation 21 is given by

$$C = \int \sum_{\boldsymbol{z}} L_{\mathrm{c}}(\boldsymbol{\Psi}) p(\boldsymbol{z} \mid \boldsymbol{\Psi}) p(\boldsymbol{\Psi}) \, \mathrm{d}\boldsymbol{\Psi}. \qquad 22.$$

If a conjugate prior is specified, then the posterior density of $\boldsymbol{\Psi}$ can be written in closed form.

In Equation 22, the sum is over all possible values of $\boldsymbol{z}$ defining the component membership of $\boldsymbol{y}_j$ ($j = 1, \ldots, n$) with respect to the g components. Thus, its direct use is only feasible with small sample sizes. We can approximate posterior quantities of interest through the use of Markov chain Monte Carlo (MCMC) methods. Such methods allow the construction of an ergodic Markov chain with stationary distribution equal to the posterior distribution of the parameter of interest, here $\boldsymbol{\Psi}$, containing the parameters in the mixture model. Gibbs sampling achieves this by simulating directly from the conditional distribution of a subvector of $\boldsymbol{\Psi}$ given all the other parameters in $\boldsymbol{\Psi}$ (and $\boldsymbol{y}_{\mathrm{obs}}$).

The unobservable indicator-vector $\boldsymbol{z}$ is introduced, and $\boldsymbol{\Psi}$ is augmented by $\boldsymbol{z}$ during the Gibbs sampling. Thus, samples for the missing-data vector $\boldsymbol{z}$ and the parameter vector $\boldsymbol{\Psi}$ are alternately generated, producing a missing-data chain and a parameter chain. Among other sampling methods, there is the Metropolis–Hastings algorithm, which, in contrast to the Gibbs sampler, simulates from a convenient proposal distribution and then accepts the proposed value with some defined probability. Diebolt & Robert (1994) took an approach based on the data augmentation ideas of Tanner & Wong (1987) (see also Lavine & West 1992, Escobar & West 1995). A key feature of Lavine & West (1992) is the use of the Dirichlet hyperpriors on the prior of $\boldsymbol{\pi}$ in contexts where the number of components is not specified. The number of components eventually chosen is dictated by the values of the hyperparameters. Robert (1996) provides a useful survey of the Bayesian approach to mixtures; see also Mengersen et al. (2011).

7. NORMAL MIXTURES

7.1. Invariance of Component Distributions

Frequently, in practice, the clusters in the case of Euclidean 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. With the application of the EM algorithm, the updates of the component means and covariance matrices exist in closed form on each M-step.

One attractive feature of adopting mixture models with elliptically symmetric components such as the normal is that the implied clustering is invariant under affine transformations of the data, that is, invariant under transformations of the vector $\boldsymbol{y}$ of the form, $\boldsymbol{y} \rightarrow \boldsymbol{C}\boldsymbol{y} + \boldsymbol{a}$, where $\boldsymbol{C}$ is a nonsingular matrix. If the clustering of a procedure is invariant under only diagonal $\boldsymbol{C}$, then it is invariant under change of measuring units but not rotations. But, as commented upon by Hartigan (1975), this form of invariance is more compelling than affine invariance.

7.2. Restrictions on Covariance Matrices

In practice, consideration has to be given to the problem of relatively large local maxima that occur as a consequence of a fitted component having a very small (but nonzero) estimate of the variance for univariate data, or estimated generalized variance (that is, the determinant of the estimated covariance matrix) for multivariate data. Such a component corresponds to a cluster containing a few data points either relatively close together or, in the case of multivariate data, almost lying in a lower-dimensional subspace. There is thus a need to monitor the relative size of the fitted mixing proportions and of the component generalized variances to identify these spurious local maximizers and to avoid the EM sequence not converging at all if the likelihood is unbounded, as in the case of unrestricted component-covariance matrices. Another approach is to constrain the generalized variances of the component-covariance matrices, or equivalently their eigenvalues, as reviewed by García-Escudero et al. (2018).

Under the homoscedasticity assumption of equal $\boldsymbol{\Sigma}_i$, the likelihood function $L(\boldsymbol{\Psi})$ will be bounded. A further simplification is to take the common component-covariance matrix $\boldsymbol{\Sigma}$ to be spherical; that is, $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$, where $\boldsymbol{I}$ denotes the $p \times p$ identity matrix. In this case, the normal mixture model is no longer invariant under change of scale. The latter constraint means that the clusters produced tend to be spherical in shape. If we also take the mixing proportions to be in equal proportions $1/g$, then it is equivalent to a soft version of k-means clustering.

7.3. Mixtures of Factor Analyzers

The normal mixture model with unrestricted component-covariance matrices in its normal component distributions is a highly parameterized one, with $\frac{1}{2}p(p+1)$ parameters for each component-covariance matrix $\boldsymbol{\Sigma}_i$ $(i = 1, \ldots, g)$. As an alternative to taking the component-covariance matrices to be the same or diagonal, one might wish to adopt some model for the component-covariance matrices that is intermediate between homoscedasticity and the unrestricted model. To this end, Banfield & Raftery (1993) introduced a parameterization of the component-covariance matrices $\boldsymbol{\Sigma}_i$ based on a variant of the standard spectral decomposition of $\boldsymbol{\Sigma}_i$ $(i = 1, \ldots, g)$ (see also Fraley & Raftery 2002). However, if p is large relative to the sample size n, it may not be possible to use this decomposition to infer an appropriate model for the component-covariance matrices. Even if it is possible, the results may not be reliable due to potential problems with near-singular estimates of the component-covariance matrices when p is large relative to n.

A common approach to reducing the number of dimensions is to perform a principal component analysis (PCA). But, as is well known, projections of the data $\boldsymbol{y}_j$ onto the first few principal axes are not always useful in portraying the group structure. 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 & Peel (2000b). This model was originally proposed by Ghahramani & Hinton (1997) and Hinton et al. (1997). With the mixture of factor analyzers model, the ith component-covariance matrix $\boldsymbol{\Sigma}_i$ has the form $\boldsymbol{\Sigma}_i = \boldsymbol{B}_i + \boldsymbol{D}_i$ $(i = 1, \ldots, g)$, where $\boldsymbol{B}_i$ is a $p \times q$ matrix of factor loadings and $\boldsymbol{D}_i$ is a diagonal matrix. It assumes that the component correlations between the observations can be explained by the conditional linear dependence of the latter on q latent or unobservable variables specific to the given component. Unlike the PCA model, the factor analysis model enjoys a powerful invariance property: Changes in the scales of the variables in $\boldsymbol{y}$ appear only as scale changes in the appropriate rows of the matrix $\boldsymbol{B}_i$ of factor loadings. If the number of factors q is chosen to be sufficiently smaller than p, the factor-analytic representation of the component-covariance matrices reduces the number of free parameters to be estimated. The mixtures of factor analyzers model can be fitted using the alternating expectation–conditional maximization (Meng & van Dyk 1997).

In practice, consideration has to be given to the number of factors q in the mixture of factor analyzers model. One obvious approach is to use BIC. An alternative approach is to use the likelihood-ratio test statistic (LRTS). For tests on g, it is well known that regularity conditions do not hold for the usual chi-squared approximation to the asymptotic null distribution of the LRTS to be valid. This is also the case for tests on q at a given level of g (see Drton & Plummer 2017).

Baek et al. (2010) considered how this factor-analytic approach can be modified to provide a greater reduction in the number of parameters. They termed their approach mixtures of common factor analyzers because the matrix of factor loadings is common to the components before the component-specific rotation of the component factors to make them white noise (see also Montanari & Viroli 2010, Viroli 2010).

García-Escudero et al. (2016) considered the joint role of trimming and constraints in robust estimation for mixtures of normal factor analyzers. More recently, Viroli & McLachlan (2017) considered deep normal mixture models by introducing layers of factors into the model.

7.4. High-Dimensional Data

In situations where the sample size n is small relative to the dimension p, it might not be practical to fit mixtures of factor analyzers, as it would involve a considerable amount of computation time. Thus, initially, some of the variables in the observation vector may have to be removed. Indeed,

the simultaneous use of too many variables in the cluster analysis may serve only to create noise that masks the effect of a smaller number of variables. Also, the intent of the cluster analysis may not be to produce a clustering of the observations on the basis of all the available variables, but rather to discover and study different clusterings of the observations corresponding to different subsets of the variables.

Therefore, McLachlan et al. (2002) developed the so-called EMMIX-GENE procedure, which has two optional steps before the final step of clustering the observations. The first step considers the selection of a subset of relevant variables from the available set of variables by screening the variables on an individual basis to eliminate those that are of little use in clustering the observations. Even after this step has been completed, too many variables may still remain. Thus, there is a second step in EMMIX-GENE, in which the retained variables are clustered (after standardization) into a number of groups on the basis of Euclidean distance so that variables with similar profiles are put into the same group.

Another way to proceed with the fitting of mixture models to high-dimensional data is to use a penalized approach, as adopted by Pan & Shen (2007) and Zhou & Pan (2009). Recently, Witten & Tibshirani (2010) provided a framework for variable selection in a clustering context; readers are also directed to the references therein, including Raftery & Dean (2006) and Maugis et al. (2009), who considered the variable selection problem in terms of model selection. Xie et al. (2010) considered a penalized version of mixtures of factor analyzers.

8. EXTENSIONS OF NORMAL MIXTURE MODELS

8.1. *t*-Mixtures

McLachlan & Peel (1998) first suggested the use of mixtures of t-distributions to provide a robust extension to mixtures of normals (see also Peel & McLachlan 2000). The t-distribution adopted for the ith component density of $\boldsymbol{Y}_j$ can be characterized as being distributed $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i/u_j)$, given the realization u_j of the latent random variable U_j having a two-parameter gamma distribution with equal parameters $\nu_i/2$ $(i = 1, \ldots, g)$. The value u_j is declared to be missing, as well as the component-indicator variables $\boldsymbol{z}_j$ in the EM framework. McLachlan & Peel (1998) 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 & Rubin (1993) replaces the M-step of the EM algorithm by a number of computationally simpler conditional maximization steps. 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. McLachlan et al. (2007) and Baek & McLachlan (2011) considered mixtures of t-factor and t-common factor analyzers, respectively.

8.2. Some Other Robust Mixture Models

Coretto & Hennig (2017) proposed the optimally tuned robust improper ML estimator (OTRIMLE) for robust clustering based on the multivariate normal model for clusters. It is inspired by the addition of a uniform noise component to a normal mixture (Banfield & Raftery 1993). The OTRIMLE uses an improper constant density for modeling outliers and noise. This is chosen optimally so that the nonnoise part of the data looks as close to a Gaussian mixture as possible.

García-Escudero et al. (2008) proposed a general robust approach to robust cluster analysis via trimming, which they considered further in the context of mixture models in a series of papers as referenced in García-Escudero et al. (2018).

8.3. Skew Normal/*t*-Mixture Models

The discussions so far have been focusing on normal component densities. However, in many applications, for example, finance and cytometry, the clusters within the data are asymmetric and exhibit other nonnormal features. In recent years, substantial progress has been made in the literature to explore the use of nonnormal distributions for mixture models. In particular, the asymmetric or skew distributions have received considerable attention, from the classical skew normal distribution by Azzalini & Dalla Valle (1996) to various different characterizations and generalizations (Arellano-Valle & Azzalini 2006). These distributions have an additional vector/matrix of parameters compared with their symmetric counterparts for regulating the skewness of their densities. For example, the classical skew normal distribution has a p-dimensional skewness vector $\boldsymbol{\delta}$. Its density, after reparametrization, can be expressed as

$$f(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\delta}) = 2\phi_p(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Omega})\,\Phi_1\left(\boldsymbol{\delta}^\top\Omega^{-1}(\boldsymbol{y}-\boldsymbol{\mu});0,\lambda\right), \qquad 23.$$

where $\boldsymbol{\Omega} = \boldsymbol{\Sigma}+\boldsymbol{\delta}\boldsymbol{\delta}^\top$ and $\lambda = 1-\boldsymbol{\delta}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{\delta}$. A skew t analog of the skew normal density (Equation 23) was adopted in the mixture models considered by Pyne et al. (2009) and Cabral et al. (2012), which is equivalent to the classical skew t-distribution by Azzalini & Capitanio (2003) (see Lee & McLachlan 2013, 2014). The skew normal distribution (Equation 23) and its skew t-analog, along with other instances of the (restricted) skew elliptical class, are suitable for modeling data where skewness is concentrated along a single direction in the sample space (McLachlan & Lee 2016).

A more general class of skew distribution is the canonical fundamental skew distributions (Arellano-Valle & Genton 2005), which include the canonical fundamental skew t (CFUST) distribution recently adopted by Lee & McLachlan (2016) for their mixture model. This latter distribution has a $p \times q$ matrix of skewness parameters $\boldsymbol{\Delta}$, allowing it to model skewness along multiple directions simultaneously. Its density can be expressed as

$$f(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\Delta},\nu) = 2^q t_p(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Omega},\nu)\,T_q\left(\boldsymbol{\Delta}^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\sqrt{\frac{\nu+p}{\nu+d(\boldsymbol{y})}};\boldsymbol{0},\boldsymbol{\Lambda},\nu+p\right), \qquad 24.$$

where $\boldsymbol{\Omega} = \boldsymbol{\Sigma}+\boldsymbol{\Delta}\boldsymbol{\Delta}^\top$, $d(\boldsymbol{y}) = (\boldsymbol{y}-\boldsymbol{\mu})^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})$, and $\boldsymbol{\Lambda} = \boldsymbol{I}_q - \boldsymbol{\Delta}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{\Delta}$.

8.4. Examples

To illustrate the fits provided by the aforementioned variants of the normal mixture model, we consider the well-known *Iris* data set collected by Anderson (1935) and first analyzed by Fisher (1936). The data set consists of measurements of the length and width of the sepals and petals of 150 *Iris* plants (comprising 50 samples from each of the three species *setosa*, *versicolor*, and *virginica*). The data can be clustered by fitting a three-component normal mixture model (**Figure 1*a***), achieving a misclassification rate (MCR) of 0.033 (corresponding to the misclassification of five *versicolor* observations). Adopting the more flexible skew normal or skew t-component densities (see **Figure 1*b*** and ***c***, respectively) provides improved clustering results, with the latter model achieving a MCR of 0.0067 (corresponding to one misclassified *versicolor* observation).

In cytometric data analysis, mixture distributions have been (implicitly or explicitly) used by many computational algorithms to model different cell populations within the data. An example analyzed by Lee et al. (2018) is given in **Figure 2**, which shows a direct application of a normal mixture model and a skew t-mixture model to segment a hematopoietic stem cell transplant sample of approximately 6,000 cells into four clusters. As is typical for cytometric data, the clusters often exhibit nonnormal characteristics. In this case, the skew t-mixture model provides a clustering in

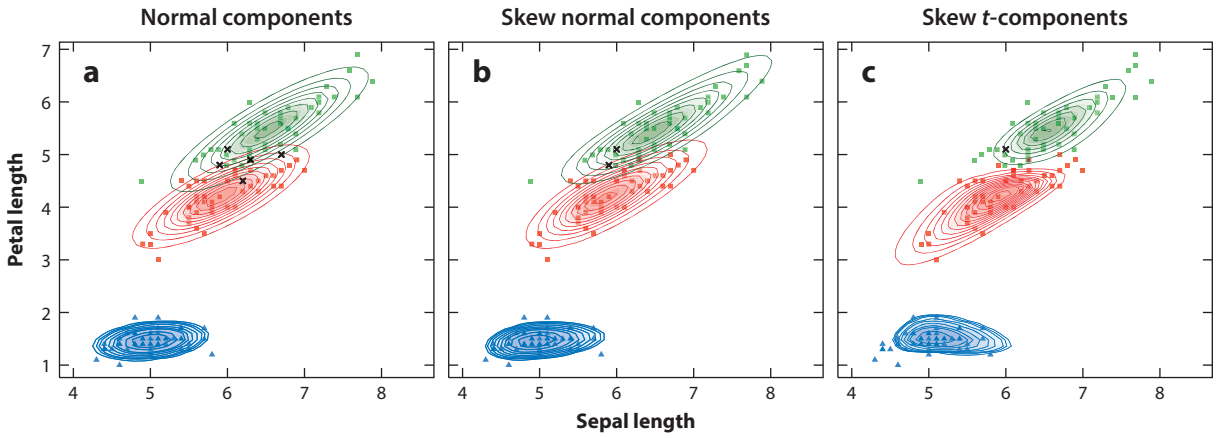

Figure 1

Clustering the *Iris* data set with mixture models; bivariate plots of the variables sepal length and petal length are shown. Dots are colored and shaped according to the cluster labels given by the mixture model, with misclassified observations shown as black crosses. Contours of the fitted three-component mixture models with (*a*) normal components, (*b*) skew normal components, and (*c*) skew *t*-components are overlaid on the scatter plots.

closer agreement with the manual analysis, achieving a MCR of 0.00277 versus 0.0041 for the normal mixture model.

8.5. Some Other Nonnormal Components

Apart from the above-mentioned mixture components of skew normal and skew *t*-distributions, particular instances of the family of generalized hyperbolic distributions of Barndorff-Nielsen (1977) have been considered as component distributions, including the normal inverse Gaussian distribution (Karlis & Santourian 2009), the asymmetric Laplace distribution (Franczak et al.

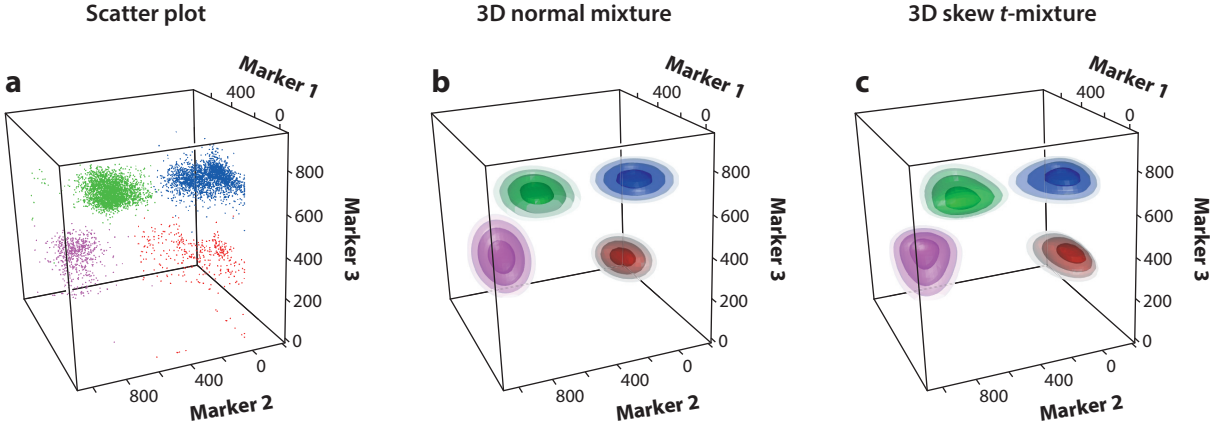

Figure 2

Automated segmentation of cell populations from a hematopoietic stem cell transplant sample using mixture models, showing (*a*) scatter plot of the four cell populations identified by manual gating, (*b*) 3D contours of the fitted four-component normal mixture model, and (*c*) 3D contours of the fitted four-component skew *t*-mixture model.

2014), and a particular generalized hyperbolic distribution (Browne & McNicholas 2015). Furthermore, Wraith & Forbes (2015) studied a multiple-scaled version of some of these distributions and Spurek (2017) considered a general split Gaussian distribution. Other approaches for dealing with skewness in the component distributions have been considered, including data transformation such as the Box-Cox transformation (Lo & Gottardo 2012) and the Manly transformation (Zhu & Melnykov 2018).

More recently, some factor-analytic analogs of some of the above-mentioned asymmetric distributions have been considered. These include, for example, mixtures of skew normal factor analyzers (Lin et al. 2016), mixtures of skew t-factor analyzers (Murray et al. 2014, Lin et al. 2015), and mixtures of generalized hyperbolic factor analyzers (Tortora et al. 2016).

9. CLUSTERING OF DEPENDENT DATA

9.1. Hidden Markov Models

The hidden Markov model (HMM) is increasingly being adopted in applications because it provides a convenient way of formulating an extension of a mixture model to allow for dependent data. To see this, it follows from Equation 12 that we can write the complete-data likelihood function for $\boldsymbol{\Psi}$ in the case of independent data $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ with marginal density Equation 2 as

$$L_c(\boldsymbol{\Psi}) = \left\{ \prod_{j=1}^{n} \prod_{i=1}^{g} f_i(\boldsymbol{y}_j; \boldsymbol{\Psi})^{z_{ij}} \right\} \times \left\{ \prod_{j=1}^{n} f(\boldsymbol{z}_j) \right\}, \qquad 25.$$

where $f(\boldsymbol{z}_j) = \prod_{i=1}^{g} \pi_i^{z_{ij}}$.

Still assuming that the vectors $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ are conditionally independent given $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$, dependence between them is introduced by taking their component-indicator variables $\boldsymbol{Z}_j$ to be dependent. Usually, a stationary Markovian model is formulated for the distribution of the hidden vectors $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$. In one dimension, this Markovian model is a Markov chain and, in two and higher dimensions, it is a Markov random field (see Besag 1986). With the relaxation of the independence assumption for the $\boldsymbol{Z}_j$, the marginal density of $\boldsymbol{Y}_j$ will not have its simple representation (Equation 2) of a mixture density as in the independence case.

The application of the EM algorithm to the hidden Markov chain model is known in the HMM literature as the Baum–Welch algorithm. Baum and his collaborators formulated this algorithm before the appearance of the EM algorithm and established the convergence properties for this algorithm (see Baum & Petrie 1966 and the references in McLachlan & Peel 2000a, chapter 13).

9.2. Mixtures of Time Series

Another application of mixture models where the observed data are not assumed to be independent concerns time series. Recent applications include those by Nguyen et al. (2016, 2017, 2018) on spatial clustering of time series via mixtures of autoregressive models and Markov random fields, maximum pseudolikelihood estimation for a model-based clustering of time-series data, and whole-volume clustering of time series data from zebrafish brain calcium images via mixture model–based functional data analysis, respectively.

9.3. Mixtures of Linear Mixed Models

Ng et al. (2006) have developed the procedure called EMMIX-WIRE (EM-based mixture analysis with random effects) to handle the clustering of correlated data that may be replicated. They

adopted conditionally a mixture of linear mixed models to specify the correlation structure between the variables and to allow for correlations among the observations. It also enables covariate information to be incorporated into the clustering process. In the analysis of gene expressions of thousands of genes from microarray experiments, Ng et al. (2015) applied this procedure to cluster the gene profiles into a small number of clusters from which contrasts were formed for the detection of genes that were differentially expressed between two types of disease.

9.4. Mixtures of Experts

Mixtures of experts models (Jacobs et al. 1991) and their hierarchical extensions are being used widely to improve the flexibility of the regression model with Gaussian errors for modeling nonlinear regression data. In recent work, Nguyen & McLachlan (2016a) demonstrated the robustness of the Laplace mixture of linear experts (MoLE) model over the Gaussian MoLE model, and an application of the Laplace MoLE model to the analysis of a climate science data set is described.

10. NUMBER OF COMPONENTS

10.1. Order of a Mixture Model

Arguably the most obdurate methodological problem associated with mixture distributions is that of identifying the number of components involved in the distribution underlying a set of data. In a cluster analysis, the choice of the number of components arises with the question of how many clusters there are in the data. Additional references may be found in the review by McLachlan & Rathnayake (2014).

The order g_o of a g-component mixture model, as in Equation 2, is defined to be the smallest value of g such that the model is compatible with the data, with the model having all components different and all the associated mixing proportions π_i nonzero. The estimation of the order of a mixture model has been considered mainly by consideration of the likelihood, using two main ways. One way is based on a penalized form of the log likelihood. As the likelihood increases with the addition of a component to the mixture model, the likelihood (usually, the log likelihood) is penalized by the subtraction of a term that penalizes the model for the number of parameters in it. This leads to a penalized log likelihood, yielding what are called information criteria for the choice of g.

The other main way for deciding on the order of a mixture model is to carry out a hypothesis test, using a likelihood ratio test (LRT). Unfortunately, the standard regularity conditions do not hold for the null distribution of the LRTS to have its usual chi-squared distribution with degrees of freedom equal to the difference between the number of parameters under the null and alternative hypotheses.

In practice, the null distribution of the LRTS is often estimated by a resampling approach in order to produce a p-value. Thus penalized likelihood criteria, like AIC (Akaike information criterion) and BIC, are less demanding than the LRT. However, they produce no number that quantifies the confidence in the result, such as a p-value.

In an attempt to overcome the shortcomings of the LRT for the number of components in a mixture model in a frequentist framework, Bayesian approaches have been suggested. For example, Aitkin & Rubin (1985) adopted an approach that places a prior distribution on the vector of mixing proportions $\boldsymbol{\pi}$. An advantage of this proposal is that any null hypothesis about the number of components is specified in the interior of the parameter space. However, Quinn et al. (1987) showed that the asymptotic null distribution of $-2 \log \lambda$ will not necessarily be chi-squared, as regularity conditions still do not hold.

Several of the information-based criteria have been derived within a Bayesian framework for model selection but can be applied also in a non-Bayesian framework. Hence, they can be applied to choose the number of components in mixture models considered from either a Bayesian or frequentist perspective. There are also approaches that apply only within a Bayesian framework, such as the procedure of Richardson & Green (1997), who used reversible jump Markov chain Monte Carlo methods to handle the case where the parameter space is of varying dimension. The effect of the prior structure, especially with respect to the mixing proportions and to g itself, is an important aspect of a Bayesian analysis of mixtures. The reader is referred to Richardson & Green (1997), and the contributions of the many discussants of their paper, on this issue.

10.2. Bayesian Information Criterion and Related Methods

The main Bayesian-based information criteria use an approximation to the integrated likelihood, as in the original proposal by Schwarz (1978), which led to his BIC. Available general theoretical justifications of this approximation rely on the same regularity conditions that break down for inference on the number of components in a frequentist framework (see McLachlan & Peel 2000a). Under certain conditions, Keribin (2000) has shown that BIC performs consistently in choosing the true number of components in a mixture model.

In practice, it is often observed that BIC tends to favor models with enough components in order to provide a good estimate of the mixture density. Hence, it tends to overestimate the number of clusters (Biernacki et al. 2000). This led Biernacki et al. (2000) to develop the integrated classification (ICL) criterion. An approximation to this criterion is given by

$$-2\log L(\hat{\boldsymbol{\Psi}}) + \mathrm{d}\log n + EN(\hat{\boldsymbol{\tau}}),$$

where $EN(\hat{\boldsymbol{\tau}}) = -\sum_{i=1}^{g}\sum_{j=1}^{n}\hat{\tau}_i(\boldsymbol{y}_j)\log\hat{\tau}_i(\boldsymbol{y}_j)$ is the entropy of the fuzzy classification matrix $((\hat{\tau}_i(\boldsymbol{y}_j)))$. Here $\hat{\tau}_i(\boldsymbol{y}_j) = \tau_i(\boldsymbol{y}_j;\hat{\boldsymbol{\Psi}})$ and $\hat{\boldsymbol{\tau}} = (\hat{\boldsymbol{\tau}}_1^\top, \ldots, \hat{\boldsymbol{\tau}}_n^\top)^\top$, where

$$\hat{\boldsymbol{\tau}}_j = (\hat{\tau}_1(\boldsymbol{y}_j), \ldots, \hat{\tau}_g(\boldsymbol{y}_j))^\top$$

is the vector of the estimated posterior probabilities of component membership of $\boldsymbol{y}_j$ ($j = 1, \ldots, n$). That is, the ICL criterion uses the entropy term $EN(\hat{\boldsymbol{\tau}})$ to penalize the model for its complexity (too many components and hence clusters).

Another approach to refining the number of clusters was given by Baudry et al. (2010) and Hennig (2010), who suggested ways in which the components can be recombined.

10.3. Resampling Approach

A formal test of the null hypothesis $H_0 : g = g_0$ versus the alternative $H_1 : g = g_1$ ($g_1 > g_0$) can be undertaken using a resampling method, as described in McLachlan (1987). With this approach, bootstrap samples are generated from the mixture model fitted under the null hypothesis of g_0 components. That is, the bootstrap samples are generated from the g_0-component mixture model with the vector $\boldsymbol{\Psi}$ of unknown parameters replaced by its ML estimate $\hat{\boldsymbol{\Psi}}_{g_0}$ computed by consideration of the log likelihood formed from the original data under H_0. The value of $-2\log\lambda$, where λ is the likelihood ratio statistic, is computed for each bootstrap sample after fitting mixture models for $g = g_0$ and g_1 to it in turn. The process is repeated independently B times, and the replicated values of $-2\log\lambda$ formed from the successive bootstrap samples provide an assessment of the bootstrap, and hence of the true, null distribution of $-2\log\lambda$. An account of

other resampling approaches including the gap statistic of Tibshirani et al. (2001) may be found in McLachlan & Khan (2004) and McLachlan & Rathnayake (2011).

10.4. Some Distributional Results for the Likelihood Ratio Test Statistic

Over the years, a number of theoretical and simulation-based results have been published on the null distribution of the LRTS, $-2\log\lambda$, for inference on the number of components in a finite mixture model. We very briefly consider here some of the theoretical results that have been derived; a fuller account may be found in McLachlan & Peel (2000a, chapter 6).

Ghosh & Sen (1985) provided a comprehensive account of the breakdown in regularity conditions for the classical asymptotic theory to hold for the LRTS, $-2\log\lambda$. For a mixture of two known but general univariate densities in unknown proportions, Titterington (1981) and Titterington et al. (1985) considered the LRT of $H_0 : g = 1$ ($\pi_1 = 1$) versus $H_1 : g = 2$ ($\pi_1 < 1$). They showed asymptotically under H_0 that $-2\log\lambda$ is zero with probability 0.5 and, with the same probability, is distributed as chi-squared with one degree of freedom. Another way of expressing this is that the asymptotic null distribution of $-2\log\lambda$ is the same as the distribution of $\{\max(0, W)\}^2$, where W is a standard normal random variable. A further way of expressing this is to say that $-2\log\lambda \sim \frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ under H_0, where χ_0^2 denotes the degenerate distribution that puts mass 1 at zero. In his monograph, Lindsay (1995, section 4.2) referred to this distribution as a chi-bar squared, that is, a mixture of chi-squared distributions.

Hartigan (1985a,b) obtained the same result for the asymptotic null distribution of $-2\log\lambda$ in the case of the two-component normal mixture with unspecified π_1 but known common variance and known means μ_1 and μ_2 where, as in the previous example, the null hypothesis $H_0 : g = 1$ was specified by $\pi_1 = 1$. This example was also considered by Ghosh & Sen (1985) in the course of their development of asymptotic theory for the distribution of the LRTS for mixture models. They were able to derive the limiting null distribution of $-2\log\lambda$ for unknown but identifiable μ_1 and μ_2, where μ_2 lies in a compact set. They showed, in the limit, that $-2\log\lambda$ is distributed as a certain functional,

$$\left[\max\left\{0, \sup_{\mu_2} W(\mu_2)\right\}\right]^2,$$

where $W(\cdot)$ is a Gaussian process with zero mean and covariance kernel depending on the true value of μ_1 under H_0, and the variance of $W(\mu_2)$ is unity for all μ_2.

Hartigan (1985a,b) showed that if μ_2 is unknown with no restrictions on it, then $-2\log\lambda$ is asymptotically unbounded above in probability at a very slow rate $[\frac{1}{2}\log(\log n)]$ when H_0 is true. Also, Bickel & Chernoff (1993) investigated the null behavior of the LRTS for this model.

Ghosh & Sen (1985) established a similar result for component densities from a general parametric family under certain conditions. For the case where the vector of parameters $\boldsymbol{\Psi}_g$ was not assumed to be identifiable, they imposed a separation condition on the values of $\boldsymbol{\Psi}_g$ under H_0 and H_1. The removal of the separation condition imposed in Ghosh & Sen (1985) presented a major challenge to researchers; see, for example, Dacunha-Castelle & Gassiat (1997), Chen & Chen (2001), and Liu & Shao (2004). Garel (1995) subsequently showed it was possible to remove the separation condition with assumptions that involve only the second derivatives of the mixture density.

Chen et al. (2001, 2004) modified the LRTS and derived its limiting distribution. Li et al. (2009) and Chen & Li (2009) proposed an EM test in the case of $g_0 = 1$ (that is, a single normal distribution under the null hypothesis), while it was further developed by Li & Chen (2010) and Chen et al. (2012), including an extension to the case of $g_0 > 1$.

11. SOFTWARE

A range of software for fitting mixture models is available for many commonly used mathematical and statistics platforms such as MATLAB, R, C, C++, Java, and Python. An account of some earlier implementations can be found in the Appendix in McLachlan & Peel (2000a). There are also modules in software such as Stata, SAS, Latent GOLD, and Mplus for fitting latent class and mixture models.

For the R platform (R Development Team 2012) there is the well-known `mclust` package by Scrucca et al. (2016) and the `EMMIX` suite of packages (McLachlan et al. 1999). The `mclust` package and some others such as `pgmm` (McNicholas & Murphy 2008) and `Rmixmod` (Lebret et al. 2015) provide functions to fit normal mixture models with various covariance structures (such as those discussed in Section 7). Apart from normal mixture models, the `mixtools` package (Benaglia et al. 2009) can fit nonparametric models and mixtures of regressions. The latter is also considered by the `flexmix` package, which implements a general framework for fitting mixtures of regression models (Grün & Leisch 2008). For more resources on relevant R packages, see the CRAN (Comprehensive R Archive Network) Task View webpage on Cluster Analysis & Finite Mixtrue models (**http://cran.r-project.org/web/views/Cluster.html**).

Developed initially by McLachlan et al. (1999) for fitting normal mixture models, the suite of `EMMIX` packages has since expanded to include mixture models with nonnormal distributions (including t, skew normal, and skew t-distributions), mixture of factor analyzers, and linear mixed models. They are available, with further details, from **https://people.smp.uq.edu.au/GeoffMcLachlan/mix_soft/index.html**.

Concerning mixtures of factor analyzers, there are several R packages available freely, such as the specialized versions of the EMMIX software `EMMIXmfa` and `EMMIXmcfa`, that are developed for (normal) mixtures of factor analyzers and mixtures of common factor analyzers, respectively. For mixtures of t-factor analyzers, there is the R package `mmtfa`.

For asymmetric mixture modeling (Section 8.3), there are specialized versions of the EMMIX program for fitting mixtures of skew normal and skew t-distributions, including `EMMIXskew` and `EMMIXcskew` (Lee & McLachlan 2018), both available on CRAN, for fitting mixtures of (restricted) skew normal distributions (Equation 23) and its skew t-analog, and mixtures of CFUST distributions (Equation 24). The R package `mixsmsn` (Prates et al. 2013) provides functions to fit several instances of the mixtures of (restricted) skew elliptical distributions.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors' research was supported by research grants from the Australian Research Council.

LITERATURE CITED

Aitkin M, Rubin DB. 1985. Estimation and hypothesis testing in finite mixture models. *J. R. Stat. Soc. B* 47:67–75

Anderson E. 1935. The irises of the Gaspé Peninsula. *Bull. Am. Iris Soc.* 59:2–5

Arellano-Valle RB, Azzalini A. 2006. On the unification of families of skew–normal distributions. *Scand. J. Stat.* 33:561–74

Arellano-Valle RB, Genton M. 2005. On fundamental skew distributions. *J. Multivar. Anal.* 96:93–116

Azzalini A, Capitanio A. 2003. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *J. R. Stat. Soc. B* 65:367–89

Azzalini A, Dalla Valle A. 1996. The multivariate skew-normal distribution. *Biometrika* 83:715–26

Baek J, McLachlan GJ. 2011. Mixtures of common t-factor analyzers for clustering high-dimensional microarray data. *Bioinformatics* 27:1269–76

Baek J, McLachlan GJ, Flack L. 2010. Mixtures of factor analyzers with common factor loadings: applications to the clustering and visualization of high-dimensional data. *IEEE Trans. Pattern Anal. Mach. Intell.* 32:1298–309

Banfield JD, Raftery AE. 1993. Model-based Gaussian and non-Gaussian clustering. *Biometrics* 49:803–21

Barndorff-Nielsen O. 1977. Exponentially decreasing distributions for the logarithm of particle size. *Proc. R. Soc. Lond. A* 353:401–19

Baudry JP, Raftery AE, Celeux G, Lo K, Gottardo R. 2010. Combining mixture components for clustering. *J. Comput. Graph. Stat.* 19:332–53

Baum LE, Petrie T. 1966. Statistical inference for probabilistic functions of finite Markov chains. *Ann. Math. Stat.* 37:1554–63

Benaglia T, Chauveau D, Hunter D, Young D. 2009. Mixtools: an R package for analyzing mixture models. *J. Stat. Softw.* 32:1–29

Besag J. 1986. On the statistical analysis of dirty pictures (with discussion). *J. R. Stat. Soc. B* 48:259–302

Bickel PJ, Chernoff H. 1993. Asymptotic distribution of the likelihood ratio statistic in a prototypical non regular problem. In *Statistics and Probability: A Raghu Raj Bahadur Festschrift*, ed. JK Ghosh, SK Mitra, BP Rao, pp. 83–96. New Delhi: Wiley Eastern

Biernacki C, Celeux G, Govaert G. 2000. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Mach. Intell.* 22:719–25

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

Browne RP, McNicholas PD. 2015. A mixture of generalized hyperbolic distributions. *Can. J. Stat.* 43:176–98

Cabral CRB, Lachos VH, Prates MO. 2012. Multivariate mixture modeling using skew-normal independent distributions. *Comput. Stat. Data Anal.* 56:126–42

Chen J. 2017. Consistency of the MLE under mixture models. *Stat. Sci.* 7:1–26

Chen H, Chen J. 2001. The likelihood ratio test for homogeneity in finite mixture models. *Can. J. Stat.* 29:201–15

Chen H, Chen J, Kalbfleisch JD. 2001. A modified likelihood ratio test for homogeneity in finite mixture models. *J. R. Stat. Soc. B* 63:19–29

Chen H, Chen J, Kalbfleisch JD. 2004. Testing for a finite mixture model with two components. *J. R. Stat. Soc. B* 66:95–115

Chen J, Li P. 2009. Hypothesis test for normal mixture models: the EM approach. *Ann. Stat.* 37:2523–42

Chen J, Li P, Fu Y. 2012. Inference on the order of a normal mixture. *J. Am. Stat. Assoc.* 107:1096–105

Coleman D, Dong X, Hardin J, Rocke DM, Woodruff DL. 1999. Some computational issues in cluster analysis with no a priori metric. *Comput. Stat. Data Anal.* 31:1–11

Coretto P, Hennig C. 2017. Robust improper maximum likelihood: tuning, computation, and a comparison with other methods for robust Gaussian clustering. *J. Am. Stat. Assoc.* 111:1648–59

Dacunha-Castelle D, Gassiat E. 1997. The estimation of the order of a mixture model. *Bernoulli* 3:279–99

Day NE. 1969. Estimating the components of a mixture of normal distributions. *Biometrika* 56:463–74

Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B* 39:1–38

Diebolt J, Robert CP. 1994. Estimation of finite mixture distributions through Bayesian sampling. *J. R. Stat. Soc. B* 56:363–75

Drton M, Plummer M. 2017. A Bayesian information criterion for singular models (with discussion). *J. R. Stat. Soc. B* 79:323–80

Escobar MD, West M. 1995. Bayesian density estimation and inference using mixtures. *J. Am. Stat. Assoc.* 90:577–88

Everitt B, Hand D. 1981. *Finite Mixture Distributions*. New York: Chapman & Hall

Fisher RA. 1936. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* 7:179–88

Fraley C, Raftery AE. 2002. Model-based clustering, discriminant analysis, and density estimation. *J. Am. Stat. Assoc.* 97:611–31

Franczak BC, Browne RP, McNicholas PD. 2014. Mixtures of shifted asymmetric Laplace distributions. *IEEE Trans. Pattern Anal. Mach. Intell.* 36:1149–57

Frühwirth-Schnatter S. 2006. *Finite Mixture and Markov Switching Models*. New York: Springer

Furman WD, Lindsay BG. 1994a. Measuring the effectiveness of moment estimators as starting values in maximizing mixture likelihoods. *Comput. Stat. Data Anal.* 17:493–507

Furman WD, Lindsay BG. 1994b. Testing for the number of components in a mixture of normal distributions using moment estimators. *Comput. Stat. Data Anal.* 17:473–92

Galton F. 1869. *Hereditary Genius: An Inquiry into Its Laws and Consequences*. London: Macmillan

Ganesalingam S, McLachlan GJ. 1978. The efficiency of a linear discriminant function based on unclassified initial samples. *Biometrika* 65:658–65

García-Escudero LA, Gordaliza A, Greselin F, Ingrassia I, Mayo-Iscar A. 2016. The joint role of trimming and constraints in robust estimation for mixtures of Gaussian factor analyzers. *Comput. Stat. Data Anal.* 99:131–47

García-Escudero LA, Gordaliza A, Greselin F, Ingrassia I, Mayo-Iscar A. 2018. Eigenvalues and constraints in mixture modeling: geometric and computational issues. *Adv. Data Anal. Classif.* 12:203–33

García-Escudero LA, Gordaliza A, Matrán C, Mayo-Iscar A. 2008. A general trimming approach to robust cluster analysis. *Ann. Stat.* 36:1324–45

Garel B. 2005. Asymptotic theory of the likelihood ratio test for the identification of a mixture. *J. Stat. Plan. Inference* 131:271–96

Ghahramani Z, Hinton G. 1997. *The EM algorithm for factor analyzers*. Tech. Rep. CRG-TR-96-1, Dep. Comput. Sci., Univ. Toronto

Ghosh JK, Sen PK. 1985. On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, Vol. 2, ed. L LeCam, R Olshen, pp. 789–806. Monterey, CA: Wadsworth

Grün B, Leisch F. 2008. FlexMix Version 2: finite mixtures with concomitant variables and varying and constant parameters. *J. Stat. Softw.* 28:1–35

Hall P, Zhou XH. 2003. Nonparametric estimation of component distributions in a multivariate mixture. *Ann. Stat.* 31:201–24

Hartigan JA. 1975. *Clustering Algorithms*. New York: Wiley

Hartigan JA. 1985a. A failure of likelihood asymptotics for normal mixtures. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, Vol. 2, ed. L LeCam, R Olshen, pp. 807–10. Monterey, CA: Wadsworth

Hartigan JA. 1985b. Statistical theory in clustering. *J. Classif.* 2:63–76

Hennig C. 2010. Methods for merging Gaussian mixture components. *Adv. Data. Anal. Classif.* 4:3–34

Hinton GE, Dayan P, Revow M. 1997. Modeling the manifolds of images of handwritten digits. *IEEE Trans. Neural. Netw.* 8:65–74

Holmes GK. 1892. Measures of distribution. *J. Am. Stat. Assoc.* 3:141–57

Hunter DR, Lange K. 2004. A tutorial on MM algorithms. *Am. Stat.* 58:30–37

Jacobs RA, Jordan MI, Nowlan SJ, Hinton GE. 1991. Adaptive mixtures of local experts. *Neur. Comput.* 3:79–87

Jeffreys SH. 1932. An alternative to the rejection of observations. *Proc. R. Soc. Lond. A* 137:78–87

Kadane JB. 1974. The role of identification in Bayesian theory. In *Studies in Bayesian Econometrics and Statistics*, ed. S Fienberg, A Zellner, pp. 175–91. New York: Elsevier

Karlis D, Santourian A. 2009. Model-based clustering with non-elliptically contoured distributions. *Stat. Comput.* 19:73–83

Keribin C. 2000. Consistent estimation of the order of mixture models. *Sankhya A* 62:49–66

Lange K. 2013. *Optimization*. New York: Springer

Lavine M, West M. 1992. A Bayesian method of classification and discrimination. *Can. J. Stat.* 20:451–61

Lebret R, Iovleff S, Langrognet F, Biernacki C, Celeux G, Govaert G. 2015. Rmixmod: The R package of the model-based unsupervised, supervised, and semi-supervised classification Mixmod library. *J. Stat. Softw.* 67:1–29

Lee SX, Leemaqz KL, McLachlan GJ. 2018. A block EM algorithm for multivariate skew normal and skew t-mixture models. *IEEE T. Neur. Net. Learn.* 29:5581–91

Lee SX, McLachlan GJ. 2013. On mixtures of skew normal and skew t-distributions. *Adv. Data. Anal. Classif.* 7:241–66

Lee SX, McLachlan GJ. 2014. Finite mixtures of multivariate skew t-distributions: some recent and new results. *Stat. Comput.* 24:181–202

Lee SX, McLachlan GJ. 2016. Finite mixtures of canonical fundamental skew t-distributions. *Stat. Comput.* 26:573–89

Lee SX, McLachlan GJ. 2018. EMMIXcskew: an R package for the fitting of a mixture of canonical fundamental skew t-distributions. *J. Stat. Softw.* 83:1–32

Leroux B. 1992. Consistent estimation of a mixing distribution. *Ann. Stat.* 20:1350–60

Li JQ, Barron AR. 1999. Mixture density estimation. In *Advances in Neural Information Processing Systems 12 (NIPS 1999)*, ed. SA Solla, TK Leen, K Müller, pp. 279–85. Cambridge, MA: MIT Press

Li P, Chen J. 2010. Testing the order of a finite mixture. *J. Am. Stat. Assoc.* 105:1084–92

Li P, Chen J, Marriott P. 2009. Non-finite Fisher information and homogeneity: an EM approach. *Biometrika* 96:411–26

Lin TI, McLachlan GJ, Lee SX. 2016. Extending mixtures of factor models using the restricted multivariate skew-normal distribution. *J. Multivar. Anal.* 143:398–413

Lin TI, Wu PH, McLachlan GJ, Lee SX. 2015. A robust factor analysis model using the restricted skew t-distribution. *TEST* 24:510–31

Lindsay BG. 1995. *Mixture Models: Theory, Geometry and Applications*. Hayward, CA: Inst. Math. Stat.

Liu X, Shao Y. 2004. Asymptotics for the likelihood ratio test in a two-component normal mixture model. *J. Stat. Plan. Inference* 123:61–81

Lo K, Gottardo R. 2012. Flexible mixture modeling via the multivariate t distribution with the Box-Cox transformation: an alternative to the skew-t distribution. *Stat. Comput.* 22:33–52

Maugis C, Celeux G, Martin-Magniette ML. 2009. Variable selection for clustering with Gaussian mixture models. *Biometrics* 65:701–9

McLachlan GJ. 1975. Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis. *J. Am. Stat. Assoc.* 70:365–69

McLachlan GJ. 1982. The classification and mixture maximum likelihood approaches to cluster analysis. In *Handbook of Statistics*, Vol. 2, ed. PR Krishnaiah, L Kanal, pp. 199–208. Amsterdam: North-Holland

McLachlan GJ. 1987. On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Appl. Stat.* 36:318–24

McLachlan GJ. 1992. *Discriminant Analysis and Statistical Pattern Recognition*. New York: Wiley

McLachlan GJ. 2016. Mixture distributions—further developments. In *Wiley StatsRef: Statistics Reference Online*, ed. N Balakrishnan, P Brandimarte, B Everitt, G Molenberghs, F Ruggeri, W Piegorsch. Chichester, UK: Wiley. **https://doi.org/10.1002/9781118445112.stat00947.pub2**

McLachlan GJ, Basford K. 1988. *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker

McLachlan GJ, Bean RW, Ben-Tovim Jones L. 2007. Extension of the mixture of factor analyzers model to incorporate the multivariate t-distribution. *Comput. Stat. Data Anal.* 51:5327–38

McLachlan GJ, Bean R, Peel D. 2002. A mixture model–based approach to the clustering of microarray expression data. *Bioinformatics* 18:413–22

McLachlan GJ, Lee SX. 2016. Comment on "On nomenclature for, and the relative merits of, two formulations of skew distributions" by A Azzalini, R Browne, M Genton, and P McNicholas. *Stat. Probab. Lett.* 116:1–5

McLachlan GJ, Khan N. 2004. On a resampling approach for tests on the number of clusters with mixture model-based clustering of tissue samples. *J. Multivar. Anal.* 90:90–105

McLachlan GJ, Krishnan T. 2008. *The EM Algorithm and Extensions*. Hoboken, NJ: Wiley. 2nd ed.

McLachlan GJ, Peel D. 1998. Robust cluster analysis via mixtures of multivariate t-distributions. In *Advances in Pattern Recognition*, ed. A Amin, D Dori, P Pudil, H Freeman. Berlin: Springer

McLachlan GJ, Peel D. 2000a. *Finite Mixture Models*. New York: Wiley

McLachlan GJ, Peel D. 2000b. Mixtures of factor analyzers. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 599–606. Burlington, MA: Morgan Kaufmann

McLachlan GJ, Peel D, Basford KE, Adams P. 1999. The EMMIX algorithm for the fitting of normal and *t*-components. *J. Stat. Softw.* 4:1–14

McLachlan GJ, Rathnayake SI. 2011. Testing for group structure in high-dimensional data. *J. Biopharm. Stat.* 21:1113–25

McLachlan GJ, Rathnayake SI. 2014. On the number of components in a Gaussian mixture model. *WIREs Data Min. Knowl. Discov.* 4:341–55

McNicholas PD. 2017. *Mixture Model-Based Classification*. Boca Raton, FL: CRC

McNicholas PD, Murphy TB. 2008. Parsimonious Gaussian mixture models. *Stat. Comput.* 18:285–96

Meng XL, Rubin DB. 1993. Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* 80:267–78

Meng XL, van Dyk D. 1997. The EM algorithm—an old folk-song sung to a fast new tune (with discussion). *J. R. Stat. Soc. B* 59:511–67

Mengersen K, Robert C, Titterington D, eds. 2011. *Mixtures: Estimation and Applications*. New York: Wiley

Montanari A, Viroli C. 2010. Heteroscedastic factor mixture analysis. *Stat. Model.* 10:441–60

Murray P, Browne R, McNicholas P. 2014. Mixtures of skew-*t* factor analyzers. *Comput. Stat. Data Anal.* 77:326–35

Newcomb S. 1886. A generalized theory of the combination of observations so as to obtain the best result. *Am. J. Math.* 8:343–66

Ng SK, McLachlan GJ, Wang K, Ben-Tovim Jones L, Ng SW. 2006. A mixture model with random-effects components for clustering correlated gene-expression profiles. *Bioinformatics* 22:1745–52

Ng SK, McLachlan GJ, Wang K, Nagymanyoki Z, Liu S, Ng SW. 2015. Inference on differential expression using cluster-specific contrasts of mixed effects. *Biostatistics* 16:98–112

Nguyen HD, McLachlan GJ. 2016a. Laplace mixtures of linear experts. *Comput. Stat. Data Anal.* 93:177–91

Nguyen HD, McLachlan GJ. 2016b. Maximum likelihood estimation of triangular and polygonal distributions. *Comput. Stat. Data Anal.* 102:23–36

Nguyen HD, McLachlan GJ, Orban P, Bellec P, Janke AL. 2017. Maximum pseudolikelihood estimation for a model-based clustering of time-series data. *Neural Comput.* 29:990–1020

Nguyen HD, McLachlan GJ, Ullmann JFP, Janke AL. 2016. Spatial clustering of time-series via mixtures of autoregressive models and Markov random fields. *Stat. Neerl.* 70:414–39

Nguyen HD, Ullmann JFP, McLachlan GJ, Voleti V, Li W, et al. 2018. Whole-volume clustering of time series data from zebrafish brain calcium images via mixture model-based functional data analysis. *Stat. Anal. Data Min.* 11:5–16

O'Neill TJ. 1978. Normal discrimination with unclassified observations. *J. Am. Stat. Assoc.* 73:821–26

Pan W, Shen X. 2007. Penalized model-based clustering with application to variable selection. *J. Mach. Learn. Res.* 8:1145–64

Panel on Nonstandard Mixtures of Distributions. 1989. Statistical models and analysis in auditing. *Stat. Sci.* 4:2–33

Pearson K. 1894. Contributions to the mathematical theory of evolution. *Phil. Trans. R. Soc. Lond. A* 185:71–110

Peel D, McLachlan GJ. 2000. Robust mixture modelling using the *t* distribution. *Stat. Comput.* 10:339–48

Prates MO, Cabral CRB, Lachos VH. 2013. Mixsmsn: fitting finite mixture of scale mixture of skew-normal distributions. *J. Stat. Softw.* 54:1–20

Pyne S, Hu X, Wang K, Rossin E, Lin TI, et al. 2009. Automated high-dimensional flow cytometric data analysis. *PNAS* 106:8519–24

Quetelet A. 1846. Lettres à S.A.R. Le Duc Régnant de Saxe-Coburg et Gotha: sur la Théorie des Probabilités, Appliquée aux Sciences Morales Et Politiques. Brussels: Hayez

Quetelet A. 1852. Sur quelques propritiétés curieuses que présentent les résultats d'une serie d'observations, faites dans la vue de déterminer une constante, lorsque les chances de rencontrer des écarts en plus et en moins sont égales et indépendantés les unes des autres. *Bull. Acad. R. Sci. Lett. Beaux-Arts Belg.* 19:303–17

Quinn BG, McLachlan GJ, Hjort NL. 1987. A note on the Aitkin-Rubin approach to hypothesis testing in mixture models. *J. R. Stat. Soc. B* 49:311–14

R Development Team. 2012. *R: A language and environment for statistical computing*. Vienna: R Found. Stat. Comput.

Raftery A, Dean N. 2006. Variable selection for model-based clustering. *J. Am. Stat. Assoc.* 101:168–78

Rao CR. 1948. The utilization of multiple measurements in problems of biological classification. *J. R. Stat. Soc. B* 10:159–203

Redner R. 1981. Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. *Ann. Stat.* 9:225–28

Richardson S, Green PJ. 1997. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. R. Stat. Soc. B* 59:731–92

Robert CP. 1996. Mixtures of distributions: inference and estimation. In *Markov Chain Monte Carlo in Practice*, ed. WR Gilks, S Richardson, DJ Spiegelhalter, pp. 441–64. London: Chapman & Hall

Roeder K, Wasserman L. 1997. Practical Bayesian density estimation using mixtures of normals. *J. Am. Stat. Assoc.* 92:894–902

Schwarz G. 1978. Estimating the dimension of a model. *Ann. Stat.* 6:461–64

Scrucca L, Fop M, Murphy TB, Raftery AE. 2016. Mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *R J.* 8:205–33

Spurek P. 2017. General split Gaussian cross-entropy clustering. *Expert Syst. Appl.* 68:58–68

Stigler SM. 1986. *The History of Statistics: The Measurement of Uncertainty Before 1900*. Cambridge, MA: Harvard Univ. Press

Tanner MA, Wong WH. 1987. The calculation of posterior distributions by data augmentation (with discussion). *J. Am. Stat. Assoc.* 82:528–50

Teicher H. 1960. On the mixture of distributions. *Ann. Math. Stat.* 31:55–73

Tibshirani R, Walther G, Hastie T. 2001. Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. B* 63:411–23

Titterington DM. 1981. Contribution to the discussion of paper by M. Aitkin, D. Anderson, and J. Hinde. *J. R. Stat. Soc. A* 144:459

Titterington DM, Smith A, Makov U. 1985. *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley

Tortora C, McNicholas P, Browne R. 2016. A mixture of generalized hyperbolic factor analyzers. *Adv. Data Anal. Classif.* 10:423–40

Viroli C. 2010. Dimensionally reduced model-based clustering through mixtures of factor mixture analyzers. *J. Classif.* 27:363–88

Viroli C, McLachlan GJ. 2017. Deep Gaussian mixture models. *Stat. Comput*. **https://doi.org/10.1007/s11222-017-9793-z**

Wald A. 1949. Note on the consistency of the maximum likelihood estimate. *Ann. Math. Stat.* 20:595–601

Weldon WFR. 1892. Certain correlated variations in *Crangon vulgaris*. *Proc. R. Soc. Lond.* 51:1–21

Weldon WFR. 1893. On certain correlated variations in *Carcinus maenas*. *Proc. R. Soc. Lond.* 54:318–29

Witten D, Tibshirani R. 2010. A framework for feature selection in clustering. *J. Am. Stat. Assoc.* 105:713–26

Wolfe JH. 1970. Pattern clustering by multivariate mixture analysis. *Multivar. Behav. Res.* 5:329–50

Wu CFJ. 1983. On the convergence properties of the EM algorithm. *Ann. Math. Stat.* 11:95–103

Wraith D, Forbes F. 2015. Location and scale mixtures of Gaussians with flexible tail behaviour: properties, inference and application to multivariate clustering. *Comput. Stat. Data Anal.* 90:61–73

Xie B, Pan W, Shen X. 2010. Penalized mixtures of factor analyzers with application to clustering high-dimensional microarray data. *Bioinformatics* 26:501–8

Yakowitz SJ, Spragins JD. 1968. On the identifiability of finite mixtures. *Ann. Math. Stat.* 39:209–14

Zhou H, Pan W. 2009. Penalized model-based clustering with unconstrained covariance matrices. *Electron. J. Stat.* 3:1473–96

Zhu X, Melnykov V. 2018. Manly transformation in finite mixture modeling. *Comput. Stat. Data Anal.* 121:190–208