Journal of Classification 33:331-373 (2016)
DOI: 10.1007/s00357-016-9211-9

# Model-Based Clustering

Paul D. McNicholas

McMaster University, Canada

**Abstract:** The notion of defining a cluster as a component in a mixture model was put forth by Tiedeman in 1955; since then, the use of mixture models for clustering has grown into an important subfield of classification. Considering the volume of work within this field over the past decade, which seems equal to all of that which went before, a review of work to date is timely. First, the definition of a cluster is discussed and some historical context for model-based clustering is provided. Then, starting with Gaussian mixtures, the evolution of model-based clustering is traced, from the famous paper by Wolfe in 1965 to work that is currently available only in preprint form. This review ends with a look ahead to the next decade or so.

**Keywords:** Cluster; Cluster analysis; Mixture models; Model-based clustering.

## 1. Defining a Cluster

The best place to start is at the beginning, which consists in a question: what is a cluster? Before positing an answer, some historical context is helpful. The oldest citation given pertaining to mixture models and clustering is usually the thesis by Wolfe (1963). McNicholas (2016) explains that while Wolfe (1963) uses the idea of a mixture model to define a cluster, he does not use a mixture model to perform clustering. More specifically, the clustering procedures developed by Wolfe (1963) are not based on maximiz-

The author is grateful to Chapman & Hall/CRC Press for allowing some text and figures from his monograph (McNicholas 2016) to be used in this review paper. The author is thankful for the helpful comments of an anonymous reviewer and the Editor. The work is partly supported by the Canada Research Chairs program.

Corresponding Author's Address: Department of Mathematics and Statistics, McMaster University, 1280 Main Street West, Hamilton, Ontario, Canada L8S4L8, email: mcnicholas@math.mcmaster.ca.

Published online: 8 November 2016

ing the likelihood—or otherwise exploiting the likelihood—of a Gaussian mixture model. Of this clustering methodology, Wolfe (1963, p. 76) writes:

> The methods described in this thesis are not only bad, they have been rendered obsolete by the author's own subsequent work.

The subsequent work referred to here is the paper by Wolfe (1965), which seems to be the first published example of Gaussian model-based clustering. Wolfe (1963) gives the following definition of a cluster, or type:

> A type is a distribution which is one of the components of [a] mixture of distributions.

McNicholas (2016) points out that Tiedeman (1955) uses a similar definition in a prescient paper that builds on famous works by Pearson (1894) and Rao (1952). As McNicholas (2016) explains, a driving force behind the work of Tiedeman (1955) is to encourage work on what we now know as clustering. Because the idea of defining a cluster in terms of a component in a mixture model goes back to Tiedeman (1955), it is worth noting how he formulated the problem:

> Consider $G$ observation matrices each of which generates a density function of the form given by equation [1]. Throw away the type identification of each observation set and you have a mixed series of unknown density form.

Here, [1] is the density of a Gaussian random variable. The objective, as laid down by Tiedeman (1955), is then

> . . . to solve the problem of reconstructing the $G$ density functions of original types.

Over the subsequent two decades, much energy was invested in its solution, led by Wolfe (1963;1965).

Wolfe (1963, Chapter I.D) discusses two alternative definitions of a cluster. One defines a cluster as a mode in a distribution, while the other focuses on similarity (cf. McNicholas 2016, Chapter 2). The principal problem with defining a cluster in terms of a mode can be seen by generating two overlapping Gaussian components such that there are clearly three modes, e.g., Figure 1.

Definitions based on similarity have long been popular and Wolfe (1963) gives an example of such a definition:

> A type is a set of objects which are more similar to each other than they are to objects not members of the set.

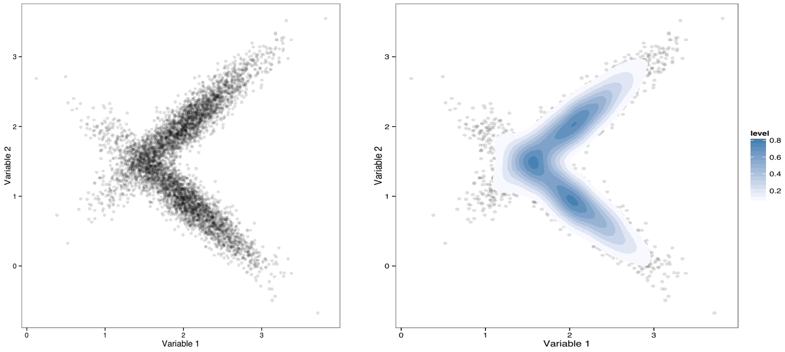

Figure 1. Scatter plots, with semi-transparent points, for data simulated from two overlapping Gaussian components, where the density is illustrated on the right-hand plot.

Wolfe (1963) cites a host of other work that uses similar definitions, e.g., Tryon (1939, 1955), Catell (1949), Stephensen (1953), and McQuitty (1956), and such definitions remain popular today. Wolfe (1963) points out several problems with definitions based on similarity. One of the issues that he raises concerns the difficulty around quantifying similarity. Wolfe (1963) writes, *inter alia*, that

> . . . most definitions of similarity are arbitrary.

Beyond the issues he raises, the fact that definitions based on similarity are often satisfied by a solution that sees each point assigned to its own cluster is highly problematic.

McNicholas (2016) proffers a mixture model-based definition that is a little more specific than those used by Tiedeman (1955) and Wolfe (1963):

> A cluster is a unimodal component within an appropriate finite mixture model.

McNicholas (2016) explains that an "appropriate" mixture model here is one that is appropriate in light of the data under consideration. What does it mean for a mixture model to be "appropriate" in light of the data? It means that the model has the necessary flexibility, or parameterization, to fit the data; e.g., if the data contain skewed clusters, then the mixture model should be able to accommodate skewed components. In many cases, being appropriate in light of the data will also mean that each component has convex contours so that each cluster is convex (cf. McNicholas 2016, Section 7.6). The unimodal requirement in the definition of McNicholas (2016) is important because if the component is not unimodal, then one of two things is almost certainly happening: the wrong mixture distribution is being fitted or not enough components are being used. An example of the former—specifically, multiple Gaussian components being used to model one skewed cluster—is given in Section 4.2. The position taken herein is that the definition given by McNicholas (2016) should be used.

That the definition of McNicholas (2016) ties the notion of a cluster to the data under consideration is essential because a cluster really only has meaning in the context of data. While this definition insists that clusters are unimodal, it is not at all the same as asserting that a cluster is a mode. Interestingly, Gordon (1981, Sec. 1.1) reports two desiderata, or desired characteristics, of a cluster that are stated as "basic ideas" by Cormack (1971):

> Two possible desiderata for a cluster can thus be stated as internal cohesion and external isolation.

Of course, complete external isolation will not be possible in many real analyses; however, the idea of internal cohesion seems quite compatible with the idea of a cluster corresponding to a unimodal component in an appropriate finite mixture. Interestingly, when referring to a situation where external isolation may not be possible, Gordon (1981, Sec. 1.1) highlights the fact that

> ...the conclusion reached will in general depend on the nature of the data.

This vital link with the data under consideration is along similar lines to the requirement of an "appropriate" finite mixture model in the definition of McNicholas (2016).

Everitt et al. (2011, Section 1.4) point out that dissection, as opposed to clustering, might be necessary in some circumstances, and Gordon (1981, Section 1.1) argues along similar lines. Everitt et al. (2011, Section 1.4) define dissection as

> ...the process of dividing a homogenous data set into different parts.

Of course, it is true that there are situations where one might wish to carry out dissection rather than clustering. In fact, there may even be cases where a departure from the definition of a cluster offered by McNicholas (2016) is desirable in light of the data under consideration. In general, however, I do not feel comfortable reporting clustering results to scientists, or other collaborators, unless the clusters can be framed in terms of the (unimodal) components of an appropriate mixture model.

A reviewer pointed out that the definition of McNicholas (2016) may be perceived as somewhat strident. While there might be situations where the data demand a departure from this definition, alternative definitions such as those based on similarity, modes, or ideas such as internal cohesion and external isolation necessarily require substantial refinement. Furthermore, such refinement seems to almost inevitably lead back to a mixture model-based definition such as that given by McNicholas (2016). For example, to refine a definition based on modes, consideration should be given to how data disperse from the modes, which begins the seemingly inescapable march back to a mixture model-based definition.

## 2. Model-Based Clustering

"Model-based clustering" refers to the use of (finite) mixture models to perform clustering and is the focus of the present review. A random vector $\mathbf{X}$ arises from a parametric finite mixture distribution if, for all $\mathbf{x} \subset \mathbf{X}$, its density can be written

$$f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g f_g(\mathbf{x} \mid \boldsymbol{\theta}_g), \tag{1}$$

where $\pi_g > 0$, such that $\sum_{g=1}^{G} \pi_g = 1$, are called mixing proportions, $f_g(\mathbf{x} \mid \boldsymbol{\theta}_g)$ is the $g$th component density, and $\boldsymbol{\vartheta} = (\boldsymbol{\pi}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G)$, with $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_G)$, is the vector of parameters. Note that $f(\mathbf{x} \mid \boldsymbol{\vartheta})$ in (1) is called a $G$-component finite mixture density. In clustering applications, the component densities $f_1(\mathbf{x} \mid \boldsymbol{\theta}_1), f_2(\mathbf{x} \mid \boldsymbol{\theta}_2), \ldots, f_G(\mathbf{x} \mid \boldsymbol{\theta}_G)$ are usually taken to be of the same type, i.e., $f_g(\mathbf{x} \mid \boldsymbol{\theta}_g) = f(\mathbf{x} \mid \boldsymbol{\theta}_g)$ for all $g$. Extensive details on finite mixture models and their applications are given in the well-known texts by Everitt and Hand (1981), Titterington, Smith and Makov (1985), McLachlan and Basford (1988), McLachlan and Peel (2000a), and Frühwirth-Schnatter (2006).

Let $\mathbf{z}_i = (z_{i1}, \ldots, z_{iG})$ denote the component membership of observation $i$, so that $z_{ig} = 1$ if observation $i$ belongs to component $g$ and $z_{ig} = 0$ otherwise. Suppose $n$ $p$-dimensional data vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are observed and all $n$ are unlabelled or treated as unlabelled. Continuing the notation from (1), and using $f_g(\mathbf{x} \mid \boldsymbol{\theta}_g) = f(\mathbf{x} \mid \boldsymbol{\theta}_g)$ for all $g$, the likelihood is

$$\mathcal{L}(\boldsymbol{\vartheta} \mid \mathbf{x}) = \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_g f(\mathbf{x}_i \mid \boldsymbol{\theta}_g).$$

After the parameters have been estimated, the predicted classification results are given by the *a posteriori* probabilities

$$\hat{z}_{ig} := \frac{\hat{\pi}_g f(\mathbf{x}_i \mid \hat{\boldsymbol{\theta}}_g)}{\sum_{h=1}^{G} \hat{\pi}_h f(\mathbf{x}_i \mid \hat{\boldsymbol{\theta}}_h)},$$

for $i = 1, \ldots, n$. The fact that these *a posteriori* predicted classifications are soft, i.e., $\hat{z}_{ig} \in [0, 1]$, under the fitted model is often considered an advantage of the mixture model-based approach. However, in some applications it is desirable to harden the *a posteriori* classifications and the most popular way to do this is via maximum *a posteriori* (MAP) classifications, i.e., MAP$\{\hat{z}_{ig}\}$, where MAP$\{\hat{z}_{ig}\} = 1$ if $g = \arg\max_h\{\hat{z}_{ih}\}$, and MAP$\{\hat{z}_{ig}\} = 0$ otherwise.

Wolfe (1965) presents software for computing maximum likelihood estimates for Gaussian model-based clustering. This software includes four different parameter estimation techniques, including an iterative scheme, and it is effective for up to five variables and six components. Day (1969) introduces an iterative technique for finding maximum likelihood estimates when the covariance matrices are held equal, and discusses clustering applications. Wolfe (1970) develops iterative approaches for finding maximum likelihood estimates in the cases of common and differing covariance matrices, respectively, and illustrates these approaches for clustering. Interestingly, Wolfe (1970) draws an analogy between his approach for Gaussian mixtures with common covariance matrices and one of the criteria described by Friedman and Rubin (1967). This and other work on parameter estimation in Gaussian model-based clustering—e.g., Edwards and Cavalli-Sforza (1965), Baum et al. (1970), Scott and Symons (1971), Orchard and Woodbury (1972), and Sundberg (1974)—effectively culminated in the landmark paper by Dempster, Laird and Rubin (1977), wherein the expectation-maximization (EM) algorithm is introduced; see Titterington et al. (1985, Section 4.3.2) and McNicholas (2016, Chapter 2). The EM algorithm is an iterative procedure for finding maximum likelihood estimates when data are incomplete. Extensive details on the EM algorithm are given by McLachlan and Krishnan (2008), and a discussion on stopping rules, with some focus on criteria based on Aitken's acceleration (Aitken 1926), is given by McNicholas (2016, Section 2.2.5).

A family of mixture models arises when various constraints are imposed upon component densities, typically upon the covariance structure. Consider a Gaussian mixture model so that the $g$th component density is $\phi(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$, where $\boldsymbol{\mu}_g$ is the mean and $\boldsymbol{\Sigma}_g$ is the covariance matrix. Some straightforward, but not necessarily useful, constraints on $\boldsymbol{\Sigma}_g$ are $\boldsymbol{\Sigma}_g = \mathbf{I}_p$, $\boldsymbol{\Sigma}_g = \sigma_g\mathbf{I}_p$, $\boldsymbol{\Sigma}_g = \sigma\mathbf{I}_p$, and $\boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$ (see Gordon 1981; Banfield and Raftery 1993, amongst others). The four corresponding mixture models, together with the unconstrained model, could be viewed as a family of five Gaussian mixture models. Banfield and Raftery (1993) consider eigen-decompositions of the component covariance matrices and study several resulting models. These models arise by first considering an eigen-decomposition of the component covariance matrices, i.e.,

$$\mathbf{\Sigma}_g = \lambda_g \mathbf{\Gamma}_g \mathbf{\Delta}_g \mathbf{\Gamma}'_g, \tag{2}$$

where $\lambda_g = |\mathbf{\Sigma}_g|^{1/p}$, $\mathbf{\Gamma}_g$ is the matrix of eigenvectors of $\mathbf{\Sigma}_g$, and $\mathbf{\Delta}_g$ is a diagonal matrix, such that $|\mathbf{\Delta}_g| = 1$, containing the normalized eigenvalues of $\mathbf{\Sigma}_g$ in decreasing order. Note that the columns of $\mathbf{\Gamma}_g$ are ordered to correspond to the elements of $\mathbf{\Delta}_g$. As Banfield and Raftery (1993) point out, the constituent elements of (2) can be viewed in the context of the geometry of the $g$th component, where $\lambda_g$ represents the volume in $p$-space, $\mathbf{\Delta}_g$ the shape, and $\mathbf{\Gamma}_g$ the orientation.

Celeux and Govaert (1995) build on the models of Banfield and Raftery (1993), resulting in a family of 14 Gaussian parsimonious clustering models (GPCMs; Table 1). The fourteen GPCM models can be thought of as belonging to one of three categories: spherical, diagonal, and general. Of these three categories, only the eight general models have flexibility in their orientation, i.e., do not assume that the variables are independent. The eight general models have $\mathcal{O}(p^2)$ covariance parameters, limiting their applicability to data with lower values of $p$. Near the end of the last century, a subset of eight of the GPCMs was made available as the MCLUST family, with accompanying S-PLUS software (Fraley and Raftery 1999). The availability of this software, together with the well-known review paper by Fraley and Raftery (2002b), played an important role in popularizing model-based clustering. In fact, such was the impact of these works that the term model-based clustering became synonymous with MCLUST for several years. Another key component to the popularity of the MCLUST family is the release of an accompanying R (R Core Team 2015) package and, perhaps most notably, the release of `mclust` version 2 (Fraley and Raftery 2002a).

Browne and McNicholas (2014c) point out that the algorithms Celeux and Govaert (1995) use for the EVE and VVE models are computationally infeasible in higher dimensions. They develop alternative algorithms for these models, based on an accelerated line search on the orthogonal Stiefel manifold (see Browne and McNicholas 2014c, for details). Browne and McNicholas (2014a) develop another approach, using fast majorization-minimization algorithms, for the EVE and VVE models and it is this approach that is implemented in the `mixture` package (Browne and McNicholas 2014b) for R. Details on this latter approach are given in Browne and McNicholas (2014a).

In a typical application involving the GPCM or some other family of mixture models, one would fit all models in a family and the best one would be selected via some criterion. A typical application of the GPCM family of models consists of running each of the models (Table 1) for a range of values of $G$. Then, the best of these models is selected using some criterion and the associated classifications are reported. The most popular criterion

Table 1. The type, nomenclature, and covariance structure for each member of the GPCM family.

| | **Model** | **Volume** | **Shape** | **Orientation** | $\mathbf{\Sigma}_g$ |
|---|---|---|---|---|---|
| Spherical | EII | Equal | Spherical | | $\lambda \boldsymbol{I}$ |
| | VII | Variable | Spherical | | $\lambda_g \boldsymbol{I}$ |
| Diagonal | EEI | Equal | Equal | Axis-Aligned | $\lambda \mathbf{\Delta}$ |
| | VEI | Variable | Equal | Axis-Aligned | $\lambda_g \mathbf{\Delta}$ |
| | EVI | Equal | Variable | Axis-Aligned | $\lambda \mathbf{\Delta}_g$ |
| | VVI | Variable | Variable | Axis-Aligned | $\lambda_g \mathbf{\Delta}_g$ |
| General | EEE | Equal | Equal | Equal | $\lambda \mathbf{\Gamma} \mathbf{\Delta} \mathbf{\Gamma}'$ |
| | VEE | Variable | Equal | Equal | $\lambda_g \mathbf{\Gamma} \mathbf{\Delta} \mathbf{\Gamma}'$ |
| | EVE | Equal | Variable | Equal | $\lambda \mathbf{\Gamma} \mathbf{\Delta}_g \mathbf{\Gamma}'$ |
| | EEV | Equal | Equal | Variable | $\lambda \mathbf{\Gamma}_g \mathbf{\Delta} \mathbf{\Gamma}_g'$ |
| | VVE | Variable | Variable | Equal | $\lambda_g \mathbf{\Gamma} \mathbf{\Delta}_g \mathbf{\Gamma}'$ |
| | VEV | Variable | Equal | Variable | $\lambda_g \mathbf{\Gamma}_g \mathbf{\Delta} \mathbf{\Gamma}_g'$ |
| | EVV | Equal | Variable | Variable | $\lambda \mathbf{\Gamma}_g \mathbf{\Delta}_g \mathbf{\Gamma}_g'$ |
| | VVV | Variable | Variable | Variable | $\lambda_g \mathbf{\Gamma}_g \mathbf{\Delta}_g \mathbf{\Gamma}_g'$ |

for this purpose is the Bayesian information criterion (BIC; Schwarz 1978), i.e.,

$$\text{BIC} = 2l(\hat{\boldsymbol{\vartheta}}) - \rho \log n, \tag{3}$$

where $\hat{\boldsymbol{\vartheta}}$ is the maximum likelihood estimate of $\boldsymbol{\vartheta}$, $l(\hat{\boldsymbol{\vartheta}})$ is the maximized log-likelihood, and $\rho$ is the number of free parameters. Leroux (1992) and Keribin (2000) give theoretical results that, under certain regularity conditions, support the use of the BIC for choosing the number of components in a mixture model. Dasgupta and Raftery (1998) discuss the BIC in the context of selecting the number of components in a Gaussian mixture model. Its application to this problem, i.e., selection of $G$, is part of the reason why the BIC has become so popular for mixture model selection in general. Despite its popularity, the model selected by the BIC does not necessarily give the best classification performance from among the candidate models. To this end, alternatives such as the integrated completed likelihood (ICL; Biernacki, Celeux, and Govaert 2000) are sometimes considered. Writing the BIC as in (3), the ICL can be calculated via

$$\text{ICL} \approx \text{BIC} + 2\sum_{i=1}^{n}\sum_{g=1}^{G} \text{MAP}\{\hat{z}_{ig}\} \log \hat{z}_{ig}, \tag{4}$$

where the term

$$2\sum_{i=1}^{n}\sum_{g=1}^{G} \text{MAP}\{\hat{z}_{ig}\} \log \hat{z}_{ig}$$

is typically described as an entropy penalty that reflects the uncertainty in the classification of observations into components. An interesting perspective on the ICL is presented by Baudry (2015), who also discusses the conditional classification likelihood and related ideas.

## 3. Mixture of Factor Analyzers and Extensions

The most popular way to handle high-dimensional data in model-based clustering applications is via the mixture of factor analyzers model or some variation thereof. Consider independent $p$-dimensional random variables $\mathbf{X}_1, \ldots, \mathbf{X}_n$. First, consider the factor analysis model (Spearman 1904, 1927; Bartlett 1953; Lawley and Maxwell 1962), which can be written $\mathbf{X}_i = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{U}_i + \boldsymbol{\varepsilon}_i$, for $i = 1, \ldots, n$, where $\boldsymbol{\Lambda}$ is a $p \times q$ matrix of factor loadings, the latent factor $\mathbf{U}_i \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_q)$, and $\boldsymbol{\varepsilon}_i \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Psi})$, where $\boldsymbol{\Psi} = \mathrm{diag}(\psi_1, \psi_2, \ldots, \psi_p)$. Note that the $\mathbf{U}_i$ are independently distributed, and are independent of the $\boldsymbol{\varepsilon}_i$, which are also independently distributed. Considering the joint distribution

$$\begin{bmatrix} \mathbf{X}_i \\ \mathbf{U}_i \end{bmatrix} \backsim \mathrm{N}\left( \begin{bmatrix} \boldsymbol{\mu} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi} & \boldsymbol{\Lambda} \\ \boldsymbol{\Lambda}' & \mathbf{I}_q \end{bmatrix} \right),$$

it follows that $\mathbb{E}[\mathbf{U}_i \mid \mathbf{x}_i] = \boldsymbol{\beta}(\mathbf{x}_i - \boldsymbol{\mu})$ and $\mathbb{E}[\mathbf{U}_i\mathbf{U}_i' \mid \mathbf{x}_i] = \mathbf{I}_q - \boldsymbol{\beta}\boldsymbol{\Lambda} + \boldsymbol{\beta}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})'\boldsymbol{\beta}'$, where $\boldsymbol{\beta} = \boldsymbol{\Lambda}'(\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi})^{-1}$. Given these expected values, it is straightforward to use an EM algorithm for parameter estimation. The choice of the number of factors $q < p$ is an important consideration in factor analysis. One approach is to choose the number of factors that captures a certain proportion of the variance in the data. Lopes and West (2004) carry out simulation studies to demonstrate that the BIC can be effective for selection of the number of factors. Another well-known approach for selecting the number of factors is parallel analysis (Horn 1965; Humphreys and Ilgen 1969; Humphreys and Montanelli 1975; Montanelli and Humphreys 1976). Different approaches for selecting the number of factors are discussed, *inter alia*, by Fabrigar et al. (1999).

Analogous to the factor analysis model, the mixture of factor analyzers model assumes that

$$\mathbf{X}_i = \boldsymbol{\mu}_g + \boldsymbol{\Lambda}_g\mathbf{U}_{ig} + \boldsymbol{\varepsilon}_{ig} \tag{5}$$

with probability $\pi_g$, for $i = 1, \ldots, n$ and $g = 1, \ldots, G$, where $\boldsymbol{\Lambda}_g$ is a $p \times q$ matrix of factor loadings, the $\mathbf{U}_{ig}$ are independently $\mathrm{N}(\mathbf{0}, \mathbf{I}_q)$, and are independent of the $\boldsymbol{\varepsilon}_{ig}$, which are independently $\mathrm{N}(\mathbf{0}, \boldsymbol{\Psi}_g)$, where $\boldsymbol{\Psi}_g$ is a $p \times p$ diagonal matrix. It follows that the density of the mixture of factor analyzers model is

$$f(\mathbf{x}_i \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g \phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g), \tag{6}$$

where $\boldsymbol{\vartheta}$ denotes the model parameters. Ghahramani and Hinton (1997) were the first to introduce a mixture of factor analyzers model. In their model, they constrain $\boldsymbol{\Psi}_g = \boldsymbol{\Psi}$ to facilitate an interpretation of $\boldsymbol{\Psi}$ as sensor noise; however, they note that it is possible to relax this constraint. Tipping and Bishop (1997, 1999) introduce the closely related mixture of probabilistic principal component analyzers (MPPCA) model, where the $\boldsymbol{\Psi}_g$ matrix in each component is isotropic, i.e., $\boldsymbol{\Psi}_g = \psi_g\mathbf{I}_p$, so that $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \psi_g\mathbf{I}_p$. McLachlan and Peel (2000b) use the unconstrained mixture of factor analyzers, i.e., with $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g$.

One can view the mixture of factor analyzers models and the probabilistic principal component analyzers model, collectively, as a family of three models, where two members arise from imposing constraints on the most general model, i.e., the model with $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g$. This family can easily be extended to a four-member family by considering the model with component covariance $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \psi\mathbf{I}_p$. A greater level of parsimony can be attained by constraining the component factor loading matrices to be equal, i.e., $\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$. McNicholas and Murphy (2005, 2008) develop a family of eight parsimonious Gaussian mixture models (PGMMs) for clustering by imposing, or not, each of the constraints $\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$, $\boldsymbol{\Psi}_g = \boldsymbol{\Psi}$, and $\boldsymbol{\Psi}_g = \psi_g\mathbf{I}_p$ upon the component covariance structure in the most general mixture of factor analyzers model, i.e., $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g$. Members of the PGMM family have between $pq - q(q-1)/2 + 1$ and $G[pq - q(q-1)/2] + Gp$ free parameters in the component covariance matrices (cf. Table 2), i.e., all members have $\mathcal{O}(p)$ covariance parameters.

Note that McNicholas and Murphy (2010b) further parameterize the mixture of factor analyzers component covariance structure by writing $\boldsymbol{\Psi}_g = \omega_g\boldsymbol{\Delta}_g$, where $\omega_g \in \mathbb{R}^+$ and $\boldsymbol{\Delta}_g$ is a diagonal matrix with $|\boldsymbol{\Delta}_g| = 1$. The result is the addition of four more models to the PGMM family; again, all have $\mathcal{O}(p)$ covariance parameters. Parameter estimation for members of the PGMM family can be carried out using alternating expectation-conditional maximization (AECM) algorithms (Meng and van Dyk 1997). The expectation-conditional maximization (ECM) algorithm (Meng and Rubin 1993) is a variant of the EM algorithm that replaces the M-step by a series of conditional maximization steps. The AECM algorithm allows a different specification of complete-data for each conditional maximization step. This makes it a convenient approach for the PGMM models, where there are two sources of missing data: the unknown component membership labels $z_{ig}$ and the latent factors $\mathbf{u}_{ig}$. Details of fitting the AECM algorithm for the more general mixture of factor analyzers model are given by McLachlan and Peel

Table 2. The nomenclature, covariance structure, and number of free covariance parameters for each member of the PGMM family, where “C” denotes “constrained”, i.e., the constraint is imposed, and “U” denotes “unconstrained”, i.e., the constraint is not imposed.

| $\mathbf{\Lambda}_g = \mathbf{\Lambda}$ | $\mathbf{\Psi}_g = \mathbf{\Psi}$ | $\mathbf{\Psi}_g = \psi_g \mathbf{I}_p$ | $\mathbf{\Sigma}_g$ | Free Cov. Paras. |
|---|---|---|---|---|
| C | C | C | $\mathbf{\Lambda}\mathbf{\Lambda}' + \psi \mathbf{I}_p$ | $pq - q(q-1)/2 + 1$ |
| C | C | U | $\mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}$ | $pq - q(q-1)/2 + p$ |
| C | U | C | $\mathbf{\Lambda}\mathbf{\Lambda}' + \psi_g \mathbf{I}_p$ | $pq - q(q-1)/2 + G$ |
| C | U | U | $\mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}_g$ | $pq - q(q-1)/2 + Gp$ |
| U | C | C | $\mathbf{\Lambda}_g\mathbf{\Lambda}_g' + \psi \mathbf{I}_p$ | $G[pq - q(q-1)/2] + 1$ |
| U | C | U | $\mathbf{\Lambda}_g\mathbf{\Lambda}_g' + \mathbf{\Psi}$ | $G[pq - q(q-1)/2] + p$ |
| U | U | C | $\mathbf{\Lambda}_g\mathbf{\Lambda}_g' + \psi_g \mathbf{I}_p$ | $G[pq - q(q-1)/2] + G$ |
| U | U | U | $\mathbf{\Lambda}_g\mathbf{\Lambda}_g' + \mathbf{\Psi}_g$ | $G[pq - q(q-1)/2] + Gp$ |

(2000a), and parameter estimation for other members of the PGMM family is discussed by McNicholas (2016). The pgmm package (McNicholas, ElSherbiny, McDaid and Murphy 2015) for R implements all twelve PGMM models for model-based clustering and classification.

Much other work has been carried out around, and building on, the mixture of factor analyzers model (e.g., Galimberti, Montanari, and Viroli 2009; Viroli 2010; Montanari and Viroli 2010a, 2011). Baek, McLachlan, and Flack (2010) argue that there may be situations where the mixture of factor analyzers model is not sufficiently parsimonious. They postulate that this might happen when $p$, $G$, or both are not small. The same concern might also apply to other members of the PGMM family. To counter this concern, Baek and McLachlan (2008) and Baek et al. (2010) build on the work of Yoshida et al. (2004, 2006) to introduce a mixture of common factor analyzers (MCFA) model. This model assumes that $\mathbf{X}_i$ can be modelled as

$$\mathbf{X}_i = \mathbf{\Lambda}\mathbf{U}_{ig} + \boldsymbol{\varepsilon}_{ig} \tag{7}$$

with probability $\pi_g$, for $i = 1, \ldots, n$ and $g = 1, \ldots, G$, where $\mathbf{\Lambda}$ is a $p \times q$ matrix of factor loadings, the $\mathbf{U}_{ig}$ are independently $\mathrm{N}(\boldsymbol{\xi}_g, \mathbf{\Omega}_g)$ and are independent of the $\boldsymbol{\varepsilon}_{ig}$, which are independently $\mathrm{N}(\mathbf{0}, \mathbf{\Psi})$, where $\mathbf{\Psi}$ is a $p \times p$ diagonal matrix. Note that $\boldsymbol{\xi}_g$ is a $q$-dimensional vector and $\mathbf{\Omega}_g$ is a $q \times q$ covariance matrix. It follows that the density of the MCFA model is given by

$$f(\mathbf{x}_i \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g \phi(\mathbf{x}_i \mid \mathbf{\Lambda}\boldsymbol{\xi}_g, \mathbf{\Lambda}\mathbf{\Omega}_g\mathbf{\Lambda}' + \mathbf{\Psi}), \tag{8}$$

where $\boldsymbol{\vartheta}$ denotes the model parameters. Noting that

$$\left[\begin{array}{c} \mathbf{X}_i \\ \mathbf{U}_{ig} \end{array}\right] \Bigg|_{z_{ig}=1} \backsim \mathrm{N}\left(\left[\begin{array}{c} \mathbf{\Lambda}\boldsymbol{\xi}_g \\ \boldsymbol{\xi}_g \end{array}\right], \left[\begin{array}{cc} \mathbf{\Lambda}\mathbf{\Omega}_g\mathbf{\Lambda}' + \mathbf{\Psi} & \mathbf{\Lambda}\mathbf{\Omega}_g \\ \mathbf{\Omega}_g\mathbf{\Lambda}' & \mathbf{\Omega}_g \end{array}\right]\right),$$

an EM algorithm can be developed for the MCFA model; see Baek et al. (2010). The MCFA model places additional restrictions on the component means and covariance matrices compared to the mixture of factor analyzers model, thereby further reducing the number of parameters to be estimated. Consequently, the model is quite restrictive and, notably, is much more restrictive than the mixture of factor analyzers model. In fact, the MCFA model can be cast as a special case of the mixture of factor analyzers model (cf. Baek et al. 2010). Other than situations in which the number of components $G$, the number of variables $p$, or both are very large, the mixture of factor analyzers model, or another member of the PGMM family, will almost certainly be preferable to the MCFA model.

Bhattacharya and McNicholas (2014) observe that for even moderately large values of $p$, the BIC can fail to select the number of components and the number of latent factors for the members of the PGMM family. Recognizing this problem, Bhattacharya and McNicholas (2014) consider a LASSO-penalized likelihood approach and proceed to show that the LASSO-penalized BIC (LPBIC) can be used to effectively select the number of components in high dimensions, where the BIC fails. Specifically, they use the penalized log-likelihood

$$\log \mathcal{L}_{\text{pen}}(\boldsymbol{\vartheta}) = \log \left\{ \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_g \phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \right\} - n\lambda_n \sum_{g=1}^{G} \pi_g \sum_{j=1}^{p} |\mu_{gj}|, \tag{9}$$

where $\mu_{gj}$ is the $j$th element in $\boldsymbol{\mu}_g$ and $\lambda_n$ is a tuning parameter that depends on $n$. Following Heiser (1995) and others, Bhattacharya and McNicholas (2014) locally approximate the penalty using a quadratic function. Details on the derivation of the LPBIC and on parameter estimation from the associated penalized likelihood are given in Bhattacharya and McNicholas (2014).

## 4. Departure from Gaussian Mixtures

### 4.1 Mixtures of Components with Varying Tailweight

The first, and perhaps most natural, departure from the Gaussian mixture model is the mixture of multivariate $t$-distributions. McLachlan and Peel (1998) and Peel and McLachlan (2000) motivate the $t$-distribution as a heavy-tailed alternative to the Gaussian distribution. The component density for the mixture of $t$-distributions is

$$f_t(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \nu_g) = \frac{\Gamma\left([\nu_g + p]/2\right) |\boldsymbol{\Sigma}_g|^{-1/2}}{(\pi\nu_g)^{p/2}\Gamma\left(\nu_g/2\right) \left[1 + \delta(\mathbf{x}, \boldsymbol{\mu}_g \mid \boldsymbol{\Sigma}_g)/\nu_g\right]^{(\nu_g+p)/2}}, \tag{10}$$

with mean $\boldsymbol{\mu}_g$, scale matrix $\boldsymbol{\Sigma}_g$, and degrees of freedom $\nu_g$, and where $\delta(\mathbf{x}, \boldsymbol{\mu}_g \mid \boldsymbol{\Sigma}_g) = (\mathbf{x} - \boldsymbol{\mu}_g)'\boldsymbol{\Sigma}_g^{-1}(\mathbf{x} - \boldsymbol{\mu}_g)$ is the squared Mahalanobis distance between $\mathbf{x}$ and $\boldsymbol{\mu}_g$. The mixture of $t$-distributions model has only $G$ more free parameters than the mixture of Gaussian distributions. Andrews and McNicholas (2011a) consider the option to constrain degrees of freedom to be equal across components, i.e., $\nu_g = \nu$, which can lead to improved classification performance by effectively allowing the borrowing of information across components to estimate the degrees of freedom. Andrews and McNicholas (2012) introduce a $t$-analogue of 12 members of the GPCM family of models by imposing the constraints in Table 1 on the component scale matrices $\boldsymbol{\Sigma}_g$, while also allowing the constraint $\nu_g = \nu$. Analogues of all 14 GPCMs, with the option to constrain $\nu_g = \nu$, are implemented in the `teigen` package (Andrews et al. 2015) for R.

While mixtures of $t$-distributions have been the most popular approach for clustering with heavier tail weight, mixtures of multivariate power exponential (MPE) distributions have emerged as an alternative and are used for clustering by Dang, Browne, and McNicholas (2015). In addition to allowing heavier tails, the MPE distribution also permits lighter tails compared to the Gaussian distribution. Dang et al. (2015) use the parametrization of the MPE distribution given by Gómez, Gómez-Viilegas, and Marin (1998), so that the component density for the mixture of MPEs is given by

$$f(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \beta_g) = k|\boldsymbol{\Sigma}_g|^{-1/2} \exp\left\{-\frac{1}{2}\left[(\mathbf{x}-\boldsymbol{\mu}_g)'\boldsymbol{\Sigma}_g^{-1}(\mathbf{x}-\boldsymbol{\mu}_g)\right]^{\beta_g}\right\}, \tag{11}$$

where

$$k = p\Gamma\left(\frac{p}{2}\right)\left[\pi^{p/2}\Gamma\left(1+\frac{p}{2\beta_g}\right)2^{1+p/2\beta_g}\right]^{-1},$$

$\boldsymbol{\mu}_g$ is the mean, $\boldsymbol{\Sigma}_g$ is the scale matrix, and $\beta_g$ determines the kurtosis. Depending on the value of $\beta_g$, two kinds of distributions can be obtained. For $0 < \beta_g < 1$, a leptokurtic distribution is obtained, which is characterized by a thinner peak and heavy tails compared to the Gaussian distribution. In this case, i.e., $\beta_g \in (0,1)$, the MPE distribution is a scale mixture of Gaussian distributions (Gómez-Sánchez-Manzano, Gómez-Viilegas, and Marin 2008). For $\beta_g > 1$, a platykurtic distribution is obtained, which is characterized by a flatter peak and thin tails compared to the Gaussian distribution. Some well-known distributions arise as special or limiting cases of the MPE distribution, e.g., a double-exponential distribution ($\beta_g = 0.5$), a Gaussian distribution ($\beta_g = 1$), or a multivariate uniform distribution ($\beta_g \to \infty$). Dang et al. (2015) use analogues of some members of the GPCM family, which, along with the option to constrain $\beta_g = \beta$, leads to a family of sixteen mixture models.

Contour plots for the bivariate power exponential distribution illustrate some of the flexibility available for different values of $\beta$ (Figure 2).

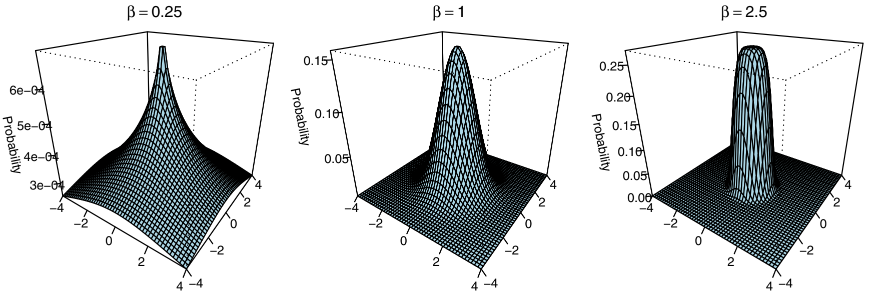

Figure 2. Density plots for the bivariate power exponential distribution for different values of $\beta$.

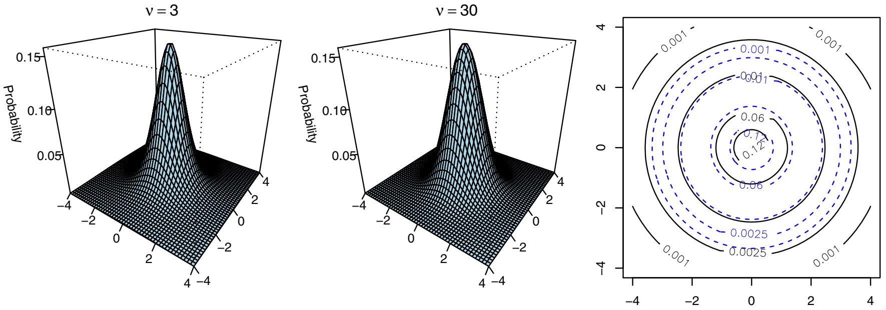

Figure 3. Density plots for the bivariate $t$-distribution for different values of $\nu$ (left and centre) as well as a contour plot reflecting both of these densities (right), where the broken contours represent the density for $\nu = 30$.

Less flexibility is engendered by changing the degrees of freedom parameter $\nu$ in the $t$-distribution, as illustrated by the bivariate density plots in Figure 3. Because the difference in the density plots in Figure 3—which have $v = 3$ and $\nu = 30$ degrees of freedom, respectively—is difficult to discern, a contour plot is also given in Figure 3 to illustrate the heavier tails for $\nu = 3$ degrees of freedom.

## 4.2 Mixtures of Asymmetric Components

Before the turn of the century, almost all work on clustering and classification using mixture models had been based on Gaussian mixture models. A little beyond the turn of the century, work on $t$-mixtures burgeoned into a substantial subfield of mixture model-based classification (e.g., McLachan, Bean, and Jones 2007; Greselin and Ingrassia 2010; Andrews, McNicholas, and Subedi 2011; Andrews and McNicholas 2011a,b, 2012; Baek and McLachlan 2011; McNicholas and Subedi 2012; Steane, McNicholas, and Yada 2012; McNicholas 2013; Lin, McNicholas, and Hsin 2014). Around the same time, work on mixtures of skewed distributions took off, including work on multivariate normal-inverse gamma mixtures (Karlis and Santourian 2009; Subedi and McNicholas 2014; O'Hagen et al. 2016), skew-normal mixtures (e.g., Lin 2009; Montanari and Viroli 2010b; Vrbik and McNicholas 2014), skew-$t$ mixtures (e.g., Lin 2010; Lee and McLachlan 2011; Vrbik and McNicholas 2012, 2014; Murray, McNicholas, and Browne 2014a), shifted asymmetric Laplace mixtures (e.g., Franczak, Browne, and McNicholas 2014), variance-gamma mixtures (McNicholas, McNicholas, and Browne 2014), and generalized hyperbolic mixtures (Browne and McNicholas 2015).

The decision about which mixtures of skewed distributions to focus on herein was partly influenced by the review paper of Lee and McLachlan (2014), who focus on certain formulations of skew-normal and skew-$t$ distributions. A little about these will be said at the end of this section; however, the focus here will be on mixtures of distributions that arise as special or limiting cases of the generalized hyperbolic distribution. Franczak et al. (2014) use a mixture of shifted asymmetric Laplace (SAL) distributions for clustering. The density of a random variable $\mathbf{X}$ from a $p$-dimensional SAL distribution is given by

$$f_{\text{SAL}}\left(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}\right) = \frac{2\exp\{(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}\}}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \left(\frac{\delta\left(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma}\right)}{2+\boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}\right)^{\lambda/2} K_{\lambda}\left(u\right), \tag{12}$$

where $\lambda = (2-p)/2$, $\boldsymbol{\Sigma}$ is a scale matrix, $\boldsymbol{\mu} \in \mathbb{R}^p$ is a location parameter, $\boldsymbol{\alpha} \in \mathbb{R}^p$ is a skewness parameter, $u = \sqrt{(2+\boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha})\delta\left(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma}\right)}$, $K_\lambda$ is the modified Bessel function of the third kind with index $\lambda$, and $\delta\left(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma}\right)$ is as defined before. Crucially, the random variable $\mathbf{X}$ can be generated via

$$\mathbf{X} = \boldsymbol{\mu} + W\boldsymbol{\alpha} + \sqrt{W}\mathbf{V}, \tag{13}$$

where $W \backsim \text{Exp}(1)$ and $\mathbf{V} \backsim \text{N}(\mathbf{0}, \boldsymbol{\Sigma})$ is independent of $W$ (Kotz, Kozubowski, and Podgorski 2001; Franczak et al. 2014). It follows that $W \mid \mathbf{x}$ has a generalized inverse Gaussian distribution (Barndorff-Nielsen 1997); accordingly, the E-steps in the associated EM algorithm are highly tractable (see Franczak et al. 2014, for details). Note that $\text{Exp}(1)$ signifies an exponential distribution with rate 1.

Before proceeding to mixtures of more flexible asymmetric distributions, it is useful to consider the relative performance of SAL mixtures and Gaussian mixtures when clusters are asymmetric. First, consider one component from a SAL distribution (Figure 4). Fitting a Gaussian mixture to this component, a mechanism emerges by which a Gaussian mixture can be used to capture an asymmetric cluster, via multiple components

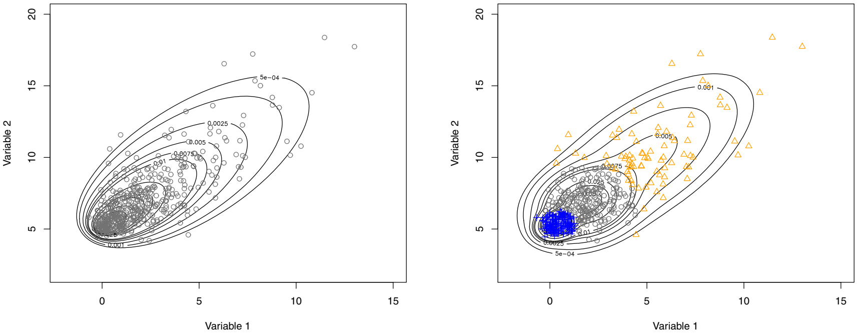

Figure 4. Scatter plots of data from a SAL distribution, with contours from a fitted SAL distribution (left) and contours from a fitted $G = 3$ component Gaussian mixture model, where plotting symbols represent component memberships (right).

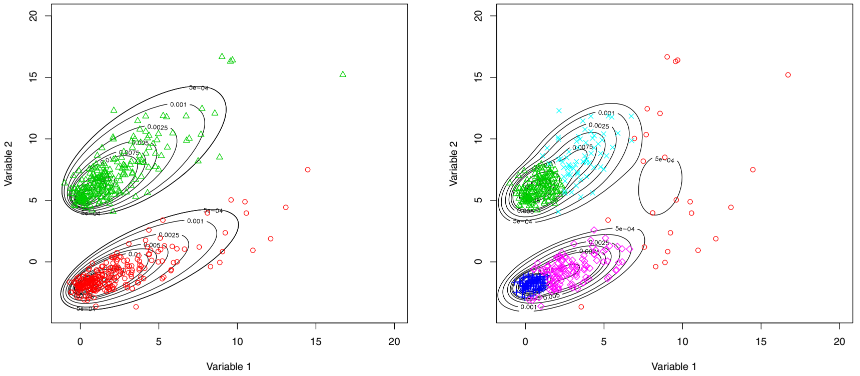

Figure 5. Scatter plots of a two-component mixture of SAL distributions, with contours from a fitted $G = 2$ component SAL mixture (left) and contours from a $G = 5$ component Gaussian mixture (right), where plotting symbols represent predicted classifications in each case.

(Figure 4). As McNicholas (2016) points out, situations such as this are reminiscent of the flame on a candle. Whether this mechanism will work for multiple asymmetric clusters will depend, *inter alia*, on how well the clusters are separated.

Consider the data in Figure 5, where there are two asymmetric clusters that can be separated by a straight line. These data are generated from a $G = 2$ component SAL mixture and so it is not surprising that fitting SAL mixtures to these data leads to the selection of a $G = 2$ component SAL mixture with perfect class agreement (Figure 5). Gaussian mixtures are fitted to these data for $G = 1, \ldots, 6$ components and the BIC selects a $G = 5$ component model; here, the Gaussian components cannot be merged to return the correct clusters because one Gaussian component has been used to capture all points that do not better fit within one of the other four components (Figure 5). While this is obvious by inspection in two dimensions, it would be difficult to detect in higher dimensions. The unsuitability of Gaussian mixtures for capturing asymmetric clusters via *a posteriori* merging has been noted previously (e.g., Franczak et al. 2014; Murray et al. 2014a). This is one reason why it has been said that merging Gaussian components is not a "get out of jail free" card (McNicholas and Browne 2013).

The SAL distribution is a special case of the generalized hyperbolic distribution. McNeil, Frey, and Embrechts (2005) note that a random variable $\mathbf{X}$ following the generalized hyperbolic distribution can be represented via the relationship in (13) with $W$ following a generalized inverse Gaussian distribution. Because of an identifiability issue (cf. Hu 2005), Browne and McNicholas (2015) use a re-parameterization (see Browne and McNicholas 2015, for details), under which the density of the generalized hyperbolic distribution is

$$f_{\mathrm{H}}(\mathbf{x} \mid \boldsymbol{\theta}) = \left[\frac{\omega + \delta\left(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma}\right)}{\omega + \boldsymbol{\beta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}}\right]^{(\lambda - p/2)/2} \times \frac{K_{\lambda - p/2}\left(\sqrt{\left[\omega + \boldsymbol{\beta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}\right]\left[\omega + \delta\left(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma}\right)\right]}\right)}{(2\pi)^{p/2}\left|\boldsymbol{\Sigma}\right|^{1/2} K_{\lambda}\left(\omega\right) \exp\left\{-\left(\mathbf{x} - \boldsymbol{\mu}\right)'\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}\right\}}, \tag{14}$$

where $\lambda$ is an index parameter, $\omega$ is a concentration parameter, $\boldsymbol{\Sigma}$ is a scale matrix, $\boldsymbol{\mu}$ is a location parameter, $\boldsymbol{\beta}$ is a skewness parameter, and $K_\lambda$ is the modified Bessel function of the third kind with index $\lambda$. Similar to the SAL distribution, $W \mid \mathbf{x}$ follows a generalized inverse Gaussian distribution, which facilitates the calculation of expected values in the E-step.

The mixture of factor analyzers model can be extended to the generalized hyperbolic distribution, or any of its special or limiting cases. For the mixture of generalized hyperbolic distributions, the first step is to consider that $\mathbf{V}$ in (13) can be decomposed using a factor analysis model, i.e., $\mathbf{V} = \boldsymbol{\Lambda}\mathbf{U} + \boldsymbol{\varepsilon}$, where $\mathbf{U} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_q)$ and $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Psi})$ in the usual way. The resulting model can be represented as

$$\mathbf{X} = \boldsymbol{\mu} + W\boldsymbol{\alpha} + \sqrt{W}(\boldsymbol{\Lambda}\mathbf{U} + \boldsymbol{\varepsilon}), \tag{15}$$

where $W$ follows a generalized inverse Gaussian distribution. Following this approach, Tortora, McNicholas, and Browne (2015) arrive at a mixture of generalized hyperbolic factor analyzers model. In doing so, they follow the same approach used by Murray et al. (2014a), who develop a mixture of skew-$t$ factor analyzers; the principal difference is the distribution of $W$, which is inverse gamma in the case of the skew-$t$ distribution. Note that there is no skew-normal distribution nested within this formulation of the skew-$t$ distribution because the skewness parameter goes to zero as the degrees of freedom parameter goes to infinity. Interestingly, Murray et al. (2014b) develop a mixture of common skew-$t$ factor analyzers in a similar fashion.

There has been a plethora of work on clustering using non-elliptical distributions beyond the mixture of generalized hyperbolic distributions and special and/or limiting cases thereof. SAL mixtures are attractive as a first departure from symmetric components because they are quite simple models, i.e., only location, scale, and skewness are parameterized in each component. The mixture of generalized hyperbolic distribution, also parameterizing concentration (as well as having an index parameter), is a natural extension. Of course, skew-normal mixtures are just as simple as SAL mixtures; however, as mentioned earlier in this section, Lee and McLachlan (2014) focus on certain formulations of the skew-$t$ and skew-normal distributions in their review of mixtures of non-elliptical distributions. One of these skew-normal formulations is given by Azzalini and Valle (1996) and examined further by Azzalini and Capitanio (1999) and others. Branco and Dey (2001) and Azzalini and Capitanio (2003) introduce an analogous skew-$t$ distribution. The other formulation is given by Sahu, Dey, and Branco (2003), for both skew-normal and skew-$t$ distributions. Extensive details on skew-normal and skew-$t$ distributions are given by Azzalini and Capitanio (2014). Mixtures of these formulations have been used for clustering and classification in several contexts, including work by Lin (2009, 2010), Vrbik and McNicholas (2012, 2014), and Lee and McLachlan (2013a,b, 2014). Vrbik and McNicholas (2014) introduce skew-normal and skew-$t$ analogues of the GPCM family and show that they can give superior clustering and classification performance when compared with their Gaussian counterparts. Azzalini et al. (2016) discuss nomenclature and some other considerations for the formulations used by Lee and McLachlan (2013a,b). Lin, McLachlan, and Lee (2016) discuss a mixture of skew-normal factor analyzers model using the formulation of Azzalini and Valle (1996).

## 5. Dimension Reduction

Suppose $p$-dimensional $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are observed and $p$ is large enough that dimension reduction is required. Note that the vagueness around how large $p$ needs to be before it is considered "large" is intentional and necessary because the answer will depend on several factors including the modelling process and the number of observations. Note also that dimension reduction is often required, or at least helpful, even when $p$ is not large because the presence of variables that are not helpful in discriminating groups can have a deleterious effect on clustering, or classification, performance.

Broadly, there are two ways to carry out dimension reduction: a subset of the $p$ variables can be selected or the data can be mapped to a (much) lower dimensional space. For reasons that will be apparent, the former approach can be referred to as explicit dimension reduction whereas the latter is implicit (McNicholas 2016). The mixture of factor analyzers model, the other members of the PGMM family, and the MCFA model are examples of implicit dimension reduction. However, as mentioned in Section 3, the MCFA approach is not recommended for general use and the PGMM family can be ineffective for larger values of $p$. The latter problem can be (partly) addressed by using a LASSO-penalized likelihood approach and model selection criterion, as discussed in Section 3. There are at least two other implicit dimension reduction techniques that deserve mention (GMMDR and HD-GMM, which will be discussed herein) and, similar to the latent factor-based approach, these methods carry out simultaneous clustering and dimension reduction. As Bouveyron and Brunet-Saumard (2014) point out in their excellent review, carrying out these two elements—clustering and dimension reduction—sequentially does not typically work; they give the particular example of clustering after principal component analysis.

There are a number of explicit approaches by which variables can be selected. Raftery and Dean (2006) propose a variable selection method that utilizes a greedy search of the model space. Their approach is based on Bayes factors. Given data $\mathbf{x}$, the Bayes factor $B_{12}$ for model $M_1$ versus model $M_2$ is

$$B_{12} = \frac{p(\mathbf{x} \mid M_1)}{p(\mathbf{x} \mid M_2)},$$

where

$$p(\mathbf{x} \mid M_k) = \int p(\mathbf{x} \mid \boldsymbol{\theta}_k, M_k) p(\boldsymbol{\theta}_k \mid M_k) d\boldsymbol{\theta}_k,$$

$\boldsymbol{\theta}_k$ is the vector of parameters for model $M_k$, and $p(\boldsymbol{\theta}_k \mid M_k)$ is the prior distribution of $M_k$ (Kass and Raftery 1995). The approach of Raftery and Dean (2006) simultaneously selects a variable subset, the number of components, and the model, i.e., the GPCM covariance structure (Table 1). This approach is implemented within the `clustvarsel` package (Dean, Raftery, and Scrucca 2012) for R, and it can work well in some situations. However, because the number of free model parameters for some of the GPCM models is quadratic in data-dimensionality, `clustvarsel` is largely ineffective in high dimensions. A related approach is described by Maugis, Celeux, and Martin-Magniette (2009a,b) and implemented as the `selvarclust` software (Maugis 2009), which is a command-line addition to the MIXMOD software (Biernacki et al. 2006). This approach relaxes the assumptions on the role of variables with the potential benefit of avoiding the over-penalization of independent variables.

More recently, the VSCC (variable selection for clustering and classification) approach (Andrews and McNicholas 2014) has been developed and used in the same situation. The VSCC technique finds a subset of variables that simultaneously minimizes the within-group variance and maximizes the between-group variance, thereby resulting in variables that show separation between the desired groups. The within-group variance for variable $j$ can be written

$$\mathcal{W}_j = \frac{\sum_{g=1}^{G} \sum_{i=1}^{n} z_{ig}(x_{ij} - \mu_{gj})^2}{n},$$

where $x_{ij}$ is the value of variable $j$ for observation $i$, $\mu_{gj}$ is the mean of variable $j$ in component $g$, and $n$ and $z_{ig}$ have the usual meanings. The variance within variable $j$ that is not accounted for by $\mathcal{W}_j$, i.e., $\sigma_j^2 - \mathcal{W}_j$, provides an indication of the variance between groups. In general, calculation of this residual variance is needed; however, if the data have been standardized to have equal variance across variables, then any variable minimizing the within-group variance is also maximizing the leftover variance. Accordingly, Andrews and McNicholas (2014) describe the VSCC method in terms of data where the variables have been standardized to have zero mean and unit variance. The VSCC approach also uses the correlation between variables, which is denoted $\rho_{jk}$ for variables $j$ and $k$. If $V$ represents the space of currently selected variables, then variable $j$ is selected if $|\rho_{jr}| < 1 - \mathcal{W}_j^m$ for all $r \in V$, where $m \in \{1, \ldots, 5\}$ is fixed. When VSCC is used for clustering, it is necessary to choose between these subsets without specific knowledge of which subset produces the best classifier. Andrews and McNicholas (2014) choose the subset that minimizes

$$\sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} - \sum_{i=1}^{n} \max_g \{\hat{z}_{ig}\} = n - \sum_{i=1}^{n} \max_g \{\hat{z}_{ig}\},$$

which is equivalent to maximizing $\sum_{i=1}^{n} \max_g \{\hat{z}_{ig}\}$. When VSCC is used for clustering, the first step is to carry out an initial clustering using a model-based or other method. VSCC is a step-wise approach and further details are given by Andrews and McNicholas (2014). An implementation of the VSCC approach is given in the `vscc` package (Andrews and McNicholas 2013) for R.

Similar to `clustvarsel` and `selvarclust`, the Gaussian mixture modelling and dimension reduction (GMMDR) approach (Scrucca 2010) is based on the GPCM family of models, and builds on the sliced inverse regression work of Li (1991, 2000). The idea behind GMMDR is to find the smallest subspace that captures the clustering information contained within the data. To do this, GMMDR seeks those directions where the cluster means $\boldsymbol{\mu}_g$ and the cluster covariances $\boldsymbol{\Sigma}_g$ vary the most, provided that each direction is $\boldsymbol{\Sigma}$-orthogonal to the others. These directions can be found via the generalized eigen-decomposition of the kernel matrix $\mathbf{M}\mathbf{v}_i = l_i \boldsymbol{\Sigma}\mathbf{v}_i$, where $l_1 \geq l_2 \geq \cdots \geq l_d > 0$, and $\mathbf{v}_i'\boldsymbol{\Sigma}\mathbf{v}_j = 1$ if $i = j$ and $\mathbf{v}_i'\boldsymbol{\Sigma}\mathbf{v}_j = 0$ otherwise (Scrucca 2010). Note that there are $d \leq p$ directions that span the subspace. The kernel matrix contains the variations in cluster means

$$\mathbf{M}_{\mathrm{I}} = \sum_{g=1}^{G} \pi_g (\boldsymbol{\mu}_g - \boldsymbol{\mu})(\boldsymbol{\mu}_g - \boldsymbol{\mu})'$$

and the variations in cluster covariances

$$\mathbf{M}_{\mathrm{II}} = \sum_{g=1}^{G} \pi_g (\boldsymbol{\Sigma}_g - \bar{\boldsymbol{\Sigma}})\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}_g - \bar{\boldsymbol{\Sigma}})',$$

such that $\mathbf{M} = \mathbf{M}_{\mathrm{I}}\boldsymbol{\Sigma}^{-1}\mathbf{M}_{\mathrm{I}} + \mathbf{M}_{\mathrm{II}}$. Note that

$$\boldsymbol{\mu} = \sum_{g=1}^{G} \pi_g \boldsymbol{\mu}_g \quad \text{and} \quad \boldsymbol{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})'$$

are the global mean and global covariance matrix, respectively, and $\bar{\boldsymbol{\Sigma}} = \sum_{g=1}^{G} \pi_g \boldsymbol{\Sigma}_g$ is the pooled within-cluster covariance matrix.

The GMMDR directions are the eigenvectors $(\mathbf{v}_1, \ldots, \mathbf{v}_d) \equiv \boldsymbol{\beta}$. These eigenvectors, ordered according to the eigenvalues, form the basis of the dimension reduction subspace $\mathcal{S}(\boldsymbol{\beta})$. The projections of the mean and covariance onto $\mathcal{S}(\boldsymbol{\beta})$ are given by $\boldsymbol{\beta}'\boldsymbol{\mu}_g$ and $\boldsymbol{\beta}'\boldsymbol{\Sigma}_g\boldsymbol{\beta}$, respectively. The GMMDR variables are the projections of the $p$-dimensional data $(\mathbf{x}_1', \ldots, \mathbf{x}_n')'$ onto the subspace $\mathcal{S}(\boldsymbol{\beta})$ and can be computed as $(\mathbf{x}_1', \ldots, \mathbf{x}_n')'\boldsymbol{\beta}$. This estimation of GMMDR variables is a sort of feature extraction. Moreover, some of the estimated GMMDR variables may provide no clustering information and need to be removed. Scrucca (2010) removes them via a modified version of the variable selection method of Raftery and Dean (2006). Scrucca (2014) extends the GMMDR approach to model-based discriminant analysis, and Morris and McNicholas (2016) apply GMMDR for model-based classification and model-based discriminant analysis. Morris, McNicholas, and Scrucca (2013) and Morris and McNicholas (2013, 2016) extend the GMMDR to mixtures of non-Gaussian distributions.

Bouveyron, Girard, and Schmid (2007a,b) introduce a family of 28 parsimonious, flexible Gaussian models specifically designed for high-dimensional data. This family, called HD-GMM, can be applied for

Table 3. Nomenclature and the number of free covariance parameters for 16 members of the HD-GMM family.

| Model | Number of Free Covariance Parameters |
|---|---|
| $[a_{gj}b_g\mathbf{\Gamma}_g d_g]$ | $\sum_{g=1}^{G} d_g[p-(d_g+1)/2]+\sum_{g=1}^{G} d_g+2G$ |
| $[a_{gj}b\mathbf{\Gamma}_g d_g]$ | $\sum_{g=1}^{G} d_g[p-(d_g+1)/2]+\sum_{g=1}^{G} d_g+1+G$ |
| $[a_g b_g\mathbf{\Gamma}_g d_g]$ | $\sum_{g=1}^{G} d_g[p-(d_g+1)/2]+3G$ |
| $[ab_g\mathbf{\Gamma}_g d_g]$ | $\sum_{g=1}^{G} d_g[p-(d_g+1)/2]+1+2G$ |
| $[a_g b\mathbf{\Gamma}_g d_g]$ | $\sum_{g=1}^{G} d_g[p-(d_g+1)/2]+1+2G$ |
| $[ab\mathbf{\Gamma}_g d_g]$ | $\sum_{g=1}^{G} d_g[p-(d_g+1)/2]+2+G$ |
| $[a_{gj}b_g\mathbf{\Gamma}_g d]$ | $Gd[p-(d+1)/2]+Gd+G+1$ |
| $[a_j b_g\mathbf{\Gamma}_g d]$ | $Gd[p-(d+1)/2]+d+G+1$ |
| $[a_{gj}b\mathbf{\Gamma}_g d]$ | $Gd[p-(d+1)/2]+Gd+2$ |
| $[a_j b\mathbf{\Gamma}_g d]$ | $Gd[p-(d+1)/2]+d+2$ |
| $[a_g b_g\mathbf{\Gamma}_g d]$ | $Gd[p-(d+1)/2]+2G+1$ |
| $[ab_g\mathbf{\Gamma}_g d]$ | $Gd[p-(d+1)/2]+G+2$ |
| $[a_g b\mathbf{\Gamma}_g d]$ | $Gd[p-(d+1)/2]+G+2$ |
| $[ab\mathbf{\Gamma}_g d]$ | $Gd[p-(d+1)/2]+3$ |
| $[a_j b\mathbf{\Gamma} d]$ | $d[p-(d+1)/2]+d+2$ |
| $[ab\mathbf{\Gamma} d]$ | $d[p-(d+1)/2]+3$ |

clustering or classification. The HD-GMM family is based on an eigen-decomposition of the component covariance matrices $\mathbf{\Sigma}_g$, which can be written

$$\mathbf{\Sigma}_g = \mathbf{\Gamma}_g\mathbf{\Delta}_g\mathbf{\Gamma}'_g,$$

where $\mathbf{\Gamma}_g$ is a $p \times p$ orthogonal matrix of eigenvectors of $\mathbf{\Sigma}_g$ and $\mathbf{\Delta}_g$ is a $p \times p$ diagonal matrix containing the corresponding eigenvalues, in decreasing order. The idea behind the HD-GMM family is to re-parametrize $\mathbf{\Delta}_g$ such that $\mathbf{\Sigma}_g$ has only $d_g + 1$ distinct eigenvalues. This is achieved via

$$\mathbf{\Delta}_g = \text{diag}\{a_{g1},\ldots,a_{gd_g},b_g,\ldots,b_g\},$$

where the first $d_g < p$ values $a_{g1},\ldots,a_{gd_g}$ represent the variance in the component-specific subspace and the other $p - d_g$ values $b_g$ are the variance of the noise. The key assumption is that, conditional on the components, the noise variance for component $g$ is isotropic and is within a subspace that is orthogonal to the subspace of the $g$th component. Although there are 28 HD-GMM models, the 16 with closed form estimators are often focused upon (Table 3).

As Bouveyron and Brunet-Saumard (2014) point out, the HD-GMM family can be regarded as a generalization of the GPCM family or as a generalization of the MPPCA model. For instance, if $d_g = p - 1$ then the HD-GMM model $[a_{gj}b_g\mathbf{\Gamma}_g d_g]$ is the same as the GPCM model VVV. Further, the HD-GMM model $[a_{gj}b_g\mathbf{\Gamma}_g d]$ is equivalent to the MPPCA model.

For the HD-GMM model $[a_g b_g \mathbf{\Gamma}_g d_g]$, Bouveyron et al. (2011) show that the maximum likelihood estimate of the $d_g$ is asymptotically consistent, a fact that has consequences for inference for isotropic PPCA (cf. Bouveyron et al. 2011).

## 6. Robust Clustering

In real applications, one may encounter data that are contaminated by outliers, noise, or generally spurious points. Borrowing the terminology used by Aitkin and Wilson (1980), these types of observations shall be collectively referred to as "bad" while all others will be called "good". When bad points are present, they can have a deleterious effect on mixture model parameter estimation. Accordingly, it is generally desirable to account for bad points when present. One way to do this is to use a mixture of distributions with component concentration parameters. Some such mixtures have already been considered herein and include $t$-mixtures and power exponential mixtures; however, Hennig (2004) points out that $t$-mixtures are vulnerable to "very extreme outliers" and the same is probably true for robustness-via-component concentration parameter approaches in general.

Within the Gaussian mixture paradigm, Campbell (1984), McLachlan and Basford (1988), Kharin (1996), and De Veaux and Krieger (1990) achieve a similar effect by using M-estimators (Huber 1964, 1981) of the means and covariance matrices of the Gaussian components of the mixture model. In a similar vein, Markatou (2000) utilizes a weighted likelihood approach to obtain robust parameter estimates. Banfield and Raftery (1993) add a uniform component on the convex hull of the data to accommodate outliers in a Gaussian mixture model, and Fraley and Raftery (1998) and Schroeter et al. (1998) further consider approaches in this direction. Hennig (2004) suggests adding an improper uniform distribution as an additional mixture component. Browne, McNicholas, and Sparling (2012) also make use of uniform distributions but they do so by making each component a mixture of a Gaussian and a uniform distribution. Rather than specifically accommodating bad points, this approach allows for what they call "bursts" of probability as well as locally heavier tails—this might have the effect of dealing with bad points for some data sets. Coretto and Hennig (2015) use an optimally tuned improper maximum likelihood estimator for robust clustering.

García-Escudero et al. (2008) outline a trimmed clustering approach that gives robust parameter estimates by allowing for a pre-specified proportion of bad points. They achieve this by imposing restrictions on the ratio between the maximum and minimum eigenvalues of the component covariance matrices. These constraints can be viewed as a multivariate extension of the univariate work of Hathaway (1985). The trimmed clustering approach of García-Escudero et al. (2008) has been applied for Gaussian mixtures and is implemented as such in the R package `tclust` (Fritz, García-Escudero, and Mayo-Iscar 2012). The approach can be very effective when the number of variables $p$ is sufficiently small so that the proportion of bad points can be accurately pre-specified. Although work to date has focused somewhat on Gaussian mixtures, a similar approach could be taken to mixtures with non-elliptical components.

Punzo and McNicholas (2016) use a mixture of contaminated Gaussian distributions, with density of the form

$$f\left(\mathbf{x} \mid \boldsymbol{\vartheta}\right) = \sum_{g=1}^{G} \pi_g \left[\alpha_g \phi\left(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\right) + \left(1-\alpha_g\right) \phi\left(\mathbf{x} \mid \boldsymbol{\mu}_g, \eta_g \boldsymbol{\Sigma}_g\right)\right], \tag{16}$$

where $\alpha_g \in (0,1)$ is the proportion of good points in the $g$th component and $\eta_g > 1$ is the degree of contamination. Because $\eta_g > 1$ is an inflation parameter, it can be interpreted as the increase in variability due to the bad observations. This contaminated Gaussian mixture approach, i.e., (16), can be viewed as a special case of the multi-layer mixture of Gaussian distributions of Li (2005), where each of the $G$ components at the top layer is itself a mixture of two components, with equal means and proportional covariance matrices at the secondary layer. One advantage of the mixture of contaminated Gaussian distributions approach is that the proportion of bad points does not need to be specified *a priori* (cf. Punzo and McNicholas 2016). As a result, it is possible to apply this approach to higher dimensional data and even to high-dimensional data, e.g., via a mixtures of contaminated Gaussian factor analyzers model (Punzo and McNicholas 2014b).

## 7. Clustering Longitudinal Data

McNicholas and Murphy (2010a) use a Gaussian mixture model with a modified Cholesky-decomposed covariance structure to cluster longitudinal data. The Cholesky decomposition is a well-known method for decomposing a matrix into the product of a lower triangular matrix and its transpose. Let $\mathbf{A}$ be a real, positive definite matrix, then the Cholesky decomposition of $\mathbf{A}$ is given by $\mathbf{A} = \mathbf{L}\mathbf{L}'$, where $\mathbf{L}$ is a unique lower triangular matrix. This decomposition is popular in numerical analysis applications, where it can be used to simplify the solution to a linear system of equations.

A modified Cholesky decomposition can be applied to a covariance matrix, and Pourahmadi (1999, 2000) exploits the fact that covariance matrix $\boldsymbol{\Sigma}$ of a random variable can be decomposed using the relation $\mathbf{T}\boldsymbol{\Sigma}\mathbf{T}' = \mathbf{D}$, where $\mathbf{T}$ is a unique unit lower triangular matrix and $\mathbf{D}$ is a unique diagonal matrix with positive diagonal entries. This relationship can also be written $\mathbf{\Sigma}^{-1} = \mathbf{T}'\mathbf{D}^{-1}\mathbf{T}$. The values of $\mathbf{T}$ and $\mathbf{D}$ have interpretations as generalized autoregressive parameters and innovation variances, respectively (Pourahmadi 1999), such that the linear least-squares predictor of $X_t$, based on $X_{t-1}, \ldots, X_1$, can be written

$$\hat{X}_t = \mu_t + \sum_{s=1}^{t-1}(-\phi_{ts})(X_s - \mu_s) + \sqrt{d_t}\epsilon_t, \qquad (17)$$

where $\epsilon_t \sim \mathrm{N}(0,1)$, the $\phi_{ts}$ are the sub-diagonal elements of $\mathbf{T}$, and the $d_t$ are the diagonal elements of $\mathbf{D}$. Pan and MacKenzie (2003) use the modified Cholesky decomposition to jointly model the mean and covariance in longitudinal studies. Pourahmadi, Daniels, and Park (2007) develop a method of simultaneously modelling several covariance matrices based on this decomposition, thereby giving an alternative to common principal components analysis (Flury 1988) for longitudinal data.

McNicholas and Murphy (2010a) consider a Gaussian mixture model with a modified Cholesky-decomposed covariance structure for each mixture component, so that the $g$th component density is

$$\phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g, (\mathbf{T}_g'\mathbf{D}_g^{-1}\mathbf{T}_g)^{-1}) = \frac{1}{\sqrt{(2\pi)^p|\mathbf{D}_g|}} \exp\left\{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_g)'\mathbf{T}_g'\mathbf{D}_g^{-1}\mathbf{T}_g(\mathbf{x}_i - \boldsymbol{\mu}_g)\right\},$$

where $\mathbf{T}_g$ and $\mathbf{D}_g$ are the $p \times p$ lower triangular matrix and the $p \times p$ diagonal matrix, respectively, that follow from the modified Cholesky decomposition of $\mathbf{\Sigma}_g$.

A family of eight Gaussian mixture models arises from the option to constrain $\mathbf{T}_g$ and/or $\mathbf{D}_g$ to be equal across components together with the option to impose the isotropic constraint $\mathbf{D}_g = \delta_g\mathbf{I}_p$. This family is known as the Cholesky-decomposed Gaussian mixture model (CDGMM) family. Each member of the CDGMM family (Table 4) has a natural interpretation for longitudinal data. Constraining $\mathbf{T}_g = \mathbf{T}$ suggests that the autoregressive relationship between time points, cf. (17), is the same across components. The constraint $\mathbf{D}_g = \mathbf{D}$ means that the variability at each time point is taken to be the same for each component, and the isotropic constraint $\mathbf{D}_g = \delta_g\mathbf{I}_p$ suggests that the variability is the same at each time point in component $g$. From a clustering point of view, two of the CDGMMs have equivalent GPCM models; however, even though this equivalence exists, the GPCM models in question (EEE and VVV) do not explicitly account for the longitudinal correlation structure. McNicholas and Murphy (2010a) also consider cases where elements below a given sub-diagonal of $\mathbf{T}_g$ are

Table 4. The nomenclature, covariance structure, and number of free covariance parameters for each member of the CDGMM family.

| Model | $\mathbf{T}_g$ | $\mathbf{D}_g$ | $\mathbf{D}_g$ | Free Cov. Parameters |
|---|---|---|---|---|
| EEA | Equal | Equal | Anisotropic | $p(p-1)/2+p$ |
| VVA | Variable | Variable | Anisotropic | $G[p(p-1)/2]+Gp$ |
| VEA | Variable | Equal | Anisotropic | $G[p(p-1)/2]+p$ |
| EVA | Equal | Variable | Anisotropic | $p(p-1)/2+Gp$ |
| VVI | Variable | Variable | Isotropic | $G[p(p-1)/2]+G$ |
| VEI | Variable | Equal | Isotropic | $G[p(p-1)/2]+1$ |
| EVI | Equal | Variable | Isotropic | $p(p-1)/2+G$ |
| EEI | Equal | Equal | Isotropic | $p(p-1)/2+1$ |

set to zero. This constrained correlation structure can be used to remove autocorrelation over large time lags.

The CDGMM models have been used effectively in real data analyses (e.g., Humbert et al. 2013) and they have been extended in a number of directions. McNicholas and Subedi (2012) consider a $t$-analogue of the CDGMM family. They also consider a linear model for the mean but another model could be implemented in a similar framework; these models, together with the CDGMM family, are available in the `longclust` package (McNicholas, Jampani, and Subedi 2015) for R. Anderlucci and Viroli (2015) extend the methodology of McNicholas and Murphy (2010a) to the situation where there are multiple responses for each individual at each time point. Their approach is nicely illustrated with data from a health and retirement study. The notion of constraining sub-diagonals of $\mathbf{T}_g$ deserves some further attention, both within the single- and multiple-response paradigms. It will also be interesting to explore the use of mixtures of MPEs as an alternative to $t$-mixtures; whereas $t$-mixtures essentially allow more dispersion about the mean when compared with Gaussian mixtures, mixtures of MPEs would allow both more and less dispersion.

## 8. Clustering Categorical and Mixed Type Data

Latent class analysis has been widely used for clustering of categorical data and data of mixed type (e.g. Goodman 1974; Celeux and Govaert 1991; Biernacki, Celeux, and Govaert 2010). Much work on refinement and extension has been carried out. For example, Vermunt (2003, 2007) develop a multilevel latent class models to account for conditional dependency between the response variables, and Marbac, Biernacki, and Vanderwalle (2014) propose a conditional modes model that assigns response variables into conditionally independent blocks. Besides latent class analysis, mixture model-based approaches for categorical data have received relatively little attention within the literature. Browne and McNicholas (2012) develop a mixture of latent variables model for clustering of data with mixed type, and a data set comprising only categorical (including binary) variables fits within their modelling framework as a special case. Browne and McNicholas (2012) draw on the deterministic annealing approach of Zhou and Lange (2010) in their parameter estimation scheme. This approach can increase the chance of finding the global maximum but Gauss-Hermite quadrature is required to approximate the likelihood. Gollini and Murphy (2014) use a mixture of latent trait analyzers (MLTA) model to cluster categorical data. They also apply their approach to binary data, where a categorical latent variable identifies clusters of observations and a latent trait is used to accommodate within-cluster dependency. A lower bound approximation to the log-likelihood is used, which is straightforward to implement and converges relatively quickly compared with other numerical approximations to the likelihood.

A mixture of item response models (Muthen and Asparouhov 2006; Vermunt 2007) has very similar structure to the MLTA model; however, it is highly parameterized, uses a probit structure, and numerical integration is required to compute the likelihood. A similar approach has also been discussed by Cagnone and Viroli (2012), who use Gauss-Hermite quadrature to approximate the likelihood; they also assume a semi-parametric distributional form for the latent variables by adding extra parameters to the model. Repeatedly sampled binary data can be clustered using multilevel mixture item response models (Vermunt 2007). McParland et al. (2014) use a mixture model approach for mixed categorical data (binary, ordinal, and nominal), where each component is effectively a hybrid of an item response model and a factor analysis model.

Tang, Browne, and McNicholas (2015) propose two mixtures of latent traits models with common slope parameters for model-based clustering of binary data. One is a general model that supposes that the dependence among the response variables within each observation is wholly explained by a low-dimensional continuous latent variable in each component. The other is specifically designed for repeatedly sampled data and supposes that the response function in each component is composed of two continuous latent variables by adding a blocking latent variable. Their proposed mixture of latent trait models with common slope parameters (MCLT) model is a categorical analogue of the MCFA model of Baek et al. (2010). The MCLT model allows for significant reduction in the number of free parameters when estimating the slope. Moreover, it facilitates a low-dimensional visual representation of the clusters, where posterior means of the continuous latent variables correspond to the manifest data.

In the mixture of latent traits model, the likelihood function involves an integral that is intractable. Tang et al. (2015) propose using a variational approximation to the likelihood, as proposed by Jaakkola and Jordan (2000), Tipping (1999), and Attias (2000). For a fixed set of values for the variational parameters, the transformed problem has a closed-form solution, providing a lower bound approximation to the log-likelihood. The variational parameters are optimized in a separate step.

Ranalli and Rocci (2016) develop an approach for clustering ordinal data. They use an underlying response variable approach, which treats ordinal variables as categorical realizations of underlying continuous variables (cf. Jöreskog 1990).

## 9. Cluster-Weighted Models

Consider data of the form $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ so that each observation is a realization of the pair $(\mathbf{X}, Y)$ defined on some space $\Omega$, where $Y \in \mathbb{R}$ is a response variable and $\mathbf{X} \in \mathbb{R}^p$ is a vector of covariates. Suppose that $\Omega$ can be partitioned into $G$ groups, say $\Omega_1, \ldots, \Omega_G$, and let $p\,(\mathbf{x}, y)$ be the joint density of $(\mathbf{X}, Y)$. In general, the density of a cluster-weighted model (CWM) can be written

$$p(\mathbf{x}, y \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g p(y \mid \mathbf{x}, \boldsymbol{\theta}_g) p(\mathbf{x} \mid \boldsymbol{\Phi}_g),$$

where $\boldsymbol{\vartheta} = (\boldsymbol{\pi}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G, \boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_G)$ denotes the model parameters. More specifically, the density of a linear Gaussian cluster-weighted model (CWM) is

$$p(\mathbf{x}, y \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g \phi_1\big(y \mid \beta_{0g} + \boldsymbol{\beta}_{1g}'\mathbf{x}, \sigma_g^2\big) \phi_p\left(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\right), \tag{18}$$

where $\beta_{0g} \in \mathbb{R}$, $\boldsymbol{\beta}_{1g} \in \mathbb{R}^p$, and $\phi_j()$ is the density of a $j$-dimensional random variable from a Gaussian distribution. The linear Gaussian CWM in (18) has been studied by Gershenfeld (1997) and Schöner (2000). CWMs are burgeoning into a vibrant subfield of model-based clustering and classification. For example, Ingrassia, Minotti, and Vittadini (2012) consider an extension to $t$-distribution that leads to the linear $t$-CWM. Ingrassia, Minotti, and Punzo (2014) introduce a family of 12 parsimonious linear $t$-CWMs, Punzo (2014) introduces the polynomial Gaussian CWM, Punzo and Ingrassia (2015a) propose CWMs for bivariate data of mixed type, and Punzo and Ingrassia (2015b) propose a family of 14 parsimonious linear Gaussian CWMs. Punzo and McNicholas (2014a) use a contamination approach for linear Gaussian CWMs. Ingrassia et al. (2015) consider CWMs with categorical responses and also consider identifiability under the assumption of Gaussian covariates.

In the mosaic of work around the use of mixture models for clustering and classification, CWMs have their place in applications with random covariates. Indeed, as distinct from finite mixtures of regressions (e.g., Leisch 2004; Früwirth-Schnatter 2006), which are examples of mixture models with fixed covariates, CWMs allow for assignment dependence, i.e., the covariate in each component can also be distinct. From a clustering and classification perspective, this implies that the covariates $\mathbf{X}$ can directly affect the clustering results—for most applications, this represents an advantage over the fixed covariates approach (Hennig 2000). A comparison of the fixed and random covariate approaches is given by Ingrassia and Punzo (2015).

Applying model (18) in high dimensions is infeasible for the same reason that using the Gaussian mixture model with unconstrained $\mathbf{\Sigma}_g$ in high dimensions is infeasible, i.e., the number of free covariance parameters is $\mathcal{O}(p^2)$. To overcome this issue, a latent Gaussian factor structure for $\mathbf{X}$ could be assumed within each mixture component—this is closely related to the factor regression model (FRM) of $Y$ on $\mathbf{X}$ (cf. West 2003; Wang et al. 2007; Carvalho 2008). Subedi et al. (2013) introduce the linear Gaussian cluster-weighted factor analyzers (CWFA) model, which has density

$$p(\mathbf{x}, y \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g \phi_1 \left( y \mid \beta_{0g} + \boldsymbol{\beta}'_{1g}\mathbf{x}, \sigma_g^2 \right) \phi_p \left( \mathbf{x} \mid \boldsymbol{\mu}_g, \mathbf{\Lambda}_g \mathbf{\Lambda}'_g + \mathbf{\Psi}_g \right).$$

where $\mathbf{\Lambda}_g$ is a $p \times q$ matrix of factor loadings, with $q < p$, and $\mathbf{\Psi}_g$ is a $p \times p$ diagonal matrix with strictly positive diagonal entries. A family of 16 CWFA models follows by applying the PGMM covariance constraints in Table 2 as well as allowing the constraint $\sigma_g^2 = \sigma^2$. As was the case for members of the PGMM family, each CWFA model has a number of free covariance parameters that is linear in $p$. Note that Subedi et al. (2015) extend the CWFA model to $t$-mixtures.

## 10. Discussion

Debate around how to define a cluster is sure to continue into the future. In addition to the discussion in Section 1 and in McNicholas (2016), and within the papers cited therein, there has been other relevant work. The paper by Hennig (2015) covers some of this other work and also raises interesting points about "true clusters". As a field of endeavour, mixture model-based approaches to clustering, classification, and discriminant analysis have made tremendous strides forward in the past decade or so. However, some important challenges and open questions remain. For one, the matter of model selection is still not resolved to a satisfactory extent. Although it is much maligned, the BIC remains the model selection criterion of choice. This point is reinforced by many applications within the literature as well as some dedicated comparative studies (e.g. Steele and Raftery 2010). The model averaging approach used by Wei and McNicholas (2015) provides an alternative to the 'single best model' paradigm; however, it too depends on the BIC. Furthermore, as the field moves away from Gaussian mixture models, questions around the efficacy of the BIC for mixture model selection will only grow in their frequency and intensity—although there is theoretical justification for using the BIC to compare non-nested models (cf. Raftery 1995), more work is needed to determine its efficacy for choosing between different mixture distributions, e.g., between a mixture of multivariate $t$-distributions and a mixture of MPEs. The search for a more effective criterion, perhaps one dedicated to clustering and classification, is perhaps the single greatest challenge within the field. As is the case for model selection, the choice of starting values is not a new problem (several strategies are discussed by Biernacki, Celeux and Govaert 2003; Shireman, Steinley and Brusco 2015, among others) but it is persistent, and efforts in this direction are sure to continue. It is quite likely that the increasing ease of access to high-performance computing equipment will help dictate the direction this work takes. The increasing dimensionality and complexity of modern data sets raise issues that demand answers. For instance, there has been a paucity of work on clustering mixed type data, ordinal data, and binary data (cf. Section 8). Another example is clickstream, and similar, data for which there has also been relatively little work (e.g., Melnykov 2016).

There has also been some interesting work on alternatives to the EM algorithm, and its variants, for parameter estimation. Of the alternatives tried to date, variational Bayes approximations, which are iterative Bayesian alternatives to the EM algorithm, perhaps hold the most promise. Their fast and deterministic nature has made the variational Bayes approach increasingly popular over the past decade or two (e.g., Waterhouse, MacKay and Robinson 1996; Jordan et al. 1999; Corduneanu and Bishop 2001). The tractability of the variational approach allows for simultaneous model selection and parameter estimation, thus removing the need for a criterion to select the number of components $G$ and potentially reducing the associated computational overhead. The variational Bayes algorithm has already been applied to Gaussian mixture models (e.g., Teschendorff et al. 2005; McGrory and Titterington 2007; Subedi and McNicholas 2016) as well as non-Gaussian mixtures (e.g., Subedi and McNicholas 2014). Interestingly, Bensmail et al. (1997) discuss exact Bayesian inference for some members of the GPCM family. Although beyond the scope of this review, there has been much work on Dirichlet process mixtures (e.g., Jain and Neal 2004; Bdiri, Bougouli, and Ziou 2016) and this is sure to continue.

The pursuit of more flexible models will continue and has the potential to provide more useful tools; however, it is very important that such methods are accompanied by effective software. This reflects a general problem: there are far more promising methods for model-based clustering than there are effective software packages. Beyond what is mentioned in Section 7, only minimal work has been done on model-based approaches to longitudinal data (e.g., De la Cruz-Mesá, Quintana, and Marshall 2008, use a mixture of non-linear hierarchical models) and this area also merits further investigation. Some recent work on fractionally-supervised classification (Vrbik and McNicholas 2015) is sure to spawn further work in similar directions. The use of copulas in mixture model-based approaches has already received some attention (e.g., Jajuga and Papla 2006; Di Lascio and Giannerini 2012; Vrac et al. 2012; Kosmidis and Karlis 2015; Marbac, Biernacki and Vandewalle 2015) and this sure to continue. Finally, there are some specific data types—both recently emerged and yet to emerge—that deserve their own special attention. One such type is next-generation sequencing data, which have already driven some interesting work within the field (e.g. Rau et al. 2015) and will surely continue to do so for some time.

## References

AITKEN, A.C. (1926), "A Series Formula for the Roots of Algebraic and Transcendental Equations", *Proceedings of the Royal Society of Edinburgh, 45*, 14–22.

AITKIN, M., and WILSON, G.T. (1980), "Mixture Models, Outliers, and the EM Algorithm", *Technometrics, 22(3)*, 325–331.

ANDERLUCCI, L., and VIROLI, C. (2015), "Covariance Pattern Mixture Models for Multivariate Longitudinal Data", *The Annals of Applied Statistics, 9(2)*, 777–800.

ANDREWS, J.L., and MCNICHOLAS, P.D. (2011a), "Extending Mixtures of Multivariate t-Factor Analyzers", *Statistics and Computing, 21(3)*, 361–373.

ANDREWS, J.L., and MCNICHOLAS, P.D. (2011b), "Mixtures of Modified t-Factor Analyzers for Model-Based Clustering, Classification, and Discriminant Analysis", *Journal of Statistical Planning and Inference, 141(4)*, 1479–1486.

ANDREWS, J.L., and MCNICHOLAS, P.D. (2012), "Model-Based Clustering, Classification, and Discriminant Analysis Via Mixtures of Multivariate $t$-Distributions: The $t$EIGEN Family", *Statistics and Computing, 22(5)*, 1021–1029.

ANDREWS, J.L., and MCNICHOLAS, P.D. (2013), *vscc: Variable Selection for Clustering and Classification*, R Package Version 0.2.

ANDREWS, J.L., and MCNICHOLAS, P.D. (2014), "Variable Selection for Clustering and Classification", *Journal of Classification, 31(2)*, 136–153.

ANDREWS, J.L., MCNICHOLAS, P.D., and SUBEDI, S. (2011), "Model-Based Classification Via Mixtures of Multivariate t-Distributions", *Computational Statistics and Data Analysis, 55(1)*, 520–529.

ANDREWS, J.L., WICKINS, J.R., BOERS, N.M., and MCNICHOLAS, P.D. (2015), *teigen: Model-Based Clustering and Classification with the Multivariate t Distribution*, R Package Version 2.1.0.

ATTIAS, H. (2000), "A Variational Bayesian Framework for Graphical Models", in *Advances in Neural Information Processing Systems*, Volume 12, MIT Press, pp. 209–215.

AZZALINI, A., BROWNE, R.P., GENTON, M.G., and MCNICHOLAS, P.D. (2016), "On Nomenclature for, and the Relative Merits of, Two Formulations of Skew Distributions", *Statistics and Probability Letters, 110*, 201–206.

AZZALINI, A., and CAPITANIO, A. (1999), "Statistical Applications of the Multivariate Skew Normal Distribution", *Journal of the Royal Statistical Society: Series B, 61(3)*, 579–602.

AZZALINI, A., and CAPITANIO, A. (2003), "Distributions Generated by Perturbation of Symmetry with Emphasis on a Multivariate Skew $t$ Distribution", *Journal of the Royal Statistical Society: Series B, 65(2)*, 367–389.

AZZALINI, A. (2014), *The Skew-Normal and Related Families*, with the collaboration of A. Capitanio, IMS monographs, Cambridge: Cambridge University Press.

AZZALINI, A., and VALLE, A.D. (1996), "The Multivariate Skew-Normal Distribution", *Biometrika / 83*, 715–726.

BAEK, J., and MCLACHLAN, G.J. (2008), "Mixtures of Factor Analyzers with Common Factor Loadings for the Clustering and Visualisation of High-Dimensional Data", Technical Report NI08018-SCH, Preprint Series of the Isaac Newton Institute for Mathematical Sciences, Cambridge.

BAEK, J., and MCLACHLAN, G.J. (2011), "Mixtures of Common t-Factor Analyzers for Clustering High-Dimensional Microarray Data", *Bioinformatics, 27*, 1269–1276.

BAEK, J., MCLACHLAN, G.J., and FLACK, L.K. (2010), "Mixtures of Factor Analyzers with Common Factor Loadings: Applications to the Clustering and Visualization of High-Dimensional Data", *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32*, 1298–1309.

BANFIELD, J.D., and RAFTERY, A.E. (1993), "Model-Based Gaussian and Non-Gaussian Clustering", *Biometrics, 49(3)*, 803–821.

BARNDORFF-NIELSEN, O.E. (1997), "Normal Inverse Gaussian Distributions and Stochastic Volatility Modelling", *Scandinavian Journal of Statistics, 24(1)*, 1–13.

BARTLETT, M.S. (1953), "Factor Analysis in Psychology as a Statistician Sees It", in *Uppsala Symposium on Psychological Factor Analysis*, Number 3 in Nordisk Psykologi's Monograph Series, Copenhagen: Ejnar Mundsgaards, pp. 23–34.

BAUDRY, J.-P. (2015), "Estimation and Model Selection for Model-Based Clustering with the Conditional Classification Likelihood", *Electronic Journal of Statistics, 9*, 1041–1077.

BAUM, L.E., PETRIE, T., SOULES, G., and WEISS, N. (1970), "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains", *Annals of Mathematical Statistics, 41*, 164–171.

BDIRI, T., BOUGUILA, N., and ZIOU, D. (2016), "Variational Bayesian Inference for Infinite Generalized Inverted Dirichlet Mixtures with Feature Selection and Its Application to Clustering", *Applied Intelligence, 44(3)*, 507–525.

BENSMAIL, H., CELEUX, G., RAFTERY, A.E., and ROBERT, C.P. (1997), "Inference in Model-Based Cluster Analysis", *Statistics and Computing, 7(1)*, 1–10.

BHATTACHARYA, S., and MCNICHOLAS, P.D. (2014), "A LASSO-Penalized BIC for Mixture Model Selection", *Advances in Data Analysis and Classification, 8(1)*, 45–61.

BIERNACKI, C., CELEUX, G., and GOVAERT, G. (2000), "Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood", *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(7)*, 719–725.

BIERNACKI, C., CELEUX, G., and GOVAERT, G. (2003), “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.

BIERNACKI, C., CELEUX, G., and GOVAERT, G. (2010), “Exact and Monte Carlo Calculations of Integrated Likelihoods for the Latent Class Model”, *Journal of Statistical Planning and Inference, 140(11)*, 2991–3002.

BIERNACKI, C., CELEUX, G., GOVAERT, G., and LANGROGNET, F. (2006), “Model-Based Cluster and Discriminant Analysis with the MIXMOD Software”, *Computational Statistics and Data Analysis, 51(2)*, 587–600.

BOUVEYRON, C., and BRUNET-SAUMARD, C. (2014), “Model-Based Clustering of High-Dimensional Data: A Review”, *Computational Statistics and Data Analysis, 71*, 52–78.

BOUVEYRON, C., CELEUX, G., and Girard, S. (2011), “Intrinsic Dimension Estimation by Maximum Likelihood in Isotropic Probabilistic PCA”, *Pattern Recognition Letters, 32(14)*, 1706–1713.

BOUVEYRON, C., GIRARD, S., and SCHMID, C. (2007a), “High-Dimensional Data Clustering”, *Computational Statistics and Data Analysis, 52(1)*, 502–519.

BOUVEYRON, C., GIRARD, S., and SCHMID, C. (2007b), “High Dimensional Discriminant Analysis”, *Communications in Statistics – Theory and Methods, 36(14)*, 2607–2623.

BRANCO, M.D., and DEY, D.K. (2001), “A General Class of Multivariate Skew-Elliptical Distributions”, *Journal of Multivariate Analysis, 79*, 99–113.

BROWNE, R.P., and MCNICHOLAS, P.D. (2012), “Model-Based Clustering and Classification of Data with Mixed Type”, *Journal of Statistical Planning and Inference, 142(11)*, 2976–2984.

BROWNE, R.P., and MCNICHOLAS, P.D. (2014a), “Estimating Common Principal Components in High Dimensions”, *Advances in Data Analysis and Classification, 8(2)*, 217–226.

BROWNE, R.P., and MCNICHOLAS, P.D. (2014b), *mixture: Mixture Models for Clustering and Classification*, R Package Version 1.1.

BROWNE, R.P., and P. D. MCNICHOLAS, P.D. (2014c), “Orthogonal Stiefel Manifold Optimization for Eigen-Decomposed Covariance Parameter Estimation in Mixture Models”, *Statistics and Computing, 24(2)*, 203–210.

BROWNE, R.P., and MCNICHOLAS, P.D. (2015), “A Mixture of Generalized Hyperbolic Distributions”, *Canadian Journal of Statistics, 43(2)*, 176–198.

BROWNE, R.P., MCNICHOLAS, P.D., and SPARLING, M.D. (2012), “Model-Based Learning Using a Mixture of Mixtures of Gaussian and Uniform Distributions”, *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(4)*, 814–817.

CAGNONE, S., and VIROLI, C. (2012), “A Factor Mixture Analysis Model for Multivariate Binary Data”, *Statistical Modelling, 12(3)*, 257–277.

CAMPBELL, N.A. (1984), “Mixture Models and Atypical Values”, *Mathematical Geology, 16(5)*, 465–477.

CARVALHO, C., CHANG, J., LUCAS, J., NEVINS, J., WANG, Q., and WEST, M. (2008), “High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics”, *Journal of the American Statistical Association, 103(484)*, 1438–1456.

CATELL, R.B. (1949), “‘R’ and Other Coefficients of Pattern Similarity”, *Psychometrika, 14*, 279–298.

CELEUX, G., and GOVAERT, G. (1991), "Clustering Criteria for Discrete Data and Latent Class Models", *Journal of Classification, 8(2)*, 157–176.

CELEUX, G., and GOVAERT, G. (1995), "Gaussian Parsimonious Clustering Models", *Pattern Recognition, 28(5)*, 781–793.

CORDUNEANU, A., and BISHOP, C.M. (2001), "Variational Bayesian Model Selection for Mixture Distributions", in *Artificial Intelligence and Statistics*, Los Altos, CA: Morgan Kaufmann, pp. 27–34.

CORETTO, P., and HENNIG, C. (2015), "Robust Improper Maximum Likelihood: Tuning, Computation, and a Comparison with Other Methods for Robust Gaussian Clustering", arXiv preprint arXiv:1405.1299v3.

CORMACK, R.M. (1971), "A Review of Classification (With Discussion)", *Journal of the Royal Statistical Society: Series A, 34*, 321–367.

DANG, U.J., BROWNE, R.P., and MCNICHOLAS, P.D. (2015), "Mixtures of Multivariate Power Exponential Distributions", *Biometrics, 71(4)*, 1081–1089.

DASGUPTA, A., and RAFTERY, A.E. (1998), "Detecting Features in Spatial Point Processes with Clutter Via Model-Based Clustering", *Journal of the American Statistical Association, 93*, 294–302.

DAY, N.E. (1969), "Estimating the Components of a Mixture of Normal Distributions", *Biometrika, 56*, 463–474.

DE LA CRUZ-MESÍA, R., QUINTANA, R.A., and MARSHALL, G. (2008), "Model-Based Clustering for Longitudinal data", *Computational Statistics and Data Analysis, 52(3)*, 1441–1457.

DE VEAUX, R.D., and KRIEGER, A.M. (1990), "Robust Estimation of a Normal Mixture", *Statistics and Probability Letters, 10(1)*, 1–7.

DEAN, N., RAFTERY, A.E., and SCRUCCA, L. (2012), *clustvarsel: Variable Selection for Model-Based Clustering*, R package version 2.0.

DEMPSTER, A.P., LAIRD, N.M., and RUBIN, D.B. (1977), "Maximum Likelihood from Incomplete Data Via the EM Algorithm", *Journal of the Royal Statistical Society: Series B, 39(1)*, 1–38.

DI LASCIO, F.M.L., and GIANNERINI, S. (2012), "A Copula-Based Algorithm for Discovering Patterns of Dependent Observations", *Journal of Classification, 29(1)*, 50–75.

EDWARDS, A.W.F., and CAVALLI-SFORZA, L.L. (1965), "A Method for Cluster Analysis", *Biometrics, 21*, 362–375.

EVERITT, B.S., and HAND, D.J. (1981), *Finite Mixture Distributions*, Monographs on Applied Probability and Statistics, London: Chapman and Hall.

EVERITT, B.S., LANDAU, S., LEESE, M., and STAHL, D. (2011), *Cluster Analysis* (5th ed.), Chichester: John Wiley & Sons.

FABRIGAR, L.R., WEGENER, D.T., MACCALLUM, R.C., and STRAHAN, E.J. (1999), "Evaluating the Use of Exploratory Factor Analysis in Psychological Research", *Psychological Methods, 4(3)*, 272–299.

FLURY, B. (1988), *Common Principal Components and Related Multivariate Models*, New York: Wiley.

FRALEY, C., and RAFTERY, A.E. (1998), "How Many Clusters? Which Clustering Methods? Answers Via Model-Based Cluster Analysis", *The Computer Journal, 41(8)*, 578–588.

FRALEY, C., and RAFTERY, A.E. (1999), "MCLUST: Software for Model-Based Cluster Analysis", *Journal of Classification, 16*, 297–306.

FRALEY, C.,and RAFTERY,A.E. (2002a), "MCLUST: Software for Model-Based Clustering, Density Estimation, and Discriminant Analysis", Technical Report 415, University of Washington, Department of Statistics.

FRALEY, C., and RAFTERY, A.E. (2002b), "Model-Based Clustering, Discriminant Analysis, and Density Estimation", *Journal of the American Statistical Association, 97(458)*, 611–631.

FRANCZAK, B.C., BROWNE, R.P., and MCNICHOLAS, P.D. (2014), "Mixtures of Shifted Asymmetric Laplace Distributions", *IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(6)*, 1149–1157.

FRIEDMAN, H.P., and RUBIN, J. (1967), "On Some Invariant Criteria for Grouping Data", *Journal of the American Statistical Association, 62*, 1159–1178.

FRITZ, H., GARCÍA-ESCUDERO, L.A., and MAYO-ISCAR, A. (2012), "tclust: An R Package for a Trimming Approach to Cluster Analysis", *Journal of Statistical Software, 47(12)*, 1–26.

FRÜHWIRTH-SCHNATTER, S. (2006), *Finite Mixture and Markov Switching Models*, New York: Springer-Verlag.

GALIMBERTI, G., MONTANARI, A., and VIROLI, C. (2009), "Penalized Factor Mixture Analysis for Variable Selection in Clustered Data", *Computational Statistics and Data Analysis, 53*, 4301–4310.

GARCÍA-ESCUDERO, L.A., GORDALIZA, A., MATRN, C., and MAYO-ISCAR, A. (2008), "A General Trimming Approach to Robust Cluster Analysis", *The Annals of Statistics, 36(3)*, 1324–1345.

GERSHENFELD, N. (1997), "Nonlinear Inference and Cluster-Weighted Modeling", *Annals of the New York Academy of Sciences, 808(1)*, 18–24.

GHAHRAMANI, Z., and HINTON, G.E. (1997), "The EM Algorithm for Factor Analyzers", Technical Report CRG-TR-96-1, University of Toronto, Toronto, Canada.

GOLLINI, I., and MURPHY, T.B. (2014), "Mixture of Latent Trait Analyzers for Model-Based Clustering of Categorical Data", *Statistics and Computing, 24(4)*, 569–588.

GÓMEZ, E., GÓMEZ-VIILEGAS, M.A., and MARIN, J.M. (1998), "A Multivariate Generalization of the Power Exponential Family of Distributions", *Communications in Statistics – Theory and Methods, 27(3)*, 589–600.

GÓMEZ-SÁNCHEZ-MANZANO, E., GÓMEZ-VILLEGAS, M.A., and Marín, J.M. (2008), "Multivariate Exponential Power Distributions as Mixtures of Normal Distributions with Bayesian Applications", *Communications in Statistics – Theory and Methods, 37(6)*, 972–985.

GOODMAN, L. (1974), "Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models", *Biometrika, 61(2)*, 215–231.

GORDON, A.D. (1981), *Classification*, London: Chapman and Hall.

GRESELIN, F., and INGRASSIA, S. (2010), "Constrained Monotone EM Algorithms for Mixtures of Multivariate t-Distributions", *Statistics and Computing, 20(1)*, 9–22.

HATHAWAY, R.J. (1985), "A Constrained Formulation of Maximum Likelihood Estimation for Normal Mixture Distributions", *The Annals of Statistics, 13(2)*, 795–800.

HEISER, W.J. (1995), "Recent Advances in Descriptive Multivariate Analysis", in *Convergent Computation by Iterative Majorization: Theory and Applications in Multidimensional Data Analysis*, ed. W.J. Krzanowski, Oxford: Oxford University Press, pp. 157–189.

HENNIG, C. (2000), "Identifiablity of Models for Clusterwise Linear Regression", *Journal of Classification, 17(2)*, 273–296.

HENNIG, C. (2004), "Breakdown Points for Maximum Likelihood Estimators of Location-Scale Mixtures", *The Annals of Statistics, 32(4)*, 1313–1340.

HENNIG, C. (2015), "What are the True Clusters?", *Pattern Recognition Letters, 64*, 53–62.

HORN, J.L. (1965), "A Rationale and Technique for Estimating the Number of Factors in Factor Analysis", *Psychometrika, 30*, 179–185.

HU, W. (2005), *Calibration of Multivariate Generalized Hyperbolic Distributions Using the EM Algorithm, with Applications in Risk Management, Portfolio Optimization and Portfolio Credit Risk*, Ph. D. thesis, The Florida State University, Tallahassee.

HUBER, P.J. (1964), "Robust Estimation of a Location Parameter", *The Annals of Mathematical Statistics, 35*, 73–101.

HUBER, P.J. (1981), *Robust Statistics*, New York: Wiley.

HUMBERT, S., SUBEDI, S., COHN, J., ZENG, B., BI, Y.-M., CHEN, X., ZHU, T., MCNICHOLAS, P.D., and ROTHSTEIN, S.J. (2013), "Genome-Wide Expression Profiling of Maize in Response to Individual and Combined Water and Nitrogen Stresses", *BMC Genetics, 14(3)*.

HUMPHREYS, L.G., and ILGEN, D.R. (1969), "Note on a Criterion for the Number of Common Factors", *Educational and Psychological Measurements, 29*, 571–578.

HUMPHREYS, L.G., and MONTANELLI, R.G. JR. (1975), "An Investigation of the Parallel Analysis Criterion for Determining the Number of Common Factors", *Multivariate Behavioral Research, 10*, 193–205.

INGRASSIA, S., MINOTTI, S.C., and PUNZO, A. (2014), "Model-Based Clustering Via Linear Cluster-Weighted Models", *Computational Statistics and Data Analysis, 71*, 159–182.

INGRASSIA, S., MINOTTI, S.C., PUNZO, A., and VITTADINI, G. (2015), "The Generalized Linear Mixed Cluster-Weighted Model", *Journal of Classification, 32(1)*, 85–113.

INGRASSIA, S., MINOTTI, S.C., and VITTADINI, G. (2012), "Local Statistical Modeling Via the Cluster-Weighted Approach with Elliptical Distributions", *Journal of Classification, 29(3)*, 363–401.

INGRASSIA, S., and PUNZO, A. (2015), "Decision Boundaries for Mixtures of Regressions", *Journal of the Korean Statistical Society, 44(2)*, 295–306.

JAAKKOLA, T.S., and JORDAN, M.I. (2000), "Bayesian Parameter Estimation Via Variational Methods", *Statistics and Computing, 10(1)*, 25–37.

JAIN, S., and NEAL, R.M. (2004), "A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model", *Journal of Computational and Graphical Statistics, 13(1)*, 158–182.

JAJUGA, K., and PAPLA, D. (2006), "Copula Functions in Model Based Clustering", in *From Data and Information Analysis to Knowledge Engineering*, Studies in Classification, Data Analysis, and Knowledge Organization, eds. M. Spiliopoulou, R. Kruse, C. Borgelt, A.Nürnberger, and W. Gaul, Berlin, Heidelberg: Springer, pp. 603–613.

JORDAN, M.I., ZGHAHRAMANI, Z., JAAKKOLA, T.S., and SAUL, L.K. (1999), "An Introduction to Variational Methods for Graphical Models", *Machine Learning, 37*, 183–233.

JÖRESKOG, K.G. (1990), "New Developments in LISREL: Analysis of Ordinal Variables Using Polychoric Correlations and Weighted Least Squares", *Quality and Quantity, 24(4)*, 387–404.

KARLIS, D., and SANTOURIAN, A. (2009), "Model-Based Clustering with Non-Elliptically Contoured Distributions", *Statistics and Computing, 19(1)*, 73–83.

KASS, R.E., and RAFTERY, A.E. (1995), "Bayes Factors", *Journal of the American Statistical Association, 90(430)*, 773–795.

KERIBIN, C. (2000), "Consistent Estimation of the Order of Mixture Models", *Sankhyā. The Indian Journal of Statistics. Series A, 62(1)*, 49–66.

KHARIN, Y. (1996), *Robustness in Statistical Pattern Recognition*, Dordrecht: Kluwer.

KOSMIDIS, I., and KARLIS, D. (2015), "Model-Based Clustering Using Copulas with Applications", arXiv preprint arXiv:1404.4077v5.

KOTZ, S., KOZUBOWSKI, T.J., and PODGORSKI, K. (2001), *The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance* (1st ed.), Boston: Burkhäuser.

LAWLEY, D.N., and MAXWELL, A.E. (1962), "Factor Analysis as a Statistical Method", *Journal of the Royal Statistical Society: Series D, 12(3)*, 209–229.

LEE, S., and MCLACHLAN, G.J. (2011), "On the Fitting of Mixtures of Multivariate Skew $t$-distributions Via the EM Algorithm", arXiv:1109.4706.

LEE, S., and MCLACHLAN, G.J.(2014), "Finite Mixtures of Multivariate Skew t-Distributions: Some Recent and New Results", *Statistics and Computing, 24*, 181–202.

LEE, S.X., and MCLACHLAN, G.J. (2013a), "Model-Based Clustering and Classification with Non-Normal Mixture Distributions", *Statistical Methods and Applications, 22(4)*, 427–454.

LEE, S.X., and MCLACHLAN, G.J. (2013b), "On Mixtures of Skew Normal and Skew t-Distributions", *Advances in Data Analysis and Classification, 7(3)*, 241–266.

LEISCH, F. (2004), "Flexmix: A General Framework For Finite Mixture Models And Latent Class Regression in R", *Journal of Statistical Software, 11(8)*, 1–18.

LEROUX, B.G. (1992), "Consistent Estimation of a Mixing Distribution", *The Annals of Statistics, 20(3)*, 1350–1360.

LI, J. (2005), "Clustering Based on a Multi-Layer Mixture Model", *Journal of Computational and Graphical Statistics, 14(3)*, 547–568.

LI, K.C. (1991), "Sliced Inverse Regression for Dimension Reduction (With Discussion)", *Journal of the American Statistical Association, 86*, 316–342.

LI, K.C. (2000), "High Dimensional Data Analysis Via the SIR/PHD Approach", Unpublished manuscript.

LIN, T.-I. (2009), "Maximum Likelihood Estimation for Multivariate Skew Normal Mixture Models", *Journal of Multivariate Analysis, 100*, 257–265.

LIN, T.-I. (2010), "Robust Mixture Modeling Using Multivariate Skew t Distributions", *Statistics and Computing, 20(3)*, 343–356.

LIN, T.-I., MCLACHLAN, G.J., and LEE, S.X. (2016), "Extending Mixtures of Factor Models Using the Restricted Multivariate Skew-Normal Distribution", *Journal of Multivariate Analysis, 143*, 398–413.

LIN, T.-I., MCNicholas, P.D., and HSIU, J.H. (2014), "Capturing Patterns Via Parsimonious t Mixture Models", *Statistics and Probability Letters, 88*, 80–87.

LOPES, H.F., and WEST, M. (2004), "Bayesian Model Assessment in Factor Analysis", *Statistica Sinica, 14*, 41–67.

MARBAC, M., BIERNACKI, C., and VANDEWALLE,V. (2014), "Finite Mixture Model of Conditional Dependencies Modes to Cluster Categorical Data", arXiv preprint arXiv:1402.5103.

MARBAC, M., BIERNACKI, C., and VANDEWALLE, V. (2015), "Model-Based Clustering of Gaussian Copulas for Mixed Data", arXiv preprint arXiv:1405.1299v3.

MARKATOU, M. (2000), "Mixture Models, Robustness, and the Weighted Likelihood Methodology", *Biometrics, 56(2)*, 483–486.

MAUGIS, C. (2009), "The Selvarclust Software", www.math.univ-toulouse.fr/~maugis/SelvarClustHomepage.html.

MAUGIS, C., CELEUX, G., and MARTIN-MAGNIETTE, M.-L. (2009a), "Variable Selection for Clustering with Gaussian Mixture Models", *Biometrics, 65(3)*, 701–709.

MAUGIS, C., CELEUX, G., and MARTIN-MAGNIETTE, M.-L. (2009b), "Variable Selection in Model-Based Clustering: A General Variable Role Modeling", *Computational Statistics and Data Analysis, 53(11)*, 3872–3882.

MCGRORY, C., and TITTERINGTON, D. (2007), "Variational Approximations in Bayesian Model Selection for Finite Mixture Distributions", *Computational Statistics and Data Analysis, 51(11)*, 5352–5367.

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

MCLACHLAN, G.J., BEAN, R.W., and JONES, L.B.-T. (2007), "Extension of the Mixture of Factor Analyzers Model to Incorporate the Multivariate t-Distribution", *Computational Statistics and Data Analysis, 51(11)*, 5327–5338.

MCLACHLAN, G.J., and KRISHNAN, T. (2008), *The EM Algorithm and Extensions* (2nd ed.), 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*, Volume 1451, Berlin: Springer-Verlag, pp. 658–666.

MCLACHLAN, G.J., and PEEL, D. (2000a), *Finite Mixture Models*, New York: John Wiley & Sons.

MCLACHLAN, G.J., and PEEL, D. (2000b), "Mixtures of Factor Analyzers", in *Proceedings of the Seventh International Conference on Machine Learning*, San Francisco, Morgan Kaufmann, pp. 599–606.

MCNEIL, A.J., FREY, R., and EMBRECHTS, P. (2005), *Quantitative Risk Management: Concepts, Techniques and Tools.*, Princeton: Princeton University Press.

MCNICHOLAS, P.D. (2013), "Model-Based Clustering and Classification Via Mixtures of Multivariate t-Distributions", in *Statistical Models for Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization, eds. P. Giudici, S. Ingrassia, and M. Vichi, Switzerland: Springer International Publishing.

MCNICHOLAS, P.D. (2016), *Mixture Model-Based Classification*, Boca Raton FL: Chapman & Hall/CRC Press.

MCNICHOLAS, P.D., and BROWNE, R.P. (2013), "Discussion of 'How to Find an Appropriate Clustering for Mixed-Type Variables with Application to Socio-Economic Stratification' by Hennig and Liao", *Journal of the Royal Statistical Society: Series C, 62(3)*, 352–353.

MCNICHOLAS, P.D., ELSHERBINY, A., MCDAID, A.F., and MURPHY, T.B. (2015), *pgmm: Parsimonious Gaussian Mixture Models*, R Package Version 1.2.

MCNICHOLAS, P.D., JAMPANI, K.R., and SUBEDI, S. (2015), *longclust: Model-Based Clustering and Classification for Longitudinal Data*, R Package Version 1.2.

MCNICHOLAS, P.D., and MURPHY, T.B. (2005), "Parsimonious Gaussian Mixture Models", Technical Report 05/11, Department of Statistics, Trinity College Dublin, Dublin, Ireland.

MCNICHOLAS, P.D., and MURPHY, T.B. (2008), “Parsimonious Gaussian Mixture Models”, *Statistics and Computing, 18(3)*, 285–296.

MCNICHOLAS, P.D., and MURPHY, T.B. (2010a), “Model-Based Clustering of Longitudinal Data”, *Canadian Journal of Statistics, 38(1)*, 153–168.

MCNICHOLAS, P.D., and MURPHY, T.B. (2010b), “Model-Based Clustering of Microarray Expression Data Via Latent Gaussian Mixture Models”, *Bioinformatics, 26(21)*, 2705–2712.

MCNICHOLAS, P.D., and SUBEDI, S. (2012), “Clustering Gene Expression Time Course Data Using Mixtures of Multivariate t-Distributions”, *Journal of Statistical Planning and Inference, 142(5)*, 1114–1127.

MCNICHOLAS, S.M., MCNICHOLAS, P.D., and BROWNE, R.P. (2014), “Mixtures of Variance-Gamma Distributions”, arxiv preprint arXiv:1309.2695v2.

MCPARLAND, D., GORMLEY, I.C., MCCORMICK, T.H., CLARK, S.J., KABUDULA, C.W., and COLLINSON, M.A. (2014), “Clustering South African Households Based on Their Asset Status Using Latent Variable Models”, *The Annals of Applied Statistics, 8(2)*, 747–776.

MCQUITTY, L.L. (1956), “Agreement Analysis: A Method of Classifying Subjects According to Their Patterns of Responses”, *British Journal of Statistical Psychology, 9*, 5–16.

MELNYKOV, V. (2016), “Model-Based Biclustering of Clickstream Data”, *Computational Statistics and Data Analysis, 93*, 31–45.

MENG, X.-L., and RUBIN, D.B. (1993), “Maximum Likelihood Estimation Via the ECM Algorithm: A General Framework”, *Biometrika, 80*, 267–278.

MENG, X.-L., and VAN DYK, D. (1997), “The EM Algorithm—An Old Folk Song Sung to a Fast New Tune (With Discussion)”, *Journal of the Royal Statistical Society: Series B, 59(3)*, 511–567.

MONTANARI, A., and VIROLI, C. (2010a), “Heteroscedastic Factor Mixture Analysis”, *Statistical Modelling, 10(4)*, 441–460.

MONTANARI, A., and VIROLI, C. (2010b), “A Skew-Normal Factor Model for the Analysis of Student Satisfaction Towards University Courses”, *Journal of Applied Statistics, 43*, 473–487.

MONTANARI, A., and VIROLI, C. (2011), “Maximum Likelihood Estimation of Mixture of Factor Analyzers”, *Computational Statistics and Data Analysis, 55*, 2712–2723.

MONTANELLI, R.G., JR., and HUMPHREYS, L.G. (1976), “Latent Roots of Random Data Correlation Matrices with Squared Multiple Correlations on the Diagonal: A Monte Carlo Study”, *Psychometrika, 41*, 341–348.

MORRIS, K., and MCNICHOLAS, P.D. (2013), “Dimension Reduction for Model-Based Clustering Via Mixtures of Shifted Asymmetric Laplace Distributions”, *Statistics and Probability Letters, 83(9)*, 2088–2093, Erratum 2014, *85*,168.

MORRIS, K., and MCNICHOLAS, P.D. (2016), “Clustering, Classification, Discriminant Analysis, and Dimension Reduction Via Generalized Hyperbolic Mixtures”, *Computational Statistics and Data Analysis, 97*, 133–150.

MORRIS, K., MCNICHOLAS, P.D., and SCRUCCA, L. (2013), “Dimension Reduction for Model-Based Clustering Via Mixtures of Multivariate t-Distributions”, *Advances in Data Analysis and Classification, 7(3)*, 321–338.

MURRAY, P.M., BROWNE, R.B., and MCNICHOLAS, P.D. (2014a), “Mixtures of Skew-t Factor Analyzers”, *Computational Statistics and Data Analysis, 77*, 326–335.

MURRAY, P.M., MCNICHOLAS, P.D., and BROWNE, R.B. (2014b), "A Mixture of Common Skew-$t$ Factor Analyzers", *Stat, 3(1)*, 68–82.

MUTHEN, B., and ASPAROUHOV, T. (2006), "Item Response Mixture Modeling: Application to Tobacco Dependence Criteria", *Addictive Behaviors, 31*, 1050–1066.

O'HAGAN, A., MURPHY, T.B., GORMLEY, I.C., MCNICHOLAS, P.D., and KARLIS, D. (2016), "Clustering with the Multivariate Normal Inverse Gaussian Distribution", *Computational Statistics and Data Analysis, 93*, 18–30.

ORCHARD, T., and WOODBURY, M.A. (1972), "A Missing Information Principle: Theory and Applications", in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Theory of Statistics*, eds. L.M. Le Cam, J. Neyman, and E.L. Scott, Berkeley: University of California Press, pp. 697–715.

PAN, J., and MACKENZIE, G. (2003), "On Modelling Mean-Covariance Structures in Longitudinal Studies", *Biometrika, 90(1)*, 239–244.

PEARSON, K. (1894), "Contributions to the Mathematical Theory of Evolution", *Philosophical Transactions of the Royal Society, Part A, 185*, 71–110.

PEEL, D., and MCLACHLAN, G.J. (2000), "Robust Mixture Modelling Using the t Distribution", *Statistics and Computing, 10(4)*, 339–348.

POURAHMADI, M. (1999), "Joint Mean-Covariance Models with Applications to Longitudinal Data: Unconstrained Parameterisation", *Biometrika, 86(3)*, 677–690.

POURAHMADI, M. (2000), "Maximum Likelihood Estimation of Generalised Linear Models for Multivariate Normal Covariance Matrix", *Biometrika, 87(2)*, 425–435.

POURAHMADI, M., DANIELS, M., and PARK, T. (2007), "Simultaneous Modelling of the Cholesky Decomposition of Several Covariance Matrices", *Journal of Multivariate Analysis, 98*, 568–587.

PUNZO, A. (2014), "Flexible Mixture Modeling with the Polynomial Gaussian Cluster-Weighted Model", *Statistical Modelling, 14(3)*, 257–291.

PUNZO, A., and INGRASSIA, S. (2015a), "Clustering Bivariate Mixed-Type Data Via the Cluster-Weighted Model", *Computational Statistics*. To appear.

PUNZO, A., and INGRASSIA, S. (2015b), "Parsimonious Generalized Linear Gaussian Cluster-Weighted Models", in , *Advances in Statistical Models for Data Analysis*, Studies in Classification, Data Analysis and Knowledge Organization, Switzerland, eds. I. Morlini, T. Minerva, and M. Vichi, Springer International Publishing, pp. 201–209.

PUNZO, A., and MCNICHOLAS, P.D. (2014a), "Robust Clustering in Regression Analysis Via the Contaminated Gaussian Cluster-Weighted Model", arXiv preprint arXiv:1409.6019v1.

PUNZO, A., and MCNICHOLAS, P.D. (2014b), "Robust High-Dimensional Modeling with the Contaminated Gaussian Distribution", arXiv preprint arXiv:1408.2128v1.

PUNZO, A., and MCNICHOLAS, P.D. (2016), "Parsimonious Mixtures of Multivariate Contaminated Normal Distributions", *Biometrical Journal*. To appear.

R CORE TEAM (2015), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing.

RAFTERY, A.E. (1995), "Bayesian Model Selection in Social Research (With Discussion)", *Sociological Methodology, 25*, 111–193.

RAFTERY, A.E., and DEAN, N. (2006), "Variable Selection for Model-Based Clustering", *Journal of the American Statistical Association, 101(473)*, 168–178.

RANALLI, M., and ROCCI, R. (2016),"Mixture Methods for Ordinal Data: A Pairwise Likelihood Approach", *Statistics and Computing, 26(1)*, 529–547.

RAO, C.R. (1952), *Advanced Statistical Methods in Biometric Research*, New York: John Wiley and Sons, Inc.

RAU, A., MAUGIS-RABUSSEAU, C., MARTIN-MAGNIETTE, M.-L, and CELEUX, G. (2015), "Co-expression Analysis of High-Throughput Transcriptome Sequencing Data with Poisson Mixture Models", *Bioinformatics, 31(9)*, 1420–1427.

SAHU, K., DEY, D.K., and BRANCO, M.D. (2003), "A New Class of Multivariate Skew Distributions with Applications to Bayesian Regression Models", *Canadian Journal of Statistics, 31(2)*, 129–150. Corrigendum: Vol. 37 (2009), 301–302.

SCHÖNER, B. (2000), *Probabilistic Characterization and Synthesis of Complex Data Driven Systems*, Ph. D. thesis, Cambridge MA: MIT.

SCHROETER, P., VESIN, J., LANGENBERGER, T., and MEULI, R. (1998), "Robust Parameter Estimation of Intensity Distributions for Brain Magnetic Resonance Images", *IEEE Transactions on Medical Imaging, 17(2)*, 172–186.

SCHWARZ, G. (1978), "Estimating the Dimension of a Model", *The Annals of Statistics, 6(2)*, 461–464.

SCOTT, A.J., and SYMONS, M.J. (1971), "Clustering Methods Based on Likelihood Ratio Criteria", *Biometrics, 27*, 387–397.

SCRUCCA, L. (2010), "Dimension Reduction for Model-Based Clustering", *Statistics and Computing, 20(4)*, 471–484.

SCRUCCA, L. (2014), "Graphical Tools for Model-Based Mixture Discriminant Analysis", *Advances in Data Analysis and Classification, 8(2)*, 147–165.

SHIREMAN, E., STEINLEY, D., and BRUSCO, M.J. (2015), "Examining the Effect of Initialization Strategies on the Performance of Gaussian Mixture Modeling", *Behavior Research Methods*.

SPEARMAN, C. (1904), "The Proof and Measurement of Association Between Two Things", *American Journal of Psychology, 15*, 72–101.

SPEARMAN, C. (1927), *The Abilities of Man: Their Nature and Measurement*, London: MacMillan and Co., Limited.

STEANE, M.A., MCNICHOLAS, P.D., and YADA,R. (2012), "Model-Based Classification Via Mixtures of Multivariate t-Factor Analyzers", *Communications in Statistics – Simulation and Computation, 41(4)*, 510–523.

STEELE, R.J., and RAFTERY, A.E. (2010), "Performance of Bayesian Model Selection Criteria for Gaussian Mixture Models", in *Frontiers of Statistical Decision Making and Bayesian Analysis*, Vol, 2, New York: Springer, pp. 113–130.

STEPHENSEN, W. (1953), *The Study of Behavior*, Chicago: University of Chicago Press.

SUBEDI, S., and MCNICHOLAS, P.D. (2014), "Variational Bayes Approximations for Clustering Via Mixtures of Normal Inverse Gaussian Distributions", *Advances in Data Analysis and Classification, 8(2)*, 167–193.

SUBEDI, S., and MCNICHOLAS, P.D. (2016), "A Variational Approximations-DIC Rubric for Parameter Estimation and Mixture Model Selection Within a Family Setting", arXiv preprint arXiv:1306.5368v2.

SUBEDI, S., PUNZO, A., INGRASSIA, S., and MCNICHOLAS, P.D. (2013), "Clustering and Classification Via Cluster-Weighted Factor Analyzers", *Advances in Data Analysis and Classification, 7(1)*, 5–40.

SUBEDI, S., PUNZO, A., INGRASSIA, S., and MCNICHOLAS, P.D. (2015), "Cluster-Weighted t-Factor Analyzers for Robust Model-Based Clustering and Dimension Reduction", *Statistical Methods and Applications, 24(4)*, 623–649.

SUNDBERG, R. (1974), "Maximum Likelihood Theory for Incomplete Data from an Exponential Family", *Scandinavian Journal of Statistics, 1(2)*, 49–58.

TANG, Y., BROWNE, R.P., and MCNICHOLAS, P.D. (2015), "Model-Based Clustering of High-Dimensional Binary Data", *Computational Statistics and Data Analysis, 87*, 84–101.

TESCHENDORFF, A., WANG, Y., BARBOSA-MORAIS, J., BRENTON, N., and CALDAS, C. (2005), "A Variational Bayesian Mixture Modelling Framework for Cluster Analysis of Gene-Expression Data", *Bioinformatics, 21(13)*, 3025–3033.

TIEDEMAN, D.V. (1955), "On the Study of Types", in *Symposium on Pattern Analysis*, ed. S.B. Sells, Randolph Field, Texas: Air University, U.S.A.F. School of Aviation Medicine, pp. 1–14.

TIPPING, M.E. (1999), "Probabilistic Visualization of High-Dimensional Binary Data", *Advances in Neural Information Processing Systems (11)*, 592–598.

TIPPING, M.E., and BISHOP, C.M. (1997), "Mixtures of Probabilistic Principal Component Analysers", Technical Report NCRG/97/003, Aston University (Neural Computing Research Group), Birmingham, UK.

TIPPING, M.E., and BISHOP,C.M. (1999), "Mixtures of Probabilistic Principal Component Analysers", *Neural Computation, 11(2)*, 443–482.

TITTERINGTON, D.M., SMITH, A.F.M, and MAKOV, U.E. (1985), *Statistical Analysis of Finite Mixture Distributions*, Chichester: John Wiley & Sons.

TORTORA, C., MCNICHOLAS, P.D., and BROWNE, R.P. (2015), "A Mixture of Generalized Hyperbolic Factor Analyzers", *Advances in Data Analysis and Classification*. To appear.

TRYON, R.C. (1939), *Cluster Analysis*, Ann Arbor: Edwards Brothers.

TRYON, R.C. (1955), "Identification of Social Areas by Cluster Analysis", in *University of California Publications in Psychology*, Volume 8, Berkeley: University of California Press.

VERMUNT, J.K. (2003), "Multilevel Latent Class Models", *Sociological Methodology, 33(1)*, 213–239.

VERMUNT, J.K. (2007), "Multilevel Mixture Item Response Theory Models: An Application in Education Testing", in *Proceedings of the 56th Session of the International Statistical Institute*, Lisbon, Portugal, pp. 22–28.

VIROLI, C. (2010), "Dimensionally Reduced Model-Based Clustering Through Mixtures of Factor Mixture Analyzers", *Journal of Classification, 27(3)*, 363–388.

VRAC, M., BILLARD, L., DIDAY, E., and CHEDIN, A. (2012), "Copula Analysis of Mixture Models", *Computational Statistics, 27(3)*, 427–457.

VRBIK, I., and MCNICHOLAS, P.D. (2012), "Analytic Calculations for the EM Algorithm for Multivariate Skew-t Mixture Models", *Statistics and Probability Letters, 82(6)*, 1169–1174.

VRBIK, I., and MCNICHOLAS, P.D. (2014), "Parsimonious Skew Mixture Models for Model-Based Clustering and Classification", *Computational Statistics and Data Analysis, 71*, 196–210.

VRBIK, I., and MCNICHOLAS, P.D. (2015), "Fractionally-Supervised Classification", *Journal of Classification, 32(3)*, 359–381.

WANG, Q., CARVALHO, C., LUCAS, J., and WEST, M. (2007), "BFRM: Bayesian Factor Regression Modelling", *Bulletin of the International Society for Bayesian Analysis, 14(2)*, 4–5.

WATERHOUSE, S., MACKAY, D., and ROBINSON, T. (1996), "Bayesian Methods for Mixture of Experts", in *Advances in Neural Information Processing Systems*, Vol. 8. Cambridge, MA: MIT Press.

WEI, Y., and MCNICHOLAS, P.D. (2015), "Mixture Model Averaging for Clustering", *Advances in Data Analysis and Classification, 9(2)*, 197–217.

WEST, M. (2003), "Bayesian Factor Regression Models in the 'Large $p$, Small $n$' Paradigm", in *Bayesian Statistics*, Volume 7, eds. J.M. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West, Oxford: Oxford University Press, pp. 723–732.

WOLFE, J.H. (1963), "Object Cluster Analysis of Social Areas", Master's thesis, University of California, Berkeley.

WOLFE, J.H. (1965), "A Computer Program for the Maximum Likelihood Analysis of Types", Technical Bulletin 65-15, U.S. Naval Personnel Research Activity.

WOLFE, J.H. (1970), "Pattern Clustering by Multivariate Mixture Analysis", *Multivariate Behavioral Research, 5*, 329–350.

YOSHIDA, R., HIGUCHI, T., and IMOTO, S. (2004), "A Mixed Factors Model for Dimension Reduction and Extraction of a Group Structure in Gene Expression Data", in *Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference*, pp. 161–172.

YOSHIDA, R., HIGUCHI, T., IMOTO, S., and MIYANO, S. (2006), "ArrayCluster: An Analytic Tool for Clustering, Data Visualization and Module Finder on Gene Expression Profiles", *Bioinformatics, 22*, 1538–1539.

ZHOU, H., and LANGE, K.L. (2010), "On the Bumpy Road to the Dominant Mode", *Scandinavian Journal of Statistics, 37(4)*, 612–631.