JIRSS (2011)
Vol. 10, No. 2, pp 181-199

On Model-Based Clustering, Classification, and Discriminant Analysis

Paul D. McNicholas

Department of Mathematics and Statistics, University of Guelph, Ontario, Canada.

Abstract. The use of mixture models for clustering and classification has burgeoned into an important subfield of multivariate analysis. These approaches have been around for a half-century or so, with significant activity in the area over the past decade. The primary focus of this paper is to review work in model-based clustering, classification, and discriminant analysis, with particular attention being paid to two techniques that can be implemented using respective R packages. Parameter estimation and model selection are also discussed. The paper concludes with a summary, discussion, and some thoughts on future work.

Keywords. Classification, clustering, discriminant analysis, `mclust`, mixture models, model-based clustering, model selection, parameter estimation, `pgmm`.

MSC: 62H30, 62F99.

1 Introduction

Clustering is performed on data to partition observations into subsets that are, in some respect, similar. When clustering data, we do not know the true group membership for any of the observations. Sometimes the number of clusters sought is known *a priori*, but when this is not the case we then have a selection problem with respect to the

Paul D. McNicholas (✉)(paul.mcnicholas@uoguelph.ca)
Received: April, 2011; June, 2011

number of clusters. The most simple and perhaps intuitive clustering approaches are hierarchical and, hierarchical clustering aside, the most famous approaches are probably partitioning clustering techniques, such as k-means clustering (cf. Hartigan and Wong, 1979) and k-medoids clustering (cf. Kaufman and Rousseeuw, 1990, Chapter 2). When performing classification, we know the group memberships of some observations and we try to use those to learn the memberships of the remaining or new observations. We can think of discriminant analysis as a type of classification, where some 'rule' is developed based on observations with known group memberships and this rule is used to classify the remaining or new observations. In machine learning parlance, one might consider that classification is either semi-supervised or supervised whereas discriminant analysis is supervised.

In this paper, selected clustering approaches that use finite mixture models are outlined, along with related methods for classification and discriminant analysis. Some focus is placed on two techniques that use Gaussian mixture models but other approaches are also discussed. Note that the use of mixture models for clustering and classification is now well established within the literature and so the arguments in their favour are not rehashed herein. The brief commentary by McLachlan (2011) is a good starting point if one wishes to review these arguments.

2 Model-Based Approaches

2.1 Finite Mixture Models

We say that a random vector $\mathbf{X}$ arises from a parametric finite mixture distribution if, for all $\mathbf{x} \subset \mathbf{X}$, we can write its density as

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

where $\pi_g > 0$, such that $\sum_{g=1}^{G} \pi_g = 1$ are called mixing proportions, the $f_g(\mathbf{x} \mid \boldsymbol{\theta}_g)$ are called component densities, and $\boldsymbol{\vartheta} = (\pi_1, \ldots, \pi_G, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G)$ is the vector of parameters. Note that $f(\mathbf{x} \mid \boldsymbol{\vartheta})$ in Equation 1 is called a G-component finite mixture density. 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 often taken to be of the same type; the Gaussian distribution is popular due to its mathematical tractability. When the component densities are multivariate Gaussian,

the mixture model density is

$$f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g \phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g), \tag{2}$$

where $\phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ is the density of a multivariate Gaussian random variable with mean $\boldsymbol{\mu}_g$ and covariance matrix $\boldsymbol{\Sigma}_g$. Note that several alternatives to the multivariate Gaussian distribution continue to be explored (see Section 3.3 for examples).

A family of mixture models is said to arise when various constraints are imposed upon the component densities, and most often upon the covariance structure; the result is a flexible modelling paradigm that incorporates more and less parsimonious models. Extensive details on finite mixture models and their applications are given by Titterington et al. (1985), McLachlan and Basford (1988), McLachlan and Peel (2000a), and Frühwirth-Schnatter (2006).

2.2 Model-Based Approaches to Clustering, Classification, and Discriminant Analysis

The idiom 'model-based clustering' has customarily been used when each member of a family of mixture models is fitted to data for clustering and the best model from amongst this family is selected by some criterion, most often the Bayesian information criterion (BIC; Schwarz, 1978). Herein, we shall use 'model-based clustering' in the customary fashion but note that the term can also be used more generally for the application of any model for clustering. 'Model-based classification' (e.g. McNicholas, 2010), or 'partial classification' (cf. McLachlan, 1992, Section 2.7), can be regarded as a semi-supervised version of model-based clustering, while model-based discriminant analysis (Hastie and Tibshirani, 1996) is a supervised version of model-based clustering. Model-based clustering, classification, and discriminant analysis are perhaps best explained through their respective likelihoods.

Let $\mathbf{z}_i$ denote the component membership of observation i, so that $z_{ig} = 1$ if observation i belongs to component g and $z_{ig} = 0$ otherwise. For convenience, we will assume finite Gaussian mixture models (Equation 2). First, we focus on clustering and so suppose that we observe n p-dimensional data vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$, all of which have unknown group

memberships. The likelihood for the mixture model can be written

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

Next, suppose that we are operating within the model-based classification paradigm; we have n observations, of which k have known group memberships. Suppose that we order these n observations so that the first k have known group memberships; we can do this without loss of generality. Then, the likelihood can be written

$$\mathcal{L}(\boldsymbol{\vartheta} \mid \mathbf{x}, \mathbf{z}) = \prod_{i=1}^{k} \prod_{g=1}^{G} \left[\pi_g \phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)\right]^{z_{ig}} \prod_{j=k+1}^{n} \sum_{h=1}^{H} \pi_h \phi(\mathbf{x}_j \mid \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h), \quad (3)$$

for $H \geq G$. The fact that we can search for a number of groups (H) greater than that already observed (G) gives model-based classification a flexibility not always present in classification approaches. From this likelihood (Equation 3), model-based clustering is clearly a special case of model-based classification that arises upon setting $k = 0$ and $H = G$ within the latter paradigm.

Finally, let us consider model-based discriminant analysis. As before, order the n observations so that the first k have known group memberships. Now, rather than using all n observations to estimate the unknown component memberships, we use only k, as follows. Form the likelihood

$$\mathcal{L}(\boldsymbol{\vartheta} \mid \mathbf{x}, \mathbf{z}) = \prod_{i=1}^{k} \prod_{g=1}^{G} \left[\pi_g \phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)\right]^{z_{ig}},$$

from the k observations with known group memberships; using the maximum likelihood estimates arising from this likelihood, we then compute the expected values

$$\hat{z}_{jg} := \frac{\hat{\pi}_g \phi(\mathbf{x}_j \mid \hat{\boldsymbol{\mu}}_g, \hat{\boldsymbol{\Sigma}}_g)}{\sum_{h=1}^{G} \hat{\pi}_h \phi(\mathbf{x}_j \mid \hat{\boldsymbol{\mu}}_h, \hat{\boldsymbol{\Sigma}}_h)},$$

for $j = k+1, \ldots, n$. These expected values play the role of a discriminant rule and the predicted group memberships are given by the maximum *a posteriori* classifications $\text{MAP}\{\hat{z}_{jg}\}$, where $\text{MAP}\{\hat{z}_{jg}\} = 1$ if $\max_g\{\hat{z}_{jg}\}$ occurs at component g, and $\text{MAP}\{\hat{z}_{jg}\} = 0$ otherwise, for $j = k + 1, \ldots, n$.

3 Two Families of Gaussian Mixture Models

3.1 The MCLUST Family

The best known family of mixture models is the MCLUST family (Banfield and Raftery, 1993; Celeux and Govaert, 1995; Fraley and Raftery, 2002), which is supported by the `mclust` package (Fraley and Raftery, 2006) for the R software (R Development Core Team, 2010). Members of the MCLUST family exhibit eigen-decomposed component covariance matrices, so that in the most general case (VVV; cf. Table 1) the component covariance structure is $\boldsymbol{\Sigma}_g = \lambda_g \mathbf{D}_g \mathbf{A}_g \mathbf{D}_g'$, where $\mathbf{D}_g$ is the matrix of eigenvectors, $\mathbf{A}_g$ is the diagonal matrix with entries proportional to the eigenvalues, and λ_g is the relevant constant of proportionality. The MCLUST family is a subset of the Gaussian parsimonious clustering models (GPCM; Celeux and Govaert, 1995) wherein valid combinations of the following constraints were used: $\lambda_g = \lambda$, $\mathbf{A}_g = \mathbf{A}$, $\mathbf{D}_g = \mathbf{D}$, $\mathbf{D}_g = \mathbf{I}_p$, and $\mathbf{A}_g = \mathbf{I}_p$, where $\mathbf{I}_p$ is the identity matrix of appropriate dimension (p). The MCLUST family (Table 1) comprises ten of the GPCM models and the nomenclature of each model derives from the cluster shapes.

Table 1: Nomenclature, covariance structure, and number of free covariance parameters for each member of the MCLUST family.

Model	Volume	Shape	Orientation	$\boldsymbol{\Sigma}_g$	Free covariance parameters
EII	Equal	Spherical	–	$\lambda\mathbf{I}$	1
VII	Variable	Spherical	–	$\lambda_g\mathbf{I}$	G
EEI	Equal	Equal	Axis-Aligned	$\lambda\mathbf{A}$	p
VEI	Variable	Equal	Axis-Aligned	$\lambda_g\mathbf{A}$	$p + G - 1$
EVI	Equal	Variable	Axis-Aligned	$\lambda\mathbf{A}_g$	$pG - G + 1$
VVI	Variable	Variable	Axis-Aligned	$\lambda_g\mathbf{A}_g$	pG
EEE	Equal	Equal	Equal	$\lambda\mathbf{D}\mathbf{A}\mathbf{D}'$	$p(p+1)/2$
EEV	Equal	Equal	Variable	$\lambda\mathbf{D}_g\mathbf{A}\mathbf{D}_g'$	$Gp(p+1)/2 - (G-1)p$
VEV	Variable	Equal	Variable	$\lambda_g\mathbf{D}_g\mathbf{A}\mathbf{D}_g'$	$Gp(p+1)/2 - (G-1)(p-1)$
VVV	Variable	Variable	Variable	$\lambda_g\mathbf{D}_g\mathbf{A}_g\mathbf{D}_g'$	$Gp(p+1)/2$

Fraley and Raftery (2002) use MCLUST for discriminant analysis (MCLUST DA); their approach facilitates multiple components per known group by using the BIC to choose the number of components as well as the best model for each group. The MCLUST family was used by Dean et al. (2006) for both classification and discriminant analysis. Within `mclust`, parameter estimation is carried out within the expectation-maximization (EM) algorithm framework (cf. Section 4.1) and so starting values are very important; Fraley and Raftery (2002) utilize a Gaussian model-based agglomerative hierarchical clustering procedure to give starting values for MCLUST (cf. Murtagh and Raftery, 1984; Banfield and Raftery, 1993). Members of the MCLUST family

with non-diagonal covariance structures (EEE, EEV, VEV, and VVV) have $\mathcal{O}(p^2)$ covariance parameters; i.e., the number of covariance parameters is quadratic in the dimensionality of the data. These members are therefore unsuited to the analysis of high-dimensional data.

3.2 Factor Analysis, Mixtures of Factor Analyzers, and the PGMM Family

Factor analysis (Spearman, 1904; Bartlett, 1953) is a well established multivariate statistical model wherein a p-dimensional random vector $\mathbf{X}$ is modelled using a q-dimensional vector of latent factors $\mathbf{U}$, where $q \ll p$. The model can be written $\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{U} + \boldsymbol{\epsilon}$, where $\boldsymbol{\Lambda}$ is a $p \times q$ matrix of factor loadings, the latent factors $\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_q)$, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$, where $\boldsymbol{\Psi}$ is a $p \times p$ diagonal matrix. The marginal distribution of $\mathbf{X}$ for this model is $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi})$. Note that the matrix of factor loadings $\boldsymbol{\Lambda}$ is not unique; if $\boldsymbol{\Lambda}$ is replaced by $\boldsymbol{\Lambda}^* = \boldsymbol{\Lambda}\mathbf{D}$ where $\mathbf{D}$ is orthonormal, then $\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi} = (\boldsymbol{\Lambda}^*)(\boldsymbol{\Lambda}^*)' + \boldsymbol{\Psi}$. This is one of the reasons that factor analysis has been, and perhaps still is, viewed with suspicion; Lawley and Maxwell (1962) outline how factor analysis "became the black sheep of statistical theory".

Ghahramani and Hinton (1997) developed a mixture of factor analyzers model; this model has the same density as a finite Gaussian mixture model (Equation 2) but with $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}$. Tipping and Bishop (1997, 1999) proposed the mixture of probabilistic principal component analyzers model, where the $\boldsymbol{\Psi}_g$ matrix in each component is isotropic so that $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \psi_g\mathbf{I}_p$; McLachlan and Peel (2000b) proposed a more general mixture of factor analyzers model with $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g$. McNicholas and Murphy (2005, 2008) later built upon these models to develop a family of eight Gaussian mixture models 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 $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g$. This family of models is known as the parsimonious Gaussian mixture model (PGMM) family (cf. Table 2). McNicholas (2010) used the PGMM family for model-based classification.

McNicholas and Murphy (2010b) further parameterized the factor analysis 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 $|\boldsymbol{\Delta}_g| = 1$. They call the resulting covariance structure $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \omega_g\boldsymbol{\Delta}_g$ the modified factor analysis covariance structure. In addition to the constraint $\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$, they impose all legitimate combinations of the constraints $\omega_g = \omega$, $\boldsymbol{\Delta}_g = \boldsymbol{\Delta}$, and

Table 2: The nomenclature and covariance structure of each PGMM.

$\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$	$\boldsymbol{\Psi}_g = \boldsymbol{\Psi}$	$\boldsymbol{\Psi}_g = \psi_g \mathbf{I}_p$	Covariance Structure
C	C	C	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \psi\mathbf{I}_p$
C	C	U	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$
C	U	C	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \psi_g\mathbf{I}_p$
C	U	U	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}_g$
U	C	C	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \psi\mathbf{I}_p$
U	C	U	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}$
U	U	C	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \psi_g\mathbf{I}_p$
U	U	U	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g$

C = constrained, U = unconstrained.

Table 3: The covariance structure and nomenclature for each member of the EPGMM family, along with the name of the equivalent member of the PGMM family where applicable.

EPGMM Nomenclature					
$\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$	$\boldsymbol{\Delta}_g = \boldsymbol{\Delta}$	$\omega_g = \omega$	$\boldsymbol{\Delta}_g = \mathbf{I}_p$	PGMM Equivalent	Covariance Structure
C	C	C	C	CCC	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \omega\mathbf{I}_p$
C	C	U	C	CUC	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \omega_g\mathbf{I}_p$
U	C	C	C	UCC	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \omega\mathbf{I}_p$
U	C	U	C	UUC	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \omega_g\mathbf{I}_p$
C	C	C	U	CCU	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \omega\boldsymbol{\Delta}$
C	C	U	U	–	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \omega_g\boldsymbol{\Delta}$
U	C	C	U	UCU	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \omega\boldsymbol{\Delta}$
U	C	U	U	–	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \omega\boldsymbol{\Delta}_g$
C	U	C	U	–	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \omega\boldsymbol{\Delta}_g$
C	U	U	U	CUU	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \omega_g\boldsymbol{\Delta}_g$
U	U	C	U	–	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \omega\boldsymbol{\Delta}_g$
U	U	U	U	UUU	$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \omega_g\boldsymbol{\Delta}_g$

C = constrained, U = unconstrained.

$\boldsymbol{\Delta}_g = \mathbf{I}_p$, giving a family of twelve Gaussian mixture models (Table 3) that they called the expanded PGMM (EPGMM) family. We shall use the term 'PGMM family' hereafter to mean this family of twelve models.

The `pgmm` package (McNicholas et al., 2011) for R provides an implementation of all twelve PGMM models for model-based clustering and classification; the classification is in the fashion described by McNicholas (2010). A key feature of the PGMM family is that all members have $\mathcal{O}(p)$ covariance parameters; i.e., the number of covariance parameters is linear in the dimensionality of the data under consideration (Table 4). This is one of the reasons that this family of models is well suited to the analysis of high-dimensional data.

Parameter estimation for the members of the PGMM family is carried out using alternating expectation-conditional maximization algorithms (cf. Section 4.1). The Woodbury identity (Woodbury, 1950) can

Table 4: The number of free covariance parameters for each member of the PGMM family.

Model	Number of Covariance Parameters
CCCC	$[pq - q(q-1)/2] + 1$
CCUC	$[pq - q(q-1)/2] + G$
UCCC	$G[pq - q(q-1)/2] + 1$
UCUC	$G[pq - q(q-1)/2] + G$
CCCU	$[pq - q(q-1)/2] + p$
CCUU	$[pq - q(q-1)/2] + [G + (p-1)]$
UCCU	$G[pq - q(q-1)/2] + p$
UCUU	$G[pq - q(q-1)/2] + [G + (p-1)]$
CUCU	$[pq - q(q-1)/2] + [1 + G(p-1)]$
CUUU	$[pq - q(q-1)/2] + Gp$
UUCU	$G[pq - q(q-1)/2] + [1 + G(p-1)]$
UUUU	$G[pq - q(q-1)/2] + Gp$

be used to avoid inversion of any non-diagonal $p \times p$ matrices on the iterations of these AECM algorithms; this is another advantage of the PGMM family for the analysis of high-dimensional data. For an $m \times m$ matrix $\mathbf{A}$, an $m \times k$ matrix $\mathbf{U}$, a $k \times k$ matrix $\mathbf{C}$, and a $k \times m$ matrix $\mathbf{V}$, the Woodbury identity is

$$(\mathbf{A} + \mathbf{UCV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{C}^{-1} + \mathbf{VA}^{-1}\mathbf{U})^{-1}\mathbf{VA}^{-1},$$

and setting $\mathbf{U} = \boldsymbol{\Lambda}_g$, $\mathbf{V} = \boldsymbol{\Lambda}_g'$, $\mathbf{A} = \omega_g\boldsymbol{\Delta}_g$, and $\mathbf{C} = \mathbf{I}_q$ gives

$$\begin{aligned}&(\omega_g\boldsymbol{\Delta}_g + \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g')^{-1} \\ &\quad = (\omega_g\boldsymbol{\Delta}_g)^{-1} - (\omega_g\boldsymbol{\Delta}_g)^{-1}\boldsymbol{\Lambda}_g(\mathbf{I}_q + \boldsymbol{\Lambda}_g'(\omega_g\boldsymbol{\Delta}_g)^{-1}\boldsymbol{\Lambda}_g)^{-1}\boldsymbol{\Lambda}_g'(\omega_g\boldsymbol{\Delta}_g)^{-1}. \quad (4)\end{aligned}$$

The left-hand-side of Equation 4 involves inversion of a $p \times p$ matrix but the right-hand-side has only diagonal and $q \times q$ matrices to be inverted. This gives an especially significant computational advantage when p is large and $q \ll p$; this arises, for example, in bioinformatics applications (e.g. McNicholas and Murphy, 2010b). A related identity for the determinant of the covariance matrix, $|\boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \omega_g\boldsymbol{\Delta}_g| = |\omega_g\boldsymbol{\Delta}_g|/|\mathbf{I}_q - \boldsymbol{\Lambda}_g'(\boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \omega_g\boldsymbol{\Delta}_g)^{-1}\boldsymbol{\Lambda}_g|$, is also helpful. These formulae are used by McLachlan and Peel (2000a) for the mixtures of factor analyzers model and by McNicholas and Murphy (2008, 2010b) and McNicholas (2010) for the PGMMs.

Another advantage of the PGMM family for the analysis of high-dimensional data is one that is shared by MCLUST or any other family

of mixture models; model-based approaches are 'trivially parallelizable' and so their implementation in parallel is relatively straightforward. With the decreasing cost of high-performance computing equipment, this advantage is becoming more and more real. McNicholas et al. (2010) describe how the PGMM family of models can be implemented in parallel; they include an illustration of the speed-up that can be achieved.

Andrews and McNicholas (2011b) use the PGMM family for model-based discriminant analysis while also introducing a t-analogue of the PGMM family. This latter comment leads nicely into Section 3.3.

3.3 Other Families of Mixture Models

The two families of models described thus far in Section 3 by no means form an exhaustive list of families of mixture models for clustering, classification, and discriminant analysis. These two families of models were chosen for review because of the associated R packages that enable the reader to use them. In the remainder of this section, some other families of models and approaches are mentioned but anything beyond a cursory nod is beyond the scope of this work.

Bouveyron et al. (2007) and McNicholas and Murphy (2010a) introduce families of Gaussian mixture models; in the former case, the family is designed for high-dimensional data and in the latter case, the family is used for clustering longitudinal data. McLachlan et al. (2007), building on work by McLachlan and Peel (1998) and Peel and McLachlan (2000), introduce a mixtures of t-factor analyzers model. Andrews et al. (2011) use a four-member family of mixtures of multivariate t-distributions for model-based classification. Andrews and McNicholas (2011a,b) introduce t-analogues of the PGMM family for model-based clustering, classification, and discriminant analysis. Baek et al. (2010) develop a variant of the mixture of factor analyzers model.

Raftery and Dean (2006) and Maugis et al. (2009a,b) discuss variable selection for clustering within the Gaussian mixture modelling framework, while Scrucca (2010) considers dimension reduction. Gormley and Murphy (2006) consider mixtures of Plackett-Luce models, Handcock et al. (2007) apply model-based clustering to social networks, and Karlis and Santourian (2009) discuss non-elliptically contoured distributions.

4 Parameter Estimation and Model Selection

4.1 Parameter Estimation

The expectation-maximization (EM) algorithm is an iterative procedure used to find maximum likelihood estimates when data are incomplete or are treated as being incomplete. The consummate citation for the EM algorithm is the famous paper by Dempster et al. (1977); however, Titterington et al. (1985, Section 4.3.2) also cite similar approaches used by Baum et al. (1970), Orchard and Woodbury (1972), and Sundberg (1974). In an EM algorithm, E- and M-steps are iterated until convergence is reached. The EM algorithm is based on the 'complete-data'; i.e., the observed data plus the missing data. In E-step, the expected value of the complete-data log-likelihood, $\mathcal{Q}$ say, is computed; in the M-step, $\mathcal{Q}$ is maximized with respect to the model parameters.

The EM algorithm and variants are commonly used for parameter estimation in model-based clustering, classification, and discriminant analysis. One such variant is the expectation-conditional maximization (ECM) algorithm (Meng and Rubin, 1993), wherein the M-step is replaced by a few conditional maximization steps that are typically more computationally efficient. When extra E-steps are added so that an E-step occurs before each of the CM-steps, the resulting algorithm is called a multicycle ECM algorithm (Meng and Rubin, 1993). The alternating ECM (AECM) algorithm (Meng and van Dyk, 1997) extends the ECM algorithm by allowing different specification of the complete-data at each stage. The AECM algorithm is used for parameter estimation for the members of the PGMM family because there are two sources of missing data: the latent factors and the group memberships. Problems with the EM algorithm include a sensitivity to starting values; the EM algorithm and various shortcomings are discussed by Titterington et al. (1985) and, along with several variations, by McLachlan and Krishnan (2008).

There are other, albeit less popular, approaches to parameter estimation in model-based clustering, classification, and discriminant analysis. Hunter and Lange (2000) formally introduced 'MM algorithms' but they, like the EM algorithm, had existed in various forms before they were systematically laid out. MM algorithms are a blueprint for a broad class of algorithms that can be summarized as minorization followed by maximization or majorization followed by minimization. The EM algorithm is an MM algorithm of the minorization-maximization variety. Drawing on a result from Hunter (2004), Gormley and Murphy (2006) used

an MM algorithm for parameter estimation in a model-based clustering context that used mixtures of Plackett-Luce models. However, there has been a general dearth of work on the use of non-EM MM algorithms for parameter estimation in model-based clustering, classification, and discriminant analysis.

Variational approximations, on the other hand, have been used for mixture model parameter estimation for some time now (e.g. Corduneanu and Bishop, 2001). Suppose we have observed data $\mathbf{x}$, missing data $\mathbf{z}$, and parameters $\boldsymbol{\varphi}$; the idea is to approximate the joint conditional distribution of $\boldsymbol{\varphi}$ and $\mathbf{z}$ given the observed data. McGrory and Titterington (2007) developed a variational Bayes algorithm for Gaussian mixture model parameter estimation. In addition to parameter estimation, their approach facilitates estimation of the number of mixture components G. Other Bayesian approaches to finite mixture model parameter estimation have also been utilized; e.g., Richardson and Green (1997), Zhang et al. (2004), and Dellaportas and Papageorgiou (2006) use reversible jump Markov chain Monte Carlo (MCMC) whereas Stephens (2000) describes an alternative MCMC approach.

4.2 Model Selection

When the term 'model selection' is used within the model-based clustering, classification, and discriminant analysis contexts, the meaning is usually two- or three-fold. First is the selection of the parametric structure, i.e., the selection of the best member of a family; second is the selection of the number of mixture components, if necessary; and third is the selection of the number of latent factors, if applicable. In certain circumstances, one may also need to choose between families of mixture models.

The Bayesian information criterion (BIC, Schwarz, 1978) remains the most prevalent mixture model selection technique within the literature; $\text{BIC} = 2l(\mathbf{x}, \hat{\boldsymbol{\Phi}}) - m \log n$, where m is the number of free parameters, n is the sample size, $\hat{\boldsymbol{\Phi}}$ is the maximum likelihood estimate of $\boldsymbol{\Phi}$, and $l(\mathbf{x}, \hat{\boldsymbol{\Phi}})$ is the maximized log-likelihood. The BIC is used to select the model and number of components in both `mclust` and `pgmm`, and is also used to select the number of factors in `pgmm`. Some theoretical and applied bases are often cited as supporting the use of the BIC (Leroux, 1992; Kass and Raftery, 1995; Kass and Wasserman, 1995; Keribin, 1998, 2000). However, the assumptions upon which these theoretical results are based may not hold, and the model selected by the BIC does not necessarily

give the best predicted classifications (see Andrews and McNicholas, 2011a, Section 5.3, for an example).

Credible alternatives to the BIC have been proposed, perhaps most notably the integrated completed likelihood (ICL; Biernacki et al., 2000), which is approximated by penalizing the BIC using the estimated mean entropy. However, like the BIC, the model selected by the ICL does not necessarily give the best predicted classifications. Despite the fact that the theoretical assumptions underpinning its use cannot necessarily be assumed to hold, and the fact that the most accurate estimated classifications do not necessarily emerge, the BIC remains the best of the approaches that have been tried. However, alternative approaches continue to be sought and some interesting work on information criteria and model selection has been carried out by Meila (2007) and Vinh et al. (2010), amongst others. Melnykov and Maitra (2010) give a nice review of work in model selection, focusing on selection of the number of mixture components.

5 Discussion

A review of work in model-based clustering, classification, and discriminant analysis has been presented, with a particular focus on two families of Gaussian mixture models: the MCLUST and PGMM families. Some space was also devoted to a brief discussion of other work, including non-Gaussian approaches. Although the Gaussian mixture model has historically been the most popular approach, a survey of the model-based clustering, classification, and discriminant analysis literature over the past few years demonstrates that much of the work has been non-Gaussian; examples include the work of Karlis and Santourian (2009), Lin (2010), and Andrews and McNicholas (2011b). Work has also been carried out on mixtures of different types of distributions; e.g., Coretto and Hennig (2011) consider mixtures of Gaussian and uniform distributions. Parameter estimation and model selection were also briefly discussed; a more in-depth discussion of model selection is given by Celeux (2007, Section 3).

Future work in the area will include efforts towards improved parameter estimation and model selection techniques. The departure from the Gaussian distribution is almost certain to become more pronounced, with non-symmetric models likely to feature. Work on the efficient implementation of model-based approaches in parallel is in its infancy and more work is expected in that direction, especially with the dimension-

ality of many modern data sets. A fully Bayesian approach to mixture modelling might well gain popularity; see Medvedovic and Sivaganesan (2002) and Dellaportas and Papageorgiou (2006) for examples. Finally, work on merging mixture components (cf. Hennig, 2010; Baudry et al., 2010) will surely continue in the coming years.

Acknowledgements

The author gratefully acknowledges the helpful comments of an anonymous reviewer. This work was supported by the University Research Chair in Computational Statistics at the University of Guelph, which is held by the author.

References

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., 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.

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.

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, pp. 23–43.

Baudry, J.-P., Raftery, A. E., Celeux, G., Lo, K., and Gottardo, R. (2010), Combining mixture components for clustering. Journal of Computational and Graphical Statistics, **19**(2), 332–353.

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.

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.

Bouveyron, C., Girard, S., and Schmid, C. (2007), High-dimensional data clustering. Computational Statistics and Data Analysis, **52**(1), 502–519.

Celeux, G. (2007), Mixture models for classification. In R. Decker and H. J. Lenz (Eds.), Advances in Data Analysis, Studies in Classification, Data Analysis, and Knowledge Organization, pp. 3–14. Berlin Heidelberg: Springer.

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, pp. 27–34. Los Altos, CA: Morgan Kaufmann.

Coretto, P. and Hennig, C. (2011), Maximum likelihood estimation of heterogeneous mixtures of Gaussian and uniform distributions. Journal of Statistical Planning and Inference, **141**(1), 462–473.

Dean, N., Murphy, T. B., and Downey, G. (2006), Using unlabelled data to update classification rules with applications in food authenticity studies. Journal of Royal Statistical Society, Series. C, **55**(1), 1–14.

Dellaportas, P. and Papageorgiou, I. (2006), Multivariate mixtures of normals with unknown number of components. Statistics and Computing, **16**(1), 57–68.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), Maximum likelihood from incomplete data via the EM algorithm. Journal of Royal Statistical Society, Series. B, **39**(1), 1–38.

Fraley, C. and Raftery, A. E. (2002), Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association, **97**(458), 611–631.

Fraley, C. and Raftery, A. E. (2006, September), MCLUST version 3 for R: Normal mixture modeling and model-based clustering. Technical Report 504, Department of Statistics, University of Washington. Minor revisions January 2007 and November 2007.

Frühwirth-Schnatter, S. (2006), Finite Mixture and Markov Switching Models. New York: Springer-Verlag.

Ghahramani, Z. and Hinton, G. E. (1997), The EM algorithm for factor analyzers. Technical Report CRG-TR-96-1, University Of Toronto, Toronto.

Gormley, I. C. and Murphy, T. B. (2006), Analysis of Irish third-level college applications data. Journal of Royal Statistical Society, Series. A, **169**(2), 361–379.

Handcock, M. S., Raftery, A. E., and Tantrum, J. M. (2007), Model-based clustering for social networks. Journal of Royal Statistical Society, Series. A, **170**(2), 301–354.

Hartigan, J. A. and Wong, M. A. (1979), A k-means clustering algorithm. Applied Statistics, **28**(1), 100–108.

Hastie, T. and Tibshirani, R. (1996), Discriminant analysis by Gaussian mixtures. Journal of Royal Statistical Society, Series. B, **58**(1), 155–176.

Hennig, C. (2010), Methods for merging Gaussian mixture components. Advances in Data Analysis and Classification, **4**, 3–34.

Hunter, D. L. and Lange, K. (2000), Rejoinder to discussion of "Optimization transfer using surrogate objective functions". Journal of Computational and Graphical Statistics, **9**, 52–59.

Hunter, D. R. (2004), MM algorithms for generalized Bradley-Terry models. The Annals of Statistics, **32**(1), pp. 384–406.

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.

Kass, R. E. and Wasserman, L. (1995), A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. Journal of the American Statistical Association, **90**(431), 928–934.

Kaufman, L. and Rousseeuw, P. J. (1990), Finding Groups in Data: An Introduction to Cluster Analysis. New York: Wiley.

Keribin, C. (1998), Estimation consistante de l'ordre de modèles de mélange. Comptes Rendus de l'Académie des Sciences. Série I. Mathématique, **326**(2), 243–248.

Keribin, C. (2000), Consistent estimation of the order of mixture models. Sankhyā. The Indian Journal of Statistics. Series A, **62**(1), 49–66.

Lawley, D. N. and Maxwell, A. E. (1962), Factor analysis as a statistical method. Journal of Royal Statistical Society, Series. D, **12**(3), 209–229.

Leroux, B. G. (1992), Consistent estimation of a mixing distribution. The Annals of Statistics, **20**(3), 1350–1360.

Lin, T.-I. (2010), Robust mixture modeling using multivariate skew t distributions. Statistics and Computing, **20**(3), 343–356.

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. (1992), Discriminant Analysis and Statistical Pattern Recognition. New Jersey: John Wiley & Sons.

McLachlan, G. J. (2011), Commentary on Steinley and Brusco (2011): Recommendations and cautions. Psychological Methods, **16**(1), 80–81.

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 edition), 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, pp.658–666. Berlin: Springer-Verlag.

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, pp.599–606. Morgan Kaufmann.

McNicholas, P. D. (2010), Model-based classification using latent Gaussian mixture models. Journal of Statistical Planning and Inference, **140**(5), 1175–1181.

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. The 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., Murphy, T. B., McDaid, A. F., and Banks, L. (2011), pgmm: Software for model-based clustering and classification via latent Gaussian mixture models. Technical Report 2011-319, Department of Mathematics and Statistics, University of Guelph, Guelph, Ontario, Canada.

McNicholas, P. D., Murphy, T. B., McDaid, A. F., and Frost, D. (2010), Serial and parallel implementations of model-based clustering via parsimonious Gaussian mixture models. Computational Statistics and Data Analysis, **54**(3), 711–723.

Medvedovic, M. and Sivaganesan, S. (2002), Bayesian infinite mixture model based clustering of gene expression profiles. Bioinformatics, **18**(9), 1194–1206.

Meila, M. (2007), Comparing clusterings — an information based distance. Journal of Multivariate Analysis, **98**(5), 873–895.

Melnykov, V. and Maitra, R. (2010), Finite mixture models and model-based clustering. Statistics Surveys, **4**, 80–116.

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, V. (1997), The EM algorithm — an old folk song sung to a fast new tune (with discussion), Journal of Royal Statistical Society, Series. B, **59**(3), 511–567.

Murtagh, F. and Raftery, A. E. (1984), Fitting straight lines to point patterns. Pattern Recognition, **17**(5), 479–483.

Orchard, T. and Woodbury, M. A. (1972), A missing information principle: Theory and applications. In L. M. Le. Cam, J. Neyman, and E. L. Scott (Eds.), Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Theory of Statistics, pp.697–715. Berkeley: University of California Press.

Peel, D. and McLachlan, G. J. (2000), Robust mixture modelling using the t distribution. Statistics and Computing, **10**(4), 339–348.

R Development Core Team (2010), R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Raftery, A. E. and Dean, N. (2006), Variable selection for model-based clustering. Journal of the American Statistical Association, **101**(473), 168–178.

Richardson, S. and Green, P. J. (1997), On Bayesian analysis of mixtures with an unknown number of components. (with discussion), Journal of Royal Statistical Society, Series. B, **59**(4), 731–792.

Schwarz, G. (1978), Estimating the dimension of a model. The Annals of Statistics, **6**(2), 461–464.

Scrucca, L. (2010), Dimension reduction for model-based clustering. Statistics and Computing, **20**(4), 471–484.

Spearman, C. (1904), The proof and measurement of association between two things. American Journal of Psychology, **15**, 72–101.

Stephens, M. (2000), Bayesian analysis of mixture models with an unknown number of components — an alternative to reversible jump methods. The Annals of Statistics, **28**(1), 40–74.

Sundberg, R. (1974), Maximum likelihood theory for incomplete data from an exponential family. Scandinavian Journal of Statistics, **1**(2), 49–58.

Tipping, T. 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, T. 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.

Vinh, N. X., Epps, J., and Bailey, J. (2010), Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. Journal of Machine Learning Research, **11**, 2837–2854.

Woodbury, M. A. (1950), Inverting modified matrices. Statistical Research Group, Memorandum Report 42. Princeton University, Princeton, New Jersey.

Zhang, Z., Chan, K. L., Wu, Y., and Chen, C. (2004), Learning a multivariate Gaussian mixture model with the reversible jump MCMC algorithm. Statistics and Computing, **14**(4), 343–355.