Stat Comput (2011) 21:361–373
DOI 10.1007/s11222-010-9175-2

Extending mixtures of multivariate t-factor analyzers

Jeffrey L. Andrews · Paul D. McNicholas

Received: 2 July 2009 / Accepted: 12 March 2010 / Published online: 10 April 2010

Abstract Model-based clustering typically involves the development of a family of mixture models and the imposition of these models upon data. The best member of the family is then chosen using some criterion and the associated parameter estimates lead to predicted group memberships, or clusterings. This paper describes the extension of the mixtures of multivariate t-factor analyzers model to include constraints on the degrees of freedom, the factor loadings, and the error variance matrices. The result is a family of six mixture models, including parsimonious models. Parameter estimates for this family of models are derived using an alternating expectation-conditional maximization algorithm and convergence is determined based on Aitken's acceleration. Model selection is carried out using the Bayesian information criterion (BIC) and the integrated completed likelihood (ICL). This novel family of mixture models is then applied to simulated and real data where clustering performance meets or exceeds that of established model-based clustering methods. The simulation studies include a comparison of the BIC and the ICL as model selection techniques for this novel family of models. Application to simulated data with larger dimensionality is also explored.

Keywords Factor analysis · Latent variables · Mixture models · Model-based clustering · Multivariate t-distributions · t-Factor analyzers

J.L. Andrews · P.D. McNicholas (✉)
Department of Mathematics and Statistics, University of Guelph, Guelph, Ontario, N1G 2W1, Canada
e-mail: pmcnicho@uoguelph.ca

J.L. Andrews
e-mail: andrewsj@uoguelph.ca

1 Introduction

1.1 Model-based clustering

Finite mixture models assume that data are collected from a finite collection of sub-populations and that the data within each sub-population can be modeled using some statistical model (cf. Sect. 1.2). The growing popularity of the use of mixture models for clustering is due, at least in part, to its intuitive appeal. Fraley and Raftery (2002) traced the use of finite mixture models for clustering back to the 1960s and 70s, citing the work of Wolfe (1963, 1970), Day (1969), and Binder (1978), amongst others.

The potential of such approaches became more and more apparent in subsequent decades. The provision of an exhaustive list of important work over the past three decades is not feasible here. However, the following works, amongst many others, would form part of such a list: McLachlan (1982), McLachlan and Basford (1988), Banfield and Raftery (1993), Celeux and Govaert (1995), Dasgupta and Raftery (1998), McLachlan and Peel (2000b), Fraley and Raftery (2002), Raftery and Dean (2006), McLachlan et al. (2007), McNicholas and Murphy (2008), and Gormley and Murphy (2008).

Nowadays, the term 'model-based clustering' is particularly common when a family of mixture models is fitted to data (see Sects. 2.1 and 2.3.2, for examples) and the best model is selected using some criterion, often the Bayesian information criterion (BIC Schwarz 1978). In this paper, the mixtures of multivariate t-factor analyzers (MMtFA) model (McLachlan et al. 2007) is developed into a family of six mixture models.

1.2 Finite mixture models

1.2.1 General framework

A random vector $\mathbf{X}$ is said to arise from a (parametric) finite mixture distribution if its density has the form

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

where $\pi_g \in [0, 1]$, such that $\sum_{g=1}^{G} \pi_g = 1$, are called mixing proportions and the $p_g(\mathbf{x} \mid \boldsymbol{\theta}_g)$ are referred to as component densities. Comprehensive reviews of finite mixture models are given by McLachlan and Peel (2000a) and Frühwirth-Schnatter (2006).

1.2.2 Gaussian mixture models

Gaussian mixture models have received the bulk of the attention in the mixture modeling literature due to their mathematical tractability. The model density can be written

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

where $\phi(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ is the density of a multivariate Gaussian distribution with mean $\boldsymbol{\mu}_g$ and covariance matrix $\boldsymbol{\Sigma}_g$. For more information on Gaussian mixture models and their use in clustering and dimensionality reduction see Celeux and Govaert (1995), Fraley and Raftery (2002), McLachlan and Peel (2000b), Raftery and Dean (2006) and McNicholas and Murphy (2008, 2010).

1.2.3 Mixtures of multivariate t-distributions

Mixtures of multivariate t-distributions have received considerably less attention to date. The model density for mixtures of multivariate t-distributions has the form

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

where π_g are the mixing proportions and

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

is the density of a p-dimensional multivariate t-distribution with mean $\boldsymbol{\mu}_g$, covariance matrix $\boldsymbol{\Sigma}_g$ and degrees of freedom ν_g, 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 Mahalanobis distance between $\mathbf{x}$ and $\boldsymbol{\mu}_g$. Parameter estimation can be carried out using the expectation-maximization algorithm, as described by McLachlan and Peel (1998).

Table 1 The covariance structure, and related nomenclature, of each member of the MMtFA family of models

Model	$\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$	$\boldsymbol{\Psi}_g = \psi_g \mathbf{I}_p$	$\nu_g = \nu$
CCC	Constrained	Constrained	Constrained
CCU	Constrained	Constrained	Unconstrained
UCC	Unconstrained	Constrained	Constrained
UCU	Unconstrained	Constrained	Unconstrained
UUC	Unconstrained	Unconstrained	Constrained
UUU	Unconstrained	Unconstrained	Unconstrained

2 Model-based clustering techniques

2.1 MCLUST

Some model-based clustering techniques are available within the R computing environment (R Development Core Team 2009). The most famous such package, `mclust` (Fraley and Raftery 2003; Fraley and Raftery 2006), uses a family of Gaussian mixture models with an eigen-decomposed covariance structure (Banfield and Raftery 1993; Celeux and Govaert 1995; Fraley and Raftery 1998, 2002). Specifically, the component covariance matrix is parameterized by $\boldsymbol{\Sigma}_g = \lambda_g \mathbf{D}_g \mathbf{A}_g \mathbf{D}_g'$, where λ_g is a constant, $\mathbf{D}_g$ is the matrix of eigenvectors and $\mathbf{A}_g$ is the diagonal matrix with entries proportional to the eigenvalues of $\boldsymbol{\Sigma}_g$. The family of MCLUST models is developed by imposing, or not, various combinations of the following constraints: $\lambda_g = \lambda$, $\mathbf{D}_g = \mathbf{D}$, $\mathbf{A}_g = \mathbf{A}$, $\mathbf{A} = \mathbf{I}_p$ and $\mathbf{D} = \mathbf{I}_p$. The MCLUST family of models is summarized in Fraley and Raftery (2006, Table 1). The MCLUST models with non-diagonal covariance structure have the property that the number of covariance parameters is quadratic in data dimensionality. Therefore, the `mclust` software is not well suited to the analysis of very high-dimensional data.

2.2 The variable selection technique

When clustering, it is often the case that some of the variables are unhelpful. The variable selection technique (Raftery and Dean 2006) was introduced to perform both data reduction and clustering. Variable selection involves repeated application of MCLUST to subsets of the variable space. Different models are compared using approximate Bayes factors (Kass and Raftery 1995). The technique is supported by the `clustvarsel` package (Dean and Raftery 2006) for R. Examples are given by Raftery and Dean (2006) to demonstrate that variable selection can give better clustering performance than MCLUST. However, McNicholas and Murphy (2008) show that variable selection can result in inferior clustering performance when compared to MCLUST. Furthermore, despite the fact that variable selection does result in a reduced set of variables it is, by its nature, not suitable for the analysis of high-dimensional data.

2.3 Mixtures of factor analyzers

2.3.1 *Mixtures of factor analyzers and mixtures of probabilistic principal components analyzers*

The factor analysis model assumes that a p-dimensional random vector $\mathbf{X}$ can be modeled using a q-dimensional vector of real-valued factors $\mathbf{U}$, such that $q < p$. Mathematically, this can be expressed as $\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{U} + \boldsymbol{\epsilon}$, where $\boldsymbol{\Lambda}$ is a $p \times q$ matrix of factor loadings, $\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_q)$ is the vector of factors and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$ with $\boldsymbol{\Psi} = \operatorname{diag}(\psi_1, \psi_2, \ldots, \psi_p)$. Note that the probabilistic principal component analysis (PPCA) model (Tipping and Bishop 1999b) is a special case of the factor analysis model with isotropic covariance structure $\boldsymbol{\Psi} = \psi \mathbf{I}_p$, where $\mathbf{I}_p$ is the p-dimensional identity matrix.

The factor analysis model was extended to the mixtures of factor analyzers model by Ghahramani and Hinton (1997). Under this model, the mixture model density given in (1) is modified by parameterizing the Gaussian covariance matrix with the factor analysis covariance structure $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g \boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}$. Tipping and Bishop (1999a) used a mixture of PPCAs model, where $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g \boldsymbol{\Lambda}_g' + \psi_g \mathbf{I}_p$, and McLachlan and Peel (2000b) introduced the fully unconstrained covariance structure for the mixtures of factor analyzers model, namely $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g \boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g$.

2.3.2 *Parsimonious Gaussian mixture models*

McNicholas and Murphy (2005, 2008) further generalized these models by allowing constraints across groups on the $\boldsymbol{\Lambda}_g$ and $\boldsymbol{\Psi}_g$ matrices as well as utilizing the isotropic constraint. The result was a family of eight parsimonious Gaussian mixture models (PGMMs), five of which were novel. The models with constrained loading matrices ($\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$) present a particularly parsimonious covariance structure, allowing for between $pq - q(q-1)/2 + 1$ and $pq - q(q-1)/2 + Gp$ covariance parameters. Furthermore, all eight models have the property that the number of covariance parameters is linear in p, making them much more suitable for the analysis of high-dimensional data than MCLUST. Extensive details of the implementation of this family of mixture models, both in serial and in parallel, are given by McNicholas et al. (2010).

2.4 Mixtures of multivariate t-factor analyzers

2.4.1 *Mixtures of probabilistic principal t-component analyzers*

Zhao and Jiang (2006) introduced the mixtures of probabilistic principal t-component analyzers (MPPtCA) model. Similarly to the mixtures of PPCAs in the Gaussian case, the mixture model density given in (2) is modified by parameterizing the component covariance matrix with the PPCA covariance structure $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g \boldsymbol{\Lambda}_g' + \psi_g \mathbf{I}_p$. Zhao and Jiang (2006) used a mixed approach to parameter estimation, which is discussed further in Sect. 3.2.4. The application for which they used the MPPtCA model was image compression and their work did not contain any clustering applications.

2.4.2 *Mixtures of multivariate t-factor analyzers*

McLachlan et al. (2007) introduced the MMtFA model. Here, the mixture model density given in (2) is modified by parameterizing the covariance matrix with the most general factor analysis covariance structure $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g \boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g$. In contrast to the application of Zhao and Jiang (2006), McLachlan et al. (2007) used the MMtFA model for clustering. Details of model fitting are given in Sect. 3.2.2.

2.5 The GMMDR technique

Scrucca (2009) introduced a model-based clustering technique that combines Gaussian mixture models with dimensionality reduction (GMMDR). Instead of imposing a structure with underlying latent variables, GMMDR looks for the subspace of the data that contains the most relevant clustering information. The vectors that span this subspace are estimated through an eigen-decomposition of a kernel matrix. Details of the methodology, visualization techniques, and an efficient greedy search algorithm are given by Scrucca (2009). The Italian wine data that were analyzed by Scrucca (2009, Sect. 5.1) are also analyzed in Sect. 5.3.2 herein. The results of these analyses are compared in Sect. 5.3.4.

3 Extending the MMtFA model

3.1 The models

In the MMtFA and MPPtCA models, the degrees of freedom for each component ν_g can be estimated from the data within the maximum likelihood framework (see Sect. 3.2.2). Herein, the MMtFA and MPPtCA models are extended by imposing combinations of the constraints: $\nu_g = \nu$, $\boldsymbol{\Psi}_g = \psi_g \mathbf{I}$, and $\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$. The result is a family of six mixture models, which are described in Table 1. Hereafter, these six models will be collectively referred to as the MMtFA family of models. Note that, like the PGMM models, all six of the MMtFA models have the property that the number of covariance parameters is linear in p.

The notion of constraining the degrees of freedom may seem an unnecessary one since the parsimony gained will be relatively small unless G is very large, which is not common in clustering applications. In practice, however, the models with constrained degrees of freedom can give better clustering performance than the unconstrained models. Examples of this phenomenon are given in Sect. 5.

3.2 Parameter estimation

3.2.1 *The alternating expectation-conditional maximization algorithm*

The expectation-maximization (EM) algorithm (Dempster et al. 1977) is an iterative technique for finding maximum likelihood estimates when data is incomplete. Furthermore, problems are often framed as incomplete-data problems in order to achieve efficient solutions using the EM algorithm. In the expectation step (E-step), the expected value of the complete-data log-likelihood, Q say, is computed. Then in the maximization step (M-step), Q is maximized with respect to the model parameters. Note that the complete-data is the missing data plus the observed data.

The expectation-conditional maximization (ECM) algorithm (Meng and Rubin 1993) sees the replacement of the M-step by a number of conditional maximization steps that are more computationally efficient. The alternating expectation-conditional maximization (AECM) algorithm (Meng and van Dyk 1997) is an extension of the ECM algorithm that permits different specification of the complete-data at each stage. Extensive details on the EM algorithm and variants are given by McLachlan and Krishnan (2008).

3.2.2 *The AECM algorithm for the MMtFA family*

Proceeding as outlined by McLachlan et al. (2007), the multivariate t-distribution can be characterized by introducing a random variable $W_{ig} \sim \text{gamma}(\nu_g/2, \nu_g/2)$. This means that all six members of the family will have three types of missing data: these w_{ig}, the latent factors $\mathbf{u}_{ig}$ and the z_{ig}, where $z_{ig} = 1$ if observation i is in component g and $z_{ig} = 0$ otherwise. Therefore, the AECM algorithm is appropriate.

For the MMtFA model, McLachlan et al. (2007) deduce that the distribution of observation $\mathbf{x}_i$ in component g, conditional on the missing data, is given by

$$\mathbf{X}_i \mid \mathbf{u}_{ig}, w_{ig}, z_{ig} = 1 \sim \mathcal{N}(\boldsymbol{\mu}_g + \boldsymbol{\Lambda}_g \mathbf{u}_{ig}, \boldsymbol{\Psi}_g / w_{ig}).$$

At the first stage of the AECM algorithm, when estimating π_g, $\boldsymbol{\mu}_g$ and ν_g, the missing data consist of the w_{ig} and the z_{ig}. At the second stage of the algorithm, when estimating $\boldsymbol{\Lambda}_g$ and $\boldsymbol{\Psi}_g$, the missing data are the $\mathbf{u}_{ig}$, w_{ig} and z_{ig}. For an observed random sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$, the complete-data log-likelihood for the MMtFA model is given by

$$l_c(\boldsymbol{\vartheta}) = \sum_{g=1}^{G} \sum_{i=1}^{n} z_{ig} \log a_{ig},$$

where

$$a_{ig} = \pi_g \gamma\left(w_j \mid \nu_g/2, \nu_g/2\right) \phi(\mathbf{u}_{ig} \mid \mathbf{0}, \mathbf{I}_q/w_{ig}) \times \phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g + \boldsymbol{\Lambda}_g \mathbf{u}_{ig}, \boldsymbol{\Psi}_g/w_{ig}),$$

where $\gamma(\cdot)$ is the gamma density (cf. McLachlan et al. 2007, (29)). In order to compute the expected value of the complete-data log-likelihood, the conditional expectations of terms like $Z_{ig} W_{ig} \mathbf{U}_{ig}$ and $Z_{ig} W_{ig} \mathbf{U}_{ig} \mathbf{U}_{ig}'$ are needed. McLachlan et al. (2007) give the conditional expectations of these terms as

$$\mathbb{E}[Z_{ig} W_{ig} \mathbf{U}_{ig} \mid \mathbf{x}_i, w_{ig}] = w_{ig} \boldsymbol{\beta}_g (\mathbf{x}_i - \boldsymbol{\mu}_g),$$

$$\mathbb{E}[Z_{ig} W_{ig} \mathbf{U}_{ig} \mathbf{U}_{ig}' \mid \mathbf{x}_i, w_{ig}] = \mathbf{I}_q - \boldsymbol{\beta}_g \boldsymbol{\Lambda}_g + w_{ig} \boldsymbol{\beta}_g (\mathbf{x}_i - \boldsymbol{\mu}_g)(\mathbf{x}_i - \boldsymbol{\mu}_g)' \boldsymbol{\beta}_g',$$

where $\boldsymbol{\beta}_g = \boldsymbol{\Lambda}_g' (\boldsymbol{\Lambda}_g \boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g)^{-1}$. These results, which are very similar to those derived in the Gaussian case, arise from theory around the expected value of sufficient statistics for the exponential family. In fact, much of the mathematics relating to the MMtFA model is analogous to the Gaussian case. McLachlan and Peel (2000a) give extensive details in the Gaussian case, and McLachlan et al. (2007) provide expansive details in the case of the multivariate t-distribution.

E-step

At each E-step, the indicator variables z_{ig} and the weights w_{ig} are updated by their conditional expected values

$$\hat{z}_{ig} = \frac{\pi_g f_t(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \nu_g)}{\sum_{h=1}^{G} \pi_h f_t(\mathbf{x} \mid \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h, \nu_h)},$$

and

$$\hat{w}_{ig} = \frac{\nu_g + p}{\nu_g + \delta(\mathbf{x}_i, \boldsymbol{\mu}_g \mid \boldsymbol{\Sigma}_g)},$$

respectively, where $\delta(\mathbf{x}_i, \boldsymbol{\mu}_g \mid \boldsymbol{\Sigma}_g)$ is the Mahalanobis distance between $\mathbf{x}_i$ and $\boldsymbol{\mu}_g$.

CM-step 1

On the first CM-step, the mixing proportions π_g and component means $\boldsymbol{\mu}_g$ are updated by

$$\hat{\pi}_g = \frac{n_g}{n} \quad \text{and} \quad \hat{\boldsymbol{\mu}}_g = \frac{\sum_{i=1}^{n} \hat{z}_{ig} \hat{w}_{ig} \mathbf{x}_i}{\sum_{i=1}^{n} \hat{z}_{ig} \hat{w}_{ig}},$$

respectively, where $n_g = \sum_{i=1}^{n} \hat{z}_{ig}$. The estimates for the ν_g do not exist in closed form but estimates can be found by setting

$$1 - \varphi\left(\frac{\hat{\nu}_g^{\text{new}}}{2}\right) + \log\left(\frac{\hat{\nu}_g^{\text{new}}}{2}\right) + \varphi\left(\frac{\hat{\nu}_g^{\text{old}} + p}{2}\right) + \frac{1}{n_g} \sum_{i=1}^{n} \hat{z}_{ig} (\log \hat{w}_{ig} - \hat{w}_{ig}) - \log\left(\frac{\hat{\nu}_g^{\text{old}} + p}{2}\right)$$

equal to zero and solving for $\hat{\nu}_g^{\text{new}}$, where $\hat{\nu}_g^{\text{old}}$ is the previous estimate of ν_g and $\varphi(\cdot)$ is the digamma function.

CM-Step 2

On the second CM-step, the factor loadings $\boldsymbol{\Lambda}_g$ and the diagonal error variance matrix $\boldsymbol{\Psi}_g$ are updated by $\hat{\boldsymbol{\Lambda}}_g^{\text{new}} = \mathbf{S}_g\hat{\boldsymbol{\beta}}_g'\boldsymbol{\Theta}_g^{-1}$ and $\hat{\boldsymbol{\Psi}}_g^{\text{new}} = \text{diag}\{\mathbf{S}_g - \hat{\boldsymbol{\Lambda}}_g\hat{\boldsymbol{\beta}}_g\mathbf{S}_g\}$, respectively, where

$$\mathbf{S}_g = \frac{1}{n_g}\sum_{i=1}^{n}\hat{z}_{ig}\hat{w}_{ig}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)',$$

and $\boldsymbol{\Theta}_g = \mathbf{I}_p - \hat{\boldsymbol{\beta}}_g\hat{\boldsymbol{\Lambda}}_g + \hat{\boldsymbol{\beta}}_g\mathbf{S}_g\hat{\boldsymbol{\beta}}_g'$. The notational convenience of $\boldsymbol{\Theta}_g$ was employed by McNicholas and Murphy (2005, 2008) and it is interesting to note that $\boldsymbol{\Theta}_g = \mathbf{I}_q$ if $\mathbf{S}_g = \boldsymbol{\Sigma}_g$.

3.2.3 Constraining the degrees of freedom parameter

The degrees of freedom parameter ν_g is effectively a shape parameter. The UUC and UCC models are the MMtFA (or UUU) and the MPPtCA (or UCU) models, respectively, with the constraint that $\nu_g = \nu$.

In order to find the maximum likelihood estimate for ν, the relevant complete-data log-likelihood function is

$$l_c(\nu) = \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\Big[K - \log\Gamma\left(\frac{\nu}{2}\right) + \frac{\nu}{2}\log\left(\frac{\nu}{2}\right) + \frac{\nu}{2}\left(\log w_{ig} - w_{ig}\right)\Big],$$

where K is constant with respect to ν. Upon differentiating the expected value of $l_c(\nu)$ with respect to ν, it can be determined that the update for ν is found by setting the equation

$$1 - \varphi\left(\frac{\hat{\nu}^{\text{new}}}{2}\right) + \log\left(\frac{\hat{\nu}^{\text{new}}}{2}\right) + \varphi\left(\frac{\hat{\nu}^{\text{old}} + p}{2}\right) + \frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n}\hat{z}_{ig}(\log\hat{w}_{ig} - \hat{w}_{ig}) - \log\left(\frac{\hat{\nu}^{\text{old}} + p}{2}\right)$$

equal to zero and solving for $\hat{\nu}^{\text{new}}$, where $\hat{\nu}^{\text{old}}$ is the previous estimate of ν.

3.2.4 Estimates under the isotropic constraint

As mentioned in Sect. 2.4.1, Zhao and Jiang (2006) imposed the isotropic constraint $\boldsymbol{\Psi}_g = \psi_g\mathbf{I}_p$ in an image compression application. In their parameter estimation, a "special AECM" algorithm was used and ν_g was numerically approximated using the method of Shoham (2002). Herein, the AECM algorithm is used, as outlined in Sect. 3.2.2, and the estimates for the degrees of freedom are updated as detailed in Sects. 3.2.2 and 3.2.3. When the isotropic constraint is imposed, the update for $\hat{\psi}_g$ is given by

$$\hat{\psi}_g^{\text{new}} = \frac{1}{p}\text{tr}\{\mathbf{S}_g - \hat{\boldsymbol{\Lambda}}_g^{\text{new}}\hat{\boldsymbol{\beta}}_g\mathbf{S}_g\}.$$

Note that this is identical to the update in the Gaussian case except that here the weights $\hat{w}_{ig}$ come in.

3.2.5 Estimates under the constrained loading matrix

Imposing the constraint $\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$ gives the largest reduction in the number of free parameters. The relevant complete-data log-likelihood, with the isotropic constraint also imposed, is

$$l_c(\boldsymbol{\Lambda}) = K - \sum_{i=1}^{n}\sum_{g=1}^{G}\frac{w_{ig}}{2\psi_g}\parallel\mathbf{x}_i - \boldsymbol{\mu}_g - \boldsymbol{\Lambda}\mathbf{u}_{ig}\parallel^2,$$

where K is constant with respect to $\boldsymbol{\Lambda}$. The expected value of this complete-data log-likelihood is

$$Q(\boldsymbol{\Lambda}) = K + \sum_{g=1}^{G}\frac{n_g}{2\psi_g}\left[2\,\text{tr}\left\{\hat{\boldsymbol{\beta}}_g\mathbf{S}_g\boldsymbol{\Lambda}\right\} - \text{tr}\left\{\boldsymbol{\Theta}_g\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\right\}\right].$$

Differentiating $Q(\boldsymbol{\Lambda})$ with respect to $\boldsymbol{\Lambda}$, and solving the equation that results from setting the resulting score function equal to zero, results in the update

$$\hat{\boldsymbol{\Lambda}}^{\text{new}} = \left[\sum_{g=1}^{G}\frac{n_g}{\hat{\psi}_g}\mathbf{S}_g\hat{\boldsymbol{\beta}}_g'\right]\left[\sum_{g=1}^{G}\frac{n_g}{\hat{\psi}_g}\boldsymbol{\Theta}_g\right]^{-1}.$$

Of course, by constraining $\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$ the updates for ψ_g are effected. In an analogous fashion to the derivation of $\hat{\boldsymbol{\Lambda}}^{\text{new}}$, it can be shown that

$$\hat{\psi}_g^{\text{new}} = \frac{1}{p}\text{tr}\{\mathbf{S}_g - 2\hat{\boldsymbol{\Lambda}}^{\text{new}}\hat{\boldsymbol{\beta}}_g\mathbf{S}_g + \hat{\boldsymbol{\Lambda}}^{\text{new}}\boldsymbol{\Theta}_g(\hat{\boldsymbol{\Lambda}}^{\text{new}})'\}.$$

These updates for $\hat{\boldsymbol{\Lambda}}^{\text{new}}$ and $\hat{\psi}^{\text{new}}$ are similar to updates for one of the PGMM models.

3.3 Mixture model selection and performance

3.3.1 The Bayesian information criterion

The BIC is a commonly used method for model selection in model-based clustering applications involving a family of mixture models (Fraley and Raftery 2002; McNicholas and Murphy 2008, 2010). The use of the BIC in mixture model selection was proposed by Dasgupta and Raftery (1998), based on an approximation to Bayes factors (Kass and Raftery 1995).

For a model with parameters $\boldsymbol{\Phi}$, the BIC is given by

$$\text{BIC} = 2l(\mathbf{x}, \hat{\boldsymbol{\Phi}}) - m\log n,$$

where $l(\mathbf{x}, \hat{\boldsymbol{\Phi}})$ is the maximized log-likelihood, $\hat{\boldsymbol{\Phi}}$ is the maximum likelihood estimate of $\boldsymbol{\Phi}$, m is the number of free parameters in the model and n is the number of observations.

Leroux (1992) and Keribin (2000) present theoretical results that, under certain regulatory conditions, support the use of the BIC for the estimation of the number of components in a mixture model. Furthermore, Lopes and West (2004) report the results of a simulation study that shows that the BIC is very effective at selecting the number of factors in a factor analysis model.

3.3.2 The integrated completed likelihood

One issue with using the BIC in model-based clustering applications is that a mixture component (group) is not necessarily the same as a true cluster. In an attempt to focus model selection on clusters rather than mixture components, Biernacki et al. (2000) introduced the integrated completed likelihood (ICL). The ICL, or the approximate ICL to be precise, is just the BIC penalized for estimated mean entropy and it is given by

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

where $\text{MAP}\{\hat{z}_{ig}\}$ is the maximum *a posteriori* classification given $\hat{z}_{ig}$, that is

$$\text{MAP}\{\hat{z}_{ig}\} = \begin{cases} 1 & \text{if } \max_g\{z_{ig}\} \text{ occurs at component } g, \\ 0 & \text{otherwise.} \end{cases}$$

Note that the MAP classification is used to give the predicted classifications (clusterings) in the analyses in Sect. 5.

The estimated mean entropy is a measure of the uncertainty in the classification of observation i into component g and so the ICL should be less likely, compared to the BIC, to split one cluster into two mixture components. Comparisons of the BIC and the ICL in model selection for the family of models introduced herein are given in Sect. 5.

3.3.3 Rand and adjusted Rand indices

Although the data analysis examples of Sect. 5 are conducted as clustering examples, the true classifications are actually known for these data. In these examples, the adjusted Rand index (Hubert and Arabie 1985) is used to measure class agreement. The Rand (1971) index can be expressed as

$$\frac{\text{number of agreements}}{\text{number of agreements} + \text{number of disagreements}},$$

where the number of agreements and the number of disagreements are based on pairwise comparisons. The Rand index is calculated on the interval [0, 1], where '0' indicates no pairwise agreements between the MAP classification and true group membership and '1' indicates perfect agreement. One criticism of the Rand index is that its expected value is greater than 0, making smaller values of the Rand index difficult to interpret.

The adjusted Rand index corrects the Rand index for chance by accounting for the fact that classification performed randomly would probably correctly classify some cases. The adjusted Rand index has an expected value of 0 under random classification and perfect classification would result in a value of 1.

4 Computational issues

4.1 Initialization

Before running the AECM algorithms for the MMtFA family of models, the $\hat{z}_{ig}$ must be initialized. One option is to initialize the $\hat{z}_{ig}$ randomly, however this opens the algorithm up to possible failure due to bad starting values. Another option, one that the `mclust` package utilizes, is to use an agglomerative hierarchical clustering procedure. This latter option is used in the analyses herein for which $n > p$.

The covariance matrices must also be initialized. Following McNicholas and Murphy (2008), an eigen-decomposition of the sample covariance matrix is used to get $\mathbf{S}_g = \mathbf{P}_g\mathbf{D}_g\mathbf{P}_g^{-1}$, where $\mathbf{D}_g$ is the diagonal matrix of eigenvalues. The $\hat{\mathbf{\Lambda}}_g$ are initialized by $\hat{\mathbf{\Lambda}}_g = \mathbf{d}_g\mathbf{P}_g$, where $\mathbf{d}_g$ is a vector comprising the square roots of the diagonal elements of $\mathbf{D}_g$. The $\hat{\mathbf{\Psi}}_g$ are then initialized as $\hat{\mathbf{\Psi}}_g = \text{diag}\{\mathbf{S}_g - \hat{\mathbf{\Lambda}}_g\hat{\mathbf{\Lambda}}_g'\}$ and the degrees of freedom are each initialized at 50.

4.2 Estimating the degrees of freedom

Code for all of the analyses herein was written in R and a numerical search for the estimates of the degrees of freedom was carried out using the `uniroot` command in the `stats` package. The `uniroot` command is based on the Fortran subroutine `zeroin` described by Brent (1973). The `lgamma` function was used to facilitate efficient calculation of the log-likelihood in the log scale. Note that the range of values for $\hat{\nu}_g$ was restricted to a maximum of 200 in order to facilitate faster convergence. An analysis of simulated data, given in Sect. 5.1.1, shows that this upper limit of 200 does not appear to hamper recovery of an underlying Gaussian structure—further evidence is provided in an analysis of higher dimensional data in Sect. 5.1.3.

4.3 Convergence of AECM algorithms

Aitken's acceleration is used to estimate the asymptotic maximum of the log-likelihood at each iteration. Based on this estimate, a decision can be made on whether or not an

AECM algorithm has converged. Aitken's acceleration at iteration k is given by

$$a^{(k)} = \frac{l^{(k+1)} - l^{(k)}}{l^{(k)} - l^{(k-1)}},$$

where $l^{(k+1)}$, $l^{(k)}$ and $l^{(k-1)}$ are the log-likelihood values from iterations $k+1$, k and $k-1$, respectively. Then an asymptotic estimate of the log-likelihood at iteration $k+1$ is given by

$$l_\infty^{(k+1)} = l^{(k)} + \frac{1}{1-a^{(k)}}(l^{(k+1)} - l^{(k)})$$

(Böhning et al. 1994). Lindsay (1995) suggests that the algorithm can be stopped when $l_\infty^{(k+1)} - l^{(k+1)} < \epsilon$ and McNicholas et al. (2010) modify this slightly and stop the algorithm when $l_\infty^{(k+1)} - l^{(k)} < \epsilon$. This modified version has the advantage that it is necessarily at least as strict as the lack of progress, $l^{(k+1)} - l^{(k)} < \epsilon$, which is used by `mclust`. The criterion given by McNicholas et al. (2010) is used for the analysis herein, with $\epsilon = 0.05$.

4.4 The Woodbury identity

One computational advantage of the MM*t*FA family of models—the fact that the number of covariance parameters is linear in p—has already been mentioned. However, these models, along with the PGMMs, have another significant advantage that is particularly important in applications involving high-dimensional data. When running the AECM algorithm for these models, it is advantageous to make use of the Woodbury identity (Woodbury 1950) to avoid inverting any non-diagonal $p \times p$ matrices.

Given an $n \times n$ matrix $\mathbf{A}$, an $n \times k$ matrix $\mathbf{U}$, a $k \times k$ matrix $\mathbf{C}$ and a $k \times n$ matrix $\mathbf{V}$, the Woodbury identity states that

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

Now, setting $\mathbf{U} = \boldsymbol{\Lambda}$, $\mathbf{V} = \boldsymbol{\Lambda}$, $\mathbf{A} = \boldsymbol{\Psi}$ and $\mathbf{C} = \mathbf{I}_q$ gives

$$(\boldsymbol{\Psi} + \boldsymbol{\Lambda}\boldsymbol{\Lambda}')^{-1} = \boldsymbol{\Psi}^{-1} - \boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}(\mathbf{I}_q + \boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1},$$

and although the left hand side of this equation involves inversion of a $p \times p$ matrix, the right hand side leaves only diagonal and $q \times q$ matrices to be inverted. This presents a major computational advantage, especially when p is large and $q \ll p$. A useful identity for the determinant of the covariance matrix follows from this, namely that

$$|\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}| = |\boldsymbol{\Psi}|/|\mathbf{I}_q - \boldsymbol{\Lambda}'(\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi})^{-1}\boldsymbol{\Lambda}|.$$

These formulae for $(\boldsymbol{\Psi} + \boldsymbol{\Lambda}\boldsymbol{\Lambda}')^{-1}$ and $|\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}|$ are used by McLachlan and Peel (2000a) for the mixtures of factor analyzers model, and by McNicholas and Murphy (2008) for the PGMMs.

Table 2 Classification table for the best `mclust` model on the third data set that was generated to assess the ability of the MM*t*FA family to recover an underlying Gaussian structure

	1	2	3	4	5	6
A	30	20				
B			22	23		
C					46	24

5 Data analyses

5.1 Simulated data

5.1.1 Recovering a Gaussian structure

In order to determine whether the MM*t*FA family can recover underlying mixtures of multivariate Gaussian distributions, and whether it can do so given the upper bound placed on the ν_g (cf. Sect. 4.2), three data sets were simulated. These data, which were generated using the `rnorm` command in R, had $G = 3$ components and $p = 6$ variables. Three data sets were generated with differing covariances. The clustering performance of the MM*t*FA family of models was then compared to that of MCLUST.

In the first simulation, both techniques underestimated the number of components, choosing $G = 2$ component models, and had identical adjusted Rand indices (0.67). In the second simulation, both methods clustered the data perfectly. In the third simulation, the results were very different. From the MM*t*FA family, the best model according to both the BIC (−1039.9) and ICL (−1040.0) was the UUC model with $q = 2$ factors, $G = 3$ components and $\hat{\nu} = 200$. This model clustered the data perfectly. Running `mclust` on the same data resulted in a $G = 6$ component model with an adjusted Rand index of 0.59. The classification table for this `mclust` model is given in Table 2. From Table 2 it is apparent that the best `mclust` model has split each of the three true clusters (A, B, and C) across two components.

In this small simulation study, an upper limit of 200 for the estimated degrees of freedom did not seem to hamper the performance of the MM*t*FA family of models in the recovery of underlying Gaussian components. In fact, the MM*t*FA family of models performed at least as well as MCLUST on all three simulated data sets. Furthermore, on the third data set one of the four new models (UUC) was selected and it returned much better clustering performance than MCLUST.

5.1.2 Comparing the BIC and the ICL

Three-dimensional Gaussian data sets were simulated in order to investigate differencies between the model selection performance of the BIC and the ICL. Two simulations

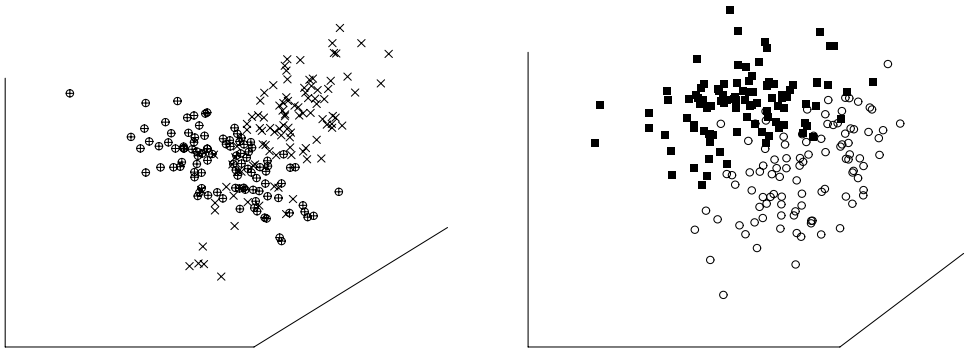

Fig. 1 Sample plots of the simulated data in the X-shaped case (left) and the 8-shaped case (right)

are presented: one where the two groups intersect (the 'X-shaped' case) and one where they simply overlap (the '8-shaped' case). In each of these two cases, 100 data sets were simulated with $n = 200$ observations divided equally among two groups. An example of each case is given in Fig. 1.

In the first study, the X-shaped case, the BIC outperformed the ICL as a model selection criterion for the MMtFA family. The BIC correctly chose a $G = 2$ group model on 98 out of 100 runs (selecting the UCC model on 90 runs), whereas the ICL selected a model with the correct number of components on 78 runs. The `mclust` software, which uses the BIC for model selection, performed better than the MMtFA family with the ICL criterion but overestimated the number of components more often than the MMtFA family with the BIC. The performance of all three techniques in the X-shaped case is summarized in Table 3.

Summary results for the 8-shaped simulations are given in Table 4. The MMtFA family with the BIC chose two-component models on 96 of the 100 runs (the CCC model was selected on 96 runs). MCLUST also chose the correct number of groups on 96 runs. However, the MMtFA family with the ICL criterion chose a $G = 1$ group model on 55 of the 100 runs.

In both of these simulation studies, the BIC outperformed the ICL in selecting the number of groups for the MMtFA family. Over all 200 runs, the ICL incorrectly estimated the number of components on 77 occasions and, in fact, underestimated the number of components on 76 of these 77 occasions. In these simulation studies, data were simulated from a three-dimensional Gaussian distribution — simulated data with higher dimensionality ($p = 500$) are analyzed in Sect. 5.1.3.

5.1.3 High-dimensional data

Given the size of many modern data sets, it is important for the MMtFA family of models to be effective in the analysis of high-dimensional data. Of particular interest is the ability of the models to analyze high-dimensional data even when the sample size n is small. A simulation study is presented in this section to demonstrate the ability of the MMtFA family to deal with data sets of this nature. This simulation study also lends further support to the claim that the upper limit of 200 that was imposed on the $\hat{\nu}_g$ does not impair the recovery of an underlying Gaussian distribution.

One-hundred data sets were generated for $G = 2$ groups from a multivariate Gaussian distribution with a factor analysis covariance structure for each component ($\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g$), where $q = 1$, $p = 500$ and $n = 50$. Again, the observations were divided equally amongst the two groups. Since $p > n$, the agglomerative hierarchical initialization procedure from the `mclust` package could not be used. Instead, the $\hat{z}_{ig}$ were initialized using k-means clustering. The initialization of $\boldsymbol{\Lambda}_g$ was difficult because the eigen-decomposition of the 500×500 sample covariance matrix could not be computed reliably.

The MMtFA family was run for $q = 1$ factor and $G = 1, 2, 3$ components. Perfect classification was obtained on all 100 runs, with the CCC model being selected every time. Running `mclust` on these simulated data sets resulted in very poor performance: G was overestimated on the vast majority of runs. However, a comparison with MCLUST is not entirely fair here because the data were simulated from a Gaussian distribution with a factor analysis covari-

Table 3 Summary of the number of groups selected by each method for the X-shaped simulations

	1	2	≥3
MMtFA BIC	0	98	2
MMtFA ICL	21	78	1
MCLUST (BIC)	0	89	11

Table 4 Summary of the number of groups selected by each method for the 8-shaped simulations

	1	2	3
MMtFA BIC	2	96	2
MMtFA ICL	55	45	0
MCLUST (BIC)	1	96	3

Table 5 Summary of the best within-model MM*t*FAs for the wine data: that is, a summary of the best combination of G and q for each of the six models in Table 1 when applied to the wine data

Model	G	q	BIC	ICL	$\hat{\nu}_1$	$\hat{\nu}_2$	$\hat{\nu}_3$	Adj. rand	Iterations
UUU	3	2	−5294.4	−5296.3	117.4	8.4	152.6	0.96	329
UUC	3	2	−5298.8	−5300.6	17.4	17.4	17.4	0.98	365
UCU	3	2	−5504.0	−5505.8	33.5	6.6	27.9	0.93	25
UCC	3	2	−5498.1	−5501.5	10.9	10.9	10.9	0.90	20
CCU	3	4	−5442.0	−5443.3	77.7	7.0	45.7	0.95	31
CCC	4	4	−5444.2	−5446.9	21.2	21.2	21.2	0.84	25

ance structure and, since $n > p$, the hierarchical initialization procedure was not available.

5.2 Real data

In addition to the simulated data analyses of Sect. 5.1, the MM*t*FA family of models was applied to two real data sets (Sects. 5.3 and 5.4). In Sect. 5.3, the models are applied to data on thirteen chemical and physical properties of Italian wines. These data were previously analyzed by McNicholas and Murphy (2008), who demonstrated that `mclust`, `clustvarsel` and the PGMMs all gave quite poor clustering performance (adjusted Rand indices < 0.8). However, Scrucca (2009) used the GMMDR technique to analyze the same data, resulting in an adjusted Rand index of 0.85. In Sect. 5.4, the MM*t*FA family is applied to data on flea beetles. The flea beetles data were chosen because both `mclust` and `clustvarsel` give perfect clustering on these data.

5.3 Italian wine data

5.3.1 The data

Forina et al. (1986) give twenty-eight chemical and physical properties of three types of Italian wine (Barolo, Grignolino and Barbera). A subset of thirteen of these variables (Table 9, Appendix) is available as part of the `gclus` package (Hurley 2004) for R.

5.3.2 The MMtFA family

All six models were fitted to the data for $q = 1, \ldots, 5$ factors and $G = 1, \ldots, 5$ components. The largest BIC resulting from all models was −5294.4 from the fully unconstrained model (UUU). The largest ICL, also from the UUU model, was −5296.3. The second largest BIC and ICL values were from the UUC model (−5298.8 and −5300.6, respectively). Both criteria show a notable decrease in value for the other four models, suggesting that the isotropic constraint may not be appropriate for these data. All but one of the models contained $G = 3$ components in the maximum within-model BIC and ICL. A summary for all six models is given in Table 5.

Table 6 Classification tables for the UUU and UUC models on the wine data

	UUU			UUC		
	1	2	3	1	2	3
Barolo	58	1		58	1	
Grignolino	1	70			71	
Barbera			48			48

Classification tables for the UUU and the UUC models with $G = 3$ and $q = 2$ are given in Table 6: the adjusted Rand indices for these models are 0.96 and 0.98, respectively. The estimates of the degrees of freedom for the UUU model are $\hat{\nu}_1 = 117.4$, $\hat{\nu}_2 = 8.4$, and $\hat{\nu}_3 = 152.6$, whereas $\hat{\nu} = 17.4$ for the constrained model. In the case of the UUU model, the differences in estimated degrees of freedom across components are not trivial: it seems that the multivariate Gaussian distribution may be suitable for both the Barolo and Barbera wines, while the multivariate t-distribution seems more appropriate for the Grignolino wines. Furthermore, it is interesting to note that while the fit of the UUU model may be better than that of the UUC model (greater BIC and yet more free parameters, which reflects a greater log-likelihood), the UUC model with $\hat{\nu} = 17.4$ gave slightly better clustering performance.

5.3.3 MCLUST, variable selection, PGMMs & GMMDR

McNicholas and Murphy (2008) analyzed the same data using PGMMs, MCLUST and variable selection. The best PGMM model had $G = 4$ components and the resulting clustering had an adjusted Rand index of 0.79 (Table 10, Appendix). McNicholas and Murphy (2008) also reported that `mclust` selected a $G = 8$ component model, resulting in an adjusted Rand index of 0.48 (Table 11, Appendix), and `clustvarsel` chose a three-group model, giving an adjusted Rand index of 0.78 (Table 12, Appendix).

Table 7 Model comparison for the wine data

Model	Adjusted rand index
UUC	0.98
UUU	0.96
CCU	0.95
UCU	0.93
UCC	0.90
GMMDR	0.85
CCC	0.84
PGMMs	0.79
Variable Selection	0.78
MCLUST	0.48

Scrucca (2009) also analyzed these data using the GMMDR technique. The GMMDR model that was chosen had $G = 3$ components and an adjusted Rand index of 0.85. A classification table for this model is given in Table 13 (Appendix).

5.3.4 Discussion

When applied to the wine data, both the UUU and UUC models give better clustering results than the techniques mentioned in Sect. 5.3.3. In fact, the other four members of the MMtFA family also clustered with greater accuracy than `mclust`, `clustvarsel` and the PGMM family. Furthermore, all of the MMtFA models, except for the CCC model, give better clustering performance than the GMMDR technique. Adjusted Rand indices for all of the models that were applied to these data are given in Table 7.

The fact that the UUC model gave better clustering performance than the UUU model, despite smaller BIC and ICL values, illustrates the fact that these criteria do not necessarily choose the best classifier. Note also that the four isotropic models (UCU, UCC, CCU and CCC) reach convergence in under 35 iterations, whereas the other two models take over 300 iterations each (see Table 5). Given that an iteration takes a similar amount of time for each model, this represents a vast difference when it comes to computation time. However, this difference in the number of iterations to convergence seems to be data-specific (cf. Sect. 5.4).

5.4 Flea beetles data

The flea beetles data were introduced by Lubischew (1962) and are available within the GGobi software (Swayne et al. 2006). The data contain six measurement variables on 72 fleas and each flea is a member of one of three species: *concinna*, *heptapotamica*, or *heikertingeri*. These variables are described in Table 14 (Appendix). Running `mclust` and `clustvarsel` on these data leads to perfect classifications in each case. As mentioned in Sect. 5.2, these data were chosen because of this fact.

The MMtFA family of models was fitted to these data for $q = 1, 2, 3$ factors and $G = 1, \ldots, 4$ mixture components. The best model according to both the BIC and the ICL was the CCC model with $q = 2$ and $G = 3$. A summary of the best combination of G and q for each model is given in Table 8. Each of the six MMtFA models gave perfect clustering for these data. Note that most of the models with the isotropic constraint took more iterations to convergence for these data than those without this constraint (see Table 8). This suggests that relative computation times for the members of the MMtFA family are indeed data-dependent.

6 Summary & future work

A family of mixture models was developed, based on the MMtFA model, for model-based clustering. Four members of this family are novel (UUC, UCC, CCU, and CCC), while a fifth (UCU) was used for clustering for the first time and was subject to different parameter estimation than was previously employed. Parameter estimation for the MMtFA family was carried out within the AECM algorithm framework, and the BIC and the ICL were used for model selection. This family of six models, which includes models with constrained degrees of freedom and constrained loading matrices, was then applied to real and simulated data. In all of these analyses, excellent clustering performance was achieved when compared with existing model-based clustering techniques.

The fact that the models with constrained degrees of freedom gave better clustering performance than those without this constraint, on more than one occasion, is interesting. One possible explanation of this phenomenon is that the estimate of the degrees of freedom is more reliable in the constrained case since it is based on more data, representing a sort of averaging across groups.

Considering that the PGMM family consists of eight members, there is scope for the development of a family of sixteen MMtFA models by imposing additional constraints on the covariance structure. However, the benefit of such development will be tied to ongoing work on the search for a better model selection technique. With just the six models used in this work, it was demonstrated that the BIC and ICL do not necessarily choose the best classifier. One may therefore be loath to run a sixteen member family unless a more effective model selection procedure were to become available.

While these models are very well suited to the analysis of high-dimensional data, as described herein, such applications will be most effective under parallelization. The message passing interface (MPI) would facilitate coarse-grain parallelization of the MMtFA family in an analogous

fashion to the parallel implementation of the PGMM family described by McNicholas et al. (2010). A hybrid parallelization procedure incorporating OpenMP within this MPI framework might provide even greater speed-up and will be the subject of future work.

Finally, the MMtFA family could be extended to the model-based classification framework wherein some of the observations have known group memberships while others do not. This extension would come about in an analogous fashion to that of the PGMM family, which has recently (McNicholas 2010) been extended to the model-based classification framework.

Acknowledgements The authors wish to thank Professor Gilles Celeux, an anonymous associate editor, and two anonymous referees for their helpful comments on this work. This work was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada, by the Canada Foundation for Innovation—Leaders Opportunity Fund and by the Ontario Research Fund—Research Infrastructure Program.

Appendix: Tables 8–14

Table 8 Summary of the best within-model MM*t*FAs for the flea beetles data: that is, a summary of the best combination of G and q for each of the six models in Table 1 when applied to the flea beetles data

Model	G	q	BIC	ICL	$\hat{\nu}_1$	$\hat{\nu}_2$	$\hat{\nu}_3$	Adj. rand	Iterations
UUU	3	1	-1017.9	-1017.9	200.0	189.4	162.7	1.0	73
UUC	3	1	-1009.2	-1009.2	200.0	–	–	1.0	84
UCU	3	1	-1012.7	-1012.8	119.8	43.2	200.0	1.0	106
UCC	3	1	-1004.3	-1004.4	153.2	–	–	1.0	163
CCU	3	2	-967.5	-967.5	200	59.5	90.4	1.0	52
CCC	3	2	-958.7	-958.8	200	–	–	1.0	143

Table 9 The thirteen chemical and physical properties of the Italian wines available in the `gclus` package

Alcohol	Proline
Malic acid	Ash
Hue	Total phenols
Flavonoids	Nonflavonoid phenols
Color intensity	Alcalinity of ash
OD_{280}/OD_{315} of diluted wines	Proanthocyanins
Magnesium	

Table 10 Classification table for the best PGMM model for the wine data

	1	2	3	4
Barolo	59			
Grignolino		38	31	2
Barbera				48

Table 11 Classification table for the best `mclust` model for the wine data

	1	2	3	4	5	6	7	8
Barolo	40	18	1					
Grignolino			21	22		27	1	
Barbera					4		17	27

Table 12 Classification table for the best `clustvarsel` model for the wine data

	1	2	3
Barolo	51	8	0
Grignolino	3	67	1
Barbera	0	1	47

Table 13 Classification table for the selected GMMDR model for the wine data

	1	2	3
Barolo	57	2	
Grignolino	2	64	5
Barbera			48

Table 14 Measurement variables, with units, for the flea beetles data

Variable	1 Unit
The width of the first joint of the first tarsus.	1 micron
The width of the second joint of the first tarsus.	1 micron
The maximal width of the head between the external edges of the eyes.	0.01 mm
The maximal width of the aedeagus in the fore-part.	1 micron
The front angle of the aedeagus.	7.5 degrees
The aedeagus width from the side.	1 micron

References

Banfield, J.D., Raftery, A.E.: Model-based Gaussian and non-Gaussian clustering. Biometrics **49**(3), 803–821 (1993)

Biernacki, C., Celeux, G., Govaert, G.: Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Trans. Pattern Anal. Mach. Intell. **22**(7), 719–725 (2000)

Binder, D.A.: Bayesian cluster analysis. Biometrika **65**, 31–38 (1978)

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P., Lindsay, B.: The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. Ann. Ins. Stat. Math. **46**, 373–388 (1994)

Brent, R.: Algorithms for Minimization without Derivatives. Prentice Hall, New Jersey (1973)

Celeux, G., Govaert, G.: Gaussian parsimonious clustering models. Pattern Recogn. **28**, 781–793 (1995)

Dasgupta, A., Raftery, A.E.: Detecting features in spatial point processes with clutter via model-based clustering. J. Am. Stat. Assoc. **93**, 294–302 (1998)

Day, N.E.: Estimating the components of a mixture of normal distributions. Biometrika **56**, 463–474 (1969)

Dean, N., Raftery, A.E.: The clustvarsel package: R package version 0.2-4 (2006)

Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B **39**(1), 1–38 (1977)

Forina, M., Armanino, C., Castino, M., Ubigli, M.: Multivariate data analysis as a discriminating method of the origin of wines. Vitis **25**, 189–201 (1986)

Fraley, C., Raftery, A.E.: How many clusters? Which clustering methods? Answers via model-based cluster analysis. Comput. J. **41**(8), 578–588 (1998)

Fraley, C., Raftery, A.E.: Model-based clustering, discriminant analysis, and density estimation. J. Am. Stat. Assoc. **97**(458), 611–631 (2002)

Fraley, C., Raftery, A.E.: Enhanced software for model-based clustering, density estimation, and discriminant analysis: MCLUST. J. Classif. **20**, 263–286 (2003)

Fraley, C., Raftery, A.E.: 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 (2006)

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

Ghahramani, Z., Hinton, G.E.: The EM algorithm for factor analyzers. Tech. Rep. CRG-TR-96-1. University of Toronto, Toronto (1997)

Gormley, I.C., Murphy, T.B.: A mixture of experts model for rank data with applications in election studies. Ann. Appl. Stat. **2**(4), 1452–1477 (2008)

Hubert, L., Arabie, P.: Comparing partitions. J. Classif. **2**, 193–218 (1985)

Hurley, C.: Clustering visualizations of multivariate data. J. Comput. Graph. Stat. **13**(4), 788–806 (2004)

Kass, R.E., Raftery, A.E.: Bayes factors. J. Am. Stat. Assoc. **90**, 773–795 (1995)

Keribin, C.: Consistent estimation of the order of mixture models. Sankhyā Indian J. Stat. Ser. A **62**(1), 49–66 (2000)

Leroux, B.G.: Consistent estimation of a mixing distribution. Ann. Stat. **20**, 1350–1360 (1992)

Lindsay, B.G.: Mixture models: theory, geometry and applications. In: NSF-CBMS Regional Conference Series in Probability and Statistics, vol. 5. Institute of Mathematical Statistics, Hayward (1995)

Lopes, H.F., West, M.: Bayesian model assessment in factor analysis. Stat. Sinica **14**, 41–67 (2004)

Lubischew, A.A.: On the use of discriminant functions in taxonomy. Biometrics **18**(4), 455–477 (1962)

McLachlan, G.J.: The classification and mixture maximum likelihood approaches to cluster analysis. In: Handbook of Statistics, vol. 2, pp. 199–208. North-Holland, Amsterdam (1982)

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

McLachlan, G.J., Krishnan, T.: The EM Algorithm and Extensions, 2nd edn. Wiley, New York (2008)

McLachlan, G.J., Peel, D.: Robust cluster analysis via mixtures of multivariate t-distributions. In: Lecture Notes in Computer Science, vol. 1451, pp. 658–666. Springer, Berlin (1998)

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

McLachlan, G.J., Peel, D.: Mixtures of factor analyzers. In: Proceedings of the Seventh International Conference on Machine Learning, pp. 599–606. Morgan Kaufmann, San Francisco (2000b)

McLachlan, G.J., Bean, R.W., Jones, L.B.T.: Extension of the mixture of factor analyzers model to incorporate the multivariate t-distribution. Comput. Stat. Data Anal. **51**(11), 5327–5338 (2007)

McNicholas, P.D.: Model-based classification using latent Gaussian mixture models. J. Stat. Plan. Inference **140**(5), 1175–1181 (2010)

McNicholas, P.D., Murphy, T.B.: Parsimonious Gaussian mixture models. Tech. Rep. 05/11, Department of Statistics, Trinity College Dublin (2005)

McNicholas, P.D., Murphy, T.B.: Parsimonious Gaussian mixture models. Stat. Comput. **18**, 285–296 (2008)

McNicholas, P.D., Murphy, T.B.: Model-based clustering of longitudinal data. Can. J. Stat. **38**(1), 153–168 (2010)

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

Meng, X.L., van Dyk: The EM algorithm—an old folk song sung to a fast new tune (with discussion). J. R. Stat. Soc. Ser. B **59**, 511–567 (1997)

Meng, X.L., Rubin, D.B.: Maximum likelihood estimation via the ECM algorithm: a general framework. Biometrika **80**, 267–278 (1993)

R Development Core Team (2009) R: a language and environment for statistical computing: R foundation for statistical computing, Vienna, Austria. http://www.R-project.org

Raftery, A.E., Dean, N.: Variable selection for model-based clustering. J. Am. Stat. Assoc. **101**(473), 168–178 (2006)

Rand, W.M.: Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. **66**, 846–850 (1971)

Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**, 31–38 (1978)

Scrucca, L.: Dimension reduction for model-based clustering. Stat. Comput. (2009, in press). doi:10.1007/s11222-009-9138-7

Shoham, S.: Robust clustering by deterministic agglomeration em of mixtures of multivariate t-distributions. Pattern Recogn. **35**(5), 1127–1142 (2002)

Swayne, D., Cook, D., Buja, A., Lang, D., Wickham, H., Lawrence, M.: (2006) GGobi Manual. Sourced from www.ggobi.org/docs/manual.pdf

Tipping, T.E., Bishop, C.M.: Mixtures of probabilistic principal component analysers. Neural Comput. **11**(2), 443–482 (1999a)

Tipping, T.E., Bishop, C.M.: Probabilistic principal component analysers. J. R. Stat. Soc. Ser. B **61**, 611–622 (1999b)

Wolfe, J.H.: Object cluster analysis of social areas. Master's thesis, University of California, Berkeley (1963)

Wolfe, J.H.: Pattern clustering by multivariate mixture analysis. Multiv. Behav. Res. **5**, 329–350 (1970)

Woodbury, M.A.: Inverting modified matrices. Statistical Research Group, Memo. Rep. No. 42, Princeton University, Princeton, New Jersey (1950)

Zhao, J., Jiang, Q.: Probabilistic PCA for t distributions. Neurocomputing **69**(16–18), 2217–2226 (2006)