Journal of Classification (2021) 38:89–108
https://doi.org/10.1007/s00357-019-09351-3

A Variational Approximations-DIC Rubric for Parameter Estimation and Mixture Model Selection Within a Family Setting

Sanjeena Subedi[1] · Paul D. McNicholas[2]

Published online: 4 March 2020

Abstract

Mixture model-based clustering has become an increasingly popular data analysis technique since its introduction over fifty years ago, and is now commonly utilized within a family setting. Families of mixture models arise when the component parameters, usually the component covariance (or scale) matrices, are decomposed and a number of constraints are imposed. Within the family setting, model selection involves choosing the member of the family, i.e., the appropriate covariance structure, in addition to the number of mixture components. To date, the Bayesian information criterion (BIC) has proved most effective for model selection, and the expectation-maximization (EM) algorithm is usually used for parameter estimation. In fact, this EM-BIC rubric has virtually monopolized the literature on families of mixture models. Deviating from this rubric, variational Bayes approximations are developed for parameter estimation and the deviance information criteria (DIC) for model selection. The variational Bayes approach provides an alternate framework for parameter estimation by constructing a tight lower bound on the complex marginal likelihood and maximizing this lower bound by minimizing the associated Kullback-Leibler divergence. The framework introduced, which we refer to as VB-DIC, is applied to the most commonly used family of Gaussian mixture models, and real and simulated data are used to compared with the EM-BIC rubric.

Keywords BIC · Clustering · DIC · EM algorithm · GPCM · Mixture models · Model-based clustering · Variational approximations · Variational Bayes · VB-DIC

✉ Sanjeena Subedi
sdang@binghamton.edu

Paul D. McNicholas
mcnicholas@math.mcmaster.ca

[1] Department of Mathematical Sciences, Binghamton University, State University of New York, 4400 Vestal Parkway East, Binghamton, NY, 13902, USA

[2] Department of Mathematics & Statistics, McMaster University, 1280 Main St. W., Hamilton, ON, L8S 4K1, Canada

1 Introduction

Most early clustering algorithms were based on heuristic approaches and some such methods, including hierarchical agglomerative clustering and k-means clustering (MacQueen 1967; Hartigan and Wong 1979), are still widely used. The use of mixture models to account for population heterogeneity has been very well established for over a century (e.g., Pearson 1894), but it was the 1960s before mixture models were used for clustering (Wolfe 1965; Hasselblad 1966; Day 1969). Because of the lack of suitable computing equipment, it was much later before the use of mixture models (e.g., Banfield and Raftery 1993; Celeux and Govaert 1995) and, more generally, probability models (e.g., Bock 1996, 1998a, b) for clustering became commonplace. Since the turn of the century, the use of mixture models for clustering has burgeoned into a popular subfield of cluster analysis and recent examples include Franczak et al. (2014), Vrbik and McNicholas (2014), Murray et al. (2014a, b), Lee and McLachlan (2014), Lin et al. (2014), Subedi et al. (2015), Morris and McNicholas (2016), O'Hagan et al. (2016), Dang et al. (2015), Lin et al. (2016), Lee and McLachlan (2016), Dang et al. (2017), Cheam et al. (2017), Melnykov and Zhu (2018), Zhu and Melnykov (2018), Gallaugher and McNicholas (2019b), Tortora et al. (2019), Biernacki and Lourme (2019), Murray et al. (2019), Morris et al. (2019), and Punzo et al. (2020). The reader may consult Bouveyron and Brunet-Saumard (2014) and McNicholas (2016b) for relatively recent reviews of model-based clustering work.

A d-dimensional random vector $\mathbf{Y}$ is said to arise from a parametric finite mixture distribution if, for all $\mathbf{y} \subset \mathbf{Y}$, we can write its density as

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

where $\rho_g > 0$ such that $\sum_{i=1}^{G} \rho_g = 1$ are the mixing proportions, $p_g(\mathbf{y} \mid \boldsymbol{\theta}_g)$ are component densities, and $\boldsymbol{\vartheta} = (\rho_1, \ldots, \rho_G, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G)$ is the vector of parameters. When the component parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G$ are decomposed and constraints are imposed on the resulting decompositions, the result is a family of mixture models. Typically, each component probability density is of the same type and, when they are Gaussian, the mixture density function is

$$f(\mathbf{y} \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \rho_g \phi_d(\mathbf{y} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g),$$

where $\phi_d(\mathbf{y} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ is the d-dimensional Gaussian density with mean $\boldsymbol{\mu}_g$ and covariance $\boldsymbol{\Sigma}_g$, and the likelihood is

$$\mathcal{L}(\boldsymbol{\vartheta} \mid \mathbf{y}_1, \ldots, \mathbf{y}_n) = \prod_{i=1}^{n} \sum_{g=1}^{G} \rho_g \phi_d(\mathbf{y}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g),$$

where $\boldsymbol{\vartheta}$ denotes the model parameters. In Gaussian families, it is usually the component covariance matrices $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_G$ that are decomposed (see Section 2).

The expectation-maximization (EM) algorithm (Dempster et al. 1977) is often used for mixture model parameter estimation but its efficacy is questionable. As discussed by Titterington et al. (1985) and others, the nature of the mixture likelihood surface leaves the EM algorithm open to failure. Although this weakness can be mitigated by using multiple re-starts, there is no way to completely overcome it. Besides its heavy reliance on starting values, convergence of the EM algorithm can be very slow. When families of mixture models are used, the EM algorithm approach must be employed in conjunction with a model selection criterion to select the member of the family and, in many cases, the number of components. There are many model selection criteria to choose from, such as the Bayesian information criterion (BIC; Schwarz 1978), the integrated completed likelihood (ICL; Biernacki et al. 2000), and the Akaike information criterion (AIC; Akaike 1974). All of these model selection criteria have some merit and various shortcomings, but the BIC remains by far the most popular (McNicholas 2016a, Chp. 2). There has been interest in the use of Bayesian approaches to mixture model parameter estimation, via Markov chain Monte Carlo (MCMC) methods (e.g., Diebolt and Robert 1994; Richardson and Green 1997; Bensmail et al. 1997; Stephens 1997, 2000; Casella et al. 2002); however, difficulties have been encountered with, inter alia, computational overhead and convergence (see Celeux et al. 2000; Jasra et al. 2005). Variational Bayes approximations present an alternative to MCMC algorithms for mixture model parameter estimation and are gaining popularity due to their fast and deterministic nature (see Jordan et al. 1999; Corduneanu and Bishop 2001; Ueda and Ghahramani 2002; McGrory and Titterington 2007, 2009; McGrory et al. 2009; Subedi and McNicholas 2014).

With the use of a computationally convenient approximating density in place of a more complex "true" posterior density, the variational algorithm overcomes the hurdles of MCMC sampling. For observed data $\mathbf{y}$, the joint conditional distribution of parameters $\boldsymbol{\theta}$ and missing data $\mathbf{z}$ is approximated by using another computationally convenient distribution $q(\boldsymbol{\theta}, \mathbf{z})$. This distribution $q(\boldsymbol{\theta}, \mathbf{z})$ is obtained by minimizing the Kullback-Leibler (KL) divergence between the true and the approximating densities, where

$$\mathrm{KL}(q(\boldsymbol{\theta}, \mathbf{z}) \mid p(\boldsymbol{\theta}, \mathbf{z} \mid \mathbf{y})) = \int_{\boldsymbol{\Theta}} \sum_{z} q(\boldsymbol{\theta}, \mathbf{z}) \log \left\{ \frac{q(\boldsymbol{\theta}, \mathbf{z})}{p(\boldsymbol{\theta}, \mathbf{z} \mid \mathbf{y})} \right\} d\boldsymbol{\theta}.$$

The approximating density is restricted to have a factorized form for computational convenience, so that $q(\boldsymbol{\theta}, \mathbf{z}) = q_\theta(\boldsymbol{\theta}) q_z(\mathbf{z})$. Upon choosing a conjugate prior, the appropriate hyper-parameters of the approximating density $q_\theta(\boldsymbol{\theta})$ can be obtained by solving a set of coupled non-linear equations.

The variational Bayes algorithm is initialized with more components than expected. As the algorithm iterates, if two components have similar parameters then one component dominates the other causing the dominated component's weighting to be zero. If a component's weight becomes sufficiently small, less than or equal to two observations in our analyses, the component is removed from consideration. Therefore, the variational Bayes approach allows for simultaneous parameter estimation and selection of the number of components.

2 Methodology

2.1 Introducing Parsimony

If d-dimensional data $\mathbf{y}_1, \ldots, \mathbf{y}_n$ arise from a finite mixture of Gaussian distributions, then the log-likelihood is

$$\log p(\mathbf{y}_1, \ldots, \mathbf{y}_n \mid \boldsymbol{\theta}) = \sum_{i=1}^{n} \log \left[\sum_{g=1}^{G} \rho_g \frac{|\boldsymbol{\Sigma}_g^{-1}|}{2\pi^{\frac{d}{2}}} \exp\left\{ \frac{1}{2} (\mathbf{y}_i - \boldsymbol{\mu}_g)' \boldsymbol{\Sigma}_g^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_g) \right\} \right].$$

The number of parameters in the component covariance matrices of is $Gd(d+1)/2$, which is quadratic in d. When dealing with real data, the number of free parameters to be estimated can very easily exceed the sample size n by an order of magnitude. Hence, the introduction of parsimony through the imposition of additional structure on the covariance matrices is desirable. Banfield and Raftery (1993) exploited geometrical constraints on the covariance matrices of a mixture of Gaussian distributions using the eigen-decomposition of the covariance matrices, such that $\mathbf{\Sigma}_g = \lambda_g \mathbf{D}_g \mathbf{A}_g \mathbf{D}'_g$, where $\mathbf{D}_g$ is the orthogonal matrix of eigenvectors and $\mathbf{A}_g$ is a diagonal matrix proportional to the eigenvalues of $\mathbf{\Sigma}_g$, such that $|\mathbf{A}_g| = 1$, and λ_g is the associated constant of proportionality. This decomposition has an advantage in terms of its interpretation, i.e., the parameter λ_g controls the cluster volume, $\mathbf{A}_g$ controls the cluster shape, and $\mathbf{D}_g$ controls the cluster orientation. This allows for imposition of several constraints on the covariance matrix that have geometrical interpretation giving rise to a family of 14 models known as Gaussian Parsimonious clustering models (GPCM; Celeux and Govaert 1995) (see Table 1).

The `mclust` package (Scrucca et al. 2016) for R (R Core Team 2018) implements 12 of these 14 GPCM models in an EM framework, with the MM framework of Browne and McNicholas (2014) used for the other two models (EVE and VVE). Bensmail et al. (1997) used Gibbs sampling to carry out Bayesian inference for eight of the GPCM models. Bayesian regularization of some of the GPCM models has been considered by Fraley and Raftery (2007). After assigning a highly dispersed conjugate prior, they replace the maximum likelihood estimator of the group membership obtained using the EM algorithm by a maximum a posteriori probability (MAP) estimator. Note that $\text{MAP}(\hat{z}_{ig}) = 1$ if $g = \arg\max_h(\hat{z}_{ih})$ and $\text{MAP}(\hat{z}_{ig}) = 0$ otherwise, where $\hat{z}_{ig}$ denotes the a posteriori expected value of Z_{ig} and

$$z_{ig} = \begin{cases} 1 & \text{if } \mathbf{x}_i \text{ belongs to component } g, \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

A modified BIC using the maximum a posteriori probability is then used for model selection. Herein, we implement 12 of those 14 GPCM models using variational Bayes approximations—conjugate priors are not available for the EVE and VVE models.

Table 1 Nomenclature, interpretation, and covariance structure for each member of the GPCM family

Model	Volume	Shape	Orientation	Σ_g
EII	Equal	Spherical		$\lambda \boldsymbol{I}$
VII	Variable	Spherical		$\lambda_g \boldsymbol{I}$
EEI	Equal	Equal	Axis-aligned	$\lambda \mathbf{A}$
VEI	Variable	Equal	Axis-aligned	$\lambda_g \mathbf{A}$
EVI	Equal	Variable	Axis-aligned	$\lambda \mathbf{A}_g$
VVI	Variable	Variable	Axis-aligned	$\lambda_g \mathbf{A}_g$
EEE	Equal	Equal	Equal	$\lambda \mathbf{D}\mathbf{A}\mathbf{D}'$
VEE	Variable	Equal	Equal	$\lambda_g \mathbf{D}\mathbf{A}\mathbf{D}'$
EVE	Equal	Variable	Equal	$\lambda \mathbf{D}\mathbf{A}_g\mathbf{D}'$
EEV	Equal	Equal	Variable	$\lambda \mathbf{D}_g\mathbf{A}\mathbf{D}'_g$
VVE	Variable	Variable	Equal	$\lambda_g \mathbf{D}\mathbf{A}_g\mathbf{D}'$
VEV	Variable	Equal	Variable	$\lambda_g \mathbf{D}_g\mathbf{A}\mathbf{D}'_g$
EVV	Equal	Variable	Variable	$\lambda \mathbf{D}_g\mathbf{A}_g\mathbf{D}'_g$
VVV	Variable	Variable	Variable	$\lambda_g \mathbf{D}_g\mathbf{A}_g\mathbf{D}'_g$

2.2 Priors and Approximating Densities

As suggested by McGrory and Titterington (2007), the Dirichlet distribution is used as the conjugate prior for the mixing proportion, such that

$$p(\boldsymbol{\rho}) = \text{Dir}(\boldsymbol{\rho}; \alpha_1^{(0)}, \ldots, \alpha_G^{(0)}),$$

where $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_G)$ are the mixing proportions and $\alpha_1^{(0)}, \ldots, \alpha_G^{(0)}$ are the hyperparameters. Conditional on the precision matrix $\mathbf{T}_g$, independent normal distributions were used as the conjugate priors for the means such that

$$p(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G \mid \mathbf{T}_1, \ldots, \mathbf{T}_G) = \prod_{g=1}^{G} \phi_d(\boldsymbol{\mu}_g; \mathbf{m}_g^{(0)}, (\beta_g^{(0)} \mathbf{T}_g)^{-1}),$$

where $\{\mathbf{m}_g^{(0)}, \beta_g^{(0)}\}_{g=1}^{G}$ are the hyper-parameters.

Fraley and Raftery (2007) assigned priors on the parameters for the covariance matrix and its components in a Bayesian regularization application. However, we assign priors on the precision matrix with the hyperparameters given in Table 2. Note that it was not possible to put a suitable (i.e., determinant one) prior on the matrix $\mathbf{A}_g$ for the models EVI and VVI or on $\mathbf{A}$ for models VEV and VEI; accordingly, we instead put a prior on $c_g\mathbf{A}_g^{-1}$ or $c\mathbf{A}^{-1}$, respectively, where c_g or c is a positive constant. Using the expected value of $c_g\mathbf{A}_g^{-1}$ (or $c\mathbf{A}^{-1}$), the expected value of $\mathbf{A}_g^{-1}$ (or $\mathbf{A}^{-1}$) is determined to satisfy the constraint that

Table 2 The precision parameter upon which a prior is placed, as well as the corresponding prior distribution and hyperparameters, for 12 of the 14 members of the GPCM family

Model	Σ_g	Precision parameter for prior	Prior and hyperparameters
EII	$\lambda\mathbf{I}_d$	λ^{-1}	Gamma $(a^{(0)}, b^{(0)})$
VII	$\lambda_g\mathbf{I}_d$	λ_g^{-1}	Gamma $(a_g^{(0)}, b_g^{(0)})$
EEI	$\lambda\mathbf{A}$	kth diagonal element of $(\lambda\mathbf{A})^{-1}$	Gamma $(a_k^{(0)}, b_k^{(0)})$
VEI	$\lambda_g\mathbf{A}$	λ_g^{-1}	Gamma $(a_g^{(0)}, b_g^{(0)})$
		kth diagonal elements of $c\mathbf{A}^{-1}$	Gamma $(al_k^{(0)}, be_k^{(0)})$
EVI	$\lambda\mathbf{A}_g$	λ^{-1}	Gamma $(a^{(0)}, b^{(0)})$
		kth diagonal elements of $c_g\mathbf{A}_g^{-1}$	Gamma $(al_{gk}^{(0)}, be_{gk}^{(0)})$
VVI	$\lambda_g\mathbf{A}_g$	λ_g^{-1}	Gamma $(a_g^{(0)}, b_g^{(0)})$
		kth diagonal elements of $c_g\mathbf{A}_g^{-1}$	Gamma $(al_{gk}^{(0)}, be_{gk}^{(0)})$
EEE	$\lambda\mathbf{DAD}'$	$\mathbf{T} = (\lambda\mathbf{DAD}')^{-1}$	Wishart $(\upsilon^{(0)}, \boldsymbol{\Sigma}^{(0)-1})$
VEE	$\lambda_g\mathbf{DAD}'$	λ_g^{-1}	Gamma $(a_g^{(0)}, b_g^{(0)})$
		$\mathbf{T} = (\mathbf{DAD}')^{-1}$	Wishart $(\upsilon^{(0)}, \boldsymbol{\Sigma}^{(0)})$
EEV	$\lambda\mathbf{D}_g\mathbf{AD}_g'$	kth diagonal elements of $(\lambda\mathbf{A})^{-1}$	Gamma $(a_k^{(0)}, b_k^{(0)})$
		$\mathbf{D}_g$	Bingham matrix $(\mathbf{A}_g^{(0)}, \mathbf{B}_g^{(0)})$
VEV	$\lambda_g\mathbf{D}_g\mathbf{AD}_g'$	λ_g^{-1}	Gamma $(a_g^{(0)}, b_g^{(0)})$
		kth diagonal element of $c\mathbf{A}^{-1}$	Gamma $(al_k^{(0)}, be_k^{(0)})$
		$\mathbf{D}_g$	Bingham matrix $(\mathbf{A}_g^{(0)}, \mathbf{B}_g^{(0)})$
EVV	$\lambda\mathbf{D}_g\mathbf{A}_g\mathbf{D}_g'$	λ^{-1}	Gamma $(a^{(0)}, b^{(0)})$
		$\mathbf{T}_g = (\mathbf{D}_g\mathbf{A}_g\mathbf{D}_g')^{-1}$	Wishart $(\upsilon_g^{(0)}, \boldsymbol{\Sigma}_g^{(0)})$
VVV	$\lambda_g\mathbf{D}_g\mathbf{A}_g\mathbf{D}_g'$	$\mathbf{T}_g = (\lambda_g\mathbf{D}_g\mathbf{A}_g\mathbf{D}_g')^{-1}$	Wishart $(\upsilon_g^{(0)}, \boldsymbol{\Sigma}_g^{(0)-1})$

the determinant is 1. Because $\mathbf{D}_g$ is an orthogonal matrix of eigenvectors, the Bingham matrix distribution is used as the conjugate prior for $\mathbf{D}_g$. The Bingham distribution, first introduced by Bingham (1974), is a probability distribution on a set of orthonormal vectors $\{\mathbf{u} : \mathbf{u}'\mathbf{u} = 1\}$ and has antipodal symmetry thus making it ideal for random axes.

The Bingham matrix distribution (Gupta and Nagar 2000) is the matrix analogue, on the Steifel manifold, of the Bingham distribution and has been used in multivariate analysis and matrix decomposition methods (Hoff 2009). The density of the Bingham matrix distribution, as defined by Gupta and Nagar (2000), is

$$p(\mathbf{D}) = b(\mathbf{A}, \mathbf{B}) \exp(\mathrm{tr}\{\mathbf{BDAD}'\})[d\mathbf{D}],$$

for $\mathbf{D} \in O(n, d)$, where $O(n, d)$ is the Stiefel manifold of $n \times d$ matrices, $[d\mathbf{D}]$ is the unit invariant measure on $O(n, d)$, and $\mathbf{A}$ and $\mathbf{B}$ are symmetric and diagonal matrices, respectively. Samples from the Bingham matrix distribution can be obtained using the Gibbs sampling algorithm implemented in the R package `rstiefel` (Hoff 2012).

The approximating densities that minimize the KL divergence are as follows. For the mixing proportions, $q_\rho(\boldsymbol{\rho}) = \mathrm{Dir}(\boldsymbol{\rho}; \alpha_1, \ldots, \alpha_G)$, where $\alpha_g = \alpha_g^{(0)} + \sum_{i=1}^n \hat{z}_{ig}$. For the mean,

$$q_{\boldsymbol{\mu}}(\boldsymbol{\mu} \mid \mathbf{T}_1, \ldots, \mathbf{T}_G) = \prod_{g=1}^G \phi_d(\boldsymbol{\mu}_g; \mathbf{m}_g, (\beta_g \mathbf{T}_g)^{-1}),$$

where $\beta_g = \beta_g^{(0)} + \sum_{i=1}^n \hat{z}_{ig}$ and

$$\mathbf{m}_g = \frac{1}{\beta_g}\left(\beta_g^{(0)}\mathbf{m}_g^{(0)} + \sum_{i=1}^n \hat{z}_{ig}\mathbf{y}_i\right).$$

The probability that the ith observation belongs to a group g is then given by

$$\hat{z}_{ig} = \frac{\varphi_{ig}}{\sum_{j=1}^G \varphi_{ij}},$$

where

$$\begin{aligned}\varphi_{ig} &= \frac{1}{\sum_{g=1}^G \varphi_{ij}} \exp\left(\mathbb{E}[\log \rho_g] + \frac{1}{2}\mathbb{E}[\log |\mathbf{T}_g|] - \frac{1}{2}\mathrm{tr}\left\{\mathbb{E}[\mathbf{T}_g](\mathbf{y}_i - \mathbb{E}[\boldsymbol{\mu}_g])\right.\right.\\ &\qquad \left.\left.\times(\mathbf{y}_i - \mathbb{E}[\boldsymbol{\mu}_g])' + \frac{1}{\beta_g}\mathbf{I}_d\right\}\right),\\ \mathbb{E}[\log(\rho_g)] &= \Psi(\hat{\alpha}_g) - \Psi\left(\sum_{g=1}^G \hat{\alpha}_g\right),\end{aligned}$$

$\mathbb{E}[\boldsymbol{\mu}_g] = \mathbf{m}_g$, and $\Psi(\cdot)$ is the digamma function. The values of $\mathbb{E}[\mathbf{T}_g]$ and $\mathbb{E}[\log |\mathbf{T}_g|]$ vary depending on the model (see Table 6, Appendix A for details). The posterior distribution of the parameters λ_g^{-1} and $\mathbf{A}_g$ are gamma distributions and, therefore, the expected value of $\mathbb{E}[\lambda_g^{-1}]$, $\mathbb{E}[\log |\lambda_g^{-1}|]$, $\mathbb{E}[\mathbf{A}_g]$, and $\mathbb{E}[\log |\mathbf{A}_g|]$ all have a closed form. The posterior distribution for $\mathbf{D}_g\mathbf{A}_g\mathbf{D}_g'$ is a Wishart distribution and so there is a closed form solution for $\mathbb{E}[\mathbf{D}_g\mathbf{A}_g\mathbf{D}_g']$ and $\mathbb{E}[\log |\mathbf{D}_g\mathbf{A}_g\mathbf{D}_g'|]$. The posterior distribution of the parameter $\mathbf{D}_g$ is a Bingham matrix distribution (see Appendix C for details) and, hence, Monte Carlo integration was used to find the expected values of $\mathbb{E}[\mathbf{T}_g]$ and $\mathbb{E}[\log |\mathbf{T}_g|]$. The estimated model parameters maximize the lower bound of the marginal log-likelihood.

2.3 Convergence

The posterior log-likelihood of the observed data obtained using the posterior expected values of the parameters is

$$\log p(\mathbf{y}_1, \ldots, \mathbf{y}_n \mid \tilde{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \log \left[\sum_{g=1}^{G} \frac{\tilde{\rho}_g |\tilde{\mathbf{T}}_g|}{2\pi^{d/2}} \exp \left\{ \frac{1}{2} (\mathbf{y}_i - \tilde{\boldsymbol{\mu}}_g)' \tilde{\mathbf{T}}_g (\mathbf{y}_i - \tilde{\boldsymbol{\mu}}_g) \right\} \right],$$

where $\tilde{\boldsymbol{\mu}}_g = \mathbf{m}_g$ and

$$\tilde{\rho}_g = \frac{\alpha_g}{\sum_{j=1}^{G} \alpha_j}.$$

The expected precision matrix $\tilde{\mathbf{T}}_g$ varies according to the model. Convergence of the algorithm for these models is determined using a modified Aitken acceleration criterion. The Aitken acceleration (Aitken 1926) is given by

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

where $l^{(m)}$ is the value of the posterior log-likelihood at iteration m. Convergence can be considered to have been achieved when

$$\left| l_{\infty}^{(m+1)} - l_{\infty}^{(m)} \right| < \epsilon,$$

where $l_{\infty}^{(m+1)}$ is an asymptotic estimate of the log-likelihood given by

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

(Böhning et al. 1994).

The VEV and EEV models utilize Gibbs sampling and Monte Carlo integration to find both the expected value of the parameter $\mathbf{T}_g$ and the expectations of functions of $\mathbf{T}_g$. As the Gibbs sampling chain approaches the stationary posterior distribution, the posterior log-likelihood oscillates rather than monotonically increasing at every new iteration. Hence, an alternate convergence criteria was used for these models. When the relative change in the parameter estimates from successive iterations is small, convergence is assumed. Hence, for the VEV and EEV models, the algorithm is stopped when

$$\max_i \left\{ \frac{|\psi_i^{(m+1)} - \psi_i^{(m)}|}{|\psi_i^{(m)}| + \delta_1} \right\} < \delta_2, \tag{1}$$

where δ_1 and δ_2 are predetermined constants, $\psi_i^{(m)}$ is the estimate of the ith parameter on the mth iteration, and i indexes over every parameter in the model—note that, for matrix- or vector-valued parameters, $\psi_i^{(m)}$ corresponds to an individual element so that i indexes over all parameter elements and the comparison in (1) is element-wise. In the analyses herein, we use $\delta_1 = 0.001$ and $\delta_2 = 0.05$ for three consecutive iterations. A detailed discussion of the convergence of Monte Carlo EM algorithm is provided in Neath et al. (2013).

2.4 Model Selection

Despite the benefits of simultaneously obtaining parameter estimates along with the number of components, a model selection criterion is needed to determine the covariance structure. For the selection of the model with the best fit, the deviance information criterion (DIC; Spiegelhalter et al. 2002) is used as suggested by McGrory and Titterington (2007). The DIC is given by

$$\text{DIC} = -2 \log p(\mathbf{y}_1, \ldots, \mathbf{y}_n \mid \tilde{\boldsymbol{\theta}}) + 2p_D,$$

where

$$2p_D \approx -2\int q_\theta(\boldsymbol{\theta})\log\left\{\frac{q_\theta(\boldsymbol{\theta})}{p(\boldsymbol{\theta})}\right\}d\boldsymbol{\theta} + 2\log\left\{\frac{q_\theta(\tilde{\boldsymbol{\theta}})}{p(\tilde{\boldsymbol{\theta}})}\right\}$$

and $\log p(\mathbf{y}_1,\ldots,\mathbf{y}_n \mid \tilde{\boldsymbol{\theta}})$ is the posterior log-likelihood of the data.

Hereafter, the variational Bayes approach that uses the variational Bayes algorithms introduced herein together with the DIC to select the model (i.e., covariance structure) will be referred to as the VB-DIC approach.

2.5 Performance Assessment

The adjusted Rand index (ARI; Hubert and Arabie 1985) is used to assess the performance of the clustering techniques applied in Section 3. The Rand index (Rand 1971) is based on the pairwise agreement between two partitions, e.g., predicted and true classifications. The ARI corrects the Rand index to account for agreement by chance: a value of 1 indicates perfect agreement, the expected value under random class assignment is 0, and negative values indicate a classification that is worse than would be expected by guessing.

3 Results

3.1 Simulation Study 1

The VB-DIC approach is run on 50 simulated two-dimensional Gaussian data sets with three components and known mean and covariance structures $\boldsymbol{\Sigma}_g = \lambda_g \mathbf{I}_d$ (VII, see Table 3 for λ_g values). For each dataset, we use five random starts for each of 12 members of the GPCM family, and we set the maximum number of components to ten each time. For each dataset, the model with the smallest DIC is selected as the final model. A $G = 3$ component model is selected on 46 out of 50 occasions with an ARI of 1 each time while a $G = 4$ component model was selected, with an average ARI of 0.96 and a standard deviation of 0.044, on the other four occasions.

A VII model is selected on 47 out of 50 occasions, with VEE and VVV models selected twice and once, respectively. When VEE and VVV are selected, the average difference between the DIC values of the model selected and VII is 1.553 with a standard deviation of 1.984 (the range of the difference is 0.470–3.049). This shows that, although the model selected was different than VII in four cases, the chosen model has similar DIC value to the VII model in each case. In all, there are 43 cases where a $G = 3$ component VII model is selected and the true and average estimated values (with standard deviations) for $\boldsymbol{\mu}_g$ and λ_g in these cases are given in Table 3—in all cases, the estimates are very close to the true values.

Table 3 Summary of the average and standard errors of the estimated parameters from the cases where a $G = 3$ component VII model is selected in Simulation Study 1

			$\hat{\boldsymbol{\mu}}_g$			$\hat{\lambda}_g$	
g	n_g	$\boldsymbol{\mu}_g$	Mean	Standard Error	λ_g	Mean	Standard Error
1	250	$(-7,-7)'$	$(-6.985,-6.990)'$	$(0.081, 0.097)'$	2.2	2.199	0.042
2	100	$(-2,2)'$	$(-1.999, 1.985)'$	$(0.073, 0.066)'$	0.5	0.489	0.101
3	150	$(8,0)'$	$(7.980, 0.031)'$	$(0.090, 0.090)'$	1.2	1.220	0.130

One advantage of using a variational Bayes approach is that, at every iteration, the hyperparameters of the variational posterior are updated to further minimize the Kullback-Leibler divergence between the approximate variational posterior density and the true posterior density. Hence, 95% credible intervals can be created using the variational posterior distribution for all the component means $\boldsymbol{\mu}_g$ and the component precision parameter $1/\sigma_g^2$ for each run (see Fig. 1). A Bayesian credible interval provides an interval within which the unobserved parameter value falls with a certain probability. Similar to Wang et al. (2005), we also evaluated the frequentist coverage probability of the intervals, i.e., the number of times the true value of the parameter is contained within the credible interval. Across the nine parameters, the mean coverage probability was 0.927 (range 0.860–0.976), which is slightly lower than 0.95. Blei et al. (2017) point out that variational inference tends to underestimate the variance of the posterior density.

For completeness, the EM algorithm together with the BIC to select the model (i.e., covariance structure and G)—referred to as the EM-BIC framework hereafter—was also applied to these data using the `mclust` package for R. In all 50 cases, a $G = 3$ component VII model is chosen and gives perfect classification results for all 50 datasets.

3.2 Simulation Study 2

We ran another simulation study with 50 different three-component, three-dimensional Gaussian distributions with known mean and covariance structure $\boldsymbol{\Sigma}_g = \boldsymbol{\Sigma} = \lambda\mathbf{DAD}'$. Again, five different runs with different random starts are used and the maximum number of components is set to ten. In 41 out of 50 datasets, a three-component model is selected by the VB-DIC approach. Out of these 41 cases, an EEE model is selected 39 times and an EEV model is selected twice. These 41 cases give an average ARI of 1.000 (sd 0.001). When an EEV model is selected, the difference in

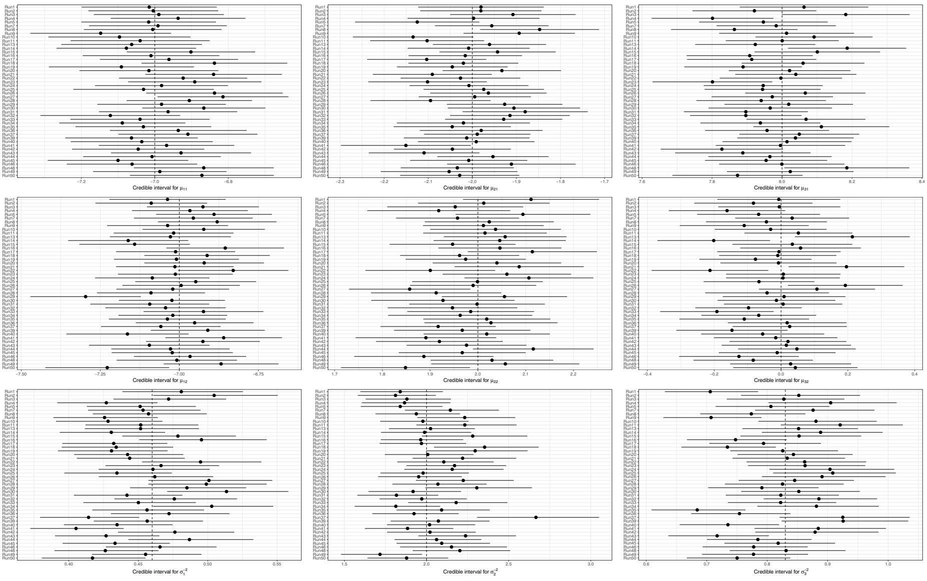

Fig. 1 95% credible intervals for the component means $\boldsymbol{\mu}_g$ (top two rows) and the component precision parameter $1/\sigma^2$ (bottom row) for the 43 runs where a $G = 3$ component VII model is selected in Simulation Study 1, where vertical lines denote the values of the parameters used to generate the data

DIC between the EEV and EEE models is 3.256 and 8.095, respectively, indicating that these two models were close in their fits. Four- and five-component models were selected for 8 and 1 of the datasets, respectively, with an average ARI of 0.923 (sd 0.097). The true and estimated mean parameters using VB-DIC for the EEE model are given in Table 4, and the true and estimated covariance parameters using VB-DIC for the EEE model are:

$$\boldsymbol{\Sigma} = \begin{bmatrix} 0.50 & 0.35 & 0.25 \\ 0.35 & 1.00 & 0.45 \\ 0.25 & 0.45 & 1.20 \end{bmatrix}, \quad \hat{\boldsymbol{\Sigma}} = \begin{bmatrix} 0.494 \text{ (sd } 0.049) & 0.346 \text{ (sd } 0.044) & 0.235 \text{ (sd } 0.046) \\ 0.346 \text{ (sd } 0.044) & 0.995 \text{ (sd } 0.076) & 0.445 \text{ (sd } 0.069) \\ 0.235 \text{ (sd } 0.046) & 0.445 \text{ (sd } 0.069) & 1.204 \text{ (sd } 0.099) \end{bmatrix}.$$

The EM-BIC framework, via `mclust`, was also used for these data. An EEE model was chosen for all 50 datasets with an average ARI of 1.0 (sd 0.001).

3.3 Clustering of Benchmark Datasets

To demonstrate the performance of the VB-DIC approach, we applied our algorithm on several benchmark datasets and compared its performance with the widely used EM-BIC framework via the `mclust` package.

Crabs Data The *Leptograpsus* crab data set, publicly available in the package `MASS` (Venables and Ripley 2002) for R, consists of biological measurements on 100 crabs from two different species (orange and blue) with 50 males and 50 females of each species. The biological measurements (in millimeters) include frontal lobe size, rear width, carapace length, carapace width, and body depth. Although this data set has been analyzed quite often in the literature, using several different clustering approaches, the correlation among the variables makes it difficult to cluster (Fig. 2). Due to this known issue with the data set, we perform an initial step of processing using principal component analysis to convert these correlated variables into principal components (Fig. 2). Finally, the VB-DIC approach was run on these uncorrelated principal components with a maximum of $G = 10$ components.

SRBCT Data The `SRBCT` dataset, available in the R package `plsgenomics` (Boulesteix et al. 2018), is a gene expression data from the microarray experiments of small round blue cell tumors (SRBCT) of childhood cancer. It contains measurements on 2,308 genes from 83 samples comprising of 29 cases of Ewing sarcoma (EWS), 11 cases of Burkitt lymphoma (BL), 18 cases of neuroblastoma (NB), and 25 cases of rhabdomyosarcoma (RMS). Note that our proposed variational Bayes algorithm is not designed for high-dimensional, low sample size (i.e., large p, small N) problems. Dang et al. (2015) performed a differential expression analysis on the gene expression data using an ANOVA across the known groups and selected the top ten genes, ranked using the obtained p-values, to represent a potential set of measurements that contain information on group identification. Hence, we preprocessed the `SRBCT` dataset in a similar manner to Dang et al. (2015) and implemented the VB-DIC approach with a maximum of $G = 10$ components.

Table 4 Summary of the average and standard errors of the estimated parameters from 39 out of the 50 three-dimensional simulated datasets where an EEE model was selected along the true parameters used to generate the data

			$\hat{\boldsymbol{\mu}}_g$	
g	n_g	$\boldsymbol{\mu}_g$	Mean	Standard Error
1	150	$(-2, -2, -2)'$	$(-2.007, -2.025, -2.002)'$	$(0.077, 0.103, 0.119)'$
2	100	$(4, 0, 0)'$	$(4.002, 0.013, 0.017)'$	$(0.051, 0.068, 0.076)'$
3	75	$(-5, 0, 2)'$	$(-5.005, -0.009, 1.976)'$	$(0.087, 0.117, 0.108)'$

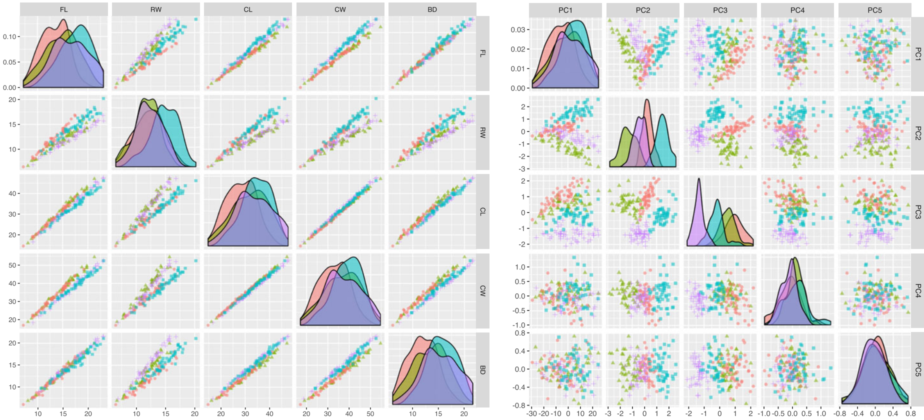

Fig. 2 Scatter plot matrix showing the relationships among the variables in the Leptograpsus crabs dataset (left) and showing the relationships among the uncorrelated principal components (right), where the colors/symbols represent the different groups

Iris Data The `Iris` data set available in the R datasets package contains measurements in centimeters of the variables sepal length, sepal width, petal length, and petal width of 50 flowers from each of the three species of Iris: *Iris setosa*, *Iris versicolor*, and *Iris virginica*.

Diabetes Data The `diabetes` dataset available in the R package `mclust` contains measurements on three variables on 145 non-obese adult patients classified into three groups (Normal, Overt and Chemical):

- `glucose`: Area under plasma glucose curve after a three hour oral glucose tolerance test.
- `insulin`: Area under plasma insulin curve after a three hour oral glucose tolerance test.
- `sspg`: Steady state plasma glucose.

Banknote Data The `banknote` dataset, available in the R package `mclust`, contains six measurements of 100 genuine and 100 counterfeit old Swiss 1000-franc bank notes. Measurements are available for the following variables:

- `Length`: Length of the bill in mm.
- `Left`: Width of left edge in mm.
- `Right`: Width of right edge in mm.
- `Bottom`: Bottom margin width in mm.
- `Top`: Top margin width in mm.
- `Diagonal`: Length of diagonal in mm.

A summary of the performance of the VB-DIC approach and the EM-BIC approach is given in Table 5, where the ARI of the approach that gives the best performance is in italics. For three out of five benchmark datasets, our VB-DIC approach outperforms the EM-BIC framework as implemented via `mclust`. For one of the five datasets, our VB-DIC approach gives the same ARI as the EM-BC framework and, in the other one of the five datasets, the EM-BIC framework yields a slightly larger ARI compared to our VB-DIC approach.

Table 5 Summary of the performance of VB-DIC approach on the benchmark datasets along with the performance using the EM-BIC framework

		VB-DIC		EM-BIC	
Dataset	Classes	G	ARI	G	ARI
Banknote	2	3	*0.842*	3	*0.842*
Crabs	4	5	*0.788*	6	0.600
Diabetes	3	4	0.645	3	*0.664*
Iris	3	3	*0.732*	2	0.568
SRBCT	4	4	*0.760*	4	0.736

4 Discussion

A variational Bayes approach for parameter estimation for the well-known GPCM family has been proposed. As stated before, an advantage of using a variational Bayes algorithm is that, because the hyperparameters of the approximating posterior densities are updated at every iteration, we are indeed updating the approximating variational posterior density of a parameter as opposed to the point estimate of a parameter as in an EM framework. This essentially leads to a natural framework to extract interval estimates (i.e., credible intervals) for every run similar to a fully Bayesian approach but without the need to create a confidence interval via bootstrapping like in an EM framework. We also preserve the monotonicity property of the log-likelihood function, like an EM algorithm, which is lost in a fully Bayesian MCMC based-approach. Additionally, the variational Bayes approach allows for simultaneously obtaining parameter estimates and the number of components. However, a model selection criterion still needs to be utilized while selecting the covariance structure. Herein, we used the DIC for the selection of the covariance structure and so the resulting variational Bayes approach was called the VB-DIC approach. As can be seen from the simulation studies, the correct covariance structure is often selected using the DIC. That said, it may well be the case that another criterion is more suitable for selecting the model (i.e., the covariance structure). Notably, starting values play a different role for variational Bayes than for the EM algorithm—because the former gradually reduces G as the algorithm iterates, the "starting values" for all but the initial G are not the values used to actually start the algorithm. Accordingly, direct comparison of the VB-DIC and EM-BIC approaches is not entirely straightforward.

In the simulation studies, the parameters estimated using variational Bayes approximations were very close to the true parameters (when the correct model was chosen), and excellent classification was obtained using the model selected by DIC. In many of the simulated and real analyses, the performance of the VB-DIC approach was very similar or the same as the EM-BIC approach. This is not surprising. As noted by McLachlan and Krishnan (2008) and Gelman et al. (2013), the EM algorithm can be thought of as a special case of variational Bayes in which the parameters are partitioned into two parts, ϕ and γ, and the approximating distribution of ϕ is required to be a point mass and the approximating distribution of γ is unconstrained conditional on the last update of ϕ. Across all the simulations, EM-BIC framework outperformed the VB-DIC approach; however, the VB-DIC approach outperformed the EM-BIC framework on three of the five benchmark real datasets considered.

In summary, we have explored a Bayesian alternative for parameter estimation for the most widely used family of Gaussian mixture models, i.e., the GPCM family. The use of variational Bayes in conjunction with the DIC for a family of mixture models is a novel idea and lends itself nicely to further research. Moreover, the DIC provides an alternative model selection criterion to the almost ubiquitous BIC. There are several possible avenues for further research, one of which is extension to the semi-supervised (e.g., McNicholas 2010) or, more generally, fractionally supervised paradigm (see Vrbik and McNicholas 2015; Gallaugher and McNicholas 2019b). Another avenue is extension to other families of Gaussian mixture models (e.g., the PGMM family of McNicholas and Murphy 2008, 2010) and to non-Gaussian families of mixture models (e.g., Vrbik and McNicholas 2014; Lin et al. 2014). Further consideration is needed vis-à-vis the approach used to selected the model (i.e., covariance structure) in the variational Bayes approach, e.g., one could conduct a detailed comparison of VB-DIC and, inter alia, VB-BIC. Finally, an analogous variational Bayes approach could be taken to parameter estimation for mixtures of matrix variate distributions (see, e.g., Viroli 2011; Gallaugher and McNicholas 2018a, b, 2019a).

Funding Information This work was supported by a Postgraduate Scholarship from the Natural Science and Engineering Research Council of Canada (Subedi); the Canada Research Chairs program (McNicholas); and an E.W.R. Steacie Memorial Fellowship (McNicholas).

Appendix A: Posterior distributions for the parameters of eigen-decomposed covariance matrix

Table 6 Posterior distributions of the precision parameters as well as their corresponding parameters for 12 of the members of the GPCM family

Model	Posterior distributions	Parameters
EII	Gamma (a, b)	$a = a^{(0)} + d\sum_{g=1}^{G}\sum_{i=1}^{n}\hat{z}_{ig} = a^{(0)} + dn$ $b = b^{(0)} + \sum_{g=1}^{G}(\sum_{i=1}^{n}\hat{z}_{ig}\mathbf{y}_i'\mathbf{y}_i + \beta_g^{(0)}\mathbf{m}_g^{(0)T}\mathbf{m}^{(\mathbf{0})} - \beta_g\mathbf{m}_g'\mathbf{m})$
VII	Gamma (a_g, b_g)	$a_g = a_g^{(0)} + d\sum_{i=1}^{n}\hat{z}_{ig}$ $b_g = b_g^{(0)} + \sum_{i=1}^{n}\hat{z}_{ig}\mathbf{y}_i'\mathbf{y}_i + \beta_g^{(0)}\mathbf{m}_g^{(0)T}\mathbf{m}^{(\mathbf{0})} - \beta_g\mathbf{m}_g'\mathbf{m}$
EEI	Gamma (a_k, b_k)	$a_k = a_k^{(0)} + \sum_{g=1}^{G}\sum_{i=1}^{n}\hat{z}_{ig} = a_k^{(0)} + n$ $b_k = b_k^{(0)} + \sum_{g=1}^{G}\sum_{i=1}^{n}(\hat{z}_{ig}y_{ik}^2 + \beta_g^{(0)}m_{gk}^{(0)2} - \beta_g m_{gk}^2)$
VEI	Gamma (a_g, b_g)	$a_g = a_g^{(0)} + d\sum_{i=1}^{n}\hat{z}_{ig}$ $b_g = b_g^{(0)} + \sum_{i=1}^{n}\hat{z}_{ig}\mathbf{y}_i'\mathbf{y}_i + \beta_g^{(0)}\mathbf{m}_g^{(0)T}\mathbf{m}^{(\mathbf{0})} - \beta_g\mathbf{m}_g'\mathbf{m}$
	Gamma (al_k, be_k)	$al_k = al_k^{(0)} + \sum_{g=1}^{G}\sum_{i=1}^{n}\hat{z}_{ig} = al_k^{(0)} + n$ $be_k = be_k^{(0)} + \sum_{g=1}^{G}\sum_{i=1}^{n}(\hat{z}_{ig}y_{ik}^2 + \beta_g^{(0)}m_{gk}^{(0)2} - \beta_g m_{gk}^2)$
EVI	Gamma (a, b)	$a = a^{(0)} + d\sum_{g=1}^{G}\sum_{i=1}^{n}\hat{z}_{ig} = a^{(0)} + dn$ $b = b^{(0)} + \sum_{g=1}^{G}\sum_{i=1}^{n}\hat{z}_{ig}\mathbf{y}_i'\mathbf{y}_i + \beta_g^{(0)}\mathbf{m}_g^{(0)T}\mathbf{m}^{(\mathbf{0})} - \beta_g\mathbf{m}_g'\mathbf{m}$
	Gamma (al_{gk}, be_{gk})	$al_{gk} = al_{gk}^{(0)} + \sum_{i=1}^{n}\hat{z}_{ig}$ $be_{gk} = be_{gk}^{(0)} + \sum_{i=1}^{n}\hat{z}_{ig}y_{ik}^2 + \beta_g^{(0)}m_{gk}^{(0)2} - \beta_g m_{gk}^2$
VVI	Gamma (a_g, b_g)	$a_g = a_g^{(0)} + d\sum_{i=1}^{n}\hat{z}_{ig}$ $b_g = b_g^{(0)} + \sum_{i=1}^{n}\hat{z}_{ig}\mathbf{y}_i'\mathbf{y}_i + \beta_g^{(0)}\mathbf{m}_g^{(0)T}\mathbf{m}^{(\mathbf{0})} - \beta_g\mathbf{m}_g'\mathbf{m}$
	Gamma (al_{gk}, be_{gk})	$al_{gk} = al_{gk}^{(0)} + \sum_{i=1}^{n}\hat{z}_{ig}$ $be_{gk} = be_{gk}^{(0)} + \sum_{i=1}^{n}\hat{z}_{ig}y_{ik}^2 + \beta_g^{(0)}m_{gk}^{(0)2} - \beta_g m_{gk}^2$
EEE	Wishart (v, Σ^{-1})	$v = v^{(0)} + \sum_{g=1}^{G}\sum_{i=1}^{n}\hat{z}_{ig} = v^{(0)} + n$ $\mathbf{\Sigma}^{-1} = \mathbf{\Sigma}^{(0)-1} + \sum_{g=1}^{G}(\sum_{i=1}^{n}\hat{z}_{ig}\mathbf{y}_i'\mathbf{y}_i + \beta_g^{(0)}\mathbf{m}_g^{(0)T}\mathbf{m}^{(\mathbf{0})} - \beta_g\mathbf{m}_g'\mathbf{m})$
VEE	Gamma (a_g, b_g)	$a_g = a_g^{(0)} + d\sum_{i=1}^{n}\hat{z}_{ig}$ $b_g = b_g^{(0)} + \sum_{i=1}^{n}\hat{z}_{ig}\mathbf{y}_i'\mathbf{y}_i + \beta_g^{(0)}\mathbf{m}_g^{(0)T}\mathbf{m}^{(\mathbf{0})} - \beta_g\mathbf{m}_g'\mathbf{m}$
	Wishart (v, Σ)	$v = v^{(0)} + \sum_{g=1}^{G}\sum_{i=1}^{n}\hat{z}_{ig} = v^{(0)} + n$ $\mathbf{\Sigma} = \mathbf{\Sigma}^{(0)} + \sum_{g=1}^{G}(\sum_{i=1}^{n}\hat{z}_{ig}\mathbf{y}_i'\mathbf{y}_i + \beta_g^{(0)}\mathbf{m}_g^{(0)T}\mathbf{m}^{(\mathbf{0})} - \beta_g\mathbf{m}_g'\mathbf{m})$
EEV	Gamma (a_k, b_k)	$a_k = a_k^{(0)} + d\sum_{g=1}^{G}\sum_{i=1}^{n}\hat{z}_{ig} = a_k^{(0)} + dn$ $b_k = b_k^{(0)} + \sum_{g=1}^{G}(\sum_{i=1}^{n}\hat{z}_{ig}y_{ik}^2 + \beta_g^{(0)}m_{gk}^2 - \beta_g m_{gk}^2)$
	Bingham matrix (P_g, Q)	See mathematical details for the EEV model below.

Table 6 (continued)

Model	Posterior distributions	Parameters
VEV	Gamma (a_g, b_g)	$a_g = a_g^{(0)} + d\sum_{i=1}^{n}\hat{z}_{ig}$
		$b_g = b_g^{(0)} + \sum_{i=1}^{n}\hat{z}_{ig}\mathbf{y}_i'\mathbf{y}_i + \beta_g^{(0)}\mathbf{m}_g^{(0)T}\mathbf{m}^{(\mathbf{0})} - \beta_g\mathbf{m}_g'\mathbf{m}$
	Gamma (al_k, be_k)	$al_k = al_k^{(0)} + \sum_{g=1}^{G}\sum_{i=1}^{n}\hat{z}_{ig} = al_k^{(0)} + n$
		$be_k = be_k^{(0)} + \sum_{g=1}^{G}(\sum_{i=1}^{n}\hat{z}_{ig}y_{ik}^2 + \beta_g^{(0)}m_{gk}^{(0)2} - \beta_g m_{gk}^2)$
	Bingham matrix (P_g, Q_g)	See posterior for $\mathbf{D}_g$ in the VEV Model below.
EVV	Gamma (a, b)	$a = a^{(0)} + d\sum_{g=1}^{G}\sum_{i=1}^{n}\hat{z}_{ig} = a^{(0)} + dn$
		$b = b^{(0)} + \sum_{g=1}^{G}(\sum_{i=1}^{n}\hat{z}_{ig}\mathbf{y}_i'\mathbf{y}_i + \beta_g^{(0)}\mathbf{m}_g^{(0)T}\mathbf{m}^{(\mathbf{0})} - \beta_g\mathbf{m}_g'\mathbf{m})$
	Wishart (v_g, Σ_g^{-1})	$v_g = v_g^{(0)} + \sum_{i=1}^{n}\hat{z}_{ig}$
		$\mathbf{\Sigma}_g^{-1} = \mathbf{\Sigma}_g^{(0)-1} + \sum_{i=1}^{n}\hat{z}_{ig}\mathbf{y}_i\mathbf{y}_i' + \beta_g^{(0)}\mathbf{m}_\mathbf{g}^{(\mathbf{0})}\mathbf{m}_g^{(0)T} - \beta_g\mathbf{m}_g\mathbf{m}_g'$
VVV	Wishart (v_g, Σ_g^{-1})	$v_g = v_g^{(0)} + \sum_{i=1}^{n}\hat{z}_{ig}$
		$\mathbf{\Sigma}_g^{-1} = \mathbf{\Sigma}_g^{(0)-1} + \sum_{i=1}^{n}\hat{z}_{ig}\mathbf{y}_i\mathbf{y}_i' + \beta_g^{(0)}\mathbf{m}_\mathbf{g}^{(\mathbf{0})}\mathbf{m}_g^{(0)T} - \beta_g\mathbf{m}_g\mathbf{m}_g'$

Appendix B: Posterior expected value of the precision parameters of the eigen-decomposed covariance matrix

Table 7 Posterior expected value of the precision parameters of the eigen-decomposed covariance matrix for 12 of the members of the GPCM family

Model	Parameters	Expected values
EII	$\lambda\mathbf{I}_d$	$\mathbb{E}[(\lambda)^{-1}] = a/b$
		$\mathbb{E}[\log\lvert(\lambda)^{-1}\rvert] = \Psi(a/2) - \log(b/2)$
VII	$\lambda_g\mathbf{I}_d$	$\mathbb{E}[(\lambda_g)^{-1}] = a_g/b_g$
		$\mathbb{E}[\log\lvert(\lambda_g)^{-1}\rvert] = \Psi(a_g/2) - \log(b_g/2)$
EEI	$\lambda\mathbf{A}$	$\mathbb{E}[(\lambda\mathbf{A})_{k,k}^{-1}] = a_k/b_k$
		$\mathbb{E}[\log\lvert(\lambda\mathbf{A})_{k,k}^{-1}\rvert] = \Psi(a_k/2) - \log(b_k/2)$
VEI	$\lambda_g\mathbf{A}$	$\mathbb{E}[\lambda_g^{-1}] = a_g/b_g$
		$\mathbb{E}[\log\lvert\lambda_g^{-1}\rvert] = \Psi(a_g/2) - \log(b_g/2)$
		$\mathbb{E}[(c\mathbf{A}^{-1})_{k,k}] = (al_k)/(be_k)$
		$\mathbb{E}[\log\lvert(c\mathbf{A}^{-1})_{k,k}\rvert] = \Psi(al_k/2) - \log(be_k/2)$
EVI	$\lambda\mathbf{A}_g$	$\mathbb{E}[\lambda^{-1}] = a/b$
		$\mathbb{E}[\log\lvert\lambda^{-1}\rvert] = \Psi(a/2) - \log(b/2)$
		$\mathbb{E}[(c_g\mathbf{A}_g^{-1})_{k,k}] = a_{gk}/b_{gk}$
		$\mathbb{E}[\log\lvert(c_g\mathbf{A}_g^{-1})_{k,k}\rvert] = \Psi(a_{gk}/2) - \log(b_{gk}/2)$
VVI	$\lambda_g\mathbf{A}_g$	$\mathbb{E}[\lambda_g^{-1}] = a_g/b_g$
		$\mathbb{E}[\log\lvert\lambda_g^{-1}\rvert] = \Psi(a_g/2) - \log(b_g/2)$
		$\mathbb{E}[(c_g\mathbf{A}_g^{-1})_{k,k}] = a_{gk}/b_{gk}$
		$\mathbb{E}[\log\lvert(c_g\mathbf{A}_g^{-1})_{k,k}\rvert] = \Psi(a_{gk}/2) - \log(b_{gk}/2)$
EEE	$\lambda\mathbf{DAD}'$	$\mathbb{E}[(\lambda\mathbf{DAD}')^{-1}] = v\Sigma^{-1}$
		$\mathbb{E}[\log\lvert(\lambda\mathbf{DAD}')^{-1}\rvert] = \sum_{k=1}^{d}\Psi((v+1-k)/2) + d\log(2) - \log\lvert\Sigma\rvert$
VEE	$\lambda_g\mathbf{DAD}'$	$\mathbb{E}[\lambda_g^{-1}] = a_g/b_g$
		$\mathbb{E}[\log\lvert\lambda_g^{-1}\rvert] = \Psi(a_g/2) - \log(b_g/2)$
		$\mathbb{E}[(\mathbf{DAD}')^{-1}] = v\Sigma^{-1}$
		$\mathbb{E}[\log\lvert(\mathbf{DAD}')^{-1}\rvert] = \sum_{k=1}^{d}\Psi((v+1-k)/2) + d\log(2) - \log\lvert\Sigma\rvert$

Table 7 (continued)

Model	Parameters	Expected values
EEV	$\lambda \mathbf{D}_g \mathbf{A} \mathbf{D}_g'$	$\mathbb{E}[(\lambda \mathbf{A})_{k,k}] = a_k/b_k$
		$\mathbb{E}[\log \lvert(\lambda \mathbf{A})_{k,k}\rvert] = \Psi(a_k/2) - \log(b_k/2)$
		$\mathbb{E}[(\lambda \mathbf{D}_g \mathbf{A} \mathbf{D}_g')^{-1} \lvert (\lambda \mathbf{A})^{-1}]$ via Monte Carlo integration
VEV	$\lambda_g \mathbf{D}_g \mathbf{A} \mathbf{D}_g'$	$\mathbb{E}[\lambda_g^{-1}] = a_g/b_g$
		$\mathbb{E}[\log \lvert\lambda_g^{-1}\rvert] = \Psi(a_g/2) - \log(b_g/2)$
		$\mathbb{E}[(c\mathbf{A}^{-1})_{k,k}] = (al_k)/(be_k)$
		$\mathbb{E}[\log \lvert(\mathbf{A}^{-1})_{k,k}\rvert] = \Psi(al_k/2) - \log(be_k/2)$
		$\mathbb{E}[(\lambda_g^{-1} \mathbf{D}_g \mathbf{A} \mathbf{D}_g')^{-1} \lvert (\lambda_g \mathbf{A})^{-1}]$ via Monte Carlo integration
EVV	$\lambda \mathbf{D}_g \mathbf{A}_g \mathbf{D}_g'$	$\mathbb{E}[\lambda] = a/b$
		$\mathbb{E}[\log \lambda] = \Psi(a/2) - \log(b/2)$
		$\mathbb{E}[(\mathbf{D}_g \mathbf{A}_g \mathbf{D}_g')^{-1}] = v\Sigma_g^{-1}$
		$\mathbb{E}[\log \lvert(\mathbf{D}_g \mathbf{A}_g \mathbf{D}_g')^{-1}\rvert] = \sum_{k=1}^{d} \Psi((v_g + 1 - k)/2) + d\log(2) - \log \lvert\Sigma_g\rvert$
VVV	$\lambda_g \mathbf{D}_g \mathbf{A}_g \mathbf{D}_g'$	$\mathbb{E}[(\mathbf{D}_g \mathbf{A}_g \mathbf{D}_g')^{-1}] = v\Sigma_g^{-1}$
		$\mathbb{E}[\log \lvert(\mathbf{D}_g \mathbf{A}_g \mathbf{D}_g')^{-1}\rvert] = \sum_{k=1}^{d} \Psi((v_g + 1 - k)/2) + d\log(2) - \log \lvert\Sigma_g\rvert$

Appendix C: Mathematical details for the EEV and VEV Models

C.1 EEV Model

The mixing proportions were assigned a Dirichlet prior distribution, such that

$$q_\rho(\boldsymbol{\rho}) = \mathrm{Dir}(\boldsymbol{\rho}; \alpha_1^{(0)}, \ldots, \alpha_G^{(0)}).$$

For the mean, a Gaussian distribution conditional on the covariance matrix was used, such that

$$q_{\boldsymbol{\mu}}(\boldsymbol{\mu} \mid \lambda, \mathbf{A}, \mathbf{D}_1, \ldots, \mathbf{D}_G) = \prod_{g=1}^{G} \phi_d(\boldsymbol{\mu}_g; \mathbf{m}_g^{(0)}, (\beta_g^{(0)-1} \lambda \mathbf{D}_g \mathbf{A} \mathbf{D}_g')).$$

For the parameters of the covariance matrix, the following priors were used: the kth diagonal elements of $(\lambda\mathbf{A})^{-1}$ were assigned a Gamma $(a_k^{(0)}, b_k^{(0)})$ distribution and $\mathbf{D}_g$ was assigned a matrix von Mises-Fisher $(\mathbf{C}_g^{(0)})$ distribution. By setting $\boldsymbol{\tau} = (\lambda\mathbf{A})^{-1}$, its prior can be written

$$p_\tau(\boldsymbol{\tau}) \propto \prod_{k=1}^{K} \tau_k^{\frac{a_k^{(0)}}{2} - 1} \exp\left\{-\frac{b_k^{(0)}}{2} \tau_k\right\},$$

where τ_k is the kth diagonal element of $\boldsymbol{\tau} = (\lambda\mathbf{A})^{-1}$.

The matrix $\mathbf{D}$ has a density as defined by Gupta and Nagar (2000):

$$p(\mathbf{D}) = b(\mathbf{Q}^{(0)}, \mathbf{P}_g^{(0)}) \exp(\mathrm{tr}\{\mathbf{Q}^{(0)} \mathbf{D} \mathbf{P}_g^{(0)} \mathbf{D}'\})[d\mathbf{D}],$$

for $\mathbf{D} \in O(d, d)$, where $O(d, d)$ is the Stiefel manifold of $d \times d$ matrices, $[d\mathbf{D}]$ is the unit invariant measure on $O(d, d)$, and $\mathbf{A}^{(0)}$ and $\mathbf{B}_g^{(0)}$ are symmetric and diagonal matrices, respectively.

The joint distribution of $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G$, $\boldsymbol{\tau}$, and $\mathbf{D}$ is

$$p(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G, \boldsymbol{\tau}, \mathbf{D}) \propto \prod_{g=1}^{G} |\beta_g^{(0)}\boldsymbol{\tau}|^{\frac{1}{2}} \exp\left\{\frac{-(\boldsymbol{\mu}_g - \mathbf{m}_g^{(0)})\beta_g^{(0)}\mathbf{D}_g'\boldsymbol{\tau}\mathbf{D}_g(\boldsymbol{\mu}_g - \mathbf{m}_g^{(0)})'}{2}\right\}$$
$$\times \exp\left\{\mathrm{tr}(\mathbf{Q}^{(0)}\mathbf{D}\mathbf{P}_g^{(0)}\mathbf{D}')\right\} \prod_{k=1}^{K} \tau_k^{\frac{a_k^{(0)}}{2}-1} \exp\left\{-\frac{b_k^{(0)}}{2}\tau_k\right\}.$$

The likelihood of the data can be written

$$\mathcal{L}(\boldsymbol{\mu}_1,\ldots, \boldsymbol{\mu}_G, \boldsymbol{\tau}, \mathbf{D} \mid \mathbf{y}_1,\ldots, \mathbf{y}_n) \propto \prod_{i=1}^{n}\prod_{g=1}^{G} |\boldsymbol{\tau}|^{\hat{z}_{ig}/2} \exp\left\{-\frac{\hat{z}_{ig}}{2}(\mathbf{y}_i - \boldsymbol{\mu}_g)\mathbf{D}_g'\boldsymbol{\tau}\mathbf{D}_g(\mathbf{y}_i - \boldsymbol{\mu}_g)'\right\}.$$

Therefore, the joint posterior distribution of $\boldsymbol{\mu}$, $\boldsymbol{\tau}$, and $\mathbf{D}$ can be written

$$p(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G, \boldsymbol{\tau}, \mathbf{D} \mid \mathbf{y}_1, \ldots, \mathbf{y}_n) \propto p(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G, \boldsymbol{\tau}, \mathbf{D})\mathcal{L}(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G, \boldsymbol{\tau}, \mathbf{D} \mid \mathbf{y}_1, \ldots, \mathbf{y}_n).$$

Thus, the posterior distribution of mean becomes

$$q_{\mu}(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G \mid \boldsymbol{\tau}, \mathbf{D}_1, \ldots, \mathbf{D}_G) = \prod_{g=1}^{G} \phi_d(\boldsymbol{\mu}_g; \mathbf{m}_g, (\beta_g\mathbf{D}_g'\boldsymbol{\tau}\mathbf{D}_g)^{-1}),$$

where $\beta_g = \beta_g^{(0)} + \sum_{i=1}^{n} \hat{z}_{ig}$ and

$$\mathbf{m}_g = \frac{1}{\beta_g}\left(\beta_g^{(0)}\mathbf{m}_g^{(0)} + \sum_{i=1}^{n} \hat{z}_{ig}\mathbf{y}_i\right).$$

The posterior distribution for the kth diagonal element of $\boldsymbol{\tau} = (\lambda\mathbf{A})^{-1}$ is

$$q_{\tau}(\tau_k) = \mathrm{Gamma}(a_k, b_k)$$

where $a_k = a_k^{(0)} + d\sum_{g=1}^{G}\sum_{i=1}^{n} \hat{z}_{ig} = a_k^{(0)} + dn$ and

$$b_k = b_k^{(0)} + \sum_{g=1}^{G}\left(\sum_{i=1}^{n} \hat{z}_{ig}y_{ik}^2 + \beta_g^{(0)}m_{gk}^2 - \beta_g m_{gk}^2\right).$$

We have

$$q(\mathbf{D}_g|\mathbf{y}; \boldsymbol{\mu}_g, \boldsymbol{\tau}) \propto \exp\left\{\mathrm{tr}\left(-\frac{1}{2}(\boldsymbol{\mu}_g - \mathbf{m}_g^{(0)})\beta_g^{(0)}\mathbf{D}_g'\boldsymbol{\tau}\mathbf{D}_g(\boldsymbol{\mu}_g - \mathbf{m}_g^{(0)})'\right)\right\}$$
$$\times \exp\left\{\mathrm{tr}\left(-\frac{1}{2}\sum_{i=1}^{n} z_{ig}(\mathbf{y} - \boldsymbol{\mu}_g)\mathbf{D}_g'\boldsymbol{\tau}\mathbf{D}_g(\mathbf{y} - \boldsymbol{\mu}_g)' + \mathbf{Q}_g^{(0)}\mathbf{D}_g\mathbf{P}_g^{(0)}\mathbf{D}_g'\right)\right\},$$

which has the functional form of a Bingham matrix distribution, i.e., the form

$$\exp\left\{\mathrm{tr}(\mathbf{Q}_g\mathbf{D}_g\mathbf{P}_g\mathbf{D}_g')\right\},$$

where $\mathbf{Q}_g = \mathbf{Q}_g^{(0)} + \boldsymbol{\tau}$ and

$$\mathbf{P}_g = \mathbf{P}_g^{(0)} - \frac{1}{2}\left[\sum_{i=1}^{n} z_{ig}(\mathbf{y} - \boldsymbol{\mu}_g)(\mathbf{y} - \boldsymbol{\mu}_g)' + (\boldsymbol{\mu}_g - \mathbf{m}_g^{(0)})\beta_g^{(0)}(\boldsymbol{\mu}_g - \mathbf{m}_g^{(0)})'\right].$$

C.2 VEV Model

Similarly, the posterior distribution of $\mathbf{D}_g$ for the VEV model has the form

$$q(\mathbf{D}_g|\mathbf{y};\boldsymbol{\mu}_g,\boldsymbol{\tau}_g)\propto\exp\left\{\mathrm{tr}\left(-\frac{1}{2}(\boldsymbol{\mu}_g-\mathbf{m}_g^{(0)})\beta_g^{(0)}\mathbf{D}_g'\boldsymbol{\tau}_g\mathbf{D}_g(\boldsymbol{\mu}_g-\mathbf{m}_g^{(0)})'\right)\right\}$$
$$\times\exp\left\{\mathrm{tr}\left(-\frac{1}{2}\sum_{i=1}^{n}\hat{z}_{ig}(\mathbf{y}-\boldsymbol{\mu}_g)\mathbf{D}_g'\boldsymbol{\tau}_g\mathbf{D}_g(\mathbf{y}-\boldsymbol{\mu}_g)'+\mathbf{Q}_g^{(0)}\mathbf{D}_g\mathbf{P}_g^{(0)}\mathbf{D}_g'\right)\right\},$$

which has the functional form of a Bingham matrix distribution, i.e., the form

$$\exp\left\{\mathrm{tr}(\mathbf{Q}_g\mathbf{D}_g\mathbf{P}_g\mathbf{D}_g')\right\},$$

where $\mathbf{Q}_g=\mathbf{Q}_g^{(0)}+\boldsymbol{\tau}_g$ and

$$\mathbf{P}_g=-\frac{1}{2}\left[\sum_{i=1}^{n}\hat{z}_{ig}(\mathbf{y}-\boldsymbol{\mu}_g)(\mathbf{y}-\boldsymbol{\mu}_g)'+(\boldsymbol{\mu}_g-\mathbf{m}_g^{(0)})\beta_g^{(0)}(\boldsymbol{\mu}_g-\mathbf{m}_g^{(0)})'\right].$$

References

Aitken, A.C. (1926). A series formula for the roots of algebraic and transcendental equations. *Proceedings of the Royal Society of Edinburgh*, *45*, 14–22.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

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

Bensmail, H., Celeux, G., Raftery, A.E., Robert, C.P. (1997). Inference in model-based cluster analysis. *Statistics and Computing*, *7*, 1–10.

Biernacki, C., Celeux, G., Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(7), 719–725.

Biernacki, C., & Lourme, A. (2019). Unifying data units and models in (co-)clustering. *Advances in Data Analysis and Classification*, *13*(1), 7–31.

Bingham, C. (1974). An antipodally symmetric distribution on the sphere. *The Annals of Statistics*, *2*(6), 1201–1225.

Blei, D.M., Kucukelbir, A., McAuliffe, J.D. (2017). Variational inference: a review for statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877.

Bock, H.H. (1996). Probabilistic models in cluster analysis. *Computational Statistics and Data Analysis*, *23*, 5–28.

Bock, H.H. (1998a). *Data science, classification and related methods*, (pp. 3–21). New York: Springer-Verlag.

Bock, H.H. (1998b). Probabilistic approaches in cluster analysis. *Bulletin of the International Statistical Institute*, *57*, 603–606.

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P., Lindsay, B. (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics*, *46*, 373–388.

Boulesteix, A.-L., Durif, G., Lambert-Lacroix, S., Peyre, J., Strimmer, K. (2018). plsgenomics: PLS Analyses for Genomics. R package version 1.5-2.

Bouveyron, C., & Brunet-Saumard, C. (2014). Model-based clustering of high-dimensional data: a review. *Computational Statistics and Data Analysis*, *71*, 52–78.

Browne, R.P., & McNicholas, P.D. (2014). Estimating common principal components in high dimensions. *Advances in Data Analysis and Classification*, *8*(2), 217–226.

Casella, G., Mengersen, K., Robert, C., Titterington, D. (2002). Perfect samplers for mixtures of distributions. *Journal of the Royal Statistical Society: Series B*, *64*, 777–790.

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

Celeux, G., Hurn, M., Robert, C. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, *95*, 957–970.

Cheam, A.S.M., Marbac, M., McNicholas, P.D. (2017). Model-based clustering for spatiotemporal data on air quality monitoring. *Environmetrics*, *93*, 192–206.

Corduneanu, A., & Bishop, C. (2001). Variational Bayesian model selection for mixture distributions. In *Artificial intelligence and statistics* (pp. 27–34). Los Altos: Morgan Kaufmann.

Dang, U.J., Browne, R.P., McNicholas, P.D. (2015). Mixtures of multivariate power exponential distributions. *Biometrics*, *71*(4), 1081–1089.

Dang, U.J., Punzo, A., McNicholas, P.D., Ingrassia, S., Browne, R.P. (2017). Multivariate response and parsimony for Gaussian cluster-weighted models. *Journal of Classification*, *34*(1), 4–34.

Day, N.E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika*, *56*(3), 463–474.

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

Diebolt, J., & Robert, C. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society: Series B*, *56*, 363–375.

Fraley, C., & Raftery, A.E. (2007). Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification*, *24*, 155–181.

Franczak, B.C., Browne, R.P., McNicholas, P.D. (2014). Mixtures of shifted asymmetric Laplace distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(6), 1149–1157.

Gallaugher, M.P.B., & McNicholas, P.D. (2018a). Finite mixtures of skewed matrix variate distributions. *Pattern Recognition*, *80*, 83–93.

Gallaugher, M.P.B., & McNicholas, P.D. (2018b). A mixture of matrix variate bilinear factor analyzers. In: Proceedings of the Joint Statistical Meetings. Alexandria, VA: American Statistical Association. Also available as arXiv preprint. arXiv:1712.08664v3.

Gallaugher, M.P.B., & McNicholas, P.D. (2019a). Mixtures of skewed matrix variate bilinear factor analyzers. Advances in Data Analysis and Classification. To appear. https://doi.org/10.1007/s11634-019-00377-4.

Gallaugher, M.P.B., & McNicholas, P.D. (2019b). On fractionally-supervised classification: weight selection and extension to the multivariate t-distribution. *Journal of Classification*, *36*(2), 232–265.

Gelman, A., Stern, H.S., Carlin, J.B., Dunson, D.B., Vehtari, A., Rubin, D.B. (2013). *Bayesian data analysis*. Boca Raton: Chapman and Hall/CRC Press.

Gupta, A., & Nagar, D. (2000). *Matrix variate distributions*. Boca Raton: Chapman & Hall/CRC Press.

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

Hasselblad, V. (1966). Estimation of parameters for a mixture of normal distributions. *Technometrics*, *8*(3), 431–444.

Hoff, P. (2012). rstiefel: random orthonormal matrix generation on the Stiefel manifold. R package version 0.9.

Hoff, P.D. (2009). Simulation of the matrix Bingham-von Mises-Fisher distribution, with applications to multivariate and relational data. *Journal of Computational and Graphical Statistics*, *18*(2), 438–456.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*, 193–218.

Jasra, A., Holmes, C.C., Stephens, D.A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Journal of the Royal Statistical Society: Series B*, *10*(1), 50–67.

Jordan, M., Ghahramani, Z., Jaakkola, T., Saul, L. (1999). An introduction to variational methods for graphical models. *Machine Learning*, *37*, 183–233.

Lee, S., & McLachlan, G.J. (2014). Finite mixtures of multivariate skew t-distributions: some recent and new results. *Statistics and Computing*, *24*, 181–202.

Lee, S.X., & McLachlan, G.J. (2016). Finite mixtures of canonical fundamental skew t-distributions – the unification of the restricted and unrestricted skew t-mixture models. *Statistics and Computing*, *26*(3), 573–589.

Lin, T., McLachlan, G.J., Lee, S.X. (2016). Extending mixtures of factor models using the restricted multivariate skew-normal distribution. *Journal of Multivariate Analysis*, *143*, 398–413.

Lin, T.-I., McNicholas, P.D., Hsiu, J.H. (2014). Capturing patterns via parsimonious t mixture models. *Statistics and Probability Letters*, *88*, 80–87.

MacQueen, J.B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press.

McGrory, C., & Titterington, D. (2007). Variational approximations in Bayesian model selection for finite mixture distributions. *Computational Statistics and Data Analysis*, *51*, 5352–5367.

McGrory, C., & Titterington, D. (2009). Variational Bayesian analysis for hidden Markov models. *Australian and New Zealand Journal of Statistics*, *51*, 227–244.

McGrory, C., Titterington, D., Pettitt, A. (2009). Variational Bayes for estimating the parameters of a hidden Potts model. *Computational Statistics and Data Analysis*, *19*(3), 329–340.

McLachlan, G.J., & Krishnan, T. (2008). *The EM algorithm and extensions*, 2nd edn. New York: Wiley.

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. (2016a). *Mixture model-based classification*. Boca Raton: Chapman & Hall/CRC Press.

McNicholas, P.D. (2016b). Model-based clustering. *Journal of Classification*, *33*(3), 331–373.

McNicholas, P.D., & Murphy, T.B. (2008). Parsimonious Gaussian mixture models. *Statistics and Computing*, *18*, 285–296.

McNicholas, P.D., & Murphy, T.B. (2010). Model-based clustering of microarray expression data via latent Gaussian mixture models. *Bioinformatics*, *26*(21), 2705–2712.

Melnykov, V., & Zhu, X. (2018). On model-based clustering of skewed matrix data. *Journal of Multivariate Analysis*, *167*, 181–194.

Morris, K., & McNicholas, P.D. (2016). Clustering, classification, discriminant analysis, and dimension reduction via generalized hyperbolic mixtures. *Computational Statistics and Data Analysis*, *97*, 133–150.

Morris, K., Punzo, A., McNicholas, P.D., Browne, R.P. (2019). Asymmetric clusters and outliers: mixtures of multivariate contaminated shifted asymmetric Laplace distributions. *Computational Statistics and Data Analysis*, *132*, 145–166.

Murray, P.M., Browne, R.B., McNicholas, P.D. (2014a). Mixtures of skew-t factor analyzers. *Computational Statistics and Data Analysis*, *77*, 326–335.

Murray, P.M., Browne, R.P., McNicholas, P.D. (2019). Mixtures of hidden truncation hyperbolic factor analyzers. Journal of Classification. To appear. https://doi.org/10.1007/s00357-019-9309-y.

Murray, P.M., McNicholas, P.D., Browne, R.P. (2014b). A mixture of common skew-t factor analyzers. *Stat*, *3*(1), 68–82.

Neath, R.C. et al. (2013). On convergence properties of the Monte Carlo EM algorithm. In: Advances in modern statistical theory and applications: a Festschrift in Honor of Morris L. Eaton, pp.43–62. Institute of Mathematical Statistics.

O'Hagan, A., Murphy, T.B., Gormley, I.C., McNicholas, P.D., Karlis, D. (2016). Clustering with the multivariate normal inverse Gaussian distribution. *Computational Statistics and Data Analysis*, *93*, 18–30.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London A*, *185*, 71–110.

Punzo, A., Blostein, M., McNicholas, P.D. (2020). High-dimensional unsupervised classification via parsimonious contaminated mixtures. *Pattern Recognition*, *98*, 107031.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Rand, W.M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*, 846–850.

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

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

Scrucca, L., Fop, M., Murphy, T.B., Raftery, A.E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, *8*(1), 205–233.

Spiegelhalter, D., Best, N., Carlin, B., Van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society: Series B*, *64*, 583–639.

Stephens, M. (1997). *Bayesian methods for mixtures of normal distributions*. Oxford: Ph.D. thesis University of Oxford.

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*, 40–74.

Subedi, S., & McNicholas, P.D. (2014). Variational Bayes approximations for clustering via mixtures of normal inverse Gaussian distributions. *Advances in Data Analysis and Classification*, *8*(2), 167–193.

Subedi, S., Punzo, A., Ingrassia, S., McNicholas, P.D. (2015). Cluster-weighed t-factor analyzers for robust model-based clustering and dimension reduction. *Statistical Methods and Applications*, *24*(4), 623–649.

Titterington, D.M., Smith, A.F.M., Makov, U.E. (1985). *Statistical analysis of finite mixture distributions*. Chichester: John Wiley & Sons.

Tortora, C., Franczak, B.C., Browne, R.P., McNicholas, P.D. (2019). A mixture of coalesced generalized hyperbolic distributions. *Journal of Classification*, *36*(1), 26–57.

Ueda, N., & Ghahramani, Z. (2002). Bayesian model search for mixture models based on optimizing variational bounds. *Neural Networks*, *15*, 1223–1241.

Venables, W.N., & Ripley, B.D. (2002). *Modern applied statistics with S*, 4th edn. New York: Springer.

Viroli, C. (2011). Finite mixtures of matrix normal distributions for classifying three-way data. *Statistics and Computing*, *21*(4), 511–522.

Vrbik, I., & McNicholas, P.D. (2014). Parsimonious skew mixture models for model-based clustering and classification. *Computational Statistics and Data Analysis*, *71*, 196–210.

Vrbik, I., & McNicholas, P.D. (2015). Fractionally-supervised classification. *Journal of Classification*, *32*(3), 359–381.

Wang, X., He, C.Z., Sun, D. (2005). Bayesian inference on the patient population size given list mismatches. *Statistics in Medicine*, *24*(2), 249–267.

Wolfe, J.H. (1965). A computer program for the maximum likelihood analysis of types. Technical Bulletin 65–15, U.S.Naval Personnel Research Activity.

Zhu, X., & Melnykov, V. (2018). Manly transformation in finite mixture modeling. *Computational Statistics & Data Analysis*, *121*, 190–208.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.