Variational Bayes Approximations for Clustering via Mixtures of Normal Inverse Gaussian Distributions

Sanjeena Subedi* and Paul D. McNicholas

Department of Mathematics & Statistics, University of Guelph

Abstract

Parameter estimation for model-based clustering using a finite mixture of normal inverse Gaussian (NIG) distributions is achieved through variational Bayes approximations. Univariate NIG mixtures and multivariate NIG mixtures are considered. The use of variational Bayes approximations here is a substantial departure from the traditional EM approach and alleviates some of the associated computational complexities and uncertainties. Our variational algorithm is applied to simulated and real data. The paper concludes with discussion and suggestions for future work.

Keywords: Clustering, MNIG, NIG, normal inverse Gaussian, variational approximations, variational Bayes

1 Introduction

The use of mixture models for clustering, referred to as model-based clustering, has become increasingly popular since the work of Wolfe (1963). A wide variety of finite mixture models has been studied extensively within the literature to date. Amongst these, the Gaussian mixture model has received special attention due to its mathematical tractability and the relative computational simplicity associated with parameter estimation. However, the Gaussian mixture model is not without limitations; for instance, the component densities are restricted to being symmetric. Over the past few years, there has been a notable increase in the preponderance of non-Gaussian mixture modelling within the literature (e.g., Lin, 2009, 2010; Andrews et al., 2011; Baek and McLachlan, 2011; Steane et al., 2012; McNicholas and Subedi, 2012; Vrbik and McNicholas, 2012; Browne et al., 2012; Morris et al., 2013; Morris and McNicholas, 2013a,b; Lee and McLachlan, 2013; Murray et al., 2013). Karlis and Santourian (2009) proposed a mixture of univariate normal inverse

*Department of Mathematics & Statistics, University of Guelph, Guelph, Ontario, N1G 2W1, Canada. E-mail: ssubedi@uoguelph.ca.

Gaussian (UNIG) distributions and a mixture of multivariate normal inverse Gaussian (MNIG) distributions; these models have the flexibility to represent both skewed and symmetric populations as well as mixtures thereof. As is typical within the field, parameter estimation for the NIG mixtures has heretofore been carried out using an expectation-maximization (EM) algorithm (Baum et al., 1970; Orchard and Woodbury, 1972; Sundberg, 1974; Dempster et al., 1977); see Karlis and Santourian (2009) for details.

The EM algorithm is an iterative procedure used to find maximum likelihood estimates for incomplete data. In the clustering context, the group memberships are missing and latent variables may also be present. One major drawback to the EM approach is its dependency on starting values. This and other problems arise because of the unpleasant nature of the likelihood surface, which leads to a very slow rate of convergence and, in some cases, convergence to local minima (cf. Titterington et al., 1985). As reported by Karlis and Santourian (2009), the EM algorithm can be very slow when dealing with complicated distributions, such as the MNIG. Furthermore, when the number of components in a mixture model is unknown, the computational cost increases further because the EM algorithm must be used in conjunction with a model-selection criterion so that every possible number of components is explored (e.g., Fraley and Raftery, 2002; Bouveyron et al., 2007; McNicholas and Murphy, 2008, 2010). Beyond the increased computational cost, this problem is further compounded by the fact that using different model selection criteria on the same data can result in selection of a different set of models (see Andrews and McNicholas, 2011, for examples).

Variational Bayes approximations have been explored by many researchers, including Waterhouse et al. (1996), Jordan et al. (1999), Corduneanu and Bishop (2001), and McGrory and Titterington (2007). Variational Bayes approximations are an iterative Bayesian alternative to the EM algorithm, and their fast and deterministic nature has made the approach increasingly popular over the past decade or so. The tractability of the variational approach allows for simultaneous model selection and parameter estimation, thus removing the need for a model selection criterion and reducing the associated computational overhead. The variational Bayes algorithm has been applied to Gaussian mixture models (cf. Teschendorff et al., 2005; McGrory and Titterington, 2007). For observed data $\mathbf{y}$, the joint conditional distribution of parameters $\boldsymbol{\theta}$ and missing data $\mathbf{w}$ is approximated by constructing a tight lower bound on the complex data marginal likelihood using a computationally convenient density $q_{\boldsymbol{\theta},\mathbf{w}}(\boldsymbol{\theta},\mathbf{w})$. The approximating density $q_{\boldsymbol{\theta},\mathbf{w}}(\boldsymbol{\theta},\mathbf{w})$ is obtained by minimizing the Kullback-Leibler (KL) divergence between the true density $h(\boldsymbol{\theta},\mathbf{w}|\mathbf{y})$ and the approximating density (Beal, 2003; McGrory and Titterington, 2007). Due to the non-negative property of the KL divergence, minimizing the KL divergence is equivalent to maximizing the lower bound. The algorithm is initialized with more components than expected, and estimation of the parameters and the number of components is performed simultaneously.

In this paper, we develop a variational Bayes framework for parameter estimation for UNIG mixtures and MNIG mixtures. Using the variational Bayes framework reduces the computational cost associated with this complex modelling framework by simultaneously estimating the parameters and the number of components. We show that variational Bayes approximations can be very effective for non-Gaussian mixture model-based clustering. The remainder of this paper is laid

out as follows. Variational approximations are developed and illustrated for UNIG mixture models (Section 2). They are then developed and illustrated for MNIG mixtures in Section 3. The paper concludes with discussion and suggestions for future work (Section 4).

2 Mixture of Univariate Normal Inverse Gaussian Distributions

In this section, we introduce a variational Bayes framework for parameter estimation for the UNIG mixture model.

2.1 The Model

A mean-variance mixture of a univariate normal distribution with the inverse Gaussian (IG) distribution (Barndorff-Nielsen, 1997), i.e.,

$$Y \mid u \sim \mathrm{N}(\mu + \beta u, u), \qquad U \sim \mathrm{IG}(\delta, \gamma),$$

results in a UNIG distribution with density

$$f(y;\boldsymbol{\theta}) = \frac{\alpha}{\pi} \exp\left\{\delta\sqrt{\alpha^2-\beta^2} - \beta\mu\right\} \phi(y)^{-\frac{1}{2}} K_1(\delta\alpha\phi(y)^{\frac{1}{2}}) \exp\{\beta y\},$$

where $\boldsymbol{\theta} = (\alpha, \beta, \mu, \delta)$ are the model parameters such that $\alpha^2 = \gamma^2 + \beta^2$, $\phi(y) = 1 + [(y-\mu)/\delta]^2$, and $K_1(y)$ is the modified Bessel function of the third kind of order 1 evaluated at y (Abramowitz and Stegun, 1972). The expected value and variance of Y are $\mathbb{E}(Y) = \mu + \delta\beta/\gamma$ and $\mathrm{Var}(Y) = \delta\alpha^2/\gamma^3$, respectively. Here, δ is a scaling parameter, μ is a location parameter, β controls the asymmetry, and $\alpha \pm \beta$ determines the heaviness of the tails. The density of the IG distribution with parameters δ and γ is

$$f(u) = (2\pi)^{-1/2}\delta u^{-3/2} \exp\left\{\delta\gamma - \frac{1}{2}(\delta^2 u^{-1} + \gamma^2 u)\right\}. \tag{1}$$

The expected value and variance of U are $\mathbb{E}[U] = \delta/\gamma$ and $\mathrm{Var}[U] = \delta/\gamma^3$, respectively. Note that this is different from the parameterization of the IG distribution used by Seshadri (1993), and can be obtained as a special case of generalized inverse Gaussian distribution (Chhikara and Folks, 1989). See Barndorff-Nielsen (1997) and Karlis and Santourian (2009) for more details on the UNIG distribution.

2.2 Parameter Estimation

From Karlis and Lillestol (2004), the joint probability density is given by $f(y,u) = f(u)f(y|u)$, where

$$f(y|u) = (2\pi)^{-1/2} u^{-1/2} \exp\left\{-\frac{1}{2u}(y-(\mu+\beta u))^2\right\},$$

and $f(u)$ is as defined in (1). Therefore,

$$f(y,u) \propto \delta \exp\{\delta\gamma - \beta\mu\} u^{-2} \exp\left\{\beta y + \mu\frac{y}{u} - \frac{1}{2}(\beta^2+\gamma^2)u - \frac{1}{2}(\mu^2+\delta^2)u^{-1}\right\}.$$

The likelihood of the complete UNIG data, i.e. the observed $\mathbf{y}$ and the latent $\mathbf{u}$ such that $(\mathbf{y},\mathbf{u}) = (y_1,\ldots,y_n,u_1,\ldots,u_n)$, has the form

$$L(\boldsymbol{\theta}) = r(\boldsymbol{\theta})^n \left[\prod_{i=1}^{n} h(y_i,u_i)\right] \exp\left\{\sum_{j=1}^{4} \Phi_j(\boldsymbol{\theta}) t_j(\mathbf{y},\mathbf{u})\right\},$$

which is within the exponential family. Here, $r(\boldsymbol{\theta})$ is the normalization constant that depends on $\boldsymbol{\theta}$, $h(y,u)$ is a continuous function of (y,u), $\Phi_j(\boldsymbol{\theta})$ is the jth natural parameter, and $t_j(\mathbf{y},\mathbf{u})$ is the jth sufficient statistic with $\mathbf{u} = (u_1,\ldots,u_n)$. For the UNIG model: $\Phi_1 = \beta$, $\Phi_2 = \mu$, $\Phi_3 = \beta^2+\gamma^2$, and $\Phi_4 = \mu^2+\delta^2$; and $t_1(\mathbf{y},\mathbf{u}) = \sum_{i=1}^n y_i$, $t_2(\mathbf{y},\mathbf{u}) = \sum_{i=1}^n y_i u_i$, $t_3(\mathbf{y},\mathbf{u}) = \frac{1}{2}\sum_{i=1}^n u_i$, and $t_4(\mathbf{y},\mathbf{u}) = \frac{1}{2}\sum_{i=1}^n u_i^{-1}$. If the conjugate prior distribution of $\boldsymbol{\theta}$ is of the form

$$h(\boldsymbol{\theta}) \propto r(\boldsymbol{\theta})^{a_0} \exp\left\{\sum_{j=1}^{4} \Phi_j(\boldsymbol{\theta}) a_j\right\},$$

then the posterior distribution is of the form

$$h(\boldsymbol{\theta} \mid \mathbf{y}) \propto r(\boldsymbol{\theta})^{(a_0+n)} \exp\left\{\sum_{j=1}^{4} \Phi_j(\boldsymbol{\theta}) \left(a_j + t_j(\mathbf{y},\mathbf{u})\right)\right\}.$$

Because (μ,β) is independent of (δ,γ), a bivariate prior normal distribution can be assigned to (μ,β), a gamma prior distribution can be assigned to δ^2, and a truncated normal prior conditional on δ can be assigned to γ (Karlis and Lillestol, 2004). The values a_j will be discussed shortly, in the mixture context.

Now consider n independent random variables $Y_1,\ldots,Y_n$ from a G-component mixture of UNIG distributions. The likelihood of the observed data $\mathbf{y} = (y_1,\ldots,y_n)$ from this mixture will have the form

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n}\sum_{g=1}^{G} \pi_g f(y_i;\boldsymbol{\theta}_g),$$

where $\pi_g > 0$ such that $\sum_{g=1}^G \pi_g = 1$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_g)$, and $f(\mathbf{y};\boldsymbol{\theta}_g)$ is the gth component density given by a UNIG distribution with parameters $\boldsymbol{\theta}_g = (\alpha_g,\beta_g,\mu_g,\delta_g)$. Note that $\pi_1,\ldots,\pi_G$ are called mixing proportions.

Define a component indicator variable Z_{ig} such that $z_{ig} = 1$ if the observation i belongs to component g and $z_{ig} = 0$ otherwise. The complete-data, i.e., the observed y_i, the latent u_{ig}, and the

missing z_{ig}, likelihood of a G-component mixture of UNIG distributions can be written

$$\begin{aligned} L_{\mathrm{c}}(\boldsymbol{\theta}) &= \prod_{g=1}^{G}\prod_{i=1}^{n}[\pi_g\ f(y_i|u_{ig};\mu_g,\beta_g)\ f(u_{ig};\delta_g,\gamma_g)]^{z_{ig}} \\ &= \prod_{g=1}^{G}\left[r(\boldsymbol{\theta}_g)^{\sum_{i=1}^{n} z_{ig}}\left(\prod_{i=1}^{n} h(y_i,u_{ig})\right)\exp\left\{\sum_{j=1}^{4}\Phi_j(\boldsymbol{\theta}_g)t_j(\mathbf{y},\mathbf{u}_g)\right\}\right], \end{aligned}$$

where $\mathbf{u}_g = (u_{1g},\ldots,u_{ng})$. If the conjugate prior distribution of $\boldsymbol{\theta}_g = (\pi_g,\mu_g,\beta_g,\delta_g,\gamma_g)$ is of the form

$$h(\boldsymbol{\theta}_g) \propto r(\boldsymbol{\theta}_g)^{a_{g,0}^{(0)}}\exp\left\{\sum_{j=1}^{4}\Phi_j(\boldsymbol{\theta}_g)a_{g,j}^{(0)}\right\},$$

with hyperparameters taking initial values $(a_{g,0}^{(0)},a_{g,1}^{(0)},\ldots,a_{g,4}^{(0)})$, then the posterior distribution is of the form

$$h(\boldsymbol{\theta}_g \mid \mathbf{y}) \propto r(\boldsymbol{\theta}_g)^{(a_{g,0}^{(0)}+\sum_{i=1}^{n} z_{ig})}\exp\left\{\sum_{j=1}^{4}\Phi_j(\boldsymbol{\theta}_g)\left(a_{g,j}^{(0)}+t_j(\mathbf{y},\mathbf{u}_g)\right)\right\},$$

where

$$\begin{aligned} a_{g,0} &= a_{g,0}^{(0)}+\textstyle\sum_{i=1}^{n} z_{ig}, & a_{g,1} &= a_{g,1}^{(0)}+\displaystyle\sum_{i=1}^{n} z_{ig}y_i, \\ a_{g,2} &= a_{g,2}^{(0)}+\textstyle\sum_{i=1}^{n} z_{ig}u_{ig}^{-1}y_i, & a_{g,3} &= a_{g,3}^{(0)}+0.5\displaystyle\sum_{i=1}^{n} z_{ig}u_{ig}, \\ a_{g,4} &= a_{g,4}^{(0)}+0.5\textstyle\sum_{i=1}^{n} z_{ig}u_{ig}^{-1}. \end{aligned}$$

The approximating density in the variational Bayes framework is restricted to a factorized form for computational convenience, so that $q_{\boldsymbol{\theta},\mathbf{w}}(\boldsymbol{\theta},\mathbf{w}) = q_{\boldsymbol{\theta}}(\boldsymbol{\theta})q_{\mathbf{w}}(\mathbf{w})$. For the mixture of UNIG distributions, the missing data are $\mathbf{w} = (\mathbf{z},\mathbf{u})$, where $\mathbf{z} = (\mathbf{z}_1,\ldots,\mathbf{z}_n)$ with $\mathbf{z}_i = (z_{i1},\ldots,z_{iG})$, and $\mathbf{u}$ is defined similarly. Therefore, the approximating density is $q_{\boldsymbol{\theta},\mathbf{w}}(\boldsymbol{\theta},\mathbf{w}) = q_{\boldsymbol{\theta}}(\boldsymbol{\theta})q_{\mathbf{z},\mathbf{u}}(\mathbf{z},\mathbf{u})$. Upon choosing a conjugate prior, the appropriate hyperparameters for the approximating density $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ for data from an exponential-family model can easily be obtained.

A Dirichlet prior with initial hyperparameters $(a_{1,0}^{(0)},\ldots,a_{G,0}^{(0)})$ is assigned to the mixing components $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_g)$ and results in a Dirichlet posterior distribution with hyperparameters $(a_{1,0},\ldots,a_{G,0})$. A bivariate normal prior distribution is assigned to (μ_g,β_g) such that

$$\begin{pmatrix} \mu_g \\ \beta_g \end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix} \bar{\mu}_g^{(0)} \\ \bar{\beta}_g^{(0)} \end{pmatrix}, \begin{pmatrix} \sigma_{\mu_g}^{(0)\,2} & \rho_g^{(0)}\sigma_{\mu_g}^{(0)}\sigma_{\beta_g}^{(0)} \\ \rho_g^{(0)}\sigma_{\mu_g}^{(0)}\sigma_{\beta_g}^{(0)}, & \sigma_{\beta_g}^{(0)\,2} \end{pmatrix}\right],$$

where

$$\rho_g^{(0)} = -\frac{a_{g,0}^{(0)}}{2\sqrt{a_{g,3}^{(0)} a_{g,4}^{(0)}}}, \qquad \bar{\mu}_g^{(0)} = \frac{1}{2(1-\rho_g^{(0)\,2}) a_{g,4}^{(0)}} \left(a_{g,2}^{(0)} - \frac{a_{g,0}^{(0)} a_{g,1}^{(0)}}{2a_{g,3}^{(0)}} \right),$$

$$\sigma_{\mu_g}^{(0)\,2} = \frac{1}{2(1-\rho_g^{(0)\,2}) a_{g,4}^{(0)}}, \qquad \bar{\beta}_g^{(0)} = \frac{1}{2(1-\rho_g^{(0)\,2}) a_{g,3}^{(0)}} \left(a_{g,1}^{(0)} - \frac{a_{g,0}^{(0)} a_{g,2}^{(0)}}{2a_{g,4}^{(0)}} \right),$$

$$\sigma_{\beta_g}^{(0)\,2} = \frac{1}{2(1-\rho_g^{(0)\,2}) a_{g,3}^{(0)}}.$$

The resulting posterior distribution for (μ_g, β_g) is

$$\begin{pmatrix} \mu_g \\ \beta_g \end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix} \bar{\mu}_g \\ \bar{\beta}_g \end{pmatrix}, \begin{pmatrix} \sigma_{\mu_g}^2 & \rho_g \sigma_{\mu_g} \sigma_{\beta_g} \\ \rho_g \sigma_{\mu_g} \sigma_{\beta_g}, & \sigma_{\beta_g}^2 \end{pmatrix} \right],$$

where

$$\rho_g = -\frac{a_{g,0}}{2\sqrt{a_{g,3} a_{g,4}}}, \qquad \bar{\mu}_g = \frac{1}{2(1-\rho_g^2) a_{g,4}} \left(a_{g,2} - \frac{a_{g,0} a_{g,1}}{2a_{g,3}} \right),$$

$$\sigma_{\mu_g}^2 = \frac{1}{2(1-\rho_g^2) a_{g,4}}, \qquad \bar{\beta}_g = \frac{1}{2(1-\rho_g^2) a_{g,3}} \left(a_{g,1} - \frac{a_{g,0} a_{g,2}}{2a_{g,4}} \right),$$

$$\sigma_{\beta_g}^2 = \frac{1}{2(1-\rho_g^2) a_{g,3}}.$$

A gamma prior distribution is assigned to δ^2 and a truncated normal prior distribution conditional on δ_g is assigned to γ_g, i.e.,

$$\delta_g^2 \sim \mathrm{Gamma}\left(\frac{a_{g,0}^{(0)}}{2} + 1, a_{g,4}^{(0)} - \frac{a_{g,0}^{(0)\,2}}{4a_{g,3}^{(0)}} \right)$$

and

$$\gamma_g \mid \delta_g \sim \mathrm{N}\left(\frac{a_{g,0}^{(0)} \delta_g}{2a_{g,3}^{(0)}}, \frac{1}{2a_{g,3}^{(0)}} \right) I(\gamma_g > 0).$$

The resulting posterior distribution for (δ_g, γ_g) is given by

$$\delta_g^2 \sim \mathrm{Gamma}\left(\frac{a_{g,0}}{2} + 1, a_{g,4} - \frac{a_{g,0}^2}{4a_{g,3}} \right)$$

and

$$\gamma_g \mid \delta_g \sim \mathrm{N}\left(\frac{a_{g,0}\delta_g}{2a_{g,3}}, \frac{1}{2a_{g,3}}\right) I(\gamma_g > 0).$$

For the variational approximation, $h(\boldsymbol{\theta}, \mathbf{w} \mid \mathbf{y})$ is taken to have a factorized form, $q_{\boldsymbol{\theta},\mathbf{w}}(\boldsymbol{\theta}, \mathbf{w}) = q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) q_{\mathbf{w}}(\mathbf{w}) = q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) q_{\mathbf{z},\mathbf{u}}(\mathbf{z}, \mathbf{u})$. Following Beal (2003), $q_{\mathbf{z}_g,\mathbf{u}_g}(\mathbf{z}_g = \mathbf{1}, \mathbf{u}_g)$ for the conjugate-exponential models can be obtained as

$$q_{\mathbf{z}_g,\mathbf{u}_g}(\mathbf{z}_g = \mathbf{1}, \mathbf{u}_g) = \prod_{i=1}^{n} q_{z_{ig},u_{ig}}(z_{ig} = 1, u_{ig})$$

and

$$\begin{aligned} q_{z_{ig},u_{ig}}(z_{ig} = 1, u_{ig}) &= \frac{1}{\mathscr{Z}_{z_{ig},u_{ig}}} \exp\left\{ \int_{\theta} \log\ p(y_i, z_{ig} = 1, u_{ig} | \boldsymbol{\theta}_g) q_{\boldsymbol{\theta}}(\boldsymbol{\theta}_g) d\boldsymbol{\theta}_g \right\} \\ &= \frac{1}{\mathscr{Z}_{z_{ig},u_{ig}}} \exp\left\{ \mathbb{E}[\log\ p(y_i, z_{ig} = 1, u_{ig} | \boldsymbol{\theta}_g)]_{q_{\boldsymbol{\theta}}(\boldsymbol{\theta}_g)} \right\}, \end{aligned}$$

where $\mathscr{Z}_{z_{ig},u_{ig}}$ is a constant.

The log of the mixture density given the parameters $\boldsymbol{\theta}_g$ is

$$\begin{aligned} \log\ p(y_i, z_{ig} = 1, u_{ig} | \boldsymbol{\theta}_g) &= \log(\pi_g) - \log(2\pi) - 2\log(u_{ig}) + \log(\delta_g) \\ &\quad + \delta_g \gamma_g + (y_i - \mu_g)\beta_g - \frac{1}{2}\left[\left(\delta_g^2 + (y_i - \mu_g)^2\right) u^{-1} + \left(\gamma_g^2 + \beta_g^2\right) u \right]. \end{aligned}$$

Setting $A_{ig} = \delta_g^2 + (y_i - \mu_g)^2$, $B_g = \gamma_g^2 + \beta_g^2$ and $C_{ig} = \delta_g \gamma_g + (y_i - \mu_g)\beta_g$, we can write

$$\begin{aligned} \log\ p(y_i, z_{ig} = 1, u_{ig} | \boldsymbol{\theta}) &= \log(\pi_g) - \log(2\pi) - 2\log(u_{ig}) + \log(\delta_g) + C_{ig} \\ &\quad - \frac{1}{2}\left[A_{ig} u_{ig}^{-1} + B_g u_{ig} \right]. \end{aligned}$$

Hence,

$$\begin{aligned} \mathbb{E}[\log\ p(y_i, z_{ig} = 1, u_{ig} | \boldsymbol{\theta})] &= \mathbb{E}[\log(\pi_g)] - \log(2\pi) - 2\log(u_{ig}) + \mathbb{E}[\log(\delta_g)] + \mathbb{E}[C_{ig}] \\ &\quad - \frac{1}{2}\left[\mathbb{E}[A_{ig}] u_{ig}^{-1} + \mathbb{E}[B_g] u_{ig} \right]. \end{aligned}$$

Therefore,

$$\begin{aligned} q_{z_{ig},u_{ig}}(z_{ig} = 1, u_{ig}) &\propto \exp\left\{ \mathbb{E}[\log(\pi_g)] - \log(2\pi) - 2\log(u_{ig}) + \mathbb{E}[\log(\delta_g)] + \mathbb{E}[C_{ig}] \right\} \\ &\quad + \exp\left\{ -\frac{1}{2}\left[\mathbb{E}[A_{ig}] u_{ig}^{-1} + \mathbb{E}[B_g] u_{ig} \right] \right\} \\ &= (2\pi)^{-1} \exp\left\{ \mathbb{E}[\log(\pi_g)] + \mathbb{E}[\log(\delta_g)] + \mathbb{E}[C_{ig}] \right\} u_{ig}^{-2} \exp\left\{ -\frac{1}{2}\left[\mathbb{E}[A_{ig}] u_{ig}^{-1} + \mathbb{E}[B_g] u_{ig} \right] \right\} \\ &= (2\pi)^{-1} \exp\left\{ \mathbb{E}[\log(\pi_g)] + \frac{1}{2}\mathbb{E}[\log(\delta_g^2)] + \mathbb{E}[C_{ig}] \right\} \mathrm{GIG}\left(u_{ig} \,\middle|\, -1, \sqrt{\mathbb{E}[A_{ig}]}, \sqrt{\mathbb{E}[B_g]} \right). \end{aligned}$$

Here, GIG$(\cdot)$ is the probability density function of generalized inverse Gaussian distribution (Jørgensen, 1982) and

$$
\begin{aligned}
&\mathbb{E}[\log(\pi_g)] = \Psi(a_{g,0}) - \Psi(n),\\
&\mathbb{E}[\log(\delta_g^2)] = \Psi\left((a_{g,0}/2)+1\right) - \log\left(a_{g,4} - (a_{g,0}^2/4a_{g,3})\right),\\
&\mathbb{E}[A_{ig}] = \mathbb{E}[\delta_g^2] + \mathbb{E}[(y_i-\mu_g)^2] = \frac{(a_{g,0}/2)+1}{a_{g,4}-(a_{g,0}^2/4a_{g,3})} + y_i^2 - 2y_i\mathbb{E}[\mu_g] + \mathbb{E}[\mu_g^2],\\
&\mathbb{E}[B_g] = \mathbb{E}[\gamma_g^2] + \mathbb{E}[\beta_g^2] = (\mathbb{E}[\gamma_g^2])^2 + \mathrm{Var}(\gamma_g) + (\mathbb{E}[\beta_g^2])^2 + \mathrm{Var}(\beta_g),\\
&\mathbb{E}[C_{ig}] = \mathbb{E}[\delta_g\gamma_g] + \mathbb{E}[(y_i-\mu_g)\beta_g] = \mathbb{E}[\delta_g\gamma_g] + y_i\mathbb{E}[\beta_g] - (\mathbb{E}[\mu_g]\mathbb{E}[\beta_g] + \mathrm{Cov}(\mu_g,\beta_g)),
\end{aligned}
$$

where $\Psi(\cdot)$ is the digamma function.

The approximating density $q_{z_{ig}}(z_{ig}=1)$ is

$$
\begin{aligned}
q_{z_{ig}}(z_{ig}=1) &= \int_{u_{ig}} q_{z_{ig},u_{ig}}(z_{ig}=1,u_{ig})du_{ig}\\
&\propto \int_{u_{ig}} (2\pi)^{-1}\exp\left\{\mathbb{E}[\log(\pi_g)] + \frac{1}{2}\mathbb{E}[\log(\delta_g^2)] + \mathbb{E}[C_{ig}]\right\}\\
&\qquad\qquad \times \mathrm{GIG}\left(u_{ig}\;\middle|\; -1, \sqrt{\mathbb{E}[A_{ig}]}, \sqrt{\mathbb{E}[B_g]}\right)du_{ig}\\
&= (2\pi)^{-1}\exp\left\{\mathbb{E}[\log(\pi_g)] + \frac{1}{2}\mathbb{E}[\log(\delta_g^2)] + \mathbb{E}[C_{ig}]\right\}\\
&\qquad\qquad \times 2\left(\frac{\mathbb{E}[A_{ig}]}{\mathbb{E}[B_g]}\right)^{-1/2} K_{-1}\left(\sqrt{\mathbb{E}[A_{ig}]\mathbb{E}[B_g]}\right).
\end{aligned}
$$

Using the approximating density $q_{z_{ig}}(z_{ig}=1)$, the probability that observation i belongs to component g is

$$
\hat{z}_{ig} = \frac{q_{z_{ig}}(z_{ig}=1)}{\sum_{g=1}^{G} q_{z_{ig}}(z_{ig}=1)}.
$$

The approximating density $q_{u_{ig}}(u_{ig} \mid z_{ig}=1)$ is

$$
\begin{aligned}
q_{u_{ig}}(u_{ig} \mid z_{ig}=1) \;&\propto\; (2\pi)^{-1}\exp\left\{\mathbb{E}[\log(\pi_g)] + \frac{1}{2}\mathbb{E}[\log(\delta_g^2)] + \mathbb{E}[C_{ig}]\right\}\\
&\times\; 2\left(\frac{\mathbb{E}[A_{ig}]}{\mathbb{E}[B_g]}\right)^{-1/2} K_{-1}\left(\sqrt{\mathbb{E}[A_{ig}]\mathbb{E}[B_g]}\right),
\end{aligned}
$$

and so $U_{ig} \mid (z_{ig} = 1) \backsim \text{GIG}(-1, \sqrt{\mathbb{E}[A_{ig}]}, \sqrt{\mathbb{E}[B_g]})$. Therefore,

$$\begin{aligned}
\mathbb{E}[U_{ig} \mid z_{ig} = 1]_{q_{u_{ig}}(u_{ig}|z_{ig}=1)} &= \left(\frac{\mathbb{E}[A_{ig}]}{\mathbb{E}[B_g]}\right)^{1/2} \frac{K_0(\sqrt{\mathbb{E}[A_{ig}]\mathbb{E}[B_g]})}{K_{-1}(\sqrt{\mathbb{E}[A_{ig}]\mathbb{E}[B_g]})}, \\
\mathbb{E}[U_{ig}^{-1} \mid z_{ig} = 1]_{q_{u_{ig}}(u_{ig}|z_{ig}=1)} &= \left(\frac{\mathbb{E}[A_{ig}]}{\mathbb{E}[B_g]}\right)^{-1/2} \frac{K_{-2}(\sqrt{\mathbb{E}[A_{ig}]\mathbb{E}[B_g]})}{K_{-1}(\sqrt{\mathbb{E}[A_{ig}]\mathbb{E}[B_g]})}.
\end{aligned}$$

The variational Bayes algorithm proceeds in the following manner:

- For the observed data $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, the algorithm is initialized with more components than expected, say G. The $\hat{z}_{ig}$ can be initialized by either randomly assigning the observations to one of the G components or by using the results from another clustering method (e.g., k-means clustering).
- Using the initialized values of $\hat{z}_{ig}$, the parameters from the gth component are initialized as follows:
  - The component's sample mean is used to initialize the parameter μ_g,
  - β_g is set to 0, and
  - γ and δ are set to 1.
- Using these values of the parameters, the expected values of U_{ig}^{-1} and U_{ig} are initialized.
- The hyperparameters of the prior distributions are initialized to give a flat distribution over the possible values of the parameters. In our case, we chose $a_{g,j}^{(0)} = 10^{-8}$, for $j = 0, \ldots, 4$; see Section 2.3.3 for a simulation study that investigates sensitivity to our choice of 10^{-8}.

1. Using the $\hat{z}_{ig}$ and the expected values of U_{ig}^{-1} and U_{ig}, the hyperparameters of the approximating density $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ are updated. Using these updated hyperparameters, the expected values $\mathbb{E}[\log r(\boldsymbol{\theta})]$ and $\mathbb{E}[\phi_j(\boldsymbol{\theta})]$ are updated.
2. Using these updated $\mathbb{E}[\log r(\boldsymbol{\theta})]$ and $\mathbb{E}[\phi_j(\boldsymbol{\theta})]$, the $\hat{z}_{ig}$, $\mathbb{E}[U_{ig}^{-1}|z_{ig} = 1]$, and $\mathbb{E}[U_{ig}|z_{ig} = 1]$ are updated.
3. Components with too few observations are eliminated. Specifically, for each component g' we do the following. If the estimated number of observations in component g', i.e., $\sum_{i=1}^{n} \hat{z}_{ig'}$, is sufficiently small (less than one in our case), then component g' is eliminated.

Steps 1, 2, and 3 are repeated until convergence.

Once convergence is achieved, the observations are assigned to clusters using maximum *a posteriori* probability (MAP), such that $\text{MAP}(\hat{z}_{ig}) = 1$ if $\max_g(\hat{z}_{ig})$ occurs in component g and $\text{MAP}(\hat{z}_{ig}) = 0$ otherwise. If the true class is known, as in our analysis, the performance of the

algorithm can be assessed using the adjusted Rand index (ARI; Hubert and Arabie, 1985). The ARI is based on the pairwise agreement between the predicted and true classifications after adjusting for agreement by chance: a value of '1' indicates a perfect classification and a value of '0' would be expected under random classification.

2.3 Simulated Data

2.3.1 Simulation Study 1

We simulated one-hundred data sets from a UNIG mixture with two components ($n_1 = 150$ and $n_2 = 150$). We chose the parameters so that the components are well separated. We ran our variational Bayes algorithm starting off with $G = 10$ components. In all one-hundred cases, our approach gave a two-component model and classification was excellent (mean ARI $= 0.99$ with std. dev. $= 0.01$). Looking at the predicted density for ten of the simulated data sets (Figure 1), it is clear that the fitted densities are capturing the data very well.

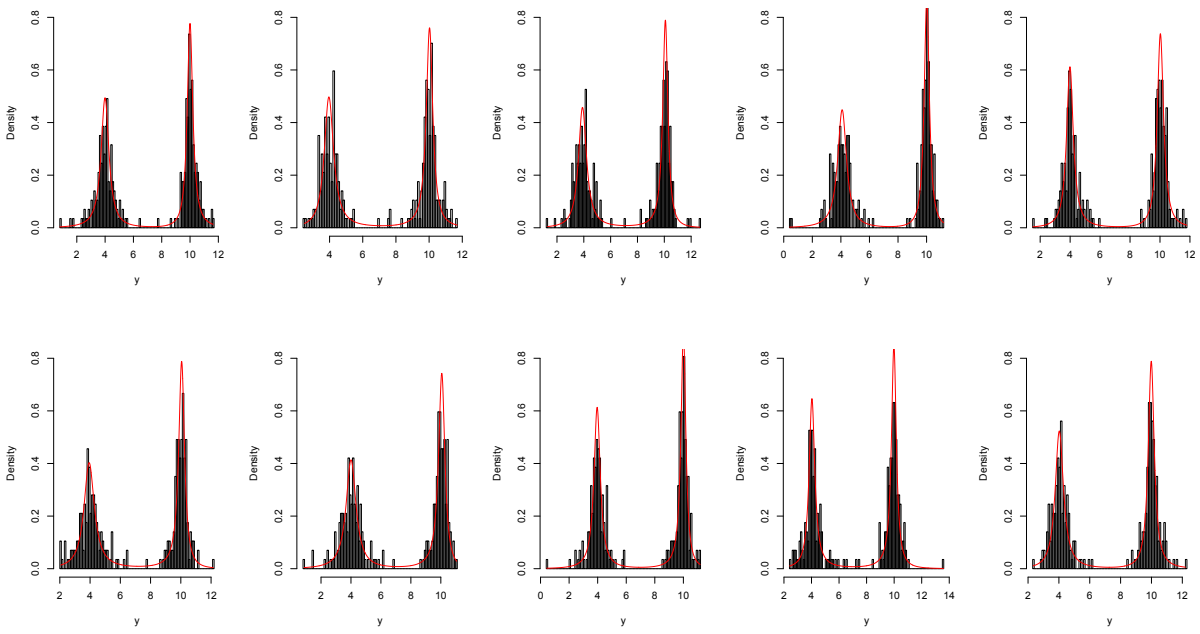

Figure 1: Histograms with fitted densities for ten data sets from Simulation 1.

2.3.2 Simulation Study 2

We simulated another one-hundred data sets from a UNIG mixture with two components ($n_1 = 150$ and $n_2 = 155$). In this case, the components were not as well separated. We again ran our

variational Bayes algorithm starting off with $G = 10$ components. Out of the one-hundred data sets, a two-component model was selected on 92 occasions and the mean ARI over all one-hundred data sets is 0.92 (with standard deviation 0.03). Figure 2 shows the fitted densities for ten of the simulated data sets; again, the fitted densities are capturing the data very well.

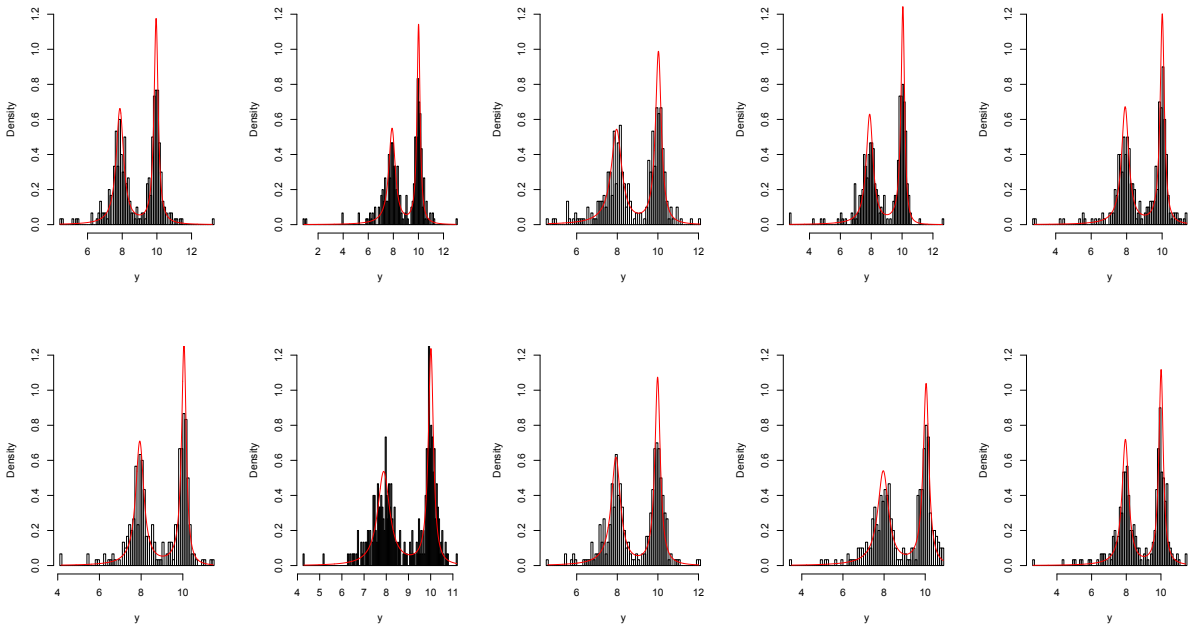

Figure 2: Histograms with fitted densities for ten data sets from Simulation 2.

2.3.3 Simulation Study 3

Recall that we initialize hyperparameters for $\boldsymbol{\theta}_g = (\pi_g, \mu_g, \beta_g, \delta_g, \gamma_g)$ so that the prior distribution of $\boldsymbol{\theta}_g$ is relatively flat. To evaluate the effect of the choice of initial values for these hyperparameters, we ran our algorithm on simulated data using 10 different initializations for these hyperparameters. Specifically, we used initial values

$$a_{0g} = a_{1g} = a_{2g} = a_{3g} = a_{4g} \in \{10^{-6}, 10^{-7}, \ldots, 10^{-15}\},$$

for $g = 1, \ldots, G$. For each of the ten runs, the data and the initial $\hat{z}_{ig}$ were the same so that only the initial values of the hyperparameters differed. The classification results obtained from all ten different initial values for the hyperparameters are identical (ARI $= 0.99$), and the fitted densities for all ten runs are virtually identical (Figure 3).

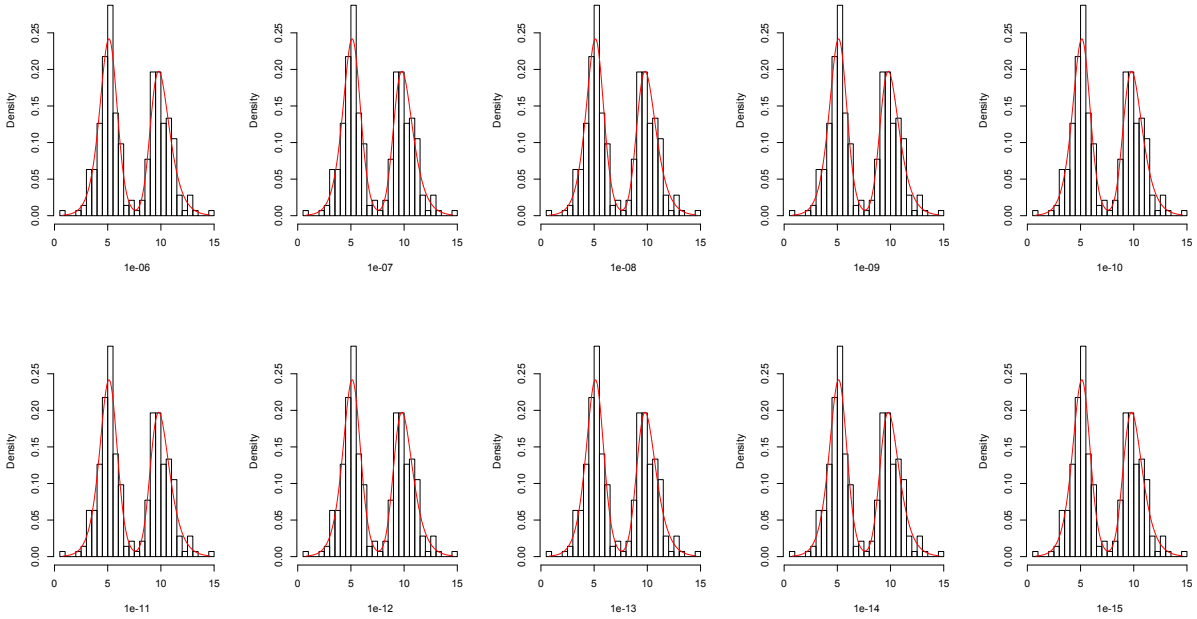

Figure 3: Histograms with fitted densities for the ten data sets from Simulation 3, where the label of each x-axis reflects the initial values for the hyperparameters.

2.4 Enzyme Data Set

We considered the enzyme data set, which is a benchmark data set for a mixture of univariate distributions with a skewed component (Bechtel et al., 1993; Karlis and Santourian, 2009). The data consist of measurements of the activity of an enzyme in the blood of 245 individuals. These data were used by Karlis and Santourian (2009) to illustrate fitting of the UNIG models within an EM algorithm framework. Their EM algorithm, in conjunction with a model selection criterion, resulted in the selection of a two-component UNIG model. We used our variational Bayes approach to fit the UNIG models, initializing at $G = 5$ components. Akin to Karlis and Santourian (2009), we obtained a two-component model that clearly gives a good fit to the data (Figure 4).

3 Mixture of Multivariate Normal Inverse Gaussian Distributions

3.1 The Model

A mean-variance mixture of a d-dimensional multivariate normal distribution with the inverse Gaussian distribution, i.e.,

$$\mathbf{Y} \mid w \sim \mathrm{N}(\boldsymbol{\mu} + w\boldsymbol{\beta}\boldsymbol{\Delta}, w\boldsymbol{\Delta}), \qquad W \sim \mathrm{IG}(\delta, \gamma)$$

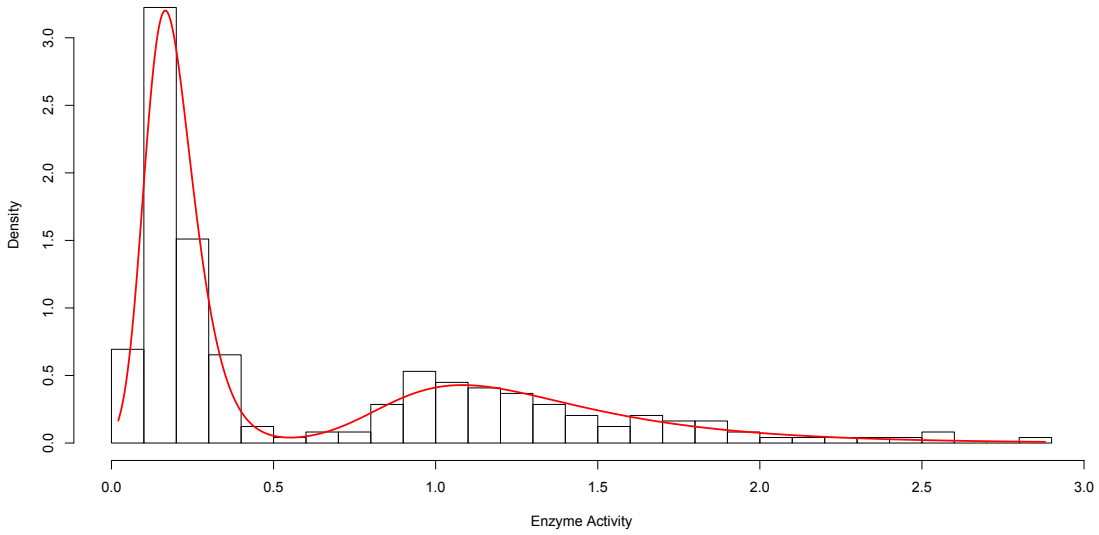

Figure 4: Histogram with fitted density for the enzyme data.

results in a MNIG distribution with density

$$f(\mathbf{y};\theta) = \frac{\delta}{2^{\frac{d-1}{2}}} \exp\left\{\delta\gamma + (\mathbf{y}-\boldsymbol{\mu})\boldsymbol{\beta}'\right\} \left[\frac{\alpha}{\pi q(\mathbf{y})}\right]^{\frac{d+1}{2}} K_{\frac{d+1}{2}}(\alpha q(\mathbf{y})),$$

where $\alpha^2 = \gamma^2 + \boldsymbol{\beta}\boldsymbol{\Delta}\boldsymbol{\beta}'$, $q(\mathbf{y})^2 = \delta^2 + (\mathbf{y}-\boldsymbol{\mu})\boldsymbol{\Delta}^{-1}(\mathbf{y}-\boldsymbol{\mu})'$, and $K_r(\mathbf{y})$ is the modified Bessel function of the third kind of order r evaluated at $\mathbf{y}$. Similar to the univariate case, the parameters contribute to the different shapes the MNIG can have. Here, $\boldsymbol{\Delta}$ is a $d \times d$ symmetric positive definite matrix that relates to the covariance matrix via

$$\text{Cov}(\mathbf{Y}) = \frac{\delta}{\gamma^3}(\gamma^2\boldsymbol{\Delta} + \boldsymbol{\Delta}\boldsymbol{\beta}'\boldsymbol{\beta}\boldsymbol{\Delta}),$$

and the restriction $|\boldsymbol{\Delta}| = 1$ is needed to ensure identifiability. Conjugate priors are unavailable for $\boldsymbol{\Delta}$ with this restriction ($|\boldsymbol{\Delta}| = 1$). An alternative re-parameterization, as discussed in Karlis and Santourian (2009), arises from

$$\mathbf{Y} \mid u \sim \text{N}(\tilde{\boldsymbol{\mu}} + u\tilde{\boldsymbol{\beta}}, u\tilde{\boldsymbol{\Sigma}}), \qquad U \sim \text{IG}(1, \tilde{\gamma}),$$

where $\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu}$, $\tilde{\gamma} = \gamma\delta$, $\tilde{\boldsymbol{\Sigma}} = \delta^2\boldsymbol{\Delta}$, and $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}\tilde{\boldsymbol{\Sigma}}$. Here, $\tilde{\boldsymbol{\Sigma}}$ is not restricted and conjugate priors exist for all of the model parameters.

3.2 Parameter Estimation

The joint probability density is $f(\mathbf{y},u) = f(u)f(\mathbf{y}|u)$, where

$$f(\mathbf{y}|u) = (2\pi)^{-1/2}u^{-d/2}|\tilde{\boldsymbol{\Sigma}}|^{-d/2}\exp\left\{\frac{-1}{2u}(\mathbf{y}-\tilde{\boldsymbol{\mu}}-\tilde{\boldsymbol{\beta}}u)'\tilde{\boldsymbol{\Sigma}}^{-1}(\mathbf{y}-\tilde{\boldsymbol{\mu}}-\tilde{\boldsymbol{\beta}}u)\right\}$$

and

$$f(u) = (2\pi)^{-1/2}u^{-3/2}\exp\left\{\tilde{\gamma}-\frac{1}{2}(2u^{-1}+\tilde{\gamma}^2u)\right\}.$$

Therefore,

$$\begin{aligned} f(\mathbf{y},u) &\propto u^{-\frac{d+3}{2}}|\tilde{\boldsymbol{\Sigma}}|^{-d/2}\exp\left\{-\frac{1}{2u}(\mathbf{y}-\tilde{\boldsymbol{\mu}}-\tilde{\boldsymbol{\beta}}u)'\tilde{\boldsymbol{\Sigma}}^{-1}(\mathbf{y}-\tilde{\boldsymbol{\mu}}+\tilde{\boldsymbol{\beta}}u)\right\} \\ &\times \exp\left\{-\frac{1}{2}(2u^{-1}+\tilde{\gamma}^2u-2\tilde{\gamma})\right\}. \end{aligned}$$

If $\boldsymbol{\theta} = (\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\Sigma}},\tilde{\boldsymbol{\beta}},\tilde{\gamma})$, then the likelihood of the complete MNIG data, i.e. the observed $\mathbf{y}$ and the latent $\mathbf{u}$ such that $(\mathbf{y},\mathbf{u}) = (\mathbf{y}_1,\ldots,\mathbf{y}_n,u_1,\ldots,u_n)$, has the form

$$L(\boldsymbol{\theta}) = r(\boldsymbol{\theta})^n\left(\prod_{i=1}^n h(\mathbf{y}_i,u_i)\right)\exp\left\{\sum_{j=1}^4\Phi_j(\boldsymbol{\theta})t_j(\mathbf{y},\mathbf{u})\right\},$$

which is within the exponential family. Note that, as before, $\mathbf{u} = (u_1,\ldots,u_n)$. Here, $r(\boldsymbol{\theta})$ is the normalization constant that depends on $\boldsymbol{\theta}$, $h(\mathbf{y},u)$ is a continuous function of $(\mathbf{y},u)$, $\Phi_j(\boldsymbol{\theta})$ is the jth natural parameter, and $t_j(\mathbf{y},\mathbf{u})$ is the jth sufficient statistic.

Now, consider n independent observations $\mathbf{y}_1,\ldots,\mathbf{y}_n$ from a G-component mixture of MNIG distributions. The likelihood is given by

$$L(\boldsymbol{\theta}) = \prod_{i=1}^n\sum_{g=1}^G\pi_g f(\mathbf{y};\tilde{\boldsymbol{\Sigma}}_g,\tilde{\boldsymbol{\beta}}_g,\tilde{\boldsymbol{\mu}}_g,\tilde{\gamma}_g),$$

where $f(\mathbf{y};\cdot)$ is the density of the MNIG distribution and the $\pi_g > 0$, such that $\sum_{i=1}^G\pi_g = 1$, are the mixing proportions. In this case, $(\tilde{\boldsymbol{\mu}}_g,\tilde{\boldsymbol{\beta}}_g)$ is independent of $\tilde{\gamma}_g$.

The complete-data likelihood for a G-component mixture of MNIG distributions can be written

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{g=1}^G\prod_{i=1}^n\left[\pi_g\, f(\mathbf{y}_i|u_{ig};\tilde{\boldsymbol{\mu}}_g,\tilde{\boldsymbol{\beta}}_g,\tilde{\boldsymbol{\Sigma}}_g)\, f(u_{ig};1,\tilde{\gamma}_g)\right]^{z_{ig}} \\ &= \prod_{g=1}^G\left[r(\boldsymbol{\theta}_g)^{(\sum_{i=1}^n z_{ig})}\left(\prod_{i=1}^n h(\mathbf{y}_i,u_{ig})\right)\exp\left\{\sum_{j=1}^4\Phi_j(\boldsymbol{\theta}_g)t_j(\mathbf{y},\mathbf{u}_g)\right\}\right], \end{aligned}$$

where $\mathbf{u}_g = (u_{1g}, \ldots, u_{ng})$. If the conjugate prior distribution of $\boldsymbol{\theta}_g = (\pi_g, \tilde{\boldsymbol{\mu}}_g, \tilde{\Sigma}_g, \tilde{\boldsymbol{\beta}}_g, \tilde{\gamma}_g)$ is of the form

$$h(\boldsymbol{\theta}_g) \propto r(\boldsymbol{\theta}_g)^{a_{g,0}^{(0)}} \exp\left\{ \sum_{j=1}^{4} \Phi_j(\boldsymbol{\theta}_g) a_{g,j}^{(0)} \right\},$$

with initial hyperparameters $\{a_{g,0}^{(0)}, \mathbf{a}_{g,1}^{(0)}, \mathbf{a}_{g,2}^{(0)}, a_{g,3}^{(0)}, a_{g,4}^{(0)}\}$, then the posterior distribution is of the form

$$h(\boldsymbol{\theta}_g \mid \mathbf{y}) \propto r(\boldsymbol{\theta}_g)^{(a_{g,0}^{(0)} + \sum_{i=1}^{n} z_{ig})} \exp\left\{ \sum_{j=1}^{4} \Phi_j(\boldsymbol{\theta}_g) \left(\mathbf{a}_{g,j}^{(0)} + t_j(\mathbf{y}, \mathbf{u}_g) \right) \right\},$$

where the hyperparameters of the posterior distributions of the mixtures of MNIG models are

$$\begin{aligned}
a_{g,0} &= a_{g,0}^{(0)} + \textstyle\sum_{i=1}^{n} z_{ig}, & \mathbf{a}_{g,1} &= \mathbf{a}_{g,1}^{(0)} + \sum_{i=1}^{n} z_{ig} \mathbf{y}_i, \\
\mathbf{a}_{g,2} &= \mathbf{a}_{g,2}^{(0)} + \textstyle\sum_{i=1}^{n} z_{ig} u_{ig}^{-1} \mathbf{y}_i, & a_{g,3} &= a_{g,3}^{(0)} + \sum_{i=1}^{n} z_{ig} u_{ig}, \\
a_{g,4} &= a_{g,4}^{(0)} + \textstyle\sum_{i=1}^{n} z_{ig} u_{ig}^{-1}.
\end{aligned}$$

Note that $\mathbf{a}_{g,j}^{(0)}$ and $\mathbf{a}_{g,j}$ for $j = 1,2$ are vectors, and $a_{g,j}^{(0)}$ and $a_{g,j}$ for $j \in \{0,3,4\}$ are scalars. When referring to a value $j \in \{0,1,2,3,4\}$, we write $a_{g,j}^{(0)}$ and $a_{g,j}$.

A Dirichlet prior with initial hyperparameters $(a_{1,0}^{(0)}, \ldots, a_{G,0}^{(0)})$ is assigned to the mixing proportions and results in a Dirichlet posterior distribution with hyperparameters $(a_{1,0}, \ldots, a_{G,0})$.

A Wishart prior was assigned to the precision matrix of the gth component, i.e., $\tilde{\boldsymbol{\Sigma}}_g^{-1} \backsim$ Wishart $(a_{g,0}, \mathbf{V}_g^{(0)})$, resulting in a Wishart posterior $\tilde{\boldsymbol{\Sigma}}_g^{-1} \backsim$ Wishart $(a_{g,0}, \mathbf{V}_g')$ with

$$\begin{aligned}
\mathbf{V}_g &= \mathbf{V}_g^{(0)} + \sum_{i=1}^{n} \hat{z}_{ig} u_{ig}^{-1} \mathbf{y}'\mathbf{y} - \mathbf{a}_{g,2}' \tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}' \mathbf{a}_{g,2} + a_{g,4} \tilde{\boldsymbol{\mu}}' \tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{\beta}}' \mathbf{a}_{g,1} + a_{g,0} \tilde{\boldsymbol{\beta}}' \tilde{\boldsymbol{\mu}} - \mathbf{a}_{g,1}' \tilde{\boldsymbol{\beta}} \\
&+ a_{g,0} \tilde{\boldsymbol{\mu}}' \tilde{\boldsymbol{\beta}} + a_{g,3} \tilde{\boldsymbol{\beta}}' \tilde{\boldsymbol{\beta}}.
\end{aligned}$$

A correlated multivariate Gaussian prior distirbution conditional on the precision matrix was assigned to $(\tilde{\boldsymbol{\mu}}_g, \tilde{\boldsymbol{\beta}}_g)$ such that

$$\begin{pmatrix} \tilde{\boldsymbol{\mu}}_g \\ \tilde{\boldsymbol{\beta}}_g \end{pmatrix} \Bigg| \tilde{\boldsymbol{\Sigma}}^{-1} \sim \mathrm{N}\left[\begin{pmatrix} \bar{\boldsymbol{\mu}}_g \\ \bar{\boldsymbol{\beta}}_g \end{pmatrix}, \begin{pmatrix} \tilde{\boldsymbol{\Sigma}}_{\mu_g}^{-1} & \tilde{\boldsymbol{\Sigma}}_{\mu_g \beta_g}^{-1} \\ \tilde{\boldsymbol{\Sigma}}_{\mu_g \beta_g}^{-1}, & \tilde{\boldsymbol{\Sigma}}_{\beta_g}^{-1} \end{pmatrix}^{-1} \right],$$

where

$$\tilde{\boldsymbol{\Sigma}}_{\mu_g}^{-1} = a_{g,4}^{(0)}\tilde{\boldsymbol{\Sigma}}_g^{-1}, \qquad \bar{\boldsymbol{\mu}}_g = \frac{a_{g,3}^{(0)}}{(a_{g,3}^{(0)}a_{g,4}^{(0)} - a_{g,0}^{(0)\,2})}\left(\mathbf{a}_{g,2}^{(0)} - \frac{\mathbf{a}_{g,1}^{(0)}a_{g,0}^{(0)}}{a_{g,3}^{(0)}}\right),$$

$$\tilde{\boldsymbol{\Sigma}}_{\beta_g}^{-1} = a_{g,3}^{(0)}\tilde{\boldsymbol{\Sigma}}_g^{-1}, \qquad \bar{\boldsymbol{\beta}}_g = \frac{a_{g,4}^{(0)}}{(a_{g,3}^{(0)}a_{g,4}^{(0)} - a_{g,0}^{(0)\,2})}\left(\mathbf{a}_{g,1}^{(0)} - \frac{\mathbf{a}_{g,2}^{(0)}a_{g,0}^{(0)}}{a_{g,4}^{(0)}}\right),$$

$$\tilde{\boldsymbol{\Sigma}}_{\mu_g\beta_g}^{-1} = a_{g,0}^{(0)}\tilde{\boldsymbol{\Sigma}}_g^{-1}.$$

The resulting posterior distribution for $(\tilde{\boldsymbol{\mu}}_g, \tilde{\boldsymbol{\beta}}_g)$ is given by

$$\begin{pmatrix} \tilde{\boldsymbol{\mu}}_g \\ \tilde{\boldsymbol{\beta}}_g \end{pmatrix} \Bigg| \tilde{\boldsymbol{\Sigma}}^{-1} \sim \mathrm{N}\left[\begin{pmatrix} \bar{\boldsymbol{\mu}}_g \\ \bar{\boldsymbol{\beta}}_g \end{pmatrix}, \begin{pmatrix} \tilde{\boldsymbol{\Sigma}}_{\mu_g}^{-1} & \tilde{\boldsymbol{\Sigma}}_{\mu_g\beta_g}^{-1} \\ \tilde{\boldsymbol{\Sigma}}_{\mu_g\beta_g}^{-1}, & \tilde{\boldsymbol{\Sigma}}_{\beta_g}^{-1} \end{pmatrix}^{-1}\right],$$

where

$$\tilde{\boldsymbol{\Sigma}}_{\mu_g}^{-1} = a_{g,4}\tilde{\boldsymbol{\Sigma}}_g^{-1}, \qquad \bar{\boldsymbol{\mu}}_g = \frac{a_{g,3}}{(a_{g,3}a_{g,4} - a_{g,0}^2)}\left(\mathbf{a}_{g,2} - \frac{\mathbf{a}_{g,1}a_{g,0}}{a_{g,3}}\right),$$

$$\tilde{\boldsymbol{\Sigma}}_{\beta_g}^{-1} = a_{g,3}\tilde{\boldsymbol{\Sigma}}_g^{-1}, \qquad \bar{\boldsymbol{\beta}}_g = \frac{a_{g,4}}{(a_{g,3}a_{g,4} - a_{g,0}^2)}\left(\mathbf{a}_{g,1} - \frac{\mathbf{a}_{g,2}a_{g,0}}{a_{g,4}}\right),$$

$$\tilde{\boldsymbol{\Sigma}}_{\mu_g\beta_g}^{-1} = a_{g,0}\tilde{\boldsymbol{\Sigma}}_g^{-1}.$$

A truncated normal prior distribution was assigned to $\tilde{\gamma}_g$ such that

$$\tilde{\gamma}_g \sim \mathrm{N}(a_{g,0}^{(0)}/a_{g,3}^{(0)}, 1/2a_{g,3}^{(0)})\mathrm{I}(\tilde{\gamma}_g > 0),$$

and so the posterior distribution for $\tilde{\gamma}_g$ is given by

$$\tilde{\gamma}_g \sim \mathrm{N}(a_{g,0}/a_{g,3}, 1/2a_{g,3})\mathrm{I}(\tilde{\gamma}_g > 0).$$

For the MNIG model,

$$\begin{aligned}\mathbb{E}[\log\ p(\mathbf{y}_i, z_{ig} = 1, u_{ig} \mid \boldsymbol{\theta})]_{q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} = \mathbb{E}[\log(\pi_g)] - \frac{d+1}{2}\log(2\pi) - \frac{d+3}{2}\log(u_{ig}) \\ + \frac{1}{2}\mathbb{E}[\log|\tilde{\boldsymbol{\Sigma}}_g^{-1}|)] + \mathbb{E}[\tilde{\gamma}_g] - \frac{1}{2}\left[\mathbb{E}[A_{ig}]u_{ig}^{-1} + \mathbb{E}[B_g]u_{ig}\right],\end{aligned}$$

where $A_{ig} = 1 + (\mathbf{y}_i - \tilde{\boldsymbol{\mu}}_g)'\tilde{\boldsymbol{\Sigma}}_g^{-1}(\mathbf{y}_i - \tilde{\boldsymbol{\mu}}_g)$, $B_g = \tilde{\gamma}_g^2 + \tilde{\boldsymbol{\beta}}_g\tilde{\boldsymbol{\Sigma}}_g^{-1}\tilde{\boldsymbol{\beta}}_g'$, and $C_{ig} = \tilde{\gamma}_g + (\mathbf{y}_i - \tilde{\boldsymbol{\mu}}_g)'\tilde{\boldsymbol{\Sigma}}_g^{-1}\tilde{\boldsymbol{\beta}}_g$.

Following Beal (2003), the approximating joint density of missing variables (z_{ig}, u_{ig}) for the conjugate-exponential models can be obtained as

$$\begin{aligned}
q_{z_{ig},u_{ig}}(z_{ig}=1,u_{ig}) \propto \exp\Bigg\{ & \mathbb{E}[\log(\pi_g)] - \frac{d+1}{2}\log(2\pi) - \frac{d+3}{2}\log(u_{ig}) + \frac{1}{2}\mathbb{E}[\log|\tilde{\boldsymbol{\Sigma}}_g^{-1}|)] \\
& + \mathbb{E}[C_{ig}] - \frac{1}{2}\left[\mathbb{E}[A_{ig}]u_{ig}^{-1} + \mathbb{E}[B_g]u_{ig}\right]\Bigg\} \\
= (2\pi)^{-\frac{d+1}{2}} \exp\Bigg\{ & \mathbb{E}[\log(\pi_g)] + \frac{1}{2}\mathbb{E}[\log|\tilde{\boldsymbol{\Sigma}}_g^{-1}|)] + \mathbb{E}[C_{ig}]\Bigg\} \\
& \times u_{ig}^{-\frac{d+3}{2}} \exp\left\{-\frac{1}{2}\left[\mathbb{E}[A_{ig}]u_{ig}^{-1} + \mathbb{E}[B_g]u_{ig}\right]\right\} \\
= (2\pi)^{-\frac{d+1}{2}} \exp\Bigg\{ & \mathbb{E}[\log(\pi_g)] + \frac{1}{2}\mathbb{E}[\log|\tilde{\boldsymbol{\Sigma}}_g^{-1}|)] + \mathbb{E}[C_{ig}]\Bigg\} \\
& \times \mathrm{GIG}\left(u_{ig} \,\middle|\, -\frac{d+1}{2}, \sqrt{\mathbb{E}[A_{ig}]}, \sqrt{\mathbb{E}[B_g]}\right).
\end{aligned}$$

Here, $\mathrm{GIG}(\cdot)$ is the probability density function of the generalized inverse Gaussian distribution and

$$\begin{aligned}
\mathbb{E}[\log(\pi_g)] &= \Psi(a_{g,0}) - \Psi(n), \\
\mathbb{E}[\log|\tilde{\boldsymbol{\Sigma}}_g^{-1}|] &= \sum_{s=1}^{d}\Psi\left(\frac{a_{g,0}+1-s}{2}\right) + d\log(2) - \log|\mathbf{V}_g|, \\
\mathbb{E}[A_{ig}] &= \mathbb{E}[1 + (\mathbf{y}_i - \tilde{\boldsymbol{\mu}}_g)'\tilde{\boldsymbol{\Sigma}}_g^{-1}(\mathbf{y}_i - \tilde{\boldsymbol{\mu}}_g)] \\
&= 1 + (\mathbf{y}_i - \mathbb{E}[\tilde{\boldsymbol{\mu}}_g])'\mathbb{E}[\tilde{\boldsymbol{\Sigma}}_g^{-1}](\mathbf{y}_i - \mathbb{E}[\tilde{\boldsymbol{\mu}}_g]) + \mathrm{tr}\left\{\mathbb{E}[\tilde{\boldsymbol{\Sigma}}_g^{-1}]\mathrm{Var}(\tilde{\boldsymbol{\mu}}_g)\right\}, \\
\mathbb{E}[B_g] &= \mathbb{E}[\tilde{\gamma}_g^2 + \tilde{\boldsymbol{\beta}}_g\tilde{\boldsymbol{\Sigma}}_g^{-1}\tilde{\boldsymbol{\beta}}_g'] \\
&= (\mathbb{E}[\tilde{\gamma}_g])^2 + \mathrm{Var}(\tilde{\gamma}_g) + \mathbb{E}[\tilde{\boldsymbol{\beta}}_g]\mathbb{E}[\tilde{\boldsymbol{\Sigma}}_g^{-1}]\mathbb{E}[\tilde{\boldsymbol{\beta}}_g] + \mathrm{tr}\left\{\mathbb{E}[\tilde{\boldsymbol{\Sigma}}^{-1}]\mathrm{Var}(\tilde{\boldsymbol{\beta}}_g)\right\}, \\
\mathbb{E}[C_{ig}] &= \mathbb{E}[\tilde{\gamma}_g + (\mathbf{y}_i - \tilde{\boldsymbol{\mu}}_g)'\tilde{\boldsymbol{\Sigma}}_g^{-1}\tilde{\boldsymbol{\beta}}_g] \\
&= \mathbb{E}[\tilde{\gamma}_g] + \mathbf{y}_i\mathbb{E}[\tilde{\boldsymbol{\Sigma}}_g^{-1}]\mathbb{E}[\tilde{\boldsymbol{\beta}}_g] - \mathbb{E}[\tilde{\boldsymbol{\mu}}_g']\mathbb{E}[\tilde{\boldsymbol{\Sigma}}_g^{-1}]\mathbb{E}[\tilde{\boldsymbol{\beta}}_g] + \mathrm{tr}\left\{\mathbb{E}[\tilde{\boldsymbol{\Sigma}}^{-1}]\mathrm{Cov}(\tilde{\boldsymbol{\mu}}_g, \tilde{\boldsymbol{\beta}}_g)\right\},
\end{aligned}$$

where $\Psi(\cdot)$ is the digamma function.

The approximating density $q_{z_{ig}}(z_{ig}=1)$ is

$$\begin{aligned}
q_{z_{ig}}(z_{ig}=1) &= \int_{u_{ig}} q_{z_{ig},u_{ig}}(z_{ig}=1,u_{ig})du_{ig} \\
&\propto \int_{u_{ig}} (2\pi)^{-\frac{d+1}{2}} \exp\left\{\mathbb{E}[\log(\pi_g)] + \frac{1}{2}\mathbb{E}[\log|\tilde{\boldsymbol{\Sigma}}_g^{-1}|] + \mathbb{E}[C_{ig}]\right\} \\
&\qquad \times \text{GIG}\left(u_{ig} \;\middle|\; -\frac{d+1}{2}, \sqrt{\mathbb{E}[A_{ig}]}, \sqrt{\mathbb{E}[B_g]}\right) du_{ig} \\
&= (2\pi)^{-\frac{d+1}{2}} \exp\left\{\mathbb{E}[\log(\pi_g)] + \frac{1}{2}\mathbb{E}[\log|\tilde{\boldsymbol{\Sigma}}_g^{-1}|] + \mathbb{E}[C_{ig}]\right\} \\
&\qquad \times 2\left(\frac{\mathbb{E}[A_{ig}]}{\mathbb{E}[B_g]}\right)^{-\frac{d+1}{2}} K_{-\frac{d+1}{2}}\left(\sqrt{\mathbb{E}[A_{ig}]\mathbb{E}[B_g]}\right).
\end{aligned}$$

The probability that $z_{ig}=1$ is

$$\hat{z}_{ig} = \frac{q_{z_{ig}}(z_{ig}=1)}{\sum_{g=1}^{G} q_{z_{ig}}(z_{ig}=1)}.$$

The density $q_{u_{ig}}(u_{ig} \mid z_{ig}=1)$ is

$$\begin{aligned}
q_{u_{ig}}(u_{ig} \mid z_{ig}=1) \;\propto\; & (2\pi)^{-\frac{d+1}{2}} \exp\left\{\mathbb{E}[\log(\pi_g)] + \frac{1}{2}\mathbb{E}[\log|\tilde{\boldsymbol{\Sigma}}_g^{-1}|] + \mathbb{E}[C_{ig}]\right\} \\
& \times 2\left(\frac{\mathbb{E}[A_{ig}]}{\mathbb{E}[B_g]}\right)^{-(d+1)/2} K_{-\frac{d+1}{2}}\left(\sqrt{\mathbb{E}[A_{ig}]\mathbb{E}[B_g]}\right),
\end{aligned}$$

and so $U_{ig} \mid (z_{ig}=1) \backsim \text{GIG}(-\frac{d+1}{2}, \sqrt{\mathbb{E}[A_{ig}]}, \sqrt{\mathbb{E}[B_g]})$. Therefore,

$$\begin{aligned}
\mathbb{E}[U_{ig}|z_{ig}=1]_{q_{u_{ig}}(u_{ig}|z_{ig}=1)} &= \left(\frac{\mathbb{E}[A_{ig}]}{\mathbb{E}[B_g]}\right)^{\frac{d+1}{2}} \frac{K_{-\frac{d-1}{2}}(\sqrt{\mathbb{E}[A_{ig}]\mathbb{E}[B_g]})}{K_{-\frac{d+1}{2}}(\sqrt{\mathbb{E}[A_{ig}]\mathbb{E}[B_g]})}, \\
\mathbb{E}[U_{ig}^{-1}|z_{ig}=1]_{q_{u_{ig}}(u_{ig}|z_{ig}=1)} &= \left(\frac{\mathbb{E}[A_{ig}]}{\mathbb{E}[B_g]}\right)^{-\frac{d+1}{2}} \frac{K_{-\frac{d+3}{2}}(\sqrt{\mathbb{E}[A_{ig}]\mathbb{E}[B_g]})}{K_{-\frac{d+1}{2}}(\sqrt{\mathbb{E}[A_{ig}]\mathbb{E}[B_g]})}.
\end{aligned}$$

Similar to the univariate approach, the variational Bayes algorithm proceeds in the following manner:

- For the observed data $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)$, the algorithm is initialized with more components than expected, say G. The $\hat{z}_{ig}$ can be initialized by either randomly assigning the observations to one of the G components or by using the results from another clustering method (e.g., k-means clustering).
- Using the initialized values of $\hat{z}_{ig}$, the parameter of the gth component is initialized as follows:

- The component's sample mean is used to initialize the parameter $\boldsymbol{\mu}_g$,
- the component's sample covariance is used to initialize $\boldsymbol{\Sigma}_g$,
- $\boldsymbol{\beta}$ is set to $\mathbf{0}$, and
- γ is set to 1.

- Using these values of the parameters, the expected values of U_{ig}^{-1} and U_{ig} are initialized.
- The hyperparameters of the prior distribution are initialized to have a flat distribution over the possible values of the parameters.

1. Using the $\hat{z}_{ig}$ and expected values of U_{ig}^{-1} and U_{ig}, the hyperparameters of the approximating density $q_\theta(\boldsymbol{\theta})$ are updated. Using these updated hyperparameters, the expected values $\mathbb{E}[\log r(\boldsymbol{\theta})]$ and $\mathbb{E}[\phi_j(\boldsymbol{\theta})]$ are updated.
2. Using these updated $\mathbb{E}[\log r(\boldsymbol{\theta})]$ and $\mathbb{E}[\phi_j(\boldsymbol{\theta})]$, the $\hat{z}_{ig}$, $\mathbb{E}[U_{ig}^{-1}|z_{ig}=1]$, and $\mathbb{E}[U_{ig}|z_{ig}=1]$ are updated.
3. Components with too few observations are eliminated. Specifically, for each component g' we do the following. If the estimated number of observations in component g', i.e., $\sum_{i=1}^{n}\hat{z}_{ig'}$, is sufficiently small (less than one in our case), then component g' is eliminated.

Steps 1, 2, and 3 are repeated until convergence.

As in the univariate case, after convergence is achieved, the observations are assigned to components using the MAP.

3.3 Simulated Data

3.3.1 Simulation Study 4

To demonstrate the recovery of underlying parameters, our variational Bayes algorithm was applied to a simulated two-dimensional data set (Figure 5) with two symmetric components ($n_1 = 150$ and $n_2 = 200$). Our algorithm was initialized with $G = 5$ components and, after running to convergence, gave a two-component model with one misclassified observation (ARI $= 0.99$). The estimated parameters are very close to the true values, as can be seen in Table 1 and below:

$$\tilde{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 1.2 & 0 \\ 0 & 1.2 \end{pmatrix}, \hat{\tilde{\boldsymbol{\Sigma}}}_1 = \begin{pmatrix} 0.75 & 0.05 \\ 0.05 & 0.68 \end{pmatrix}; \tilde{\boldsymbol{\Sigma}}_2 = \begin{pmatrix} 1 & 0.4 \\ 0.4 & 1 \end{pmatrix}, \hat{\tilde{\boldsymbol{\Sigma}}}_2 = \begin{pmatrix} 1.32 & -0.35 \\ -0.35 & 1.28 \end{pmatrix}.$$

Table 1: Estimated and true values for the parameters of the MNIG model in Simulation Study 4.

		$\tilde{\boldsymbol{\beta}}$	$\tilde{\gamma}$	$\tilde{\boldsymbol{\mu}}$
Component 1	True	$(0.1, 0.2)$	1.2	$(-2, -10)$
	Estimated	$(0.17, 0.49)$	1.63	$(-2.39, -10.35)$
Component 2	True	$(0.2, 0.75)$	0.8	$(-10, -12)$
	Estimated	$(0.53, 1.00)$	0.69	$(-10.00, -11.89)$

Our model clearly fits the data very well, with the contours capturing the shape of the two components (Figure 5). The parameter estimates must be considered in context with the actual fit of the model (Figure 5) because it is known that different parameter sets can give very similar densities for these models (Lillestol, 2000).

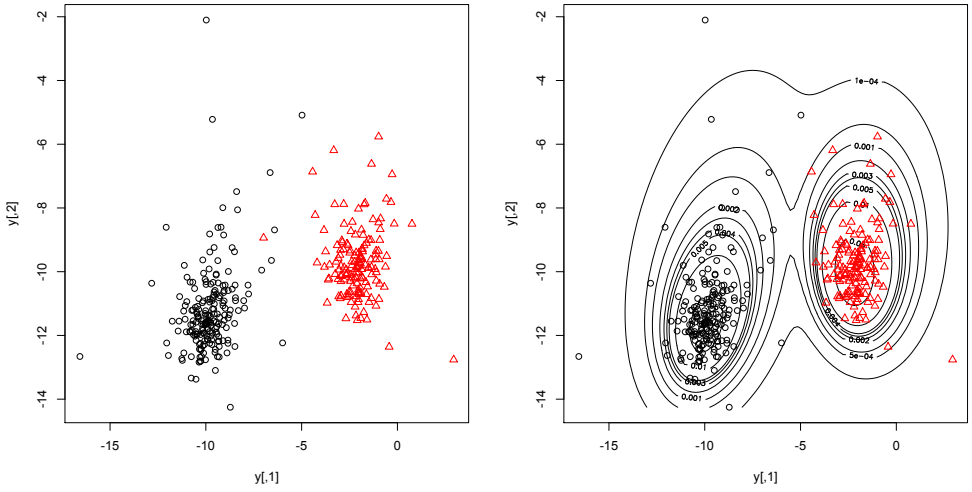

Figure 5: Scatter plot highlighting the true labels for the simulated data from Simulation Study 4 (left) and a contour plot showing the predicted classifications (right).

3.3.2 Simulation Study 5

To present a more challenging and higher dimensional example, we generated a ten-dimensional data set with two components ($n_1 = 150$ and $n_2 = 200$) that are not well separated (Figure 6). The variational Bayes algorithm was run starting with $G = 10$ components, resulting in a two-component model with perfect classification (ARI $= 1$).

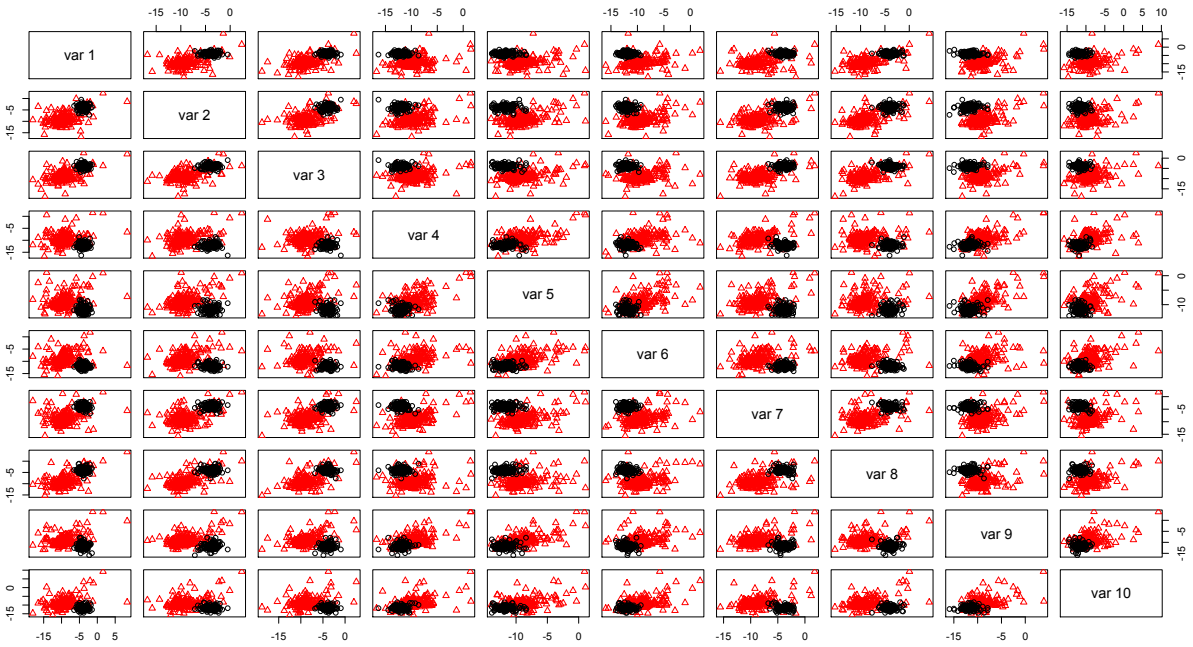

Figure 6: Pairs plot showing the true classifications for the data from Simulation 5.

3.4 Old Faithful Data

The Old Faithful data are available in the R package `MASS` (Venables and Ripley, 2002). These data comprise the waiting time between and the duration of 272 eruptions of the Old Faithful geyser in Yellowstone National Park, Wyoming, USA. These data do not contain true labels; however, upon visual inspection, the data seem to have two classes: shorter, more frequent eruptions and longer, less frequent eruptions. We ran our variational Bayes algorithm on the Old Faithful data starting with $G = 7$ components. The resulting $G = 2$ component model fits the data very well (Figure 7).

Several others have used non-Gaussian model-based clustering on these data, via variants of the EM algorithm, and obtained similar results (e.g., Franczak et al., 2012; Vrbik and McNicholas, 2012).

3.5 Crabs data

The crabs data, available in the R package `MASS` (Venables and Ripley, 2002), contain morphological measurements of 50 male and 50 female crabs (*Leptograpsus variegatus*) in each of the two colour forms: blue and orange. The measured morphological variables are frontal lobe size (FL), rear width (RW), carapace length (CL), carapace width (CW), and body depth (BD), all measured in mm. We ran our variational Bayes algorithm on the crab data starting with $G = 10$ components. This resulted in a $G = 4$ component model with an associated ARI of 0.79 (Table 2). The variables

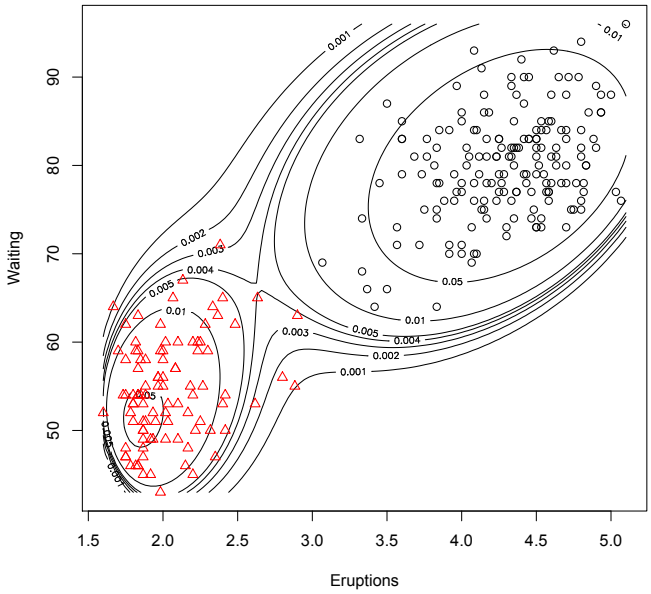

Figure 7: Contour plot of the Old Faithful data using the fitted two-component mixture model.

in these data are highly correlated and so visualizing classification performance is easier in the principal component space (Figure 8).

Table 2: The new merged "true" classification and the estimated classification.

True \ Estimates	1	2	3	4
Blue&Male	41	9		
Blue&Female	1	48		1
Orange&Male			50	
Orange&Female			6	44

Because the variables in these data are so highly correlated, they are very difficult to cluster. In fact, they are notoriously difficult to cluster and our variational Bayes MNIG performs very well when compared to many other approaches. To find methods in the literature that produce a higher ARI than 0.79, one needs to look at families of mixture models (e.g., McNicholas et al., 2010; Andrews and McNicholas, 2012; Vrbik and McNicholas, ress) or methods that combine variable selection and clustering (e.g., Morris et al., 2013). Of course, if our variational Bayes MNIG approach was extended to incorporate a family of models, an ARI above 0.79 might be achieved (cf. Section 4).

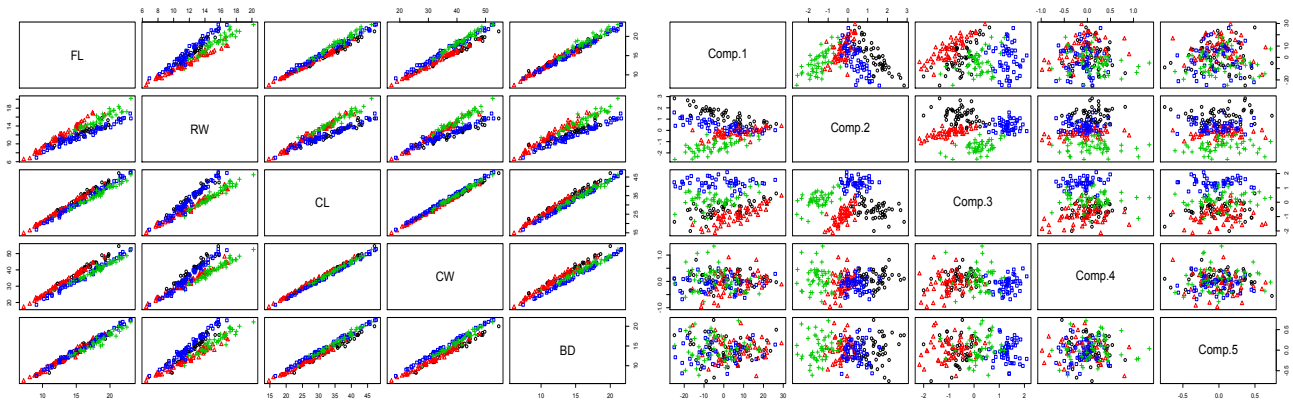

Figure 8: Pairs plots showing estimated classifications for the crabs data using the variables (left) and the principal component (right), respectively.

3.6 Fish Catch Data

The fish catch data, available in the R package `rrcov` (Todorov and Filzmoser, 2009), contain different measurements on the body size and weight of seven different fish species (bream, whitewish, roach, parkki, smelt, pike, and perch). The variable `Weight` gives the weight of the fish in grams, `Length1` is the length from the nose to the beginning of the tail, `Length2` is the length from the nose to the notch of the tail, `Length3` is the length from the nose to the end of the tail, `Height` is the maximal height as a percentage of `Length3`, and `Width` is maximal width as a percentage of `Length3`.

As expected, all of the length measurements are very highly correlated with each other (correlation > 0.99) and with the weight measurements (correlation > 0.91), cf. Figure 9. Therefore, the highly correlated variables `Length1`, `Length2`, and `Weight` were dropped from further analysis. These data were explored by Karlis and Santourian (2009), who dropped `Length1`, `Length2`, and `Height` before their analysis.

We ran our variational Bayes algorithm on the resulting three-dimensional data set, starting with $G = 10$ components. This resulted in a $G = 4$ component model with the classifications shown in Figure 10. By inspection of Figure 10, we can see that a four-component solution is not unreasonable based on the three variables used (`Length2`, `Height`, and `Weight`). Classification results for our four-component model (Table 3) correspond exactly to a merging of Species 2, 3, and 7 (whitewish, roach, and perch), and Species 1 and 4 (bream and parkki). Karlis and Santourian (2009) who used different variables — `Length3`, `Weight`, and `Width` — in an EM framework, obtained a seven-component model using the Akaike information criterion (AIC; Akaike, 1973) for model selection and a four-component model using the Bayesian information criterion (BIC; Schwarz, 1978).

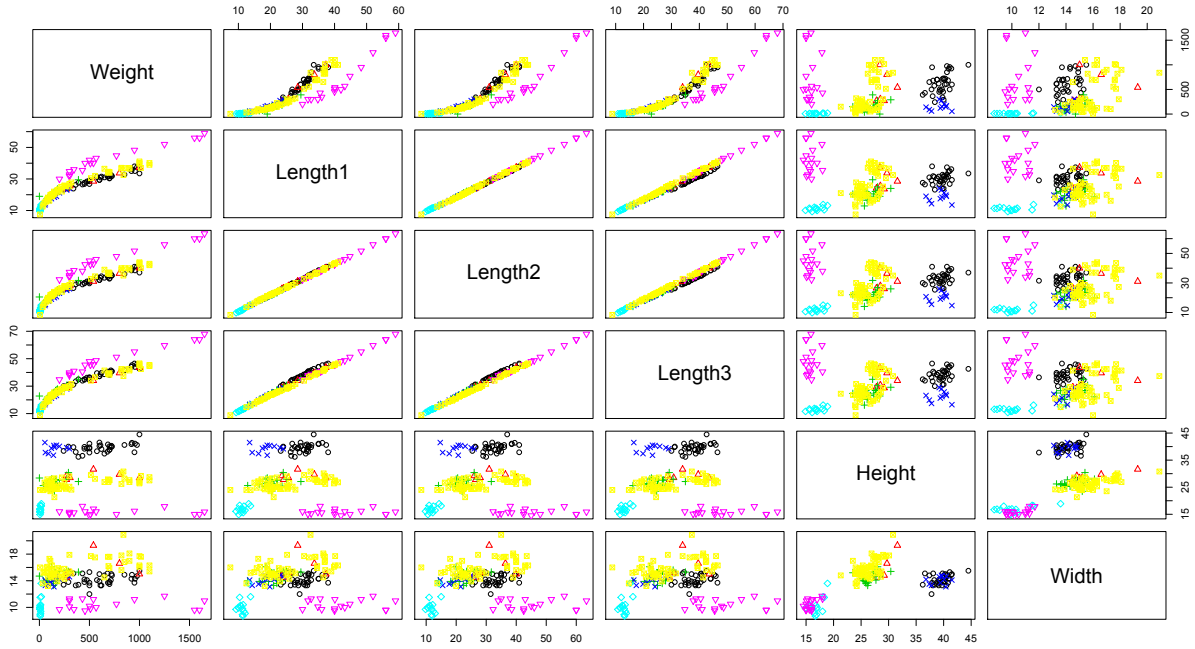

Figure 9: Matrix scatter plot of all variables in the fish catch data set, where different colours represent different species.

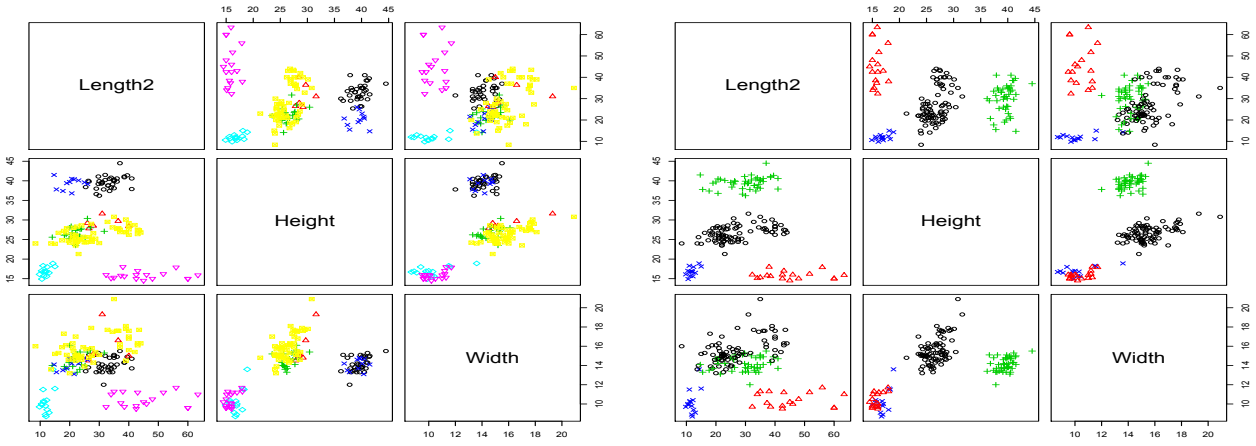

Figure 10: True (left) and estimated (right) classifications for the fish catch data based on the variables `Length2`, `Height`, and `Width`.

Table 3: Cross-tabulation of true versus predicted classifications for the fish catch data.

True\ Estimates	1	2	3	4
Bream	34			
Parkki	11			
Whitewish		6		
Roach		20		
Perch		56		
Smelt			14	
Pike				17

4 Conclusion

Variational Bayes approximations are presented as an effective alternative to the EM algorithm for parameter estimation for UNIG and MNIG mixtures. They have been used for Gaussian mixture models in the past; however, this is their first application for non-Gaussian models. Furthermore, it is the first time variational approximations have been used for non-symmetric distributions. Although we illustrated our variational Bayes approach through model-based clustering, it could be applied to model-based classification (e.g., McNicholas, 2010) or discriminant analysis (Hastie and Tibshirani, 1996) in an analogous fashion. In this paper, we illustrate that variational Bayes approximations can be very effective for non-Gaussian mixture model-based clustering, classification, and discriminant analysis. Accordingly, this paper may well be the forerunner to several others detailing the application of variational Bayes approximations in the complex modelling situations that can arise in non-Gaussian model-based clustering.

As reported by Karlis and Santourian (2009), the EM algorithm for MNIG takes a very long time to converge. Therefore, running multiple EM algorithms to cover a range of values for G, which is needed when the true number of components is unknown, adds to an already heavy computational burden. Variational Bayes approximations, on the other hand, start off with more components than expected, and once the number of observations in a component becomes sufficiently small, it is removed. This allows for simultaneous parameter estimation and estimation of the number of components, and is far more computationally efficient than running an EM algorithm for each of several possible values of G.

We demonstrated the efficacy of our approach by clustering real and simulated data for both the UNIG and MNIG mixtures. Some possible avenues for further research include extending these models, and the associated parameter estimation approach, to achieve parsimony. The could be carried out via imposing constraints upon an eigen-decomposition of the component scale matrices by analogy with the work of Celeux and Govaert (1995) on Gaussian mixtures. Our variational approximations could be extended to mixtures of factor analyzers (Ghahramani and Hinton, 1997; McLachlan and Peel, 2000), or a variatiants thereof (McNicholas and Murphy, 2008, 2010; Baek et al., 2010; Murray et al., 2013), where it may be possible to select the number of latent factors

in addition to the number of components. Applications aimed at longitudinal data analysis (e.g. McNicholas and Murphy, 2010; McNicholas and Subedi, 2012) and contaminated mixtures (Punzo and McNicholas, 2013) will also be considered within a variational framework.

Acknowledgements

This work was supported by a Postgraduate Scholarship from the Natural Sciences and Engineering Research Council of Canada, an Early Researcher Award from the Ontario Ministry of Research and Innovation, and the University Research Chair in Computational Statistics.

References

Abramowitz, M. and I. Stegun (1972). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables* (9th edition ed.). New York: NY: Dover Press.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second international symposium on information theory*, Volume 1, pp. 267–281. Springer Verlag.

Andrews, J. L. and P. D. McNicholas (2011). Extending mixtures of multivariate t-factor analyzers. *Statistics and Computing 21*(3), 361–373.

Andrews, J. L. and P. D. McNicholas (2012). Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions. *Statistics and Computing 22*(5), 1021–1029.

Andrews, J. L., P. D. McNicholas, and S. Subedi (2011). Model-based classification via mixtures of multivariate t-distributions. *Computational Statistics and Data Analysis 55*, 520–529.

Baek, J. and G. J. McLachlan (2011). Mixtures of common t-factor analyzers for clustering high-dimensional microarray data. *Bioinformatics 27*, 1269–1276.

Baek, J., G. J. McLachlan, and L. K. Flack (2010). Mixtures of factor analyzers with common factor loadings: Applications to the clustering and visualization of high-dimensional data. *IEEE Transactions on Pattern Analysis and Machine Intelligence 32*(7), 1298–1309.

Barndorff-Nielsen, O. E. (1997). Normal inverse Gaussian distributions and stochastic volatility modelling. *Scandinavian Journal of Statistics 24*(1), 1–13.

Baum, L. E., T. Petrie, G. Soules, and N. Weiss (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics 41*, 164–171.

Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference*. Ph. D. thesis, University of London.

Bechtel, Y., C. Bonaiti-Pellie, N. Poisson, J. Magnette, and P. Bechtel (1993). A population and family study of N-acetyltransferase using caffeine urinary metabolites. *Clinical Pharmacology & Therapeutics 54*(2), 134–141.

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

Browne, R. P., P. D. McNicholas, and M. D. Sparling (2012). Model-based learning using a mixture of mixtures of Gaussian and uniform distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence 34*(4), 814–817.

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

Chhikara, R. S. and J. L. Folks (1989). *The Inverse Gaussian Distribution: Theory, Methodology, and Applications*, Volume 95 of *Statistics: Textbooks and Monographs*. New York: Marcel Dekker, Inc.

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

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

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

Franczak, B. C., R. P. Browne, and P. D. McNicholas (2012). Mixtures of shifted asymmetric Laplace distributions. arXiv:1207.1727v3.

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

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

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

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

Jørgensen, B. (1982). *Statistical Properties of the Generalized Inverse Gaussian Distribution*, Volume 21. New York: Springer.

Karlis, D. and J. Lillestol (2004). Bayesian estimation of NIG models via Markov chain Monte Carlo methods. *Applied Stochastic Models in Business and Industry 20*, 323–338.

Karlis, D. and A. Santourian (2009). Model-based clustering with non-elliptically contoured distributions. *Statistics and Computing 19*(1), 73–83.

Lee, S. X. and G. J. McLachlan (2013). On mixtures of skew normal and skew t-distributions. *Advances in Data Analysis and Classification 7*(3), 241–266.

Lillestol, J. (2000). Risk analysis and the NIG distribution. *Journal of Risk 2*, 41–56.

Lin, T.-I. (2009). Maximum likelihood estimation for multivariate skew normal mixture models. *Journal of Multivariate Analysis 100*, 257–265.

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

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

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

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

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

McNicholas, P. D. and T. B. Murphy (2010). Model-based clustering of longitudinal data. *The Canadian Journal of Statistics 38*(1), 153–168.

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

McNicholas, P. D. and S. Subedi (2012). Clustering gene expression time course data using mixtures of multivariate t-distributions. *Journal of Statistical Planning and Inference 142*(5), 1114–1127.

Morris, K. and P. D. McNicholas (2013a). Dimension reduction for model-based clustering via mixtures of shifted asymmetric Laplace distributions. *Statistics and Probability Letters 83*(9), 2088–2093.

Morris, K. and P. D. McNicholas (2013b). Non-Gaussian mixtures for dimension reduction, clustering, classification, and discriminant analysis. arXiv:1308.6315.

Morris, K., P. D. McNicholas, and L. Scrucca (2013). Dimension reduction for model-based clustering via mixtures of multivariate t-distributions. *Advances in Data Analysis and Classification 7*(3), 321–338.

Murray, P. M., R. P. Browne, and P. D. McNicholas (2013). Mixtures of skew-*t* factor analyzers. arXiv:1305.4301v2.

Murray, P. M., P. D. McNicholas, and R. P. Browne (2013). Mixtures of common skew-*t* factor analyzers. arXiv: 1307.5558v2.

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

Punzo, A. and P. D. McNicholas (2013). Outlier detection via parsimonious mixtures of contaminated Gaussian distributions. arXiv:1305.4669.

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

Seshadri, V. (1993). *The Inverse Gaussian Distribution: A Case Study in Exponential Families*. New York: Oxford University Press.

Steane, M. A., P. D. McNicholas, and R. Yada (2012). Model-based classification via mixtures of multivariate t-factor analyzers. *Communications in Statistics – Simulation and Computation 41*(4), 510–523.

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

Teschendorff, A., Y. Wang, N. Barbosa-Morais, J. Brenton, and C. Caldas (2005). A variational Bayesian mixture modelling framework for cluster analysis of gene-expression data. *Bioinformatics 21*(13), 3025–3033.

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

Todorov, V. and P. Filzmoser (2009). An object-oriented framework for robust multivariate analysis. *Journal of Statistical Software 32*(3), 1–47.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S* (Fourth ed.). New York: Springer.

Vrbik, I. and P. D. McNicholas (2012). Analytic calculations for the EM algorithm for multivariate skew-t mixture models. *Statistics and Probability Letters 82*(6), 1169–1174.

Vrbik, I. and P. D. McNicholas (in press). Parsimonious skew mixture models for model-based clustering and classification. *Computational Statistics and Data Analysis*. To appear, DOI: 10.1016/j.csda.2013.07.008.

Waterhouse, S., D. MacKay, and T. Robinson (1996). Bayesian methods for mixture of experts. In *Advances in Neural Information Processing Systems*, Volume 8. Cambridge, MA: MIT Press.

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