A Mixture of Generalized Hyperbolic Distributions

Ryan P. Browne* and Paul D. McNicholas

Department of Mathematics & Statistics, University of Guelph

Abstract

We introduce a mixture of generalized hyperbolic distributions as an alternative to the ubiquitous mixture of Gaussian distributions as well as their near relatives of which the mixture of multivariate t and skew-t distributions are predominant. The mathematical development of our mixture of generalized hyperbolic distributions model relies on its relationship with the generalized inverse Gaussian distribution. The latter is reviewed before our mixture models are presented along with details of the aforesaid reliance. Parameter estimation is outlined within the expectation-maximization framework before the performance of our mixture models is illustrated in clustering applications on simulated and real data. In particular, the ability of our models to recover parameters for data from underlying Gaussian, and skew-t distributions is demonstrated. Finally, the role of Generalized hyperbolic mixtures within the wider model-based clustering, classification, and density estimation literature is discussed.

1 Introduction

Finite mixture models are based on the underlying assumption that a population is a convex combination of a finite number of densities. They therefore lend themselves quite naturally to classification and clustering problems. Formally, a random vector $\mathbf{X}$ arises from a parametric finite mixture distribution if, for all $\mathbf{x} \subset \mathbf{X}$, its density can be written

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

where $\pi_g > 0$ such that $\sum_{g=1}^{G} \pi_g = 1$ are the mixing proportions, $f_1(\mathbf{x} \mid \boldsymbol{\theta}_g), \ldots, f_G(\mathbf{x} \mid \boldsymbol{\theta}_g)$ are called component densities, and $\boldsymbol{\vartheta} = (\boldsymbol{\pi}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G)$ is the vector of parameters with $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_G)$. The component densities $f_1(\mathbf{x} \mid \boldsymbol{\theta}_1), \ldots, f_G(\mathbf{x} \mid \boldsymbol{\theta}_G)$ are usually taken to

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

be of the same type, most commonly multivariate Gaussian. The popularity of the multivariate Gaussian distribution is due to its mathematical tractability and flexibility for density estimation. In the event that the component densities are multivariate Gaussian, the density of the mixture model is $f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g \phi(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$, where $\phi(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ is the multivariate Gaussian density with mean $\boldsymbol{\mu}_g$ and covariance matrix $\boldsymbol{\Sigma}_g$. The idiom 'model-based clustering' is used to connote clustering using mixture models. Model-based classification (e.g., Dean et al., 2006; McNicholas, 2010), or partial classification (cf. McLachlan, 1992, Section 2.7), can be regarded as a semi-supervised version of model-based clustering, while model-based discriminant analysis is supervised (cf. Hastie and Tibshirani, 1996).

The recent burgeoning of non-Gaussian approaches to model-based clustering includes work on the multivariate t-distribution (Peel and McLachlan, 2000), the skew-normal distribution (Lin, 2009), the skew-t distribution (Lin, 2010; Lee and McLachlan, 2011; Vrbik and McNicholas, 2012; Murray et al., 2013), the variance-gamma distribution (McNicholas et al., 2013), as well as other approaches (Karlis and Meligkotsidou, 2007; Handcock et al., 2007; Browne et al., 2012). In this paper, we add to the richness of the pallet of non-Gaussian mixture model-based approaches to clustering and classification by introducing a mixture of generalized hyperbolic distributions, which contains the aforementioned models as special or limiting cases.

In Section 2, our methodology is developed drawing on connections with the generalized inverse Gaussian distribution. Parameter estimation is described (Section 3) before both simulated and real data analyses are used to illustrate our approach (Section 4). The paper concludes with a summary and suggestions for future work in Section 5.

2 Methodology

2.1 Generalized Inverse Gaussian Distribution

The generalized inverse Gaussian (GIG) distribution was introduced by Good (1953) and its statistical properties were laid down by Barndorff-Nielsen and Halgreen (1977), Blæsild (1978), Halgreen (1979), and Jørgensen (1982). Write $Y \backsim \text{GIG}(\psi, \chi, \lambda)$ to indicate that the random variable Y follows a generalized inverse Gaussian (GIG) distribution with parameters (ψ, χ, λ) and density

$$p(y \mid \psi, \chi, \lambda) = \frac{(\psi/\chi)^{\lambda/2}\, y^{\lambda-1}}{2K_\lambda\left(\sqrt{\psi\chi}\right)} \exp\left\{-\frac{\psi y + \chi/y}{2}\right\}, \tag{1}$$

for $y > 0$, where $\psi, \chi \in \mathbb{R}^+$, $\lambda \in \mathbb{R}$, and K_λ is the modified Bessel function of the third kind with index λ. There are several special cases of the GIG distribution, such as the gamma distribution ($\chi = 0$, $\lambda > 0$) and the inverse Gaussian distribution ($\lambda = -1/2$).

Setting $\chi = \omega\eta$ and $\psi = \omega/\eta$ or $\omega = \sqrt{\psi\chi}$ and $\eta = \sqrt{\chi/\psi}$, we obtain a different but for our purposes, more meaningful parameterization of the GIG density,

$$h(y \mid \omega, \eta, \lambda) = \frac{(y/\eta)^{\lambda-1}}{2\eta K_\lambda(\omega)} \exp\left\{-\frac{\omega}{2}\left(\frac{y}{\eta} + \frac{\eta}{y}\right)\right\}, \tag{2}$$

where $\eta > 0$ is a scale parameter, $\omega > 0$ is a concentration parameter, and λ is an index parameter. Herein, we write $Y \backsim \mathcal{I}(\omega, \eta, \lambda)$ to indicate that a random variable Y has the GIG density as parameterized in (2). The GIG distribution has some attractive properties including the tractability of the following expected values:

$$\begin{aligned}
\mathbb{E}[Y] &= \eta \frac{K_{\lambda+1}(\omega)}{K_\lambda(\omega)}, \\
\mathbb{E}[1/Y] &= \frac{1}{\eta}\frac{K_{\lambda-1}(\omega)}{K_\lambda(\omega)} = \frac{1}{\eta}\frac{K_{\lambda+1}(\omega)}{K_\lambda(\omega)} - \frac{2\lambda}{\omega\eta}, \\
\mathbb{E}[\log Y] &= \log\eta + \frac{\partial}{\partial v}\log K_\lambda(\omega) = \log\eta + \frac{1}{K_\lambda(\omega)}\frac{\partial}{\partial v}K_\lambda(\omega).
\end{aligned} \tag{3}$$

2.2 Generalized Hyperbolic Distribution

McNeil et al. (2005) give the density of a random variable $\mathbf{X}$ following the generalized hyperbolic distribution,

$$\begin{aligned}
f(\mathbf{x} \mid \boldsymbol{\vartheta}) = &\left[\frac{\chi + \delta(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Delta})}{\psi + \boldsymbol{\alpha}'\boldsymbol{\Delta}^{-1}\boldsymbol{\alpha}}\right]^{(\lambda-p/2)/2} \\
&\times \frac{[\psi/\chi]^{\lambda/2} K_{\lambda-p/2}\left(\sqrt{[\psi + \boldsymbol{\alpha}'\boldsymbol{\Delta}^{-1}\boldsymbol{\alpha}][\chi + \delta(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Delta})]}\right)}{(2\pi)^{p/2}|\boldsymbol{\Delta}|^{1/2} K_\lambda\left(\sqrt{\chi\psi}\right)\exp\left\{(\boldsymbol{\mu} - \mathbf{x})'\boldsymbol{\Delta}^{-1}\boldsymbol{\alpha}\right\}},
\end{aligned} \tag{4}$$

where $\delta(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Delta}) = (\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Delta}^{-1}(\mathbf{x} - \boldsymbol{\mu})$ is the squared Mahalanobis distance between $\mathbf{x}$ and $\boldsymbol{\mu}$, and $\boldsymbol{\vartheta} = (\lambda, \chi, \psi, \boldsymbol{\mu}, \boldsymbol{\Delta}, \boldsymbol{\alpha})$ is the vector of parameters. Herein, we use the notation $\mathbf{X} \backsim \mathcal{G}_p(\lambda, \chi, \psi, \boldsymbol{\mu}, \boldsymbol{\Delta}, \alpha)$ to indicate that a p-dimensional random variable $\mathbf{X}$ has the generalized hyperbolic density in (4). Note that we use $\boldsymbol{\Delta}$ to denote the scale because, in this parameterization, we need to hold $|\boldsymbol{\Delta}| = 1$ to ensure idenitifiability (cf. Section 2.3).

A generalized hyperbolic random variable $\mathbf{X}$ can be generated by combining a random variable $Y \backsim \text{GIG}(\psi, \chi, \lambda)$ and a latent multivariate Gaussian random variable $\mathbf{U} \backsim \mathcal{N}(\mathbf{0}, \boldsymbol{\Delta})$ using the relationship

$$\mathbf{X} = \boldsymbol{\mu} + Y\boldsymbol{\alpha} + \sqrt{Y}\mathbf{U}, \tag{5}$$

and it follows that $\mathbf{X} \mid (Y = y) \backsim \mathcal{N}(\boldsymbol{\mu} + y\boldsymbol{\alpha}, y\boldsymbol{\Delta})$. Therefore, from Bayes' theorem,

$$\begin{aligned}
f(y \mid \mathbf{x}) = &\frac{f(\mathbf{x} \mid y)h(y)}{f(\mathbf{x})} = \\
&\left[\frac{\psi + \boldsymbol{\alpha}'\boldsymbol{\Delta}^{-1}\boldsymbol{\alpha}}{\chi + \delta(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Delta})}\right]^{(\lambda-p/2)/2} \frac{y^{\lambda-p/2-1}\exp\left\{-\left[y\left(\psi + \boldsymbol{\alpha}'\boldsymbol{\Delta}^{-1}\boldsymbol{\alpha}\right) + \left(\chi + \delta(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Delta})\right)/y\right]/2\right\}}{2K_{\lambda-p/2}\left(\sqrt{[\psi + \boldsymbol{\alpha}'\boldsymbol{\Delta}^{-1}\boldsymbol{\alpha}][\chi + \delta(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Delta})]}\right)},
\end{aligned}$$

and so we have $Y \mid (\mathbf{X}=\mathbf{x}) \backsim \text{GIG}(\psi+\boldsymbol{\alpha}'\boldsymbol{\Delta}^{-1}\boldsymbol{\alpha}, \chi+\delta\left(\mathbf{x},\boldsymbol{\mu}\mid\boldsymbol{\Delta}\right), \lambda-p/2)$.

McNeil et al. (2005) describe a variety of limiting cases for the generalized hyperbolic distribution. For $\lambda=1$, we obtain the multivariate generalized hyperbolic distribution such that its univariate margins are one-dimensional hyperbolic distributions, for $\lambda=(p+1)/2$, we obtain the p-dimensional hyperbolic distribution, and for $\lambda=-1/2$, we obtain the inverse Gaussian distribution. If $\lambda>0$ and $\chi\to 0$, we obtain a limiting case of the distribution known as the generalized, Bessel function or variance-gamma distribution (Barndorff-Nielsen, 1978). If $\lambda=1$, $\psi=2$ and $\chi\to 0$, then we obtain the asymmetric Laplace distribution (cf. Kotz et al., 2001) and if $\boldsymbol{\alpha}=\mathbf{0}$, we have the symmetric generalized hyperbolic distribution (Barndorff-Nielsen, 1978). Other special and limiting cases include the multivariate normal-inverse Gaussian (MNIG) distribution (Karlis and Meligkotsidou, 2007), the skew-t distribution as well as the multivariate t-, skew-normal, and Gaussian distributions.

2.3 Identifiability and Re-Parameterizations

Suppose we relax the condition that $|\boldsymbol{\Delta}|=1$, in which case we use $\boldsymbol{\Sigma}$ to denote the scale matrix. An identifiability issue arises because the density of $\mathbf{X}_1 \backsim \mathcal{G}_p(\lambda,\chi/c,c\psi,\boldsymbol{\mu},c\boldsymbol{\Sigma},c\boldsymbol{\alpha})$ and $\mathbf{X}_2 \backsim \mathcal{G}_p(\lambda,\chi,\psi,\boldsymbol{\mu},\boldsymbol{\Sigma},\alpha)$ is identical for any $c\in\mathbb{R}^+$. Using $\boldsymbol{\Delta}$, with $|\boldsymbol{\Delta}|=1$, instead of $\boldsymbol{\Sigma}$, solves this problem but would be prohibitively restrictive for model-based clustering and classification applications. An alternative approach is to use the relationship in (5) to set the scale parameter $\eta=1$. This relationship is equivalent to $\mathbf{X}=\boldsymbol{\mu}+Y\eta\boldsymbol{\alpha}+\sqrt{Y\eta}\mathbf{U}=\boldsymbol{\mu}+Y\boldsymbol{\beta}+\sqrt{Y}\mathbf{U}$, where $\boldsymbol{\beta}=\eta\boldsymbol{\alpha}$, $Y\backsim\mathcal{I}(\omega,1,\lambda)$ and $\mathbf{U}\backsim\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})$. Under this parameterization, the density of the generalized hyperbolic distribution is

$$f(\mathbf{x}\mid\boldsymbol{\theta})=\left[\frac{\omega+\delta\left(\mathbf{x},\boldsymbol{\mu}|\boldsymbol{\Sigma}\right)}{\omega+\boldsymbol{\beta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}}\right]^{(\lambda-p/2)/2}\frac{K_{\lambda-p/2}\Big(\sqrt{\left[\omega+\boldsymbol{\beta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}\right]\left[\omega+\delta\left(\mathbf{x},\boldsymbol{\mu}|\boldsymbol{\Sigma}\right)\right]}\Big)}{(2\pi)^{p/2}\left|\boldsymbol{\Sigma}\right|^{1/2}K_{\lambda}\left(\omega\right)\exp\big\{-\left(\mathbf{x}-\boldsymbol{\mu}\right)'\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}\big\}}, \tag{6}$$

and $Y\mid(\mathbf{X}=\mathbf{x})\backsim\text{GIG}(\omega+\boldsymbol{\beta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta},\omega+\delta\left(\mathbf{x},\boldsymbol{\mu}|\boldsymbol{\Sigma}\right),\lambda-p/2)$. We use $\mathcal{G}_p^*(\lambda,\omega,\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\beta})$ to denote the density in (6) and we use this parameterization when we describe parameter estimation (Section 3).

2.4 Comments

The presence of the index parameter λ in the generalized hyperbolic density engenders a flexibility that is not found in its special and limiting cases. As an illustration of this point, consider the log-densities for three of these special and limiting cases as well as the generalized hyperbolic distributions for two different values of λ. Looking at Figure 1, it is striking that the Gaussian, variance-gamma, and t-distributions have relatively similar log-densities, whereas the generalized hyperbolic distributions have markedly different log-densities. This illustrates the extra modelling flexibility induced by the index parameter λ. Furthermore, with appropriate parameterizations, the generalized hyperbolic distribution can replicate the log-density of any of the other distributions in Figure 1. Illustrations of the generalized hyperbolic distribution capturing special and limiting cases are presented in Section 4.2. Note that the same location and scale parameters were used for each distribution, and, where relevant, skewness was set to $\mathbf{0}$.

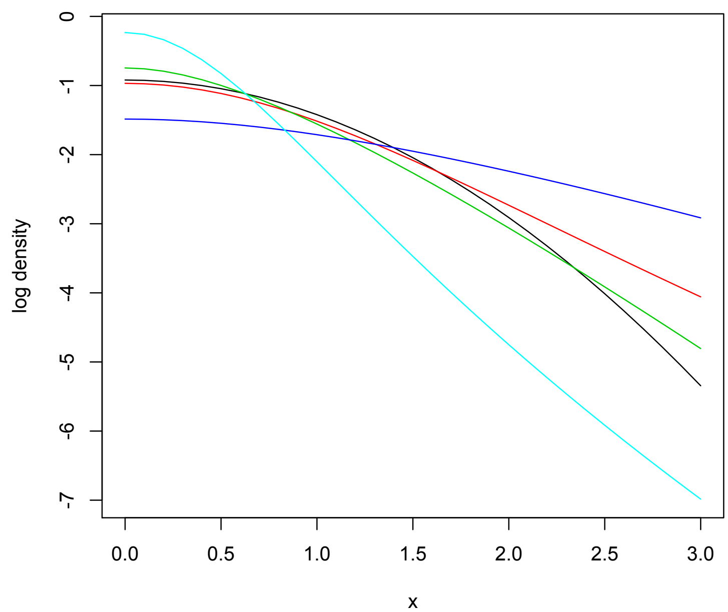

Figure 1: Log-density plots for the Gaussian distribution (black), t-distribution with 5 degrees of freedom (red), variance-gamma distribution with 5 degrees of freedom (green), and the generalized hyperbolic distribution with $\lambda = 2$ (blue) and $\lambda = -2$ (turquoise), respectively.

3 Parameter Estimation

Parameter estimation is carried out using an expectation-maximization (EM) algorithm (Dempster et al., 1977). The EM algorithm is an iterative technique that facilitates maximum likelihood estimation when data are incomplete or treated as being incomplete. In our case, the missing data comprise the group memberships and the latent variable. We assume a clustering paradigm so that none of the group membership labels are known. Denote group memberships by z_{ig}, where $z_{ig} = 1$ if observation i is in component g and $z_{ig} = 0$ otherwise. The latent variables $Y_1, \ldots, Y_n$ are assumed to follow GIG distributions and the complete-data log-likelihood is given by

$$
\begin{aligned}
l_c(\boldsymbol{\vartheta} \mid \mathbf{x}, \mathbf{y}, \mathbf{z}) &= \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\left[\log \pi_g + \sum_{j=1}^{p} \log \phi\left(\mathbf{x}_i \mid \boldsymbol{\mu}_g + y_i\boldsymbol{\beta}_g, y_i\boldsymbol{\Sigma}_g\right) + \log h(y_i \mid \omega_g, \lambda_g)\right] \\
&= C - \frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\log\left|\boldsymbol{\Sigma}_g^{-1}\right| + \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\log h(y_i \mid \omega_g, \lambda_g) \\
&\quad - \frac{1}{2}\mathrm{tr}\left\{\sum_{g=1}^{G}\boldsymbol{\Sigma}_g^{-1}\sum_{i=1}^{n} z_{ig}\left[\frac{1}{y_i}(\mathbf{x}_i - \boldsymbol{\mu}_g)(\mathbf{x}_i - \boldsymbol{\mu}_g)' - (\mathbf{x}_i - \boldsymbol{\mu}_g)\boldsymbol{\beta}_g' - \boldsymbol{\beta}_g(\mathbf{x}_i - \boldsymbol{\mu}_g)' + y_i\boldsymbol{\beta}\boldsymbol{\beta}'\right]\right\}.
\end{aligned}
$$

where C does not depend on the model parameters.

In the E-step, the expected value of the complete-data log-likelihood is computed. Because our model is from the exponential family, this is equivalent to replacing the sufficient statistics of the missing data by their expected values in $l_c(\boldsymbol{\vartheta} \mid \mathbf{x}, \mathbf{w}, \mathbf{z})$; here the missing data are the latent variables and the group membership labels. These two sources of missing data are independent and so we are only required to calculate the marginal conditional distribution for the latent variable and group memberships given the observed data. We require following expectations:

$$
\begin{aligned}
\mathbb{E}\left[Z_{ig} \mid \mathbf{x}_i\right] &= \frac{\pi_g f(\mathbf{x}_i \mid \boldsymbol{\theta}_g)}{\sum_{h=1}^{G}\pi_h f(\mathbf{x}_i \mid \boldsymbol{\theta}_h)} =: \hat{z}_{ig}, \\
\mathbb{E}\left[W_i \mid \mathbf{x}_i, Z_{ig} = 1\right] &= \eta\frac{K_{\lambda+1}(\omega)}{K_\lambda(\omega)} =: a_{ig}, \\
\mathbb{E}\left[1/W_i \mid \mathbf{x}_i, Z_{ig} = 1\right] &= \frac{1}{\eta}\frac{K_{\lambda+1}(\omega)}{K_\lambda(\omega)} - \frac{2\lambda}{\omega\eta} =: b_{ig}, \\
\mathbb{E}\left[\log(W_i) \mid \mathbf{x}_i, Z_{ig} = 1\right] &= \log\eta + \frac{1}{K_\lambda(\omega)}\frac{\partial}{\partial v}K_\lambda(\omega) =: c_{ig},
\end{aligned}
$$

and we use the notation $n_g = \sum_{i=1}^{n}\hat{z}_{ig}$, $\bar{a}_g = (1/n_g)\sum_{i=1}^{n}\hat{z}_{ig}a_i$, $\bar{a}_g = (1/n_g)\sum_{i=1}^{n}\hat{z}_{ig}b_i$, and $\bar{c}_g = (1/n_g)\sum_{i=1}^{n}\hat{z}_{ig}c_i$ hereafter.

In the M-step, we maximize the expected value of the complete-data log-likelihood to get the updates for the parameter estimates. The update for the mixing proportions is $\hat{\pi}_g = n_g/n$, where $n_g = \sum_{i=1}^{n}\hat{z}_{ig}$. Updates for $\boldsymbol{\mu}_g$ and $\boldsymbol{\beta}_g$ are given by

$$
\hat{\boldsymbol{\mu}}_g = \frac{\sum_{i=1}^{n}\hat{z}_{ig}\mathbf{x}_i(\bar{a}_g b_{ig} - 1)}{\sum_{i=1}^{n}\hat{z}_{ig}(\bar{a}_g b_{ig} - 1)} \quad \text{and} \quad \hat{\boldsymbol{\beta}}_g = \frac{\sum_{i=1}^{n}\hat{z}_{ig}\mathbf{x}_i(\bar{b}_g - b_{ig})}{\sum_{i=1}^{n}\hat{z}_{ig}(\bar{a}_g b_{ig} - 1)},
$$

respectively. The update for $\boldsymbol{\Sigma}_g$ is given by

$$
\hat{\boldsymbol{\Sigma}}_g = \frac{1}{n_g}\sum_{i=1}^{n}\hat{z}_{ig}b_{ig}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)' - \hat{\boldsymbol{\beta}}_g\left(\bar{\mathbf{x}}_g - \hat{\boldsymbol{\mu}}_g\right)' - \left(\bar{\mathbf{x}}_g - \hat{\boldsymbol{\mu}}_g\right)(\hat{\boldsymbol{\beta}}_g)' + \bar{a}_g\hat{\boldsymbol{\beta}}_g(\hat{\boldsymbol{\beta}}_g)', \quad (7)
$$

where $\bar{\mathbf{x}}_g = (1/n_g)\sum_{i=1}^n z_{ig}\mathbf{x}_i$. To demonstrate that $\hat{\mathbf{\Sigma}}_g$ is positive-definite, first note that, from Jensen's inequality: $1/\mathbb{E}[W_i] \leq \mathbb{E}[1/W_i]$, for all $i = 1, \ldots, n$. It follows that $1/a_{ig} \leq b_{ig}$ and so

$$\bar{a}_g = \frac{1}{n_g}\sum_{i=1} \hat{z}_{ig}a_{ig} \geq \frac{1}{n_g}\sum_{i=1} \frac{\hat{z}_{ig}}{b_{ig}}.$$

Now, replacing $\bar{a}_g$ with $(1/n_g)\sum_{i=1}^n(\hat{z}_{ig}/b_{ig})$ in (7), we obtain

$$\mathbf{\Sigma}_g^* = \frac{1}{n_g}\sum_{i=1}^n z_{ig}b_{ig}\left(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g - \frac{1}{b_{ig}}\hat{\boldsymbol{\beta}}_g\right)\left(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g - \frac{1}{b_{ig}}\hat{\boldsymbol{\beta}}_g\right)'$$

and the inequality $\hat{\mathbf{\Sigma}}_g \succeq \mathbf{\Sigma}_g^* \succ 0$ holds, ensuring that $\hat{\mathbf{\Sigma}}_g$ is positive-definite.

To update ω_g and λ_g we maximize the function

$$q_g(\omega_g, \lambda_g) = -\log\left(K_{\lambda_g}(\omega_g)\right) + (\lambda_g - 1)\bar{c}_g - \frac{\omega_g}{2}\left(\bar{a}_g + \bar{b}_g\right)$$

via conditional maximization, i.e., we maximize the function with respect to one parameter while holding the other parameter fixed. Baricz (2010) show that $K_\lambda(\omega)$ is strictly log-convex with respect to λ and ω. This implies that $q_g(\omega_g, \lambda_g)$ is strictly concave. The conditional maximization updates are

$$\lambda_g^{(t+1)} = \bar{c}_g\lambda_g^{(t)}\left[\frac{1}{K_\lambda\left(\omega_g^{(t)}\right)}\frac{\partial}{\partial\lambda}K_\lambda\left(\omega_g^{(t)}\right)\right]^{-1} = \bar{c}_g\lambda_g^{(t+1)}\left[\frac{\partial}{\partial\lambda}\log K_\lambda\left(\omega_g^{(t)}\right)\Big|_{\lambda=\lambda^{(t)}}\right]^{-1}$$

and

$$\omega_g^{(t+1)} = \omega_g^{(t)} - \frac{\partial_{\omega_g}q_g}{\partial^2_{\omega_g}q_g} = \omega_g^{(t)} - \frac{\frac{\partial}{\partial\omega_g}q_g(\omega_g, \lambda_g^{(t)})}{\frac{\partial^2}{\partial\omega_g^2}q_g(\omega_g, \lambda_g^{(t)})} = \omega_g^{(t)} - \frac{\frac{\partial}{\partial\omega_g}q_g}{\frac{\partial^2}{\partial\omega_g^2}q_g}.$$

Details on their derivations are given in the appendix.

The Aitken acceleration (Aitken, 1926) can used to estimate the asymptotic maximum of the log-likelihood at each iteration of an EM algorithm and thence to determine convergence. The Aitken acceleration at iteration k is

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

where $l^{(k)}$ is the log-likelihood at iteration k. An asymptotic estimate of the log-likelihood at iteration $k+1$ is

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

and the algorithm can be considered to have converged when $l_\infty^{(k)} - l^{(k)} < \epsilon$ (Böhning et al., 1994; Lindsay, 1995). This criterion is used for the analyses in Section 4.

In many practical applications, the number of mixture components G is unknown. In our illustrative data analyses (Section 4), G is treated as unknown and the Bayesian information criterion (BIC; Schwarz, 1978) is used for model selection. The BIC can be written BIC $= 2l(\mathbf{x} \mid \hat{\boldsymbol{\theta}}) - \rho \log n$, where $l(\mathbf{x} \mid \hat{\boldsymbol{\theta}})$ is the maximized log-likelihood, $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of $\boldsymbol{\theta}$, and ρ is the number of free parameters. The use of the BIC for mixture model selection can be motivated through Bayes factors (Kass and Raftery, 1995; Kass and Wasserman, 1995; Dasgupta and Raftery, 1998) and has become popular due to its widespread use within the Gaussian mixture modelling literature. While many alternatives have been proffered, none have yet proved superior.

4 Data analyses

4.1 Overview

The mixture of generalized hyperbolic distributions model is illustrated on simulated and real data. We consider cluster analyses, but these mixture models could equally well be applied for semi-supervised classification, discriminant analysis, or density estimation. In each of our clustering analyses, the true classifications are known but treated as unknown for illustration. While this sort of synthetic clustering example may not be considered quite akin to real clustering, it is representative of what has become the norm within the model-based clustering literature. Furthermore, the real data sets that we use are selected because of their popularity as benchmark data sets within the aforesaid literature. The crabs data (Section 4.3.1), in particular, are notoriously difficult to cluster.

Because we know the true group memberships, we can assess the performance of these mixture models in terms of classification accuracy, which we measure using the adjusted Rand index (ARI; Rand, 1971; Hubert and Arabie, 1985). The ARI has expected value 0 under random classification and takes the value 1 for perfect class agreement. Negative values of the ARI indicate classification that is worse than would be expected under random classification.

4.2 Simulated data analyses

Data are simulated to illustrate the effectiveness of our mixture of generalized hyperbolic distributions, with our parameter estimation approach, for modelling data from its special and limiting cases. Data are simulated from a mixture of Gaussian distributions and a mixture of skew-t distributions, respectively, and true and estimated parameters are compared. In each case, 100 two-component data sets are simulated with $n_1 = n_2 = 250$ and the models are fitted within the model-based clustering paradigm for $G = 1, \ldots, 5$. In all cases, a $G = 2$ component model is selected, perfect classification results are obtained, and the parameter estimates are close to the true values (Tables 1 and 2). Note that the results are reported in the ψ_g, χ_g parameterization for ease of interpretation; however, they were run with the ω_g parameterization.

It is notable that, in these limiting and special cases, the parameters in question are estimated very well by our mixture of generalized hyperbolic distributions. In the Gaussian mixture simulation (Table 1), both components should and do have large χ_g as well as small ψ_g and $\boldsymbol{\alpha}_g$. For the skew-t mixture example (Table 2), we generated data with degrees of freedom $\chi_1 = 8, \chi_2 = 20$ and the average estimates were very close ($\hat{\chi}_1 = 8.24, \hat{\chi}_2 = 21.02$). Again, ψ_g should be and is small. Note that λ_g is not free to vary for either the Gaussian or skew t-distributions, i.e., it is constrained so that $\lambda_g = -\chi_g/2$.

Table 1: Mean parameter estimates from the application of our mixture of generalized hyperbolic distributions to 100 simulated data sets from a two-component mixture of Gaussian distributions.

	$g=1$		$g=2$	
	True	Estimated	True	Estimated
$\boldsymbol{\mu}_g$	$(3.00, 3.00)$	$(2.91, 2.97)$	$(-3.00, -3.00)$	$(-2.74, -3.14)$
$\boldsymbol{\alpha}_g$	$(0.00, 0.00)$	$(0.09, 0.03)$	$(0.00, 0.00)$	$(-0.26, 0.14)$
$\boldsymbol{\Sigma}_g$	$(1.00, -0.75, 1.00)$	$(0.96, -0.72, 0.98)$	$(1.00, -0.75, 1.00)$	$(0.98, -0.74, 0.99)$
ψ_g	$\to 0$	0.00	$\to 0$	0.00
χ_g	$\to \infty$	192.76	$\to \infty$	189.96
λ_g	$\to -\infty$	-96.38	$\to -\infty$	-94.98

Table 2: Mean parameter estimates from the application of our mixture of generalized hyperbolic distributions to 100 simulated data sets from a two-component mixture of skew-t distributions.

	$g=1$		$g=2$	
	True	Estimated	True	Estimated
$\boldsymbol{\mu}_g$	$(3.00, 3.00)$	$(2.95, 3.04)$	$(-3.00, -3.00)$	$(-2.89, -3.11)$
$\boldsymbol{\alpha}_g$	$(2.00, -2.00)$	$(2.05, -2.06)$	$(-1.00, 1.00)$	$(-1.30, 1.36)$
$\boldsymbol{\Sigma}_g$	$(1.00, -0.75, 1.00)$	$(0.99, -0.74, 0.98)$	$(1.00, -0.75, 1.00)$	$(1.01, -0.76, 1.01)$
ψ_g	$\to 0$	0.00	$\to 0$	0.00
χ_g	8	8.24	20	21.02
λ_g	-4	-4.12	-10	-10.51

4.3 Real data analyses

4.3.1 Leptograpus crabs data

Campbell and Mahon (1974) reported data on five biological measurements of 200 crabs from the genus *leptograpus*. The data were collected in Fremantle, Western Australia and comprise 50 male and 50 female crabs for each of two species: orange and blue. The data were sourced from the `MASS` library for R (R Core Team, 2013), which contains data sets from Venables and Ripley (1999). These data are used by Raftery and Dean (2006) to illustrate the performance of a variable selection technique for model-based clustering. Mixtures of generalized hyperbolic distributions are fitted to these data for $G = 1, \ldots, 9$. The model chosen by the BIC had $G = 4$ components and the resulting MAP classifications leave only 15 crabs misclassified (ARI = 0.82). For a Gaussian mixture model fitted over the same range of G, the BIC chose a $G = 2$ component model with classification performance akin to guessing (ARI = 0.03; cf. Table 3).

In addition to outperforming the Gaussian mixture model, the performance of our mixture of generalized hyperbolic distributions on the crabs data compares favourably with many other analyses throughout the literature. For example, applying the famous MCLUST models (Fraley and Raftery, 2002), using the `mclust` software (Fraley et al., 2013) for R, results in a $G = 4$ component model that misclassifies 80 crabs (ARI = 0.31; Table 3). While merging Gaussian components can sometimes improve classification performance, it will not help with either our Gaussian mixture model or our MCLUST model for the crabs data. The extent of the difference in performance between our generalized hyperbolic mixtures and MCLUST is especially significant when one considers that many people consider MCLUST to be a sort of gold standard approach for model-based clustering.

Table 3: Classifications for the chosen mixture of generalized hyperbolic distributions and Gaussian mixture model for the crabs data.

		Gen. Hyperbolic				Gaussian		MCLUST			
		1	2	3	4	1	2	1	2	3	4
Blue	Male	39	11			21	29	31		19	
	Female		50			26	24		33	16	1
Orange	Male			50		24	26	25		25	
	Female			4	46	9	41		14	5	31

4.3.2 Italian wine data

Forina et al. (1986) reported chemical and physical measurements on three varieties (Barolo, Grignolino, Barbera) of wine from the Piedmont region of Italy. There are 178 samples and 13 measurements available within the `gclus` package (Hurley, 2004) for R. Mixtures of generalized hyperbolic distributions were fitted to these data for $G = 1, \ldots, 9$. The BIC selected a $G = 3$ component model with just three misclassified samples (ARI = 0.95; Table 4). Mixtures of Gaussian distributions were fitted over the same range of G and the BIC selected a $G = 2$ component model with ARI = 0.55 (Table 4).

Applying the famous MCLUST family to these data results in a $G = 8$ component model with ARI = 0.48. Again, our generalized hyperbolic mixtures have outperformed the well

Table 4: Classifications for the chosen mixture of generalized hyperbolic distributions and Gaussian mixture model for the Italian wine data.

	Gen. Hyperbolic			Gaussian	
	1	2	3	1	2
Barolo	58	1		59	
Grignolino	1	70		3	68
Barbera		1	47		48

established MCLUST family. When analyzing the MCLUST results for the crabs data (Section 4.3.1), we pointed out that merging components could not improve the classification results. Here, however, the most favourable merging scenario results in a model that misclassifies just two samples (cf. Table 5), which is a comparable result to our generalized hyperbolic mixtures.

Table 5: Classifications for the chosen MCLUST model for the Italian wine data.

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

5 Discussion

A mixture of generalized hyperbolic distributions has been introduced. Parameter estimation, via an EM algorithm, was enabled by exploitation of the relationship with the GIG distribution. The mixture models were illustrated in two real clustering applications where they outperformed Gaussian mixture models and performed favourably when compared to the well-established MCLUST approach. Although illustrated for clustering, mixtures of generalized hyperbolic distributions can also be applied for semi-supervised classification, discriminant analysis, and density estimation. They represent perhaps the most flexible in a series of alternatives to the Gaussian mixture models for clustering and classification. What sets the mixture of generalized hyperbolic distributions apart from other alternatives is the presence of an index parameters, and the fact that it has many common models as special or limiting cases. In a simulation, we showed that the generalized hyperbolic mixtures with our EM algorithm perform very well when recovering parameters for limiting and special cases.

The precise impact of our generalized hyperbolic mixtures within the wider model-based clustering literature is difficult to predict at present. Certainly, we do not suggest that one should use our mixtures of generalized hyperbolic approaches exclusively, completely ignoring more well-established approaches such as Gaussian mixtures. However, results obtained to date suggest that application of mixtures of generalized hyperbolic in real cluster analyses can outperform its special cases and this should not be ignored. One downside of our generalized hyperbolic mixtures is longer run time, e.g., fitting the wine data for $G = 3$ takes just over 216 seconds using serial code written in R, while fitting a Gaussian mixture takes less than one second. Implementing the code in a lower level language and parallel implementation can be used to greatly reduce the runtime, and both are topics of ongoing work.

Future work will also focus on the introduction of parsimony to mixtures of generalized hyperbolic distributions and a detailed study comparing the resulting families to its special and limiting cases. Parsimony can be achieved by imposing constraints on the $\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_G$, which is a somewhat natural approach because, for all but very small p, most of the model parameters are there. These constraints could be in the form of an eigen-decomposition, as used by MCLUST. There is also the possibility of considering a generalized hyperbolic analogue of the mixture of factor analyzers (Ghahramani and Hinton, 1997; McLachlan and Peel, 2000) or mixture of common factors (Baek et al., 2010).

A Updates for λ_g and ω_g

For this section we drop the subscript g to ease the notational burden, so we have

$$q(\omega, \lambda) = -\log\left(K_\lambda\left(\omega\right)\right) + (\lambda - 1)\bar{c} - \frac{\omega}{2}\left(\bar{a} + \bar{b}\right).$$

Baricz (2010) shows $K_\lambda\left(\omega\right)$ is strictly log-convex is with respect to λ and ω. Thus, we use conditional maximization to obtain updating equations. For ω, we use Newton's method using the first and second derivative of q with respect to ω.

$$\frac{\partial}{\partial\omega}q(\omega, \lambda) = -\frac{K'_\lambda(\omega)}{K_\lambda(\omega)} - \frac{1}{2}\left(\bar{a} + \bar{b}\right) = \frac{1}{2}\left[R_\lambda(\omega) + R_{-\lambda}(\omega) - \left(\bar{a} + \bar{b}\right)\right],$$

$$\frac{\partial^2}{\partial\omega^2}q(\omega, \lambda) = \frac{1}{2}\left[R_\lambda(\omega)^2 - \frac{1+2\lambda}{\omega}R_\lambda(\omega) - 1 + R_{-\lambda}(\omega)^2 - \frac{1-2\lambda}{\omega}R_{-\lambda}(\omega) - 1\right],$$

where $R_\lambda(\omega) = K_{\lambda+1}(\omega)/K_\lambda(\omega)$.

To update λ, we construct a surrogate function by deriving a bound on the second derivative via the following integral representation (Watson, 1944, pg. 181) of the modified Bessel function of the third kind. The second derivative of $\cosh(t)$, is itself $\cosh(t)$ which is bounded below by 1. Thus, we can construct a quadratic function such that

$$\begin{aligned}
\cosh(t) &\geq 1 + t^2 \\
e^{-\omega \cosh(t)} &\leq e^{-\omega(1+t^2)} \quad \text{for } \omega > 0 \\
e^{-\omega \cosh(t)} \cosh(\lambda t) &\leq e^{-\omega(1+t^2)} \cosh(\lambda t) \\
\int_0^\infty e^{-\omega \cosh(t)} \cosh(\lambda t) \mathrm{d}t &\leq \int_0^\infty e^{-\omega(1+t^2)} \cosh(\lambda t) \mathrm{d}t \\
K_\lambda(\omega) &\leq \sqrt{\frac{\pi}{2\omega}} \exp\left\{\frac{\lambda^2 - 2\omega^2}{2\omega}\right\} \\
\log K_\lambda(\omega) &\leq \log\sqrt{\frac{\pi}{2}} - \log\omega + \frac{\lambda^2 - 2\omega^2}{2\omega}
\end{aligned}$$

Because both these functions are convex in $\lambda \in \mathbb{R}$, the following inequality holds

$$0 \leq \frac{\partial^2}{\partial \lambda^2} \log K_\lambda(\omega) \leq \frac{1}{\omega}$$

for all $\lambda \in \mathbb{R}$. Because the second derivative is bounded and $\log K_\lambda(\omega)$, is an even function with respect to λ we follow de Leeuw and Lange (2009) and construct the majorizing surrogate function

$$g(\lambda|\lambda_0) = \log K_{\lambda_0}(\omega) + \frac{\lambda^2 - \lambda_0^2}{2\lambda_0}\left(\frac{\partial}{\partial\lambda} \log K_\lambda(\omega)\Big|_{\lambda=\lambda_0}\right).$$

Applying this surrogate function to $q(\omega, \lambda)$, we obtain the desired update for λ.

References

Aitken, A. C. (1926). On Bernoulli's numerical solution of algebraic equations. *Proceedings of the Royal Society of Edinburgh 46*, 289–305.

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*, 1298–1309.

Baricz, A. (2010). Turn type inequalities for some probability density functions. *Studia scientiarum mathematicarum Hungarica 47*, 175–189.

Barndorff-Nielsen, O. (1978). Hyperbolic distributions and distributions on hyperbolae. *Scandinavian Journal of Statistics 5*, 151–157.

Barndorff-Nielsen, O. and C. Halgreen (1977). Infinite divisibility of the hyperbolic and generalized inverse Gaussian distributions. *Z. Wahrscheinlichkeitstheorie Verw. Gebiete 38*, 309–311.

Blæsild, P. (1978). The shape of the generalized inverse Gaussian and hyperbolic distributions. Research Report 37, Department of Theoretical Statistics, Aarhus University, Denmark.

Böhning, D., E. Dietz, R. Schaub, P. Schlattmann, and B. Lindsay (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.

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.

Campbell, N. A. and R. J. Mahon (1974). A multivariate study of variation in two species of rock crab of genus leptograpsus. *Australian Journal of Zoology 22*, 417–425.

Dasgupta, A. and A. E. Raftery (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association 93*, 294–302.

de Leeuw, J. and K. Lange (2009). Sharp quadratic majorization in one dimension. *Computational Statistics and Data Analysis 53*(7), 2471–2484.

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

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.

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

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.

Fraley, C., A. E. Raftery, and L. Scrucca (2013). *mclust: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation.* R package version 4.2.

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

Good, I, J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika 40*, 237–260.

Halgreen, C. (1979). Self-decomposibility of the generalized inverse Gaussian and hyperbolic distributions. *Z. Wahrscheinlichkeitstheorie Verw. Gebiete 47*, 13–18.

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

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*(1), 193–218.

Hurley, C. (2004). Clustering visualizations of multivariate data. *Journal of Computational and Graphical Statistics 13*(4), 788–806.

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

Karlis, D. and L. Meligkotsidou (2007). Finite mixtures of multivariate Poisson distributions with application. *Journal of Statistical Planning and Inference 137*(6), 1942–1960.

Kass, R. E. and A. E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*(430), 773–795.

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

Kotz, S., T. J. Kozubowski, and K. Podgorski (2001). *The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance*. Boston: Birkhauser.

Lee, S. and G. J. McLachlan (2011). On the fitting of mixtures of multivariate skew t-distributions via the EM algorithm. arXiv:1109.4706.

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

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

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

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

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

McNeil, A. J., R. Frey, and P. Embrechts (2005). *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press.

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

McNicholas, S. M., P. D. McNicholas, and R. P. Browne (2013). Mixtures of variance-gamma distributions. Arxiv preprint arXiv:1309.2695.

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

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

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

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

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association 66*(336), 846–850.

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

Venables, W. N. and B. D. Ripley (1999). *Modern Applied Statistics with S-PLUS*. Springer.

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

Watson, G. N. (1944). *A Treatise on the Theory of Bessel Functions*. Cambridge: Cambridge University Press.