HAL
open science

Location and scale mixtures of Gaussians with flexible tail behaviour: properties, inference and application to multivariate clustering

Darren Wraith, Florence Forbes

To cite this version:

Darren Wraith, Florence Forbes. Location and scale mixtures of Gaussians with flexible tail behaviour: properties, inference and application to multivariate clustering. Computational Statistics and Data Analysis, 2015, 90, pp.61-73. 10.1016/j.csda.2015.04.008 . hal-01970565

HAL Id: hal-01970565

https://hal.science/hal-01970565

Submitted on 5 Jan 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Location and scale mixtures of Gaussians with flexible tail behaviour: properties, inference and application to multivariate clustering

Darren Wraith[1], Florence Forbes[1]

INRIA, Laboratoire Jean Kuntzman, Mistis team
655 avenue de l'Europe, Montbonnot
38334 Saint-Ismier Cedex, France

Abstract

The family of location and scale mixtures of Gaussians has the ability to generate a number of flexible distributional forms. The family nests as particular cases several important asymmetric distributions like the Generalised Hyperbolic distribution. The Generalised Hyperbolic distribution in turn nests many other well known distributions such as the Normal Inverse Gaussian. In a multivariate setting, an extension of the standard location and scale mixture concept is proposed into a so called *multiple scaled* framework which has the advantage of allowing different tail and skewness behaviours in each dimension with arbitrary correlation between dimensions. Estimation of the parameters is provided via an EM algorithm and extended to cover the case of mixtures of such multiple scaled distributions for application to clustering. Assessments on simulated and real data confirm the gain in degrees of freedom and flexibility in modelling data of varying tail behaviour and directional shape.

Keywords: Covariance matrix decomposition, EM algorithm, Gaussian location and scale mixture, Multivariate Generalised Hyperbolic distribution, Robust clustering

1. Introduction

A popular approach to identify groups or clusters within data is via a parametric finite mixture model (Fruwirth-Schnatter, 2006). While the vast majority

of work on such mixtures has been based on Gaussian mixture models (see *e.g.* Fraley and Raftery, 2002). In many applications the tails of Gaussian distributions are shorter than appropriate and the Gaussian shape is not suitable for highly asymmetric data. A natural extension to the Gaussian case is to consider families of distributions which can be represented as *location* ***and*** *scale Gaussian mixtures* of the form,

$$p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta}, \boldsymbol{\theta}) = \int_0^\infty \mathcal{N}_M(\mathbf{y}; \boldsymbol{\mu} + w\boldsymbol{\beta}\boldsymbol{\Sigma}, w\boldsymbol{\Sigma})\ f_W(w; \boldsymbol{\theta})\ \mathrm{d}w, \tag{1}$$

where $\mathcal{N}_M(\ .\ ; \boldsymbol{\mu} + w\boldsymbol{\beta}\boldsymbol{\Sigma}, w\boldsymbol{\Sigma})$ denotes the M-dimensional Gaussian distribution with mean $\boldsymbol{\mu} + w\boldsymbol{\beta}\boldsymbol{\Sigma}$ and covariance $w\boldsymbol{\Sigma}$ and f_W is the probability distribution of a univariate positive variable W referred to hereafter as the weight variable. The parameter $\boldsymbol{\beta}$ is an additional M-dimensional vector parameter for skewness. When $\boldsymbol{\beta} = 0$ and W^{-1} follows a Gamma distribution $\mathcal{G}(\nu/2, \nu/2)$ *i.e.* f_W is an Inverse Gamma distribution $inv\mathcal{G}(\nu/2, \nu/2)$ where ν denotes the degrees of freedom, we recover the well known multivariate t-distribution (Kotz and Nadarajah, 2004). The weight variable W in this case effectively acts to govern the tail behaviour of the distributional form from light tails ($\nu \to \infty$) to heavy tails ($\nu \to 0$) depending on the value of ν.

In the more general case of, for example, allowing $\boldsymbol{\beta} \neq 0$ and f_W being a Generalised Inverse Gaussian (GIG) distribution, we recover the family of Generalised Hyperbolic (GH) distributions (Barndorff-Nielsen, 1997) which is able to represent a particularly large number of distributional forms.

Due to the flexibility of the GH family, recent interest has focussed on applications for mixture models and factor analysis (Browne and McNicholas, 2013; Tortora et al., 2013). For mixture model applications, semi-parametric or non-parametric approaches can also be used. However, to maintain tractability or identifiability in a multivariate setting, most approaches appear to restrict the type of dependence structures between the coordinates of the multidimensional variable. Typically, conditional independence (on the mixture components) is assumed in (Benaglia et al., 2009a,b) while a Gaussian copula is used in (Chang and Walther, 2007).

For applied problems, the most popular form of the GH family appears to be the Normal Inverse Gaussian (NIG) distribution (Barndorff-Nielsen et al., 1982; Protassov, 2004; Karlis and Santourian, 2009). The NIG distribution has been used extensively in financial applications, see Protassov (2004); Barndorff-Nielsen (1997) or Aas et al. (2005); Aas and Hobaek Haff (2006) and references therein, but also in geoscience and signal processing (Gjerde et al., 2011; Oigard et al., 2004). Another popular distributional form allowing for skewness and heavy or light tails includes different forms of the multivariate skew-t. As presented by Lee and McLachlan (2013c), most formulations adopt either a *restricted* or *unrestricted* characterization. Unrestricted forms include the proposals and implementations of Sahu et al. (2003); Lee and McLachlan (2014b); Lin (2010) while restricted forms include that of Azzalini and Capitanio (2003); Basso et al. (2010); Branco and Dey (2001); Cabral et al. (2012); Pyne et al. (2009). In more recent work, Lee and McLachlan (2014a) pointed out that both restricted an unrestricted characterizations could be unified under a more general formulation referred to as *Canonical fundamental skew-t distribution*. Alternative distributional forms include those based on scale mixtures of skew-normal distributions such as in Vilca et al. (2014a) and Lin et al. (2014). In the bivariate case, this includes extensions to the Birnbaum-Saunders distribution (Vilca et al., 2014b).

Although the above approaches provide for great flexibility in modelling data of highly asymmetric and heavy tailed form, they assume f_W to be a univariate distribution and hence each dimension is governed by the same amount of tail-weight. There have been various approaches to address this issue in the statistics literature for both symmetric and asymmetric distributional forms but most of them suffer either from the non-existence of a closed-form pdf or from a difficult generalization to more than two dimensions (see Forbes and Wraith (2014) for more detailed references).

An alternative approach (Schmidt et al., 2006), which takes advantage of the property that Generalised Hyperbolic distributions are closed under affine-linear transformations, derives independent GH marginals but estimation of

parameters appears to be restricted to density estimation, and not formally generalisable to estimation settings for a broad range of applications (*e.g.* clustering, regression, *etc.*). A more general approach outside of the GH distribution setting is outlined in (Ferreira and Steel, 2007b,a) with a particular focus on regression models using a Bayesian framework.

In this paper, we build upon the multiple scaled framework of Forbes and Wraith (2014) to provide a much wider variety of distributional forms, allowing different tail and skewness behavior in each dimension of the variable space with arbitrary correlation between dimensions. A similar approach to ours has been undertaken in (as yet) unpublished work (Tortora et al., 2014b; Franczak et al., 2014). The latter work focusses on Laplace distributions while the former introduces the *Coalesced GH distribution* as a mixture of a standard GH and a multiple scaled GH distribution. This multiple scaled distribution is derived using a different parameterization and with constraints on the parameters. Both measures are undertaken to ensure identifiability but has the disadvantage of limiting the type of tail behaviors able to be modelled and results in a very different performance in terms of clustering as illustrated in Section 4.

The paper is outlined as follows. Section 2 presents the proposed new family of multiple scaled GH (and NIG) distributions. In Section 3, details are outlined for maximum likelihood estimation of the parameters for the multiple scaled NIG distribution via the EM algorithm. In Section 4 we explore the performance of the approach on simulated and real data sets in the context of clustering. Section 5 concludes with a discussion and areas for further research.

2. Multiple scaled Generalised Hyperbolic distributions

In this section we outline further details of the standard (single weight) multivariate GH distribution and then the proposed multiple scaled GH distribution.

2.1. Multivariate Generalised Hyperbolic distribution

As mentioned previously the Generalised Hyperbolic distribution can be represented in terms of a *location* ***and*** *scale Gaussian mixture* (1). Using notation

equivalent to that of Barndorff-Nielsen (1997) Section 7 and Protassov (2004), the multivariate GH density takes the following form

$$\begin{aligned}\mathcal{GH}(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\beta},\lambda,\gamma,\delta) &= \int_0^\infty \mathcal{N}_M(\mathbf{y};\boldsymbol{\mu}+w\boldsymbol{\Sigma}\boldsymbol{\beta},w\boldsymbol{\Sigma})\ \mathcal{GIG}(w;\lambda,\gamma,\delta)dw \\ &= (2\pi)^{-M/2}|\boldsymbol{\Sigma}|^{-1/2}\left(\frac{\gamma}{\delta}\right)^{\lambda}\left(\frac{q(\mathbf{y})}{\alpha}\right)^{\lambda-\frac{M}{2}}K_{\lambda-\frac{M}{2}}(q(\mathbf{y})\alpha) \\ &\quad\times\ (K_\lambda(\delta\gamma))^{-1}\exp(\boldsymbol{\beta}^T(\mathbf{y}-\boldsymbol{\mu}))\ , \qquad (2)\end{aligned}$$

where $|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$, $\delta > 0$, and $q(\mathbf{y})$ and α are positive and given by

$$q(\mathbf{y})^2 = \delta^2 + (\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\ , \qquad (3)$$

$$\gamma^2 = \alpha^2 - \boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta} \geqq 0\ . \qquad (4)$$

The parameters $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$ are column vectors of length M ($M \times 1$ vector), $K_r(x)$ is the modified Bessel function of the third kind of order r evaluated at x (see Appendix in Jorgensen (1982)), and $\mathcal{GIG}(w;\lambda,\gamma,\delta)$ is the density function of the GIG distribution which depends on three parameters,

$$\mathcal{GIG}(w;\lambda,\gamma,\delta) = \left(\frac{\gamma}{\delta}\right)^{\lambda}\frac{w^{\lambda-1}}{2K_\lambda(\delta\gamma)}\exp(-\frac{1}{2}(\delta^2/w+\gamma^2 w))\ . \qquad (5)$$

An alternative (hierarchical) representation of the multivariate GH distribution (which is useful for simulation) is

$$\begin{aligned}\boldsymbol{Y}|W=w &\sim \mathcal{N}_M(\boldsymbol{\mu}+w\boldsymbol{\Sigma}\boldsymbol{\beta},w\boldsymbol{\Sigma})\ , \\ W &\sim \mathcal{GIG}(\lambda,\gamma,\delta)\ . \qquad (6)\end{aligned}$$

By setting $\lambda = -1/2$ in the GIG distribution we recover the Inverse Gaussian (IG) distribution $\mathcal{IG}(w;\gamma,\delta)$, which (when used as the mixing distribution) leads to the NIG distribution (see section B of the Supplementary Materials for details).

Using the parameterisation of Barndorff-Nielsen (1997), an identifiability issue arises as the densities $\mathcal{GH}(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\beta},\lambda,\gamma,\delta)$ and $\mathcal{GH}(\boldsymbol{\mu},k^2\boldsymbol{\Sigma},\boldsymbol{\beta},\lambda,k\gamma,\delta/k)$ are identical for any $k > 0$. For the estimation of parameters, this problem can be solved by constraining the determinant of $\boldsymbol{\Sigma}$ to be 1.

2.2. Multiple Scaled Generalised Hyperbolic distribution (MSGH)

Following the same approach as in Forbes and Wraith (2014), the standard location and scale representation (1) is generalised into a multiple scale version

$$p(\mathbf{y};\boldsymbol{\mu},\boldsymbol{D},\boldsymbol{A},\boldsymbol{\beta},\boldsymbol{\theta}) = \int_0^\infty \dots \int_0^\infty \mathcal{N}_M(\mathbf{y};\boldsymbol{\mu}+\boldsymbol{D}\boldsymbol{\Delta}_{\mathbf{w}}\boldsymbol{A}\boldsymbol{D}^T\boldsymbol{\beta},\boldsymbol{D}\boldsymbol{\Delta}_{\mathbf{w}}\boldsymbol{A}\boldsymbol{D}^T) \times f_{\mathbf{w}}(w_1\dots w_M;\boldsymbol{\theta})\,\mathrm{d}w_1\dots dw_M\ , \tag{7}$$

where $\boldsymbol{\Sigma} = \boldsymbol{D}\boldsymbol{A}\boldsymbol{D}^T$ with $\boldsymbol{D}$ the matrix of eigenvectors of $\boldsymbol{\Sigma}$, $\boldsymbol{A}$ a diagonal matrix with the corresponding eigenvalues and $\boldsymbol{\Delta}_{\mathbf{w}} = \mathrm{diag}(w_1,\dots w_M)$. The weights are assumed independent: $f_{\mathbf{w}}(w_1\dots,w_M;\boldsymbol{\theta}) = f_{W_1}(w_1;\boldsymbol{\theta}_1)\dots f_{W_M}(w_M;\boldsymbol{\theta}_M)$.

If we set $f_{W_m}(w_m)$ to a GIG distribution $\mathcal{GIG}(w_m;\lambda_m,\gamma_m,\delta_m)$, it follows that our generalization (MSGH) of the multivariate GH distribution with $\boldsymbol{\lambda} = [\lambda_1,\dots,\lambda_M]^T$,$\boldsymbol{\gamma} = [\gamma_1,\dots,\gamma_M]^T$ and $\boldsymbol{\delta} = [\delta_1,\dots,\delta_M]^T$ as M-dimensional vectors is:

$$\begin{aligned} & \mathcal{MSGH}(\mathbf{y};\boldsymbol{\mu},\boldsymbol{D},\boldsymbol{A},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\delta}) \\ = \ & (2\pi)^{-M/2}\prod_{m=1}^{M}|A_m|^{-1/2}\left(\frac{\gamma_m}{\delta_m}\right)^{\lambda_m}\left(\frac{q_m(\mathbf{y})}{\alpha_m}\right)^{\lambda_m-1/2}K_{\lambda_m-1/2}(q_m(\mathbf{y})\alpha_m)\times \\ & (K_{\lambda_m}(\delta_m\gamma_m))^{-1}\exp([\boldsymbol{D}^T(\mathbf{y}-\boldsymbol{\mu})]_m\,[\boldsymbol{D}^T\boldsymbol{\beta}]_m)\ , \end{aligned} \tag{8}$$

with $\alpha_m^2 = \gamma_m^2 + A_m[\boldsymbol{D}^T\boldsymbol{\beta}]_m^2$ and $q_m(\mathbf{y})^2 = \delta_m^2 + A_m^{-1}[\boldsymbol{D}^T(\mathbf{y}-\boldsymbol{\mu})]_m^2$.

Alternatively, with $\mathbf{w} = [w_1,\dots,w_M]^T$ we can define it as

$$\begin{aligned} \boldsymbol{Y}|\boldsymbol{W}=\mathbf{w} \ &\sim\ \mathcal{N}_M(\boldsymbol{\mu}+\boldsymbol{D}\boldsymbol{\Delta}_{\mathbf{w}}\boldsymbol{A}\boldsymbol{D}^T\boldsymbol{\beta},\boldsymbol{D}\boldsymbol{\Delta}_{\mathbf{w}}\boldsymbol{A}\boldsymbol{D}^T)\ , \\ \boldsymbol{W} \ &\sim\ \mathcal{GIG}(\lambda_1,\gamma_1,\delta_1)\otimes\dots\otimes\mathcal{GIG}(\lambda_M,\gamma_M,\delta_M), \end{aligned} \tag{9}$$

where notation $\otimes$ means that the $\boldsymbol{W}$ components are independent. If we set $f_{W_m}(w_m)$ to an Inverse Gaussian distribution $\mathcal{IG}(w_m;\gamma_m,\delta_m)$, it follows that our generalization (MSNIG) of the multivariate NIG distribution is:

$$\begin{aligned} \mathcal{MSNIG}(\mathbf{y};\boldsymbol{\mu},\boldsymbol{D},\boldsymbol{A},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\delta}) \ = \ & \prod_{m=1}^{M}\delta_m\exp(\delta_m\gamma_m+[\boldsymbol{D}^T(\mathbf{y}-\boldsymbol{\mu})]_m\,[\boldsymbol{D}^T\boldsymbol{\beta}]_m) \\ & \times\frac{\alpha_m}{\pi q_m}K_1(\alpha_m q_m(\mathbf{y}))\ , \end{aligned} \tag{10}$$

with $\alpha_m^2 = \gamma_m^2 + A_m[\boldsymbol{D}^T\boldsymbol{\beta}]_m^2$, $q_m(\mathbf{y})^2 = \delta_m^2 + A_m^{-1}[\boldsymbol{D}^T(\mathbf{y}-\boldsymbol{\mu})]_m^2$ and K_1 is the modified Bessel function of order 1.

It is interesting to note that the multiple scaled GH distribution allows potentially each dimension to follow a particular case of the GH distribution family. For example, in a bivariate setting $\mathbf{Y} = [Y_1, Y_2]^T$, the variate Y_1 could follow a hyperboloid distribution (λ_1=0) and Y_2 a NIG distribution ($\lambda_2 = -1/2$).

2.3. Identifiability issues

In contrast to the standard multivariate GH distribution, constraining the determinant of $\boldsymbol{A}$ to be 1 is not enough to ensure identifiability in the MSGH case. Indeed, assuming the determinant $|\boldsymbol{A}| = 1$, if we set $\boldsymbol{A}', \boldsymbol{\delta}', \boldsymbol{\gamma}'$ so that $A'_m = k_m^2 A_m, \delta'_m = \delta_m/k_m$ and $\gamma'_m = k_m\gamma_m$, for all values $k_1 \dots k_m$ satisfying $\prod_{m=1}^M k_m^2 = 1$, it follows that the determinant $|\boldsymbol{A}'| = 1$ and that the $\mathcal{MSGH}(\mathbf{y}, \boldsymbol{\mu}, \boldsymbol{D}, \boldsymbol{A}', \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}', \boldsymbol{\delta}')$ and $\mathcal{MSGH}(\mathbf{y}, \boldsymbol{\mu}, \boldsymbol{D}, \boldsymbol{A}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\delta})$ expressions are equal. Identifiability can be guaranteed by adding that all δ_m's (or equivalently all γ_m's) are equal. In practice, we will therefore assume that for all $m = 1 \dots M$, $\delta_m = \delta$.

2.4. Some properties of the multiple scaled GH distributions

The MSGH distribution (as defined in (8)) provides for very flexible distributional forms. For illustration, in the bivariate case, several contour plots of the multiple scaled NIG (MSNIG), *i.e.* for all m, $\lambda_m = -1/2$, are shown in Figure 1 and compared with the standard multivariate NIG. In this two-dimensional setting, we use for $\boldsymbol{D}$ a parameterisation via an angle ξ so that $D_{11} = D_{22} = \cos\xi$ and $D_{21} = -D_{12} = \sin\xi$, where D_{md} denotes the (m, d) entry of matrix $\boldsymbol{D}$. Similar to the standard NIG the parameter $\boldsymbol{\beta}$ measures asymmetry and its sign determines the type of skewness. For the standard NIG the contours are not necessarily elliptical and this is also the case with the MSNIG. In the case of the MSNIG additional flexibility is provided by allowing the parameter $\boldsymbol{\gamma}$ to be a vector of dimension M instead of a scalar. Keeping all δ_m's equal to the same δ, this vectorisation of $\boldsymbol{\gamma}$ effectively allows each dimension to be governed by different tail behaviour depending on the values of $\boldsymbol{\gamma}$. The parameters $\boldsymbol{\gamma}$ and

$\boldsymbol{\beta}$ govern the tail behaviour of the density with smaller values of $\boldsymbol{\gamma}$ implying heavier tails, and larger values lighter tails. Other multiple scaled and standard GH distributions are also illustrated in Figure 1. As shown in Figure 1(g,i), for different values of λ_m the shape of the contours do not change significantly but larger values of λ_m tend to produce heavier tails. This can also be seen from the tail behaviour analysis of the MSGH, which is similar to that of the GH with tails governed by a combined algebraic and exponential form. Details are given in section C of the Supplementary Materials.

Moments of the MSGH distribution can be obtained using the moments of the GIG distribution (see Jorgensen, 1982), *i.e.*, if W follows a $\mathcal{GIG}(\lambda, \gamma, \delta)$ distribution, for all $r \in \mathbb{Z}_+$,

$$E[W^r] = \left(\frac{\delta}{\gamma}\right)^r \frac{K_{\lambda+r}(\delta\gamma)}{K_\lambda(\delta\gamma)}. \tag{11}$$

It follows from representation (9) that when $\mathbf{Y}$ follows a MSGH distribution,

$$\begin{aligned} E[\mathbf{Y}] &= E[E[\mathbf{Y}|\mathbf{W}]] = \boldsymbol{\mu} + \boldsymbol{D}E[\Delta_{\mathbf{w}}]\boldsymbol{A}\boldsymbol{D}^T\boldsymbol{\beta} \\ &= \boldsymbol{\mu} + \boldsymbol{D}\,\mathrm{diag}\left(\frac{\delta_m}{\gamma_m}\frac{K_{\lambda_m+1}(\delta_m\gamma_m)}{K_{\lambda_m}(\delta_m\gamma_m)}\right)\boldsymbol{A}\boldsymbol{D}^T\boldsymbol{\beta}, \end{aligned} \tag{12}$$

where for short, we denoted by $\mathrm{diag}(u_m)$ the M-dimensional diagonal matrix whose diagonal components are $\{u_1, \ldots, u_M\}$.

For the covariance matrix, we obtain,

$$\begin{aligned} \mathrm{Var}[\mathbf{Y}] =& E[\mathrm{Var}[\mathbf{Y}|\mathbf{W}]] + \mathrm{Var}[E[\mathbf{Y}|\mathbf{W}]] \\ =& \boldsymbol{D}E[\Delta_{\mathbf{W}}]\boldsymbol{A}\boldsymbol{D}^T + \boldsymbol{D}\boldsymbol{A}\,\mathrm{Var}[\Delta_{\mathbf{W}}\boldsymbol{D}^T\boldsymbol{\beta}]\boldsymbol{A}\boldsymbol{D}^T \\ =& \boldsymbol{D}\,\mathrm{diag}\Bigg(\frac{\delta_m A_m}{\gamma_m}\frac{K_{\lambda_m+1}(\delta_m\gamma_m)}{K_{\lambda_m}(\delta_m\gamma_m)}\Bigg(1 + \frac{\delta_m}{\gamma_m}[\boldsymbol{D}^T\boldsymbol{\beta}]_m^2\boldsymbol{A}_m \\ &\times\left(\frac{K_{\lambda_m+2}(\delta_m\gamma_m)}{K_{\lambda_m+1}(\delta_m\gamma_m)} - \frac{K_{\lambda_m+1}(\delta_m\gamma_m)}{K_{\lambda_m}(\delta_m\gamma_m)}\right)\Bigg)\Bigg)\boldsymbol{D}^T \end{aligned} \tag{13}$$

For details of the mean and variance for the MSNIG distribution see section B of the Supplementary Materials.

As can be seen from (13), the variance of the MSGH takes a slightly complicated form with some dependency on the skewness parameter $\boldsymbol{\beta}$. This dependency is also present in the variance, recalled below, of the standard multivariate GH as given in (2),

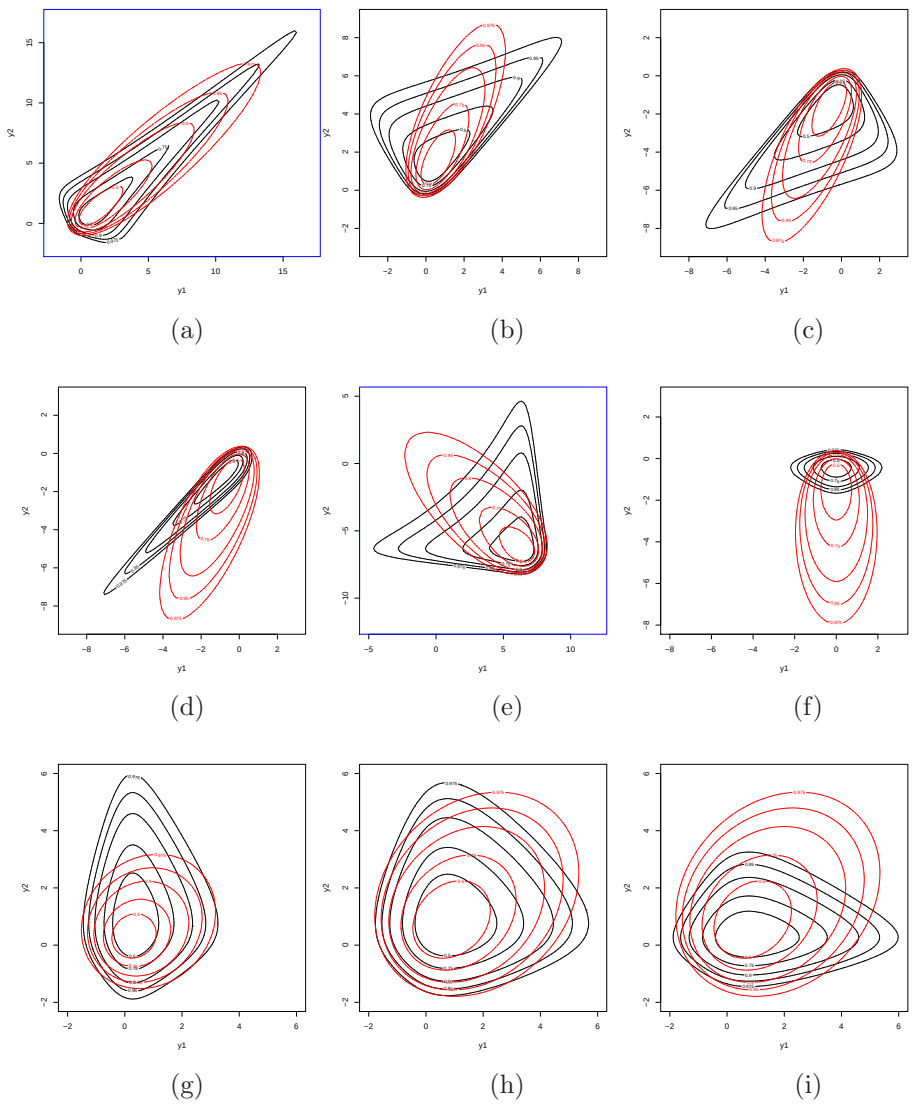


Figure 1: Top and middle panels: contour plots of bivariate MSNIG ($\lambda_1 = \lambda_2 = -1/2$) distributions (solid lines) with $\boldsymbol{\mu} = [0, 0]^T$ and $\boldsymbol{\delta} = \{1, 1\}$. The difference with the standard multivariate NIG (red dashed lines) is illustrated with univariate δ and γ values taken as the first respective component of the bivariate $\boldsymbol{\delta}$ and $\boldsymbol{\gamma}$. The (a-d) panels correspond to the same $\boldsymbol{\Sigma}$ built from $\boldsymbol{A} = diag(3/2, 2/3)$ and $\xi = \pi/4$ with (a) $\boldsymbol{\beta} = [2, 2]^T, \boldsymbol{\gamma} = [1, 1]^T$, (b) $\boldsymbol{\beta} = [0, 5]^T, \boldsymbol{\gamma} = [2, 2]^T$, (c)$\boldsymbol{\beta} = [0, -5]^T, \boldsymbol{\gamma} = [2, 2]^T$ and (d) $\boldsymbol{\beta} = [0, -5]^T, \boldsymbol{\gamma} = [2, 10]^T$. The (e,f) panels correspond to $\boldsymbol{\Sigma} = \boldsymbol{I}_2$ with (e) $\boldsymbol{\beta} = [-2, 2]^T, \boldsymbol{\gamma} = [1, 1]^T$, (f) $\boldsymbol{\beta} = [0, -5]^T, \boldsymbol{\gamma} = [1, 1]^T$. Bottom panels: contour plots of various multiple scaled (solid lines) and standard (red dashed lines) GH distributions all with $\boldsymbol{\Sigma} = \boldsymbol{I}_2$, $\boldsymbol{\beta} = [1, 1]^T$, $\boldsymbol{\gamma} = [2, 2]^T$, $\boldsymbol{\delta} = [1, 1]^T$ and (g) $\boldsymbol{\lambda} = [-1/2, 2]^T$, (h) $\boldsymbol{\lambda} = [-2, 2]^T$, (i) $\boldsymbol{\lambda} = [2, -1/2]^T$.

$$\mathrm{Var}[\mathbf{Y}_{GH}] = \frac{\delta}{\gamma}\frac{K_{\lambda+1}(\delta\gamma)}{K_\lambda(\delta\gamma)}\mathbf{\Sigma} + \frac{\delta^2}{\gamma^2}\left(\frac{K_{\lambda+2}(\delta\gamma)}{K_\lambda(\delta\gamma)} - \frac{K^2_{\lambda+1}(\delta\gamma)}{K^2_\lambda(\delta\gamma)}\right)\mathbf{\Sigma}\boldsymbol{\beta}^T\boldsymbol{\beta}\mathbf{\Sigma}\,. \quad (14)$$

In both expressions (13) and (14), the skewness parameter $\boldsymbol{\beta}$ affects the correlation structure. This is not always the case. As noted by Sahu et al. (2003), in their *unrestricted* characterization of the skew-t distribution, the skewness parameter does not affect the correlation structure. In their case, the skewness parameter acts only on the diagonal elements of the covariance matrix.

A notable difference between the covariance structure of the MSGH and the standard GH is that in the case of a diagonal scale matrix $\mathbf{\Sigma}$, variates of the MSGH are independent of each other. Interestingly, this is not the case for the standard multivariate GH where the same latent factor W is shared across dimensions, and this effectively acts to induce some degree of dependency between dimensions (although they may be uncorrelated). A similar situation arises in the case of other distributions with shared latent factors, for example the standard t-distribution. As mentioned previously, in the MSGH case the latent factor W is allowed to vary independently across dimensions.

In terms of the marginals of the MSGH distribution, they are easy to sample from but computing their pdfs involves, in general, numerical integration. Details are given in section D of the Supplementary Materials.

3. Maximum likelihood estimation of parameters

In this section, we outline an EM approach to estimate the parameters of the MSNIG distribution as it appears to be the most popular case of the GH family. As noted also by (Protassov, 2004; Barndorff-Nielsen, 1997), for the GH distribution it can be very difficult to show a significant difference between different values of λ due to the flatness of the likelihood and computational difficulties arise in some cases where the likelihood can be infinite. For these reasons we outline the particular case of allowing all λ_m's to be fixed but we note that it is relatively straightforward to extend our proposed approach to the more general case. Also for identifiability reasons, we set all δ_m's to the same δ value so

that the parameters to estimate in the MSNIG case are $\boldsymbol{\Psi} = \{\boldsymbol{\mu}, \boldsymbol{D}, \boldsymbol{A}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \delta\}$ with $|\boldsymbol{A}| = 1$.

Estimation of most of the parameters for the MSNIG distribution is relatively straightforward but the separate estimation of $\boldsymbol{D}$ and $\boldsymbol{A}$ requires an additional minimization algorithm based on the Flury and Gautschi algorithm (Flury, 1984; Flury and Gautschi, 1986).

Let us consider an *i.i.d* sample $\mathbf{y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ of the MSNIG distribution defined in (10). As in the standard NIG distribution case (Karlis, 2002), a convenient computational advantage of the EM approach is to view the weights as an additional missing variable $\boldsymbol{W}$. The observed data $\mathbf{y}$ are seen as being incomplete and additional missing weight variables $\boldsymbol{W}_1 \ldots \boldsymbol{W}_N$ with for $i \in \{1 \ldots N\}$, $\boldsymbol{W}_i = [W_{i1} \ldots W_{iM}]^T$ are introduced. These weights are defined so that $\forall i \in \{1 \ldots N\}$:

$$\begin{aligned} \mathbf{Y}_i | \boldsymbol{W}_i = \boldsymbol{w}_i &\sim \mathcal{N}_M(\boldsymbol{\mu} + \boldsymbol{D}\boldsymbol{\Delta}_{\mathbf{w}_i}\boldsymbol{A}\boldsymbol{D}^T\boldsymbol{\beta}, \boldsymbol{D}\boldsymbol{\Delta}_{\mathbf{w}_i}\boldsymbol{A}\boldsymbol{D}^T) \ , \\ \text{and} \quad \boldsymbol{W}_i &\sim \mathcal{IG}(\gamma_1, \delta) \otimes \cdots \otimes \mathcal{IG}(\gamma_M, \delta) \ , \end{aligned} \tag{15}$$

where $\boldsymbol{\Delta}_{\mathbf{w}_i} = \mathrm{diag}(w_{i1}, \ldots, w_{iM})$.

As a way of circumventing the restriction that the determinant $|\boldsymbol{A}| = 1$ in the M-step, representation (15) above can be rewritten equivalently as,

$$\begin{aligned} \mathbf{Y}_i | \boldsymbol{W}_i = \boldsymbol{w}_i &\sim \mathcal{N}_M(\boldsymbol{\mu} + \boldsymbol{D}\boldsymbol{\Delta}_{\mathbf{w}_i}\boldsymbol{D}^T\tilde{\boldsymbol{\beta}}, \boldsymbol{D}\boldsymbol{\Delta}_{\mathbf{w}_i}\tilde{\boldsymbol{A}}\boldsymbol{D}^T) \ , \\ \boldsymbol{W}_i &\sim \mathcal{IG}(\tilde{\gamma}_1, 1) \otimes \cdots \otimes \mathcal{IG}(\tilde{\gamma}_M, 1) \ , \end{aligned} \tag{16}$$

where $\tilde{\boldsymbol{A}} = \delta^2 \ \boldsymbol{A}$, $\tilde{\boldsymbol{\beta}} = \boldsymbol{D}\tilde{\boldsymbol{A}}\boldsymbol{D}^T\boldsymbol{\beta}$, $\tilde{\boldsymbol{\gamma}} = \delta \ \boldsymbol{\gamma}$ and $\tilde{\boldsymbol{A}}$ is now a general (positive definite) diagonal matrix. Note that in the location term in the definition above (16), $\boldsymbol{D}\boldsymbol{\Delta}_{\mathbf{w}_i}\boldsymbol{D}^T\tilde{\boldsymbol{\beta}} = \boldsymbol{D}\boldsymbol{\Delta}_{\mathbf{w}_i}\tilde{\boldsymbol{A}}\boldsymbol{D}^T\boldsymbol{\beta}$.

3.1. E step

At iteration (r) with $\boldsymbol{\psi}^{(r)}$ being the current parameter value, the E-step leads to the computation for all $i = 1, \ldots, N$, of the missing variables posterior distributions $p(\boldsymbol{w}_i | \mathbf{y}_i; \boldsymbol{\psi}^{(r)})$. It consists then of calculating $p(\boldsymbol{w}_i | \mathbf{y}_i; \boldsymbol{\psi}^{(r)}) \propto$

$p(\mathbf{y}_i|\boldsymbol{w}_i;\boldsymbol{\psi}^{(r)})p(\boldsymbol{w}_i;\boldsymbol{\psi}^{(r)})$ which can be shown (see Appendix of Karlis and Santourian, 2009) to follow a GIG distribution (see definition (5)). In our case, and assuming the $\boldsymbol{W}_i$'s are independent we have,

$$p(\boldsymbol{w}_i|\mathbf{y}_i;\boldsymbol{\psi}^{(r)}) = \prod_{m=1}^{M} \mathcal{GIG}(w_{im};-1,\hat{\alpha}_m^{(r)},\phi_{im}^{(r)}), \tag{17}$$

where

$$\begin{aligned}
\phi_{im}^{(r)} &= \sqrt{1+\frac{[\boldsymbol{D}^{(r)T}(\mathbf{y}_i-\boldsymbol{\mu}^{(r)})]^2_{(m)}}{\tilde{\boldsymbol{A}}_m^{(r)}}}\ , \\
\hat{\alpha}_m^{(r)} &= \sqrt{\tilde{\gamma}_m^{(r)2}+\frac{[\boldsymbol{D}^{(r)T}\tilde{\boldsymbol{\beta}}^{(r)}]^2_m}{\tilde{\boldsymbol{A}}_m^{(r)}}}\ .
\end{aligned}$$

As all moments of a GIG distribution exist (see (11)), it follows that we have closed form expressions for the following quantities needed in the E-step,

$$\begin{aligned}
s_{im}^{(r)} &= E[W_{im}|\mathbf{y}_i;\boldsymbol{\psi}^{(r)}] = \frac{\phi_{im}^{(r)}K_0(\phi_{im}^{(r)}\hat{\alpha}_m^{(r)})}{\hat{\alpha}_m^{(r)}K_{-1}(\phi_{im}^{(r)}\hat{\alpha}_m^{(r)})}\ , \\
t_{im}^{(r)} &= E[W_{im}^{-1}|\mathbf{y}_i;\boldsymbol{\psi}^{(r)}] = \frac{\hat{\alpha}_m^{(r)}K_{-2}(\phi_{im}^{(r)}\hat{\alpha}_m^{(r)})}{\phi_{im}^{(r)}K_{-1}(\phi_{im}^{(r)}\hat{\alpha}_m^{(r)})}\ .
\end{aligned}$$

Note that equivalently $K_{-1}=K_1$ and $K_{-2}=K_2$. The Bessel function can be numerically evaluated in most statistical packages. All computations in this paper were undertaken using R (Team, 2011).

3.2. M step

For the updating of $\boldsymbol{\psi}$, the M-step consists of two independent steps for $(\boldsymbol{\mu},\boldsymbol{D},\tilde{\boldsymbol{A}},\tilde{\boldsymbol{\beta}})$ and $\tilde{\boldsymbol{\gamma}}$,

$$\begin{aligned}
(\boldsymbol{\mu},\boldsymbol{D},\tilde{\boldsymbol{A}},\tilde{\boldsymbol{\beta}})^{(r+1)} &= \arg\max_{\boldsymbol{\mu},\boldsymbol{D},\boldsymbol{A},\boldsymbol{\beta}} \sum_{i=1}^{N} E[\log p(\mathbf{y}_i,|\mathbf{W}_i;\boldsymbol{\mu},\boldsymbol{D},\tilde{\boldsymbol{A}},\tilde{\boldsymbol{\beta}})|\mathbf{y}_i,\boldsymbol{\psi}^{(r)}] \qquad (18) \\
&= \arg\max_{\boldsymbol{\mu},\boldsymbol{D},\boldsymbol{A},\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} -\frac{1}{2}\log|\tilde{\boldsymbol{A}}| - \frac{1}{2}(\mathbf{y}_i-\boldsymbol{\mu}-\boldsymbol{D}\boldsymbol{S}_i^{(r)}\boldsymbol{D}^T\tilde{\boldsymbol{\beta}})^T \right. \\
&\quad \times \left. \boldsymbol{D}\tilde{\boldsymbol{A}}^{-1}\boldsymbol{T}_i^{(r)}\boldsymbol{D}^T(\mathbf{y}_i-\boldsymbol{\mu}-\boldsymbol{D}\boldsymbol{S}_i^{(r)}\boldsymbol{D}^T\tilde{\boldsymbol{\beta}}) \right\},
\end{aligned}$$

and

$$\begin{aligned}\tilde{\boldsymbol{\gamma}}^{(r+1)} &= \arg\max_{\tilde{\boldsymbol{\gamma}}} \sum_{i=1}^{N}\sum_{m=1}^{M} E[\log p(W_{im};\tilde{\gamma}_m,1)|\mathbf{y}_i,\boldsymbol{\psi}^{(r)}] \qquad (19)\\ &= \arg\max_{\tilde{\boldsymbol{\gamma}}} \left\{ \sum_{i=1}^{N}\sum_{m=1}^{M} \tilde{\boldsymbol{\gamma}}_m - \frac{1}{2}\tilde{\boldsymbol{\gamma}}_m^2 s_{im}^{(r)}) \right\},\end{aligned}$$

where $\boldsymbol{T}_i^{(r)} = \mathrm{diag}(t_{i1}^{(r)},\ldots,t_{iM}^{(r)})$ and $\boldsymbol{S}_i^{(r)} = \mathrm{diag}(s_{i1}^{(r)},\ldots,s_{iM}^{(r)})$ and ignoring constants.

The optimization of these steps leads to the following update equations.

Updating $\boldsymbol{\mu}$. It follows from (18) that for fixed $\boldsymbol{D}$ and $\boldsymbol{A}$ (ignoring constants)

$$\boldsymbol{\mu}^{(r+1)} = \arg\min_{\boldsymbol{\mu}} \left\{ \sum_{i=1}^{N} (\mathbf{y}_i - \boldsymbol{\mu} - \boldsymbol{D}\boldsymbol{S}_i^{(r)}\boldsymbol{D}^T\tilde{\boldsymbol{\beta}})^T \boldsymbol{D}\tilde{\boldsymbol{A}}^{-1}\boldsymbol{T}_i^{(r)}\boldsymbol{D}^T(\mathbf{y}_i - \boldsymbol{\mu} - \boldsymbol{D}\boldsymbol{S}_i^{(r)}\boldsymbol{D}^T\tilde{\boldsymbol{\beta}}) \right\}, \qquad (20)$$

which by fixing $\boldsymbol{D}$ to the current estimation $\boldsymbol{D}^{(r)}$, leads to

$$\begin{aligned}\boldsymbol{\mu}^{(r+1)} =& \left(\frac{\sum_{i=1}^{N}\boldsymbol{T}_i^{(r)}\boldsymbol{D}^{(r)T}}{N} - N\,(\sum_{i=1}^{N}\boldsymbol{S}_i^{(r)})^{-1} \right)^{-1}\\ &\times \left(\frac{\sum_{i=1}^{N}\boldsymbol{T}_i^{(r)}\boldsymbol{D}^{(r)T}\mathbf{y}_i}{N} - \sum_{i=1}^{N}\mathbf{y}_i\,(\sum_{i=1}^{N}\boldsymbol{S}_i^{(r)})^{-1} \right).\end{aligned}$$

Updating $\tilde{\boldsymbol{\beta}}$. To update $\tilde{\boldsymbol{\beta}}$ we have to minimize the following quantity,

$$\tilde{\boldsymbol{\beta}}^{(r+1)} = \arg\min_{\tilde{\boldsymbol{\beta}}} \left\{ \sum_{i=1}^{N} (\mathbf{y}_i - \boldsymbol{\mu} - \boldsymbol{D}\boldsymbol{S}_i^{(r)}\boldsymbol{D}^T\tilde{\boldsymbol{\beta}})^T \boldsymbol{D}\tilde{\boldsymbol{A}}^{-1}\boldsymbol{T}_i^{(r)}\boldsymbol{D}^T(\mathbf{y}_i - \boldsymbol{\mu} - \boldsymbol{D}\boldsymbol{S}_i^{(r)}\boldsymbol{D}^T\tilde{\boldsymbol{\beta}}) \right\}, \qquad (21)$$

which by fixing $\boldsymbol{D}$ and $\boldsymbol{\mu}$ to their current estimations $\boldsymbol{D}^{(r)}$ and $\boldsymbol{\mu}^{(r+1)}$, leads to

$$\tilde{\boldsymbol{\beta}}^{(r+1)} = \boldsymbol{D}^{(r)}(\sum_{i=1}^{N}\boldsymbol{S}_i^{(r)})^{-1}\boldsymbol{D}^{(r)T}\sum_{i=1}^{N}(\mathbf{y}_i - \boldsymbol{\mu}^{(r+1)}).$$

Updating $\boldsymbol{D}$. Using the equality $x^T\boldsymbol{S}x = \mathrm{trace}(\boldsymbol{S}xx^T)$ for any matrix $\boldsymbol{S}$, it follows that for fixed $\tilde{\boldsymbol{A}}$ and $\boldsymbol{\mu}$, $\boldsymbol{D}$ is obtained by minimizing

$$\begin{aligned}\boldsymbol{D}^{(r+1)} =& \arg\min_{\boldsymbol{D}} \left\{ \sum_{i=1}^{N}\mathrm{trace}(\boldsymbol{D}\boldsymbol{T}_i^{(r)}\tilde{\boldsymbol{A}}^{(r)-1}\boldsymbol{D}^T\boldsymbol{V}_i) + \sum_{i=1}^{N}\mathrm{trace}(\boldsymbol{D}\boldsymbol{S}_i^{(r)}\tilde{\boldsymbol{A}}^{(r)-1}\boldsymbol{D}^T\boldsymbol{B}_i)\right.\\ &\left. - 2\sum_{i=1}^{N}\mathrm{trace}(\boldsymbol{D}\tilde{\boldsymbol{A}}^{(r)-1}\boldsymbol{D}^T\boldsymbol{C}_i) \right\},\end{aligned}$$

where $\boldsymbol{V}_i = (\mathbf{y}_i - \boldsymbol{\mu}^{(r+1)})(\mathbf{y}_i - \boldsymbol{\mu}^{(r+1)})^T, \boldsymbol{B}_i = \tilde{\boldsymbol{\beta}}^{(r+1)}\tilde{\boldsymbol{\beta}}^{(r+1)T}$, $\boldsymbol{C}_i = (\mathbf{y}_i - \boldsymbol{\mu}^{(r+1)})\tilde{\boldsymbol{\beta}}^{(r+1)T}$.

Using current values $\boldsymbol{\mu}^{(r+1)}, \boldsymbol{\beta}^{(r+1)}$ and $\tilde{\boldsymbol{A}}^{(r)}$, the parameter $\boldsymbol{D}$ can be updated using an algorithm derived from Flury and Gautschi (see Celeux and Govaert (1995)) which is outlined in section E of the Supplementary Materials. Although not considered in this work, in a model-based clustering context, additional information for an efficient implementation can be found in Lin (2014).

Updating $\tilde{\boldsymbol{A}}$. To update $\tilde{\boldsymbol{A}}$ we have to minimize the following quantity (See section F of the Supplementary Materials).

$$\tilde{\boldsymbol{A}}^{(r+1)} = \arg\min_{\tilde{\boldsymbol{A}}} \left\{ \text{trace}((\sum_{i=1}^{N} \boldsymbol{M}_i)\ \tilde{\boldsymbol{A}}^{-1}) + N\log\tilde{\boldsymbol{A}} \right\},$$

where $\boldsymbol{M}_i = \boldsymbol{T}_i^{(r)1/2}\boldsymbol{D}^{(r+1)T}\boldsymbol{V}_i\boldsymbol{D}^{(r+1)}\boldsymbol{T}_i^{(r)1/2} + \boldsymbol{S}_i^{(r)1/2}\boldsymbol{D}^{(r+1)T}\boldsymbol{B}_i\boldsymbol{D}^{(r+1)}\boldsymbol{S}_i^{(r)1/2} - \boldsymbol{D}^{(r+1)T}(\boldsymbol{C}_i + \boldsymbol{C}_i^T)\boldsymbol{D}^{(r+1)}$ and $\boldsymbol{M}_i$ is a symmetric positive definite matrix.

Using a corollary found in Celeux and Govaert (1995) (See section F of the Supplementary Materials) and by setting $\boldsymbol{D}$ and $\boldsymbol{\mu}$ to their current estimations $\boldsymbol{D}^{(r+1)}$ and $\boldsymbol{\mu}^{(r+1)}$ we find for all m,

$$\begin{aligned}\tilde{\boldsymbol{A}}_m^{(r+1)} =& \frac{1}{N}\sum_{i=1}^{N}\Big([\boldsymbol{D}^{(r+1)T}(\mathbf{y}_i - \boldsymbol{\mu}^{(r+1)})]_m^2 t_{im}^{(r)} + [\boldsymbol{D}^{(r+1)T}\tilde{\boldsymbol{\beta}}^{(r+1)}]_m^2 s_{im}^{(r)} \\ &- 2[\boldsymbol{D}^{(r+1)T}(\mathbf{y}_i - \boldsymbol{\mu}^{(r+1)})]_m [\boldsymbol{D}^{(r+1)T}\tilde{\boldsymbol{\beta}}^{(r+1)}]_m \Big).\end{aligned} \tag{22}$$

Updating $\tilde{\boldsymbol{\gamma}}$. It follows from (19) that to update $\tilde{\boldsymbol{\gamma}}$ we have to minimize,

$$\tilde{\boldsymbol{\gamma}}_m^{(r+1)} = \arg\min_{\tilde{\boldsymbol{\gamma}}} \left\{ \sum_{i=1}^{N}\sum_{m=1}^{M} \frac{1}{2}\tilde{\boldsymbol{\gamma}}_m^2 s_{im}^{(r)} - \tilde{\boldsymbol{\gamma}}_m) \right\}, \tag{23}$$

which for all $m = 1, \ldots, M$ leads to $\tilde{\boldsymbol{\gamma}}_m^{(r+1)} = \dfrac{N}{\sum_{i=1}^{N} s_{im}^{(r)}}$.

Updating constrained $\tilde{\boldsymbol{\gamma}}$. Similar updating equations can be easily derived when $\tilde{\boldsymbol{\gamma}}$ is assumed to be equal for several dimensions. If we assume that for all

m, $\tilde{\gamma}_m = \tilde{\gamma}$ then

$$\tilde{\boldsymbol{\gamma}}^{(r+1)} \quad = \quad \frac{NM}{\sum_{i=1}^{N}\sum_{m=1}^{M} s_{im}^{(r)}} \, .$$

It is also quite easy to extend the above equation to the case where $\tilde{\boldsymbol{\gamma}}$ is assumed to be equal for only some of the dimensions. For either case, model choice criteria could be used to justify the appropriateness of the assumed parameter space for $\tilde{\boldsymbol{\gamma}}$.

Eventually, to transform the estimated parameters back to their original form we can take $\delta = |\tilde{\boldsymbol{A}}|^{\frac{1}{2M}}, \gamma_m = \tilde{\boldsymbol{\gamma}}_m/\delta, \boldsymbol{\beta} = \boldsymbol{D}\tilde{\boldsymbol{A}}^{-1}\boldsymbol{D}^T\tilde{\boldsymbol{\beta}}$ and $\boldsymbol{A} = \tilde{\boldsymbol{A}}/|\tilde{\boldsymbol{A}}|^{1/M}$.

3.3. Mixture of multiple scaled NIG distributions

The previous results can be extended to cover the case of K-component mixture of MSNIG distributions. With the usual notation for the proportions $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_K\}$ and $\boldsymbol{\psi}_k = \{\boldsymbol{\mu}_k, \boldsymbol{D}_k, \boldsymbol{A}_k, \boldsymbol{\beta}_k, \boldsymbol{\gamma}_k, \delta_k\}$ for $k = 1 \ldots K$, we consider,

$$p(\mathbf{y}; \boldsymbol{\phi}) = \sum_{k=1}^{K} \pi_k \mathcal{MSNIG}(\mathbf{y}; \boldsymbol{\mu}_k, \boldsymbol{D}_k, \boldsymbol{A}_k, \boldsymbol{\beta}_k, \boldsymbol{\gamma}_k, \delta_k) \, ,$$

where k indicates the kth component of the mixture and $\boldsymbol{\phi} = \{\boldsymbol{\pi}, \boldsymbol{\psi}\}$ with $\boldsymbol{\psi} = \{\boldsymbol{\psi}_1, \ldots \boldsymbol{\psi}_K\}$ the mixture parameters. Details on the EM framework are given in section G of the Supplementary Materials.

As the results of the EM algorithm can be particularly sensitive to initial values (Karlis and Xekalaki, 2003), for the results to follow we used a number of approaches to generate different initial values for parameters, including the use of random partitions, k-means and trimmed k-means (Garcia-Escudero and Gordaliza, 1999). Often the most successful strategy found was by estimating $\boldsymbol{\mu}_k$, $\boldsymbol{D}_k$ and $\boldsymbol{A}_k$ using the results from a trimmed k-means clustering (with $\boldsymbol{\beta}_k = 0$) and setting $\gamma_{km} = \delta_k = 1$ for all $k = 1 \ldots K$ and $m = 1 \ldots M$. The computational speed of the EM algorithm for the MSNIG distribution is comparable to the standard NIG case with the exception that the update of $\boldsymbol{D}$ can be slow for high dimensional applications as the Flury and Gautschi algorithm involves sequentially updating every pair of column vectors of $\boldsymbol{D}$.

A more global approach to the update of $\boldsymbol{D}$ has been proposed recently by Browne and McNicholas (2012) which has the potential to significantly speed up the computation time. Browne and McNicholas (2014) have also proposed the use of majorization-minimization algorithms for the same purpose.

4. Applications of multiple scaled NIG distributions

In this section we present an application of the MSNIG distribution on a real dataset to demonstrate its flexibility in analyzing skewed multivariate data. Additional results on simulated data and two more real datasets are reported in sections H, I and K of the Supplementary Materials where the performance of the MSNIG compares favourably to the standard NIG and other tested distributions (see section 4.1), with the MSNIG providing a better fit.

4.1. Lymphoma data

To illustrate some of the differences between the standard NIG and MSNIG we examine a clustering problem for a lymphoma dataset recently analysed by Lee and McLachlan (2013c). The data consists of a subset of data originally presented and collected by Maier et al. (2007). In Maier et al. (2007) blood samples from 30 subjects were stained with four fluorophore-labeled antibodies against $CD4$, $CD45RA$, $SLP76(pY128)$, and $ZAP70(pY292)$ before and after an anti-$CD3$ stimulation. In the first example we will look at clustering a subset of the data containing the variables $CD4$ and $ZAP70$ (Figure 2), which appear to be bimodal and display an asymmetric pattern. In particular, one of the modes appears to show both strong correlation between the two variables and substantial skewness.

Of interest in this example is to compare the goodness of fit from fitting mixtures of standard NIG and MSNIG distributions. For comparison, we also present the results of fitting using mixtures of skew-normal (Lachos et al., 2010) and skew-t distributions using two types of formulation: the *unrestricted* characterization (Sahu et al., 2003; Lee and McLachlan, 2014b; Lin, 2010) and the *restricted* one (Azzalini and Capitanio, 2003; Basso et al., 2010; Branco and

Dey, 2001; Cabral et al., 2012; Pyne et al., 2009). Estimation of the parameters for these distributions was undertaken using the R package **mixsmsn** (Cabral et al., 2012) and for the unrestricted skew-t case using R code available on: http://www.maths.uq.edu.au/ gjm/mix_soft/EMMIX-skew/index.html. Also available on CRAN, R package **EMMIXuskew** (Lee and McLachlan, 2013a).

We also show the results obtained with the mixture of coalesced GH distributions described in (Tortora et al., 2014b) and implemented in the R package **MixGHD** available on the CRAN (Tortora et al., 2014a). The so-called coalesced GH distribution is a mixture of two distributions, a standard GH and a multiple scaled GH distributions. Although their definitions of the GIG and GH distributions are equivalent to ours, the resulting formula for the multiple scaled GH distribution (eq. (19) in Tortora et al. (2014b)) is different. Moreover, we provide an exact EM algorithm for inference which is not the case for the algorithm provided in (Tortora et al., 2014b). For comparison, we ran the MixGHD package to fit two versions of the coalesced GH mixture. The first version corresponds to the most general model (denoted by *Coalesced GH*). In the second version, we fitted a mixture of multiple scaled GH distributions by setting the corresponding inner weights in the coalesced distributions to 0. This version is denoted by $MSGH^{TFBM}$ to distinguish it from our own MSGH distribution. Both results are different from the results obtained with our algorithm (see Figure 2 (e,f)) and do not provide realistic classification results as shown in Section J of our Supplementary Materials (See also the clearly worse likelihood and BIC values reported in Table 1). The reason for this is twofold. First, as mentioned in the introduction, the formula defining the multiple scaled GH distribution in Tortora et al. (2014b) is not equivalent to ours. Further, the constraint imposed on the parameterization, namely $\boldsymbol{\delta} = \boldsymbol{\gamma}$ in our notation, makes their multiple scaled GH not as flexible, particularly in modeling different tail behaviors ($\boldsymbol{\gamma}$ parameter).

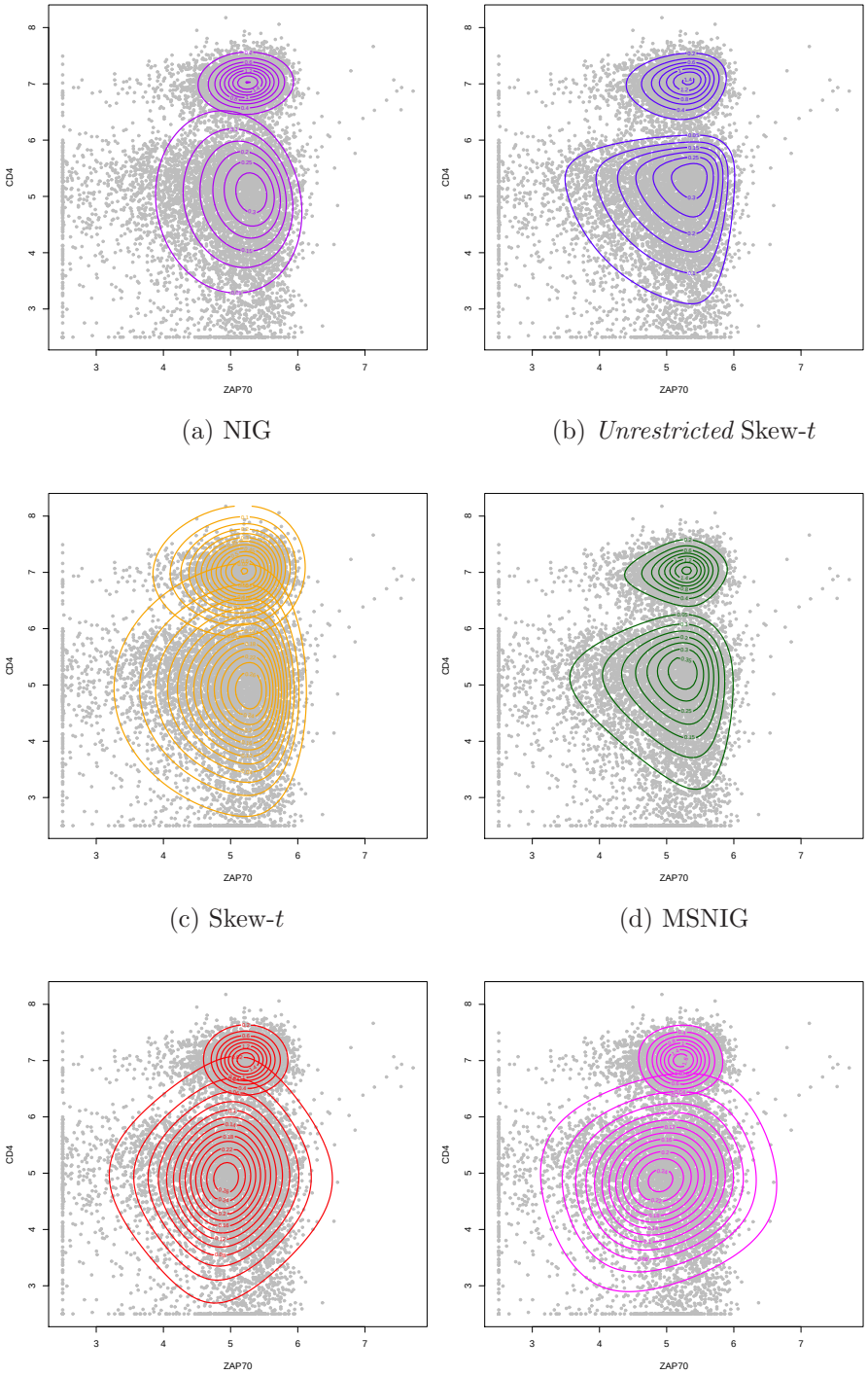
Figures 2 (a) to (d) show the separate contour lines (of each component) from fitting mixtures of: standard NIG (Karlis and Santourian, 2009)(a); unrestricted Skew-t (Sahu et al., 2003; Lee and McLachlan, 2014b; Lin, 2010) (b);

Skew-t (Azzalini and Capitanio, 2003; Basso et al., 2010; Branco and Dey, 2001; Cabral et al., 2012; Pyne et al., 2009) (c); and MSNIG (d). Likelihood values and estimates of the BIC for the different approaches are also provided in Table 1. As we can see from Figure 2 there is quite a difference in the goodness of fit between the approaches. In particular, we see a clear difference in the fitted results between the standard NIG and MSNIG with the latter providing a closer fit to the data. Similar results to the standard NIG are obtained for the *restricted* Skew-t (c) and Skew-normal (Lachos et al., 2010) (not shown) approaches. Note that in the **mixsmsn** package used, the Skew-t mixture is fitted with equal degree-of-freedom parameters for each component. Interestingly the fitted results of the *unrestricted* Skew-t (b) and the MSNIG (d) appear to be similar. BIC values for these two approaches are also similar (MSNIG = 47,266, *unrestr.* Skew-t = 47,140) but with more support for the *unrestricted* Skew-t.

Table 1: Results for Lymphoma dataset (Par. is the number of parameters)

	Example 1 ($CD4$ v. $ZAP70$)			Example 2 ($CD45$ v.$CD4$)		
Model	Log-likelihood	Par.	BIC	Log-likelihood	Par.	BIC
MSNIG	-23,545	19	47,266	-16,444	39	33,219
NIG	-23,842	17	47,841	-16,573	35	33,443
Skew-t (Unrestr.)	-23,492	17	47,140	-16,540	35	33,378
Skew-t	-23,868	16	47,672	-16,561	32	33,394
Skew-normal	-23,762	15	47,874	-16,573	31	33,410
Coalesced GH	-24,477	43	49,350	-18,319	87	37,379
MSGH^{TFBM}	-24,754	23	49,720	-17,509	47	35,418

As suggested by Lee and McLachlan (2013b) a possible reason for the difference in the results between the unrestricted Skew-t and the skew-normal and Skew-t is the differing impact of the skewness parameter on the correlation structure. As mentioned previously, in the skew-t formulation of Sahu et al. (2003) the skewness parameter acts only on diagonal elements of the covariance matrix and does not affect the correlation structure, which is not the case for the other formulations of the skew-t and skew-normal approaches.



(a) NIG

(b) *Unrestricted* Skew-t

(c) Skew-t

(d) MSNIG

(e) Coalesced GH

(f) MSGH^{TFBM}

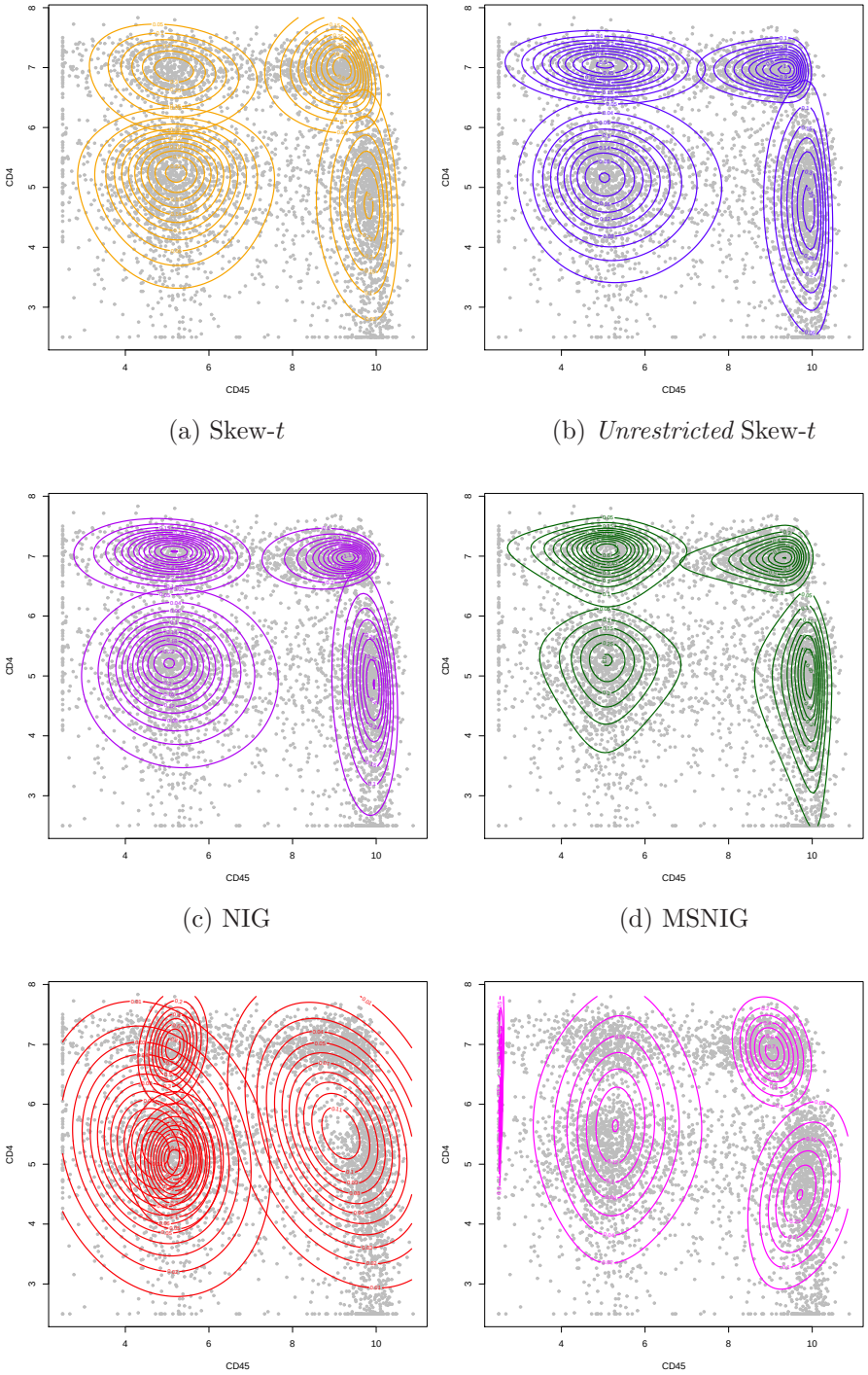
Figure 2: Lymphoma data, $CD4$ v. $ZAP70$. Fitted contour lines for: (a) Standard NIG (Karlis and Santourian, 2009); (b) *Unrestricted* Skew-t (Sahu et al., 2003); (c) Skew-t (Azzalini and Capitanio, 2003); (d) Multiple scaled NIG; (e) Coalesced GH (Tortora et al., 2014b) and (f) Multiple scaled GH (Tortora et al., 2014b).

We now consider a second example to highlight further differences between the standard NIG and MSNIG in a clustering context using the same dataset. In this example we look at a subset of the dataset containing the variables $CD45$ and $CD4$, which also appear to be highly multimodal and asymmetric in shape. The fitted results from a mixture model with four components are shown in Figure 3 with contour lines representing the fitted density of each component (see also results in Table 1). From the fitted results we can see a better fit from the MSNIG (BIC = 33,219) compared to the standard NIG (BIC = 33,443). The better fit appears to come from the increased flexibility of the MSNIG to represent non-elliptical shapes. The fitted results for the Skew-t and *unrestr.* Skew-t are slightly better than for the standard NIG (BIC = 33,394 and 33,378, respectively). Similar results to the Skew-t are found for the Skew-normal (not shown).

Results using the **MixGHD** package as specified above are also shown in Figure 3 (e) and (d). The final results are not very satisfying although we checked that the algorithm started from a good initialization (see Figure 9 in the Supplementary Materials).

In section J of the Supplementary Materials, we also provide the results found using the **MixGHD** package when λ is set to -1/2, which corresponds to NIG distributions. The resulting MSNIG^{TFBM} distribution (supplementary Figure 10) does not behave much better than its MSGH^{TFBM} generalization. Also we observed (supplementary Figure 11) that the GH parameterization proposed in (Browne and McNicholas, 2013) provided results very close to the standard NIG distribution (Figure 2 (a) and Figure 3 (c)).

In section K of the Supplementary Materials we also compare the classification performance of the different approaches on a flow cytometry problem using lymphoma data where the true group labels are known (through manual gating).



(a) Skew-t

(b) *Unrestricted* Skew-t

(c) NIG

(d) MSNIG

(e) Coalesced GH

(f) MSGH^{TFBM}

Figure 3: Lymphoma data, $CD45$ v. $CD4$. Fitted contour lines for: (a) Skew-t; (b) *Unrestricted* Skew-t; (c) Standard NIG; (d) Multiple scaled NIG; (e) Coalesced GH (Tortora et al., 2014b) and (f) Multiple scaled GH (Tortora et al., 2014b).

5. Conclusion

We have proposed a relatively simple way to extend *location and scale mixture distributions*, such as the multivariate generalised hyperbolic distribution (GH), to allow for different tail behaviour in each dimension. In contrast to existing approaches, the main advantages include closed form densities, the possibility of arbitrary correlation between dimensions and the applicability to high dimensional spaces. Various properties of the MSGH family are well defined and estimation of the parameters is also relatively straightforward using the familiar EM algorithm. Assessments of the performance of the proposed model on simulated and real data suggest that the extension provides a considerable degree of freedom and flexibility in modelling data of varying tail behaviour and directional shape.

For future research, parsimonious models could be considered using special decompositions of the scale matrix such as in the model-based clustering approach of Celeux and Govaert (1995) and Fraley and Raftery (2002), which would be straightforward to generalize to multiple scaled distributions (see O'Hagan et al. (2014) for mixtures of standard NIG distributions). Similarly, for very high dimensional data, other parsimonious models could also be considered with special modelling of the covariance matrix such as in the High Dimensional Data Clustering (HDDC) framework of Bouveyron et al. (2007). As it is natural in an EM setting, learning with missing observations could also be addressed following the work of Lin et al. (2006); Lin (2014); Wang (2015) with interesting applications including sound source separation and localization (*e.g.* Deleforge et al. (2015)).

Although we have illustrated the approach on clustering examples, the multiple scaled NIG is applicable to other contexts including, for example, regression modelling (Young and Hunter, 2010; Hunter and Young, 2012), outlier detection and modelling of spatial data (Forbes et al., 2010). An R package for the proposed approach will also be available in the near future.

References

Aas, K., Hobaek Haff, I., 2006. The generalised hyperbolic skew Student's t-distribution. Journal of Financial Econometrics 4 (2), 275–309.

Aas, K., Hobaek Haff, I., Dimakos, X., 2005. Risk estimation using the multivariate normal inverse Gaussian distribution. Journal of Risk 8 (2), 39–60.

Azzalini, A., Capitanio, A., 2003. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution. Journal of the Royal Statististical Society B 65, 367–389.

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

Barndorff-Nielsen, O., Kent, J., Sorensen, M., 1982. Normal variance-mean mixtures and z Distributions. International Statistics Review 50 (2), 145–149.

Basso, R., Lachos, V., Cabral, C., Ghosh, P., 2010. Robust mixture modelling based on scale mixtures of skew-normal distributions. Computational Statistics and Data Analysis 54, 2926–2941.

Benaglia, T., Chauveau, D., Hunter, D., 2009a. An EM-like algorithm for semi- and nonparametric estimation in multivariate mixtures. Journal of Computational and Graphical Statistics 18, 505–526.

Benaglia, T., Chauveau, D., Hunter, D., Young, D., 2009b. mixtools: An R Package for Analyzing Finite Mixture Models. Journal of Statistical Software 32 (6).

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

Branco, M., Dey, D., 2001. A general class of multivariate skew-elliptical distributions. Journal of Multivariate Analysis 79, 99–113.

Browne, R., McNicholas, P., 2012. Orthogonal Stiefel manifold optimization for eigen-decomposed covariance parameter estimation in mixture models. Statistics and Computing Published online.

Browne, R., McNicholas, P., 2013. A mixture of generalized hyperbolic distributions, arXiv:1305.1036.

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

Cabral, C., Lachos, V., Prates, M., 2012. Multivariate mixture modelling using skew-normal independent distributions. Computational Statistics and Data Analysis 56, 126–142.

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

Chang, G., Walther, G., 2007. Clustering with mixtures of log-concave distributions. Computational Statistics and Data Analysis 51, 6242–6251.

Deleforge, A., Forbes, F., Horaud, R., 2015. Acoustic Space Learning for Sound-Source Separation and Localization on Binaural Manifolds. International Journal of Neural Systems 25 (1).

Ferreira, J. T. A. S., Steel, M. F. J., 2007a. Model comparison of coordinate-free multivariate skewed distributions with an application to stochastic frontiers. Journal of Econometrics 137, 641–673.

Ferreira, J. T. A. S., Steel, M. F. J., 2007b. A new class of multivariate skew distributions with applications to regression analysis. Statistica Sinica 17, 505–529.

Flury, B. N., 1984. Common Principal Components in K Groups. Journal of the American Statistical Association 79 (388), 892–898.

Flury, B. N., Gautschi, W., 1986. An Algorithm for Simultaneous Orthogonal Transformation of Several Positive Definite Symmetric Matrices to Nearly

Diagonal Form. SIAM Journal on Scientific and Statistical Computing 7 (1), 169–184.

Forbes, F., Doyle, S., Garcia-Lorenzo, D., Barillot, C., Dojat, M., 13-15 May 2010. A Weighted Multi-Sequence Markov Model For Brain Lesion Segmentation. In: 13th International Conference on Artificial Intelligence and Statistics (AISTATS10). Sardinia, Italy.

Forbes, F., Wraith, D., 2014. A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweight: Application to robust clustering. Statistics and Computing 24 (6), 971–984.

Fraley, C., Raftery, A. E., 2002. Model-Based Clustering, Discriminant Analysis, and Density Estimation. Journal of the American Statistical Association 97, 611–631.

Franczak, B., Tortora, C., Browne, R., McNicholas, P., 2014. Mixtures of skewed distributions with hypercube contours, arXiv:1403.2285v4.

Fruwirth-Schnatter, S., 2006. Finite Mixture and Markov Switching Models. Springer Series in Statistics.

Garcia-Escudero, L., Gordaliza, A., 1999. Robustness properties of k-means and Trimmed k-means. Journal of the American Statistical Association 94 (447), 956–969.

Gjerde, T., Eidsvik, J., Nyrnes, E., Bruun, B., 2011. Positioning and Position Error of Petroleum Wells. Journal of Geodetic Science 1, 158–169.

Hunter, D., Young, D., 2012. Semiparametric Mixtures of Regressions. Journal of Nonparametric Statistics 24 (1), 19–38.

Jorgensen, B., 1982. Statistical Properties of the Generalized Inverse Gaussian Distribution. In: Lecture Notes in Statistics. Springer, New York.

Karlis, D., 2002. An EM type algorithm for maximum likelihood estimation of the normal inverse Gaussian distribution. Statistics and Probability letters 57, 43–52.

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

Karlis, D., Xekalaki, E., 2003. Choosing initial values for the EM algorithm for finite mixtures. Computational Statistics and Data Analysis 41 (3-4), 577–590.

Kotz, S., Nadarajah, S., 2004. Multivariate t Distributions and their Applications. Cambridge.

Lachos, V., Ghosh, P., Arellano-Valle, R., 2010. Likelihood based inference for skew normal independent mixed models. Statistica Sinica 20, 303–322.

Lee, S., McLachlan, G., 2013a. EMMIXuskew: an R package for fitting mixtures of multivariate skew t-distributions via the EM algorithm. Journal of Statistical Software 55 (12).

Lee, S., McLachlan, G., 2013b. Model-based clustering and classification with non-normal mixture distributions (with discussion). Statistical Methods and Applications 22, 427–479.

Lee, S., McLachlan, G., 2013c. On mixtures of skew normal and skew t-distributions. Advances in Data Analysis and Classification 7, 241–266.

Lee, S., McLachlan, G., 2014a. Finite mixtures of canonical fundamental skew t-distributions. arXiv preprint arXiv:1405.0685.

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

Lin, T.-I., 2010. Robust mixture modelling using multivariate skew-t distribution. Statistics and Computing 20, 343–356.

Lin, T.-I., 2014. Learning from incomplete data via parameterized t mixture models through eigenvalue decomposition. Computational Statistics and Data Analysis 71, 183–195.

Lin, T.-I., Ho, H. J., Lee, C.-R., 2014. Flexible mixture modelling using the multivariate skew-t-normal distribution. Statistics and Computing 24 (4), 531–546.

Lin, T.-I., Lee, J. C., Ho, H. J., 2006. On fast supervised learning for normal mixture models with missing information. Pattern Recognition 39 (6), 1177–1187.

Maier, L., Anderson, D., De Jager, P., Wicker, L., Hafler, D., 2007. Allelic variant in ctla4 alters t cell phosphorylation patterns. In: Proceedings of the National Academy of Sciences of the United States of America. Vol. 104. pp. 18607–18612.

O'Hagan, A., Murphy, T. B., Gormley, I. C., McNicholas, P., Karlis, D., 2014. Clustering with the multivariate Normal Inverse Gaussian distribution . Computational Statistics and Data Analysis.

Oigard, T. A., Hanssen, A., Hansen, R. E., 2004. The multivariate normal inverse Gaussian distribution: EM-estimation and analysis of synthetic aperture sonar data. In: XII European Signal Processing Conference, Eusipco. Vienna, Austria.

Protassov, R., 2004. EM-based maximum likelihood parameter estimation for multivariate generalized hyperbolic distributions. Statistics and Computing 14, 67–77.

Pyne, S., Hu, X., Wang, K., 2009. Automated high-dimensional flow cytometric flow analysis. Proceedings of the National Academy of Sciences of the United States of America 106, 8519–8524.

Sahu, S., Dey, D., Branco, M., 2003. A new class of multivariate skew distributions with applications to Bayesian regression models. The Canadian Journal of Statistics 31, 129–150.

Schmidt, R., Hrycej, T., Stutzle, E., 2006. Multivariate distribution models with generalized hyperbolic margins. Computational Statistics and Data Analysis 50, 2065–2096.

Team, R. D. C., 2011. R: A language and environment for statistical computing. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Tortora, C., Browne, R. P., Franczak, B. C., McNicholas, P. D., July 2014a. MixGHD: Model based clustering and classification using the mixture of generalized hyperbolic distributions. Version 1.0.

Tortora, C., Franczak, B., Browne, R., McNicholas, P., 2014b. Model-based clustering using mixtures of coalesced generalized hyperbolic distributions, arXiv:1403.2332v3.

Tortora, C., McNicholas, P., Browne, R., 2013. A mixture of generalized hyperbolic factor analyzers, arXiv:1311.6530.

Vilca, F., Balakrishnan, N., Zeller, C., 2014a. Multivariate Skew-Normal Generalized Hyperbolic distribution and its properties. Journal of Multivariate Analysis 128, 73–85.

Vilca, F., Balakrishnan, N., Zeller, C., 2014b. A robust extension of the bivariate Birnbaum-Saunders distribution and associated inference. Journal of Multivariate Analysis 124, 418–435.

Wang, W., 2015. Mixtures of common -factor analyzers for modeling high-dimensional data with missing values. Computational Statistics and Data Analysis 83 (0), 223 – 235.

Young, D., Hunter, D., 2010. Mixtures of Regressions with Predictor-Dependent Mixing Proportions. Computational Statistics and Data Analysis 54 (10), 2253–2266.