Statistics and Computing (2019) 29:415–428
https://doi.org/10.1007/s11222-018-9815-5

Robust finite mixture modeling of multivariate unrestricted skew-normal generalized hyperbolic distributions

Mohsen Maleki[1] · **Darren Wraith**[2] · **Reinaldo B. Arellano-Valle**[3]

Received: 18 June 2017 / Accepted: 9 May 2018 / Published online: 19 May 2018

Abstract
In this paper, we introduce an unrestricted skew-normal generalized hyperbolic (SUNGH) distribution for use in finite mixture modeling or clustering problems. The SUNGH is a broad class of flexible distributions that includes various other well-known asymmetric and symmetric families such as the scale mixtures of skew-normal, the skew-normal generalized hyperbolic and its corresponding symmetric versions. The class of distributions provides a much needed unified framework where the choice of the best fitting distribution can proceed quite naturally through either parameter estimation or by placing constraints on specific parameters and assessing through model choice criteria. The class has several desirable properties, including an analytically tractable density and ease of computation for simulation and estimation of parameters. We illustrate the flexibility of the proposed class of distributions in a mixture modeling context using a Bayesian framework and assess the performance using simulated and real data.

Keywords Bayesian analysis · Finite mixtures · MCMC · Unrestricted skew-normal generalized hyperbolic family · Skew-normal · Generalized hyperbolic distribution

1 Introduction

Statistical models based on finite mixtures of distributions have been extensively used in a wide variety of applications. Applying finite mixture models to real datasets allows fitting different characteristics of the empirical distribution, such as multimodality, skewness, kurtosis and heterogeneity, across observations. For general reviews of mixture models and applications, see Hogan and Laird (1997), Böhning (2000), McLachlan and Peel (2000), Frühwirth-Schnatter (2006), Lin (2010) and Mengersen et al. (2011).

While the vast majority of work on mixture models has focused on Gaussian mixture models, in many applications the tails of Gaussian distributions are shorter than appropriate and the Gaussian shape is not suitable for highly asymmetric data. Recent research has thus focused on fitting finite mixture models with more flexible distributional forms. The Student-t and the contaminated Gaussian distributions are two symmetric members of the scale mixtures of normal (SMN) family of distributions due to Andrews and Mallows (1974), which provide attractive heavy-tailed alternatives to the Gaussian distribution. Building upon this work is the class of scale mixtures of skew-normal (SMSN) distributions proposed by Branco and Dey (2001). The class of SMSN distributions provides location-scale density functions which depend on additional parameters of shape and kurtosis, and includes as special cases the normal and skew-normal (SN) densities, as well as the full SMN class of symmetric densities. Special symmetric and skew-symmetric heavier tail members of the SMSN family are, e.g., the Student-t, Cauchy, skew-t (ST), skew-Cauchy (SC), skew-contaminated normal (SCN) and skew-slash (SSL) distributions. Comprehensive coverage of the fundamental theory and new developments for SN and related distributions is given by Azzalini and Capitanio 2014; see also Genton (2004), Arellano-Valle and Genton (2005, 2010), Arellano-Valle and Azzalini (2006), Arellano-Valle and Azzalini (2006).

✉ Darren Wraith
d.wraith@qut.edu.au

[1] Department of Statistics, Shiraz University, Shiraz, Iran

[2] Institute of Health and Biomedical Innovation (IHBI), Queensland University of Technology (QUT), Brisbane, QLD, Australia

[3] Department of Statistics, Universidad Católica de Chile, Santiago, Chile

Many of the different distributions within the class of SMSN distributions have been developed, and their performance assessed in the context of mixture models. Lin et al. (2007), Lin et al. (2009) and Pyne et al. (2009) studied mixtures of skew-normal distributions. Frühwirth-Schnatter and Pyne (2010) considered Bayesian inference for finite mixtures of univariate and multivariate SN and ST distributions. Basso et al. (2010) considered the robust mixture modeling based on the SMSN family. Wang et al. (2009), Lin (2010), Lee and McLachlan (2014), Vrbik and McNicholas (2012) and Forbes and Wraith (2014) considered mixtures of multivariate ST distribution. Maleki and Arellano-Valle (2017) proposed a time series model based on finite mixtures of SMSN distributions. For a review of mixtures of SN and ST distributions, see Lee and McLachlan (2013a, b).

Other distributional forms within the SMSN family of distributions have also been examined. Karlis and Santourian (2009) developed mixtures of multivariate normal inverse Gaussian distributions. Franczak et al. (2014) examined mixtures of shifted asymmetric Laplace (SAL) distributions. Morris et al. (2014) proposed mixtures of contaminated SAL distributions. Browne and McNicholas (2015) and Wraith and Forbes (2015) examined mixtures of generalized hyperbolic distributions.

In the SMSN class of distributions, although the mixing distribution typically controls the tail behavior, it can also affect the behavior of skewness (Branco and Dey 2001). A recent work on this theme providing flexibility in both skewness and heavy tails has been considered by Vilca et al. (2014) in a class of distributions referred to as multivariate SN generalized hyperbolic (SNGH) distributions. In this setting, the mixing distribution follows a generalized inverse Gaussian distribution (GIG), which has previously been demonstrated to provide considerable flexibility in modeling heavy-tailed data (Wraith and Forbes 2015).

In other recent works, a broad class of skewed distributions has been explored by Lee and McLachlan (2016) in a mixture model context focusing on the unified ST (SUT) distribution (Arellano-Valle and Azzalini 2006) and the fundamental ST distribution (Arellano-Valle and Genton 2005), including as a special case the location-scale variant of the canonical fundamental or unrestricted ST (or skew unified t; SUT) distribution. A particular feature and advantage of the SUT distribution is that it encompasses as special cases the canonical fundamental or unrestricted SN (or skew unified normal; SUN) distribution (Arellano-Valle and Genton 2005) and other SN or ST variants (e.g., Sahu et al. 2003; Arellano-Valle et al. 2007; Lachos et al. 2007, 2010), thus providing considerable flexibility for modeling where the best fitting distribution can be chosen simply (automatically) through parameter estimation or use of model choice criteria.

In this paper, we propose a very general class of distributions which extends the previous work on the SUN and SUT distributions by considering a mixing distribution for this class of models which follows a generalized inverse Gaussian (GIG). We refer to this new family of distributions as an unrestricted skew-normal generalized hyperbolic distribution (SUNGH). The new family provides a very general framework for a large class of distributions and has several desirable properties, including an analytically tractable density and ease of computation for simulation and estimation of parameters. The family also provides a high degree of flexibility for the modeling of complex multivariate data with different degrees of asymmetry, kurtosis and heavy tails. A particular attractiveness of this family of distributions is that it encompasses as special cases all of the distributions previously considered in the SMSN family and extensions to the unrestricted classes (e.g., SUT and SUN). Thus, this class of distributions provides a much needed unified framework where the choice of the best fitting distribution can proceed quite naturally through either parameter estimation or by placing constraints on specific parameters and assessing through model choice criteria. We illustrate the advantages of this new family in the finite mixture modeling context using a Bayesian framework.

There are some computational advantages to using a Bayesian framework in a mixture model setting. First, allowing for the influence or effect of missing data on parameter estimates is quite natural in a Bayesian setting as various patterns of missing data (e.g., class-dependent missingness) can be imputed at each MCMC iteration from the posterior predictive distribution (e.g., using a mixture model defined using open-source software such as JAGS or NIMBLE). In contrast, often quite separate and complex methods are needed for maximum likelihood estimation in these settings (Lin et al. 2009 and Wang et al. 2004). Further, for the complex distributions we consider in this paper, previous work using the EM algorithm has at times relied on approximations (Lee and McLachlan 2016) or calculations of derivatives involving complex functions (Browne and McNicholas 2015) for the estimation of parameters. This difficulty also extends to the estimation of the standard errors for parameters (if they are available) using asymptotic approximations to the observed information matrix if the sample size is large or resorting to a bootstrap method which is computationally demanding (Basso et al. 2010). At times, the standard errors for the parameters are also unavailable (particularly for the GH distribution) or not mentioned (e.g., Browne and McNicholas 2015). This is not to say that estimation in the Bayesian setting is devoid of potential computational issues, in particular the issue of label switching is a more prominent issue compared to methods using ML estimation (Mengersen et al. (2011)).

The paper is organized as follows. In Sect. 2, we provide some background to the SUN and GIG distributions. Section 3 outlines the details and properties of the new SUNGH

family. In Sect. 4, we present a Bayesian analysis of a finite mixture model following a SUNGH distribution. In Sect. 5, we illustrate the performance of the proposed approach on real and simulated data. Finally, in Sect. 6, we present our main conclusions and discuss some areas of further research.

2 SUN and GIG distributions

2.1 Preliminaries

Following Arellano-Valle and Genton (2005), Arellano-Valle and Azzalini (2006) and Arellano-Valle et al. (2007), we say that a $p \times 1$ random vector $\boldsymbol{X}$ follows an unrestricted skew-normal (SUN) with $p \times 1$ location vector $\boldsymbol{\mu}$, $p \times p$ positive definite dispersion matrix $\boldsymbol{\Sigma}$ and $p \times q$ skewness parameter matrix $\boldsymbol{\Lambda}$, denoted by $\boldsymbol{X} \sim \text{SUN}_{p,q}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda})$, if its probability density function (pdf) is

$$f(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}) = 2^q \phi_p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\psi}) \Phi_q\left(\boldsymbol{\Lambda}^\top \boldsymbol{\psi}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) | \boldsymbol{\Upsilon}\right), \quad \boldsymbol{x} \in \boldsymbol{R}^p, \tag{1}$$

where $\boldsymbol{\psi} = \boldsymbol{\Sigma} + \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top$, $\boldsymbol{\Upsilon} = \boldsymbol{I}_q - \boldsymbol{\Lambda}^\top \boldsymbol{\psi}^{-1} \boldsymbol{\Lambda} = \left(\boldsymbol{I}_q + \boldsymbol{\Lambda}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\Lambda}\right)^{-1}$, and $\phi_k(\cdot|\boldsymbol{\mu}, \boldsymbol{\psi})$ and $\Phi_q(\cdot|\boldsymbol{\Upsilon})$ are, respectively, the pdf and cumulative distribution function (cdf) of the multivariate normal distributions given by $N_p(\boldsymbol{\mu}, \boldsymbol{\psi})$ and $N_q(\boldsymbol{0}, \boldsymbol{\Upsilon})$. The SUN class of multivariate distributions defined by (1) contains various special cases. For instance, we recover the multivariate normal when $\boldsymbol{\Lambda} = \boldsymbol{0}$, the multivariate SN which called here restricted SN (rMSN) when $q = 1$, and the multivariate SN of Sahu et al. (2003) when $p = q$ and $\boldsymbol{\Lambda}$ being a diagonal matrix. In fact, the SUN distribution becomes an important special case of the unified SN distribution (SUN) studied by Arellano-Valle and Azzalini (2006).

The random vector $\boldsymbol{X} \sim \text{SUN}_{p,q}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda})$ can be stochastically represented from different ways. According to Arellano-Valle et al. (2006), the SUN random vector $\boldsymbol{X}$ has selection representation given by

$$\boldsymbol{X} = \boldsymbol{\mu} + (\boldsymbol{V}_1 | \boldsymbol{V}_0 > 0), \tag{2}$$

where the condition $\boldsymbol{V}_0 > 0$ means that each element of $\boldsymbol{V}_0$ is positive, and

$$\begin{pmatrix} \boldsymbol{V}_0 \\ \boldsymbol{V}_1 \end{pmatrix} \sim N_{p+q}\left(\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{I}_q & \boldsymbol{\Lambda}^\top \\ \boldsymbol{\Lambda} & \boldsymbol{\Sigma} \end{pmatrix}\right).$$

The representation in (2) becomes a selection representation of the rMSN distribution when $q = 1$, i.e., when the latent vector $\boldsymbol{V}_0$ is replaced by a one-dimensional normal random variable V_0. Also, if we let $\boldsymbol{V}_0 = \boldsymbol{W}_0$ and $\boldsymbol{V}_1 = \boldsymbol{W}_1 + \boldsymbol{\Lambda}\boldsymbol{W}_0$, where $\boldsymbol{W}_0 \sim \boldsymbol{N}_q(\boldsymbol{0}, \boldsymbol{I}_q)$ and $\boldsymbol{W}_1 \sim \boldsymbol{N}_p(\boldsymbol{0}, \boldsymbol{I}_p)$ are independent, it follows from (2) that the stochastic representation of $\boldsymbol{X}$ is given by

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{\Lambda}|\boldsymbol{W}_0| + \boldsymbol{\Sigma}^{1/2}\boldsymbol{W}_1, \tag{3}$$

where $|\boldsymbol{W}_0|$ is the vector formed with the absolute value of each component of $\boldsymbol{W}_0$. For more details, see Arellano-Valle et al. (2006), Arellano-Valle and Azzalini (2006) and Arellano-Valle et al. (2007). In particular, the mean vector and covariance matrix of $\boldsymbol{X}$ are given by $E[\boldsymbol{X}] = \boldsymbol{\mu} + \sqrt{2/\pi}\boldsymbol{\Lambda}\boldsymbol{1}_p$ and $\text{Cov}[\boldsymbol{X}] = \boldsymbol{\psi} - \frac{2}{\pi}\boldsymbol{\Lambda}\boldsymbol{1}_q\boldsymbol{1}_q^\top\boldsymbol{\Lambda}^\top$, respectively, where $\boldsymbol{1}_q$ denotes the vector of ones with length q.

In this work, we consider the extension of the scale mixtures of rMSN (SMRSN or SMSN) distributions to the scale mixtures of SUN (SMSUN) distributions. Specifically, we consider the family of random vectors defined by

$$\boldsymbol{Y} = \boldsymbol{\mu} + \kappa(U)^{1/2}\boldsymbol{X}, \tag{4}$$

where $\boldsymbol{X} \sim \text{SUN}_{p,q}(\boldsymbol{0}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda})$, $\kappa(\cdot)$ is a positive scale function and U is a mixing random variable which is independent of $\boldsymbol{X}$. For our proposed SUNGH distribution, we consider the SMSUN class of distributions defined by (4) when the mixing random variable U follows a GIG distribution.

2.2 The family of GIG distribution

The GIG class is a rich family of flexible distributions with positive support that has been studied by several authors. For instance, see Good (1953), Barndorff-Nielsen and Halgreen (1977), Jørgensen (1982), among others. Thus, the choice of a GIG distribution for the scale mixing variable U in (4) is a natural candidate and provides a highly flexible unified class of multivariate distributions for multivariate statistical analysis.

The GIG distribution has several but equivalent representations in terms of its parameterization. In this paper, and in order to simplify and have closed-form posterior distributions in the Bayesian framework adopted here, we consider (without loss of generality) the following two representations:

First representation $\text{GIG}^*(\upsilon, \gamma, \rho)$: A random variable U has a GIG distribution, denoted by $U \sim \text{GIG}^*(\upsilon, \gamma, \rho)$, if its pdf is given by

$$\mathcal{GIG}^*(u|\upsilon, \gamma, \rho) = \left(\frac{\gamma}{\rho}\right)^\upsilon \frac{u^{\upsilon-1}}{2K_\upsilon(\rho\gamma)} \exp\left(-\frac{1}{2}\left(\frac{\rho^2}{u} + \gamma^2 u\right)\right), \quad u > 0, \tag{5}$$

where $K_r(x)$ is the modified Bessel function of the third kind of order r evaluated at x, and the parameter spaces are given by $\gamma > 0$, $\rho > 0$ and $-\infty < \upsilon < +\infty$.

Second representation $\text{GIG}_*(\upsilon, \psi, \eta)$: A random variable U follows a GIG distribution denoted by $U \sim \text{GIG}_*(\upsilon, \psi, \eta)$, if its pdf is given by

$$\mathcal{GIG}_*(u|\upsilon, \psi, \eta) = \frac{(u/\eta)^{\upsilon-1}}{2\eta K_\upsilon(\psi)} \exp\left(-\frac{\psi}{2}\left(\frac{u}{\eta}+\frac{\eta}{u}\right)\right), u > 0, \quad (6)$$

where $K_r(x)$ is defined previously and the parameter spaces are $\psi > 0, \eta > 0$ and $-\infty < \upsilon < +\infty$. This representation will be used to simplify the posterior representation of the GIG parameters. In this case, the mth moment of the random variable $U^{1/2}$ is given by

$$E\left(U^{m/2}\right) = \frac{K_{\upsilon+m/2}(\psi)}{K_\upsilon(\psi)}\eta^{m/2}, m = 1, 2, \ldots.$$

The equivalence between both representations of the GIG distribution considered in (5) and (6) is obtained by observing the one-to-one relationship between their parameters given by $\psi = \rho\gamma$ and $\eta = \rho/\gamma$. Particular members of the GIG class lead to a variety of skewed distributions belonging to the proposed family. The inverse Gaussian is one member of this class which has been extensively studied by Chhikara and Folks (1989), Seshadri (1993) and Johnson et al. (1994, chap. 15). Two additional members of the GIG class are the hyperbola and the positive hyperbolic distributions, both of which have been studied by Barndorff-Nielsen (1978) and Barndorff-Nielsen and Blaesild (1980). The exponential, gamma and inverse gamma distributions are also special members of the GIG family. For a recent study on these distributions, see Vilca et al. (2014) and references therein.

In this paper, we define the multivariate random variable $\boldsymbol{Y}$ via (4), and by considering a multivariate SUN random variable $\boldsymbol{X}$ according to (3) and a GIG scale random variable U distributed according to the second representation in (6). As mentioned previously, we refer to this proposed family as SUNGH distributions.

3 The family of SUNGH distributions

An alternative way to define SUNGH distribution follows by replacing Eq. (3) in Eq. (4). From this, we can say that a $p \times 1$ random vector $\boldsymbol{Y}$ follows a SUNGH distribution if

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{W} + \kappa(U)^{1/2}\boldsymbol{\Sigma}^{1/2}\boldsymbol{W}_1, \quad (7)$$

where $\boldsymbol{\mu}$ is a $p \times 1$ location vector, $\boldsymbol{\Sigma}$ is a $p \times p$ scale matrix, $\boldsymbol{\Lambda}$ is a $p \times q$ shape matrix, $\boldsymbol{W} = \kappa^{1/2}(U)|\boldsymbol{W}_0|$, $\boldsymbol{W}_0 \sim N_q(\boldsymbol{0}, \boldsymbol{I}_q)$, $\boldsymbol{W}_1 \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p)$ and $U \sim \text{GIG}_*(\upsilon, \psi, \eta)$, with $\boldsymbol{W}_0$, $\boldsymbol{W}_1$ and U being independent random quantities. These assumptions also imply that $\boldsymbol{W}$ is also independent of $\boldsymbol{W}_1$. Note that if we set $\boldsymbol{W} = U, \kappa(u) = u$ and $q = 1$ we obtain the GH distribution proposed by McNeil et al. (2005) and considered in the mixture model context by Browne and McNicholas (2015). For this reason, the GH distribution is more restrictive (less flexible) compared to the SUNGH distribution. Since the conditional distribution of $\boldsymbol{Y}$ given $U = u$ is given by $\boldsymbol{Y}|U = u \sim \text{SUN}_{p,q}(\boldsymbol{\mu}, \kappa(u)\boldsymbol{\Sigma}, \kappa(u)^{1/2}\boldsymbol{\Lambda})$, the marginal pdf of $\boldsymbol{Y}$ becomes the infinite mixture of the SUN pdf in (1) given by

$$f(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}, \boldsymbol{\varpi}) = 2^q \int_0^\infty \phi_p(\boldsymbol{y}|\boldsymbol{\mu}, \kappa(u)\boldsymbol{\psi})\, \Phi_q\left(\kappa(u)^{-1/2}\boldsymbol{\Lambda}^\top\boldsymbol{\psi}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})|\boldsymbol{\Upsilon}\right) \mathcal{GIG}_*(u|\boldsymbol{\varpi})\, du, \quad (8)$$

$\boldsymbol{y} \in \boldsymbol{R}^p$, where $\boldsymbol{\varpi} = (\upsilon, \psi, \eta)^\top$, and $\boldsymbol{\psi}$ and $\boldsymbol{\Upsilon}$ defined as in (1). In what follows, we will refer to the SUNGH random vector in (7) as $\boldsymbol{Y} \sim \text{SUNGH}_{p,q}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}, \boldsymbol{\varpi})$.

Note that there are some identifiability issues concerning the GIG parameters $\boldsymbol{\varpi}$ and skewness matrix $\boldsymbol{\Lambda}$. Using (8) the density is not identifiable as for any parameter $c > 0$, the parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}, \upsilon, \psi, \eta)$ and $(\boldsymbol{\mu}, c\boldsymbol{\Sigma}, c\boldsymbol{\Lambda}, \upsilon, \psi/c, c\eta)$ yield the same density. A simple fix which results in an identifiable density is to set $\eta = 1$ and so $\boldsymbol{\varpi} = (\upsilon, \psi)^\top$. An alternative parameterization which can provide for greater flexibility is discussed in Wraith and Forbes (2015). Further, any permutation matrix can be multiplied by $\boldsymbol{W}$ from the stochastic representations (7) without any changes to the distribution of $\boldsymbol{Y}$, so sorting $\boldsymbol{\Lambda}$ by the norm of the columns or some other sorting method is also needed to ensure identifiability of the proposed model.

Varying the scale mixing function $\kappa(U)$ for a given distribution of U belonging to the $\text{GIG}_*(\boldsymbol{\varpi})$ class leads to a variety of members in the SUNGH family. Alternatively, we can fix the scale function and vary the distribution of U within the $\text{GIG}_*(\boldsymbol{\varpi})$ class. In the latter case, a convenient choice for the scale function is $\kappa(u) = u$, for which the pdf in (8) becomes

$$f(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}, \boldsymbol{\varpi}) = 2^q \mathcal{GH}_p\left(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\psi}, \boldsymbol{0}, \boldsymbol{\nu}'\right) GH_q(\boldsymbol{B}|\boldsymbol{0}, \boldsymbol{\Upsilon}, \boldsymbol{0}, \boldsymbol{\nu}''), \boldsymbol{y} \in \boldsymbol{R}^p, \quad (9)$$

where $\boldsymbol{\nu}' = \left(\upsilon, \sqrt{\psi/\eta}, \sqrt{\psi\eta}\right)^\top$), $\boldsymbol{\nu}'' = \left(\upsilon - p/2, \sqrt{\psi/\eta}, q'(\boldsymbol{y})\right)^\top$, $q'(\boldsymbol{y})^2 = (\boldsymbol{y}-\boldsymbol{\mu})^\top\boldsymbol{\psi}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}) + \psi\eta$, $\boldsymbol{\psi} = \boldsymbol{\Sigma} + \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top$, $\boldsymbol{\Upsilon} = \boldsymbol{I}_q - \boldsymbol{\Lambda}^\top\boldsymbol{\psi}^{-1}\boldsymbol{\Lambda}$ and $\boldsymbol{B} = \boldsymbol{\Lambda}^\top\boldsymbol{\psi}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})$, $\mathcal{GH}_p$ and GH_q denote the p-variate pdf and q-variate cdf of the generalized Hyperbolic distribution, respectively (Wraith and Forbes 2015).

The flexibility of the SUNGH family proposed in (8) can also be observed by varying the value of the dimension q. In fact, for $q = 1$ (the restricted case) we obtain as a special

case of (8) the SN generalized hyperbolic (SNGH) distributions considered in Vilca et al. (2014), and thus some known SMSN (or SMRSN) distributions, as well the corresponding symmetrical variant for $\boldsymbol{\Lambda} = \mathbf{0}$.

A special case of the GIG distribution is the gamma distribution, so the proposed family of distributions covers the canonical fundamental unrestricted skew-normal (CFUSN) distribution of Arellano-Valle and Genton (2005), and the canonical fundamental unrestricted skew-t (CFUST) distribution of Lee and McLachlan (2016). Subsequently, a mixture model approach covering these distributions contains the finite mixtures of CFUSN and CFUST. By considering (9) in the symmetric case, the SUNGH and GH studied by Wraith and Forbes (2015) and Browne and McNicholas (2015) are similar, but in the asymmetric case these families are different. In particular, a greater degree of flexibility is available for the SUNGH family by allowing the skewness parameter to be multivariate $(p \times q)$ rather than $p \times 1$. The SUNGH family also has several desirable properties outlined in Propositions 2 to 6 below which will allow the family to be used in a variety of statistical models (e.g., mixed models and regression).

Known members of the SMSN family contained in the SNGH family are the SN, ST, SSL and skew-Laplace (SLP), and their respective symmetric versions. In the unrestricted case $(q > 1)$, the proposed family contains several subfamilies of distributions (symmetric and asymmetric) considered in the literature. For instance, if in (9) we let $q = p$, $\boldsymbol{\Lambda} = \text{diag}\left(\lambda_1, \ldots, \lambda_p\right)$ and $\kappa(u) = 1$, then the multivariate skew-normal distribution of Sahu et al. (2003) is obtained. Finally, if $\boldsymbol{\Lambda} = \mathbf{0}$ (symmetric case) and $\kappa(u) = u$, then (9) becomes the symmetric generalized hyperbolic (GH) distribution introduced by Barndorff-Nielsen and Halgreen (1977).

In the following propositions, we present some necessary and useful properties of the SUNGH family for the next sections. The proofs of these results are presented in "Appendix."

Proposition 1 *Let* $\boldsymbol{Y} \sim \text{SUNGH}_{p,q}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}, \boldsymbol{\varpi})$. *Then, the following results hold:*

(a) *if* $k_1 = E\left[\kappa(U)^{1/2}\right] < \infty$, *then* $E[\boldsymbol{Y}] = \boldsymbol{\mu} + \sqrt{\frac{2}{\pi}} k_1 \boldsymbol{\Lambda} \mathbf{1}_q$,

(b) *if* $k_2 = E[\kappa(U)] < \infty$, *then* $Var[\boldsymbol{Y}] = k_2 \boldsymbol{\psi} - \frac{2}{\pi} \boldsymbol{\Lambda} \left[k_2 \boldsymbol{I}_q - \left(k_2 - k_1^2\right) \mathbf{1}_q \mathbf{1}_q^\top\right] \boldsymbol{\Lambda}^\top$.

Proposition 2 *Let* $\boldsymbol{Y} \sim \text{SUNGH}_{p,q}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}, \boldsymbol{\varpi})$. *Then,* $\boldsymbol{Y} \sim \text{SUNGH}_{p,q+m}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}^*, \boldsymbol{\varpi}\right)$ *for each* $m = 1, 2, \ldots$, *where* $\boldsymbol{\Lambda}^* = \left(\boldsymbol{\Lambda}_{p\times q} \ \mathbf{0}_{p\times m}\right)$ *or* $\boldsymbol{\Lambda}^* = \left(\mathbf{0}_{p\times m} \ \boldsymbol{\Lambda}_{p\times q}\right)$.

Proposition 3 *Let* $\boldsymbol{Y} \sim \text{SUNGH}_{p,q}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}, \boldsymbol{\varpi})$. *Then, for each* $\boldsymbol{b} \in \boldsymbol{R}^n$ *and full row rank matrix* $\boldsymbol{B} \in \boldsymbol{R}^{n\times p}$ *we have*

$$\boldsymbol{X} = \boldsymbol{b} + \boldsymbol{B}\boldsymbol{Y} \sim \text{SUNGH}_{n,q}\left(\boldsymbol{b} + \boldsymbol{B}\boldsymbol{\mu}, \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}^\top, \boldsymbol{B}\boldsymbol{\Lambda}, \boldsymbol{\varpi}\right).$$

Proposition 4 *Let* $\boldsymbol{Y} \sim \text{SUNGH}_{p,q}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}, \boldsymbol{\varpi})$. *Partition* $\boldsymbol{Y} = \left(\boldsymbol{Y}_1^\top, \boldsymbol{Y}_2^\top\right)^\top$, *where the first and second sub-vectors are of dimensions* $p_1 \times 1$ *and* $p_2 \times 1$, *respectively, with* $p_1 + p_2 = p$. *The corresponding partition of the parameters* $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda})$ *is*

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}, \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_1 \\ \boldsymbol{\Lambda}_2 \end{pmatrix},$$

where $\boldsymbol{\mu}_i$, $\boldsymbol{\Sigma}_{ii}$ *and* $\boldsymbol{\Lambda}_i$ *have dimensions* $p_i \times 1$, $p_i \times p_i$ *and* $p_i \times q$, *respectively, for* $i = 1, 2$. *Then, the marginal distribution of* $\boldsymbol{Y}_i$ *is* $\text{SUNGH}_{p_i,q}\left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_{ii}, \boldsymbol{\Lambda}_i, \boldsymbol{\varpi}\right), i = 1, 2$.

Proposition 5 *If under the same conditions of Proposition 4, we have* $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21} = \mathbf{0}$ *then a necessary and sufficient condition to have null correlation between* $\boldsymbol{Y}_1$ *and* $\boldsymbol{Y}_2$ *is that* $\boldsymbol{\Lambda}_1 = \mathbf{0}$ *or* $\boldsymbol{\Lambda}_2 = \mathbf{0}$.

Proposition 6 *Consider the same conditions of Proposition 4 with the partition of shape matrix* $\boldsymbol{\Lambda} = \left(\boldsymbol{\Lambda}_{ij}\right)_{i,j=1,2}$, *where* $\boldsymbol{\Lambda}_{ij}$ *has dimension* $p_i \times q_j$, *with* $q_1 + q_2 = q$. *If* $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}^\top = \mathbf{0}$ *and* $\boldsymbol{\Lambda}_{12} = \mathbf{0}$ *or* $\boldsymbol{\Lambda}_{21} = \mathbf{0}$, *then* $\boldsymbol{Y}_i \sim \text{SUNGH}_{p_i,q_i}\left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_{ii}, \boldsymbol{\Lambda}_{ii}, \boldsymbol{\varpi}\right), i = 1, 2$, *and* $Cov(\boldsymbol{Y}_1, \boldsymbol{Y}_2) = -\frac{2}{\pi} k_1^2 \boldsymbol{\Lambda}_{12} \mathbf{1}_{q_1} \mathbf{1}_{q_2}^\top \boldsymbol{\Lambda}_{21}^\top$.

4 Finite mixtures of SUNGH family

4.1 FM-SUNGH model

In this section, we consider finite mixtures of the proposed SUNGH family of distributions (hereafter FM-SUNGH). To establish notation, we consider the usual mixture model defined as

$$f(\boldsymbol{y}; \boldsymbol{\Theta}, \boldsymbol{p}) = \sum_{k=1}^{K} p_k f(\boldsymbol{y}; \boldsymbol{\Theta}_k), \tag{10}$$

where $\boldsymbol{\Theta} = \left(\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_K\right)$, with $\boldsymbol{\Theta}_k = \left(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\Lambda}_k, \boldsymbol{\upsilon}_k, \boldsymbol{\psi}_k, \boldsymbol{\eta}_k\right), k = 1, \ldots, K$, $\boldsymbol{p} = (p_1, \ldots, p_K)^\top$ (for which $p_k > 0, k = 1, \ldots, K$ and $\sum_{k=1}^{K} p_k = 1$), $\boldsymbol{\upsilon}_k = \left(\upsilon_{k1}, \ldots, \upsilon_{kp}\right)^\top$, $\boldsymbol{\psi}_k = \left(\psi_{k1}, \ldots, \psi_{kp}\right)^\top$, $\boldsymbol{\eta}_k = \left(\eta_{k1}, \ldots, \eta_{kp}\right)^\top$ and $f(\boldsymbol{y}; \boldsymbol{\Theta}_k)$ given by (8). This model hereafter will be called FM-SUNGH. The identifiability of mixtures of distributions has been studied by Teicher (1963) and Holzmann et al. (2006) to ensure that the FM-SUNGH is identifiable.

The SUNGH family is a rich class of distributions, and various particular forms from this family have been considered over the last few years in the case of mixture models. In Table 1, we outline details of some of the distributions and the corresponding parameters within the SUNGH family.

Using the mixture model representation in (10), for each i.i.d. sample in the form of $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$, we can utilize

an (latent) indicator (allocation) variables $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$, to assign observations to belong to different components of the mixture ($k = 1, \ldots, K$). The standard assumption for the allocation random variables $Z_1, \ldots, Z_K$ is that they follow a multinomial joint distribution: $\boldsymbol{Z}_i = (Z_1, \ldots, Z_K) \sim Multinomial\,(K, p_1, \ldots, p_K)$, so that $P\,(Z_i = k) = p_k$; $i = 1, \ldots, n, k = 1, \ldots, K$. In terms of Z_i, we can conclude that

$$\boldsymbol{Y}_i | \, Z_i = k \overset{\text{ind.}}{\sim} \text{SUNGH}\,(\boldsymbol{\Theta}_k)\,, P\,(Z_i = k) = p_k.$$

Let $\boldsymbol{C} = \{\boldsymbol{Y}, \boldsymbol{U}, \boldsymbol{W}, \boldsymbol{Z}\}$ denote the complete data, where $\boldsymbol{Y} = \left(\boldsymbol{Y}_1^\top, \ldots, \boldsymbol{Y}_n^\top\right)^\top$ is the observed variable and $\boldsymbol{U} = (U_{11}, \ldots, U_{1K}, \ldots, U_{n1}, \ldots, U_{nK})^\top$, $\boldsymbol{W} = \left(\boldsymbol{W}_{11}^\top, \ldots, \boldsymbol{W}_{1K}^\top, \ldots, \boldsymbol{W}_{n1}^\top, \ldots, \boldsymbol{W}_{nK}^\top\right)^\top$ and $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^\top$ are the latent or unobserved variables. If we consider the SUNGH stochastic representation (7) in terms of a finite mixture model for $i = 1, \ldots, n$ and $k = 1, \ldots, K$, a hierarchical representation is

$$\boldsymbol{Y}_i | \, \boldsymbol{W}_{ik} = \boldsymbol{w}_{ik}, U_{ik} = u_{ik}, Z_i = k \overset{\text{ind.}}{\sim} N_p \left(\boldsymbol{\mu}_k + \boldsymbol{\Lambda}_k \boldsymbol{w}_{ik}, \kappa\,(u_{ik})\,\boldsymbol{\Sigma}_k\right), \tag{11}$$

$$\boldsymbol{W}_{ik} | \, U_{ik} = u_{ik}, Z_i = k \overset{\text{ind.}}{\sim} HN_q \left(\boldsymbol{0}, \kappa\,(u_{ik})\,\boldsymbol{I}_q\right), \tag{12}$$

$$U_{ik} | \, Z_i = k \overset{\text{ind.}}{\sim} \text{GIG}_*\,(\upsilon_k, \psi_k, \eta_k)\,, \tag{13}$$

$$\boldsymbol{Z} \sim \text{Multinomial}(K, p_1, \ldots, p_K), \tag{14}$$

where HN_q denotes the q-variate right half-normal distribution.

The model's complete data likelihood function is then given by

$$L\,(\boldsymbol{\Theta}\,|\boldsymbol{C}\,) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left(\phi_p\left(\boldsymbol{y}_i | \boldsymbol{\mu}_k + \boldsymbol{\Lambda}_k \boldsymbol{w}_{ik}, \kappa\,(u_{ik})\,\boldsymbol{\Sigma}_k\right) H\phi_q\left(\boldsymbol{w}_{ik} | \boldsymbol{0}, \kappa\,(u_{ik})\,\boldsymbol{I}_q\right) \mathcal{GIG}_*\,(u_{ik} | \upsilon_k, \psi_k, \eta_k)\, P\,(Z_i = k)\right), \tag{15}$$

where $H\phi_q\,(\boldsymbol{w}|\boldsymbol{0}, \cdot) = \phi_q\,(\boldsymbol{w}|\boldsymbol{0}, \cdot)\,I(\boldsymbol{w} > \boldsymbol{0})$ is the q-variate right half-normal pdf.

4.2 Bayesian analysis

4.3 Priors

In this section, we choose priors for the parameters $\boldsymbol{\Theta}$ which will be used in Applications section. By assuming independency between the different types of parameters in $\boldsymbol{\Theta}$ and that the skewness matrix of each mixture component be in the form of $\boldsymbol{\Lambda}_k = \left(\boldsymbol{\lambda}_{k1} | \ldots | \boldsymbol{\lambda}_{kq}\right)$, prior distributions for some of the FM-SUNGH model parameters are given by

$$\boldsymbol{p} \sim Dir\,(\delta_1, \ldots, \delta_K)\,, \boldsymbol{\mu}_k \sim N_p\,(\boldsymbol{m}_k, \boldsymbol{M}_k)\,, \quad \boldsymbol{\Sigma}_k \sim IW_{t_k}\,(\boldsymbol{T}_k)\,, \boldsymbol{\lambda}_{kt} \sim N_p\,(\boldsymbol{\ell}_{kt}, \boldsymbol{L}_{kt})\,, \quad t = 1, \ldots, p$$

for $k = 1, \ldots, K$, and where Dir and IW denote the Dirichlet and inverse Wishart distributions, respectively. An alternative representation of the skewness matrices priors and posteriors in the Gibbs updates is provided in "Appendix." Prior distributions of the scaled factor variables for $k = 1, \ldots, K$ are:

$$\upsilon_k \sim N\left(\mu_k, \sigma_k^2\right), \eta_k \sim \text{GIG}^*\,(\alpha_k, \chi_k, \varepsilon_k)\,, \psi_k \sim \exp\,(\varsigma_k)\,.$$

4.3.1 Posteriors

By considering the likelihood function (15) and the priors specified previously, the joint posterior of $\boldsymbol{\Theta}$ is given by

$$\pi\,(\boldsymbol{\Theta}, \boldsymbol{u}, \boldsymbol{w}, \boldsymbol{z}|\,\boldsymbol{y}) \propto L\,(\boldsymbol{\Theta}\,|\boldsymbol{y}, \boldsymbol{u}, \boldsymbol{w}, \boldsymbol{z})\,\pi\,(\boldsymbol{\Theta})\,.$$

The above joint posterior is intractable, but we can use an MCMC method such as Gibbs sampling and Metropolis–Hastings to draw samples using the conditional posterior distributions. To establish notation, let $B_k = \{i, Z_i = k\}$ be the set of observation indices for those $\boldsymbol{y}_i$ classified into the kth cluster and n_k is equal to the number of observations allocated to the kth component (cluster). Apart from the parameters for the scaled factor variables, all conditional posterior distributions have closed form and are as follows: (Note that $\boldsymbol{\Theta}_{(-\boldsymbol{\theta})}$ denotes the set of parameters without its element $\boldsymbol{\theta}$.)

$\boldsymbol{p}|\,\boldsymbol{\Theta}_{(-p)}, \boldsymbol{y}, \boldsymbol{u}, \boldsymbol{w}, z_i = k \sim Dir\left(\delta_{pos.1}, \ldots, \delta_{pos.K}\right)$, where

$$\delta_{pos.k} = \delta_k + n_k; k = 1, \ldots, K. \tag{16}$$

$\boldsymbol{\mu}_k\big|\,\boldsymbol{\Theta}_{(-\mu_k)}, \boldsymbol{y}, \boldsymbol{u}, \boldsymbol{w}, z_i = k \sim N_p\,(\boldsymbol{\mu}, \boldsymbol{\Sigma})\,, k = 1, \ldots, K$, where

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}\left(\boldsymbol{M}_k^{-1}\boldsymbol{m}_k + \sum\nolimits_{B_k} \kappa\,(u_{ik})^{-1}\,\boldsymbol{\Sigma}_k^{-1}\left(\boldsymbol{y}_i - \boldsymbol{\Lambda}_k \boldsymbol{w}_{ik}\right)\right),$$
$$\boldsymbol{\Sigma} = \left[\boldsymbol{M}_k^{-1} + \sum\nolimits_{B_k} \kappa\,(u_{ik})^{-1}\,\boldsymbol{\Sigma}_k^{-1}\right]^{-1}, \tag{17}$$

$\boldsymbol{\Sigma}_k|\,\boldsymbol{\Theta}_{(-\boldsymbol{\Sigma}_k)}, \boldsymbol{y}, \boldsymbol{u}, \boldsymbol{w}, z_i = k \sim IW_{t_k+n}\,(\boldsymbol{T})\,, k = 1, \ldots, K$, where

$$\boldsymbol{T} = \sum\nolimits_{B_k} \kappa\,(u_{ik})^{-1}\left(\boldsymbol{y}_i - \boldsymbol{\mu}_k - \boldsymbol{\Lambda}_k \boldsymbol{w}_{ik}\right)\left(\boldsymbol{y}_i - \boldsymbol{\mu}_k - \boldsymbol{\Lambda}_k \boldsymbol{w}_{ik}\right)^\top + \boldsymbol{T}_k, \tag{18}$$

$\boldsymbol{\lambda}_{kt}|\,\boldsymbol{\Theta}_{(-\lambda_{kt})}, \boldsymbol{y}, \boldsymbol{u}, \boldsymbol{w}, z_i = k \sim N_p\,(\boldsymbol{\mu}, \boldsymbol{\Sigma})\,; k = 1, \ldots, K;$ $t = 1, \ldots, p$, where

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}\left(\boldsymbol{L}_{kt}^{-1}\boldsymbol{\ell}_{kt} + \sum\nolimits_{B_k} \kappa\,(u_{ik})^{-1}\,w_{ik(t)}\boldsymbol{\Sigma}_k^{-1}\right.$$

Table 1 Summary of non-normal finite mixture models belonging to the FM-SUNGH model

Finite mixture models	Component density	FM-SUNGH	References
FM of restricted skew-normal (FM-rMSN)	$f(\boldsymbol{x}\|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\delta}) = 2\phi_p(\boldsymbol{x}\|\boldsymbol{\mu},\boldsymbol{\Sigma})\,\Phi_1\left(\boldsymbol{\delta}^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\,\|\boldsymbol{\Upsilon}\right)$, $\boldsymbol{\Upsilon} = 1-\boldsymbol{\delta}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$	$\kappa(u)=1, \quad q=1$	Pyne et al. (2009)
FM of unrestricted skew-normal (FM-CFUSN)	$f(\boldsymbol{x}\|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\Lambda}) = 2^q\phi_p(\boldsymbol{x}\|\boldsymbol{\mu},\boldsymbol{\psi})\,\Phi_q\left(\boldsymbol{\Lambda}^\top\boldsymbol{\psi}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\,\|\boldsymbol{\Upsilon}\right)$, $\boldsymbol{\psi} = \boldsymbol{\Sigma}+\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top, \quad \boldsymbol{\Upsilon} = \boldsymbol{I}_q-\boldsymbol{\Lambda}^\top\boldsymbol{\psi}^{-1}\boldsymbol{\Lambda} = \left(\boldsymbol{I}_q+\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-1}$	$\kappa(u)=1$	Lin (2009)
FM of restricted skew-t (FM-rMST)	$f(\boldsymbol{x}\|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\delta},n) = 2t_p(\boldsymbol{x}\|\boldsymbol{\mu},\boldsymbol{\Sigma},n)\,T_1\left(\boldsymbol{\delta}^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\sqrt{\frac{n+p}{n+d}}\|0,\Upsilon,n+p\right)$, $d=(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}), \Upsilon = 1-\boldsymbol{\delta}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$	$\eta\downarrow 0, \quad \upsilon=\psi=n/2, \quad q=1$	Pyne et al. (2009) and Vrbik and McNicholas (2012)
FM of unrestricted skew-t (FM-CFUST)	$f(\boldsymbol{x}\|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\delta},n) = 2t_p(\boldsymbol{x}\|\boldsymbol{\mu},\boldsymbol{\Sigma},n)\,T_q\left(\boldsymbol{\Delta}^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\sqrt{\frac{n+p}{n+d}}\|\boldsymbol{0},\boldsymbol{\Lambda},n+p\right)$, $\boldsymbol{\Delta}=\mathrm{diag}(\boldsymbol{\delta}), d=(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}), \boldsymbol{\Lambda} = \boldsymbol{I}_p-\boldsymbol{\delta}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$	$\eta\downarrow 0, \quad \upsilon=\psi=n/2$	Lin (2010), Lee and McLachlan (2014) and Lee and McLachlan (2016)
FM of multivariate student-t (FM-T)	$f(\boldsymbol{x}\|\boldsymbol{\mu},\boldsymbol{\Sigma},n) = \frac{\Gamma((n+p)/2)}{\Gamma(n/2)n^{p/2}\pi^{p/2}\|\boldsymbol{\Sigma}\|^{1/2}}\left(1+\frac{1}{n}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)^{-\frac{n+p}{2}}$	$\eta\downarrow 0, \quad \upsilon=\psi=n/2, \quad \boldsymbol{\Lambda}=\boldsymbol{0}, \quad q=1$	McLachlan and Peel (2000)
FM of scale mixtures of skew-normal (FM-SMSN)	$f(\boldsymbol{x}\|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda}) = 2\int_0^\infty \phi_p\left(\boldsymbol{y}\|\boldsymbol{\mu},u^{-1}\boldsymbol{\Sigma}\right)\Phi_1\left(u^{1/2}\boldsymbol{\lambda}^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)dH(u)$	See Section 3 in Vilca et al. (2014)	Basso et al. (2010) and Maleki and Arellano-Valle (2017)
FM of generalized hyperbolic (FM-GH)	$f(\boldsymbol{x}\|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\alpha},\lambda,\omega) = \frac{e^{-\alpha^\top\boldsymbol{\Sigma}^{-1}(x-\mu)}}{(2\pi)^{p/2}\|\boldsymbol{\Sigma}\|^{1/2}}\left(\frac{\omega+d}{\gamma}\right)^{\frac{\lambda}{2}-\frac{p}{4}}\frac{K_{\lambda-\frac{p}{2}}\left(\sqrt{\gamma(\omega+d)}\right)}{K_\lambda(\omega)}$, $\gamma=\omega+\alpha^\top\boldsymbol{\Sigma}^{-1}\alpha, d=(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$	Let $\boldsymbol{W}=U$ in Eq. (7), $\kappa(u)=u$ and q = 1	McNeil et al. (2005) and Browne and McNicholas (2015)

$$\left(\boldsymbol{y}_i - \boldsymbol{\mu}_k - \boldsymbol{\Lambda}_{k(-t)}\boldsymbol{w}_{ik(-t)}\right)\Big),$$
$$\boldsymbol{\Sigma} = \left(\boldsymbol{L}_{kt}^{-1} + \sum\nolimits_{B_k} \kappa\left(u_{ik}\right)^{-1} w_{ik(t)}^2 \boldsymbol{\Sigma}_k^{-1}\right)^{-1}, \tag{19}$$

where $\boldsymbol{\Lambda}_{k(-t)}$ denotes the $p \times (q-1)$ skewness matrix $\boldsymbol{\Lambda}_k$ with the tth column eliminated, $\boldsymbol{w}_{ik(-t)}$ denotes the $(q-1)\times 1$ vector $\boldsymbol{w}_{ik}$ vector with the tth element eliminated, and $w_{ik(t)}$ denotes the tth element of the vector $\boldsymbol{w}_{ik}$.

The full conditional posterior distribution for the latent variables Z_i, U_{ik} and $\boldsymbol{W}_{ik}$, for $i = 1, \ldots, n$; $k = 1, \ldots, K$, are given by:

$Z_i|\,\boldsymbol{\Theta}, \boldsymbol{y}, \boldsymbol{u}, \boldsymbol{w} \sim \text{Multinomial}\left(K, p_{p.1}, \ldots, p_{p.K}\right)$, where

$$p_{p.k} = \frac{p_k f\left(\boldsymbol{y}_i; \boldsymbol{\Theta}_k\right)}{\sum_{j=1}^{K} p_j f\left(\boldsymbol{y}_i; \boldsymbol{\Theta}_j\right)}, k = 1, \ldots, K, \tag{20}$$

$U_{ik}|\,\boldsymbol{\Theta}, \boldsymbol{y}, \boldsymbol{w}, z_i = k \sim \text{GIG}^*\left(\text{a}_u, \text{b}_u, \text{c}_u\right)$, where $\kappa(u) = u$ and

$$\begin{aligned}
\text{a}_u &= \upsilon_{ik} - \frac{p+q}{2},\\
\text{b}_u &= \left(\psi_{ik}/\eta_{ik}\right)^{1/2},\\
\text{c}_u &= \Big(\Big[\boldsymbol{w}_{ik}^{\top}\boldsymbol{w}_{ik} + \left(\boldsymbol{y}_i - \boldsymbol{\mu}_k - \boldsymbol{\Lambda}_k\boldsymbol{w}_{ik}\right)^{\top}\\
&\qquad \boldsymbol{\Sigma}_k^{-1}\left(\boldsymbol{y}_i - \boldsymbol{\mu}_k - \boldsymbol{\Lambda}_k\boldsymbol{w}_{ik}\right)\Big] + \psi_{ik}\eta_{ik}\Big)^{1/2}.
\end{aligned} \tag{21}$$

$\boldsymbol{W}_{ik}|\,\boldsymbol{\Theta}, \boldsymbol{y}, \boldsymbol{u}, z_i = k \sim HN_q\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$, where

$$\begin{aligned}
\boldsymbol{\mu} &= \kappa\left(u_{ik}\right)^{-1}\boldsymbol{\Sigma}\boldsymbol{\Lambda}_k^{\top}\boldsymbol{\Sigma}_k^{-1}\left(\boldsymbol{y}_i - \boldsymbol{\mu}_k\right),\\
\boldsymbol{\Sigma} &= \kappa\left(u_{ik}\right)\left(\boldsymbol{I}_q + \boldsymbol{\Lambda}_k^{\top}\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Lambda}_k\right)^{-1}.
\end{aligned} \tag{22}$$

Finally, the full conditional posterior for the scaled factor variables $\upsilon_k, \psi_k, \eta_k, k = 1, \ldots, K$, is as follows: $\eta_k|\,\boldsymbol{\Theta}_{(-\eta_k)}, \boldsymbol{u}, \boldsymbol{w}, z_i = k \sim \text{GIG}^*\left(\text{a}_\eta, \text{b}_\eta, \text{c}_\eta\right)$, where

$$\begin{aligned}
\text{a}_\eta &= \alpha_k - \upsilon_k n_k,\\
\text{b}_\eta &= \left(\chi_k^2 + \psi_k \sum\nolimits_{B_k} 1/u_{ik}\right)^{1/2},\\
\text{c}_\eta &= \left(\varepsilon_k^2 + \psi_k \sum\nolimits_{B_k} u_{ik}\right)^{1/2}.
\end{aligned} \tag{23}$$

The full conditional posterior density of $\upsilon_k, k = 1, \ldots, K$ is proportional to:

$$\pi\left(\upsilon_k|\,\boldsymbol{\Theta}_{(-\upsilon_k)}, \boldsymbol{u}, \boldsymbol{w}, Z_i = k\right) \sim \pi_1\left(\upsilon_k\right) N\left(\mu_k, \sigma_k^2\right), \tag{24}$$

where $\pi_1\left(\upsilon_k\right) = \left(K_{\upsilon_k}\left(\psi_k\right)\right)^{-n_k} \prod\limits_{B_k}\left(u_{ik}/\eta_k\right)^{\upsilon_k}$.

The full conditional posterior density of $\psi_k, k = 1, \ldots, K$ is also proportional to:

$$\begin{aligned}
&\pi\left(\psi_k|\,\boldsymbol{\Theta}_{(-\psi_k)}, \boldsymbol{u}, \boldsymbol{w}, z_i = k\right) \sim \pi_2\left(\psi_k\right)\\
&\quad E\left(\varsigma_k + \sum\nolimits_{B_k}\left(u_{ik}/\eta_k + \eta_k/u_{ik}\right)/2\right),
\end{aligned} \tag{25}$$

where $\pi_2\left(\psi_k\right) = \left(K_{\upsilon_k}\left(\psi_k\right)\right)^{-n_k}$ and $E\left(\varphi\right)$ denotes the density of the exponential distribution with rate parameter φ.

Note that (24) and (25) do not have closed forms, but a Metropolis–Hastings or rejection sampling step can be embedded in the MCMC scheme to obtain draws from them.

5 Applications

In this section, we present a simulation study and applications on two real datasets to evaluate the performance of the proposed SUNGH model for clustering problems. For illustrative purposes, we choose K to be equal to two for all models presented.

5.1 Simulated data

To illustrate some of the differences between the SUNGH family of models, we consider the case of two clusters each sampled from a four-dimensional SUNGH distribution with known parameters which are slightly separated from each other. For the first and second cluster

$$\boldsymbol{\mu}_1 = (4,4,4,4)^{\top}, \quad \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & \cdots & 0.5\\ \vdots & \ddots & \vdots\\ 0.5 & \cdots & 1 \end{pmatrix} \text{ and}$$

$$\boldsymbol{\Lambda}_1 = \begin{pmatrix} -4 & -4\\ 1 & 3\\ -4 & -4\\ 1 & 3 \end{pmatrix},$$

$$\boldsymbol{\mu}_2 = (-1,-1,-1,-1)^{\top} \text{ and } \boldsymbol{\Lambda}_2 = \begin{pmatrix} 4 & -5\\ 1 & 2\\ 4 & -5\\ 1 & 2 \end{pmatrix},$$

respectively. Both clusters shared the same parameters for the $\text{GIG}_*\left(\upsilon, \psi, \eta\right)$ distribution, where $\upsilon = -0.5$, $\psi = 1$ and $\eta = 1$. The sample size for each cluster is 300 and 450, respectively. A plot of the simulated data is shown in Fig. 1 with the observations belonging to each cluster labeled by different colors.

For estimation of the different models, largely non-informative prior distributions were used for each of the component parameters: $\boldsymbol{\mu} = \left(\mu_1, \ldots, \mu_4\right)^{\top} \sim N_4\left(\boldsymbol{0}, \boldsymbol{\Sigma}\right)$, where $\boldsymbol{\Sigma} = 10^3\boldsymbol{I}_4$, $\boldsymbol{\Sigma} \sim IW_\tau\left(\boldsymbol{T}\right)$, where $\tau = 4$ and $\boldsymbol{T} = \boldsymbol{I}_4$

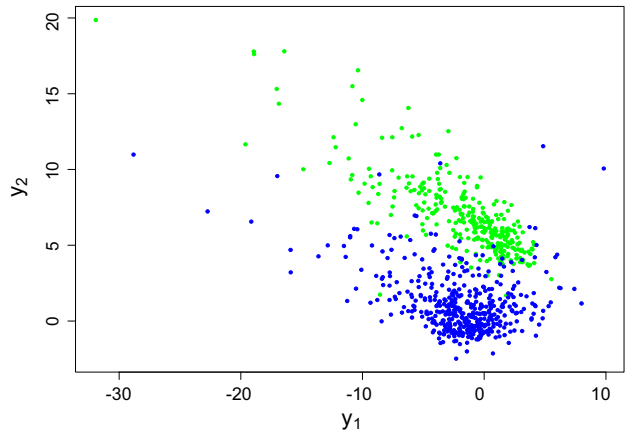

Fig. 1 Plot of results for simulated data. Colors indicate the groups to which observations belong in the first two dimensions. (Color figure online)

with skewness matrix $\boldsymbol{\Lambda}_{4\times 2}$ with priors of its columns as $\boldsymbol{\lambda}_t \sim N_4(\boldsymbol{\ell}_t, \boldsymbol{L}_t)$ for which $\boldsymbol{\ell}_t = \mathbf{0}$ and $\boldsymbol{L}_t = 10^3\boldsymbol{I}_4$ and for $t = 1, 2$ (and in the matrix variate prior of the skewness matrix we can consider that $\boldsymbol{\Lambda}_{4\times 2} \sim MN_{4,2}\left(\mathbf{0}, 10^3\boldsymbol{I}_4, 10^3\boldsymbol{I}_4\right)$), $\upsilon \sim N\left(0, 10^3\right)$, $\eta \sim \text{GIG}^*(0.001, 2000, 0)$, $\psi \sim \exp(0.1)$ and $\boldsymbol{p} \sim Dir(1, \ldots, 1)$. Also, we chose $\kappa(u) = u$ for the scale mixing function. All computations are implemented in the R software version 3.3.1 (R Core Team 2017) with a core i7 760 processor 2.8 GHz. The R and Nimble code for the models are available from the authors upon request. Gibbs sampling runs of 60,000 iterations with burn-in of 30,000 were used, and convergence criteria were established using the Gelman–Rubin statistic (Gelman and Rubin 1992) and by visual inspection. To address the issue of label switching over the MCMC iterations (Mengersen et al. (2011)), we used the *maximum a posteriori* estimate (MAP) to select one of the $k!$ modal regions and a distance based measure on the space of parameters to relabel parameters in proximity to this region (Celeux et al. 2000).

Model performance was assessed by comparing the classification accuracy and model selection criteria for different distributions within the family of SUNGH models (see Table 2). For classification accuracy, we report the adjusted rand index (ARI) (Hubert and Arabie 1985) which ranges from 0 (no match) to 1 (perfect match). We also report the EAIC and EBIC which are variations of the classical AIC and BIC criteria for use in a Bayesian setting (Carlin and Louis 2011) (lower values indicate a better fit). In a mixture setting, it is also possible to compare the DIC values using one of the measures suggested by Celeux et al. (2006).

As to be expected, from Table 2 we can see quite clearly that the classification performance of the true model (SUNGH ($q = 2$)) is very good with an ARI of 0.87 and model choice criteria all appear to favor this model. A higher log-likelihood was found for the model SUNGH ($q = 3$) with a similar ARI score to the SUNGH ($q = 2$) model, but on other criteria this model was not favored due to the extra parameters involved. In applied settings, and where the true labels are unknown, a similar trade-off will be made between choosing more complex models with extra flexibility in the skewness matrix (higher q values) and relative improvement in model choice or goodness-of-fit measures. The performance of the SN and SNGH models is also to be expected given the relative lack of flexibility for the skewness parameter to accommodate the degree of skewness in all dimensions in this application.

5.2 Real applications

5.2.1 AIS example

In this example, we consider a dataset from the Australian Institute Sports (AIS) containing measures of physical activity for 202 athletes (102 male and 100 female) based on sex, red cell count, white cell count, hematocrit, hemoglobin, plasma ferritin concentration, body mass index, sum of skin folds, body fat percentage, lean body mass and finally height and weight of the athletes (Cook and Weisberg 1994). The data are available in the R package "sn" (see Azzalini 2015).

To assess the performance of the proposed SUNGH model, we use BMI and body fat percentage (Bfat) to classify male and female athletes. Figure 2a shows the observations for male (in black) and female (in red) athletes according to these two measures, suggesting a reasonably skewed distribution for both males and females with a particularly strong skewed and heavy-tailed distribution for male athletes. Figure 2b–d also shows the fitted contours and assigned labels for each observation for several of the models examined (SNGH, SN and SUNGH).

Table 3 presents the model choice criteria for the different models examined. The results suggest that the SUNGH has

Table 2 Results for simulated data example

Model	Log-likelihood (max)	EAIC	EBIC	DIC_2	ARI
SN	−5936.9	12009.3	12198.8	11980.8	0.47
SNGH	−5692.1	11504.2	11693.6	11460.1	0.77
SUNGH ($q = 2$)	−5642.8	**11458.9**	**11676.1**	**11444.3**	**0.87**
SUNGH ($q = 3$)	**−5628.5**	11487.5	11723.2	11514.1	0.86

The best values are indicated in bold

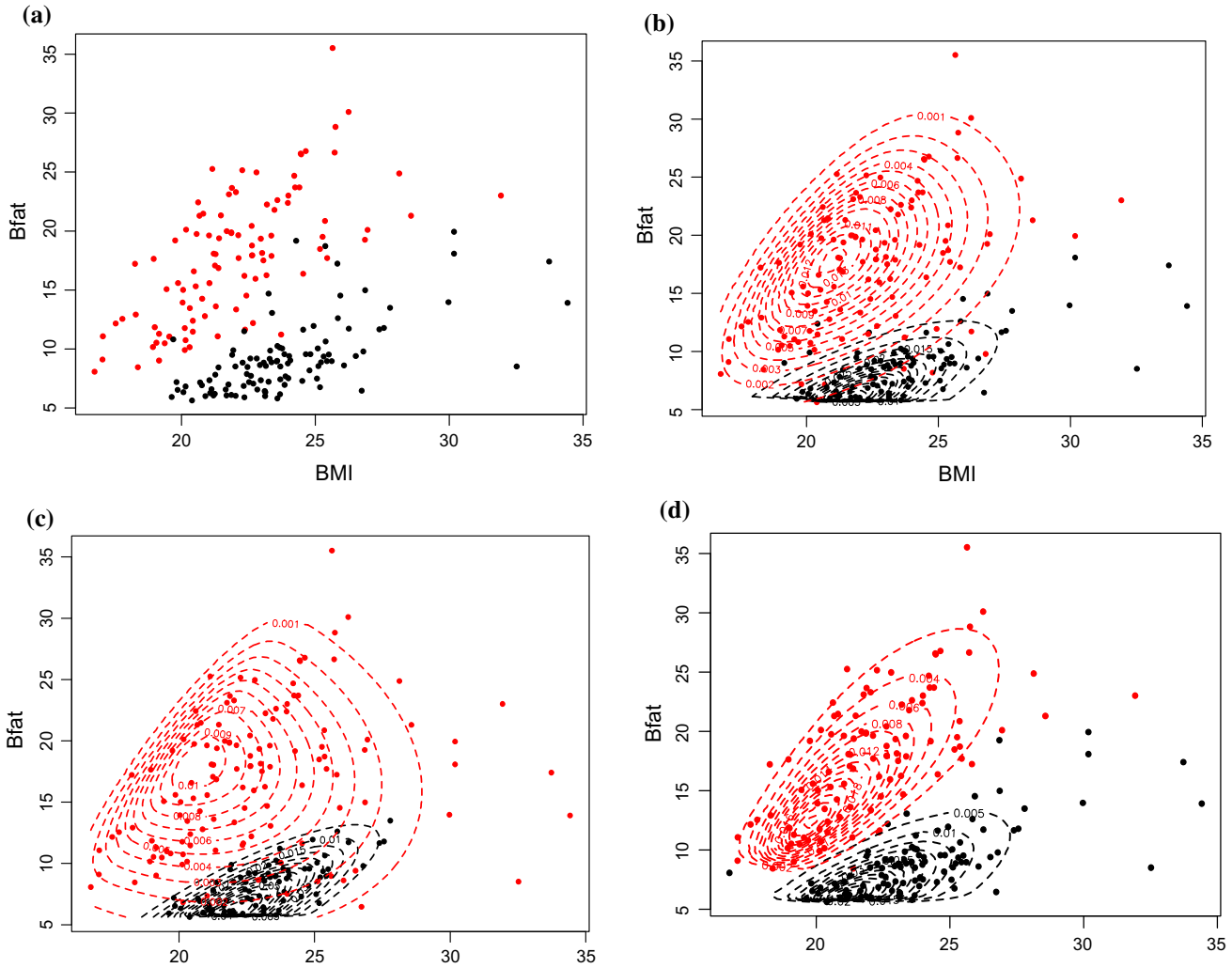

Fig. 2 Plots of results for AIS example: **a** true data, **b** SNGH, **c** SN, **d** SUNGH

Table 3 Model choice criteria for AIS data example

Model	Log-likelihood (max)	EAIC	EBIC	DIC_2	ARI
SN	−1071.2	2200.6	**2266.7**	**2176.3**	0.52
SNGH	−1071.8	2205.2	2278.3	2186.8	0.64
(CFUST) SUT ($q = 2$)	−1074.3	2229.8	2309.2	2216.1	0.69
SUNGH ($q = 2$)	−1056.4	**2197.8**	2277.2	2186.8	**0.79**

The best values are indicated in bold; SN and SUT denote the restricted skew-normal and unrestricted skew-t distributions, respectively

the highest log-likelihood and the lowest EAIC, but the SN model has the lowest values for the EBIC and DIC_2 measures. However, the ARI for the SN model (= 0.52) is considerably lower than the SUNGH model (= 0.79) suggesting more support for the SUNGH model in terms of classification accuracy. We can also see these results reflected visually in Fig. 2 with the SUNGH (Fig. 2d) able to represent the skewed nature of the distribution for the two groups, particularly for the male athletes. In contrast, the SN model (Fig. 2c) poorly represents the skewed distribution of the female athletes and the heavy-tailed nature of the distribution for the male athletes. As can be expected, the classification results for the SNGH model (Fig. 2b) are visually similar to the SUNGH; however, small differences (due to the reduced flexibility of the skewness parameter for the SNGH) can be observed which impact greatly on the classification accuracy (ARI = 0.64).

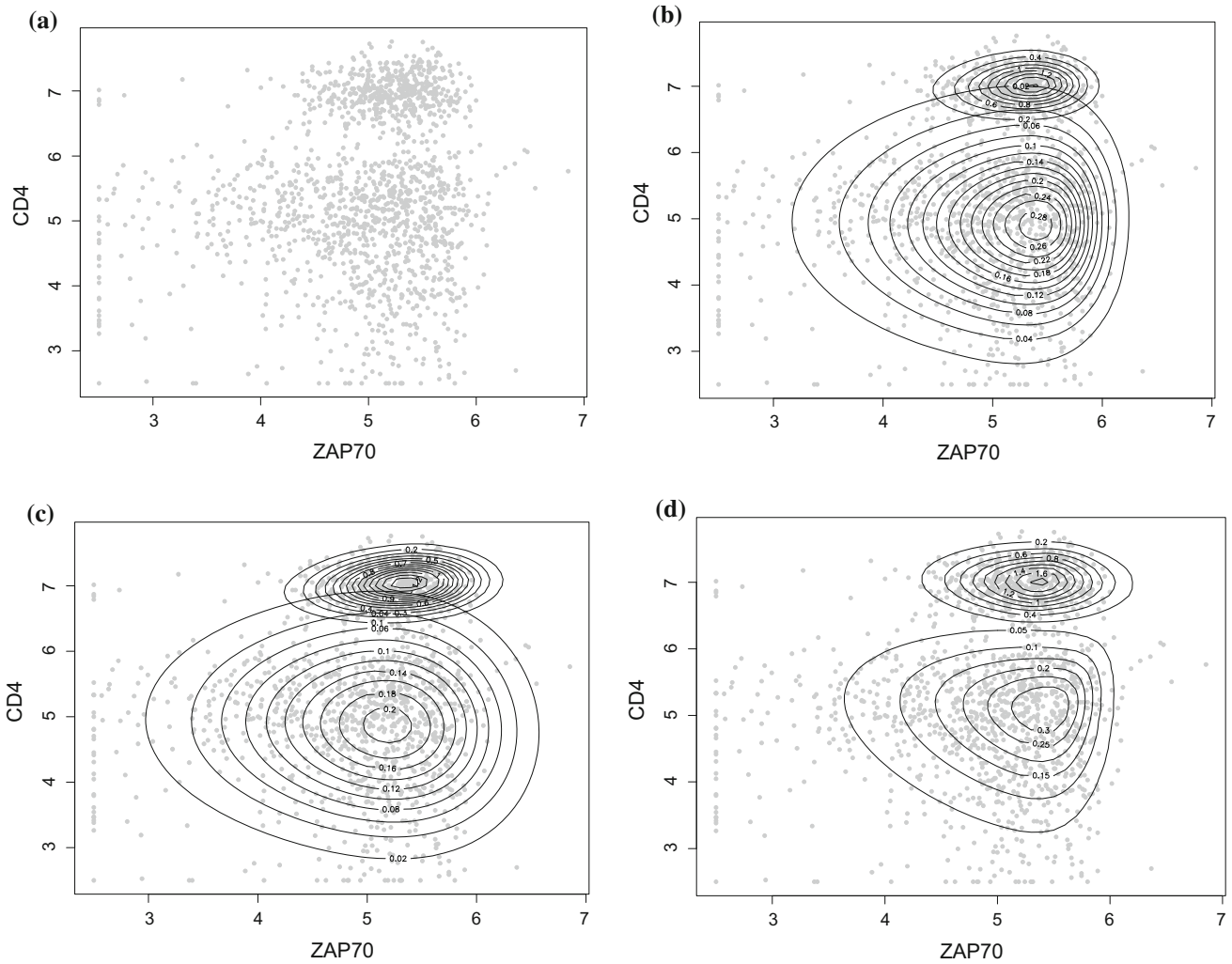

Fig. 3 Plots of observations and fitted contours for lymphoma example: **a** plot of observations for CD4 and ZAP70, **b** SNGH, **c** SN, **d** SUNGH

5.2.2 Lymphoma example

In another example, we examine a clustering problem for a lymphoma dataset analyzed by Lee and McLachlan (2013b). The data consist 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, SLP 76(pY 128) and ZAP 70 (pY 292) before and after an anti-CD3 stimulation. To illustrate the performance of different distributions within the SUNGH family, we will look at clustering a subset of the data containing the variables CD4 and ZAP70 (Fig. 3), which appear to be bimodal and display an asymmetric pattern. In particular, the largest mode appears to show both strong correlation between the two variables and substantial skewness in both dimensions.

From Fig. 3, we can see a clear difference between the SNGH and SUNGH models with the latter providing a closer fit to the two groups visible in the data. This is further supported by the model choice criteria with all three measures (EAIC, EBIC and DIC_2) favoring the SUNGH model (Table 4) compared to SNGH. The SUNGH model is also preferred over the SN model (Fig. 3c) with the latter model not appearing to fit or represent the larger component in the data. Overall, the SUNGH model appears to fit the two groups in these data quite well with the lowest values for two of the three model choice measures (EAIC and EBIC). Using the DIC_2 criteria, the lowest value appears to be for the SUT model, so there is some support for this in terms of model choice. However, the difference between DIC_2 values for this model and SUNGH is not great (SUNGH = 7201.3; SUT = 7198.5), suggesting little difference in terms of this measure.

Table 4 Model choice criteria for lymphoma example

Model	Log-likelihood (max)	EAIC	EBIC	DIC_2
SN	−3657.0	7433.6	7539.8	7473.1
(CFUSN) SUN ($q = 2$)	−3567.8	7507.1	7634.6	7782.6
ST	−3577.6	7240.7	7347.0	7246.2
(CFUST) SUT ($q = 2$)	−3404.5	7051.8	7179.3	**7198.5**
SUNGH ($q = 2$)	**−3396.4**	**7045.0**	**7172.5**	7201.3
SNGH	−3578.9	7252.4	7358.7	7266.9

The best values are indicated in bold; SN and SUN denote the restricted and unrestricted skew-normal distributions, respectively; ST and SUT denote the restricted and unrestricted skew-t distributions, respectively

6 Conclusion

We have proposed a flexible family of unrestricted skew-normal generalized hyperbolic (SUNGH) distributions for application in clustering problems which are capable of representing distributions of asymmetric and heavy-tailed forms. The family contains several other well-known symmetric and asymmetric families of distributions such as scale mixtures of the skew-normal family (SMSN) as special cases. Various properties of the SUNGH family are well defined, and estimation of the parameters is relatively straightforward in a Bayesian framework with most of the Gibbs sampling updates available in closed form. Assessments of the performance of the proposed model on simulated and real data suggest that the family provides a considerable degree of freedom and flexibility in modeling data of varying tail behavior and directional shape. As this family of distributions and the parameterization we have adopted preserves several important propositions (e.g., closed under linear combinations), the SUNGH family can be used in a variety of other statistical models (e.g., linear multilevel/mixed models and regression).

Acknowledgements The authors would like to thank the coordinating editor and anonymous reviewers for their suggestions, corrections and encouragement, which helped us to improve earlier versions of the manuscript.

Appendix

A.1. Proof of Propositions 1 to 6

In this appendix, we prove Propositions 1 to 6.

Proof of Proposition 1 By considering (7),

$$\begin{aligned}
\text{(a):}\quad \boldsymbol{\mu}_{\boldsymbol{Y}} &= E[\boldsymbol{Y}] = E_U[E_{\boldsymbol{Y}|U}[\boldsymbol{Y}]] \\
&= E_U[E_{\boldsymbol{Y}|U}[\boldsymbol{\mu} + \kappa(u)^{1/2}\boldsymbol{\Lambda}|\boldsymbol{W}_0| + \kappa(u)^{1/2}\boldsymbol{\Sigma}^{1/2}\boldsymbol{W}_1]] \\
&= E_U[\boldsymbol{\mu} + \kappa(U)^{1/2}\boldsymbol{\Lambda}E|\boldsymbol{W}_0| + \kappa(U)^{1/2}\boldsymbol{\Sigma}^{1/2}E(\boldsymbol{W}_1)] \\
&= \boldsymbol{\mu} + k_1\boldsymbol{\Lambda}E|\boldsymbol{W}_0|,
\end{aligned}$$

$$\begin{aligned}
\text{(b):}\quad \text{Var}[\boldsymbol{Y}] &= E[(\boldsymbol{Y} - \boldsymbol{\mu}_{\boldsymbol{Y}})(\boldsymbol{Y} - \boldsymbol{\mu}_{\boldsymbol{Y}})^\top] \\
&= E_U[E_{\boldsymbol{Y}|U}[(\boldsymbol{Y} - \boldsymbol{\mu}_{\boldsymbol{Y}})(\boldsymbol{Y} - \boldsymbol{\mu}_{\boldsymbol{Y}})^\top]] \\
&= E_U[E_{\boldsymbol{Y}|U}[(\boldsymbol{\Lambda}[\kappa(u)^{1/2}|\boldsymbol{W}_0| - k_1E|\boldsymbol{W}_0|] \\
&\quad + \kappa(u)^{1/2}\boldsymbol{\Sigma}^{1/2}\boldsymbol{W}_1)([\kappa(u)^{1/2}|\boldsymbol{W}_0^\top| \\
&\quad - k_1E|\boldsymbol{W}_0^\top|]\boldsymbol{\Lambda}^\top + \kappa(u)^{1/2}\boldsymbol{W}_1^\top\boldsymbol{\Sigma}^{1/2})]] \\
&= \boldsymbol{\Lambda}E_U[E_{\boldsymbol{Y}|U}\{[\kappa(U)^{1/2}|\boldsymbol{W}_0| \\
&\quad - k_1E|\boldsymbol{W}_0|][\kappa(U)^{1/2}|\boldsymbol{W}_0^\top| \\
&\quad - k_1E|\boldsymbol{W}_0^\top|]\}]\boldsymbol{\Lambda}^\top \\
&\quad + E_U[\kappa(U)\boldsymbol{\Sigma}^{1/2}E_{\boldsymbol{Y}|U}[\boldsymbol{W}_1\boldsymbol{W}_0^\top]\boldsymbol{\Sigma}^{1/2}] \\
&= \boldsymbol{\Lambda}[k_2E|\boldsymbol{W}_0||\boldsymbol{W}_0^\top| - k_1^2E|\boldsymbol{W}_0|E|\boldsymbol{W}_0^\top|]\boldsymbol{\Lambda}^\top + k_2\boldsymbol{\Sigma} \\
&= \boldsymbol{\Lambda}\left[k_2\left(\frac{2}{\pi}\mathbf{1}_q\mathbf{1}_q^\top + \left(1 - \frac{2}{\pi}\right)\boldsymbol{I}_q\right) - \frac{2}{\pi}k_1^2\mathbf{1}_q\mathbf{1}_q^\top\right] \\
&\quad \boldsymbol{\Lambda}^\top + k_2\boldsymbol{\Sigma} \\
&= k_2\boldsymbol{\Psi} + \boldsymbol{\Lambda}\left[(k_2 - k_1^2)\frac{2}{\pi}\mathbf{1}_q\mathbf{1}_q^\top - \frac{2}{\pi}k_2\boldsymbol{I}_q\right]\boldsymbol{\Lambda}^\top
\end{aligned}$$

□

Proof of Proposition 2 By considering the stochastic representation (7) and the fact that $\boldsymbol{W}_0$ (and so $\boldsymbol{W}$) are uncorrelated, this subject proved. In the case of $\boldsymbol{\Lambda}^* = \left(\boldsymbol{\Lambda}_{p\times q}\ \mathbf{0}_{p\times m}\right)$, relation (7) for $\boldsymbol{Y} \sim \text{SUNGH}_{p,q+m}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}^*, \varpi\right)$ is equivalent to $\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\Lambda}^*\boldsymbol{W} + \kappa\,(U)^{1/2}\,\boldsymbol{\Sigma}^{1/2}\boldsymbol{W}_1 = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{W}^{(1)} + \kappa\,(U)^{1/2}\,\boldsymbol{\Sigma}^{1/2}\boldsymbol{W}_1$, where $\boldsymbol{W}^{(1)}$ is the first q components of W, and in the case of $\boldsymbol{\Lambda}^* = \left(\mathbf{0}_{p\times m}\ \boldsymbol{\Lambda}_{p\times q}\right)$, relation (7) for $\boldsymbol{Y} \sim \text{SUNGH}_{p,q+m}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}^*, \boldsymbol{\varpi}\right)$ is equivalent to $\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\Lambda}^*\boldsymbol{W} + \kappa\,(U)^{1/2}\,\boldsymbol{\Sigma}^{1/2}\boldsymbol{W}_1 = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{W}^{(2)} + \kappa\,(U)^{1/2}\,\boldsymbol{\Sigma}^{1/2}\boldsymbol{W}_1$, where $\boldsymbol{W}^{(2)}$ is the last q components of $\boldsymbol{W}$ □

Proof of Proposition 3 By considering the stochastic representation (7), we have that $\boldsymbol{b} + \boldsymbol{B}\boldsymbol{Y} = \boldsymbol{b} + \boldsymbol{B}\boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{\Lambda}\boldsymbol{W} + \kappa\,(U)^{1/2}\left(\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}^\top\right)^{1/2}\boldsymbol{W}_1$ □

Proof of Proposition 4 By considering Proposition 3, with $\boldsymbol{b} = \mathbf{0}$ and the matrix $\boldsymbol{B}$ in the form of $\left(\boldsymbol{I}_{p_1}\ \mathbf{0}_{p_1\times p_2}\right)$ or $\left(\mathbf{0}_{p_2\times p_1}\ \boldsymbol{I}_{p_2}\right)$, respectively, this subject proved □

Proof of Proposition 5 Since $\boldsymbol{Y} = \left(\boldsymbol{Y}_1^\top, \boldsymbol{Y}_2^\top\right)^\top$, from part b) of the Proposition 1, we have $\text{Var}[\boldsymbol{Y}] = \left(\text{Cov}\left(\boldsymbol{Y}_i, \boldsymbol{Y}_j\right)\right)_{i,j=1,2}$

$= \left(\boldsymbol{\Sigma}_{ij} + \boldsymbol{\Lambda}_i \left[\left(k_2 - k_1^2\right)\frac{2}{\pi}\mathbf{1}_q\mathbf{1}_q^\top - \frac{2}{\pi}k_2 I_q\right]\boldsymbol{\Lambda}_j^\top\right)$. Thus, if $\boldsymbol{\Sigma}_{12} = \mathbf{0}$, then $\mathrm{Cov}\left(\boldsymbol{Y}_1, \boldsymbol{Y}_2\right) = \boldsymbol{\Lambda}_1\left[\left(k_2 - k_1^2\right)\frac{2}{\pi}\mathbf{1}_q\mathbf{1}_q^\top - \frac{2}{\pi}k_2 I_q\right]\boldsymbol{\Lambda}_2^\top$, thus following that each of the conditions $\boldsymbol{\Lambda}_1 = \mathbf{0}$ or $\boldsymbol{\Lambda}_2 = \mathbf{0}$ leads to $\mathrm{Cov}\left(\boldsymbol{Y}_1, \boldsymbol{Y}_2\right) = \mathbf{0}$ □

Proof of Proposition 6 The first part follows by applying Proposition 2 in the Proposition 4. For the proof of the second result, note from the proof of Proposition 5 that

$$\mathrm{Cov}\left(\boldsymbol{Y}_1, \boldsymbol{Y}_2\right) = \left(\boldsymbol{\Lambda}_{11}\ \mathbf{0}_{p_1\times q_2}\right)_{p_1\times q} \left[\left(k_2 - k_1^2\right)\frac{2}{\pi}\mathbf{1}_q\mathbf{1}_q^\top - \frac{2}{\pi}k_2 I_q\right]_{q\times q} \left(\mathbf{0}_{p_2\times q_1}\ \boldsymbol{\Lambda}_{22}\right)_{q\times p_2}^\top.$$

Thus, using the partitions $\boldsymbol{I}_q = \mathrm{diag}\left(\boldsymbol{I}_{q_1}, \boldsymbol{I}_{q_2}\right)$ and $\mathbf{1}_q = \left(\mathbf{1}_{q_1}^\top, \mathbf{1}_{q_2}^\top\right)^\top$ we obtain the proof □

A.2. Matrix variate priors for skewness matrix

Considering the matrix variate priors in the form of $\boldsymbol{\Lambda}_k \sim MN_{p,q}\left(\boldsymbol{N}_k, \boldsymbol{S}_k, \boldsymbol{F}_k\right), k = 1, \ldots, K$, where MN denotes the matrix normal distributions, this leads to the following posteriors instead of (19) as follows:

$\mathrm{vec}(\boldsymbol{\Lambda}_k)|\,\boldsymbol{\Theta}_{(-\boldsymbol{\Lambda}_k)}, \boldsymbol{y}, \boldsymbol{u}, \boldsymbol{w}, z_i = k \sim N_{pq}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right); k = 1, \ldots, K$, where

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}\left[\left(\mathbf{S}_k \otimes \boldsymbol{F}_k\right)^{-1}\mathrm{vec}\left(\boldsymbol{N}_k\right) + \sum_{B_k}\kappa\left(u_{ik}\right)^{-1}\left(\boldsymbol{M}_{ik}^\top \otimes \boldsymbol{\Sigma}_k^{-1}\right)\right], \tag{19a}$$

$$\boldsymbol{\Sigma} = \left[\left(\mathbf{S}_k \otimes \boldsymbol{F}_k\right)^{-1} + \sum_{B_k}\kappa\left(u_{ik}\right)^{-1}\left(\boldsymbol{\Sigma}_k^{-1} \otimes \boldsymbol{L}_{ik}\right)\right]^{-1},$$

where $\boldsymbol{L}_{ik} = \boldsymbol{w}_{ik}\boldsymbol{w}_{ik}^\top$ and $\boldsymbol{M}_{ik} = \left(\boldsymbol{y}_i - \boldsymbol{\mu}_k\right)\boldsymbol{w}_{ik}^\top$, for which $\otimes$ denotes the Kronecker product and vec denotes the vectorization of a matrix (a linear transformation which converts the matrix into a column vector).

Using these forms for the Gibbs updates may improve mixing and convergence to a stationary distribution. However, they involve the use of matrix variate distributions for which users may not be familiar; hence, a simpler (computational) update is provided in the main text.

References

Andrews, D.R., Mallows, C.L.: Scale mixture of normal distribution. J. Roy. Stat. Soc. B **36**, 99–102 (1974)

Arellano-Valle, R.B., Azzalini, A.: On the unification of families of skew-normal distributions. Scand. J. Stat. **33**, 561–574 (2006)

Arellano-Valle, R.B., Genton, M.G.: On fundamental skew distributions. J. Multivar. Anal. **96**, 93–116 (2005)

Arellano-Valle, R.B., Genton, M.G.: Multivariate unified skew-elliptical distributions. Chil. J. Stat. **2**, 17–34 (2010)

Arellano-Valle, R.B., Branco, M.D., Genton, M.G.: A unified view on skewed distributions arising from selections. Can. J. Stat. **34**, 581–601 (2006)

Arellano-Valle, R.B., Bolfarine, H., Lachos, G.H.: Bayesian inference for skew-normal linear mixed model. J. Appl. Stat. **33**, 561–574 (2007)

Azzalini, A.: Package 'sn'. http://azzalini.stat.unipd.it/SN (2015). Accessed 13 May 2017

Azzalini, A., with the collaboration of Capitanio, A.: The Skew-Normal and Related Families. IMS Monographs Series. Cambridge University Press (2014)

Barndorff-Nielsen, O.: Hyperbolic distributions and distributions on hyperbolae. Scand. J. Stat. **5**, 151–157 (1978)

Barndorff-Nielsen, O., Blaesild, P.: Hyperbolic distributions. In: Kotz, S., Johnson, N.L., Read, C. (eds.) Encyclopedia of Statistical Sciences, vol. 3. Wiley, New York (1980)

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

Basso, R.M., Lachos, V.H., Cabral, C.R.B., Ghosh, P.: Robust mixture modeling based on the scale mixtures of skew-normal distributions. Comput. Stat. Data Anal. **54**, 2926–2941 (2010)

Böhning, D.: Computer-Assisted Analysis of Mixtures and Applications. Meta-Analysis, Disease Mapping and Others. Chapman & Hall, Boca Raton (2000)

Branco, M.D., Dey, D.K.: A general class of multivariate skew-elliptical distributions. J. Multivar. Anal. **79**, 99–113 (2001)

Browne, R.P., McNicholas, P.D.: A mixture of generalized hyperbolic distributions. Can. J. Stat. **43**(2), 176–198 (2015)

Carlin, B.P., Louis, T.A.: Bayesian Methods for Data Analysis. CRC Press, Boca Raton (2011)

Celeux, G., Hurn, M., Robert, C.P.: Computational and inferential difficulties with mixture posterior distributions. J. Am. Stat. Assoc. **95**, 957–970 (2000)

Celeux, G., Forbes, F., Robert, C.P., Titterington, D.M.: Deviance information criteria for missing data models. Bayesian Anal. **1**, 651–674 (2006)

Chhikara, R.S., Folks, J.L.: The Inverse Gaussian Distribution. Marcel Dekker, New York (1989)

Cook, R.D., Weisberg, S.: An Introduction to Regression Graphics. Wiley, New York (1994)

Forbes, F., Wraith, D.: A new family of multivariate heavy-tailed distributions with variable marginal amounts of tail weight: application to robust clustering. Stat. Comput. **24**(6), 971–984 (2014)

Franczak, B.C., Browne, R.P., McNicholas, P.D.: Mixtures of shifted asymmetric laplace distributions. IEEE Trans. Pattern Anal. Mach. Intell. **36**(6), 1149–1157 (2014)

Frühwirth-Schnatter, S.: Finite Mixture and Markov Switching Models. Springer Series in Statistics. Springer, Berlin (2006)

Frühwirth-Schnatter, S., Pyne, S.: Bayesian inference for finite mixtures of skew-normal and skew-t distributions. Biostatistics **11**(2), 317–336 (2010)

Gelman, A., Rubin, D.B.: Inference from iterative simulation using multiple sequences. Stat. Sci. **7**, 457–511 (1992)

Genton, M.G.: Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality. Chapman & Hall, Boca Raton (2004)

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

Hogan, J.W., Laird, N.M.: Mixture models for the joint distribution of repeated measures and event times. Stat. Med. **16**, 239–258 (1997)

Holzmann, H., Munk, A., Gneiting, T.: Identifiability of finite mixtures of elliptical distributions. Scand. J. Stat. **33**(4), 753–763 (2006)

Hubert, L., Arabie, P.: Comparing partitions. J. Classif. **2**, 193–218 (1985)
Johnson, N.L., Kotz, S., Balakrishnan, N.: Continous Univariate Distributions, vol. 1. Wiley, New York (1994)
Jørgensen, B.: Statistical Properties of the Generalized Inverse Gaussian distribution. Springer, New York (1982)
Karlis, D., Santourian, A.: Model-based clustering with non-elliptically contoured distributions. Stat. Comput. **19**(1), 73–83 (2009)
Lachos, V.H., Bolfarine, H., Arellano-Valle, R.B.: Likelihood-based inference for multivariate skew-normal regression models. Commun. Stat. Theory Methods **36**(9), 1769–1786 (2007)
Lachos, V.H., Ghosh, P., Arellano-Valle, R.B.: Likelihood based inference for skew-normal independent linear mixed models. Stat. Sin. **20**, 303–322 (2010)
Lee, S.X., McLachlan, G.J.: Model-based clustering and classification with non-normal mixture distributions. Stat. Methods Appl. **22**(4), 427–454 (2013a)
Lee, S.X., McLachlan, G.J.: On mixtures of skew normal and skew t distributions. Adv. Data Anal. Classif. **7**(3), 241–266 (2013b)
Lee, S.X., McLachlan, G.J.: Finite mixtures of multivariate skew t distributions: some recent and new results. Stat. Comput. **24**, 181–202 (2014)
Lee, S.X., McLachlan, G.J.: Finite mixtures of canonical fundamental skew t-distributions: the unification of the restricted and unrestricted skew t-mixture models. Stat. Comput. **26**, 573–589 (2016)
Lin, T.I.: Maximum likelihood estimation for multivariate skew normal mixture models. J. Multivar. Anal. **100**(2), 257–265 (2009)
Lin, T.I.: Robust mixture modeling using multivariate skew t distributions. Stat. Comput. **20**(3), 343–356 (2010)
Lin, T.I., Lee, J.C., Yen, S.Y.: Finite mixture modeling using the skew-normal distribution. Stat. Sin. **17**(b), 909–927 (2007)
Lin, T.I., Ho, H.J., Chen, C.L.: Analysis of multivariate skew normal models with incomplete data. J. Multivar. Anal. **100**(10), 2337–2351 (2009)
Maier, L.M., Anderson, D.E., De Jager, P.L., Wicker, L.S., Hafler, D.A.: Allelic variant in CTLA4 alters t cell phosphorylation patterns. Proc. Natl. Acad. Sci. USA **104**, 18607–18612 (2007)
Maleki, M., Arellano-Valle, R.B.: Maximum a-posteriori estimation of autoregressive processes based on finite mixtures of scale-mixtures of skew-normal distributions. J. Stat. Comput. Simul. **87**(6), 1061–1083 (2017)
McLachlan, G.J., Peel, D.: Finite Mixture Models. Wiley, Chichester (2000)
McNeil, A.J., Frey, R., Embrechts, P.: Quantitative Risk Management: Concepts, Techniques and Tools. Princeton University Press, Princeton (2005)
Mengersen, K., Robert, C., Titterington, D.M.: Mixtures: Estimation and Applications. Wiley, Chichester (2011)
Morris, K., McNicholas, P.D., Punzo, A., Browne, R.P.: Robust Asymmetric Clustering. ArXiv e-print arxiv:1402.6744 (2014)
Pyne, S., Hu, X., Wang, K., Rossin, E., Lin, T.I., Maier, L.M., Baecher-Allan, C., McLachlan, G.J., Tamayo, P., Hafler, D.A., De Jager, P.L., Mesirov, J.P.: Automated high-dimensional flow cytometric data analysis. Proc. Natl. Acad. Sci. **106**(21), 8519–8524 (2009)
R Core Team.: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ (2017). Accessed 20 June 2017
Sahu, S.K., Dey, D.K., Branco, M.D.: A new class of multivariate skew distributions with applications to Bayesian regression models. Can. J. Stat. **31**(2), 129–150 (2003)
Seshadri, V.: The Inverse Gaussian Distribution: A Case Study in Exponential Families. Oxford University Press, New York (1993)
Teicher, H.: Identifiability of finite mixtures. Ann. Math. Stat. **34**(4), 1265–1269 (1963)
Vilca, F., Balakrishnan, N., Zeller, C.B.: Multivariate skew-normal generalized hyperbolic distribution and its properties. J. Multivar. Anal. **128**, 73–85 (2014)
Vrbik, I., McNicholas, P.D.: Analytic calculations for the EM algorithm for multivariate skew-t mixture models. Stat. Probab. Lett. **82**(6), 1169–1174 (2012)
Wang, H.X., Zhang, Q.B., Luo, B., Wei, S.: Robust mixture modelling using multivariate t-distribution with missing information. Pattern Recogn. Lett. **25**(6), 701–710 (2004)
Wang, K., Ng, S.K., McLachlan, G.J.: Multivariate skew t mixture models: applications to fluorescence-activated cell sorting data. In: Digital Image Computing: Techniques and Applications, Los Alamitos, California, pp. 526–531. IEEE (2009)
Wraith, D., Forbes, F.: Location and scale mixtures of Gaussians with flexible tail behaviour: properties, inference and application to multivariate clustering. Comput. Stat. Data Anal. **90**(Oct.), 61–73 (2015)