Computational Statistics (2019) 34:1039–1053
https://doi.org/10.1007/s00180-019-00870-6

ORIGINAL PAPER

Mixtures of multivariate restricted skew-normal factor analyzer models in a Bayesian framework

Mohsen Maleki[1] · **Darren Wraith**[2]

Received: 11 December 2017 / Accepted: 21 January 2019 / Published online: 31 January 2019

Abstract
The mixture of factor analyzers (MFA) model, by reducing the number of free parameters through its factor-analytic representation of the component covariance matrices, is an important statistical model to identify hidden or latent groups in high dimensional data. Recent approaches to extend the approach to skewed data or skewness in the latent groups have been examined in a frequentist setting where there are some known computational limitations. For these reasons we consider a Bayesian approach to the restricted skew-normal mixtures of factor analysis MFA model. We examine the performance and flexibility of the approach on real datasets and illustrate some of the computational advantages in a missing data setting.

Keywords Bayesian analysis · Gibbs sampling · Mixture of factor analysis model · Restricted skew-normal distribution

1 Introduction

Factor analysis (*FA*) models and finite mixture (*FM*) models are both popular statistical techniques which have wide application to the analysis of data and extraction of hidden or latent variables. In a *FA* model, the covariance relationship between variables can be explained by a fewer number of latent variables or latent factors which can be used to simplify analysis in high-dimensional settings or establish common themes or constructs (e.g., in psychometric testing). The *FA* model has wide applications in various fields such as social sciences, biology, medical sciences and epidemiological studies. An *FM* model is also a latent variable model and can represent the presence of subpopulations within an overall population, without requiring that an observed data set should identify the sub-population to which an individual

✉ Darren Wraith
d.wraith@qut.edu.au

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

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

observation belongs [for further details see e.g. McLachlan and Peel (2000) and references therein and Lee and McLachlan (2013a, b)].

A combination of the *FA* and *FM* models based on the Gaussian distribution was first studied by Hinton et al. (1997) and Ghahramani and Hinton (1997), commonly called the Gaussian mixture of factor analysis (*MFA*) model. However, in many applied problems, the data may be moderately or severely skewed which can result in seriously misleading inference even for small departures from normality (Wall et al. 2012). In practice, the imposition of symmetry in the components of the mixture models may be a fairly restrictive condition. For example Lin et al. (2007), Maleki and Arellano-Valle (2017) and Maleki et al. (2018a), argue the normal mixture model tends to over-fit when additional components are included to capture the skewness and, sometimes, increasing the number of components may lead to difficulties in computations (e.g. small number of observations belonging to a group) and interpretation of results (See discussion and results in Murray et al. 2014).

In recent years, there has been much research on asymmetrical families of distributions which contain the Gaussian family as a special (symmetrical) case. For example, the class of Skew-Normal distribution studied by Azzalini (1985), Azzalini and Dalla-Vale (1996), Azzalini and Capitanio (1999) and Arellano-Valle and Azzalini (2006), have wide application in many statistical models (for more asymmetrical distributions and their applications see Azzalini 2014). In particular, Azzalini and Dalla-Vale (1996) and Sahu et al. (2003) studied the so called restricted multivariate skew-normal (*rMSN* or *rSN*) distribution which is suitable for analyzing skewed multivariate data as well as symmetrically distributed. Recently, Lin et al. (2016) applied the *rMSN* distribution to the structure of the *MFA* model, hereafter called the mixtures of skew-normal factor analyzers (*MSNFA*) model. In the *rMSN* distribution, skewness is controlled by a vector of skewness parameters multiplied by a common skewing variable in its convolution type representation. An alternative formulation of skew distributions is through the use of so called unrestricted forms in which there is no reliance on a common skewing variable and skewness is allowed to be represented in more than one direction [in contrast to a single direction for the restricted case; see e.g. Lee and McLachlan (2013b), Maleki et al. (2018b)]. However, this extra flexibility can lead to identifiability issues (the skewness matrix is not rotation invariant) and there are also issues in terms of computational tractability. These forms and computational issues can be explored in future research.

In the *MSNFA* model the latent component factors follow the family of *rMSN* distributions in an attempt to model the data adequately in the presence of skewed sub-populations. The *MSNFA* model has a novel approach to dimension reduction and representing appropriately non-normal data. Lin et al. (2016) used an *EM*-type algorithm to obtain the maximum likelihood (*ML*) estimates of the proposed model parameters and estimated factor scores by products within the estimation procedure.

Most estimation methods for *MFA* models are classical inferences based on the maximum likelihood (*ML*) estimates. However, the likelihood function of the *MFA* and *FM* models can be unbounded for some samples and this is a problematic issue, so some researchers have considered the Bayesian approach to estimate the Gaussian *MFA* model. Bishop (1999) proposed a partial Bayesian framework to the mixture of principal component analyzer (PCA), which is an isotropic version of the

MFA based on the Gaussian distribution. Bishop (1999) used a maximum a-posteriori (*MAP*) method of estimation by using a simple Gaussian prior for the factor loadings, and also by using approximate Bayesian inference, derived an algorithm for estimation of the hyper-parameters (parameters of the prior). Ghahramani and Beal (2000) proposed an efficient and deterministic variational approximation to full Bayesian integration of Gaussian *MFA* model parameters. Ustugi and Kumagai (2001) introduced a full prior on all parameters of the Gaussian *MFA* model by using conjugate priors.

More recent extensions to the Gaussian *MFA* model in a Bayesian framework include a matrix variate t distribution for the factor scores (Ando 2009) and normal/independent distributions for the error term to provide for outliers and a robust specification (Lee and Xia 2008a). A number of other extensions have focused on semi-parametric models (Yang and Dunson 2010; Lee and Xia 2008b; Song et al. 2010; Murray et al. 2013), non-parametric approaches (Chen et al. 2010; Paisley and Carin 2009) and allowing for flexible prior distributions (Ghosh and Dunson 2009). Other extensions have focused on exploiting the use of prior distributions or information for sparse applications in high-dimensional settings (Carvalho et al. 2008; Knowles and Ghahramani 2007; Paisley and Carin 2009; Bhattacharya and Dunson 2011) and in a dynamic time series context (Chen et al. 2011). A difficulty with some of the more flexible semi-parametric or non-parametric models proposed for the factor analysis models is there is often a sacrifice that is made in terms of interpretation, parsimony and computation. This is particularly an issue in the factor analysis context where various forms of the factor loading matrix can be derived and where simplicity and interpretability often become appealing for users (see, e.g., Frühwirth-Schnatter and Lopes 2012; Conti et al. 2014).

There are several other computational advantages of using a Bayesian approach for mixtures of factor analysis models (compared to *ML* estimation) including the use of prior information or specification of prior distributions to regularize the parameter space, particularly in high dimensional settings (Carvalho et al. 2008) and/or in cases where there is considerable noise (e.g., imaging data). In particular, Suarez and Ghosal (2016) examine the performance of placing a prior distribution on the error term of a principal components approach for functional data with the degree of informativeness or smoothing determined by *apriori* knowledge or empirically derived. We note that this information is relatively easily included in a Bayesian model without the need for introducing additional computational demands or complexity. Finally, we note that the number of components and factors could be allowed to vary and be updated as part of the computational approach (Frühwirth-Schnatter and Lopes 2012).

Extensions to the more general case of structural equation modeling are also relatively easier than in the *ML* estimates setting (e.g., Lee and Xia 2008a). Further extensions to allow 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*). Computation of the standard error or uncertainty of parameter estimates also does not rely on using asymptotic approximations to the

observed information matrix if the sample size is large or resorting to a bootstrap method which requires a very large amount of computations (Basso et al. 2010).

In this paper, we consider the *MSNFA* model of Lin et al. (2016) and propose a Bayesian inference with full priors of all model parameters. This parametric model has several desirable properties, including representation of the symmetrical *MFA* model as a special case. The distribution also has a convenient hierarchical representation which leads to closed form marginal posteriors and facilitates ease of computations using a Gibbs sampler *MCMC* algorithm to estimate the model parameters. To illustrate the flexibility of the Bayesian approach in this setting, we also consider the performance of the model in missing data settings.

The paper is organized as follows. In Sect. 2, we provide a review and background to the *rSN* distribution and the *MSNFA* model. Section 3 presents a Bayesian analysis of the *MSNFA* and details of the Gibbs sampling algorithm. In Sect. 4, we illustrate the performance of the proposed model on real datasets. Finally, in Sect. 5, we present our main conclusions and discuss possible extensions and areas of further research.

2 A review of the multivariate rSN family and MSNFA model

In this part we begin with a brief review of the multivariate restricted skew normal (*rMSN*) family introduced and studied by Azzalini and Dalla-Vale (1996) and Lee and McLachlan (2013b). We then outline details for the mixtures of factor analysis model based on the *rMSN* family.

2.1 The multivariate restricted skew normal family

A q-dimensional random vector $\boldsymbol{X}$ follows an *rMSN* distribution with q-dimensional location vector $\boldsymbol{\mu}$, $q\times q$ positive definite dispersion matrix $\boldsymbol{\Sigma}$, and q-dimensional skewness parameter vector λ, denoted by $\boldsymbol{X} \sim rSN_q(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \lambda)$, can be constructed stochastically by

$$\boldsymbol{X} = \boldsymbol{\mu} + \lambda W + \boldsymbol{\Sigma}^{1/2}\boldsymbol{V}, \tag{1}$$

where $W = |V_0|$ is the absolute value of $V_0 \sim N_1(0,1)$ and independent of $\boldsymbol{V} \sim N_q(0, \boldsymbol{I}_q)$. Note that $E(\boldsymbol{X}) = \boldsymbol{\mu} + c\lambda$ and $\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma} + \left(1 - c^2\right)\lambda\lambda^\top$, where $c = \sqrt{2/\pi}$.

Considering the stochastic representation of $\boldsymbol{X} \sim rSN_q(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \lambda)$ leads to the following probability density function (pdf)

$$f(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \lambda) = 2\phi_q(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Omega})\boldsymbol{\Phi}_1\left(\sigma^{-1}\lambda^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right), \quad \boldsymbol{x} \in \boldsymbol{R}^q, \tag{2}$$

where $\boldsymbol{\Omega} = \boldsymbol{\Sigma} + \lambda\lambda^\top$, $\sigma^2 = 1 - \lambda^\top\boldsymbol{\Omega}^{-1}\lambda = \left(1 + \lambda^\top\boldsymbol{\Sigma}^{-1}\lambda\right)^{-1}$, and $\phi_q(\cdot|\boldsymbol{\mu}, \boldsymbol{\Omega})$ and $\boldsymbol{\Phi}_1(\cdot)$ are, respectively, the probability distribution function (pdf) for the multivariate normal distribution $N_q(\boldsymbol{\mu}, \boldsymbol{\Omega})$, and the cumulative distribution function (cdf) of the standard univariate normal distribution.

Also the random vector $\boldsymbol{X} \sim rSN_q(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ has the following hierarchical representation:

$$\begin{aligned} \boldsymbol{X}|W = w &\sim N_q(\boldsymbol{\mu} + \boldsymbol{\lambda} W, \boldsymbol{\Sigma}), \\ W &\sim HN_1(0, 1), \end{aligned} \tag{3}$$

where HN_1 is the univariate right-half of standard normal distribution

For more details about this family of distributions (including the mean, variance, the moment generating function, and other interesting properties), see e.g., Lee and McLachlan (2013b), Lin et al. (2016) and Maleki et al (2018a, b).

2.2 The mixture of restricted skew-normal factor model

Lin et al. (2016) introduced the generalization of traditional factor analysis (*FA*), called restricted skew-normal factor model. Given a p-dimentional random sample $\boldsymbol{Y} = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n\}$ and location vector $\boldsymbol{\mu}$, a $p \times q$ matrix of factor loadings $\boldsymbol{L}$, factor analysis finds uncorrelated symmetrical/asymmetrical q-dimensional ($q < p$) vectors of latent factors $\boldsymbol{F}_1, \ldots, \boldsymbol{F}_n$ that explain a large amount of variability in the data, and $\epsilon_1, \ldots, \epsilon_n$ are the p-dimentional vector of Gaussian errors. The factor analysis model for $j = 1, \ldots, n$ can be written as

$$\boldsymbol{Y}_j = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{F}_j + \epsilon_j, \tag{4a}$$

for which the latent factors and model errors are independently distributed as:

$$\boldsymbol{F}_j \overset{iid}{\sim} rSN_q\left(-c\boldsymbol{\Delta}^{-1/2}\boldsymbol{\lambda}, \boldsymbol{\Delta}^{-1}, \boldsymbol{\Delta}^{-1/2}\boldsymbol{\lambda}\right), \quad \boldsymbol{\epsilon}_j \overset{iid}{\sim} N_p(\boldsymbol{0}, \boldsymbol{D}), \tag{4b}$$

where $c = \sqrt{2/\pi}$, the scale matrix $\boldsymbol{\Delta} = \boldsymbol{I}_q + \left(1 - c^2\right)\boldsymbol{\lambda}\boldsymbol{\lambda}^\top$, positive diagonal matrix $\boldsymbol{D} = \text{diag}\left(D_1, \ldots, D_p\right)$ and *SN* denotes a skew-normal distribution. Note that $E\left[\boldsymbol{F}_j\right] = 0$, $\text{Cov}\left[\boldsymbol{F}_j\right] = \boldsymbol{I}_q$, $E\left[\epsilon_j\right] = 0$ and also, $E\left[\boldsymbol{Y}_j\right] = \boldsymbol{\mu}$, $\text{Cov}\left[\boldsymbol{Y}_j\right] = \boldsymbol{L}\boldsymbol{L}^\top + \boldsymbol{D}$. This model we will refer to as the *SNFA* model, and due to Proposition 3 from Lin et al. (2016),

$$\boldsymbol{Y}_j \sim rSN_q(\boldsymbol{\mu} - c\boldsymbol{\alpha}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}), \quad j = 1, \ldots, n, \tag{5}$$

where $\boldsymbol{\alpha} = \boldsymbol{L}\boldsymbol{\Delta}^{-1/2}\boldsymbol{\lambda}$ and $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{\Delta}^{-1}\boldsymbol{L}^\top + \boldsymbol{D}$. To ensure the identifiability of the *SNFA* model (4a, b), we constrain the loading matrix $\boldsymbol{L}$ so that the upper-right triangle is zero and diagonal entries are strictly positive (Fokoué and Titterington 2003; Lopes and West 2004; Lin et al. 2016). At times these conditions may be too restrictive and influence the ordering of the factors, so alternative formulations have been examined in Leung and Drton (2016), Frühwirth-Schnatter and Lopes (2012) and Conti et al. (2014)

Lin et al. (2016) generalize the *SNFA* model to its corresponding mixture model, called Mixture of restricted skew-normal factors denoted by *MSNFA* model with the following details. Let $\boldsymbol{Y}_j = \left(Y_{j1}, \ldots, Y_{jp}\right)^\top, j = 1, \ldots, n$ are p-dimentional vector of p feature variables, for which $\boldsymbol{Y}_j$ follows from finite groups. The latent membership-indicator variables $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ indicate which component each observation belongs to. In detail, $Z_{ij} = \left(\boldsymbol{Z}_j\right)_i$ for $i = 1, \ldots, \text{g}$ and $j = 1, \ldots, n$ is one or zero, according to whether

$\boldsymbol{Y}_j$ belongs or does not belong to the i-th component. These latent variables have multinomial distribution denoted by $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n \sim \mathcal{M}(1;\boldsymbol{\pi})$, for which $\boldsymbol{\pi} = \left(\pi_1, \ldots, \pi_{\mathrm{g}}\right)^\top$, with marginal probability mass function (pmf) given by

$$P\left(\boldsymbol{Z}_j;\boldsymbol{\pi}\right) = \pi_1^{z_{1j}} \pi_2^{z_{2j}} \ldots \pi_{\mathrm{g}}^{z_{\mathrm{g}j}}, \quad j = 1, \ldots, n, \quad \text{subject to} \sum_{i=1}^{\mathrm{g}} \pi_i = 1 \quad \text{and } \pi_i > 0, \quad i = 1, \ldots, \mathrm{g}.$$

So given $Z_{ij} = 1, \boldsymbol{Y}_j$ has the structure

$$Y_j = \mu_i + L_i F_{ij} + \epsilon_{ij}, \quad \text{with probability } \pi_i, \tag{6}$$

where the latent factors and model errors are independently distributed as $\boldsymbol{F}_{ij} \overset{ind}{\sim} rSN_q\left(-c\boldsymbol{\Delta}_i^{-1/2}\lambda_i, \boldsymbol{\Delta}_i^{-1}, \boldsymbol{\Delta}_i^{-1/2}\lambda_i\right), \epsilon_{ij} \overset{ind}{\sim} N_p\left(\boldsymbol{0}, \boldsymbol{D}_i\right)$, for which $\boldsymbol{\Delta}_i = \boldsymbol{I}_q + \left(1 - c^2\right)\lambda_i\lambda_i^\top$ and positive diagonal matrix $\boldsymbol{D}_i = \mathrm{diag}\left(D_{i1}, \ldots, D_{ip}\right)$ for $j = 1, \ldots, n$ and $i = 1, \ldots, \mathrm{g}$.

Also density of $\boldsymbol{Y}_j$ is

$$f\left(\boldsymbol{y}_j|\boldsymbol{\Theta}\right) = \sum_{i=1}^{\mathrm{g}} \pi_i f_i\left(\boldsymbol{y}_j|\boldsymbol{\theta}_i\right); \quad j = 1, \ldots, n, \tag{7}$$

where $f_i\left(\boldsymbol{y}_j|\boldsymbol{\theta}_i\right)$ is the pdf of each *SNFA* component (6) given by (5) and (2), and $\boldsymbol{\theta}_i = \left(\boldsymbol{\mu}_i, \boldsymbol{L}_i, \boldsymbol{D}_i, \lambda_i\right)$, for which $\boldsymbol{\Theta} = \left(\pi_1, \ldots, \pi_{\mathrm{g}-1}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{\mathrm{g}}\right)$. Therefore, the log-likelihood function due to model (6) is given by

$$\mathcal{L}(\boldsymbol{\Theta}|\boldsymbol{y}) = \sum_{j=1}^{n} \log\left(\sum_{i=1}^{\mathrm{g}} \pi_i f_i\left(\boldsymbol{y}_j|\boldsymbol{\theta}_i\right)\right). \tag{8}$$

3 Bayesian analysis

In this section we construct the augmented likelihood function based on the completed data (including the latent variables) to derive the joint posterior distribution.

3.1 Augmented likelihood function

Let $C = \{\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{Z}\}$ denote the complete data, where $\boldsymbol{Y} = \left(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n\right)$, $\boldsymbol{W} = \left(W_1, \ldots, W_n\right)$ and $\boldsymbol{Z} = \left(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n\right)$. In accordance with the hierarchical representation (3) to the model (6), the following flexible hierarchical representations are satisfied:

$$\begin{aligned}
\boldsymbol{Y}_j\Big|\boldsymbol{F}_{ij}, Z_{ij} = 1 &\overset{ind.}{\sim} N_p\left(\boldsymbol{\mu}_i + \boldsymbol{L}_i\boldsymbol{F}_{ij}, \boldsymbol{D}_i\right), \\
\boldsymbol{F}_{ij}\Big|W_j = w_j, Z_{ij} = 1 &\overset{ind.}{\sim} N_q\left(-c\boldsymbol{\Delta}_i^{-1/2}\lambda_i + w_j\boldsymbol{\Delta}_i^{-1/2}\lambda_i, \boldsymbol{\Delta}_i^{-1}\right), \\
W_j\Big|Z_{ij} = 1 &\overset{ind.}{\sim} HN_1(0, 1).
\end{aligned} \tag{9}$$

Note that the above hierarchical representation can be reformulated as

$$\begin{aligned} \boldsymbol{Y}_j \Big| \tilde{\boldsymbol{F}}_{ij}, Z_{ij} = 1 &\overset{ind.}{\sim} N_p\Big(\boldsymbol{\mu}_i + \tilde{\boldsymbol{L}}_i \tilde{\boldsymbol{F}}_{ij}, \boldsymbol{D}_i\Big), \\ \tilde{\boldsymbol{F}}_{ij} \Big| W_j = w_j, Z_{ij} = 1 &\overset{ind.}{\sim} N_q\big(w_j \boldsymbol{\lambda}_i, \boldsymbol{I}_q\big), \\ W_j \Big| Z_{ij} = 1 &\overset{ind.}{\sim} TN_1(c, 1) I\big(W_j > c\big), \end{aligned} \tag{10}$$

where $c = \sqrt{2/\pi}$, $\tilde{\boldsymbol{L}}_i = \boldsymbol{L}_i \boldsymbol{\Delta}_i^{-1/2}$, $\tilde{\boldsymbol{F}}_{ij} = \boldsymbol{\Delta}_i^{1/2} \boldsymbol{F}_{ij}$ and $TN_1(c, 1) I\big(W_j > c\big)$ denotes the truncated normal distribution with mean c and variance one before truncation on $(c, +\infty)$. So, the complete-data augmented likelihood function of $\boldsymbol{\Theta}$ is given by

$$L(\boldsymbol{\Theta}|C) = \prod_{j=1}^{n} \prod_{i=1}^{\mathrm{g}} Z_{ij} \Big[\pi_i \phi_p\Big(\boldsymbol{y}_j \Big| \boldsymbol{\mu}_i + \tilde{\boldsymbol{L}}_i \tilde{\boldsymbol{F}}_{ij}, \boldsymbol{D}_i\Big) \phi_q\Big(\tilde{\boldsymbol{F}}_{ij} \Big| w_j \boldsymbol{\lambda}_i, \boldsymbol{I}_q\Big) \phi_1\Big(W_j \Big| c, 1\Big) I\big(W_j > c\big) \Big]. \tag{11}$$

3.2 Priors and posteriors

Our Bayesian approach is based on a Gibbs sampler *MCMC* algorithm to draw the samples from the full conditional posteriors. We assign prior distributions to the unknown model parameters and consider independently weak informative proper priors for the elements of $\boldsymbol{\Theta}$. Also, we consider the loading matrix in the form of $\tilde{\boldsymbol{L}}_i = \big[\ell_{i.rt}\big]$ ($\ell_{i.rt}$ are $\boldsymbol{L}_i$ elements). So, for the unknown parameters in the *MSNFA* model, we consider priors given by

$$\boldsymbol{\pi} = \big(\pi_1, \ldots, \pi_{\mathrm{g}-1}\big) \sim Dir\big(\eta_1, \ldots, \eta_{\mathrm{g}}\big), \quad \boldsymbol{\mu}_i \sim N_p\big(\boldsymbol{m}_i, \boldsymbol{M}_i\big), \quad \boldsymbol{\lambda}_i \sim N_q\big(\mathcal{I}_i, \mathcal{G}_i\big),$$

$$\ell_{i.rt} \sim N_1\big(\mu_{\ell i}, \sigma^2_{\ell i}\big); \quad r > t, \ell_{i.rr} \sim HN_1\big(\mu_{\ell i}, \sigma^2_{\ell i}\big), \quad D_{i.r} \sim IG\big(\mathfrak{a}_i, \mathfrak{b}_i\big),$$

for $i = 1, \ldots, \mathrm{g}$, $r = 1, \ldots, p$ and $t = 1, \ldots, q$, where notations *Dir* and *IG* represent the Dirichlet and inverse Gamma distributions, respectively.

The joint posterior distribution $p(\boldsymbol{\Theta}, \boldsymbol{F}, \boldsymbol{w}, \boldsymbol{z}|\boldsymbol{y}) \propto L(\boldsymbol{\Theta}|C) p(\boldsymbol{\Theta})$ is (generally) analytically intractable and *MCMC* methods such as Gibbs sampling (Gelfand and Smith 1990) by using the full conditional posterior distributions are often needed to draw samples from this distribution. The full conditional posteriors for $i = 1, \ldots, \mathrm{g}$, $t = 1, \ldots, p$ and $r = 1, \ldots, q$ are given as follows: (in the following quantities $\boldsymbol{\Theta}_{(-\varepsilon)}$ is the set of parameters without the parameter ε, $\mathfrak{F}_i = \big\{j : z_{ij} = 1\big\}$ and n_i is equal to the number of observations allocated to the i-th *FA* component),

$$\boldsymbol{\pi} | \boldsymbol{\Theta}_{(-\boldsymbol{\pi})}, \boldsymbol{y}, \boldsymbol{F}, \boldsymbol{w}, \boldsymbol{z} \sim Dir\big(\eta_1 + n_1, \ldots, \eta_{\mathrm{g}} + n_{\mathrm{g}}\big).$$

$$\boldsymbol{\mu}_i \Big| \boldsymbol{\Theta}_{(-\boldsymbol{\mu}_i)}, \boldsymbol{y}, \boldsymbol{F}, \boldsymbol{w}, z_{ij} = 1 \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\mu} = \boldsymbol{\Sigma}\Big(\boldsymbol{M}_i^{-1} \boldsymbol{m}_i + \sum_{\mathfrak{F}_i} \boldsymbol{D}_i^{-1} \big(\boldsymbol{y}_j - \tilde{\boldsymbol{L}}_i \tilde{\boldsymbol{F}}_{ij}\big)\Big)$ and $\boldsymbol{\Sigma} = \Big(\boldsymbol{M}_i^{-1} + \sum_{\mathfrak{F}_i} \boldsymbol{D}_i^{-1}\Big)^{-1}$.

$$\ell_{i.rt}\big|\boldsymbol{\Theta}_{(-\ell_{i.rt})}, \boldsymbol{y}, \boldsymbol{F}, \boldsymbol{w}, z_{ij} = 1 \sim N_1\big(\mu, \sigma^2\big),$$

where $\mu = \sigma^2\Big(\mu_{\ell i}\sigma_{\ell i}^{-2} + D_{i.r}^{-1}\sum_{\mathfrak{J}_i} F_{ij(t)}\Big(y_{jr} - \mu_{ir} - \ell_{ir(-t)}^{\top}\tilde{\boldsymbol{F}}_{ij}\Big)\Big)$ and $\sigma^2 = \Big(\sigma_{\ell i}^{-2} + D_{i.r}^{-1}\sum_{\mathfrak{J}_i} F_{ij(t)}^2\Big)^{-1}$, for which y_{jr} and μ_{ir} be the r-th components of $\boldsymbol{y}_j$ and $\boldsymbol{\mu}_i$, respectively, $F_{ij(t)}$ be the t-th components of $\tilde{\boldsymbol{F}}_{ij}$, and ℓ_{ir} be the r-th row of $\tilde{\boldsymbol{L}}_i$ (so $\ell_{i.rt}$ is the t-th element of ℓ_{ir}), and $\ell_{ir(-t)}$ be the r-th row of $\tilde{\boldsymbol{L}}_i$ which t-th component zero.

Also $\ell_{i.rr}\big|\boldsymbol{\Theta}_{(-\ell_{i.rr})}, \boldsymbol{y}, \boldsymbol{F}, \boldsymbol{w}, z_{ij} = 1 \sim N_1\big(\mu, \sigma^2\big)I\big(\ell_{i.rr} > 0\big)$, with the above parameters for which indices t replaced by r.

$$\tilde{\boldsymbol{F}}_{ij}\Big|\boldsymbol{\Theta}, \boldsymbol{y}, \boldsymbol{w}, z_{ij} = 1 \sim N_q(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\mu} = \boldsymbol{\Sigma}\Big(w_j\lambda_i + \tilde{\boldsymbol{L}}_i^{\top}\boldsymbol{D}_i^{-1}\big(\boldsymbol{y}_j - \boldsymbol{\mu}_i\big)\Big)$ and $\boldsymbol{\Sigma} = \Big(\boldsymbol{I}_q + \tilde{\boldsymbol{L}}_i^{\top}\boldsymbol{D}_i^{-1}\tilde{\boldsymbol{L}}_i\Big)^{-1}$.

$$D_{i.r}\big|\boldsymbol{\Theta}_{(-D_{i.r})}, \boldsymbol{y}, \boldsymbol{F}, \boldsymbol{w}, z_{ij} = 1 \sim IG(a, b),$$

where $a = \mathfrak{a}_i + n_i/2$ and $b = \mathfrak{b}_i + \frac{1}{2}\sum_{\mathfrak{J}_i}\big(y_{jr} - \mu_{ir} - \ell_i^{\top}\tilde{\boldsymbol{F}}_{ij}\big)^2$.

$$\boldsymbol{\lambda}_i\big|\boldsymbol{\Theta}_{(-\boldsymbol{\lambda}_i)}, \boldsymbol{y}, \boldsymbol{F}, \boldsymbol{w}, z_{ij} = 1 \sim N_q(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\mu} = \boldsymbol{\Sigma}\Big(\mathcal{G}_i^{-1}\mathcal{I}_i + \sum_{\mathfrak{J}_i} w_j\tilde{\boldsymbol{F}}_{ij}\Big)$ and $\boldsymbol{\Sigma} = \Big(\mathcal{G}_i^{-1} + \Big[\sum_{\mathfrak{J}_i} w_j^2\Big]\boldsymbol{I}_q\Big)^{-1}$.

$$W_j\Big|\boldsymbol{\Theta}, \boldsymbol{y}, \boldsymbol{F}, z_{ij} = 1 \sim TN_1\big(\mu, \sigma^2\big)I\Big(W_j > c\Big),$$

where $\mu = \sigma^2\big(c + \sum_{i=1}^{\mathrm{g}}\lambda_i^{\top}\tilde{\boldsymbol{F}}_{ij}\big)$ and $\sigma^2 = \big(1 + \sum_{i=1}^{\mathrm{g}}\lambda_i^{\top}\lambda_i\big)^{-1}$.

$$\boldsymbol{Z}|\boldsymbol{\Theta}, \boldsymbol{y}, \boldsymbol{F}, \boldsymbol{w} \sim \mathcal{M}\left(1; \frac{\pi_1 f_1\big(\boldsymbol{y}_j|\boldsymbol{\theta}_1\big)}{\sum_{i=1}^{\mathrm{g}}\pi_i f_i\big(\boldsymbol{y}_j|\boldsymbol{\theta}_i\big)}, \ldots, \frac{\pi_{\mathrm{g}} f_{\mathrm{g}}\big(\boldsymbol{y}_j|\boldsymbol{\theta}_{\mathrm{g}}\big)}{\sum_{i=1}^{\mathrm{g}}\pi_i f_i\big(\boldsymbol{y}_j|\boldsymbol{\theta}_i\big)}\right).$$

where $f_i\big(\boldsymbol{y}_j|\boldsymbol{\theta}_i\big); i = 1, \ldots, \mathrm{g}$, are the component's pdf defined in (7).

3.3 Imputation of missing values

An advantage of the hierarchical representation in (9) or (10) is the ability to easily simulate from the model or sample from the parameters using existing Bayesian software such as Stan (Stan Development Team 2017) or NIMBLE (NIMBLE Development Team 2017) (JAGS or OpenBUGS were not able to be used due to the absence of functions to enable matrix inversion). One advantage of this is the ability to easily accommodate missing data and impute values from the model naturally as part of the parameter updates.

Let $\boldsymbol{Y}_M$ and $\boldsymbol{Y}_O$ represent the missing and observed responses, respectively. Missing data imputation in a Bayesian framework relies on the posterior predictive

distribution for the missing data,$P\left(\boldsymbol{Y}_M \middle| \boldsymbol{Y}_O\right) = \int P(\boldsymbol{Y}_M|\boldsymbol{Y}_O, \boldsymbol{\Theta})P(\boldsymbol{\Theta}|\boldsymbol{Y}_O)d\boldsymbol{\Theta}$. As for most missing data problems with an unknown missing pattern, the posterior predictive distribution cannot be simulated directly and a Gibbs sampling algorithm is often used with parameters updated in two generic steps, $\boldsymbol{y}_{i,M}^{(t+1)} \sim P\left(\boldsymbol{y}_{i,M} \middle| \boldsymbol{y}_O, \boldsymbol{\Theta}^{(t)}\right)$ for $i = 1, \ldots, N$ and $\boldsymbol{\Theta}^{(t+1)} \sim P\left(\boldsymbol{\Theta}^{(t)} \middle| \boldsymbol{y}_O, \boldsymbol{y}_{i,M}\right)$. Starting from reasonable initial values $\boldsymbol{y}_{i,M}^{(0)}$ and $\boldsymbol{\Theta}^{(0)}$ and running the algorithm for a large number of iterations provides convergence towards these limiting distributions. In this paper, we implement this approach in NIMBLE as it can be undertaken relatively easily (using only one extra line of code) and extended (if needed) to include situations where missingness may depend on other covariates (e.g. conditions relating to the experiment or particular characteristics of individuals in a survey).

4 Applications

In this section, we assess the performance and flexibility of the proposed *MSNFA* model using real data examples which display signs of skewness and are challenging to fit using the MFA model.

4.1 Priors and computation

For estimation of the different models, largely non-informative prior distributions were used for each of the component parameters: $\boldsymbol{\mu}_i \sim N_p\left(\boldsymbol{m}_i, \boldsymbol{M}_i\right)$, where $\boldsymbol{m}_i = 0$ and $\boldsymbol{M}_i = 10^3\boldsymbol{I}_p$ priors of its columns as $\boldsymbol{\lambda}_i \sim N_q\left(\mathcal{I}_i, \mathcal{G}_i\right)$, which $\mathcal{I}_i = 0$ and $\mathcal{G}_i = 10^3\boldsymbol{I}_q$, $\ell_{i.rt} \sim N_1(0, 100); r > t$, $\ell_{i.rr} \sim HN_1(0, 100)$, $D_{i.r} \sim IG(1, 1)$ for $i = 1, 2$, and $\boldsymbol{\pi} \sim Dir(1, \ldots, 1)$. All computations are implemented on the R software version 3.3.1 (R Core Team 2017) with a core i7 760 processor 2.8 GHz. Gibbs sampling runs of 50,000 iterations with burn-in of 10,000 was used and convergence criteria was established using the Gelman–Rubin statistic (Gelman and Rubin 1992) and by visual inspection. Computations were also verified and models developed using NIMBLE. 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 re-label parameters in proximity to this region (Celeux et al. 2000). A sample copy of the R and NIMBLE code used are available from the authors upon request *(and will be available on a public website shortly).* To avoid some computational issues common to factor analysis (e.g. underflow errors) we scale the datasets examined using the scale function in R. Finally note that a lot of different solutions have been proposed to boost MCMC (see, among others, Meng and Van Dyk 1999; van Dyk and Meng 2001; Yu and Meng 2011; Van Dyk 2010).

Model performance was assessed by comparing the classification accuracy and model selection criteria for *MSNFA* and *MFA* (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).

4.2 Seeds example

In the first example, we examine a clustering problem for a seeds dataset analyzed by Lin et al. (2016) and originally analyzed by Charytanowicz et al. (2010). The data consists of seven geometric features (area, perimeter, compactness, length of kernel, asymmetry coefficient, and length of kernel groove) measured from the X-ray images of 210 wheat kernels, belonging to three different wheat varieties (Kama, Rosa and Canadian). To illustrate the performance of *MSNFA* family we will focus on the case where g is *a priori* known to be 3 with q varying from 2 to 4.

For the *MFA* model (see Table 1) the best classification results were obtained for $q = 4$ with an *ARI* estimate of 0.69, however, all of the model selection criteria appeared to clearly favor $q = 3$ with a slightly lower ARI estimate of 0.66. The best classification results for the *MSNFA* model appeared to be for the $q = 3$ case and with a higher ARI estimate of 0.76. In terms of model selection criteria, estimates for all of the criteria for the MSNFA also clearly favored this particular model. Overall, the *MSNFA* model appears to better fit the three groups in this data quite well compared to the *MFA* case with significant improvements in model choice criteria estimates and classification results.

4.3 AIS data

The second example considers the Australian Institute of Sport (AIS) data containing a number of physical and hematological measurements ($p = 11$) from 100 female and 102 male athletes (n = 202). As a number of variables in the dataset (e.g.

Table 1 Results for seeds data example

Model	q	Log-likelihood (max)	m	EAIC	EBIC	DIC_2	ARI
MFA	2	−730.9	83	1739.7	2017.5	1685.6	0.46
	3	**−572.6**	**98**	**1495.4**	**1823.4**	**1453.6**	**0.66**
	4	−829.3	110	1957.1	2325.3	1815.5	0.69
MSNFA	2	−819.0	89	1891.4	2189.2	1788.6	0.29
	3	**−543.3**	107	**1384.0**	**1742.2**	**1253.5**	**0.76**
	4	−861.3	122	2033.4	2441.8	1856.3	0.63

The best values are indicated in bold

MFA and MSNFA denote the normal and restricted skew-normal factor analysis models respectively

BMI) display signs of moderate skewness, a number of previous studies have used this dataset to examine the performance of skew-normal and skew-t mixture models to correctly classify the male and female athletes into their respective groups (e.g. Murray et al. 2014; Lee and McLachlan 2013a). Similarly, we are interested in assessing the performance of the MSNFA to correctly classify male and female athletes using all of the variables available (most of the previous studies have used only two variables).

From Table 2 we can see quite clearly that the classification performance of the *MSNFA* is very good with an *ARI* of 0.96 and model choice criteria all appear to favor this model. By contrast the *MFA* model is not able to accommodate the skewness in the data and the best *ARI* was 0.85 for the $q = 5$ model.

To illustrate one of the benefits of using a Bayesian approach, we conduct an experiment on the AIS data by assessing the performance of the classification and associated errors in a missing data context. As mentioned previously, the hierarchical structure of the MNSFA allows the model to be coded and computations performed in NIMBLE (or Stan) which relatively easily facilitates the imputation of missing values from the full model (i.e. conditional means). In this experiment, we randomly delete values in the dataset under two different degrees of missingness [5% (low) and 30% (high) of the total sample ($n \times p$)] and compare the performance of imputing values using the model (conditional approach) or according to mean imputation (unconditional approach) where the missing values are replaced by their unconditional means (mean of complete values for the variable). This type of missingness is often described as missing at random (MAR) (See Little and Rubin 1987). Along with the model selection and performance measures outlined previously, we also assess the results using the mean squared errors (MSE),

$$MSE = \frac{1}{n}\sum_{j=1}^{n}\left(\boldsymbol{y}_j^m - \hat{\boldsymbol{y}}_j^m\right)^{\top}\left(\boldsymbol{y}_j^m - \hat{\boldsymbol{y}}_j^m\right),$$

Table 2 Results for AIS data

Model	q	Log-likelihood (max)	EAIC	EBIC	DIC_2	ARI
MFA	2	−2124.0	4495.5	4783.4	4395.0	0.94
	3	−1795.2	3877.3	4224.7	3744.2	0.92
	4	−1624.9	3567.8	3968.1	3401.7	0.85
	5	**−1580.5**	**3510.2**	**3956.8**	**3319.4**	**0.85**
	6	−1591.3	3578.3	4064.6	3386.1	0.85
MSNFA	2	−2126.3	4499.4	4800.5	4382.2	0.94
	3	−1776.4	3859.2	4226.4	3721.6	0.90
	4	−1578.5	3460.3	3820.9	3327.7	0.88
	5	**−1548.9**	**3439.8**	**3853.3**	**3281.5**	**0.96**
	6	−1709.3	4208.0	4667.9	4441.5	0.94

The best values are indicated in bold

MFA and MSNFA denote the normal and restricted skew-normal factor analysis models respectively

Table 3 Results for AIS data (missing data)

Model	Missing (%)	Loglike (max)	EAIC	EBIC	DIC_2	ARI	MSE
MSNFA-UC	5	− 1935.0 (28.1)	4253.1 (55.04)	4733.1 (55.05)	4056.8 (56.8)	0.68 (0.22)	0.47 (0.07)
	30	− 1767.3 (27.9)	3885.8 (55.6)	4365.5 (55.6)	3656.9 (57.0)	0.24 (0.22)	0.69 (0.06)
MSNFA-CO	5	**− 1568.2 (33.4)**	**3520.9 (69.4)**	**4000.7 (69.4)**	**3325.6 (80.0)**	**0.83 (0.08)**	**0.23 (0.10)**
	30	**− 1348.6 (34.2)**	**3055.5 (70.5)**	**3535.2 (70.4)**	**2833.9 (74.0)**	**0.79 (0.09)**	**0.32 (0.04)**

The best values are indicated in bold

Presented are the mean estimates with standard deviation in brackets. The annotations UC and CO denote the results for the unconditional (mean imputation) and conditional (full model) models respectively

where $\hat{\boldsymbol{y}}$ denotes the imputed value and $n^* = \sum_{j=1}^{n} (p - p_j^o)$ is the number of total missing values. A smaller value of MSE indicates a more accurate prediction of missing values.

Table 3 presents the results of the two models (unconditional and conditional) against the mean values of the model selection criteria (EIC, EBIC, etc.), classification performance (ARI) and the MSE over 30 replications of the dataset under each missingness rate scenario (5% or 30%).

Under both degrees of missingness, the results for the conditional model (MSNFA-CO) are clearly superior to the results for the unconditional model (MSNFA-UC) with only a relatively small decrease in performance. In contrast, the results for the unconditional model have quickly deteriorated with an average classification result for the ARI of 0.68 (compared to 0.83 for the conditional model). The extent and type of deterioration in performance obviously depend on the setting but in this setting we saw substantial deterioration for a relatively small degree of missingness (5%). An alternative to the unconditional approach includes listwise deletion where an entire record is deleted from the analysis if a single value is missing. This approach is only really applicable for large samples, which is rarely the case for most applications where factor analysis is commonly used. Thus, the conditional approach (using the full model) is often preferred and used but relies upon the ease of use and availability of the computational approach in practice.

5 Conclusion

We have outlined and assessed the performance of a *MSNFA* model within a Bayesian framework. Various properties of the *SNFA* family are well defined and estimation of the parameters is relatively straightforward in a Bayesian framework with all of the Gibbs sampling updates available in closed form. Assessments of the

performance of the proposed model on simulated and real data suggest that this distribution provides a considerable degree of flexibility in modeling data of varying directional shape. Various extensions to the *MSNFA* model are possible, including the use of this distribution in the more general setting of a structural equation model and extending existing models where sparse covariance structures are necessary for particular settings/applications. Similar to the work by Suarez and Ghosal (2016) more informative priors (known apriori or empirically derived) could be placed on the variance of the error term [diagonal matrix $\boldsymbol{D}$ in (4b)] in noisy or error prone settings to improve estimates. Such an extension is relatively easy to implement using the computational approach outlined. Further extensions relating to the incorporation of covariates, either as part of the missing data process or separately, also follow in a relatively straightforward way from the proposed model and software available (e.g. NIMBLE). Further extensions can also be made to incorporate unrestricted skew distributional forms (Maleki et al. 2018b) and asymmetric two-piece distributions belonging to the mixture distributions introduced by Maleki and Mahmoudi (2017) and Hoseinzadeh et al. (2018).

Acknowledgements The authors would like to thank the associated editor and anonymous reviewers for their suggestions, corrections and encouragement, which helped us to improve earlier versions of the manuscript. We also would like to acknowledge helpful discussions with Geoff McLachlan and Sharon Lee (UQ) in the preparation of this work.

References

Ando T (2009) Bayesian factor analysis with fat-tailed factors and its exact marginal likelihood. J Multivar Anal 100(8):1717–1726

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

Azzalini A (1985) A class of distributions which includes the normal ones. Scand J Stat 12:171–178

Azzalini A (2014) The skew-normal and related families. Institute of Mathematical Statistics Monographs, Cambridge University Press, Cambridge

Azzalini A, Capitanio A (1999) Statistical applications of the multivariate skew-normal distribution. J R Stat Soc B 61:579–602

Azzalini A, Dalla-Vale A (1996) The multivariate skew-normal distribution. Biometrika 83:715–726

Basso RM, Lachos VH, Cabral CRB, Ghosh P (2010) Robust mixture modeling based on the scale mixtures of skew-normal distributions. Comput Stat Data Anal 54:2926–2941

Bhattacharya A, Dunson DB (2011) Sparse Bayesian infinite factor models. Biometrika 98(2):291–306

Bishop CM (1999) Bayesian PCA. In: Kearns MS, Solla SA, Cohn DA (eds) Advances in neural information processing systems, vol 11. MIT Press, Cambridge, pp 382–388

Carlin BP, Louis TA (2011) Bayesian methods for data analysis, 3rd edn. Chapman & Hall, CRC Press, Boca Raton

Carvalho CM, Chang J, Lucas JE, Nevins JR, Wang Q, West M (2008) High-dimensional sparse factor modeling: applications in gene expression genomics. J Am Stat Assoc 103(484):1438–1456

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

Celeux G, Forbes F, Robert CP, Titterington DM (2006) Deviance information criteria for missing data models. Bayesian Anal 1:651–674

Charytanowicz M, Niewcazs J, Kulczycki P, Lukasik S, Zak S (2010) A complete gradient clustering algorithm for features analysis of x-ray images. In: Pietka E, Kawa J (eds) Information technologies in biomedicine. Springer, Berlin, pp 15–24

Chen M, Silva J, Paisley J, Wang C, Dunson D, Carin L (2010) Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: algorithm and performance bounds. IEEE Trans Signal Process 58(12):6140–6155
Chen M, Zaas A, Woods C, Ginsburg GS, Lucas J, Dunson D, Carin L (2011) Predicting viral infection from high-dimensional biomarker trajectories. J Am Stat Assoc 106:1259–1279
Conti G, Frühwirth-Schnatter S, Heckman JJ, Piatek R (2014) Bayesian exploratory factor analysis. J Econom 183(1):31–57
Fokoué E, Titterington DM (2003) Mixtures of factor analyzers. Bayesian estimation and inference by stochastic simulation. Mach Learn 50:73–94
Frühwirth-Schnatter S, Lopes HF (2012) Parsimonious Bayesian factor analysis when the number of factors is unknown. Unpublished Technical Report
Gelfand AE, Smith AFM (1990) Sampling based approaches to calculating marginal densities. J Am Stat Assoc 85:398–409
Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences (with discussion). Stat Sci 7:457–511
Ghahramani Z, Beal MJ (2000) Variational inference for Bayesian mixtures of factor analysers. Adv Neural Inf Process Syst 12:449–455
Ghahramani Z, Hinton GE (1997) The EM algorithm for mixtures of factor analyzers. Technical Report No. CRG-TR-96-1. University of Toronto, Department of Computer Science, Toronto
Ghosh J, Dunson DB (2009) Default prior distributions and efficient posterior computation in Bayesian factor analysis. J Comput Graph Stat 18(2):306–320
Hinton GE, Dayan P, Revow M (1997) Modeling the manifolds of images of handwritten digits. IEEE Trans Neural Netw 8:65–74
Hoseinzadeh A, Maleki M, Khodadadi Z, Contreras-Reyes JE (2018) The Skew-Reflected-Gompertz distribution for analyzing the symmetric and asymmetric data. J Comput Appl Math 349:132–141
Hubert L, Arabie P (1985) Comparing partitions. J Classif 2:193–218
Knowles D, Ghahramani Z (2007) Infinite sparse factor analysis and infinite independent components analysis. In: 7th international conference on independent component analysis and signal separation. Springer, Berlin, pp 381–388
Lee SX, McLachlan GJ (2013a) Model-based clustering and classification with non-normal mixture distributions. Stat Methods Appl 22(4):427–454
Lee SX, McLachlan GJ (2013b) On mixtures of skew normal and skew t distributions. Adv Data Anal Classif 7(3):241–266
Lee SY, Xia YM (2008a) A robust Bayesian approach for structural equation models with missing data. Psychometrika 73:343–364
Lee SY, Xia YM (2008b) Semiparametric Bayesian analysis of structural equation models with fixed covariates. Stat Med 27:2341–2360
Leung D, Drton M (2016) Order-invariant prior specification in Bayesian factor analysis. Stat Probab Lett 111:60–66
Lin TI, Lee JC, Yen SY (2007) Finite mixture modeling using the skew-normal distribution. Stat Sin 17:909–927
Lin TI, McLachlan GJ, Lee SX (2016) Extending mixtures of factor models using the restricted multivariate skew-normal distribution. J Multivar Anal 143:398–413
Little RJA, Rubin DB (1987) Statistical analysis with missing data. Wiley, New York
Lopes HF, West M (2004) Bayesian model assessment in factor analysis. Stat Sin 4:41–67
Maleki M, Arellano-Valle RB (2017) 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
Maleki M, Mahmoudi MR (2017) Two-pieces location-scale distributions based on scale mixtures of normal family. Commun Stat Theory Methods 46(24):12356–12369
Maleki M, Wraith D, Arellano-Valle RB (2018a) Robust finite mixture modeling of multivariate unrestricted skew-normal generalized hyperbolic distributions. Stat Comput. https://doi.org/10.1007/s11222-018-9815-5
Maleki M, Wraith D, Arellano-Valle RB (2018b) A flexible class of parametric distributions for Bayesian linear mixed models. Test. https://doi.org/10.1007/s11749-018-0590-6
McLachlan GJ, Peel D (2000) Finite mixture models. Wiley, New York
Meng XL, Van Dyk DA (1999) Seeking efficient data augmentation schemes via conditional and marginal augmentation. Biometrika 86:301–320

Mengersen K, Robert C, Titterington DM (2011) Mixtures: estimation and applications. Wiley, Chichester
Murray PM, Dunson DB, Carin L, Lucas JE (2013) Bayesian Gaussian copula factor models for mixed data. J Am Stat Assoc 108(502):656–665
Murray PM, Browne RP, McNicholas PD (2014) Mixtures of skew-t factor analyzers. Comput Stat Data Anal 77:326–335
NIMBLE Development Team (2017) NIMBLE: an R package for programming with BUGS models, Version 0.6-10. http://r-nimble.org. Accessed 19 Feb 2018
Paisley J, Carin L (2009) Nonparametric factor analysis with beta process priors. In: Proceedings of the 26th annual international conference on machine learning, pp 777–784
R Core Team (2017) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/. Accessed 19 Feb 2018
Sahu SK, Dey DK, Branco MD (2003) A new class of multivariate skew distributions with applications to Bayesian regression models. Can J Stat 31(2):129–150
Song XY, Pan JH, Kwok T, Vandenput L, Ohlsson C, Leung PC (2010) A semiparametric Bayesian approach for structural equation models. Biom J 52(3):314–332
Stan Development Team (2017) The stan core library, version 2.17.0. http://mc-stan.org. Accessed 19 Feb 2018
Suarez AJ, Ghosal S (2016) Bayesian estimation of principal components for functional data. Bayesian Anal 12:1–23
Ustugi A, Kumagai T (2001) Bayesian analysis of mixtures of factor analyzers. Neural Comput 13(5):993–1002
Van Dyk DA (2010) Marginal Markov chain Monte Carlo methods. Stat Sin 20:1423–1454
Van Dyk DA, Meng XL (2001) The art of data augmentation. J Comput Graph Stat 10:1–50
Wall MM, Guo J, Amemiya Y (2012) Mixture factor analysis for approximating a non-normally distributed continuous latent factor with continuous and dichotomous observed variables. Multivar Behav Res 47:276–313
Yang M, Dunson DB (2010) Bayesian semiparametric structural equation models with latent variables. Psychometrika 75(4):675–693
Yu Y, Meng XL (2011) To center or not to center: that is not the question an ancillarity sufficiency interweaving strategy (ASIS) for boosting MCMC efficiency. J Comput Graph Stat 20:531–570