Bayesian finite mixture modeling based on scale mixtures of univariate and multivariate skew-normal distributions

Marcus G. Lavagnole Nascimento†, Carlos A. Abanto-Valle† and Victor H. Lachos⋆

† Department of Statistics, Federal University of Rio de Janeiro, Caixa Postal 68530, CEP: 21945-970, Rio de Janeiro, Brazil

⋆ Department of Statistics, University of Connecticut, U-4120, Storrs, CT 06269, USA

Abstract

In this article, finite mixtures of scale mixtures of skew-normal (FM-SMSN) distributions are introduced with the aim of dealing simultaneously with asymmetric behavior and heterogeneity observed in some data sets. A Bayesian methodology based on the data augmentation principle is derived and an efficient Markov-chain Monte Carlo (MCMC) algorithm is developed. These procedures are discussed with emphasis on finite mixtures of skew-normal, skew-t and skew-slash distributions for both univariate as well as multivariate case. Univariate and bivariate data sets using FM-SMSN distributions are analyzed. According to the results FM-SMSN distributions support both data sets.

Keywords: Bayesian inference, finite mixture, scale mixture of normal distributions, Markov chain Monte Carlo.

1 Introduction

Finite mixtures of distributions have been vastly applied in different scientific fields, such as genetics (Fu et al., 2011), neural networks (Zhang et al., 2013), image processing (Bouguila et al., 2004) and social sciences (da Paz et al., 2017), just to mention a few. This class of models comprise a powerful tool to deal with heterogeneous data in which observations belong to distinct populations and to approximate complex probability densities.

The class of scale mixtures of skew-normal distributions (Branco and Dey, 2001, SMSN) deals efficiently with skewness and heavy-tails. The SMSN family encompasses the entire symmetric class of scale mixtures of normal distributions (Andrews and Mallows, 1974). Additionally, it contains as proper elements the skew-normal (Azzalini, 1985, SN), skew-t (Azzalini and Capitanio, 2003, ST) and skew-slash (Wang and Genton, 2006, SSL), among others. These distributions are characterized by heavier tails than the skew-normal, providing a reasonable alternative for robust inference in presence of skewness.

As an attempt to model properly data sets arising from a class or several classes with asymmetric observations, Lin, Lee and Yen (2007) and Lin, Lee and Hsieh (2007) proposed a methodology for maximum likelihood estimation of finite mixtures of skew-normal and skew-t distributions based on EM-type algorithm (Dempster et al., 1977). Basso et al. (2010) extended these ideas to finite mixture models based on SMSN distributions.

Frühwirth-Schnatter and Pyne (2010) introduced a Bayesian methodology for finite mixtures of univariate as well multivariate skew-normal and skew-t distributions. Based on the data augmentation principle, the authors developed a Markov chain Monte Carlo (MCMC) algorithm in order to perform the sampling from the joint posterior distribution. Furthermore, they expressed the resulting model using a stochastic representation in terms of a random-effects model(Azzalini, 1986; Henze, 1986) and the standard hierarchical representation of a finite mixture model introduced by Diebolt and Robert (1994).

In this paper, a Bayesian approach for finite mixture of scale mixtures of univariate and multivariate skew-normal models is proposed. To that end, the class of models presented by Basso et al. (2010) and a multivariate extension are described from a Bayesian perspective and the methods considered by Frühwirth-Schnatter and Pyne (2010) are employed.

The remainder of the paper is organized as follows. Section 2 introduces the scale mixtures of skew-normal distributions. Section 3 is related to modeling finite mixtures of scale mixtures of skew-normal (FM-SMSN) distributions from a Bayesian viewpoint. Section 4 is devoted to the application and model comparison among particular members of the FM–SMSN using univariate and bivariate data sets. Finally, some concluding remarks and suggestions for future developments are given in Section 5.

2 Scale mixtures of skew-normal distributions

In this section the class of scale mixtures of skew-normal distributions (Branco and Dey, 2001) is introduced. First, the univariate skew-normal distribution is described and the multivariate version is derived. Then, based on the stochastic representation, the SMSN distributions are presented, in particular, the skew-t and the skew-slash. Finally, a reparameterization is defined.

2.1 The univariate and multivariate Skew-Normal Distribution

As defined by Azzalini (1985), a univariate random variable Z follows a skew-normal distribution, $Z \sim SN(\mu, \sigma^2, \lambda)$, if its probability density function is given by

$$f_Z(z) = \frac{2}{\sigma}\phi\left(\frac{z-\mu}{\sigma}\right)\Phi\left(\lambda\left(\frac{z-\mu}{\sigma}\right)\right),\ z \in \Re, \tag{1}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are, respectively, the probability density function (pdf) and the cumulative distribution function (cdf) of the univariate standard normal distribution. Furthermore, $(\mu, \sigma^2, \lambda) \in \Re \times \Re_+ \times \Re$ are the location, scale and skewness parameters respectively.

Lemma 1 *A random variable $Z \sim SN(\mu, \sigma^2, \lambda)$ has a stochastic representation given by*

$$Z = \mu + \sigma\delta W + \sigma\sqrt{1-\delta^2}\varepsilon, \tag{2}$$

where $W \sim TN_{[0,\infty)}(0, 1)$ and $\varepsilon \sim N(0, 1)$ are independent and $\delta = \lambda/(\sqrt{1+\lambda^2})$. $TN_A(\cdot,\cdot)$ and $N(\cdot,\cdot)$ denote the truncated normal on set A and the normal distributions respectively.

Azzalini and Dalla Valle (1996) proposed a multivariate version of the skew-normal distribution as a generalization of the stochastic representation stated in equation (2). Let $\mathbf{Y} = (Y_1, \ldots, Y_p)^{'} \in \Re^p$ such that $Y_j = \delta_j W + \sqrt{1-\delta^2}\varepsilon_j$, $j = 1, \ldots, p$, where $W \sim NT_{[0,+\infty)}(0,1)$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_p)^{'} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})$ are independent and $\delta_j \in (-1,1)$. $N_p(\cdot,\cdot)$ denotes the multivariate normal distribution. Hence, the transformation $\mathbf{Z} = \boldsymbol{\mu} + \boldsymbol{\sigma}\mathbf{Y}$ with location parameter $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^{'} \in \Re^p$ and diagonal scale matrix $\boldsymbol{\sigma}$ =Diag$(\sigma_1, \ldots, \sigma_p)$, $\sigma_j > 0$, is immediately associated to the follow stochastic representation

$$Z_j = \mu_j + \sigma_j \delta_j W + \sigma_j \sqrt{1-\delta_j^2}\varepsilon_j. \tag{3}$$

The resulting distribution is denominated the basic multivariate skew normal distribution, $\mathbf{Z} \sim SN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$, with density

$$f_{\mathbf{Z}}(\mathbf{z}) = 2\phi_p(\mathbf{z}-\boldsymbol{\mu}; \boldsymbol{\Sigma})\Phi(\boldsymbol{\lambda}^{'}\boldsymbol{\sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu})), \tag{4}$$

where $\phi_p(\cdot)$ is the zero mean multivariate normal distribution probability density function. It is possible to relate $\boldsymbol{\lambda}$ and $\boldsymbol{\Sigma}$ to the parameters $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_p)^{'}$, $\boldsymbol{\sigma}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ in the stochastic representation (3) through

$$\boldsymbol{\Sigma} = \boldsymbol{\sigma}\overline{\boldsymbol{\Sigma}}\boldsymbol{\sigma}, \;\; \boldsymbol{\lambda} = \frac{1}{\sqrt{1-\boldsymbol{\delta}^{'}\boldsymbol{\delta}}}\overline{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\delta}, \tag{5}$$

where $\overline{\boldsymbol{\Sigma}} = \boldsymbol{\Delta}\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\boldsymbol{\Delta} + \boldsymbol{\delta}\boldsymbol{\delta}^{'}$ and $\boldsymbol{\Delta}$ =Diag$(\sqrt{1-\delta_1^2}, \ldots, \sqrt{1-\delta_p^2})$. Moreover, $\overline{\boldsymbol{\Sigma}}_{jj} = (1-\delta_j^2)(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})_{jj} + \delta_j^2 = 1$, thus $\boldsymbol{\Sigma}_{jj} = \omega_j^2$, consequently, $\overline{\boldsymbol{\Sigma}}$ is a correlation matrix.

Considering the parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$, the parameters $(\boldsymbol{\delta}, \boldsymbol{\sigma}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})$ in the stochastic representation (3) are obtained from

$$\boldsymbol{\delta} = \frac{1}{1+\boldsymbol{\lambda}^{'}\overline{\boldsymbol{\Sigma}}\boldsymbol{\lambda}}\overline{\boldsymbol{\Sigma}}\boldsymbol{\lambda}, \;\; \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} = \boldsymbol{\Delta}^{-1}\overline{\boldsymbol{\Sigma}}\boldsymbol{\Delta}^{-1} - \tilde{\boldsymbol{\lambda}}\tilde{\boldsymbol{\lambda}}^{'}, \tag{6}$$

where $\overline{\boldsymbol{\Sigma}} = \boldsymbol{\sigma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\sigma}^{-1}$, $\boldsymbol{\sigma}$ =Diag$(\boldsymbol{\Sigma})^{1/2}$ a diagonal matrix obtained from the diagonal elements of $\boldsymbol{\Sigma}$, $\tilde{\boldsymbol{\lambda}} = (\tilde{\lambda}_1, \ldots, \tilde{\lambda}_1)$ in which $\tilde{\lambda}_j = \delta_j/\sqrt{1-\delta_j}$ and $\boldsymbol{\Delta}$ as previously defined.

2.2 Scale mixtures of univariate and multivariate skew-normal distributions

Let Z be a random variable such that $Z \sim SN(0, \sigma^2, \lambda)$. A random variable X is in the scale mixtures of skew-normal family, $X \sim SMSN(\mu, \sigma^2, \lambda, H)$, if it can be written as

$$X = \mu + k^{1/2}(U)Z, \tag{7}$$

where μ, $k(\cdot)$ and U are, respectively, a location parameter, a positive weight function and a random variable with cumulative distribution function $H(\cdot; \boldsymbol{\nu})$ and probability density function $h(\cdot; \boldsymbol{\nu})$ where $\boldsymbol{\nu}$ is a scalar or a vector parameter indexing the distribution of U.

Lemma 2 *A random variable $X \sim SMSN(\mu, \sigma^2, \lambda, H)$ has a stochastic representation given by*

$$X = \mu + \sigma\delta k^{1/2}(U)W + k^{1/2}(U)\sigma\sqrt{1-\delta^2}\varepsilon, \tag{8}$$

where $W \sim TN_{[0,+\infty)}(0, 1)$ and $\varepsilon \sim N(0, 1)$ are independent and $\delta = \lambda/(\sqrt{1+\lambda^2})$.

A random variable $\mathbf{X}$ belongs to the scale mixtures of multivariate skew-normal family, $\mathbf{X} \sim SMSN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, H)$, if it can be written as

$$\mathbf{X} = \boldsymbol{\mu} + k^{1/2}(U)\mathbf{Z}, \tag{9}$$

where $\mathbf{Z} \sim SN_p(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$.

Along this work, we restrict our attention to the case in that $k(U) = U^{-1}$. As mentioned above, the SMSN family constitutes a class of asymmetric thick-tailed distributions including the skew normal, the skew-t and the skew-slash distributions, which are obtained by choosing the mixing variables as: $U = 1$, $U \sim G(\frac{\nu}{2}, \frac{\nu}{2})$ and $U \sim Be(\nu, 1)$, where $G(\cdot, \cdot)$ and $Be(\cdot, \cdot)$ indicate the gamma and beta distributions respectively.

2.3 Reparameterization

Following Frühwirth-Schnatter and Pyne (2010), a parameterization in terms of $\boldsymbol{\theta}^* = (\mu, \psi, \tau^2, \nu)$ is applied for the scale mixtures of skew-normal distributions, hence the stochastic representation given by the equation (8) is rewritten as

$$X = \mu + \psi k^{1/2}(U)W + k^{1/2}(U)\tau\varepsilon, \tag{10}$$

where $\psi = \sigma\delta$ and $\tau^2 = \sigma^2(1-\delta^2)$. The original parametric vector $\boldsymbol{\theta} = (\mu, \sigma^2, \lambda, \nu)$ is recovered through

$$\lambda = \frac{\psi}{\tau}, \;\; \sigma^2 = \tau^2 + \psi^2. \tag{11}$$

Introducing the new parametric vector $\boldsymbol{\theta}^* = (\boldsymbol{\mu}, \boldsymbol{\psi}, \boldsymbol{\Omega}, \nu)$, it is equally possible to find a similar representation for the scale mixtures of multivariate skew-normal distributions:

$$\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\psi} W + \boldsymbol{\varepsilon}, \tag{12}$$

where $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_p)^{'}$, $\psi_j = \sigma_j\delta_j$, $\boldsymbol{\Omega} = \boldsymbol{\Sigma} - \boldsymbol{\psi}\boldsymbol{\psi}^{'}$, $\boldsymbol{\varepsilon} \sim N_p(\mathbf{0}, k^{1/2}(U)\boldsymbol{\Omega})$ and $W|U = u \sim TN_{[0,+\infty)}(0, k^{1/2}(u))$. The original parametric vector $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu)$ is recovered through

$$\boldsymbol{\Sigma} = \boldsymbol{\Omega} + \boldsymbol{\psi}\boldsymbol{\psi}^{'}, \;\; \boldsymbol{\lambda} = \frac{1}{\sqrt{1 - \boldsymbol{\psi}^{'}\boldsymbol{\Sigma}^{-1}\boldsymbol{\psi}}}\boldsymbol{\sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\psi}, \tag{13}$$

remembering that $\boldsymbol{\sigma}$ =Diag$(\boldsymbol{\Sigma})^{1/2}$ a diagonal matrix obtained from the diagonal elements of $\boldsymbol{\Sigma}$.

3 Finite mixtures of Scale mixture of skew-normal distributions

3.1 The model

Consider a K-component mixture model ($K > 1$) in which a set $\mathbf{x}_1, \ldots, \mathbf{x}_n$ arises from a mixture of SMSN$_p$ distributions, given by

$$f(\mathbf{x}_i|\boldsymbol{\vartheta}, \boldsymbol{\eta}) = \sum_{k=1}^{K} \eta_k g(\mathbf{x}_i|\boldsymbol{\theta}_k^*), \tag{14}$$

where $\eta_k > 0$, $k = 1, \ldots, K$, $\sum_{k=1}^{K} \eta_k = 1$ and $g(\cdot|\boldsymbol{\theta}_k^*)$ denotes the SMSN$_p(\boldsymbol{\theta}_k^*)$ pdf. Here, $\boldsymbol{\vartheta}$ and $\boldsymbol{\eta}$ denote the unknown parameters with $\boldsymbol{\vartheta} = (\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_K^*)$ and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)$. According to the reparameterization introduced on subsection 2.3, $\boldsymbol{\theta}_k^* = (\mu_k, \psi_k, \tau_k^2, \nu_k)$ and $\boldsymbol{\theta}_k^* = (\boldsymbol{\mu}_k, \boldsymbol{\psi}_k, \boldsymbol{\Omega}_k, \nu_k)$ are the parameters for the component k assuming $p = 1$ and $p \geq 2$ respectively.

Introducing the allocation vector $\mathbf{S} = (\mathbf{S}_1, \ldots, \mathbf{S}_n)$, i. e., the vector containing the information about in which group the observation $\mathbf{x}_i$ of the random variable $\mathbf{X}_i$ is. The indicator variable

$\mathbf{S}_i = (S_{i1}, \ldots, S_{iK})^\top$, with

$$S_{ik} = \begin{cases} 1, & \text{if } \mathbf{X}_i \text{ belongs to component } k \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

and $\sum_{k=1}^{K} S_{ik} = 1$. Given the weights vector $\boldsymbol{\eta}$, the latent variables $\mathbf{S}_1, \ldots, \mathbf{S}_n$ are independent with multinomial densities

$$p(\mathbf{S}_i|\boldsymbol{\eta}) = \eta_1^{S_{i1}} \eta_2^{S_{i2}} \ldots (1 - \eta_1 - \cdots - \eta_{K-1})^{S_{iK}}. \tag{16}$$

The joint density of $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ and $\mathbf{S} = (\mathbf{S}_1, \ldots, \mathbf{S}_n)$ is given by

$$f(\mathbf{x}, \mathbf{s}|\boldsymbol{\vartheta}) = \prod_{k=1}^{K} \prod_{i=1}^{n} [\eta_k g(\mathbf{x}_i|\boldsymbol{\theta}_k^*)]^{S_{ik}}. \tag{17}$$

Note that from the stochastic representation and the introduction of the latent variables $\mathbf{W} = (W_1, \ldots, W_n)$ and $\mathbf{U} = (U_1, \ldots, U_n)$ described on subsection 2.2, a p-dimensional random variable $\mathbf{X}_i$ drawn from the scale mixture of skew-normal distributions has a hierarchical representation. Hence, the individual $\mathbf{X}_i$ belonging to the k-th component can be written as

$$\begin{aligned} \mathbf{X}_i|\boldsymbol{\theta}_k^*, w_i, u_i, S_{ik} = 1 \quad &\sim \quad N_p(\boldsymbol{\mu}_k + \boldsymbol{\psi}_k w_i, u_i^{-1}\boldsymbol{\Omega}_k), \\ W_i|u_i, S_{ik} = 1 \quad &\sim \quad TN_{[0,+\infty)}(0, u_i^{-1}), \\ U_i|S_{ik} = 1, \nu_k \quad &\sim \quad h(\cdot; \boldsymbol{\nu}_k). \end{aligned} \tag{18}$$

Thus, the joint density of $\mathbf{X}$ and the latent variables $\mathbf{S}$, $\mathbf{W}$ and $\mathbf{U}$ is

$$f(\mathbf{x}, \mathbf{s}, \mathbf{w}, \mathbf{u}|\boldsymbol{\vartheta}, \boldsymbol{\eta}) = \prod_{k=1}^{K} \left[\prod_{i=1}^{n} [\eta_k f(\mathbf{x}_i|\boldsymbol{\theta}_k^*, w_i, u_i) f(w_i|u_i) f(u_i|\nu_k)]^{S_{ik}} \right] p(\mathbf{s} \mid \boldsymbol{\eta}). \tag{19}$$

3.2 Bayesian Inference

To perform a Bayesian analysis, the first step is to select the priors distributions. In the context of finite mixture models, it is necessary a special attention on these choices since it is not suggested to choose an improper prior because it might implies in an improper posterior density (Frühwirth-Schnatter, 2006). Additionally, as noticed by Jennison (1997), it is recommended to avoid be as "noninformative as possible" by choosing large prior variances because the number of components is highly influenced by the prior choices. For these reasons, as in Frühwirth-Schnatter and Pyne (2010), it was adopted the hierarchical priors introduced by Richardson and Green (1997) for mixtures of normals to reduce sensitivity with respect to choosing the prior variances.

Hence, taking an arbitrary mixture component k, the prior set was specified as: $\boldsymbol{\eta} \sim D(e_0, \ldots, e_0)$, $(\mu_k, \psi_k)|\tau_k^2 \sim N_2(\mathbf{b}_0, \tau_k^2 \mathbf{B}_0)$, $\tau_k^2|C_0 \sim IG(c_0, C_0)$ and $C_0 \sim G(g_0, G_0)$, where e_0, $\mathbf{b}_0 \in \Re^2$, $\mathbf{B}_0 \in \Re^{2\times 2}$, c_0, g_0 and G_0 are known hyper parameters, $D(\cdot, \ldots, \cdot)$ and $IG(\cdot, \cdot)$ indicate the dirichlet and inverse gamma distributions, respectively. Looking for the multivariate case, extensions of the preceding priors were chosen: $\boldsymbol{\eta} \sim D(e_0, \ldots, e_0)$, $(\boldsymbol{\mu}_k, \boldsymbol{\psi}_k)|\boldsymbol{\Omega}_k \sim N_{2\times p}(\mathbf{b}_0, \mathbf{B}_0, \boldsymbol{\Omega}_k)$, $\boldsymbol{\Omega}_k|C_0 \sim IW(c_0, C_0)$, $C_0 = diag(\zeta_1, \ldots, \zeta_p)$, and $\zeta_j \sim G(g_0, G_0)$, $j = 1, \ldots, p$, where e_0, $\mathbf{b}_0 \in \Re^{2\times p}$, $\mathbf{B}_0 \in \Re^{2\times 2}$, c_0, g_0 and G_0 are known hyper parameters, $N_{q\times p}(\cdot, \cdot, \cdot)$ and $IW(\cdot, \cdot)$ indicate the matrix normal and inverse Wishart distributions respectively. Considering the parameter ν priors, $\nu_k \sim G_{(1,\infty)}(\alpha, \gamma)$ and $\nu_k \sim G_{(1,40)}(\alpha, \gamma)$, where α and γ are known hyper parameters and $G_A(\cdot, \cdot)$ denotes the truncated gamma on set A, are specified to the FM-ST and FM-SSL models respectively.

The joint posterior density of parameters and latent unobservable variables can be written as

$$p(\boldsymbol{\vartheta}, \boldsymbol{\eta}, \mathbf{w}, \mathbf{u}, \mathbf{s}|\mathbf{x}) \propto \left\{ \prod_{k=1}^{K} \left[\prod_{i=1}^{n} [\eta_k f(\mathbf{x}_i|\boldsymbol{\theta}_k^*, w_i, u_i) f(w_i|u_i) f(u_i|\nu_k)]^{S_{ik}} \right] p(\boldsymbol{\theta}_k^*) \right\} p(\mathbf{s} \mid \boldsymbol{\eta}) p(\boldsymbol{\eta}), \quad (20)$$

where $p(\boldsymbol{\theta}_k^*) = p(\boldsymbol{\mu}_k, \boldsymbol{\psi}_k|\boldsymbol{\Omega}_k)p(\boldsymbol{\Omega}_k|C_0)p(C_0)p(\nu_k)$. As expressed in Tanner and Wong (1987), in light of the data augmentation principle, conditional on the allocation vector $\mathbf{S}$, the parameter estimation may be executed independently for each parametric component $\boldsymbol{\theta}_k^*$ and for the weights distribution $\boldsymbol{\eta}$, as a consequence, the full conditionals of the parameters and the latent unobservable variables for the finite mixture of SMSN model are written as follows:

$$
\begin{aligned}
p(\boldsymbol{\eta}|\mathbf{s}) &\propto p(\mathbf{s}|\boldsymbol{\eta})p(\boldsymbol{\eta}) & (21)\\
p(w_i|S_{ik}=1,\cdots) &\propto \left[f(\mathbf{x}_i|\boldsymbol{\theta}_k^*, w_i, u_i)f(w_i|u_i)\right]^{S_{ik}}, & (22)\\
p(u_i|S_{ik}=1,\cdots) &\propto \left[f(\mathbf{x}_i|\boldsymbol{\theta}_k^*, w_i, u_i)f(w_i|u_i)f(u_i|\nu_k)\right]^{S_{ik}}, & (23)\\
p(\boldsymbol{\mu}_k, \boldsymbol{\psi}_k|\cdots) &\propto \prod_{\{i:S_{ik}=1\}} f(\mathbf{x}_i|\boldsymbol{\theta}_k^*, w_i, u_i)p(\boldsymbol{\mu}_k, \boldsymbol{\psi}_k|\boldsymbol{\Omega}_k), & (24)\\
p(\boldsymbol{\Omega}_k|\cdots) &\propto \prod_{\{i:S_{ik}=1\}} f(\mathbf{x}_i|\boldsymbol{\theta}_k^*, w_i, u_i)p(\boldsymbol{\Omega}_k|C_0), & (25)\\
p(C_0|\cdots) &\propto \prod_{k=1}^{K} p(\boldsymbol{\Omega}_k|C_0)p(C_0), & (26)\\
p(\nu_k|\cdots) &\propto \prod_{\{i:S_{ik}=1\}} f(u_i|\nu_k)p(\nu_k). & (27)
\end{aligned}
$$

Additional details about the full conditionals are available in Appendix A and B, respectively.

In furtherance of to make Bayesian analysis feasible for parameter estimation in the FM-SMSN class of models, random samples from the posterior distributions of $(\mathbf{s}, \boldsymbol{\vartheta}, \boldsymbol{\eta}, \mathbf{w}, \mathbf{u})$ given $\mathbf{x}$ are drawn through Monte Chain Monte Carlo simulation methods. Algorithm 1 describes the sampling scheme from the full conditionals distributions of the parameters and the latent unobservable variables.

Algorithm 1 *MCMC for finite mixture of scale mixtures of skew-normal.*

1 Set $t = 1$ *and get starting values for* $\mathbf{S}^{(0)}$, $(\boldsymbol{\theta}_1^{*(0)}, \ldots, \boldsymbol{\theta}_K^{*(0)})$, $\boldsymbol{\eta}^{(0)}$, $\mathbf{w}^{(0)}$ *and* $\mathbf{u}^{(0)}$*;*

2 Parameter simulation conditional on the classification $\mathbf{S}^{(t-1)}$*:*

2.1 Sample $\boldsymbol{\eta}^{(t)}$ *from* $p(\boldsymbol{\eta}|\mathbf{s}^{(t-1)})$*;*

2.2 Sample the component latent variables $w_i^{(t)}$ *and* $u_i^{(t)}$, $i = 1, \ldots, n$, *from the full conditionals (22)-(23) and the component parameters* $\boldsymbol{\mu}_k^{*(t)}, \boldsymbol{\psi}_k^{*(t)}, \boldsymbol{\Omega}_k^{*(t)}, \nu_k^{*(t)}$, $k = 1, \ldots, K$, *from the full conditionals (24)-(27).*

3 Sample $S_i^{(t)}$ *independently for each* $i = 1, \ldots, n$ *from*

$$
Pr(S_i = j|\mathbf{x}_i, \boldsymbol{\vartheta}) = \frac{g(\mathbf{x}_i|\boldsymbol{\theta}_j^*)Pr(S_i = j|\boldsymbol{\vartheta})}{\sum_{k=1}^{K} g(\mathbf{x}_i|\boldsymbol{\theta}_k^*)Pr(S_i = k|\boldsymbol{\vartheta})}. \tag{28}
$$

4 Set $t = t + 1$ and repeat the steps $2, 3$ and 4 until convergence is achieved.

Post-processed the MCMC, in order to deal with the label switching problem, the Kullback-Leibler algorithm introduced by Stephens (2000) is applied. Introduced by Redner and Walker (1984) into the mixture models background, the term *label switching* refers to the invariance of the mixture likelihood function under relabeling the components. Considering the maximum likelihood estimation, where we are looking for the corresponding modes of the likelihood function, label switching is not an object of interest. From the Bayesian point of view, however, it is a topic of concern because the labeling of the unobserved categories changes during the sample process of the mixture posterior distribution.

4 Application

In this section, MCMC methodology introduced in the previous section is applied to real data sets. Two different well known data sets in the context of finite mixture are analyzed. The first one, a univariate data set, is the body mass index data. The second one, a bivariate data set, is the Swiss Fertility and Socioeconomic Indicators (1888) data. Additionally, in order to compare the fit of the different models considered, we compute two classical comparison criteria, the Akaike Information Criterion (Akaike, 1974, AIC) and the Bayesian Information Criterion (Schwarz, 1978, BIC), and two versions proposed by Gelman et al. (2014) of the Bayesian criteria known as Watanabe-Akaike Information Criterion (Watanabe, 2010, WAIC)

4.1 The Body Mass Index Data

The body mass index (BMI) for men aged between 18 to 80 years is considered. This data set has been analyzed by Lin, Lee and Hsieh (2007), Lin, Lee and Yen (2007) and Basso et al. (2010). The data set comes from the National Health and Nutrition Examination Survey, developed by the National Center for Health Statistics (NCHS) of the Center for Disease Control (CDC) in the United States of America. The BMI, expressed in kg/m^2, is the ratio of body weight in kilograms and body height in meters squared and has been used as a standard measure for overweight and obesity.

The original sample consists of 4579 BMI records observations, however, as in Lin, Lee and Hsieh (2007), Lin, Lee and Yen (2007) and Basso et al. (2010), in order to explore the mixture characteristics, it is considered only those participants who have their weights within $[39.50kg, 70.00kg]$ and $[95.01kg, 196.80kg]$. In consequence, the remaining data set is composited by two subgroups: the first consists of 1069 participants and the second, of 1054.

The FM-SN, FM-ST and FM-SSL univariate models were fitted to the BMI data set. To that end, the priors hyperparameters were chosen as: $e_0 = 4$, $\mathbf{b}_0 = (0, 0)$, $\mathbf{B}_0 = Diag(10, 100)$, $c_0 = 2.5$, $g_0 = 0.5 + (r - 1)/2$, $r = 2$, $G_0 = g_0(\rho S_x)^{-1}$, $\rho = 0.5$, in which S_x is the sample variance. For the FM-ST and FM-SSL models, $\alpha = 2$ and $\gamma = 0.1$ (Juárez and Steel, 2010) were specified. Lastly, 50000 iterations of the algorithm MCMC were executed, the first 10000 draws were discarded as a burn-in period, and then the next 40000 were recorded. In order to reduce the autocorrelation between successive outputs of the simulated chain, only every 40th values of the chain were stored. With the resulting 1000, the posterior estimates were calculated.

Table 1 contains the parameters maximum a posteriori estimation for the models under analysis: FM-SN, FM-ST and FM-SSL, besides their corresponding 95% high posterior density credibility interval. Additionally, we computed the BIC, AIC, WAIC_1 and the WAIC_2 as models comparison criteria. The criteria values indicate that the FM-ST model has the best fitting result followed by the FM-SSL. An interesting point to underline is that our results are in line with Basso et al. (2010), as long as the authors obtained very similar results for the FM-ST and FM-SSL. It is important to mention that the Kullback-Leibler algorithm (Stephens, 2000) was applied and the label switching problem was not identified, probably because there are two well defined components. This verification validates all the previous results.

Figure 1 consists in a graphical comparison between the three models under analysis in this work when they are applied to density estimation. In order to obtain a better visualization, the fitting results were superimposed on a single set of coordinate axes. The figure analysis indicates that the heavy tailed FM-SMSN models (FM-ST and FM-SSL) have a better fit than the FM-SN model. Hence, it is possible to say that both analysis, the one based on models comparison criteria and the graphical, point that heavy tailed FM-SMSN models have more satisfactory results.

Also in figure 1, a visual analysis indicates that the first component presents a symmetric behav-

Table 1: Estimation results for fitting the distributions under analysis to the BMI data.

Parameters	FM-SN		FM-ST		FM-SSL	
	MODE	95%	MODE	95%	MODE	95%
μ_1	21.102	(19.952,22.706)	20.711	(19.732,21.939)	20.695	(19.940,21.600)
μ_2	28.296	(27.692,28.732)	29.107	(28.585,29.715)	28.775	(28.292,29.332)
σ_1^2	5.578	(4.484,8.426)	5.705	(4.508,9.434)	5.732	(4.186,8.138)
σ_2^2	64.000	(57.248,71.870)	39.810	(31.529,49.980)	35.927	(27.002,45.219)
λ_1	0.305	(-0.601,1.166)	0.647	(-0.251,1.292)	0.605	(0.030,1.159)
λ_2	3.286	(2.284,4.131)	2.377	(1.659,3.289)	2.904	(2.178,3.768)
η_1	0.484	(0.453,0.513)	0.491	(0.463,0.516)	0.487	(0.463,0.516)
η_2	0.516	(0.487,0.547)	0.509	(0.484,0.537)	0.513	(0.484,0.537)
ν_1	-	-	31.914	(12.671,78.699)	8.840	(3.450,31.672)
ν_2	-	-	7.050	(4.538,12.265)	2.588	(1.786,3.804)
BIC	13808.20		**13790.21**		13790.62	
AIC	13768.63		**13739.34**		13739.74	
WAIC_1	13765.27		**13730.95**		13734.55	
WAIC_2	13748.77		**13714.91**		13715.49	

ior. Reinforcing the results introduced on table 1, figure 2 illustrates very well this comportment in the sense that for the FM-SN and FM-ST models the first component skewness parameter credibility intervals contains the 0 and, for the FM-SSL, the credibility interval lower band is really close to 0. Another interesting point from the graphical analysis is that only the second component looks like to present heavy tails and, on table 1, these characteristics are clearly verified. To estimate a specific ν_k for each component $k = 1, 2$ is an advantage in comparison with Basso et al. (2010), it incorporates more flexibility in the modeling and makes the differentiation between heavy tails and non heavy tails components possible.

4.2 The Swiss Fertility and Socioeconomic Indicators (1888) Data

As an application of the multivariate models here proposed, the Swiss Fertility and Socioeconomic Indicators data (Mosteller and Tukey, 1977) is analysed. In 1888, Switzerland was entering a period known as the demographic transition, i. e., its fertility was beginning to fall from the high level typical of underdeveloped countries and life expectancy was rising. The data set consists on 47 French-speaking regions observations on 6 variables: fertility, males involved in agriculture as occupation, draftees receiving highest mark on army examination, education beyond primary school for draftees, catholic (as opposed to protestant) and infant mortality, each of which is in percent. For the actual analysis the variables males involved in agriculture as occupation and catholic (as

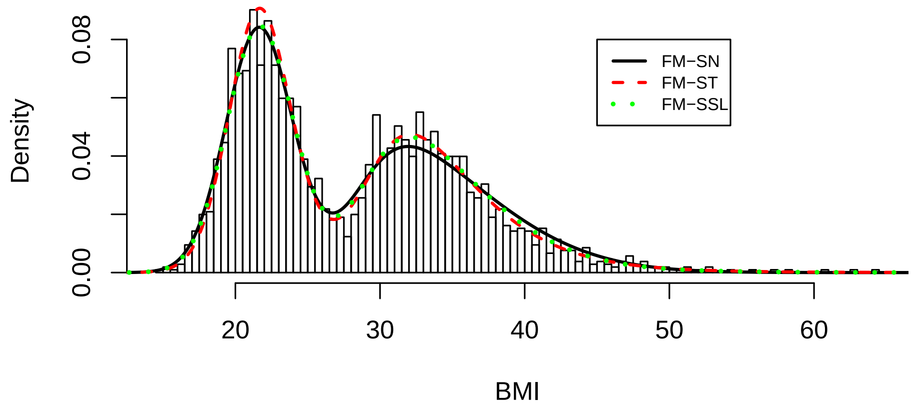

Figure 1: Histogram of the BMI data with FM-SN, FM-ST and FM-SSL fitted models.

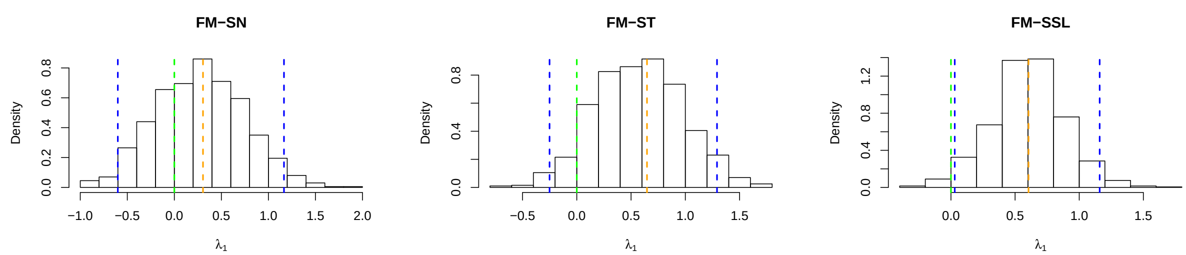

Figure 2: First Component skewness parameters full conditional samples.

opposed to protestant) was chosen.

Considering the estimation process for the FM-SN, FM-ST and FM-SSL models, the priors hyperparameters set was specified as: $e_0 = 4$, $\mathbf{b}_0 = (0,0,0,0)$, $\mathbf{B}_0 = Diag(100,50)$, $c_0 = 3$, $g_0 = 0.01$ and $G_0 = 0.01$. For the FM-ST and FM-SSL models, $\alpha = 2$ and $\gamma = 0.1$ (Juárez and Steel, 2010) were specified. A MCMC simulation for 20000 iterations was drawn, the first 10000 draws were discarded as a burn-in period, and then the next 10000 were recorded. In order to reduce the autocorrelation between successive values of the simulated chain, only every 10th values of the chain were stored. With the resulting 1000 we calculated the posterior estimates.

Table 2: Estimation results for fitting the distributions under analysis to the Swiss Fertility and Socioeconomic Indicators data.

Parameters	FM-SN		FM-ST		FM-SSL	
	MODE	95%	MODE	95%	MODE	95%
μ_{11}	86.9456	(40.1899,100.4691)	87.0765	(37.4758,98.9539)	87.8605	(30.184,97.5692)
μ_{12}	99.9113	(91.3471,104.1774)	100.0964	(91.2522,104.0726)	99.9200	(87.9491,103.5447)
μ_{21}	48.9545	(37.7709,58.8895)	54.6026	(44.0188,69.3384)	49.2419	(39.5855,60.5054)
μ_{22}	0.3929	(-2.858,2.9914)	1.4496	(-1.3172,2.8623)	0.9048	(-2.2383,2.5885)
λ_{11}	-6.4033	(-14.9332,5.8284)	-6.6017	(-15.4791,6.5275)	-6.8093	(-19.8278,14.5387)
λ_{12}	0.6630	(-1.2509,3.107)	0.7705	(-1.3708,3.0089)	0.6582	(-1.8449,3.4751)
λ_{21}	0.2034	(-1.5931,1.8167)	-0.3158	(-3.0543,1.5330)	0.2117	(-1.5624,1.7471)
λ_{22}	10.1714	(5.2164,19.1119)	7.5526	(1.9366,16.8224)	12.8892	(4.9878,24.2531)
$\Sigma_{1,11}$	829.1405	(228.4878,2320.9692)	804.7753	(258.3221,1821.5043)	817.4741	(199.3698,2202.7477)
$\Sigma_{1,12}$	155.7333	(20.4612,491.9568)	142.1053	(20.1312,373.1442)	147.8842	(10.3269,456.3256)
$\Sigma_{1,22}$	36.3575	(12.7539,129.5463)	31.8416	(10.1468,93.2204)	32.1106	(9.8144,112.6119)
$\Sigma_{2,11}$	454.4555	(260.2786,780.5774)	337.3832	(141.3357,756.5421)	413.5714	(176.3066,755.931)
$\Sigma_{2,12}$	-134.1912	(-418.8719,5.6376)	-78.6445	(-268.1597,1.2529)	-145.8333	(-366.7802,0.3296)
$\Sigma_{2,22}$	320.8791	(211.4496,567.751)	62.5551	(8.3697,320.8531)	342.3077	(54.8696,508.0684)
η_1	0.6519	(0.5165,0.7676)	0.3647	(0.2522,0.4951)	0.3611	(0.2337,0.4819)
η_2	0.3481	(0.2324,0.4835)	0.6353	(0.5049,0.7478)	0.6389	(0.5181,0.7663)
ν_1	-	-	16.3462	(1.3707,51.0251)	12.6596	(3.105,36.3747)
ν_2	-	-	2.8863	(1.0387,24.8196)	3.2718	(1.0002,32.3217)
BIC	**830.91**		831.20		839.87	
AIC	803.16		**799.74**		808.42	
WAIC_1	**813.37**		838.05		832.08	
WAIC_2	756.29		**730.75**		746.37	

As in the previous case, table 2 contains the parameters maximum a posteriori estimation for the multivariate models under analysis: FM-SN, FM-ST and FM-SSL, besides their corresponding 95% high posterior density credibility interval and the BIC, AIC, WAIC_1 and WAIC_2 to enable models comparison. The BIC and WAIC_1 comparison criteria values indicate that the FM-SN model has the best performance between the models under analysis, however, considering the AIC and WAIC_1, the FM-ST model looks like to have the best fit for the data. It is important to mention that, once again, the Kullback-Leibler algorithm (Stephens, 2000) was applied and the label switching problem was not identified.

A graphical analysis of the figure 3 permits a visualization of the results introduced by the models comparison criteria. Also in figure 3, a visual analysis indicates that only the second component looks like to present heavy tails. Illustrating the results introduced on table 2, figure 4 reinforces this perception as the second component degrees of freedom credibility intervals is in a range of lower values than the first component degrees of freedom.

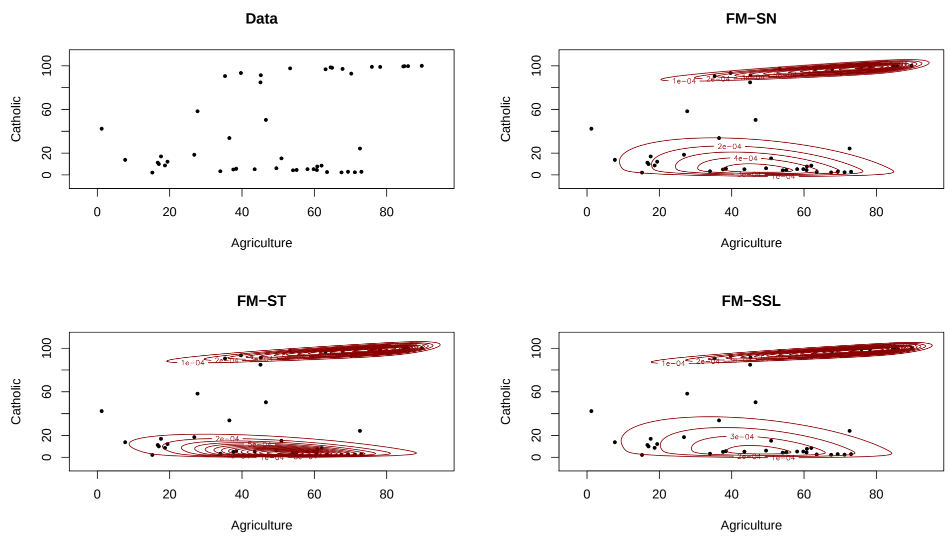

Figure 3: Plot of the Swiss Fertility and Socioeconomic Indicators data with FM-SN, FM-ST and FM-SSL fitted models.

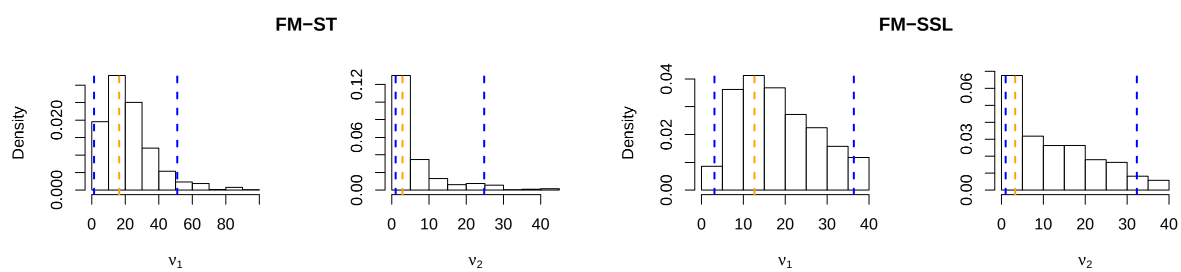

Figure 4: ν_k, $k = 1, 2$, full conditional samples.

5 Conclusion

In this work, a Bayesian finite mixture modeling based on scale mixtures of skew-normal distributions was proposed. The introduced models extend Basso et al. (2010) and Frühwirth-Schnatter and Pyne (2010) in the sense that not only the univariate case is investigated as in Basso et al. (2010) and that a wider range of SMSN distributions as the ones analyzed by Frühwirth-Schnatter and Pyne (2010) are studied. The class of models on focus here simultaneously accommodates multimodality, asymmetry and heavy tails, thus consists in a powerful tool for practitioners from different areas.

A Markov chain Monte Carlo algorithm is developed by exploring the hierarchical structure of the SMSN distributions. Additionally, two well known data sets in the finite mixture context are analyzed, the BMI data as an univariate application and the Swiss Fertility and Socioeconomic Indicators (1888) data as a multivariate application. An interesting extension which will be pursued in a future research is to develop fully Bayesian inference, i. e., to consider the number of components as an unknown quantity of interest.

A Finite mixture of Scale Mixtures of Skew-normal full conditional distributions

Considering the FM-SN model and assuming $\mathbf{F}_{n\times 2} = (\mathbf{1}\ \mathbf{w})$, for each $k = 1, \ldots, K$, construct a matrix $\mathbf{F}_k \in \Re^{N_k \times 2}$, $N_k = \sum_{i=1}^{n} S_{ik}$. Similarly, construct an observation matrix $\mathbf{x}_k \in \Re^{N_k \times p}$. Hence, by the Bayes theorem, the full conditionals are

- $\boldsymbol{\eta}|\mathbf{s} \sim D(e_0 + N_1, \ldots, e_0 + N_K)$;
- $(\mu_k, \psi_k)|\mathbf{s}, \mathbf{x}, \mathbf{w}, \tau_k^2 \sim N_2(\mathbf{b}_k, \mathbf{B}_k)$;

  $\mathbf{B}_k = \left(\frac{1}{\tau_k^2}\mathbf{B}_0^{-1} + \frac{1}{\tau_k^2}(\mathbf{F}_k^{'}\mathbf{F}_k)\right)^{-1}$

  $\mathbf{b}_k = \mathbf{B}\left(\frac{1}{\tau_k^2}\mathbf{B}_0^{-1}\mathbf{b}_0 + \frac{1}{\tau_k^2}(\mathbf{F}_k^{'}\mathbf{x}_k)\right)$
- $\tau_k^2|\mathbf{s}, \mathbf{x}, \mathbf{w}, C_0, \mu_k, \psi_k \sim IG(c_k, C_k)$;

  $c_k = c_0 + \frac{N_k}{2} + \frac{1}{2}$

  $C_k = C_0 + \frac{(\mathbf{x}_k - \mathbf{F}_k\beta_k)^{'}(\mathbf{x}_k - \mathbf{F}_k\beta_k) + (\beta_k - \mathbf{b}_0)^{'}\mathbf{B}_0^{-1}(\beta_k - \mathbf{b}_0)}{2}$
- $C_0|\tau_1^2, \ldots, \tau_K^2 \sim G(g, G)$.

  $g = g_0 + Kc_0$

  $G = G_0 + \sum_{k=1}^{K} \frac{1}{\tau_k^2}$

where $\beta_k = (\mu_k\ \psi_k)^{'}$. Considering now the latent variable $\mathbf{W}$

- $W_i|S_{ik} = 1, x_i, \mu_k, \psi_k, \tau_k^2 \sim TN_{[0,+\infty)}(a, A)$;

  $a = \frac{(x_i - \mu_k)\psi_k}{\tau_k^2 + \psi_k^2}$

  $A = \frac{\tau_k^2}{\tau_k^2 + \psi_k^2}$

For the FM-ST and the FM-SSL models the full conditionals are almost the same, the difference is that $\mathbf{F}$ is replaced by $\mathbf{F}_{n\times 2}^{w} = (\sqrt{\mathbf{u}}\ \sqrt{\mathbf{u}}\mathbf{w})$ and $\mathbf{x}$, by $\mathbf{x}^{w} = \sqrt{\mathbf{u}}\mathbf{x}$, where $\sqrt{\mathbf{u}}$ is the square root element by element. Considering now the latent variable $\mathbf{W}$

- $W_i|S_{ik}=1, x_i, u_i, \mu_k, \psi_k, \tau_k^2 \sim TN_{[0,+\infty)}(a, A/u_i)$.

Lastly, for the latent variable $\mathbf{U}$ and the degrees of freedom

- Skew-T

  $U_i|S_{ik}=1, x_i, w_i, \nu_k, \mu_k, \psi_k, \tau_k^2 \sim G\left(\frac{\nu_k}{2}+1, \frac{\nu_k}{2}+\frac{(x_i-\mu_k-\psi_k w_i)^2}{2\tau^2}+\frac{w_i^2}{2}\right)$;

- Skew-Slash

  $U_i|S_{ik}=1, x_i, w_i, \nu_k, \mu_k, \psi_k, \tau_k^2 \sim G_{(0,1)}\left(\nu_k+1, \frac{(x_i-\mu_k-\psi_k w_i)^2}{2\tau^2}+\frac{w_i^2}{2}\right)$;

  $\nu_k|\mathbf{s}, \mathbf{u} \sim G_{(1,40)}(\alpha+N_k, \gamma-\sum_{i:S_{ik}=1} u_i)$

For the degrees of freedom in skew-t is not possible to find a closed form to the full conditionals, so a Metropolis-Hastings step is required. To sample ν_k, $k=1,\dots,K$ a normal log random walk proposal was used

$$log(\nu_k^{new}-1) \sim N(log(\nu_k-1), c_{\nu_k}) \tag{29}$$

with adaptive width parameter c_{ν_k} (Shaby and Wells, 2010). The proposal was shifted away from 0, as it is advisable to avoid values for ν_k that are close to 0, see Fernández and Steel (1999).

B Finite mixture of Scale Mixture of Multivariate Skew-normal full conditionals distributions

Considering the FM-SN model and assuming $\mathbf{F}_{n\times 2}=(\mathbf{1}\ \mathbf{w})$, for each $k=1,\dots,K$, construct a matrix $\mathbf{F}_k \in \Re^{N_k\times 2}$, $N_k=\sum_{i=1}^n S_{ik}$. Similarly, construct an observation matrix $\mathbf{x}_k \in \Re^{N_k\times p}$. Hence, by the Bayes theorem, the full conditionals are

- $\boldsymbol{\eta}|\mathbf{s} \sim D(e_0+N_1,\dots,e_0+N_K)$;

- $(\boldsymbol{\mu}_k, \boldsymbol{\psi}_k)|\mathbf{s}, \mathbf{x}, \mathbf{w}, \boldsymbol{\Omega}_k \sim N_{2\times p}(\mathbf{b}_k, \mathbf{B}_k, \boldsymbol{\Omega}_k)$;

  $\mathbf{B}_k=\left(\mathbf{B}_0^{-1}+\mathbf{F}_k'\mathbf{F}_k\right)^{-1}$

$\mathbf{b}_k = \mathbf{B}\left(\mathbf{B}_0^{-1}\mathbf{b}_0 + \mathbf{F}_k'\mathbf{x}_k\right)$

- $\boldsymbol{\Omega}_k|\mathbf{s}, \mathbf{x}, \mathbf{w}, C_0, \boldsymbol{\mu}_k, \boldsymbol{\psi}_k \sim IW(c_k, C_k)$;

  $c_k = c_0 + N_k + p$

  $C_k = C_0 + (\mathbf{x}_k - \mathbf{F}_k\beta_k)'(\mathbf{x}_k - \mathbf{F}_k\beta_k) + (\beta_k - \mathbf{b}_0)'\mathbf{B}_0^{-1}(\beta_k - \mathbf{b}_0)$

- $\zeta_j|\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_K \sim G(g, G)$, $j = 1, \ldots, p$.

  $g = g_0 + K\frac{c_0}{2}$

  $G = G_0 + \frac{1}{2}\sum_{k=1}^{K}\boldsymbol{\Omega}_{k,jj}^{-1}$

where $\beta_k = (\boldsymbol{\mu}_k\ \boldsymbol{\psi}_k)'$. Considering now the latent variable $\mathbf{W}$

- $W_i|S_{ik} = 1, \mathbf{x}_i, \boldsymbol{\mu}_k, \boldsymbol{\psi}_k, \boldsymbol{\Omega}_k \sim TN_{[0,+\infty)}(a, A)$;

  $A = \frac{1}{1+\boldsymbol{\psi}'\boldsymbol{\Omega}_k^{-1}\boldsymbol{\psi}_k}$

  $a = ((\mathbf{x}_i - \boldsymbol{\mu}_k)\boldsymbol{\Omega}_k^{-1}\boldsymbol{\psi}_k)A.$

As in the univariate case, for the FM-ST and the FM-SSL models, $\mathbf{F}$ is replaced by $\mathbf{F}^w_{n\times 2} = (\sqrt{\mathbf{u}}\ \sqrt{\mathbf{u}}\mathbf{w})$, $\mathbf{x}$, by $\mathbf{x}^w = \sqrt{\mathbf{u}}\mathbf{x}$ and for latent variable $\mathbf{W}$,

- $W_i|S_{ik} = 1, \mathbf{x}_i, u_i, \boldsymbol{\mu}_k, \boldsymbol{\psi}_k, \boldsymbol{\Omega}_k \sim TN_{[0,+\infty)}(a, A/u_i)$.

Considering the latent variable $\mathbf{U}$ and the degrees of freedom,

- Skew-T

  $U_i|S_{ik} = 1, \mathbf{x}_i, w_i, \nu_k, \boldsymbol{\mu}_k, \boldsymbol{\psi}_k, \boldsymbol{\Omega}_k \sim G\left(\frac{\nu_k}{2} + 1, \frac{\nu_k}{2} + \frac{(\mathbf{x}_i - \boldsymbol{\mu}_k - \boldsymbol{\psi}_k w_i)'\boldsymbol{\Omega}_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k - \boldsymbol{\psi}_k w_i)}{2} + \frac{w_i^2}{2}\right)$;

- Skew-Slash

  $U_i|S_{ik} = 1, \mathbf{x}_i, w_i, \nu_k, \boldsymbol{\mu}_k, \boldsymbol{\psi}_k, \boldsymbol{\Omega}_k \sim G_{(0,1)}\left(\nu_k + 1, \frac{(\mathbf{x}_i - \boldsymbol{\mu}_k - \boldsymbol{\psi}_k w_i)'\boldsymbol{\Omega}_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k - \boldsymbol{\psi}_k w_i)}{2} + \frac{w_i^2}{2}\right)$;

  $\nu_k|\mathbf{s}, \mathbf{u} \sim G_{(1,40)}(\alpha + N_k, \gamma - \sum_{i:S_{ik}=1} u_i)$

As before, for the degrees of freedom in skew-t is not possible to find a closed form to the full conditionals, so a Metropolis-Hastings step was required and the same proposal distribution (29) was used.

References

Akaike, H. (1974), "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, 19, 716–723.

Andrews, D. F., and Mallows, C. L. (1974), "Scale mixtures of normal distributions," *Journal of the Royal Statistical Society, Series B*, 36, 99–102.

Azzalini, A. (1985), "A class of distributions which includes the normal ones," *Scandinavian Journal of Statistics*, 12, 171–178.

Azzalini, A. (1986), "Further results on a class of distributions which includes the normal ones," *Statistica*, 46, 199–208.

Azzalini, A., and Capitanio, A. (2003), "Distributions generated by pertubation of symmetry with emphasis on a multivariate skew-t distribution," *Journal of the Royal Statistical Society, Series B*, 65, 367–389.

Azzalini, A., and Dalla Valle, A. (1996), " The multivariate skew normal distribution," *Biometrika*, 83, 715–726.

Basso, R. M., Lachos, V. H., Cabral, C. R. B., and Gosh, P. (2010), "Robust mixture modeling based on scale mixtures of skew-normal distributions," *Computational Statistics and Data Analysis*, 54, 2926–2941.

Bouguila, N., Ziou, D., and Vaillancourt, J. (2004), "Unsupervised learning of a finite mixture model based on the Dirichlet distribution and its application," *IEEE Transactions on Image Processing*, 13, 1533–1543.

Branco, M. D., and Dey, D. K. (2001), "A general class of multivariate skew-elliptical distributions," *Journal of Multivariate Analysis*, 79, 99–113.

da Paz, R. F., Bazán, J. L., and Milan, L. A. (2017), "Bayesian estimation for a mixture of simplex distributions with an unknown number of components: HDI analysis in Brazil," *Journal of Applied Statistics*, 44, 1630–1643.

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

Diebolt, J., and Robert, C. P. (1994), "Estimation of finite mixture distributions through Bayesian sampling," *Journal of the Royal Statistical Society, Series B*, 56, 363–375.

Fernández, C., and Steel, M. F. J. (1999), "Multivariate student-t regression models: Pitfalls and inference," *Biometrika*, 86, 153–167.

Frühwirth-Schnatter, S. (2006), *Finite Mixture and Markov Switching Models*, 1 edn, New York: Springer.

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

Fu, R., Dey, D. K., and Holsinger, K. E. (2011), "A Beta-Mixture Model for Assessing Genetic Population Structure," *Biometrics*, 67, 1073–1082.

Gelman, A., Hwang, J., and Vehtari, A. (2014), "Understanding predictive information criteria for Bayesian models," *Statistics and Computing*, 24, 997–1016.

Henze, N. (1986), "A probabilistic representation of the skew-normal distribution," *Scandinavian Journal of Statistics*, 13, 271–275.

Jennison, C. (1997), "Discussion of the paper by Richardson and Green," *Journal of the Royal Statistical Society, Series B*, 59, 778–779.

Juárez, M. A., and Steel, M. F. J. (2010), "Model-based clustering of non-Gaussian panel data based on skew-t distributions," *Journal of Business & Economic Statistics*, 28, 52–66.

Lin, T., Lee, J., and Hsieh, W. (2007), "Robust mixture modelling using the skew t distribution," *Statistics and Computing*, 17, 81–92.

Lin, T., Lee, J., and Yen, S. (2007), "Finite mixture modelling using the skew normal distribution," *Statistica Sinica*, 17, 909–927.

Mosteller, F., and Tukey, J. W. (1977), *Data Analysis and Regression: A Second Course in Statistics*, 1 edn, Reading: Addison-Wesley.

Redner, R. A., and Walker, H. (1984), "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Review*, 26, 195–239.

Richardson, S., and Green, P. J. (1997), "On Bayesian analysis of mixtures with an unknown number of components," *Journal of the Royal Statistical Society, Series B*, 59, 731–792.

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

Shaby, B. A., and Wells, M. T. (2010), Exploring an Adaptive Metropolis Algorithm,, Technical report, Duke University, Department of Statistical Science.

Stephens, M. (2000), "Dealing with label switching in mixture models," *Journal of the Royal Statistical Society, Series B*, 62, 795–809.

Tanner, M. A., and Wong, W. H. (1987), "The calculation of posterior distributions by data augmentation," *Journal of the American Statistical Association*, 82, 528–540.

Wang, J., and Genton, M. G. (2006), "The multivariate skew-slash distribution," *Journal of Statistical Planning and Inference*, 136, 209–220.

Watanabe, S. (2010), "Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory," *The Journal of Machine Learning Research*, 11, 3571–3594.

Zhang, H., Wu, Q. M. J., and Nguyen, T. M. (2013), "Incorporating Mean Template Into Finite Mixture Model for Image Segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, 24, 328–335.