

# Robust Mixture Regression Modeling based on Scale Mixtures of Skew-Normal Distributions

Camila Borelli Zeller<sup>a</sup>

Celso Rômulo Barbosa Cabral<sup>b</sup>

Víctor Hugo Lachos<sup>c \*</sup>

<sup>a</sup>*Departamento de Estatística, Universidade Federal de Juiz de Fora, Brazil*

<sup>b</sup>*Departamento de Estatística, Universidade Federal do Amazonas, Brazil*

<sup>c</sup>*Departamento de Estatística, Universidade Estadual de Campinas, Brazil*

## Abstract

The traditional estimation of mixture regression models is based on the assumption of normality (symmetry) of component errors and thus is sensitive to outliers, heavy-tailed errors and/or asymmetric errors. In this work we present a proposal to deal with these issues simultaneously in the context of the mixture regression by extending the classic normal model by assuming that the random errors follow a scale mixtures of skew-normal distributions. This approach allows us to model data with great flexibility, accommodating skewness and heavy tails. The main virtue of considering the mixture regression models under the class of scale mixtures of skew-normal distributions is that they have a nice hierarchical representation which allows easy implementation of inference. We develop a simple EM-type algorithm to perform maximum likelihood inference of the parameters of the proposed model. In order to examine the robust aspect of this flexible model against outlying observations, some simulation studies are also been presented. Finally, a real data set is analyzed, illustrating the usefulness of the proposed method.

**Keywords:** EM algorithm, Finite mixtures of regression models, Scale mixtures of skew-normal distributions.

## 1 Introduction

Modeling based on finite mixture distributions is a rapidly developing area with an exploding range of applications. Finite mixture models are nowadays applied in such diverse areas as biology, biometrics, genetics, medicine and marketing, among others. There are various features of finite mixture distributions that make them useful in statistical modeling. For instance, statistical models which are based on finite mixture distributions capture many specific properties of real data such as multimodality, skewness, kurtosis, and unobserved heterogeneity. The importance of mixture distributions can be noted from the large number of books on mixtures, including Lindsay (1995), Böhning (2000), McLachlan & Peel (2000), Frühwirth-Schnatter (2006) and Mengersen *et al.*

---

\*Correspondence to: Departamento de Estatística, IMECC, Universidade Estadual de Campinas, CEP 13083-859, Campinas, São Paulo, Brazil. Tel.: +55 1935216078. E-mail address: hlachos@ime.unicamp.br (V.H. Lachos)

(2011) and the special editions of the journal *Computational Statistics and Data Analysis* (Böhning *et al.*, 2007, 2014).

On the other hand, in applied statistics, a large number of applications deal with relating a random variable  $Y_i$ , which is observed on several occasions  $i = 1, \dots, n$ , to a set of explanatory variables or covariates  $(x_{i1}, \dots, x_{id-1})$  through a regression-type model, where the conditional mean of  $Y_i$  is assumed to depend on  $\mathbf{x}_i = (1, x_{i1}, \dots, x_{id-1})^\top$  through  $E(Y_i | \boldsymbol{\beta}, \mathbf{x}_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$ , where  $\boldsymbol{\beta}$  is a vector of unknown regression coefficients of dimension  $d$ . In many circumstances, however, the assumption that the regression coefficient is fixed over all possible realizations of  $Y_1, \dots, Y_n$  is inadequate, and models where the regression coefficient changes are of great practical importance. One way to capture such changes in the parameter of a regression model is to use finite mixtures of regression models (MRM). MRM are widely used to investigate the relationship between variables coming from several unknown latent homogeneous groups. They were first introduced by Quandt (1972) under the titles *switching regression* or *clusterwise linear regression* (Späth, 1979). Comprehensive surveys are available in McLachlan & Peel (2000), and from a Bayesian point of view, in Frühwirth-Schnatter (2006, chap. 8).

The literature on maximum likelihood estimation of the parameters of the Gaussian MRM (hereafter N-MRM), is very extensive. Applications include marketing (DeSarbo & Cron, 1988; DeSarbo *et al.*, 1992; Quandt & Ramsey, 1978), economics (Cosslett & Lee, 1985; Hamilton, 1989), agriculture (Turner, 2000), nutrition (Arellano-Valle *et al.*, 2008), and psychometrics (Liu *et al.*, 2011). The standard algorithm in this case is the so-called EM (Expectation-Maximization) of Dempster *et al.* (1977), or perhaps some extension like the ECM (Meng & Rubin, 1993) or the ECME (Liu & Rubin, 1994) algorithms. Many extensions of this classic model have been proposed to broaden the applicability of linear regression analysis to situations where the Gaussian error term assumption may be inadequate, for example, because the datasets involve skewed or longer than normal tails errors. Some such extensions rely on the use of the skew-normal (Azzalini, 1985) and the Student-t distributions, respectively. A MRM based on the Student-t model (T-MRM) has been recently proposed by Yao. *et al.* (2014) to estimate the mixture regression parameters robustly. Liu & Lin (2014) proposed a robust version of the MRM by using the skew-normal model (SN-MRM), which appears to be a more theoretically compelling modeling tool for practitioners because it can investigate differential effects of covariates and accommodate moderately asymmetrical errors.

In this article, we propose a unified robust mixture regression model based on scale mixtures of skew-normal distributions (SMSN) by extending the mixture of scale mixtures of skew-normal distributions proposed by Basso *et al.* (2010) to the regression setting. The class of SMSN distributions, proposed by Branco & Dey (2001), is attractive since it simultaneously models skewness with heavy tails. Besides this, it has a stochastic representation for easy implementation of the EM algorithm and it also facilitates the study of many useful properties. This extension result in a flexible class of models for robust estimation in MRM since it contains distributions such as the skew-normal distribution and all the symmetric class of scale mixtures of normal distributions defined by Andrews & Mallows (1974). Moreover, the class of SMSN distributions is a rich class that contains proper elements such as the skew-t (Azzalini & Capitanio, 2003), skew-slash (Wang & Genton, 2006) and skew-contaminated normal distribution (Lachos *et al.*, 2010). Therefore they can be used in many types of models to infer robustness. In addition, this rich class of distributions can naturally attribute different weights to each observation and consequently control the influence of a single observation on the parameter estimates. Thus, the objectives of this study are: (i) to propose a mixture regression estimation method based on SMSN distributions, extending the recent works of Yao. *et al.* (2014) and Liu & Lin (2014); (ii) to implement and evaluate the proposed method computationally; and (iii) to apply these results to the analysis of a real life dataset.

The remainder of the paper is organized as follows. In Section 2, we briefly discuss the SMSN distributions and some of their properties. In Section 3, we present the SMSN-MRM, including the EM algorithm for maximum likelihood (ML) estimation and the observed information matrix. In Sections 4 and 5, numerical examples using both simulated and real data are given to illustrate the performance of the proposed method. Finally, some concluding remarks are presented in Section 6.

## 2 Scale mixtures of skew-normal distributions

### 2.1 Preliminaries

In order to introduce some notations, we start with the definition of SMSN distributions; see Branco & Dey (2001) for more details.

**Definition 1.** (Azzalini, 1985). *We say that a random variable  $Y$  has a skew-normal distribution with location parameter  $\mu$ , dispersion parameter  $\sigma^2 > 0$  and skewness parameter  $\lambda$ , and we write  $Y \sim \text{SN}(\mu, \sigma^2, \lambda)$ , if its density is given by*

$$f(y) = 2\phi(y; \mu, \sigma^2)\Phi(a), \quad y \in \mathbb{R},$$

where  $a = \lambda\sigma^{-1}(y - \mu)$ ,  $\phi(\cdot; \mu, \sigma^2)$  stands for the pdf of the univariate normal distribution with mean  $\mu$  and variance  $\sigma^2$ ,  $\Phi(\cdot)$  represents the distribution function of the standard univariate normal distribution.

**Definition 2.** *The distribution of the random variable  $Y$  belongs to the family of scale mixtures of skew-normal (SMSN) distributions when*

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

where  $\mu$  is a location parameter,  $X \sim \text{SN}(0, \sigma^2, \lambda)$ ,  $\kappa(\cdot)$  is a positive weight function and  $U$  is a positive random variable with a cdf  $H(u; \mathbf{v})$ , where  $\mathbf{v}$  is a (possibly multivariate) parameter indexing the distribution of  $U$ , known as the scale factor parameter, which is independent of  $X$ .

We use the notation  $Y \sim \text{SMSN}(\mu, \sigma^2, \lambda; H)$ . The name of the class becomes clear when we note that the conditional distribution of  $Y$  given  $U = u$  is skew-normal. Specifically, we have that

$$Y|U = u \sim \text{SN}(\mu, \kappa(u)\sigma^2, \lambda), \quad U \sim H(\cdot; \mathbf{v}).$$

Thus, the density of  $Y$  is given by

$$g(y) = 2 \int_0^\infty \phi(y; \mu, \kappa(u)\sigma^2)\Phi(\kappa(u)^{-1/2}a)dH(u; \mathbf{v}). \quad (2.1)$$

We also write  $Y \sim \text{SMSN}(\mu, \sigma^2, \lambda, \mathbf{v})$ , observing that here there is a little abuse of notation, by omission of  $H$ .

One particular case of this distribution is the skew-normal distribution (Azzalini, 1985), for which  $H$  is degenerate, with  $\kappa(u) = 1, u > 0$ . Also, when  $\lambda = 0$ , the SMSN distribution reduces to the scale mixtures of normal distribution (SMN) (Andrews & Mallows, 1974). In this work, without loss of generality, specifically in the numerical examples using both simulated and real data, we will concentrate on the case in which  $\kappa(u) = u^{-1}$ , i.e., the skew-t (ST) and the skew-slash

(SSL), whose properties have been widely discussed in Lachos *et al.* (2010), Basso *et al.* (2010) and Zeller *et al.* (2011), for example.

A random variable  $Y$  with a pdf as in (2.1) has a marginal stochastic representation, see Lachos *et al.* (2010), given by

$$Y \stackrel{d}{=} \mu + \Delta T + \kappa^{1/2}(U)\gamma T_1, \quad (2.2)$$

where  $\stackrel{d}{=}$  means “equal in distribution”,  $\Delta = \sigma\delta$ ,  $\delta = \frac{\lambda}{\sqrt{1+\lambda^2}}$ ,  $\gamma^2 = \sigma^2 - \Delta^2$ ,  $T_1 = \kappa^{1/2}(U)|T_0|$ ,  $|T_0|$  denotes the absolute value of  $T_0$ ,  $U \sim H(\cdot; \mathbf{v})$ ,  $T_0 \sim \mathbf{N}(0, 1)$  and  $T_1 \sim \mathbf{N}(0, 1)$  are all independent variables. The representation in (2.2) facilitates EM implementation for the maximum likelihood estimation. Another important result that will be useful in implementing the EM algorithm is given next. The statements of these results can be found in Zeller *et al.* (2011).

**Proposition 1.** *Let  $Y \sim \text{SMSN}(\mu, \sigma^2, \lambda; H)$  and let  $U \sim H$  be the mixing random scale factor. Then*

$$\begin{aligned} E[Y] &= \mu + \sqrt{\frac{2}{\pi}} K_1 \Delta, \quad \text{Var}[Y] = \sigma^2 \left( K_2 - \frac{2}{\pi} K_1^2 \delta^2 \right), \\ u_r &= E[\kappa^{-r}(U)|y] = \frac{2f_0(y)}{f(y)} \mathbf{E}\{\kappa^{-r}(U_y)\Phi(\kappa^{-1/2}(U_y)a)\} \quad \text{and} \\ \eta_r &= E[\kappa^{-r/2}(U)W_\Phi(\kappa^{-1/2}(U)a)|y] = \frac{2f_0(y)}{f(y)} \mathbf{E}\{\kappa^{-r/2}(U_y)\phi(\kappa^{-1/2}(U_y)a)\}, \end{aligned}$$

where  $W_\Phi(\cdot) = \frac{\phi_1(\cdot)}{\Phi(\cdot)}$ ,  $a = \lambda y_0$ , with  $y_0 = \frac{(y-\mu)}{\sigma}$ ,  $f_0$  is the pdf of  $Y_0 \sim \text{SMN}(\mu, \sigma^2; H)$ ,  $U_y \stackrel{d}{=} U|Y_0 = y$  and  $K_r = E[\kappa^{r/2}(U)]$ ,  $r = 1, 2, \dots$

## 2.2 Examples of SMSN distributions

In this section we consider some particular cases of SMSN distributions. Although we can deal with any  $\kappa$  function, we restrict our attention to the case where  $\kappa(u) = 1/u$ , since it leads to good mathematical properties.

- *The skew-t distribution with  $\nu$  degrees of freedom.*

In this case, we consider  $U \sim \text{Gamma}(\frac{1}{2}, \frac{1}{2})$ ,  $\nu > 0$ , in Definition 2 – where  $\text{Gamma}(a, b)$  denotes the gamma distribution with mean  $a/b$ . The density of  $Y$  takes the form

$$g(y) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}\sigma} \left(1 + \frac{d}{\nu}\right)^{-\frac{\nu+1}{2}} T\left(\sqrt{\frac{\nu+1}{d+\nu}}A; \nu+1\right), \quad y \in \mathbb{R},$$

where  $d = (y-\mu)^2/\sigma^2$  and  $T(\cdot; \nu)$  denotes the distribution function of the standard Student-t distribution, with location zero, scale one and  $\nu$  degrees of freedom, namely  $t(0, 1, \nu)$ . We use the notation  $Y \sim \text{ST}(\mu, \sigma^2, \lambda, \nu)$ . A particular case of the skew-t distribution is the skew-Cauchy distribution, when  $\nu = 1$ . Also, when  $\nu \rightarrow \infty$ , we get the skew-normal distribution as the limiting case.

- *The skew–slash distribution.*

In this case we have  $U \sim \text{Beta}(v, 1)$  – where  $\text{Beta}(a, b)$  denotes the beta distribution with parameters  $a$  and  $b$  – with positive shape parameter  $v$ , and use the notation  $Y \sim \text{SSL}(\mu, \sigma^2, \lambda, v)$ . The density of  $Y$  is given by

$$g(y) = 2v \int_0^1 u^{v-1} \phi(y; \mu, u^{-1} \sigma^2) \Phi(u^{1/2} A) du, \quad y \in \mathbb{R}.$$

The skew-slash is a heavy-tailed distribution having as limiting distribution the skew-normal one (when  $v \rightarrow \infty$ ).

### 3 The proposed model

In this section, we consider the mixture regression model where the random errors follow a scale mixtures of skew-normal distributions (SMSN-MRM). In general, a normal mixture regression model (N-MRM) is defined as: let  $Z$  be a latent class variable such that given  $Z = j$ , the response  $y$  depends on the  $p$ -dimensional predictor  $\mathbf{x}$  in a linear way

$$Y = \mathbf{x}^\top \boldsymbol{\beta}_j + \varepsilon_j, \quad j = 1, \dots, G, \quad (3.3)$$

where  $G$  is the number of groups (also called components in mixture models) in the population and  $\varepsilon_j \sim N(0, \sigma_j^2)$  is independent of  $\mathbf{x}$ . Suppose  $P(Z = j) = p_j$  and  $Z$  is independent of  $\mathbf{x}$ , then the conditional density of  $Y$  given  $\mathbf{x}$ , without observing  $Z$ , is

$$f(y|\mathbf{x}, \boldsymbol{\theta}) = \sum_{j=1}^G p_j \phi(y|\mathbf{x}^\top \boldsymbol{\beta}_j, \sigma_j^2), \quad (3.4)$$

where  $\phi(\cdot|\mu, \sigma^2)$  is the density function of  $N(\mu, \sigma^2)$  and  $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \dots, \boldsymbol{\theta}_G^\top)^\top$ , with  $\boldsymbol{\theta}_j = (p_j, \boldsymbol{\beta}_j^\top, \sigma_j^2)^\top$ . The model (3.4) is the so called normal mixture of regression models. Following Yao. *et al.* (2014) and Liu & Lin (2014), we extend the N-MRM defined above by considering the linear relationship in (3.3) with the following assumption:

$$\varepsilon_j \sim \text{SMSN}(b\Delta_j, \sigma_j^2, \lambda_j, \mathbf{v}_j), \quad j = 1, \dots, G, \quad (3.5)$$

where  $\Delta_j = \sigma_j \delta_j$ ,  $\delta_j = \frac{\lambda_j}{\sqrt{1 + \lambda_j^2}}$ ,  $b = -\sqrt{\frac{2}{\pi}} K_1$ , with  $K_r = E[\kappa^{r/2}(U)]$ ,  $r = 1, 2, \dots$ , which corresponds to the regression model where the error distribution has mean zero and hence the regression parameters are all comparable.

The mixture regression model with scale mixtures of skew-normal distributions defined above can be formulated in a similar way to the model defined in (3.4) as follows:

$$f(y|\mathbf{x}, \boldsymbol{\theta}) = \sum_{j=1}^G p_j g(y|\mathbf{x}, \boldsymbol{\theta}_j), \quad (3.6)$$

where  $g(\cdot|\mathbf{x}, \boldsymbol{\theta}_j)$  is the density function of  $\text{SMSN}(\mathbf{x}^\top \boldsymbol{\beta}_j + b\Delta_j, \sigma_j^2, \lambda_j, \mathbf{v}_j)$  and  $\boldsymbol{\theta}_j = (p_j, \boldsymbol{\beta}_j^\top, \sigma_j^2, \lambda_j, \mathbf{v}_j)^\top$ . Concerning the parameter  $\mathbf{v}_j$  of the mixing distribution  $H(\cdot; \mathbf{v}_j)$ , for  $j = 1, \dots, G$ , it can be a vector

of parameters, e.g., the contaminated normal distribution. Thus, for computational convenience we assume that  $\mathbf{v} = \mathbf{v}_1 = \mathbf{v}_2 = \dots, \mathbf{v}_G$ . This strategy works very well in the empirical studies that we have conducted and greatly simplifies the optimization problem. Observe that the model considers that the regression coefficient and the error variance are not homogeneous over all independent possible pairs  $(Y_i, \mathbf{x}_i), i = 1 \dots, n$ . In fact, they change between subgroups of observations.

In the context of classic inference, the unknown parameter  $\boldsymbol{\theta}$ , given observations  $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$ , is traditionally estimated by the maximum likelihood estimate (MLE):

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^n \log(f(y_i | \mathbf{x}_i, \boldsymbol{\theta})). \quad (3.7)$$

Note that the maximizer of (3.7) does not have an explicit solution, so we propose to use an EM-type algorithm (Dempster *et al.*, 1977). For a gentle tutorial on the EM algorithm and its applications to parameter estimation for mixture models, see McLachlan & Peel (2000).

### 3.1 Maximum likelihood estimation via EM algorithm

In this section, we present an EM algorithm for the ML estimation of the mixture regression model with scale mixtures of skew-normal distributions. To explore the EM algorithm we present the SMSN-MRM in an incomplete-data framework, using the results presented in Section 2.

In order to simplify notations, algebra and future interpretations, it is appropriate to deal with a random vector  $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{iG})^\top$  instead of the random variable  $Z_i$ , where

$$Z_{ij} = \begin{cases} 1, & \text{if the } i\text{th observation is from the } j\text{th component;} \\ 0, & \text{otherwise.} \end{cases}$$

Consequently, under this approach the random vector  $\mathbf{Z}$  has multinomial distribution considering a withdrawal into  $G$  categories, with probabilities  $p_1, \dots, p_G$ , i.e.,

$$P(\mathbf{Z}_i = \mathbf{z}_i) = p_1^{z_{i1}} p_2^{z_{i2}} \dots p_G^{z_{iG}},$$

where  $\sum_{j=1}^G p_j = 1$ , such that

$$Y_i | Z_{ij} = 1 \stackrel{\text{ind}}{\sim} \text{SMSN}(\mathbf{x}_i^\top \boldsymbol{\beta}_j + b\Delta_j, \sigma_j^2, \lambda_j, \mathbf{v}_j).$$

For the vector  $\mathbf{Z}_i$  we will use the notation  $\mathbf{Z}_i \stackrel{\text{iid}}{\sim} \text{Multinomial}(1, p_1, \dots, p_g)$ . Observe that  $Z_{ij} = 1$  if and only if  $Z_i = j$ . Thus, from (2.2), the set-up defined above can be written hierarchically as

$$Y_i | T_i = t_i, U_i = u_i, Z_{ij} = 1 \stackrel{\text{ind}}{\sim} N(\mathbf{x}_i^\top \boldsymbol{\beta}_j + \Delta_j t_i, \kappa(u_i) \gamma_j^2), \quad (3.8)$$

$$T_i | U_i = u_i, Z_{ij} = 1 \stackrel{\text{iid}}{\sim} TN_1(b, \kappa(u_i); (b, \infty)), \quad (3.9)$$

$$U_i | Z_{ij} = 1 \stackrel{\text{ind}}{\sim} H(u_i; \mathbf{v}), \quad (3.10)$$

$$\mathbf{Z}_i \stackrel{\text{iid}}{\sim} \text{Multinomial}(1, p_1, \dots, p_g), \quad (3.11)$$

for  $i = 1, \dots, n$ , all independent, where  $\gamma_j^2 = \sigma_j^2 - \Delta_j^2$  and  $TN_1(r, s; (a, b))$  denotes the univariate normal distribution  $(N(r, s))$ , truncated on the interval  $(a, b)$ . Let  $\mathbf{y} = (y_1, \dots, y_n)^\top$ ,  $\mathbf{u} = (u_1, \dots, u_n)^\top$ ,

$\mathbf{t} = (t_1, \dots, t_n)^\top$  and  $\mathbf{z} = (\mathbf{z}_1^\top, \dots, \mathbf{z}_n^\top)^\top$ . Then, under the hierarchical representation (3.8)–(3.10), it follows that the complete log-likelihood function associated with  $\mathbf{y}_c = (\mathbf{y}^\top, \mathbf{u}^\top, \mathbf{t}^\top, \mathbf{z}^\top)^\top$  is

$$\ell_c(\boldsymbol{\theta}|\mathbf{y}_c) = c + \sum_{i=1}^n \sum_{j=1}^G z_{ij} \log p_j - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^G z_{ij} \log \gamma_j^2 - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^G \frac{z_{ij} \mathbf{K}^{-1}(u_i)}{\gamma_j^2} (y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j - \Delta_j t_i)^2,$$

where  $c$  is a constant that is independent of the parameter vector  $\boldsymbol{\theta}$ . Letting  $\widehat{\boldsymbol{\theta}}_j^{(k)} = (\widehat{p}_j^{(k)}, \widehat{\boldsymbol{\beta}}_j^{(k)\top}, \widehat{\sigma}_j^2)^{\top}$ ,  $(\widehat{\lambda}_j^{(k)}, \mathbf{v}^{(k)})^\top$ , the estimates of  $\boldsymbol{\theta}$  at the  $k$ -th iteration.

It follows, after some simple algebra, that the conditional expectation of the complete log-likelihood function has the form

$$Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) = c + \sum_{i=1}^n \sum_{j=1}^G \widehat{z}_{ij}^{(k)} \log p_j - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^G \widehat{z}_{ij}^{(k)} \log \gamma_j^2 - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^G \frac{\widehat{z}u_{ij}^{(k)}}{\gamma_j^2} (y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j)^2 \quad (3.12)$$

$$+ \sum_{i=1}^n \sum_{j=1}^G \frac{\widehat{z}ut_{ij}^{(k)}}{\gamma_j^2} (y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j) \Delta_j - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^G \frac{\widehat{z}ut^2_{ij}^{(k)}}{\gamma_j^2} \Delta_j^2, \quad (3.13)$$

where  $\widehat{z}_{ij}^{(k)} = E[Z_{ij}|y_i, \widehat{\boldsymbol{\theta}}^{(k)}]$ ,  $\widehat{z}u_{ij}^{(k)} = E[Z_{ij} \mathbf{K}^{-1}(U_i)|y_i, \widehat{\boldsymbol{\theta}}^{(k)}]$ ,  $\widehat{z}ut_{ij}^{(k)} = E[Z_{ij} \mathbf{K}^{-1}(U_i) T_i | y_i, \widehat{\boldsymbol{\theta}}^{(k)}]$  and  $\widehat{z}ut^2_{ij}^{(k)} = E[Z_{ij} \mathbf{K}^{-1}(U_i) T_i^2 | y_i, \widehat{\boldsymbol{\theta}}^{(k)}]$ . By using known properties of conditional expectation, we obtain

$$\widehat{z}_{ij}^{(k)} = \frac{\widehat{p}_j^{(k)} g(y_i | \mathbf{x}_i, \widehat{\boldsymbol{\theta}}_j^{(k)})}{\sum_{j=1}^G \widehat{p}_j^{(k)} g(y_i | \mathbf{x}_i, \widehat{\boldsymbol{\theta}}_j^{(k)})}, \quad (3.14)$$

$\widehat{z}u_{ij}^{(k)} = \widehat{z}_{ij}^{(k)} \widehat{u}_{ij}^{(k)}$ ,  $\widehat{z}ut_{ij}^{(k)} = \widehat{z}_{ij}^{(k)} \widehat{ut}_{ij}^{(k)}$  and  $\widehat{z}ut^2_{ij}^{(k)} = \widehat{z}_{ij}^{(k)} \widehat{ut}^2_{ij}^{(k)}$ , with

$$\widehat{u}_{ij}^{(k)} = \widehat{u}_{ij}^{(k)} (\widehat{m}_{ij}^{(k)} + b) + \widehat{M}_j^{(k)} \widehat{\eta}_{ij}^{(k)}, \quad (3.15)$$

$$\widehat{ut}^2_{ij}^{(k)} = \widehat{u}_{ij}^{(k)} (\widehat{m}_{ij}^{(k)} + b)^2 + \widehat{M}_j^{2(k)} + \widehat{M}_j^{(k)} (\widehat{m}_{ij}^{(k)} + 2b) \widehat{\eta}_{ij}^{(k)}, \quad (3.16)$$

where  $\widehat{M}_j^2 = \frac{\widehat{\gamma}_j^2}{\widehat{\gamma}_j^2 + \widehat{\Delta}_j^2}$  and  $\widehat{m}_{ij} = \widehat{M}_j^2 \frac{\widehat{\Delta}_j}{\widehat{\gamma}_j^2} (y_i - \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_j - b \widehat{\Delta}_j)$ ,  $i = 1, \dots, n$ , with all these quantities

evaluated at  $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{(k)}$ . Since  $a_{ij} = \frac{m_{ij}}{M_{ij}} = \lambda_j \sigma_j (y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j - b \Delta_j)$ , the conditional expectations

given in (3.15) - (3.16), specifically  $\widehat{u}_{ij} = \widehat{u}_{1ij}$  and  $\widehat{\eta}_{ij} = \widehat{\eta}_{1ij}$ , can be easily derived from the result given in Section 2 (see Proposition 1). Thus, at least for the ST and SCN distributions of the SMSN class, we have a closed-form expression for the quantities  $\widehat{u}_{ij}$  and  $\widehat{\eta}_{ij}$ , as can be found in Zeller *et al.* (2011) and Basso *et al.* (2010). For the SSL, Monte Carlo integration can be employed, which yields the so-called MC-EM algorithm; see Wei & Tanner (1990), McLachlan & Krishnan (2008) and Zeller *et al.* (2011).

When the M-step turns out to be analytically intractable, it can be replaced with a sequence of conditional maximization (CM) steps. The resulting procedure is known as the ECM algorithm. The ECME algorithm, a faster extension of EM and ECM, is obtained by maximizing the constrained Q-function with some CM-steps that maximize the corresponding constrained actual marginal likelihood function, called CML-steps. Next, we describe this EM-type algorithm

(ECME) for maximum likelihood estimation of the parameters of the SMSN-MRM.

**E-step:** Given  $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{(k)}$ , compute  $\widehat{z}_{ij}^{(k)}, \widehat{zu}_{ij}^{(k)}, \widehat{z\boldsymbol{u}}_{ij}^{(k)}$  and  $\widehat{zut}_{ij}^{2(k)}$ , for  $i = 1, \dots, n$ , using (3.15)-(3.16).

**CM-step:** Update  $\widehat{\boldsymbol{\theta}}^{(k+1)}$  by maximizing  $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})$  over  $\boldsymbol{\theta}$ , which leads to the following closed form expressions:

$$\widehat{p}_j^{(k+1)} = \frac{\sum_{i=1}^n \widehat{z}_{ij}^{(k)}}{n}, \quad (3.17)$$

$$\widehat{\boldsymbol{\beta}}_j^{(k+1)} = \left( \sum_{i=1}^n \widehat{z\boldsymbol{u}}_{ij}^{(k)} \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1} \sum_{i=1}^n (\widehat{zu}_i^{(k)} y_i - \widehat{zut}_{ij}^{(k)} \widehat{\Delta}_j^{(k)}) \mathbf{x}_i, \quad (3.18)$$

$$\widehat{\gamma}_j^{2(k+1)} = \frac{\sum_{i=1}^n \left[ \widehat{z\boldsymbol{u}}_{ij}^{(k)} (y_i - \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_j^{(k)})^2 - 2\widehat{zut}_{ij}^{(k)} \widehat{\Delta}_j^{(k)} (y_i - \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_j^{(k)}) + \widehat{zut}_{ij}^{2(k)} \widehat{\Delta}_j^{2(k)} \right]}{\sum_{i=1}^n \widehat{z}_{ij}^{(k)}}, \quad (3.19)$$

$$\widehat{\Delta}_j^{(k+1)} = \frac{\sum_{i=1}^n \widehat{zut}_{ij}^{(k)} (y_i - \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_j^{(k)})}{\sum_{i=1}^n \widehat{zut}_{ij}^{2(k)}}, \quad (3.20)$$

$$\widehat{\sigma}_j^{2(k+1)} = \widehat{\gamma}_j^{2(k+1)} + \widehat{\Delta}_j^{2(k+1)}, \quad \widehat{\lambda}_j^{(k+1)} = \frac{\widehat{\Delta}_j^{(k+1)}}{\sqrt{\widehat{\gamma}_j^{2(k+1)}}}, \quad j = 1, \dots, G. \quad (3.21)$$

**CML-step:** Update  $\widehat{\mathbf{v}}^{(k)}$  by maximizing the actual marginal log-likelihood function, obtaining

$$\widehat{\mathbf{v}}^{(k+1)} = \operatorname{argmax}_{\mathbf{v}} \sum_{i=1}^n \log \left( \sum_{j=1}^G p_j^{(k)} g(y_i | \mathbf{x}_i, \boldsymbol{\beta}_j^{(k+1)}, \sigma_j^{2(k+1)}, \boldsymbol{\lambda}_j^{(k+1)}, \mathbf{v}) \right), \quad (3.22)$$

where  $g(\cdot | \mathbf{x}_i, \boldsymbol{\theta}_j)$  is defined in (3.6).

A more parsimonious model is achieved by supposing  $\gamma_1^2 = \dots = \gamma_G^2 = \gamma^2$ , which can be seen as an extension of the N-MRM with restricted variance-covariance components. In this case, the updates for  $\widehat{p}_j^{(k)}, \widehat{\boldsymbol{\beta}}_j^{(k)}$  and  $\widehat{\Delta}_j^{(k)}$  remain the same, and the update for  $\widehat{\gamma}_j^{2(k)}$  is given as

$$\widehat{\gamma}^{2(k+1)} = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^G \widehat{z}_{ij}^{(k)} \widehat{\gamma}_j^{2(k+1)}.$$

The iterations are repeated until a suitable convergence rule is satisfied, e.g.,

$$\left| \frac{\ell(\widehat{\boldsymbol{\theta}}^{(k+1)})}{\ell(\widehat{\boldsymbol{\theta}}^{(k)})} - 1 \right| < 10^{-5}. \quad (3.23)$$

Useful starting values required to implement this algorithm are those obtained under the normality assumption when  $\widehat{\lambda}_j^{(0)} = 3\operatorname{sign}(\widehat{\rho}_j)$ , where  $\widehat{\rho}_j$  is the sample skewness coefficient for group  $j$ , for  $j = 1, \dots, G$ . However, in order to ensure that the true maximum likelihood estimates are identified, we recommend running the EM algorithm using a range of different starting values. Note that when  $\lambda_j = 0$  (or  $\Delta_j = 0$ ) the M-step equations reduce to the equations obtained assuming SMN distributions. Particularly, this algorithm clearly generalizes the results found in Yao. *et al.* (2014) by taking  $\kappa(U_i) = U_i^{-1}$  and  $U_i \sim \text{Gamma}(\frac{\nu}{2}, \frac{\nu}{2})$ ,  $i = 1, \dots, n$ .

## 3.2 Notes on implementation

It is well known that mixture models may provide a multimodal log-likelihood function. In this sense, the method of maximum likelihood estimation through EM algorithm may not give maximum global solutions if the starting values are far from the real parameter values. Thus, the choice of starting values for the EM algorithm in the mixture context plays a big role in parameter estimation. In our examples and simulation studies we consider the following procedure for the SMSN-MRM

- Partition the sample into  $g$  groups using the K-means clustering algorithm (Basso *et al.*, 2010);
- Compute the proportion of data points belonging to the same cluster  $j$ , say  $p_j^{(0)}$ ,  $j = 1, \dots, g$ . This is the initial value for  $p_j$ ;
- For each group  $j$ , compute the initial values  $\boldsymbol{\beta}_j^{(0)}$  using the method of least square. Then, use the residuals of each group to compute initial values  $(\sigma_j^2)^{(0)}$ ,  $\lambda_j^{(0)}$  and  $\nu^{(0)}$  by using the R package *smsnmix()* (Prates *et al.*, 2011).

## 3.3 Model selection

Because there is no universal criterion for mixture model selection, we chose four criteria to compare the models in the SMSN family. The first three are the Akaike information criterion (AIC), the Bayesian information criterion (BIC) and the efficient determination criterion (EDC). As the more popular AIC and BIC criteria, EDC has the form

$$-2\ell(\hat{\boldsymbol{\theta}}) + \rho c_n,$$

where  $\ell(\boldsymbol{\theta})$  is the actual log-likelihood,  $\rho$  is the number of free parameters that have to be estimated in the model and the penalty term  $c_n$  is a convenient sequence of positive numbers. Here, we use  $c_n = 0.2\sqrt{n}$ , a proposal that was considered in Basso *et al.* (2010) and Cabral *et al.* (2012). We have  $c_n = 2$  for AIC and  $c_n = \log n$  for BIC, where  $n$  is the sample size. We also considered the values of ICL (Integrated Completed Likelihood) choice criterion (Basso *et al.*, 2010) for the models. This is suitable for classification purposes because it penalizes model complexity (like AIC, BIC and EDC) and the inability of the fitted mixture model to provide a reasonable partition of the data. This inability is measured by the difference between ICL and BIC, which is the estimated mean entropy.

In the next sections, simulation studies and a real dataset are presented in order to illustrate the performance of the proposed method.

## 4 Simulation Experiments

In this section, we consider three simulation experiments to show the applicability of our proposed model. Our intention is to show that the SMSN-MRM can do exactly what it is designed for, that is, satisfactorily model data that have a structure with serious departures from the normal assumption.

Table 1: Scenario 1: Mean and standard deviations (SD) for EM estimates based on 500 samples from the SMSN-MRM. True values of parameters are in parentheses.

| Parameter        | SN      |        | ST      |        | SSL     |        |
|------------------|---------|--------|---------|--------|---------|--------|
|                  | Mean    | SD     | Mean    | SD     | Mean    | SD     |
| $\beta_{01}(-1)$ | -1.0008 | 0.1593 | -1.0054 | 0.2358 | -0.9902 | 0.1961 |
| $\beta_{11}(-4)$ | -4.0075 | 0.2640 | -3.9835 | 0.3442 | -4.0213 | 0.3144 |
| $\beta_{21}(-3)$ | -3.0036 | 0.1409 | -3.0039 | 0.1697 | -3.0117 | 0.1667 |
| $\beta_{02}(3)$  | 3.0017  | 0.0607 | 2.9863  | 0.0878 | 2.9925  | 0.0794 |
| $\beta_{12}(7)$  | 6.9975  | 0.0977 | 7.0080  | 0.1199 | 7.0008  | 0.1251 |
| $\beta_{22}(2)$  | 2.0013  | 0.0470 | 2.0037  | 0.0591 | 2.0016  | 0.0560 |
| $\sigma_1^2(2)$  | 1.9416  | 0.4546 | 1.9680  | 0.5810 | 1.9397  | 0.5642 |
| $\sigma_2^2(1)$  | 0.9820  | 0.1431 | 0.9517  | 0.1717 | 0.9589  | 0.1680 |
| $\lambda_1(2)$   | 2.1293  | 1.0379 | 2.1125  | 0.8213 | 2.0707  | 0.9563 |
| $\lambda_2(4)$   | 4.1458  | 1.2514 | 3.8421  | 1.0230 | 3.7720  | 1.0586 |
| $\nu(3)$         | -       | -      | 3.0142  | 0.4777 | 3.3427  | 1.2521 |
| $p_1(0.3)$       | 0.2998  | 0.0207 | 0.3002  | 0.0205 | 0.3008  | 0.0211 |

Table 2: Scenario 2: Mean and standard deviations (SD) for EM estimates based on 500 samples from the SMSN-MRM. True values of parameters are in parentheses.

| Parameter        | SN ( $\gamma_1^2 = \gamma_2^2$ ) |        | ST ( $\gamma_1^2 = \gamma_2^2$ ) |        | SSL( $\gamma_1^2 = \gamma_2^2$ ) |        |
|------------------|----------------------------------|--------|----------------------------------|--------|----------------------------------|--------|
|                  | Mean                             | SD     | Mean                             | SD     | Mean                             | SD     |
| $\beta_{01}(-1)$ | -1.0091                          | 0.1778 | -1.0370                          | 0.2346 | -0.9973                          | 0.2055 |
| $\beta_{11}(-4)$ | -3.9832                          | 0.2958 | -4.0005                          | 0.3386 | -4.0235                          | 0.3355 |
| $\beta_{21}(-3)$ | -3.0051                          | 0.1431 | -3.0016                          | 0.1707 | -3.0057                          | 0.1814 |
| $\beta_{02}(3)$  | 2.9882                           | 0.1065 | 2.9836                           | 0.1540 | 2.9915                           | 0.1306 |
| $\beta_{12}(7)$  | 7.0193                           | 0.1812 | 7.0125                           | 0.2246 | 7.0082                           | 0.2165 |
| $\beta_{22}(2)$  | 2.0093                           | 0.0997 | 1.9978                           | 0.1127 | 1.996                            | 0.1100 |
| $\sigma_1^2(2)$  | 1.8959                           | 0.4086 | 1.8088                           | 0.5414 | 1.8030                           | 0.5421 |
| $\sigma_2^2(2)$  | 1.9306                           | 0.2674 | 1.9085                           | 0.3992 | 1.8528                           | 0.4462 |
| $\lambda_1(2)$   | 1.8414                           | 0.5898 | 1.8268                           | 0.5598 | 1.6561                           | 0.6370 |
| $\lambda_2(2)$   | 1.8951                           | 0.4690 | 1.9110                           | 0.4608 | 1.7116                           | 0.5301 |
| $\nu(3)$         | -                                | -      | 3.0070                           | 0.5075 | 3.5332                           | 1.8907 |
| $p_1(0.3)$       | 0.2997                           | 0.0216 | 0.2981                           | 0.0219 | 0.3013                           | 0.0204 |

## 4.1 Experiment 1: Parameter recovery

In this section, we consider two scenarios for simulation in order to verify if we can estimate the true parameter values accurately by using the proposed estimation method. This is the first step to ensure that the estimation procedure works satisfactorily. We fit the SMSN-MRM to data that were artificially generated from the following SMSN-MRM:

$$\begin{cases} Y_i = \mathbf{x}_i^\top \boldsymbol{\beta}_1 + \varepsilon_1, & Z_{i1} = 1, \\ Y_i = \mathbf{x}_i^\top \boldsymbol{\beta}_2 + \varepsilon_2, & Z_{i2} = 1, \end{cases}$$

where  $Z_{ij}$  is a component indicator of  $Y_i$  with  $P(Z_{ij} = 1) = p_j, j = 1, 2, \mathbf{x}_i^\top = (1, x_{i1}, x_{i2})$ , such that  $x_{i1} \sim U(0, 1)$  and  $x_{i2} \sim U(-1, 1), i = 1, \dots, n$ , and  $\varepsilon_1$  and  $\varepsilon_2$  follow a distribution in the family of

SMSN distributions, as the assumption given in (3.5).

We generated 500 random samples of size  $n = 500$  from the SN, ST and the SSL models with the following parameter values:  $\boldsymbol{\beta}_1 = (\beta_{01}, \beta_{11}, \beta_{21})^\top = (-1, -4, -3)^\top$ ,  $\boldsymbol{\beta}_2 = (\beta_{02}, \beta_{12}, \beta_{22})^\top = (3, 7, 2)^\top$ ,  $p_1 = 0.3$  and  $\nu = 3$ . In addition, we consider the following scenarios: scenario 1 :  $\sigma_1^2 = 2$ ,  $\sigma_2^2 = 1$ ,  $\lambda_1 = 2$  and  $\lambda_2 = 4$ , and scenario 2 :  $\sigma_1^2 = \sigma_2^2 = 2$  and  $\lambda_1 = \lambda_2 = 2$ , i.e.,  $\gamma_1^2 = \gamma_2^2$ . We used the maximum likelihood estimation via EM algorithm for each sample, using the stopping criterion (3.23). No existing program is available to estimate SMSN-MRM directly. Therefore, ML estimation via the EM algorithm was implemented using R.

In the mixture context, the likelihood is invariant under a permutation of the class labels in parameter vectors. Therefore, a label switching problem can occur when some labels of the mixture classes permute (McLachlan & Peel, 2000). Although the switching of class labels is not a concern in the general course of the maximum likelihood estimation via the EM algorithm for studies with only one replication, it was a serious problem in our simulation study because the same model was estimated iteratively for 500 replications per cell. To solve this problem, we chose the labels by minimizing the distance to the true parameter values. The average values and the corresponding standard deviations (SD) of the EM estimates across all samples were computed and the results are presented in Tables 1 and 2. Note that all the point estimates are quite accurate in all the considered scenarios. Thus, the results suggest that the proposed EM-type algorithm produced satisfactory estimates.

## 4.2 Experiment 2: Classification

In this section, we illustrate the ability of the SMSN-MRM to fit data with a mixture structure generated from a different family of skew distribution and we also investigate the ability of the SMSN-MRM to cluster observations, that is, to allocate them into groups of observations that are similar in some sense. We know that each data point belongs to one of  $G$  heterogeneous populations, but we do not know how to discriminate between them. Modeling by mixture models allows clustering of the data in terms of the estimated (posterior) probability that a single point belongs to a given group.

A lot of work in model-based clustering has been done using finite mixtures of normal distributions. As the posterior probabilities  $\hat{z}_{ij}$ , defined in (3.14), can be highly influenced by atypical observations, there have been efforts to develop robust alternatives, like mixtures of t-Student distributions (see McLachlan & Peel (1998) and the references herein). Our idea is to extend the flexibility of these models, by including possible skewness of the related components; see the work of Liu & Lin (2014) based on the SN-MRM.

We generated 500 samples under the following scenarios: (a) scenario 1 (Figure 1): a mixture of two skew-Birnbaum-Saunders regression models (see Santana *et al.*, 2011; Vilca *et al.*, 2011); and (b) scenario 2 (Figure 2): a mixture of two skew-normal generalized hyperbolic models (see Vilca *et al.*, 2014). The parameter values were chosen to present a considerable proportion of outliers and the skewness pattern. It can be seen from Figures 1 and 2 that the groups are poorly separated. Furthermore, note that although we have a two component mixture, the histogram need not be bimodal.

For each sample of size  $n = 500$ , we proceed with clustering ignoring the known true classification. Following the method proposed by Liu & Lin (2014), to assess the quality of the classification function of each mixture model, an index measure was used in the current study, called correct classification rate (CCR), which is based on the posterior probability assigned to each subject. The SMSN-MRM were fitted using the algorithm described in Section 3.1 in order to obtain

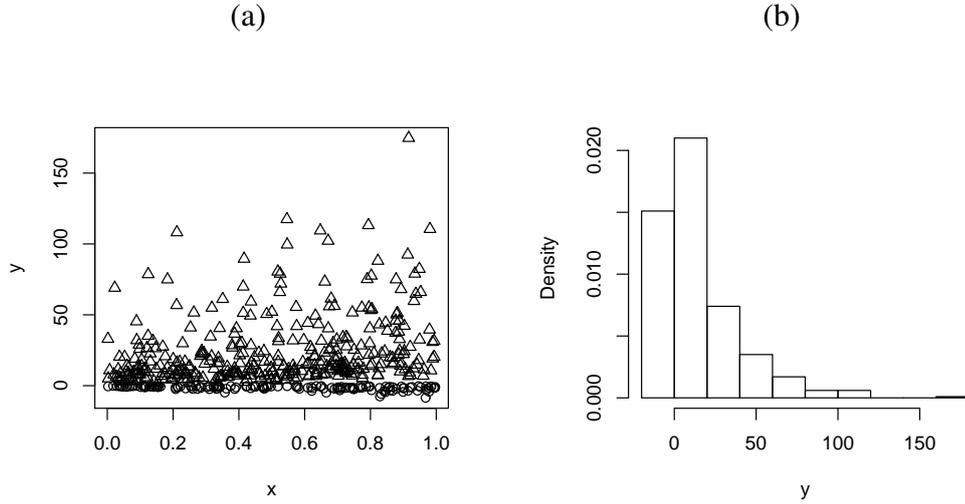


Figure 1: Experiment 2. (a) The scatter plot and (b) histogram for one of the simulated samples - scenario 1.

the estimate of the posterior probability that an observation  $y_i$  belongs to the  $j$ th component of the mixture, i.e.  $\hat{z}_{ij}$ . For sample  $l$ ,  $l = 1, \dots, 500$ , we compute the number of correct allocations (CCRs) divided by the sample size  $n$ , that is,  $ACCR = \frac{1}{500} \sum_{l=1}^{500} CCR_l$ . Table 3 shows the mean value of the correct allocation rates, where larger values indicate better classification results.

Table 3: Experiment 2. Mean of right allocation rates for fitted SMSN-MRM.

| Fitted Model | ACCR       |            |
|--------------|------------|------------|
|              | Scenario 1 | Scenario 2 |
| Normal       | 0.5211     | 0.7391     |
| SN           | 0.5432     | 0.6964     |
| ST           | 0.5747     | 0.7674     |
| SSL          | 0.5603     | 0.7634     |

Obviously, one expects the best classification rate when modeling with true components (scenario 1 and scenario 2), but it is interesting to verify what happens when we use SMSN components. Comparing with the results for the normal model, we see that modeling using the ST or SSL distribution represents an improvement in the outright clustering and has a better performance, showing their robustness to discrepant observations. Under scenario 1, the SN model showed better performance compared to the normal model, but this did not occur in scenario 2. This fact can be explained because the skew-normal distribution can still be affected by atypical observations since it does not have heavy tails as is the case of the ST and SSL models.

For each sample of size  $n = 500$ , we compare the ability of some classic model selection criteria (see Subsection 3.2) to select the appropriate model between the SMSN-MRM. Table 4 presents the percentages of models selected according to the four aforementioned criteria under five conditions, say, SN vs. Normal; ST vs. Normal; SSL vs. Normal; ST vs. SN; SSL vs. SN.

Under scenario 1 (data generated from a mixture of two skew-Birnbaum-Saunders regression models), comparing the asymmetric models SN, ST and SSL with the normal (symmetrical) model,

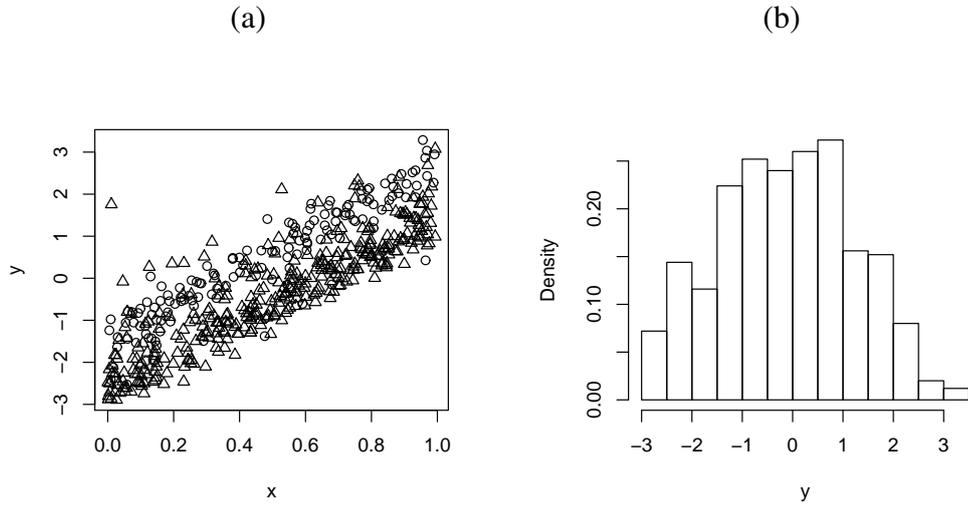


Figure 2: Experiment 2. (a) The scatter plot and (b) histogram for one of the simulated samples - scenario 2.

Table 4: Experiment 2. Percentages of preferred models under five conditions examined.

| Condition Examined | Scenario 1 |     |     |     |
|--------------------|------------|-----|-----|-----|
|                    | AIC        | BIC | EDC | ICL |
| SN vs Normal       | 94         | 77  | 87  | 96  |
| ST vs Normal       | 92         | 74  | 84  | 99  |
| SSL vs Normal      | 73         | 43  | 57  | 99  |
| ST vs SN           | 62         | 62  | 62  | 99  |
| SSL vs SN          | 17         | 17  | 17  | 99  |
| Condition Examined | Scenario 2 |     |     |     |
|                    | AIC        | BIC | EDC | ICL |
| SN vs. Normal      | 79         | 21  | 41  | 51  |
| ST vs. Normal      | 93         | 65  | 77  | 81  |
| SSL vs. Normal     | 92         | 59  | 73  | 78  |
| ST vs. SN          | 91         | 91  | 91  | 88  |
| SSL vs. SN         | 88         | 88  | 88  | 85  |

note that all criteria favor the asymmetric models (except the BIC when examining SSL vs. Normal). Moreover, note that the ICL criterion has the highest percentage since in this scenario the asymmetric models also performed better in classification (see Table 3). Comparing the asymmetric models with heavy tails (ST and SSL) to the SN model, the ST model was selected by all criteria.

Under scenario 2 (a mixture of two skew-normal generalized hyperbolic models), comparing the asymmetric models SN, ST and SSL with normal symmetrical model, note that all criteria favor the asymmetric models (excluding BIC and EDC criteria in condition when examining SN vs. Normal). Note that when comparing SN vs. Normal, the ICL favors the SN model but the number of right allocations is greater for the normal model (see Table 3). Furthermore, for all criteria, the asymmetric models with heavy tails (ST and SSL) fitted the data better than the SN

model.

### 4.3 Experiment 3

In this section, first we investigate the ability of the SMSN-MRM to cluster observations and then we compare the ability of some classic procedures to choose between the underlying SMSN-MRM. We fixed the number of components ( $G = 2$ ), sample size ( $n = 500$ ) and parameter values ( $\boldsymbol{\beta}_1 = (-1, -4)^\top$ ,  $\boldsymbol{\beta}_2 = (4, -6)^\top$ ,  $\sigma_1^2 = \sigma_2^2 = 2$ ,  $\lambda_1 = \lambda_2 = 2$ , i.e.  $\gamma_1^2 = \gamma_2^2$ , and  $p_1 = 0.3$ ), which is a restriction suggested by Basso *et al.* (2010) and Yao. *et al.* (2014). Then, without loss of generality, we artificially generated 500 samples from a mixture regression of skew-t ( $\nu = 3$ ) and, for each sample, we fitted the normal, SN, ST and the SSL models with homogeneous nature of the covariance structure. Figure 3 shows a scatter plot and a histogram for one of these simulated samples.

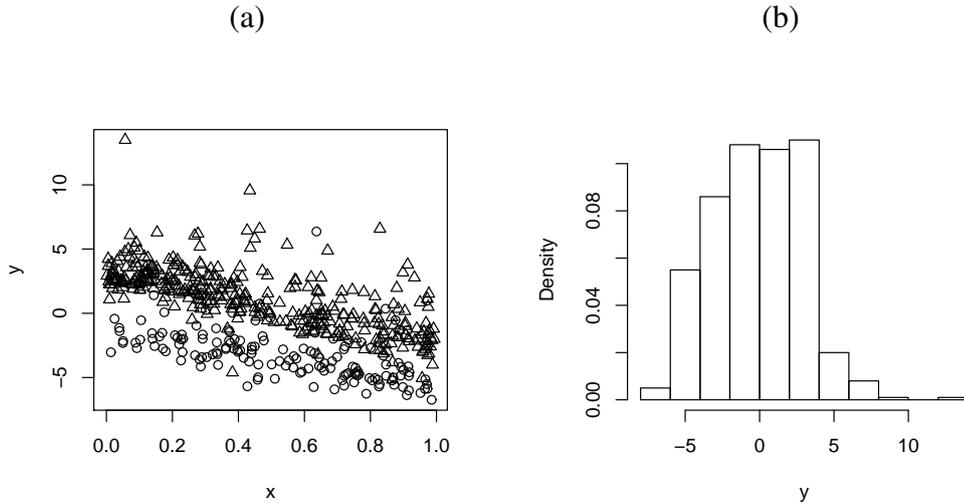


Figure 3: Experiment 3. (a) The scatter plot and (b) histogram for one of the simulated samples.

From the clustering standpoint, in order to give stronger evidence of the superiority of the method using the SMSN-MRM family, the right number of allocations was computed for each sample. The mean and standard deviation (SD) of right allocations of these samples are shown in Table 5. It can be seen that the means are greater and the standard deviations are smaller for the skewed/heavy-tailed SMSN-MRM, in particular to the true model, i.e. ST model ( $\gamma_1^2 = \gamma_2^2$ ). In addition, we present the mean value of the correct allocation rates (*ACCR*). Compared with the results for the normal model, modeling using the SN, ST or SSL distribution represents a substantial improvement in the outright clustering. Also, the ST model (true model) outperforms performance when compared with the SN and the SSL models, as expected.

For each fitted model, we computed the AIC, BIC, EDC and the ICL criterion. Table 6 shows the rates (percentages) at which the true model is chosen for each criterion. Note that all the criteria have satisfactory behavior, in that, they favor the true model, that is, the ST model with two components, except ICL which still performs poorly. Figure 4 shows the AIC values for each sample and model.

This simulation study shows similar results to those reported in Basso *et al.* (2010), in the context of mixture modeling based on scale mixtures of skew-normal distributions. We believe that this topic about model selection deserves a more detailed and extensive investigation, which is

Table 5: Experiment3. Right allocation analysis for 500 samples artificially generated from the ST model ( $\gamma_1^2 = \gamma_2^2$ ).

| Fitted Model                         | Mean of right allocations | SD of right allocations | CCR    |
|--------------------------------------|---------------------------|-------------------------|--------|
| Normal ( $\gamma_1^2 = \gamma_2^2$ ) | 412.3440                  | 109.4916                | 0.8247 |
| SN ( $\gamma_1^2 = \gamma_2^2$ )     | 442.2720                  | 58.7987                 | 0.8845 |
| ST ( $\gamma_1^2 = \gamma_2^2$ )     | 464.9040                  | 39.3982                 | 0.9298 |
| SSL ( $\gamma_1^2 = \gamma_2^2$ )    | 457.4340                  | 53.9565                 | 0.9149 |

Table 6: Experiment3. Percentages that the true model is chosen using different criteria.

| Condition Examined   | AIC | BIC | EDC | ICL |
|--|-----|-----|-----|-----|
| ST ( $\gamma_1^2 = \gamma_2^2$ ) vs Normal ( $\gamma_1^2 = \gamma_2^2$ ) | 100 | 100 | 100 | 5   |
| ST ( $\gamma_1^2 = \gamma_2^2$ ) vs SN ( $\gamma_1^2 = \gamma_2^2$ )     | 99  | 99  | 99  | 60  |
| ST ( $\gamma_1^2 = \gamma_2^2$ ) vs SSL ( $\gamma_1^2 = \gamma_2^2$ )    | 99  | 99  | 99  | 86  |

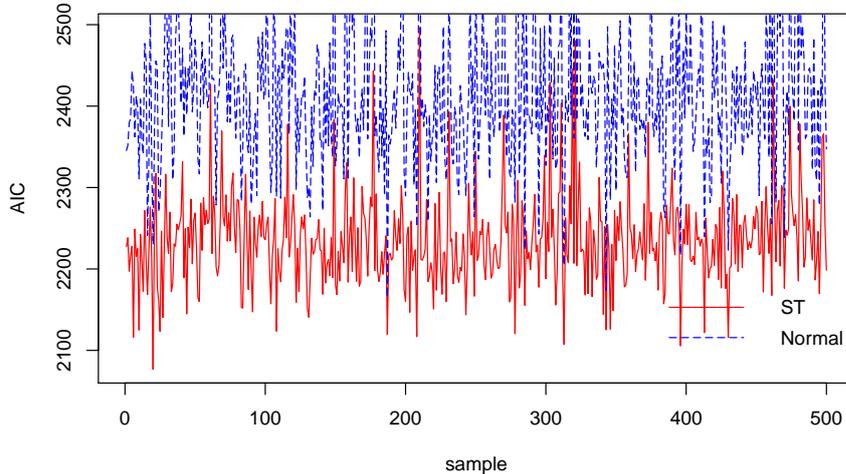


Figure 4: Experiment3. AIC values for 500 samples. Red line: ST model ( $\gamma_1^2 = \gamma_2^2$ ) and blue line: Normal model ( $\gamma_1^2 = \gamma_2^2$ ).

one of our purposes in order to extend the present paper including a study about the choice of the (possibly) unknown number of components, for example, and also treating the multivariate case. An overview of selection criteria can be found in Depraetere & Vandebroek (2014), in the context of mixture regression models based on the assumption of normality.

In addition, a real data set is analyzed, illustrating the usefulness of the proposed method. Thus, in the following application, we use those criteria as a rough guide for model choice.

## 5 Real dataset

We illustrate our proposed methods with a dataset obtained from Cohen (1984), representing the perception of musical tones by musicians. In this tone perception experiment a pure fundamental tone with electronically generated overtones added was played to a trained musician. The subjects were asked to tune an adjustable tone to one octave above the fundamental tone and their perceived tone was recorded versus the actual tone. The experiment recorded 150 trials from the same musician. The overtones were determined by a stretching ratio, which is the ratio between adjusted tone and the fundamental tone. Two separate trends clearly emerge, see Figure 5(a), which relate to two hypotheses explored in Cohen (1984), called the interval memory hypothesis and the partial matching hypothesis. Many articles have analyzed this dataset using a mixture of linear regressions framework; see DeVaux (1989), Viele & Tong (2002) and Hunter & Young (2012). These data were analyzed recently by Yao. *et al.* (2014), leading them to propose a robust mixture regression using the t-distribution. Now we revisit this dataset with the aim of expanding the inferential results to the SMSN family. Specifically, we focus on the SN, ST and the SSL distributions. To verify the existence of skewness in the data, Figure 5(b) presents a histogram of the data and shows there is apparent non-normal pattern.

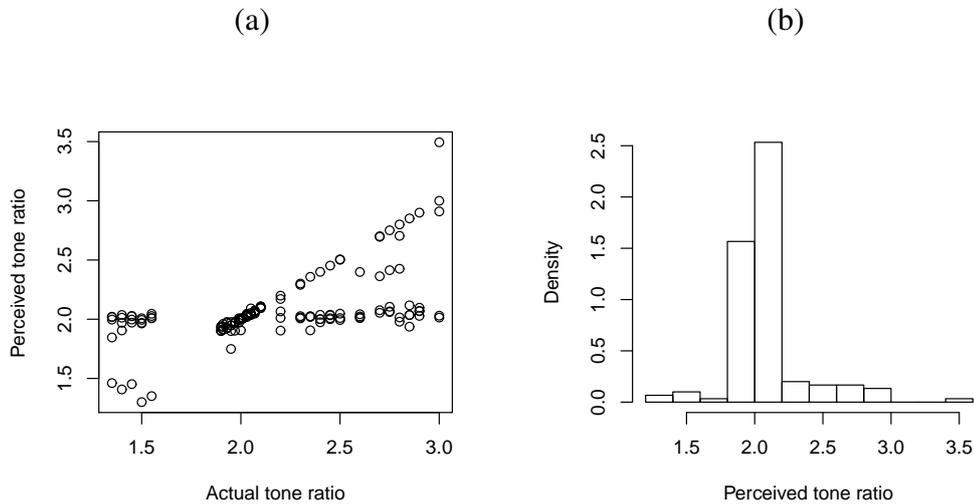


Figure 5: (a) The scatter plot and (b) histogram of the tone perception data.

Table 7 presents the ML estimates of the parameters from the normal ( $\gamma_1^2 = \gamma_2^2$ ), SN ( $\gamma_1^2 = \gamma_2^2$ ), ST ( $\gamma_1^2 = \gamma_2^2$ ) and the SSL ( $\gamma_1^2 = \gamma_2^2$ ) models, along with their corresponding standard errors (SE) calculated via the bootstrap procedure (100 replications). As in Basso *et al.* (2010), we also compare the normal, SN, ST and the SSL models by inspecting some information selection criteria. Comparing the models by looking at the values of the information criteria presented in Table 8, we observe that the SN, ST and the SSL models outperform the normal model, indicating that asymmetric distributions with heavier tails provide a better fit than the normal and the SN distributions. In addition, it appears that the ST model presents a better fit than all other models.

For this dataset we also adjusted the normal, SN, ST and the SSL models without considering the homogeneous nature of the variance parameter, but the ST ( $\gamma_1^2 = \gamma_2^2$ ) model showed the best fit compared to other models. Thus, for brevity we present only the results for the models with homogeneous nature of the scale parameter.

From the clustering point of view, in order to give evidence of the superiority of the method

Table 7: Tone perception data. ML estimation results for fitting several mixture models. SE are the estimated standard errors based on the bootstrap procedure.

| Parameter      | Normal ( $\gamma_1^2 = \gamma_2^2$ ) |        | SN ( $\gamma_1^2 = \gamma_2^2$ ) |        | ST ( $\gamma_1^2 = \gamma_2^2$ ) |        | SSL ( $\gamma_1^2 = \gamma_2^2$ ) |        |
|----------------|--------------------------------------|--------|----------------------------------|--------|----------------------------------|--------|-----------------------------------|--------|
|                | Estimates                            | SE     | Estimates                        | SE     | Estimates                        | SE     | Estimates                         | SE     |
| $\beta_{01}$   | 1.8901                               | 0.0399 | 1.9147                           | 0.0236 | 1.9103                           | 0.5490 | 1.9243                            | 0.7695 |
| $\beta_{11}$   | 0.0572                               | 0.0175 | 0.0433                           | 0.0109 | 0.0354                           | 0.2768 | 0.0383                            | 0.0163 |
| $\beta_{02}$   | -0.0442                              | 0.0648 | 0.1594                           | 0.0642 | 0.0019                           | 0.5492 | 0.2061                            | 0.0855 |
| $\beta_{12}$   | 1.0111                               | 0.0290 | 0.9044                           | 0.0298 | 0.9978                           | 0.2770 | 0.8859                            | 0.3541 |
| $\sigma_1^2$   | 0.0070                               | 0.0008 | 0.0023                           | 0.0003 | 0.0029                           | 0.0008 | 0.0022                            | 0.0010 |
| $\sigma_2^2$   | 0.0070                               | 0.0008 | 0.0793                           | 0.0114 | 0.0001                           | 0.0008 | 0.0541                            | 0.0215 |
| $\lambda_1$    | -                                    | -      | -0.0128                          | 0.0036 | -8.0653                          | 2.3243 | -1.6120                           | 0.6762 |
| $\lambda_2$    | -                                    | -      | 5.7922                           | 0.6347 | -0.7363                          | 2.1018 | 9.2738                            | 3.8378 |
| $\nu$          | -                                    | -      | -                                | -      | 2.0000                           | 0.0439 | 2.0000                            | 0.8752 |
| $\mathbf{p}_1$ | 0.6765                               | 0.0472 | 0.7310                           | 0.0441 | 0.5496                           | 0.0385 | 0.7251                            | 0.2813 |

Table 8: Tone perception data. Some information criteria.

| Information Criteria | Normal ( $\gamma_1^2 = \gamma_2^2$ ) | SN ( $\gamma_1^2 = \gamma_2^2$ ) | ST ( $\gamma_1^2 = \gamma_2^2$ ) | SSL ( $\gamma_1^2 = \gamma_2^2$ ) |
|----------------------|--------------------------------------|----------------------------------|----------------------------------|-----------------------------------|
| log-likelihood       | 106.2549                             | 134.0726                         | <b>201.2834</b>                  | 135.5021                          |
| AIC                  | -198.5097                            | -250.1451                        | <b>-382.5667</b>                 | -251.0041                         |
| BIC                  | -174.4247                            | -220.0387                        | <b>-352.4604</b>                 | -220.8977                         |
| EDC                  | -194.9138                            | -245.6502                        | <b>-378.0718</b>                 | -246.5092                         |
| ICL                  | 4274.6590                            | 6831.2690                        | <b>1299.2110</b>                 | 1318.8930                         |

using the SMSN-MRM family, we carried out a bootstrap experiment with 100 replications. For each replication, the right number of allocations was computed. The mean and standard deviation (SD) of these bootstrap replications are shown in Table 9. Also, we present the associated 95% normal asymptotic confidence intervals (IC). It can be seen that the means are greater and the standard deviations are smaller for the heavy-tailed SMSN-MRM. Figure 6 shows the scatter plots of the data set with the fitted models.

Table 9: Tone perception data. Right allocation analysis through bootstrap procedure for the dataset.

| Fitted Model | Mean of right allocations | SD of right allocations | 95% IC for right allocations |
|--------------|---------------------------|-------------------------|------------------------------|
| Normal       | 130.9500                  | 3.4275                  | [124.0095;137.8005]          |
| SN           | 136.7900                  | 6.8199                  | [123.1502;150.4298]          |
| ST           | 141.4000                  | 2.7303                  | [135.9394;146.8606]          |
| SSL          | 138.6264                  | 3.3071                  | [132.0122;145.2406]          |

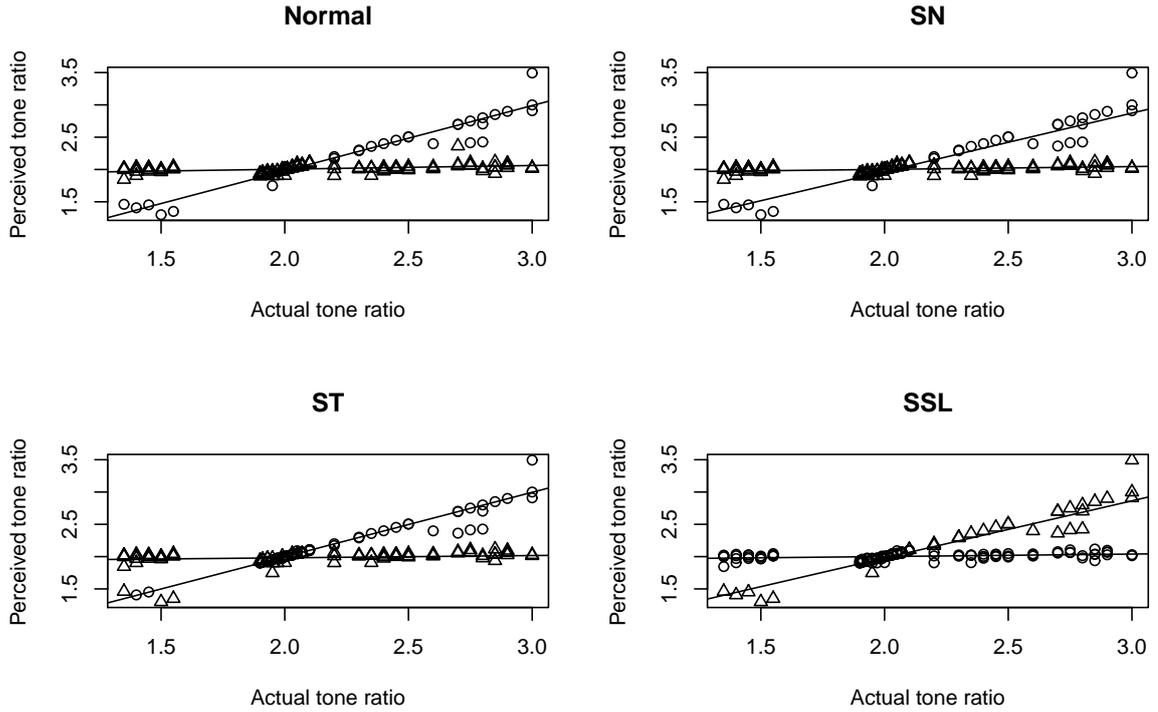


Figure 6: Tone perception data. The scatter plots of the dataset with the fitted models.

## 6 Conclusions

In this work we propose a robust approach to finite mixture of regression models based on scale mixtures of skew-normal distributions. Our proposed model generalizes the recent works of Basso *et al.* (2010), Yao. *et al.* (2014) and Liu & Lin (2014). This robust regression model simultaneously accommodates asymmetry and heavy tails, thus allowing practitioners from different areas to analyze data in an extremely flexible way. An ECME algorithm is developed by exploring the statistical properties of the class considered, which can be easily implemented and coded with existing statistical software such as the R package. Through simulation studies we have shown that this EM-type algorithm gives reasonable estimates in an asymptotic unbiased sense as well as the ability of the FM-SMSN distributions with heavy tails to cluster heterogeneous data. These results indicate that the use of the SMSN-MRM offers a better fit, protection against outliers and more precise inference. The R code is available from us upon request.

This method can be extended to multivariate settings, such as the recent proposals of Galimberti & Soffritti (2014) for mixtures of multivariate Student-t distributions, in order to model, for instance, longitudinal data as discussed in Verbeke & Lesaffre (1996). We intend to pursue this in future research. Another worthwhile task is to develop a fully Bayesian inference via the Markov chain Monte Carlo method.

## Acknowledgements

Víctor H. Lachos was supported by CNPq-Brazil (BPPesq) and São Paulo State Research Foundation (FAPESP). Celso Rômulo Barbosa Cabral was supported by CNPq (BPPesq). Camila Borelli Zeller was supported by CNPq (BPPesq) and Minas Gerais State Research Foundation (FAPEMIG, universal project).

## References

- Andrews, D. F. & Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society, Series B*, **36**, 99–102.
- Arellano-Valle, R. B., Castro, L. M., Genton, M. G., Gómez, H. W. et al. (2008). Bayesian inference for shape mixtures of skewed distributions, with application to regression analysis. *Bayesian Analysis*, **3**(3), 513–539.
- Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, **12**, 171–178.
- Azzalini, A. & Capitanio, A. (2003). Distributions generated and perturbation of symmetry with emphasis on the multivariate skew-t distribution. *Journal of the Royal Statistical Society, Series B*, **61**, 367–389.
- Basso, R. M., Lachos, V. H., Cabral, C. R. B. & Ghosh, P. (2010). Robust mixture modeling based on scale mixtures of skew-normal distributions. *Computational Statistics and Data Analysis*, **54**, 2926–2941.
- Böhning, D. (2000). *Computer-assisted analysis of mixtures and applications. Meta-analysis, disease mapping and others*. Chapman&Hall/CRC, Boca Raton.
- Böhning, D., Seidel, W., Alfó, M., Garel, B., Patilea, V. & Walther, G. (2007). Editorial: Advances in mixture models. *Computational Statistics & Data Analysis*, **51**, 5205–5210.
- Böhning, D., Hennig, C., McLachlan, G. J. & McNicholas, P. D. (2014). Editorial: The 2nd special issue on advances in mixture models. *Computational Statistics & Data Analysis*, **71**, 1–2.
- Branco, M. D. & Dey, D. K. (2001). A general class of multivariate skew-elliptical distributions. *Journal of Multivariate Analysis*, **79**, 99–113.
- Cabral, C. R. B., Lachos, V. H. & Prates, M. O. (2012). Multivariate mixture modeling using skew-normal independent distributions. *Computational Statistics and Data Analysis*, **56**, 126–142.
- Cohen, E. (1984). Some effects of inharmonic partials on interval perception. *Music Perception*, **1**, 323–349.
- Cosslett, S. R. & Lee, L.-F. (1985). Serial correlation in latent discrete variable models. *Journal of Econometrics*, **27**(1), 79–97.
- Dempster, A., Laird, N. & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

- Depraetere, N. & Vandebroek, M. (2014). Order selection in finite mixtures of linear regressions. *Statistical Papers*, **55**, 871–911.
- DeSarbo, W. S. & Cron, W. L. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, **5**, 248–282.
- DeSarbo, W. S., Wedel, M., Vriens, M. & Ramaswamy, V. (1992). Latent class metric conjoint analysis. *Marketing Letters*, **3**(3), 273–288.
- DeVeaux, R. D. (1989). Mixtures of linear regressions. *Computational Statistics and Data Analysis*, **8**(3), 227–245.
- Frühwirth-Schnatter, S. (2006). *Finite mixture and markov switching models*. Springer, New York.
- Galimberti, G. & Soffritti, G. (2014). A multivariate linear regression analysis using finite mixtures of t distributions. *Computational Statistics & Data Analysis*, **71**, 138–150.
- Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the Econometric Society*, pages 357–384.
- Hunter, D. R. & Young, D. S. (2012). Semiparametric mixtures of regressions. *Journal of Non-parametric Statistics*, **24**(1), 19–38.
- Lachos, V. H., Ghosh, P. & Arellano-Valle, R. B. (2010). Likelihood based inference for skew-normal independent linear mixed models. *Statistica Sinica*, **20**, SS-08-045.
- Lindsay, B. G. (1995). *Mixture models: theory geometry and applications*, volume 51. NSF-CBMS Regional Conference Series in Probability and Statistics, Institute of Mathematical Statistics, Hayward.
- Liu, C. & Rubin, D. B. (1994). The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika*, **80**, 267–278.
- Liu, M. & Lin, T.-I. (2014). A skew-normal mixture regression model. *Educational and Psychological Measurement*, **74**, 139–162.
- Liu, M., Hancock, G. R. & Harring, J. R. (2011). Using finite mixture modeling to deal with systematic measurement error: A case study. *Journal of Modern Applied Statistical Methods*, **10**(1), 22.
- McLachlan, G. J. & Krishnan, T. (2008). *The EM algorithm and extensions*. John Wiley & Sons, New Jersey.
- McLachlan, G. J. & Peel, D. (1998). Robust cluster analysis via mixtures of multivariate t-distributions. *Amin, A., Dori, D., Pudil, P., Freeman, H. (Eds.), Lecture Notes in Computer Science*, **1451**, 658–666.
- McLachlan, G. J. & Peel, D. (2000). *Finite mixture models*. Wiley, New York.
- Meng, X. & Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, **81**, 633–648.

- Mengersen, K., Robert, C. P. & Titterton, D. M. (2011). *Mixtures: Estimation and Applications*. John Wiley and Sons.
- Prates, M. O., Lachos, V. & Cabral, C. (2011). mixsmsn: Fitting finite mixture of scale mixture of skew-normal distributions. *R package version 0.2-9*.
- Quandt, R. E. (1972). A new approach to estimating switching regressions. *Journal of the American Statistical Association*, **67**, 306–310.
- Quandt, R. E. & Ramsey, J. B. (1978). Estimating mixtures of normal distributions and switching regressions. *Journal of the American Statistical Association*, **73**(364), 730–738.
- Santana, L., Vilca, F. & Leiva, V. (2011). Influence analysis in skew-birnbaum–saunders regression models and applications. *Journal of Applied Statistics*, **38**, 1633–1649.
- Späth, H. (1979). Algorithm 39 clusterwise linear regression. *Computing*, **22**(4), 367–373.
- Turner, T. R. (2000). Estimating the propagation rate of a viral infection of potato plants via mixtures of regressions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **49**(3), 371–384.
- Verbeke, G. & Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, **91**, 217–221.
- Viele, K. & Tong, B. (2002). Modeling with mixtures of linear regressions. *Statistics and Computing*, **12**(4), 315–330.
- Vilca, F., Santana, L., Leiva, V. & Balakrishnan, N. (2011). Estimation of extreme percentiles in birnbaum–saunders distributions. *Computational Statistics and Data Analysis*, **55**, 1665–1678.
- Vilca, F., Balakrishnan, N. & Zeller, C. B. (2014). Multivariate skew-normal generalized hyperbolic distribution and its properties. *Journal of Multivariate Analysis*, **128**, 73–85.
- Wang, J. & Genton, M. G. (2006). The multivariate skew-slash distribution. *Journal of Statistical Planning and Inference*, **136**, 209–220.
- Wei, G. C. G. & Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man’s data augmentation algorithms. *Journal of the American Statistical Association*, **85**, 699–704.
- Yao, W., Wei, Y. & Yu, C. (2014). Robust mixture regression using the t-distribution. *Computational Statistics and Data Analysis*, **71**, 116–127.
- Zeller, C. B., Lachos, V. H. & Vilca-Labra, F. E. (2011). Local influence analysis for regression models with scale mixtures of skew-normal distributions. *Journal of Applied Statistics*, **38**, 348–363.