Linear Regression Models with Finite Mixtures of Skew Heavy-Tailed Errors

Luis Benites[a], Rocío Maehara[a] and Victor H. Lachos[b]

[a] *Departamento de Estatística, Universidade de São Paulo, Brazil*
[b] *Departamento de Estatística, Universidade Estadual de Campinas, Campinas, São Paulo, Brazil*

Abstract

We consider estimation of regression models whose error terms follow a finite mixture of scale mixtures of skew-normal (SMSN) distributions, a rich class of distributions that contains the skew-normal, skew-t, skew-slash and skew-contaminated normal distributions as proper elements. This approach allows us to model data with great flexibility, accommodating simultaneously multimodality, skewness and heavy tails. We developed a simple EM-type algorithm to perform maximum likelihood (ML) inference of the parameters of the proposed model with closed-form expression at the E-step. Furthermore, the standard errors of the ML estimates can be obtained as a byproduct. The practical utility of the new method is illustrated with the analysis of real dataset and several simulation studies. The proposed algorithm and methods are implemented in the R package `FMsmsnReg()`.

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

1. Introduction

A basic assumption of the linear regression (LR) model is that the error terms follow a normal distribution. However, it is well-known that some phenomena are not always in agreement with this assumption, yielding data having distribution with heavy-tails, skewness or multimodality. These characteristics can be circumvented by data transformations (namely, Box-Cox, etc.), which can render approximate normality with reasonable empirical results. However, some possible drawbacks of these methods are: (i) transformations provide reduced information on the underlying data generation scheme; (ii) component wise transformations may not guarantee joint normality; (iii) parameters may lose interpretability in a transformed scale; and (iv) transformations may not be universal and usually vary with the dataset. Hence, from a practical perspective, there is a need to seek an appropriate theoretical model that avoids data transformation.

Many extensions of this classic model have been proposed to broaden the applicability of the Gaussian linear regression (N-LR) analysis to situations where the Gaussian error term assumption may be

*Address for correspondence: Victor Hugo Lachos Davila, Departamento de Estatistica, IMECC, Universidade Estadual de Campinas, CEP 13083-859, Campinas, São Paulo, Brazil.
Email address: `lbenitesanchez@gmail.com rmaeharaa@gmail.com hlachos@ime.unicamp.br` (Luis Benites[a], Rocío Maehara[a] and Victor H. Lachos[b])

inadequate, such as, the use of the Student-t distribution (Lange et al., 1989), which is appropriate for datasets involving errors with longer than normal tails. Other extensions include the use of the symmetrical class of scale mixtures of normal (SMN) distributions (Andrews and Mallows, 1974; Lange and Sinsheimer, 1993), as discussed in Galea et al. (1997), or even the asymmetrical class of skew-normal (SMSN) distributions proposed by Branco and Dey (2001). However, in practice when nothing is known about the true distribution of the error terms, a linear regression analysis based on any of the above models can be performed using an incorrectly specified model. Furthermore, there can be situations where a single parametric family is unable to provide a satisfactory model for local variations in the observed data. To overcome these problems, solutions that use finite mixture models have been recently proposed. For instance, Bartolucci and Scaccia (2005), Soffritti and Galimberti (2011) and Galimberti and Soffritti (2014) have developed methods for linear regression analysis by assuming a finite mixture of Gaussian (FM-N-LR) and Student-t (FM-T-LR) components for the error terms. A drawback of these proposals is that they are not appropriate when the error terms present, for instance, multimodality, heavy-tails and skewness simultaneously. Thus in this article we propose a mixture model for the random errors based on the class of scale mixtures of skew-normal distributions (FM-SMSN-LR model) by extending the mixture model based on symmetrical distributions.

The class of SMSN distributions, proposed by Branco and 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 results in a flexible class of models for robust estimation in FM-SMSN-LR models since it contains distributions such as the skew-normal distribution and all the symmetric class of scale mixtures of normal distributions defined by Andrews and Mallows (1974). Moreover, the class of SMSN distributions is a rich class that contains proper elements such as the skew- t, skew-slash and skew- contaminated normal distribution. The use of mixture models in the LR context is well-known to deal with different regression functions, the so-called *switching regression* as developed by Zeller et al. (2015). In this paper, instead mixtures of regressions, mixtures are exploited as a convenient semiparametric method, which lies between parametric models and kernel density estimators, to model the unknown distributional shape of the errors. Moreover, the proposed algorithm and methods were implemented in the R package `FMsmsnReg()`.

The remainder of the paper is organized as follows. In Section 2, we briefly discuss some properties of the univariate SMSN family. In Section 3, we present the linear regression model based on SMSN and the related maximum likelihood (ML) estimation. In Section 4, we present the FM-SMSN-LR model, including the EM-type algorithm for ML estimation, and derive the empirical information matrix analytically to obtain the standard errors. In Section 5 and 6, numerical samples using both simulated and real dataset are given to illustrate the performance of proposed model. Finally, some concluding remarks are presented in Section 7.

2. Scale mixtures of skew-normal distributions

Throughout this paper, $X \sim \mathrm{N}(\mu, \sigma^2)$ denotes a random variable X with normal distribution with mean μ and variance σ^2 and $\phi\left(\cdot|\mu, \sigma^2\right)$ denotes its probability density function (pdf). In turn, $\phi(\cdot)$ and $\Phi(\cdot)$ denote the pdf and the cumulative distribution function (cdf) of the standard normal distribution, respectively. In general, we use the traditional convention of denoting a random variable (or a random vector) by an upper-case letter and its realization by the corresponding lower-case letter. Random vectors and matrices are denoted by boldface letters. $\mathbf{X}^\top$ is the transpose of $\mathbf{X}$. $X \perp Y$ indicates that the random variables X and Y are independent.

We start by defining the skew-normal (SN) distribution and then we introduce some useful properties. Thus, as defined by Azzalini (1985), a random variable Z has a skew-normal distribution with location parameter μ, scale parameter σ^2 and skewness parameter λ, denoted by $Z \sim SN(\mu, \sigma^2, \lambda)$, if its pdf is given by:

$$\phi_{SN}(z|\mu, \sigma^2, \lambda) = 2\phi(z|\mu, \sigma^2)\Phi\left(\frac{\lambda(z-\mu)}{\sigma}\right). \tag{1}$$

Next we present the stochastic representations of a random variable with SN distribution, which is useful to generate random samples and to obtain the moments and other related properties. If $Z \sim SN(\mu, \sigma^2, \lambda)$, then a convenient stochastic representation is given by:

$$Z = \mu + \Delta|T_0| + \Gamma^{1/2}T_1, \tag{2}$$

where $\Delta = \sigma\delta$, $\Gamma = (1-\delta^2)\sigma^2$, $T_0 \perp T_1$ and $|\cdot|$ denotes the absolute value.

The relation between the SMSN class and the SN distribution is given in the next definition

Definition 1. *A random variable Y has a SMSN distribution with location parameter μ, scale parameter σ^2 and skewness parameter λ, denoted by SMSN$(\mu, \sigma^2, \lambda; H)$, if it has the following stochastic representation:*

$$Y = \mu + \kappa^{1/2}(U)Z, , \quad U \perp Z, \tag{3}$$

where $Z \sim SN(0, \sigma^2, \lambda)$, U is a positive random variable with cdf $H(\,\cdot\,|\boldsymbol{\nu})$ indexed by a scalar or vector parameter $\boldsymbol{\nu}$ and $\kappa(u)$ is a positive function of u.

The random variable U is known as *the scale factor* and its cdf $H(\,\cdot\,|\boldsymbol{\nu})$ is called the *mixing distribution function*. Note that when $\lambda = 0$, the SMSN family reduces to the symmetric class of scale mixtures of normal independent (SMN) distributions.

Although we can deal with any κ function, in this paper we restrict our attention to the case where $\kappa(u) = 1/u$, since it leads to good mathematical properties. Given $U = u$, we have that $Y|U = u \sim SN(\mu, u^{-1}\sigma^2, \lambda)$. Thus, the density of Y is given by

$$\phi_{SMSN}(y|\mu, \sigma^2, \lambda, \boldsymbol{\nu}) = 2\int_0^\infty \phi(y|\mu, u^{-1}\sigma^2)\Phi\left(\frac{u^{1/2}\lambda(y-\mu)}{\sigma}\right) dH(u|\boldsymbol{\nu}). \tag{4}$$

When H is degenerate, with $u = 1$, we obtain the $SN(\mu, \sigma^2, \lambda)$ distribution and when $\lambda = 0$, the SMSN distributions reduces to the class of scale–mixtures of the normal (SMN) distribution represented by the pdf $\phi_{SMN}(y|\mu, \sigma^2, \boldsymbol{\nu}) = \int_0^\infty \phi(y|\boldsymbol{\mu}, u^{-1}\boldsymbol{\Sigma})dH(u|\boldsymbol{\nu})$.

Another important result that will be useful in implementing the EM algorithm is given next. The statements of these results can be found in Basso et al. (2010)

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

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

where $W_\Phi(.) = \dfrac{\phi(.)}{\Phi(.)}$, $A = \lambda y_0$, with $y_0 = \dfrac{(y-\mu)}{\sigma}$, f_0 is the pdf of $Y_0 \sim SMN(\mu, \sigma^2; H)$, $U_y \overset{d}{=} U|Y_0 = y$ and $K_r = E[U^{-r/2}]$, $r = 1, 2, \ldots$

Some particular cases of the SMSN family of distributions are given next

- *The skew–t distribution with ν degrees of freedom.* In this case, the density of Y takes the form

$$\phi_T(y|\mu, \sigma^2, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}\sigma} \left(1 + \frac{d}{\nu}\right)^{-\frac{\nu+1}{2}} T\left(\sqrt{\frac{\nu+1}{d+\nu}} A|\nu+1\right), \quad y \in \mathbb{R}, \tag{5}$$

  where $d = (y-\mu)^2/\sigma^2$, $A = \lambda(y-\mu)/\sigma$ and $T(\cdot|\nu)$ denotes the distribution function of the standard Student–t distribution, with location zero, scale one and ν degrees of freedom, namely $t(0, 1, \nu)$. We use the notation $Y \sim ST(\mu, \sigma^2, \lambda; \nu)$. It is known that if $\nu \to \infty$, the skew normal distribution is obtained. In the special case where $\lambda = 0$, we obtain the cdf of the ordinary Student-t distribution with ν degrees of freedom. Although the ST distribution has nice properties (see, Azzalini and Genton, 2008), there are some inferential problems. For instance, it can be shown that the observed information matrix is singular when $\lambda = 0$ and $\nu \to \infty$. In this article, we only consider the case that the degrees of freedom ν is finite.

- *The skew–slash distribution.* It is denoted by $Y \sim SSL(\mu, \sigma^2, \lambda; \nu)$ and the associated density is given by

$$\phi_{SL}(y|\mu, \sigma^2, \nu) = 2\nu \int_0^1 u^{\nu-1} \phi(y|\mu, u^{-1}\sigma^2) \Phi(u^{1/2} A) du, \quad y \in \mathbb{R}. \tag{6}$$

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

- *The skew contaminated normal distribution.* We denote it by $Y \sim SCN(\mu, \sigma^2, \lambda; \nu, \gamma)$. Its density is given by

$$\phi_{SCN}(y|\mu, \sigma^2, \boldsymbol{\nu}) = 2\{\nu \phi(y|\mu, \gamma^{-1}\sigma^2) \Phi(\gamma^{1/2} A) + (1-\nu)\phi(y|\mu, \sigma^2)\Phi(A)\}, \ \ \nu, \gamma \in (0, 1).$$

  The parameters ν and γ can be interpreted as the proportion of outliers and a scale factor, respectively. The skew contaminated normal distribution reduces to the skew-normal distribution when $\gamma = 1$.

3. The SMSN linear regression model

The linear regression model based on SMSN distributions–hereafter SMSN-LR model– is defined as:

$$Y_i = \mathbf{x}_{ic}^\top \boldsymbol{\beta}_c + \varepsilon_i, \;\; i = 1, \ldots, n, \tag{7}$$

where the Y_i are the responses, $\mathbf{x}_{ic} = (1, x_{i1}, \ldots, x_{ip})^\top$ is a vector of explanatory variable values of dimension $(p+1) \times 1$, $\boldsymbol{\beta}_c = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ is the regression parameter vector and the random errors $\varepsilon_i \overset{iid}{\sim} SMSN(-\sqrt{\frac{2}{\pi}} K_1 \Delta, \sigma^2, \lambda; H)$, with $K_r = E\{U^{-r/2}\}$, $r = 1, 2 \ldots$, $\Delta = \sigma\delta$ and $\delta = \lambda/(1+\lambda^2)^{1/2}$, which corresponds to the regression model where the error distribution has mean zero and hence the regression parameters are all comparable. From Proposition 1, we have that

$$E[Y_i] = \mathbf{x}_{ic}^\top \boldsymbol{\beta}_c, \; Var[Y_i] = K_2 \sigma^2 - b^2 \Delta^2,$$

where $b = -\sqrt{\frac{2}{\pi}} K_1$ and $Y_i \sim SMSN(\mathbf{x}_{ic}^\top \boldsymbol{\beta}_c + b\Delta, \sigma^2, \lambda; H)$, for $i = 1, \ldots, n$.

3.1. Parameter estimation via the EM-algorithm

In this subsection we develop an Expectation-Maximization (EM) algorithm (Dempster et al., 1977) for ML estimation of the parameters of the SMSN-LR model. In order to do this, we first represent the SMSN-LR model in an incomplete data framework using Lemma 2 given in Basso et al. (2010). We consider the following hierarchical representation for Y_i:

$$Y_i | T_i = t_i, U_i = u_i \;\; \overset{ind}{\sim} \;\; N(\mathbf{x}_{ic}^\top \boldsymbol{\beta}_c + \Delta t_i, U_i^{-1}\Gamma), \tag{8}$$

$$T_i | U_i = u_i \;\; \overset{ind}{\sim} \;\; TN(b, u_i^{-1}; (b, \infty)), \tag{9}$$

$$U_i \;\; \overset{iid}{\sim} \;\; H(.|\boldsymbol{\nu}), \tag{10}$$

where $\Gamma = (1-\delta^2)\sigma^2$, $\Delta = \sigma\delta$ and $TN(r, s; (a, b))$ denotes the univariate normal distribution $(N(r, s))$, truncated on the interval (a, b). A useful straightforward result is that the conditional distribution of T_i given y_i and u_i is $TN(\mu_{Ti} + b, u_i^{-1} M_T^2; (b, \infty))$, with

$$M_T^2 = \frac{\Gamma}{\Delta^2 + \Gamma}, \;\; \mu_{T_i} = \frac{\Delta}{\Delta^2 + \Gamma}(y_i - \mathbf{x}_{ic}^\top \boldsymbol{\beta}_c - \Delta b).$$

Now we proceed for the E-step of the algorithm. To represent the estimator of the parameter $\xi = g(\boldsymbol{\theta})$, we will use the general notation $\widehat{\xi} = g(\widehat{\boldsymbol{\theta}})$, where $g(\cdot)$ is a generic function of $\boldsymbol{\theta} = (\boldsymbol{\beta}_c^\top, \sigma^2, \lambda)^\top$. Thus, let $\mathbf{y} = (y_1, \ldots, y_n)^\top$, $\mathbf{t} = (t_1, \ldots, t_n)^\top$ and $\mathbf{u} = (u_1, \ldots, u_n)^\top$. It follows that the complete log-likelihood function associated with $(\mathbf{y}, \mathbf{t}, \mathbf{u})$ is given by:

$$\ell_c(\boldsymbol{\theta} | \mathbf{y}, \mathbf{t}, \mathbf{u}) = c - \frac{n}{2} \log \Gamma - \frac{1}{2\Gamma} \sum_{i=1}^{n} u_i (y_i - \mathbf{x}_{ic}^\top \boldsymbol{\beta}_c - \Delta t_i)^2, \tag{11}$$

where c is a constant that is independent of $\boldsymbol{\theta}$. Letting $\widehat{u}_i = E[U_i | \boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, y_i]$, $\widehat{ut}_i = E[U_i T_i | \boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, y_i]$, $\widehat{ut_i^2} = E[U_i T_i^2 | \boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, y_i]$ and using known properties of conditional expectation we obtain:

$$\widehat{ut}_i = \widehat{u}_i(\widehat{\mu}_{Ti} + b) + \widehat{M}_T \widehat{\tau}_{1_i}, \quad \widehat{ut_i^2} \;\; = \;\; \widehat{u}_i(\widehat{\mu}_{T_i} + b)^2 + \widehat{M}_T^2 + \widehat{M}_T(\widehat{\mu}_{T_i} + 2b)\widehat{\tau}_{1_i},$$

where $\widehat{\tau}_{1_i} = E\left[U_i^{1/2} W_\Phi\left(\frac{U_i^{1/2}\widehat{\mu}_{T_i}}{\widehat{M}_T}\right)|\widehat{\boldsymbol{\theta}}, y_i\right]$. In each step, the conditional expectations $\widehat{u}_i = \widehat{u}_{1_i}$ and $\widehat{\tau}_{1_i}$ can be easily derived from the results given in Basso et al. (2009, Subsection 2.1). For the skew–t, skew-slash and skew contaminated normal distribution we have computationally attractive expressions that can be easily implemented (see also Lachos et al., 2010).

These expressions are quite useful in implementing the M-step, which consists of maximizing the expected complete data function or the $Q-$function over $\boldsymbol{\theta}$, given by:

$$\begin{aligned} Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) &= E[\ell_c(\boldsymbol{\theta})|\mathbf{y}, \widehat{\boldsymbol{\theta}}^{(k)}] = c - \frac{n}{2}\log(\Gamma) - \frac{1}{2\Gamma}\sum_{i=1}^{n}\Big[\widehat{u}_i^{(k)}\left(y_i - \mathbf{x}_{ic}^\top\boldsymbol{\beta}_c\right)^2 \\ &- 2\Delta(y_i - \mathbf{x}_{ic}^\top\boldsymbol{\beta}_c)\widehat{ut}_i^{(k)} + \Delta^2\widehat{ut_i^2}^{(k)}\Big], \end{aligned}$$

where $\widehat{\boldsymbol{\theta}}^{(k)}$ is an updated value of $\widehat{\boldsymbol{\theta}}$.

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 *ECM algorithm* (Meng and Rubin, 1993). Next, we describe this EM-type algorithm (ECM) for maximum likelihood estimation of the parameters of the SMSN-LR model.

E-step: Given a current estimate $\widehat{\boldsymbol{\theta}}^{(k)}$, compute $\widehat{u}_i^{(k)}$ $\widehat{ut}_i^{(k)}$, $\widehat{ut_i^2}^{(k)}$, for $i = 1, \ldots, n$.

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

$$\widehat{\boldsymbol{\beta}}_c^{(k+1)} = \left(\sum_{i=1}^{n}\widehat{u}_i^{(k)}\mathbf{x}_{ic}\mathbf{x}_{ic}^\top\right)^{-1}\left(\sum_{i=1}^{n}\widehat{u}_i^{(k)}y_i\mathbf{x}_{ic} - \widehat{\Delta}^{(k)}\mathbf{x}_{ic}\widehat{ut}_i^{(k)}\right), \tag{12}$$

$$\widehat{\Delta}^{(k+1)} = \frac{\sum_{i=1}^{n}(y_i - \mathbf{x}_{ic}^\top\widehat{\boldsymbol{\beta}}_c^{(k+1)})\widehat{ut}_i^{(k)}}{\sum_{i=1}^{n}\widehat{ut}_i^{2(k)}}, \tag{13}$$

$$\widehat{\Gamma}^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{u}_i^{(k)}\left(y_i - \mathbf{x}_{ic}^\top\widehat{\boldsymbol{\beta}}_c^{(k+1)}\right)^2 - 2\widehat{\Delta}^{(k+1)}(y_i - \mathbf{x}_{ic}^\top\widehat{\boldsymbol{\beta}}_c^{(k+1)})\widehat{ut}_i^{(k)} + \widehat{\Delta}^{2(k+1)}\widehat{ut_i^2}^{(k)}\right]. \tag{14}$$

CML-step: Update $\widehat{\boldsymbol{\nu}}^{(k)}$ by maximizing the current marginal log-likelihood function, obtaining

$$\widehat{\boldsymbol{\nu}}^{(k+1)} = \text{argmax}_{\boldsymbol{\nu}}\sum_{i=1}^{n}\log(\phi_{SMSN}(y_i|\mathbf{x}_{ic}^\top\widehat{\boldsymbol{\beta}}_c^{(k+1)} + b\widehat{\Delta}^{(k+1)}, (\widehat{\sigma}^2)^{(k+1)}, \widehat{\lambda}^{(k+1)}, \boldsymbol{\nu})), \tag{15}$$

where $\phi_{SMSN}(\cdot|\boldsymbol{\theta})$ is the SMSN density defined in (4). Note that $\widehat{\sigma}^{2(k+1)}$ and $\widehat{\lambda}^{(k+1)}$ can be recovered by using the fact that $\lambda = \Delta/\sqrt{\Gamma}$ and $\sigma^2 = \Delta^2 + \Gamma$.

This process is iterated until a suitable convergence rule is satisfied, e.g. if $|\widehat{\boldsymbol{\theta}}^{(k+1)} - \widehat{\boldsymbol{\theta}}^{(k)}|$ is sufficiently small, or until some distance involving two successive evaluations of the actual log-likelihood $\ell(\boldsymbol{\theta})$, like $|\ell(\widehat{\boldsymbol{\theta}}^{(k+1)}) - \ell(\widehat{\boldsymbol{\theta}}^{(k)})|$ or $|\ell(\widehat{\boldsymbol{\theta}}^{(k+1)})/\ell(\widehat{\boldsymbol{\theta}}^{(k)}) - 1|$, is small enough.

4. The linear regression model with FM–SMSN errors

4.1. The model

The linear regression model with FM–SMSN errors- hereafter FM-SMSN-LR model- is defined by considering that the random error ϵ_i, defined in (7), follows a g-component mixture of SMSN densities

given by:

$$f(\epsilon_i) = \sum_{j=1}^{g} p_j \phi_{SMSN}(y_i|\mu_j + b\Delta_j, \sigma_j^2, \lambda_j, \boldsymbol{\nu}_j),\ p_j \geq 0,\ \sum_{j=1}^{g} p_j = 1,\ i = 1, \ldots, n, \quad j = 1, \ldots, g, \tag{16}$$

where the $\mu_j's$ satisfy the identifiability constraint $\sum_{j=1}^{g} p_j \mu_j = 0$.

We write $\mu_{ij} = \mathbf{x}_{ic}^\top \boldsymbol{\beta}_c + \mu_j = \beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta} + \mu_j = \vartheta_j + \mathbf{x}_i^\top \boldsymbol{\beta}$, where $\vartheta_j = \beta_0 + \mu_j$, $\mathbf{x}_i^\top = (x_{i1}, \ldots, x_{ip})^\top$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$. It follows that the density of the ith observation, y_i, is

$$f(y_i|\boldsymbol{\Theta}) = \sum_{j=1}^{g} p_j \phi_{SMSN}(y_i|\mu_{ij} + b\Delta_j, \sigma_j^2, \lambda_j, \boldsymbol{\nu}_j), \tag{17}$$

where $\mu_{ij} = \mathbf{x}_i^\top \boldsymbol{\beta} + \vartheta_j$; $p_1, \ldots, p_g$ are the mixing probabilities and $\boldsymbol{\Theta} = (\boldsymbol{\beta}^\top, (p_1, \ldots, p_{g-1})^\top, \boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_g^\top)^\top$ is the vector with all parameters, with $\boldsymbol{\theta}_j = (\vartheta_j, \sigma_j^2, \lambda_j, \boldsymbol{\nu}_j^\top)^\top$ is the specific vector of parameters for the component j. Concerning the parameter $\boldsymbol{\nu}_j$ of the mixing distribution $H(.|\boldsymbol{\nu}_j)$, for $j = 1, \ldots, g$, it is worth noting that it can be a vector of parameters, e.g., the contaminated normal distribution. Thus, for computational convenience we assume that $\boldsymbol{\nu}_1 = \ldots = \boldsymbol{\nu}_g = \boldsymbol{\nu}$ and in this case, $\boldsymbol{\Theta} = (\boldsymbol{\beta}^\top, (p_1, \ldots, p_{g-1})^\top, \boldsymbol{\nu}^\top, \boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_g^\top)^\top$, with $\boldsymbol{\theta}_j = (\vartheta_j, \sigma_j^2, \lambda_j)^\top$. This strategy it works very well in the empirical studies that we have conducted and greatly simplifies the optimization problem.

For each i and j, consider the latent indicator variable Z_{ij} such that:

$$P(Z_{ij} = 1) = 1 - P(Z_{ij} = 0) = p_j, \quad \sum_{j=1}^{g} p_j = 1 \quad \text{and} \quad y_i|Z_{ij} = 1 \sim SMSN(\mu_{ij} + b\Delta_j, \sigma_j^2, \lambda_j; H(\boldsymbol{\nu})). \tag{18}$$

Note that by integrating out $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ig})^\top$ we obtain the marginal density (16). $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ are independent random vectors, each one having a multinomial distribution with probability function

$$f(\mathbf{z}_i) = p_1^{z_{i1}} p_2^{z_{i2}} ... (1 - p_1 - \ldots - p_{g-1})^{z_{ig}},$$

which we denote by $\mathbf{Z}_i \sim M(1; p_1 \ldots, p_g)$.

These latent vectors appear in the hierarchical representation given next, which is used to build the ECME algorithm. From (18) along with Definition 1, the FM-SMSN-LR model can be represented as

$$y_i|u_i, t_i, Z_{ij} = 1 \overset{ind}{\sim} N(\mu_{ij} + \Delta_j t_i, u_i^{-1}\Gamma_j), \tag{19}$$

$$T_i|u_i, Z_{ij} = 1 \overset{ind}{\sim} NT(b, u_i^{-1}, (b, \infty)), \tag{20}$$

$$U_i|Z_{ij} = 1 \overset{ind}{\sim} H(u_i; \boldsymbol{\nu}) \tag{21}$$

and

$$\mathbf{Z}_i \overset{iid}{\sim} M(1; p_1 \ldots, p_g), \qquad i = 1, \ldots, n,\ j = 1, \ldots, g, \tag{22}$$

where

$$\Gamma_j = (1 - \delta_j^2)\sigma_j^2, \quad \Delta_j = \sigma_j \delta_j, \quad \text{and } \delta_j = \frac{\lambda_j}{\sqrt{1 + \lambda_j^2}}. \tag{23}$$

4.2. Parameter estimation via the ECME algorithm

In this subsection we show how to implement the ECME algorithm for ML estimation of the parameters of a FM-SMSN-LR model. By using (19)-(22), we have that the complete-data log-likelihood function is

$$\ell_c(\boldsymbol{\Theta}) = c + \sum_{i=1}^{n}\sum_{j=1}^{g} Z_{ij}\left(\log(p_j) - \frac{1}{2}\log\Gamma_j - \frac{u_i}{2\Gamma_j}(y_i - \mu_{ij} - \Delta_j t_i)^2 + \log(h(u_i|\boldsymbol{\nu})) + \log(NT(t_i|b, u_i^{-1}, (b,\infty))\right),$$

where c is a constant that is independent of the parameter vector $\boldsymbol{\Theta}$. Defining the following quantities

$$\begin{aligned}
\widehat{z}_{ij} &= \mathrm{E}[Z_{ij}|\widehat{\boldsymbol{\Theta}}, y_i],\\
\widehat{s}_{1ij} &= \mathrm{E}[Z_{ij}U_i|\widehat{\boldsymbol{\Theta}}, y_i],\\
\widehat{s}_{2ij} &= \mathrm{E}[Z_{ij}U_iT_i|\widehat{\boldsymbol{\Theta}}, y_i]\\
&\text{and}\\
\widehat{s}_{3ij} &= \mathrm{E}[Z_{ij}U_iT_i^2|\widehat{\boldsymbol{\Theta}}, y_i]
\end{aligned}$$

and using known properties of conditional expectation, we obtain:

$$\begin{aligned}
\widehat{z}_{ij} &= \frac{\widehat{p}_j\psi(y_i;\widehat{\boldsymbol{\theta}}_j)}{\sum_{j=1}^{g}\widehat{p}_j\psi(y_i;\widehat{\boldsymbol{\theta}}_j)},\\
\widehat{s}_{1ij} &= \widehat{z}_{ij}\widehat{u}_{ij},\\
\widehat{s}_{2ij} &= \widehat{z}_{ij}\left(\widehat{u}_{ij}\widehat{\mu}_{T_{ij}} + \widehat{M}_{T_j}\widehat{\tau}_{1_{ij}}\right)\\
&\text{and}\\
\widehat{s}_{3ij} &= \widehat{z}_{ij}\left(\widehat{u}_{ij}\widehat{\mu}_{T_{ij}}^2 + \widehat{M}_{T_j}^2 + \widehat{M}_{T_j}(\widehat{\mu}_{T_{ij}} + b)\widehat{\tau}_{1_{ij}}\right),
\end{aligned} \tag{24}$$

where

$$\begin{aligned}
\widehat{\tau}_{1_{ij}} &= E\left[U_i^{1/2}W_{\Phi_1}\left(\frac{U_i^{1/2}\widehat{\mu}_{T_{ij}})}{\widehat{M}_{T_j}}\right) \mid \widehat{\boldsymbol{\Theta}}, y_i, Z_{ij} = 1\right],\\
\widehat{M}_{T_j}^2 &= \frac{\Gamma_j}{\Gamma_j + \Delta_j^2},\\
\widehat{\mu}_{T_{ij}} &= b + \frac{\Delta_j}{\Gamma_j + \Delta_j^2}(y_i - \mu_{ij} - \Delta b)\\
&\text{and}\\
\widehat{u}_{ij} &= E[U_j|\widehat{\boldsymbol{\Theta}}, y_i, Z_{ij} = 1], \qquad i = 1,\ldots,n, \quad j = 1,\ldots,g.
\end{aligned}$$

Once again, in each step, the conditional expectations $\widehat{u}_{ij}$ and $\widehat{\tau}_{1_{ij}}$ can be easily derived from the results given in Basso et al. (2010). Thus, the $Q-$function is given by

$$\begin{aligned}
Q(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}}^{(k)}) &= c + \sum_{i=1}^{n}\sum_{j=1}^{g}\left(\widehat{z}_{ij}^{(k)}(\log(p_j) - \frac{1}{2}\log|\Gamma_j|) - \frac{1}{2\Gamma_j}\left(\widehat{s}_{1ij}^{(k)}(y_i - \mu_{ij})^2 - 2(y_i - \mu_{ij})\Delta_j\widehat{s}_{2ij}^{(k)}\right.\right.\\
&\quad \left.\left. + \Delta_j^2\widehat{s}_{3ij}^{(k)}\right) + E[Z_{ij}\log(h(U_i|\boldsymbol{\nu}))|\widehat{\boldsymbol{\Theta}}^{(k)}, y_i] + E[Z_{ij}\log(TN(T_i|b, u_i^{-1}, (b,\infty)))|\widehat{\boldsymbol{\Theta}}^{(k)}, y_i]\right).
\end{aligned}$$

Also, we have adopted the same strategy used in Subsection 3.1 to update the estimate of $\boldsymbol{\nu}$, by direct maximization of the marginal log-likelihood, circumventing the computation of $\widehat{s}_{4ij} = E[Z_{ij}\log(h(U_i|\boldsymbol{\nu}))|\widehat{\boldsymbol{\Theta}}, y_i]$ and $\widehat{s}_{5ij} = E[Z_{ij}\log(TN(T_i|b, u_i^{-1}, (b,\infty)))|\widehat{\boldsymbol{\Theta}}^{(k)}, y_i]$.

Thus, the ECME algorithm for maximum likelihood estimation of $\boldsymbol{\Theta}$ is defined as follows:
E-step: Given a current estimate $\widehat{\boldsymbol{\Theta}}^{(k)}$, compute $\widehat{z}_{ij}$, $\widehat{s}_{1ij}$, $\widehat{s}_{2ij}$, $\widehat{s}_{3ij}$, for $i = 1, \ldots, n$ and $j = 1, \ldots, g$.
CM-steps: Update $\widehat{\boldsymbol{\Theta}}^{(k)}$ by maximizing $Q(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}}^{(k)}) = E[\ell_c(\boldsymbol{\Theta})|\mathbf{y}, \widehat{\boldsymbol{\Theta}}^{(k)}]$ over $\boldsymbol{\Theta}$, which leads to the following closed-form expressions:

$$\begin{aligned}
\widehat{p}_j^{(k+1)} &= n^{-1}\sum_{i=1}^{n}\widehat{z}_{ij}^{(k)}, \\
\widehat{\vartheta}_j^{(k+1)} &= \frac{\sum_{i=1}^{n}[\widehat{s}_{1ij}^{(k)}(y_i - \mathbf{x}_i^\top\widehat{\boldsymbol{\beta}}) - \widehat{\Delta}_j\widehat{s}_{2ij}^{(k)}]}{\sum_{i=1}^{n}\widehat{s}_{1ij}^{(k)}}, \\
\widehat{\boldsymbol{\beta}}^{(k+1)} &= \left(\sum_{i=1}^{n}\sum_{j=1}^{g}\frac{\widehat{s}_{1ij}^{(k)}\mathbf{x}_i\mathbf{x}_i^\top}{\widehat{\Gamma}_j}\right)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{g}\frac{1}{\widehat{\Gamma}_j}[\widehat{s}_{1ij}^{(k)}(y_i - \widehat{\vartheta}_j^{(k)}) - \widehat{\Delta}_j\widehat{s}_{2ij}^{(k)}]\mathbf{x}_i, \\
\widehat{\Delta}_j^{(k+1)} &= \frac{\sum_{i=1}^{n}(y_i - \widehat{\mu}_{ij}^{(k)})\widehat{s}_{2ij}^{(k)}}{\sum_{i=1}^{n}\widehat{s}_{3ij}^{(k)}}
\end{aligned}$$

and

$$\widehat{\Gamma}_j^{(k+1)} = \sum_{i=1}^{n}\left(\widehat{s}_{1ij}^{(k)}(y_i - \widehat{\mu}_{ij}^{(k)})^2 - 2(y_i - \widehat{\mu}_{ij}^{(k)})\widehat{\Delta}_j s_{2ij}^{(k)} + \widehat{\Delta}_j^2\widehat{s}_{3ij}^{(k)}\right)/\sum_{i=1}^{n}\widehat{z}_{ij}^{(k)}. \tag{25}$$

CML-step: Update $\widehat{\boldsymbol{\nu}}^{(k)}$ by maximizing the current marginal log-likelihood function, obtaining

$$\boldsymbol{\nu}^{(k+1)} = \text{argmax}_{\boldsymbol{\nu}}\sum_{i=1}^{n}\log\left(\sum_{j=1}^{g}p_j\phi_{SMSN}\left(y_i|\mu_{ij}^{(k+1)} + b(\boldsymbol{\nu})\Delta_j^{(k+1)}, \sigma_j^{2(k+1)}, \lambda_j^{(k+1)}, \boldsymbol{\nu}\right)\right). \tag{26}$$

We can also obtain a estimative of β_0 as

$$\widehat{\beta}_0^{(k+1)} = \sum_{j=1}^{g}\widehat{p}_j^{(k+1)}\widehat{\vartheta}_j^{(k+1)}$$

and for μ_j, $j = 1, \ldots, g$, as

$$\widehat{\mu}_j = \widehat{\vartheta}_j - \widehat{\beta}_0.$$

This process is iterated until a suitable convergence rule is satisfied, e.g., $|\ell(\widehat{\boldsymbol{\Theta}}^{(k+1)}) - \ell(\widehat{\boldsymbol{\Theta}}^{(k)})|$ or $|\ell(\widehat{\boldsymbol{\Theta}}^{(k+1)})/\ell(\widehat{\boldsymbol{\Theta}}^{(k)}) - 1|$, is small enough.

A usual criticism is that EM–type procedures tend to get stuck in local modes. A convenient way to avoid such limitations is to try several EM iterations with a variety of starting values. If there are several modes, one can find the global mode by comparing their relative masses and log-likelihood values. We suggest the following strategy. For β_0 and $\boldsymbol{\beta}$ use the ordinary least-square (OLS) estimate. Initial values for p_j, μ_j, σ_j^2 and λ_j, $j = 1, \ldots, g$, are obtained by fitting the mixture model (16) to the OLS residuals, which can be provided through the R library `mixsmsn()` (Prates et al., 2013).

4.3. Model selection

Because there is no universal criterion for mixture model selection, we chose three criteria to compare the models considered in this work. These are the Akaike information criterion (AIC) (Akaike, 1974), the Bayesian information criterion (BIC) (Schwarz, 1978) and the efficient determination criterion (EDC) (Bai et al., 1989). Like the more popular AIC and BIC, EDC has the form $-2\ell(\widehat{\boldsymbol{\theta}}) + \rho c_n$, where $\ell(\boldsymbol{\theta})$ is the actual log-likelihood, ρ 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 $c_n = 0.2\sqrt{n}$ for EDC, a proposal that was considered in Basso et al. (2010) and Cabral et al. (2012a). We have $c_n = 2$ for AIC, $c_n = \log n$ for BIC, where n is the sample size.

5. Approximated standard errors

A simple way of obtaining the standard errors of ML estimates of mixture model parameters is to approximate the asymptotic covariance matrix of $\widehat{\boldsymbol{\Theta}}$ by the inverse of the observed information matrix, where $\boldsymbol{\Theta} = (\boldsymbol{\beta}^\top, (p_1, \ldots, p_{g-1})^\top, \boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_g^\top)^\top$ with $\boldsymbol{\theta}_j = (\vartheta_j, \sigma_j^2, \lambda_j)^\top$, $j = 1, \ldots, g$. Let $\mathbf{I}_o(\boldsymbol{\Theta}) = -\partial^2\ell(\boldsymbol{\Theta})/\partial\boldsymbol{\Theta}\partial\boldsymbol{\Theta}^\top$ be the observed information matrix, where $\ell(\boldsymbol{\Theta})$ is the observed log-likelihood function in (17). In this work we use the alternative method suggested by Basford et al. (1997) and parameterization, which consists of approximating the inverse of the covariance matrix by:

$$\mathbf{I}_o(\widehat{\boldsymbol{\Theta}}) = \sum_{i=1}^{n} \widehat{\mathbf{s}}_i \widehat{\mathbf{s}}_i^\top, \quad \text{where} \quad \widehat{\mathbf{s}}_i = \frac{\partial}{\partial \boldsymbol{\Theta}} \log \left(\sum_{j=1}^{g} \phi_{SMSN}(y_i | \mu_{ij} + b\Delta_j, \sigma_j^2, \lambda_j, \boldsymbol{\nu}) \right) \Bigg|_{\boldsymbol{\Theta}=\widehat{\boldsymbol{\Theta}}}, \tag{27}$$

where

$$\widehat{\mathbf{s}}_i = (\widehat{s}_{i,\boldsymbol{\beta}}, \widehat{s}_{i,p_1}, \ldots, \widehat{s}_{i,p_{g-1}}, \widehat{s}_{i,\vartheta_1}, \ldots, \widehat{s}_{i,\vartheta_g}, , \widehat{s}_{i,\sigma_1^2}, \ldots, \widehat{s}_{i,\sigma_g^2}, \widehat{s}_{i,\lambda_1}, \ldots, \widehat{s}_{i,\lambda_g})^\top.$$

Expressions for the elements $\widehat{s}_{i,\boldsymbol{\beta}}, \widehat{s}_{i,p_j}, \widehat{s}_{i,\vartheta_j}, \widehat{s}_{i,\sigma_j^2}, \widehat{s}_{i,\lambda_j}$, $j = 1, \ldots, g$, are given as follows:

$$\begin{aligned} \widehat{s}_{i,\boldsymbol{\beta}} &= \frac{\sum_{j=1}^{G} p_j D_{\boldsymbol{\beta}}(y_i; \boldsymbol{\Theta}_j)}{f(y_i; \boldsymbol{\Theta})}, \quad \widehat{s}_{i,\vartheta_j} = \frac{p_r D_{\vartheta_j}(y_i; \boldsymbol{\Theta}_j)}{f(y_i; \boldsymbol{\Theta})}, \\ \widehat{s}_{i,\sigma_j^2} &= \frac{p_r D_{\sigma_j^2}(y_i; \boldsymbol{\Theta}_j)}{f(y_i; \boldsymbol{\Theta})}, \widehat{s}_{i,\lambda_j} = \frac{p_r D_{\lambda_j}(y_i; \boldsymbol{\Theta}_j)}{f(y_i; \boldsymbol{\Theta})} \end{aligned}$$

and

$$\widehat{s}_{i,p_j} = \frac{1}{f(y_i; \boldsymbol{\Theta})} \left(\phi_{SMSN}(y_i | \mu_{ij} + b\Delta_j, \sigma_j^2, \lambda_j, \boldsymbol{\nu}) - \phi_{SMSN}(y_i | \mu_{ig} + b\Delta_g, \sigma_g^2, \lambda_g, \boldsymbol{\nu}) \right),$$

where

$$D_{\vartheta_j}(y_i; \boldsymbol{\theta}_j) = \frac{\partial}{\partial \vartheta_j} (\phi_{SMSN}(y_i | \mu_{ij} + b\Delta_j, \sigma_j^2, \lambda_j, \boldsymbol{\nu})).$$

Let us define

$$I_{ij}^{\Phi}(w) = \int_0^\infty \kappa^{-w}(u_i) \exp\{-\frac{1}{2}\kappa^{-1}(u_i) d_{ij}\} \Phi(\kappa^{-1/2}(u_i) A_{ij}) dH(u_i)$$

and

$$I_{ij}^{\phi}(w) \quad = \quad \int_0^{\infty} \kappa^{-w}(u_i) \exp\{-\frac{1}{2}\kappa^{-1}(u_i) d_{ij}\} \phi\left(\kappa^{-1/2}(u_i) A_{ij}\right) dH(u_i),$$

where

$$d_{ij} = \frac{(y_i - \mu_{ij} - b\Delta_j)^2}{\sigma_j^2} \quad \text{and} \quad A_{ij} = \lambda_j \frac{(y_i - \mu_{ij} - b\Delta_j)^2}{\sigma_j}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, g.$$

After some algebraic manipulation, we obtain:

$$\begin{aligned}
D_{\boldsymbol{\beta}}(y_i; \Theta_j) &= \frac{2}{\sqrt{2\pi\sigma_j^2}} \left[\sigma^{-2}(y_i - \mu_{ij} - b\Delta_j) I_{ij}^{\Phi}(3/2) - \sigma_j^{-1}\lambda_j I_{ij}^{\phi}(1)\right] \mathbf{x}_i, \\
D_{\vartheta_j}(y_i; \Theta_j) &= \frac{2}{\sqrt{2\pi\sigma_j^2}} \left[\sigma_j^{-2}(y_i - \mu_{ij} - b\Delta_j) I_{ij}^{\Phi}(3/2) - \sigma_j^{-1}\lambda_j I_{ij}^{\phi}(1)\right], \\
D_{\sigma_j^2}(y_i; \Theta_j) &= \frac{1}{\sqrt{2\pi\sigma_j^2}} \left[-\sigma_j^{-2} I_{ij}^{\Phi}(1/2) + \sigma_j^{-4}(y_i - \mu_{ij} - b\Delta_j)^2 I_{ij}^{\Phi}(3/2)\right. \\
&+ \left. \sigma_j^{-4}(y_i - \mu_{ij} - b\Delta_j) b\Delta_j I_{ij}^{\Phi}(3/2) - \lambda_j \sigma_j^{-3}(y_i - \mu_{ij}) I_{ij}^{\phi}(1)\right], \\
D_{\lambda_j}(y_i; \Theta_j) &= \frac{2}{\sqrt{2\pi}\sigma_j^2} \left[\frac{(y_i - \mu_{ij} - b\Delta_j) b}{(1 + \lambda_j^2)^{(3/2)}} I_{ij}^{\Phi}(3/2) + \left((y_i - \mu_{ij} - b\Delta_j) - \frac{b\Delta_j}{1 + \lambda_j^2} I_{ij}^{\phi}(1)\right)\right].
\end{aligned}$$

Now, returning to our original parameterization, we find the Hessian matrix for the original parameter vector $\boldsymbol{\Theta}^* = (\boldsymbol{\beta}^\top, (p_1, \ldots, p_{g-1})^\top, \beta_0, \mu_1, \ldots, \mu_g, \sigma_1^2, \ldots, \sigma_g^2, \lambda_1, \ldots, \lambda_g)^\top$,

$$\mathbf{I}_o(\widehat{\boldsymbol{\Theta}}^*) = \mathbf{J}(\boldsymbol{\Theta}|\boldsymbol{\Theta}^*)^\top (\mathbf{I}_o(\widehat{\boldsymbol{\Theta}}))^{-1} \mathbf{J}(\boldsymbol{\Theta}|\boldsymbol{\Theta}^*),$$

where $\mathbf{J}(\boldsymbol{\Theta}|\boldsymbol{\Theta}^*)$ is the Jacobian matrix of order $(p + 4g - 1) \times (p + 4g)$, defined by:

$$J(\boldsymbol{\Theta}|\boldsymbol{\Theta}^*) = \frac{\partial \boldsymbol{\Theta}}{\partial \boldsymbol{\Theta}^*} = \begin{pmatrix}
\mathbf{I}_p & \mathbf{0}_{p\times(g-1)} & \mathbf{0}_p & \mathbf{0}_{p\times g} & \mathbf{0}_{p\times g} & \mathbf{0}_{p\times g} \\
\mathbf{0}_{(g-1)\times p} & \mathbf{I}_{(g-1)\times(g-1)} & \mathbf{0}_{(g-1)} & \boldsymbol{\gamma} & \mathbf{0}_{(g-1)\times g} & \mathbf{0}_{(g-1)\times g} \\
\mathbf{0}_{g\times p} & \mathbf{0}_{g\times(g-1)} & \mathbf{1}_g & \mathbf{I}_g & \mathbf{0}_{g\times g} & \mathbf{0}_{g\times g} \\
\mathbf{0}_{g\times p} & \mathbf{0}_{g\times(g-1)} & \mathbf{0}_{g-1} & \mathbf{0}_{g\times g} & \mathbf{I}_{g\times g} & \mathbf{0}_{g\times g} \\
\mathbf{0}_{g\times p} & \mathbf{0}_{g\times(g-1)} & \mathbf{0}_{g-1} & \mathbf{0}_{g\times g} & \mathbf{0}_{g\times g} & \mathbf{I}_{g\times g}
\end{pmatrix},$$

where $\boldsymbol{\gamma} = \dfrac{\partial \mathbf{p}}{\partial \boldsymbol{\mu}^\top} = \mathbf{A}\mathbf{p}^\top$ is a matrix of dimension $(g-1) \times g$, with $\mathbf{A} = \left(-\dfrac{1}{\mu_1 - \mu_g}, \ldots, -\dfrac{1}{\mu_{g-1} - \mu_g}\right)^\top$ and $\mathbf{p} = (p_1, \ldots, p_{g-1})^\top$.

6. Simulation experiments

In this section, we conduct some simulation studies to illustrate the performance of our proposed model. The computational procedures were implemented using the R software (R Core Team, 2015), through the package `FMsmsnReg()` (Benitess et al., 2016). The first simulation study shows the consistency of the approximate standard errors for the ML estimates of parameters. The second simulation study shows that our proposed ECME algorithm estimates do provide good asymptotic properties. In the third study we compare the performance of the estimates for FM-SMSN-LR models in the presence of outliers on the response variable.

6.1. Parameter recovery (simulation study 1)

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 ECME algorithm. This is the first step to ensure that the estimation procedure works satisfactorily. We fit data that were artificially generated from the following model with two components:

$$\begin{cases} Y_i = \beta_0 + \mu_1 + \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_1,\ Z_{i1} = 1, \\ Y_i = \beta_0 + \mu_2 + \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_2,\ Z_{i2} = 1, \end{cases}$$

where Z_{ij} is a component indicator of Y_i with $P(Z_{ij} = 1) = p_j$, $j = 1, 2$, $\mathbf{x}_i^\top = (x_{i1}, x_{i2})$, such that $x_{i1} \sim U(0, 1)$ and $x_{i2} \sim U(0, 1)$, $i = 1, \ldots, n$, and ε_1 and ε_2 follow a distribution as in the assumption given in (16). We consider the following parameter values: $\beta_0 = -1$, $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top = (-4, -3)^\top$, $\mu_1 = -4$, $\mu_2 = 1$, $\lambda_1 = 1$, $\lambda_2 = -4$ and $p_1 = 0.8$. In addition, we consider the following scenarios: scenario 1 (well separated components): $\sigma_1^2 = 0.2$ and $\sigma_2^2 = 0.4$, and scenario 2 (poorly separated components): $\sigma_1^2 = 2$ and $\sigma_2^2 = 2$. For each combination of parameters, we generated 500 Monte Carlo samples of size $n = 500$ from the FM-SMSN-LR models, under three different situations: FM-SN-LR, FM-ST-LR ($\nu = 3$) and FM-SCN-LR ($\boldsymbol{\nu}^\top = (0.1, 01)$). The average values (Mean) and standard deviations (MC Sd) of the estimates across the 500 Monte Carlo samples were computed, along with the average (IM SE) values of the approximate standard deviations of the estimates obtained through the method described in Section 5. The results for the first scenario presented in Table 1. Note that in both scenarios, the results suggest that the proposed FM-SMSN-LR model produced satisfactory estimates. We also see from this table that the estimation method of the standard errors provides relatively close results (IM SE and MC Sd), indicating that the proposed asymptotic approximation for the variances of the ML estimates (Equation 27) is reliable.

Table 1: Simulation Study 1: mean and MC Sd are the respective mean estimates and standard deviations from fitting a FM-SMSN-LR model based on 500 samples. IM SE is the average value of the approximate standard error obtained through the information-based method. True values of parameters are in parentheses.

		Scenario 1: ($\sigma_1^2 = 0.2$, $\sigma_2^2 = 0.4$)			Scenario 2: ($\sigma_1^2 = \sigma_2^2 = 2$)		
Parameter		SN	ST($\nu = 3$)	SCN ($\nu = 0.1$)	SN	ST($\nu = 3$)	SCN ($\nu = 0.1$)
$\beta_0(-1)$	Mean	-1.0034	-1.0059	-0.9894	-1.0119	-1.0084	-0.9491
	IM SE	0.0841	0.1225	0.1090	0.2651	0.4668	0.3744
	MC Sd	0.0962	0.1065	0.0984	0.1284	0.1569	0.1905
$\beta_1(-4)$	Mean	-3.9966	-3.9991	-4.0026	-3.9977	-4.0003	-4.0003
	IM SE	0.0522	0.0638	0.0584	0.1272	0.1481	0.1502
	MC Sd	0.0528	0.0617	0.0603	0.1263	0.1384	0.1418
$\beta_2(-3)$	Mean	-3.0000	-3.0023	-3.0031	-2.9922	-2.9928	-2.9988
	IM SE	0.0519	0.0621	0.0587	0.1283	0.1499	0.1476
	MC Sd	0.0519	0.0627	0.0584	0.1263	0.1480	0.1426
$\mu_1(-4)$	Mean	-3.9980	-4.0097	-4.0141	-4.0561	-4.1158	-4.1601
	IM SE	0.0522	0.0858	0.0755	0.1893	0.3806	0.3021
	MC Sd	0.0938	0.1120	0.1039	0.1457	0.2623	0.3068
$\mu_2(1)$	Mean	1.0030	1.0146	1.0070	0.9952	1.0317	1.0524
	IM SE	0.0415	0.0580	0.0501	0.1049	0.1556	0.1315
	MC Sd	0.0906	0.0932	0.0927	0.0948	0.1291	0.1499
σ_1^2	Mean	0.2059	0.1840	0.1836	1.8241	1.6679	1.6051
	IM SE	0.0962	0.0712	0.0907	1.5942	0.9857	1.0652
	MC Sd	0.0377	0.0550	0.0503	0.4019	0.5767	0.6200
σ_2^2	Mean	0.4039	0.3829	0.3720	2.0732	1.8898	1.6386
	IM SE	0.0392	0.0529	0.0471	0.2789	0.3463	0.2527
	MC Sd	0.0328	0.0581	0.0599	0.1945	0.3118	0.4586
$p_1(0.2)$	Mean	0.2006	0.2020	0.2006	0.1971	0.2008	0.2021
	IM SE	0.0179	0.0186	0.0184	0.0203	0.0272	0.0234
	MC Sd	0.0179	0.0185	0.0183	0.0185	0.0265	0.0267
ν	Mean	-	3.2118	0.1102	-	3.9060	0.1329
$\gamma(0.1)$	Mean	-	-	0.1214	-	-	0.1999

6.2. Asymptotic properties of the EM estimates (simulation study 2)

The main focus in this simulation study is to show the asymptotic properties of the EM estimates. Our strategy is to generate artificial samples from the FM-SMSN-LR model with $x_i^\top = (x_{1i}, x_{2i})$, such that $x_{i1} \sim U(0,1)$ and $x_{i2} \sim U(0,1)$, $i = 1, \ldots, n$. We choose sample sizes $n = 100, 200, 300, 400, 500, 600, 700,$ 800, 900 and 1000. The true values of the parameters were taken as $\beta_0 = -1$, $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top = (-4, -3)^\top$, $\mu_1 = -4$, $\mu_2 = 1$, $\sigma_1^2 = 1$ and $\sigma_2^2 = 0.5$. For each combination of parameters and sample sizes, we generated 500 random samples from the FM-SMSN-LR models, under three different situations: FM-SN-LR, FM-ST-LR ($\nu = 3$) nd FM-SCN-LR ($\boldsymbol{\nu}^\top = (0.1, 01)$). In order to analyze asymptotic properties of the EM estimates, we computed the bias and the relative root mean square error (RMSE) for each combination of sample size and parameter values. For θ_i, they are given by:

$$\begin{aligned} \text{Bias}(\theta_i) &= \frac{1}{500}\sum_{j=1}^{500}(\theta_i^{(j)} - \theta_i), \\ \text{RMSE}(\theta_i) &= \sqrt{\frac{1}{500}\sum_{j=1}^{500}(\theta_i^{(j)} - \theta_i)^2}, \end{aligned}$$

where $\widehat{\theta}_i^{(j)}$ is the estimate of θ_i for the j-th sample. The results for β_1, β_2 and σ_1^2 are shown in Figure 1 and the results for μ_2, σ_2 and λ_1 are shown in Figure 2. One can see a pattern of convergence to zero of the bias and RMSE when n increases for all the parameters. As a general rule, we can say that Bias and RMSE tend to approach zero when the sample size increases, indicating that the estimates based on the proposed EM-type algorithm under the FM-SMSN-LR model do provide good asymptotic properties.

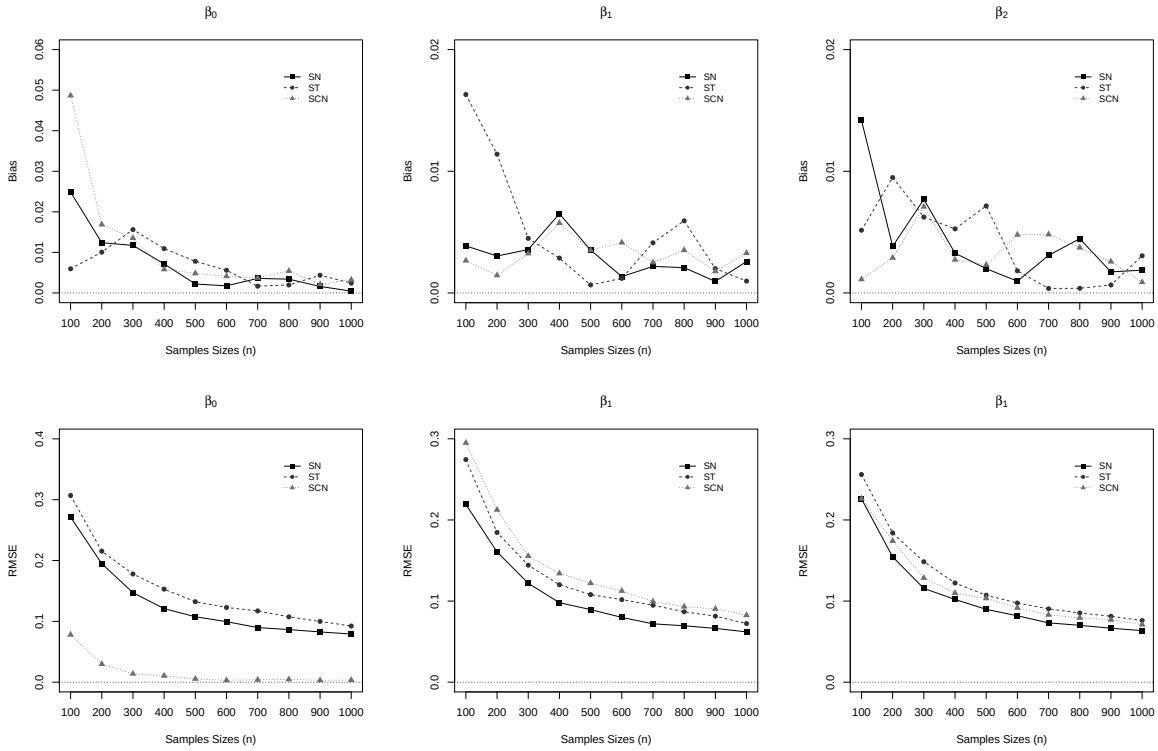

Figure 1: Simulation study 2. Average bias (first row) and average RMSE (second row) of the estimates of β_0,β_1, β_2.

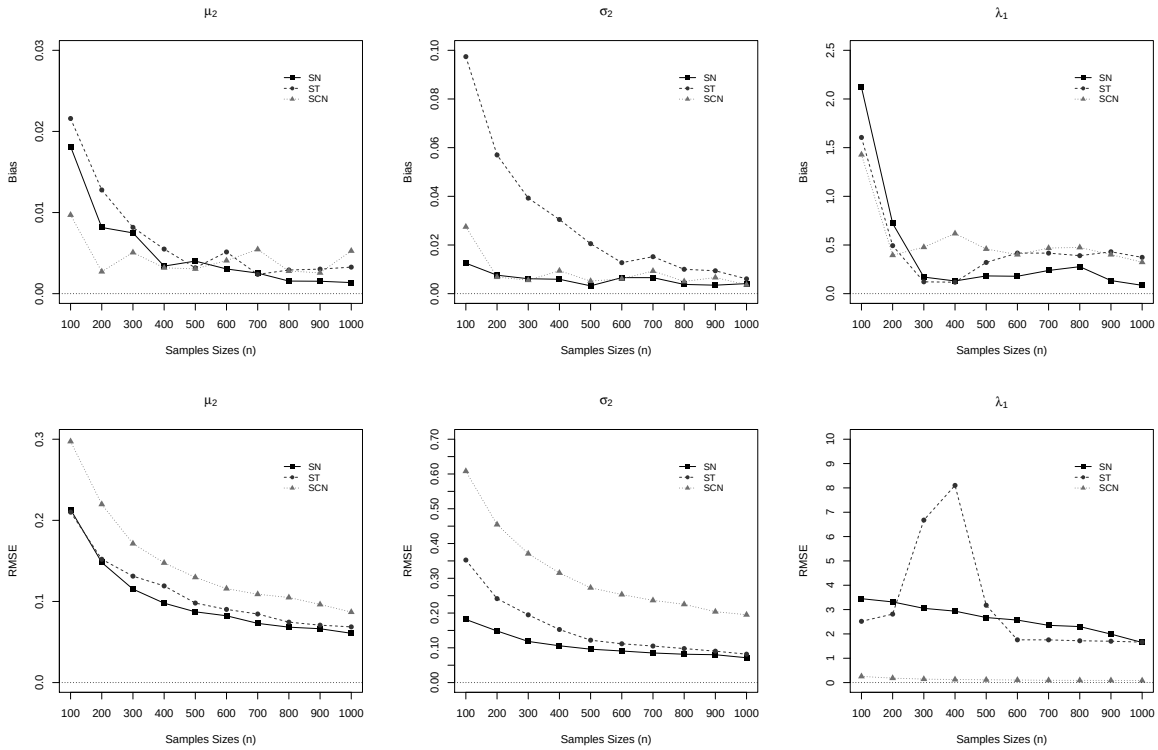

Figure 2: Simulation study 2. Average bias (first row) and average RMSE (second row) of the estimates of μ_2,σ_2, λ_1.

6.3. Robustness of the EM estimates (simulation study 3)

The purpose of this simulation study is to compare the effect of the robustness of the estimates for the FM-SMSN-LR models in the presence of outliers on the response variable. We consider the different cases of the FM-SMSN models with fixed $\boldsymbol{\nu}$, i.e., FM-SN-LR, FM-ST-LR ($\nu = 3$) and FM-CN-LR ($(\nu, \gamma) = (0.1, 0.1)$).

For this case, we generated 500 samples of size $n = 500$ under the FM-SMSN-LR model with $\varepsilon_i \sim \sum_{j=1}^{2} p_j SMSN(y_i|\mu_j + b\Delta_j, \sigma_j^2, \lambda_j, \boldsymbol{\nu}_j)$. To assess how much the EM estimates are influenced by the presence of outliers, we replaced the observation y_{150} by: $y_{150}(\vartheta) = y_{150} + \vartheta$, with $\vartheta = 1, 2, \ldots, 10$. For each replication, we obtained the parameter estimates with and without outliers, under the three FM-SMSN-LR models. We are interested in evaluating the relative change in the estimates as a ϑ function. Given $\boldsymbol{\Theta} = (\beta_1, \beta_2, p_1, p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, with $\boldsymbol{\theta}_j = (\beta_0, \mu_j, \sigma_j^2, \lambda_j)$, $j = 1, 2$, the relative change is defined by

$$RC\left(\widehat{\boldsymbol{\Theta}}_i(\vartheta)\right) = \left| \frac{\left(\widehat{\boldsymbol{\Theta}}_i(\vartheta) - \widehat{\boldsymbol{\Theta}}_i\right)}{\widehat{\boldsymbol{\Theta}}_i} \right|,$$

where $\widehat{\boldsymbol{\Theta}}_i(\vartheta)$ and $\widehat{\boldsymbol{\Theta}}_i$ denote the EM estimates of $\boldsymbol{\Theta}_i$ with and without perturbation, respectively.

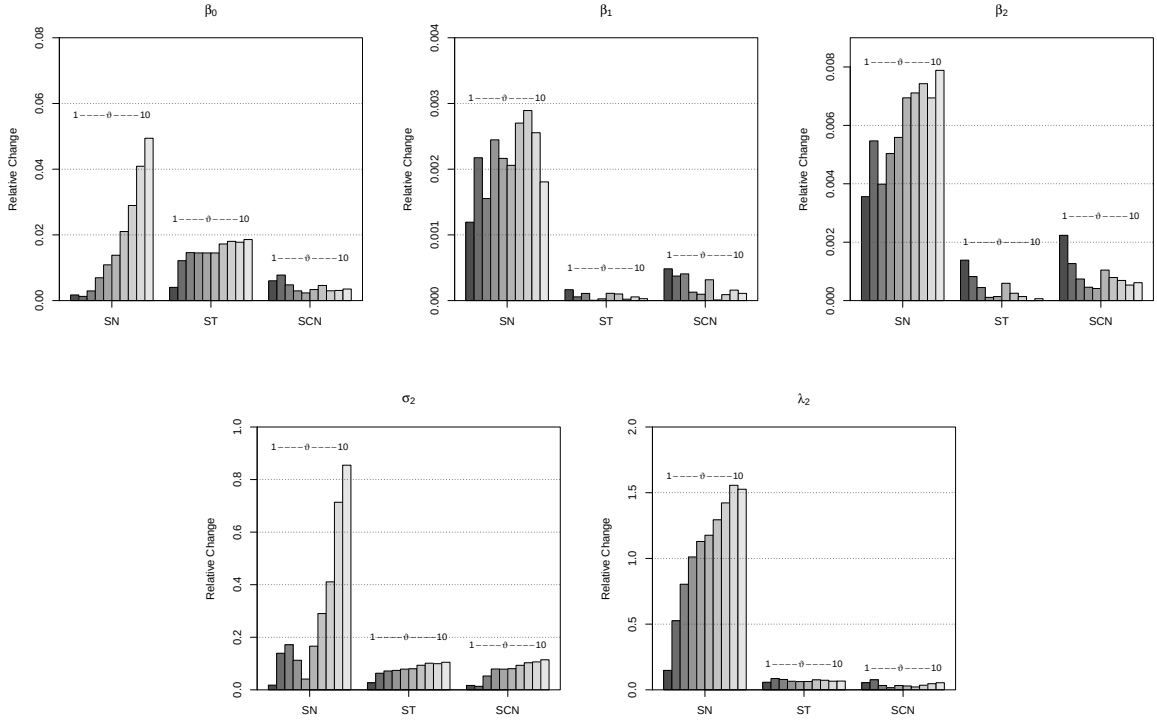

Figure 3: Simulation study 3. Average relative changes of estimates for different perturbations ϑ.

Figure 3 shows the average values of the relative changes undergone by all the parameters. We note that for all parameters, the average relative changes suddenly increase under FM-SN-LR model as the ϑ value grows. In contrast, for the FM-SMSN-LR models with heavy tails, namely the FM-ST-LR ($\nu = 3$)

and FM-SCN-LR($\nu = (0.1, 0.1)$), the measures vary little, indicating they are more robust than the FM-SN-LR model in the ability to accommodate discrepant observations.

7. Application

In this section, we consider the dataset previously analyzed by Forbes (1998) in a normal regression setting. The Horse Racing at Eagle Farm dataset contains the results for each horse in a sequence of 8 races at Eagle Farm, Brisbane, on 31 August 1998. Here, we focus on finishing position (Position), which is assumed to be explained by number of horses in race ($Starters$) and proportion of wins in previous starts ($Ratio$). Thus, we consider the following FM-SMSN-LR model:

$$Position_i = \beta_0 + \beta_1 Starters_i + \beta_2 Ratio_i + \varepsilon_i,$$

where ε_i belongs to the FM-SMSN family for $i = 1, \ldots, 102$. Using the `R` package `FMsmsnReg()` (see Appendix A), we fit the FM-SMSN-LR models as was described in Section 2. Table 2 presents the ML estimates of the parameters considering the four models, say, FM-SN-LR, FM-ST-LR, FM-SCN-LR and the FM-SSL-LR model, along with the corresponding standard errors (SE), obtained via the information-based procedure presented in Section 5. Notice from this table that the small value of the estimate of ν for the FM-ST-LR and FM-SSL-LR models indicates a lack of adequacy of the SN (and normal) assumption. Table 3 compares the fit of various mixture models with two, three and four components, using the model selection criteria discussed in Subsection 4.3 (see also Basso et al., 2010). Note from this table that, as expected, the heavy-tailed models perform significantly better than the SN (and normal) model, with mixtures of two ($g = 2$) components significantly better in all cases. Moreover, the FM-SCN-LR model fits the data substantially better.

Table 2: Horse Racing at Eagle Farm data. Parameter estimates of the FM-SMSN- LR models with $g = 2$. SE denotes the corresponding standard errors, obtained via the information-based matrix.

Parameter	FM-SN		FM-ST		FM-SCN		FM-SSL	
	ML	SE	ML	SE	ML	SE	ML	SE
β_0	30.8586	0.0672	31.2560	0.0563	29.5462	0.0810	30.8271	0.0682
β_1	0.3998	0.1103	0.4136	0.1002	0.4582	0.1349	0.4131	0.1086
β_2	-0.4813	0.0550	-0.4855	0.0564	-0.4864	0.1224	-0.4824	0.0604
p_1	0.3894	1.5633	0.36489	1.2814	0.3360	0.6980	0.3836	1.7550
μ_1	4.3664	11.676	4.7920	10.706	4.1784	13.8127	4.2467	11.4849
μ_2	-2.7843	5.8629	-2.7599	5.4037	-2.1145	6.8528	-2.6432	5.7719
σ_1^2	15.5259	5.8211	10.2435	5.3155	0.1331	6.9681	9.1715	5.7220
σ_2^2	5.6358	4.1416	2.5396	4.2291	0.0743	0.0342	2.3169	2.8320
λ_1	7.4412	2.0265	6.3088	1.5404	2.0575	0.0462	6.5139	1.6529
λ_2	-1.6843	8.5575	0.2706	6.9799	0.8547	3.2876	-0.7658	9.0087
ν	-	-	5.4983	-	0.7922	-	2.3005	-
γ	-	-	-		0.0100	-	-	

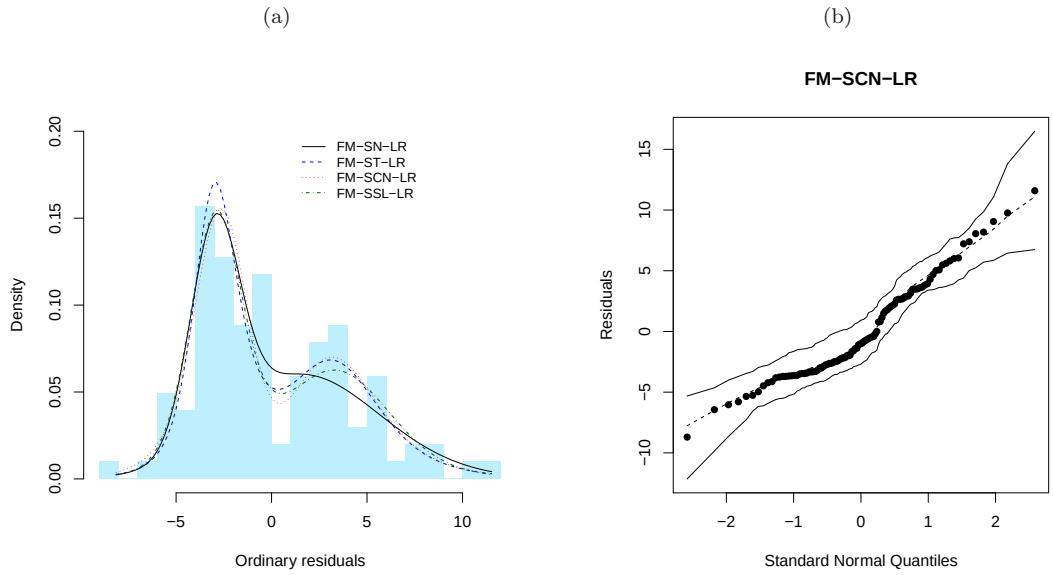

Figure 4: (a) Histogram of ordinary residuals with FM-SMSN-LR residual density model. (b) Q-Q plots and simulated envelopes for the residual.

Table 3: Horse Racing at Eagle Farm data. Model selection criteria for various FM-SMSN-LR models

Criteria	g	FM-SN	FM-ST	FM-SCN	FM-SSL
log-lik	2	-274.177	-273.448	-268.202	-274.772
	3	-271.077	-267.988	-272.834	-274.803
	4	-270.74	-269.227	-269.135	-272.664
AIC	2	568.354	568.897	560.404	571.545
	3	570.154	565.976	577.668	579.605
	4	577.479	576.454	578.27	583.329
BIC	2	594.603	597.772	591.903	600.419
	3	606.903	605.35	619.668	618.98
	4	624.729	626.329	630.77	633.203
EDC	2	568.553	569.116	560.642	571.763
	3	570.432	566.274	577.986	579.904
	4	577.837	576.832	578.668	583.707

In Figure 4 (a), we plot the histogram of OLS residuals and then display the residual densities for the four FM-SMSN-LR models superimposed on a single set of coordinate axes. Based on this a graphical representation, it appears once again that the FM-ST-LR, FT-SCN-LR and FT-SSL-LR fit have quite reasonable and better fit than the FM-SN-LR model. In order to detect incorrect specification of the error distribution for our best model (FM-SCN-LR), we present QQ-plots and simulated envelopes for the residuals $(y - \hat{y})$ in Figure 4 (b). This figure clearly indicates that the FM-SCN-LR is suitable for modeling these data, since there are no observations falling outside the envelope.

8. Conclusions

In this paper we consider a regression model whose error term follow a finite mixture of scale mixtures of skew-normal (SMSN) distributions, which is a rich class of distributions that contains the skew-normal, skew-t, skew-slash and skew-contaminated normal distributions as proper elements. This approach allows us to model data with great flexibility, accommodating simultaneously multimodality, skewness and heavy

tails for the random error in linear regression models. It is important to stress that our proposal is different from that proposed by Zeller et al. (2015), where they use a finite mixture of linear regression models, the so-called *switching regression*. In this paper, instead of mixtures of regressions, mixtures are exploited as a convenient semiparametric method, which lies between parametric models and kernel density estimators, to model the unknown distributional shape of the errors, and for this robust structure we developed a simple EM-type algorithm to perform maximum likelihood (ML) inference of the parameters with closed-form expression at the E-step. The proposed methods are implemented using the R package `FMsmsnReg()`, providing practitioners with a convenient tool for further applications in their domain. The practical utility of the new method is illustrated with the analysis of two real datasets and several simulation studies.

The proposed methods can be extended to multivariate settings using the multivariate SMSN class of distributions (Cabral et al., 2012b), such as the recent proposals of Soffritti and Galimberti (2011) and Galimberti and Soffritti (2014). Due to the popularity of Markov chain Monte Carlo techniques, another potential work is to pursue a fully Bayesian treatment in this context for producing posterior inference. The method can also be extended to mixtures of regressions with skewed and heavy-tailed censored responses based on recent approaches by Caudill (2012) and Karlsson and Laitila (2014).

Acknowledgment: Victor H Lachos acknowledges support from CNPq-Brazil (Grant 306334/2015-1) and FAPESP-Brazil (Grant 2014/02938-9). Partial support from CAPES are also Acknowledge.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Cont. 19, 716–723.

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

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

Azzalini, A., Genton, M., 2008. Robust likelihood methods based on the skew-t and related distributions. International Statistical Review 76, 1490–1507.

Bai, Z. D., Krishnaiah, P. R., Zhao, L. C., 1989. On rates of convergence of efficient detection criteria in signal processing with white noise. IEEE Trans. Info. Theory 35, 380–388.

Bartolucci, F., Scaccia, L., 2005. The use of mixtures for dealing with non-normal regression errors. Computational Statistics & Data Analysis 48 (4), 821–834.

Basford, K., Greenway, D., McLachlan, G., Peel, D., 1997. Standard errors of fitted component means of normal mixtures. Computational Statistics 12, 1–18.

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 & Data Analysis 54 (12), 2926–2941.

Benitess, L., Maehara, R., Lachos, V. H., 2016. CensMixReg: Censored Linear Mixture Regression Models. R package version 1.0.
URL `http://CRAN.R-project.org/package=FMsmsnReg`

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., Madruga, M. R., 2012a. Bayesian analysis of skew-normal independent linear mixed models with heterogeneity in the random-effects population. Journal of Statistical Planning and Inference 142, 181–200.

Cabral, C. R. B., Lachos, V. H., Prates, M. O., 2012b. Multivariate mixture modeling using skew-normal independent distributions. Computational Statistics & Data Analysis 56 (1), 126–142.

Caudill, S. B., 2012. A partially adaptive estimator for the censored regression model based on a mixture of normal distributions. Statistical Methods & Applications 21, 121–137.

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.

Forbes, D., 1998. A day at the races. MS305 Data Analysis Project, Department of Mathematics, University of Queensland.

Galea, M., Paula, G. A., Bolfarine, H., 1997. Local influence in elliptical linear regression models. Journal of the Royal Statistical Society: Series D (The Statistician) 46 (1), 71–79.

Galimberti, G., Soffritti, G., 2014. A multivariate linear regression analysis using finite mixtures of t distributions. Computational Statistics and Data Analysis 71, 138–150.

Karlsson, M., Laitila, T., 2014. Finite mixture modeling of censored regression models. Statistical Papers 55, 627–642.

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

Lange, K. L., Little, R., Taylor, J., 1989. Robust statistical modeling using t distribution. Journal of the American Statistical Association 84, 881–896.

Lange, K. L., Sinsheimer, J. S., 1993. Normal/independent distributions and their applications in robust regression. J. Comput. Graph. Stat 2, 175–198.

Meng, X., Rubin, D. B., 1993. Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika 81, 633–648.

Prates, M. O., Lachos, V. H., Cabral, C. R. B., 2013. mixsmsn: Fitting finite mixture of scale mixture of skew-normal distributions. Journal of Statistical Software 54 (12), 1–20.

R Core Team, 2015. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
URL `http://www.R-project.org/`

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

Soffritti, G., Galimberti, G., 2011. Multivariate linear regression with non-normal errors: a solution based on mixture models. Statistics and Computing 21 (4), 523–536.

Zeller, C. B., Cabral, C. R., Lachos, V. H., 2015. Robust mixture regression modeling based on scale mixtures of skew-normal distributions. TEST, 1–22.

Appendix A: Sample output from R package *FMsmsnReg()*

```
-------------------------------------------------------------
Finite Mixture of Scale Mixture Skew Normal Regression Model
-------------------------------------------------------------

Observations = 102

Family = Skew.cn
-----------
Estimates
-----------

       Estimate      SE
beta0   29.5462  0.0810
beta1    0.4582  0.1349
beta2   -0.4864  0.1224
mu1      4.1784 13.8127
mu2     -2.1145  6.8528
sigma1   0.1326  6.9681
sigma2   0.0739  0.0342
shape1   1.8704  0.0462
shape2   0.8492  3.2876
pii1     0.3388  0.6980
nu       0.7912     -
gamma    0.0100     -

------------------------
Model selection criteria
------------------------

        Loglik     AIC     BIC     EDC     ICL
Value -268.202 560.404 591.903 560.642 970.211

-------
Details
-------

Convergence reached? = TRUE
EM iterations = 267 / 500
Criteria = 8.594e-06
Processing time = 8.260466 secs
```