CHILEAN JOURNAL OF STATISTICS

Edited by Víctor Leiva

Volume 10 Number 1
April 2019
ISSN: 0718-7912 (print)
ISSN: 0718-7920 (online)

Published by the
Chilean Statistical Society
SOCHE
SOCIEDAD CHILENA DE ESTADÍSTICA

Chilean Journal of Statistics
Vol. 10, No. 1, April 2019, 21–41

Statistical Modeling
Research Paper

Linear regression models using finite mixtures of skew heavy-tailed distributions

Luis Benites[1,*], Rocío Maehara[2], Victor H. Lachos[3], and Heleno Bolfarine[4]

[1]Departamento de Ciencias, Pontificia Universidad Católica del Perú, Perú,
[2]Departamento de Ingeniería, Universidad del Pacífico, Lima, Perú,
[3]Department of Statistics, University of Connecticut, Storrs, Connecticut, USA,
[4]Departamento de Estatística, Universidade de São Paulo, Brazil

(Received: 15 March 2019 · Accepted in final form: 15 April 2019)

Abstract

In this paper, we propose a regression model based on the assumption that the error term follows a mixture of normal distributions. Specifically, we consider a finite scale mixture of skew-normal distributions, a rich family that contains the skew-normal, skew-t, skew-slash and skew-contaminated normal distributions as members. This model allows us to describe data with high flexibility, simultaneously accommodating multimodality, skewness and heavy tails. We develop a simple EM-type algorithm to perform maximum likelihood inference of the parameters of the proposed model with closed-form expressions for both E- and M-steps. Furthermore, the observed information matrix is derived analytically to account for the corresponding standard errors and a bootstrap procedure is implemented to test the number of components in the mixture. The practical utility of the new model is illustrated with a real dataset and several simulation studies. The proposed algorithm and methods are implemented in an `R` package named `FMsmsnReg`.

Keywords: ECME algorithm · Mixture model · Non-normal error distribution · Scale mixtures of skew-normal distributions

Mathematics Subject Classification: Primary 62J05 · Secondary 62J99

1. Bibliographical Review and Motivating Example

1.1 Introduction

A basic assumption of the linear regression (LR) model is that the error term follows a normal distribution. However, it is well known that data from some phenomena do not always satisfy this assumption, instead having a distribution with heavy tails, skewness or multimodality. Many extensions of this classic model have been proposed to broaden the applicability of Gaussian linear regression (N-LR) analysis to situations where the Gaussian error term assumption may be inadequate, such as, the use of the Student-t distribution (Lange et al., 1989), which is appropriate for datasets involving errors with

*Corresponding author. Email: lbenitess@pucp.edu.pe

ISSN: 0718-7912 (print)/ISSN 0718-7920 (online)

http://www.soche.cl/chjs

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), the asymmetrical class of skew-normal (SMSN) distributions proposed by Branco and Dey (2001) or the unified skew-elliptical distributions proposed by Arellano and Genton (2010). However, in practice when nothing is known about the true distribution of the error terms, a risk exists that linear regression analysis based on any of the above models will be performed using an incorrectly specified model. There can also 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 (FM-LR) models have been recently proposed. For instance, Bartolucci and Scaccia (2005), Soffritti and Galimberti (2011) and Galimberti and Soffritti (2014) 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.

The classic approach to finite mixture modeling has several challenging aspects. There are nontrivial issues, like non-identifiability and an unbounded likelihood. In this context, Holzmann and Munk (2006) established the identifiability of finite mixtures of elliptical distributions under conditions of the characteristic or probability density function (PDF) generators. More recently, Otianiano et al. (2015) established the identifiability of finite mixture of skew-normal (Azzalini, 1985) and skew-t (Azzalini and Genton, 2008) distributions.

The class of SMSN distributions, proposed by Branco and Dey (2001), is attractive since it simultaneously models skewness with heavy tails (Prates et al., 2012) and contains as proper elements distributions such as the skew-normal, skew-t, skew-slash, skew-contaminated normal and all the symmetric class of scale mixtures of normal (SMN) distributions defined by Andrews and Mallows (1974). Besides this, it has a stochastic representation for easy implementation of the Expectation-Maximization (EM) algorithm (Dempster et al., 1977) and it also facilitates the study of many useful properties. Thus, this extension results in a flexible class of models for robust estimation and inference in FM-LR models.

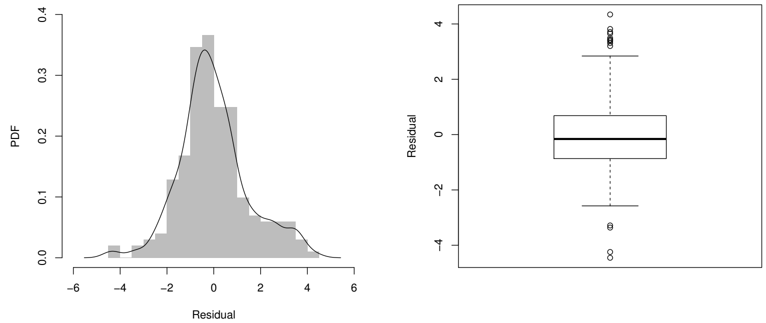
The objective of this paper is to propose a mixture regression model (and associated likelihood inference) based on the mixtures of the class of scale mixtures of skew-normal (SMSN) distributions, by extending the mixture model based on symmetrical distributions. An advantage of this model is the possibility of fitting multimodality, heavy tails and skewness simultaneously. We derive a mixture model for the random errors based on the class of SMSN distributions (FM-SMSN-LR model) and evaluate the performance of the FM-SMSN-LR model by simulations. In order to motivate our research, we describe the following example with a dataset from the Australian Institute of Sport data (AIS).

1.2 Motivating example

Before discussing the goal of this work, we present a motivating example. More specifically, we extend the linear regression model proposed by Bartolucci and Scaccia (2005), which is defined as

$$Y_i = \beta_0 + \boldsymbol{x}_i{}^\top \boldsymbol{\beta} + \varepsilon_i, \quad f(\varepsilon_i) = \sum_{j=1}^{g} p_j \phi(\varepsilon_i | \mu_j, \sigma_j^2), \quad i = 1, \ldots, n,$$

where Y_i is the response of case i, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ is a vector of explanatory variable values, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is a vector of unknown linear parameters, p_j are positive weights summing to 1, the μ_j terms satisfy the constraint $\sum_{j=1}^{g} p_j \mu_j = 0$, $\phi(.; \mu_j, \sigma_j^2)$ denotes



(a) (b)

Figure 1. Histogram with a kernel PDF estimate superimposed (a) and the boxplot of ordinary residuals (b) with AIS data.

the PDF of the normal distribution, by assuming that the distribution of the error terms follows a finite mixture of SMSN distribution, so that the FM-SMSN-LR is defined. It is important to stress that our proposal is different from that of Zeller et al. (2016), where the linear regression is modeled with different regression functions, the so-called mixture of regressions or switching regression (Quandt and Ramsey, 1978). An important question that is addressed in this paper is whether a mixture model ($g \geq 2$) is needed instead of a one-component model. Thus, we use the parametric bootstrap log-likelihood ratio statistic, which was proposed by Turner (2000).

To test our proposed model, we use the AIS data available in an `R` package named `FMsmsnReg`. Figure 1 (panels a and b) displays the histogram with a kernel PDF estimate superimposed and the boxplot of ordinary residuals, respectively, obtained by fitting a N-LR model to the AIS data. The plots reveal the existence of multimodal residuals, with evident presence of outliers. In summary, it is necessary to consider a more robust structure in the error. Therefore, this example serves as a motivation for the FM-SMSN-LR model.

1.3 Organization of the paper

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 FM-SMSN-LR model, including the EM-type algorithm for maximum likelihood (ML) estimation, and derive the empirical information matrix analytically to obtain the standard errors. In Section 4, numerical samples using both simulated and real data are given to illustrate the performance of the proposed model. Finally, some concluding remarks are presented in Section 5.

2. Background

2.1 Scale mixtures of skew-normal distributions

Next, we start by defining the skew-normal (SN) distribution and then we introduce some useful properties. 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 $\lambda \in \mathrm{R}$, denoted by $Z \sim \mathrm{SN}(\mu, \sigma^2, \lambda)$, if its PDF is given by

$$\phi_{\mathrm{SN}}(z|\mu, \sigma^2, \lambda) = 2\phi(z|\mu, \sigma^2)\Phi\big(\lambda(z-\mu)/\sigma\big).$$

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

DEFINITION 2.1 A random variable Y has an SMSN distribution with location parameter μ, scale parameter σ^2 and skewness parameter λ, denoted by $\mathrm{SMSN}(\mu, \sigma^2, \lambda; H)$, if it has the stochastic representation

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

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

The mean and variance of Y are given respectively by

$$\mathrm{E}[Y] = \mu + \sqrt{\frac{2}{\pi}}K_1\Delta, \quad \mathrm{Var}[Y] = \sigma^2\Big(K_2 - \frac{2}{\pi}K_1^2\delta^2\Big), \tag{1}$$

where $\Delta = \sigma\delta$, with $\delta = \lambda/\sqrt{1+\lambda^2}$ and $K_r = \mathrm{E}[U^{-r/2}]$, $r = 1, 2, \ldots$. Although we can deal with any $\kappa(\cdot)$ 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 $Y|U = u \sim \mathrm{SN}(\mu, u^{-1}\sigma^2, \lambda)$. Thus, the PDF of Y is expressed as

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

When H is degenerate, with $u = 1$, we obtain the $\mathrm{SN}(\mu, \sigma^2, \lambda)$ distribution, and when $\lambda = 0$, the SMSN distribution reduces to the class of scale-mixtures of the normal (SMN) distribution represented by the PDF $f_0(y) = \phi_{\mathrm{SMN}}(y|\mu, \sigma^2, \boldsymbol{\nu}) = \int_0^\infty \phi(y|\mu, u^{-1}\sigma^2)\mathrm{d}H(u|\boldsymbol{\nu})$.

2.2 Special cases of the SMSN distributions

Some special families of SMSN distributions are the following:

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

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

  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 $Y \sim \mathrm{ST}(\mu, \sigma^2, \lambda, \nu)$.
- The skew-slash distribution. It is denoted by $Y \sim \mathrm{SSL}(\mu, \sigma^2, \lambda, \nu)$ and the associated PDF is given by

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

  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 $Y \sim \mathrm{SCN}(\mu, \sigma^2, \lambda, \nu, \gamma)$. Its PDF is given by

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

  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 $\gamma = 1$.

2.3 Computational framework

The `R` software (R Core Team, 2016) produces statistical analyses, with its open source codes. This non-commercial computational program may be downloaded from `http://www.r-project.org`. Our method was implemented in `R` and its codes are available through the `FMsmsnReg` package (Benites et al., 2016). We use the `mixmsmsn` package, which allows the simulation of mixture the class of scale mixture of skew-normal distributions, see Prates et al. (2013). This computational framework is useful for conducting the simulation studies and the empirical illustration carried out in Section 4.

3. The linear regression model with FM-SMSN errors

3.1 General context

Next, we introduce the linear regression model using finite mixture of skew heavy tailed distributions where the distribution of the error terms follows a finite mixture of scale mixture of skew-normal distributions (FM-SMSN-LR), following a similar setup as that developed by Bartolucci and Scaccia (2005). Consider the linear regression model expressed as

$$Y_i = \beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{3}$$

where Y_i is the response of case i, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ is a vector of explanatory variables of dimension $(p+1) \times 1$, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is the regression parameter vector. Furthermore, we assume that

$$f(\varepsilon_i) = \sum_{j=1}^{g} p_j \phi_{\mathrm{SMSN}}\left(\varepsilon_i | \mu_j + b\Delta_j, \sigma_j^2, \lambda_j, \boldsymbol{\nu}_j\right), \quad i = 1, \ldots, n, \tag{4}$$

where p_j are positive weights summing to 1, the μ_j s satisfy the identifiability constraint $\sum_{j=1}^{g} p_j \mu_j = 0$, $b = -\sqrt{2/\pi} K_1$, $K_1 = \mathrm{E}[U^{-1/2}]$, $\Delta_j = \sigma_j \delta_j$ with $\delta_j = \lambda_j / \sqrt{1 + \lambda_j^2}$. Then from Equation (1), we have that $\mathrm{E}(\varepsilon_i) = 0$. Thus, for linearity of SMSN distributions, the PDF of Y_i is expressed as

$$f(y_i|\boldsymbol{\theta}) = \sum_{j=1}^{g} p_j \phi_{\mathrm{SMSN}}(y_i | \mu_{ij} + b\Delta_j, \sigma_j^2, \lambda_j, \boldsymbol{\nu}_j), \quad \mu_{ij} = \beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta} + \mu_j = \vartheta_j + \boldsymbol{x}_i^\top \boldsymbol{\beta}, \tag{5}$$

where $\mu_{ij} = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \vartheta_j$, $\vartheta_j = \beta_0 + \mu_j$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, (p_1, \ldots, p_{g-1})^\top, \vartheta_1, \ldots, \vartheta_g, \sigma_1^2, \ldots, \sigma_g^2, \lambda_1, \ldots, \lambda_g, \nu_1, \ldots, \nu_g)^\top$ is the vector with all parameters. Concerning the parameter $\boldsymbol{\nu}_j$ of the mixing distribution $H(.|\boldsymbol{\nu}_j)$, for $j = 1, \ldots, g$, it can be a vector of parameters, e.g., the contaminated normal distribution. Thus, for computational convenience we assume that $\boldsymbol{\nu}_1 = \ldots = \boldsymbol{\nu}_g = \boldsymbol{\nu}$. This strategy works very well in the empirical studies that we have conducted and greatly simplifies the optimization problem. For $U = 1$, Equations (3) and (4) lead to the FM-N-LR defined by Bartolucci and Scaccia (2005). Moreover, when $g = 1$ and a nonlinear function is used instead of $\boldsymbol{x}_i^\top \boldsymbol{\beta}$, the FM-SMSN-LR framework reduces to the model discussed by Garay et al. (2011). For each i and j, consider the latent indicator variable Z_{ij}, such that

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

Observe that $Z_{ij} = 1$ if and only if $Z_i = j$. Then

$$P(Z_{ij} = 1) = 1 - P(Z_{ij} = 0) = p_j \quad \text{and} \quad y_i | Z_{ij} = 1 \sim \mathrm{SMSN}\big(\mu_{ij} + b\Delta_j, \sigma_j^2, \lambda_j; H(\boldsymbol{\nu})\big). \tag{6}$$

Note that by integrating out $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{ig})^\top$, we obtain the marginal PDF presented in Equation (2) and $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ are independent random vectors, each one having a multinomial distribution with PDF defined as $f(\boldsymbol{z}_i) = p_1^{z_{i1}} p_2^{z_{i2}} ... (1 - p_1 - \ldots - p_{g-1})^{z_{ig}}$, which we denote by $\boldsymbol{Z}_i \sim \mathrm{M}(1; p_1 \ldots, p_g)$. These latent vectors appear in the hierarchical representation given next, which is used to build the Expectation Conditional Maximization Either (ECME) algorithm as proposed by Liu and Rubin (1994), which is a variant of the EM algorithm Dempster et al. (1977). From Equation (6) along with Definition 2.1, the FM-SMSN-LR model can be represented as

$$\begin{aligned} Y_i | u_i, t_i, Z_{ij} = 1 &\overset{\text{IND}}{\sim} \mathrm{N}(\mu_{ij} + \Delta_j t_i, u_i^{-1} \Gamma_j), \qquad (7) \\ T_i | u_i, Z_{ij} = 1 &\overset{\text{IND}}{\sim} \mathrm{TN}\big(b, u_i^{-1}, (b, \infty)\big), \\ U_i | Z_{ij} = 1 &\overset{\text{IND}}{\sim} H(u_i; \boldsymbol{\nu}), \\ \boldsymbol{Z}_i &\overset{\text{IID}}{\sim} \mathrm{M}(1; p_1 \ldots, p_g), \quad i = 1, \ldots, n, \ j = 1, \ldots, g, \qquad (8) \end{aligned}$$

where IND denotes independent, whereas IID stands for independent and identically distributed, with $\Gamma_j = (1 - \delta_j^2)\sigma_j^2$, $\Delta_j = \sigma_j \delta_j$ and $\delta_j = \lambda_j / \sqrt{1 + \lambda_j^2}$.

3.2 Parameter estimation via the ECME algorithm

Next, we show how to implement the ECME algorithm for ML estimation of the parameters of the FM-SMSN-LR model. By using Equations (7) to (8), we have that the complete-data log-likelihood function is given by

$$\ell_c(\boldsymbol{\theta} | \boldsymbol{y}, \boldsymbol{t}, \boldsymbol{u}, \boldsymbol{z}) = c + \sum_{i=1}^{n} \sum_{j=1}^{g} Z_{ij} \Big\{ \log(p_j) - \frac{1}{2} \log(\Gamma_j) - \frac{u_i}{2\Gamma_j} (y_i - \mu_{ij} - \Delta_j t_i)^2 + \log(h(u_i | \boldsymbol{\nu})) + \log\big[\phi_{\mathrm{TN}}(t_i | b, u_i^{-1}, (b, \infty))\big] \Big\},$$

where c is a constant that is independent of the parameter vector $\boldsymbol{\theta}$. By defining the quantities $\widehat{z}_{ij} = \mathrm{E}[Z_{ij} | \widehat{\boldsymbol{\theta}}, y_i]$, $\widehat{s}_{1ij} = \mathrm{E}[Z_{ij} U_i | \widehat{\boldsymbol{\theta}}, y_i]$, $\widehat{s}_{2ij} = \mathrm{E}[Z_{ij} U_i T_i | \widehat{\boldsymbol{\theta}}, y_i]$ and $\widehat{s}_{3ij} =$

$\mathrm{E}[Z_{ij}U_iT_i^2|\widehat{\boldsymbol{\theta}}, y_i]$, as having known properties of conditional expectation, we obtain

$$\widehat{z}_{ij} = \frac{\widehat{p}_j \phi_{\mathrm{SMSN}}(y_i|\mu_{ij} + b\Delta_j, \sigma_j^2, \lambda_j, \boldsymbol{\nu})}{\sum_{j=1}^{g} \widehat{p}_j \phi_{\mathrm{SMSN}}(y_i|\mu_{ij} + b\Delta_j, \sigma_j^2, \lambda_j, \boldsymbol{\nu})},$$

$\widehat{s}_{1ij} = \widehat{z}_{ij}\widehat{u}_{ij}$, $\widehat{s}_{2ij} = \widehat{z}_{ij}(\widehat{u}_{ij}\widehat{\mu}_{T_{ij}} + \widehat{M}_{T_j}\widehat{\tau}_{1_{ij}})$ and $\widehat{s}_{3ij} = \widehat{z}_{ij}(\widehat{u}_{ij}\widehat{\mu}^2_{T_{ij}} + \widehat{M}^2_{T_j} + \widehat{M}_{T_j}(\widehat{\mu}_{T_{ij}} + b)\widehat{\tau}_{1_{ij}})$, where

$$\widehat{\tau}_{1_{ij}} = \mathrm{E}\left[U_i^{1/2} W_{\Phi_1}\left(\frac{U_i^{1/2}\widehat{\mu}_{T_{ij}})}{\widehat{M}_{T_j}}\right) \mid \widehat{\boldsymbol{\theta}}, y_i, Z_{ij} = 1\right], \quad i = 1, \ldots, n, \quad j = 1, \ldots, g,$$

$$\widehat{M}^2_{T_j} = \frac{\Gamma_j}{\Gamma_j + \Delta_j^2}, \ \widehat{\mu}_{T_{ij}} = b + \frac{\Delta_j}{\Gamma_j + \Delta_j^2}(y_i - \mu_{ij} - \Delta b) \ \text{ and } \ \widehat{u}_{ij} = \mathrm{E}[U_j|\widehat{\boldsymbol{\theta}}, y_i, Z_{ij} = 1].$$

Once again, at 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

$$Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) = c + \sum_{i=1}^{n}\sum_{j=1}^{g}\Bigg(\widehat{z}_{ij}^{(k)}(\log(p_j) - \frac{1}{2}\log(\Gamma_j) - \frac{1}{2\Gamma_j}\Big(\widehat{s}_{1ij}^{(k)}(y_i - \mu_{ij})^2 - 2(y_i - \mu_{ij})\Delta_j\widehat{s}_{2ij}^{(k)}$$
$$+ \ \Delta_j^2\widehat{s}_{3ij}^{(k)}\Big) + \mathrm{E}[Z_{ij}\log(h(U_i|\boldsymbol{\nu}))|\widehat{\boldsymbol{\theta}}^{(k)}, y_i] + \mathrm{E}[Z_{ij}\log(\phi_{\mathrm{TN}}(T_i|b, u_i^{-1}, (b, \infty)))|\widehat{\boldsymbol{\theta}}^{(k)}, y_i]\Bigg).$$

In the CML-step we update the estimate of $\boldsymbol{\nu}$ by direct maximization of the marginal log-likelihood, circumventing the computation of the conditional expectations $\widehat{s}_{4ij} = \mathrm{E}[Z_{ij}\log(h(U_i|\boldsymbol{\nu}))|\widehat{\boldsymbol{\theta}}, y_i]$ and $\widehat{s}_{5ij} = \mathrm{E}[Z_{ij}\log(\phi_{\mathrm{TN}}(T_i|b, u_i^{-1}, (b, \infty)))|\widehat{\boldsymbol{\theta}}^{(k)}, y_i]$. Thus, the ECME algorithm for ML 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 $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) = \mathrm{E}[\ell_c(\boldsymbol{\theta})|\boldsymbol{y}, \widehat{\boldsymbol{\theta}}^{(k)}]$ over $\boldsymbol{\theta}$, which leads to the closed-form expressions given by

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

$$\widehat{\vartheta}_j^{(k+1)} = \left(\sum_{i=1}^{n}(\widehat{s}_{1ij}^{(k)}(y_i - \boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}}) - \widehat{\Delta}_j^{(k)}\widehat{s}_{2ij}^{(k)})\right) / \sum_{i=1}^{n}\widehat{s}_{1ij}^{(k)},$$

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

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

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

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

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

Through constraint $\sum_{j=1}^{g} p_j \mu_j = 0$ (Bartolucci and Scaccia, 2005), we obtain the estimates of β_0 and μ_j as

$$\widehat{\beta}_0^{(k+1)} = \sum_{j=1}^{g} \widehat{p}_j^{(k+1)} \widehat{\vartheta}_j^{(k+1)} \quad \text{and} \quad \widehat{\mu}_j^{(k+1)} = \widehat{\vartheta}_j^{(k+1)} - \widehat{\beta}_0^{(k+1)},$$

respectively, for $j = 1, \ldots, g$. This process is iterated until a suitable stopping criterion is satisfied. To avoid an indication of lack of progress of the algorithm (McNicholas et al., 2010), we adopted the Aitken acceleration method as the stopping criterion. At iteration k, we first compute the Aitken acceleration factor $c^{(k)} = (\ell^{(k+1)} - \ell^{(k)})/(\ell^{(k)} - \ell^{(k-1)})$, where following Böhning et al. (1994), the asymptotic estimate of the log-likelihood at iteration $k+1$ is given by

$$\ell_{\infty}^{(k+1)} = \ell^{(k)} + \frac{1}{1 - c^{(k)}} \left[\ell^{(k+1)} - \ell^{(k)} \right]. \tag{9}$$

As pointed out by Lindsay (1995), the algorithm is considered to reach convergence when $\ell_{\infty}^{(k+1)} - \ell^{(k+1)} < \varepsilon$, where ε is the desired tolerance (we use $\varepsilon = 10^{-6}$). A usual criticism is that EM-type procedures tend to get stuck in local modes. A convenient way to avoid this limitation 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-squares (OLS) estimate. Initial values for p_j, μ_j, σ_j^2, λ_j and $\boldsymbol{\nu}$, $j = 1, \ldots, g$, are obtained by fitting the mixture model given in Equation (3) to the OLS residuals (Bartolucci and Scaccia, 2005), which can be done through the `FMsmsnReg` package (Benites et al., 2016).

3.3 Model selection and approximate standard errors

Consider the problem of comparing several FM-SMSN-LR models, with different numbers of component PDFs. Here, we use two model selection criteria, the Akaike information criterion plus a bias correction term (Hurvich and Tsai, 1989), denoted by (AIC_c), and the adjusted Bayesian information criterion (Sclove, 1987), denoted by (BIC_a). These criteria are defined as

$$\text{AIC}_c = -2\ell(\widehat{\boldsymbol{\theta}}) + \frac{2n\rho}{n - \rho - 1} \quad \text{and} \quad \text{BIC}_a = -2\ell(\widehat{\boldsymbol{\theta}}) + \rho \log\left(\frac{n+2}{2}\right),$$

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 n is the sample size.

A simple way of obtaining the standard errors of ML estimators of mixture model parameters is to approximate the asymptotic covariance matrix of $\widehat{\boldsymbol{\theta}}$ by the inverse of the observed information matrix. Let $\boldsymbol{I}_o(\boldsymbol{\theta}) = -\partial^2 \ell(\boldsymbol{\theta}|\boldsymbol{y})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top$ be the observed information matrix, where $\ell(\boldsymbol{\theta}|\boldsymbol{y})$ is the observed log-likelihood function, which is obtained using Equation (5). In this work we use the alternative method suggested by Basford et al. (1997), which consists of approximating the inverse of the covariance matrix by

$$\boldsymbol{I}_o(\widehat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \widehat{\boldsymbol{s}}_i \widehat{\boldsymbol{s}}_i^{\top}, \quad \text{where} \quad \widehat{\boldsymbol{s}}_i = \frac{\partial}{\partial \boldsymbol{\theta}} \log\left[f(y_i|\theta)\right]\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}, \tag{10}$$

where $\widehat{\boldsymbol{s}}_i = (\widehat{s}_{i,\boldsymbol{\beta}}^{\top}, \widehat{s}_{i,p_1}, \ldots, \widehat{s}_{i,p_{g-1}}, \widehat{s}_{i,\vartheta_1}, \ldots, \widehat{s}_{i,\vartheta_g}, \widehat{s}_{i,\sigma_1^2}, \ldots, \widehat{s}_{i,\sigma_g^2}, \widehat{s}_{i,\lambda_1}, \ldots, \widehat{s}_{i,\lambda_g}, \widehat{s}_{i,\nu})^{\top}$. It is important to stress that the standard error of ν, obtained from $\widehat{s}_{i,\boldsymbol{\nu}}$, depends heavily on the calculation of conditional expectation $\mathrm{E}[\log(U_i)|y_{\mathrm{obs}_i}, \widehat{\boldsymbol{\theta}}]$, which relies on computationally intensive Monte Carlo integrations, since no analytical expression for this expected value exists. Therefore, the expressions for the elements $\widehat{s}_{i,\boldsymbol{\beta}}^{\top}, \widehat{s}_{i,p_j}, \widehat{s}_{i,\vartheta_j}, \widehat{s}_{i,\sigma_j^2}, \widehat{s}_{i,\lambda_j}$, for $j = 1, \ldots, g$, are given as

$$\widehat{s}_{i,\boldsymbol{\beta}}^{\top} = \frac{\sum_{j=1}^{G} p_j D_{\boldsymbol{\beta}}(y_i;\boldsymbol{\theta}_j)}{f(y_i;\boldsymbol{\theta})}, \widehat{s}_{i,\vartheta_j} = \frac{p_j D_{\vartheta_j}(y_i;\boldsymbol{\theta}_j)}{f(y_i;\boldsymbol{\theta})}, \widehat{s}_{i,\sigma_j^2} + \frac{p_j D_{\sigma_j^2}(y_i;\boldsymbol{\theta}_j)}{f(y_i;\boldsymbol{\theta})}, \ \widehat{s}_{i,\lambda_j} = \frac{p_j D_{\lambda_j}(y_i;\boldsymbol{\theta}_j)}{f(y_i;\boldsymbol{\theta})},$$

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

with

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

After some algebraic manipulation, we obtain

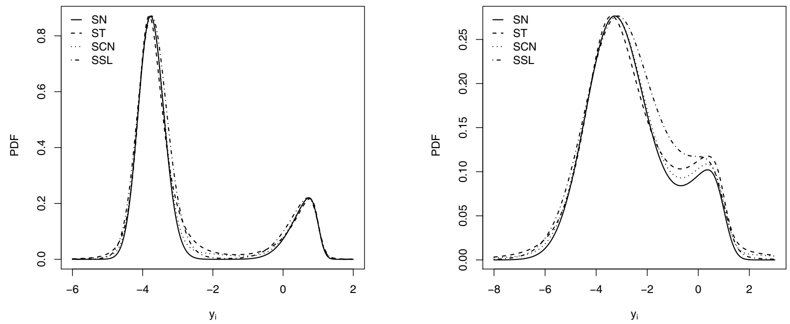
$$D_{\boldsymbol{\beta}}(y_i;\boldsymbol{\theta}_j) = \frac{2}{\sqrt{2\pi\sigma_j^2}}\Big[\sigma^{-2}(y_i - \mu_{ij} - b\Delta_j) I_{ij}^{\Phi}(3/2) - \sigma_j^{-1}\lambda_j I_{ij}^{\phi}(1)\Big]\boldsymbol{x}_i,$$

$$D_{\vartheta_j}(y_i;\boldsymbol{\theta}_j) = \frac{2}{\sqrt{2\pi\sigma_j^2}}\Big[\sigma_j^{-2}(y_i - \mu_{ij} - b\Delta_j) I_{ij}^{\Phi}(3/2) - \sigma_j^{-1}\lambda_j I_{ij}^{\phi}(1)\Big],$$

$$D_{\lambda_j}(y_i;\boldsymbol{\theta}_j) = \frac{2}{\sqrt{2\pi}\sigma_j^2}\left[\frac{(y_i - \mu_{ij} - b\Delta_j)b}{(1+\lambda_j^2)^{(3/2)}} I_{ij}^{\Phi}(3/2) + \left((y_i - \mu_{ij} - b\Delta_j) - \frac{b\Delta_j}{1+\lambda_j^2} I_{ij}^{\phi}(1)\right)\right],$$

$$\begin{aligned} D_{\sigma_j^2}(y_i;\boldsymbol{\theta}_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}}\Big[&-\sigma_j^{-2} I_{ij}^{\Phi}(1/2) + \sigma_j^{-4}(y_i - \mu_{ij} - b\Delta_j)^2 I_{ij}^{\Phi}(3/2) \\ &+ \sigma_j^{-4}(y_i - \mu_{ij} - b\Delta_j) b\Delta_j I_{ij}^{\Phi}(3/2) - \lambda_j \sigma_j^{-3}(y_i - \mu_{ij}) I_{ij}^{\phi}(1)\Big] \end{aligned}$$

where the expressions $I_{ij}^{\Phi}(w)$ and $I_{ij}^{\phi}(w)$ are given in Basso et al. (2010). The information-based approximation defined in Equation (10) is asymptotically applicable. However, it is less reliable unless the sample size is sufficiently large. Observe that the asymptotic covariance matrix of the ML estimates, that is, the inverse of Equation (10), was obtained using the parametrization $\varphi_j = \beta_0 + \mu_j$, $j = 1, \ldots, g$. We can use the traditional delta method (see Rao, 1973, Sec. 6a.2), to obtain standard errors using the original parameterization.



(a) (b)

Figure 2. Target mixture PDFs from simulated data in Scenario 1 (a) and Scenario 2 (b).

4. Numerical Studies

4.1 Parameter recovery (simulation study I)

We conduct three simulation studies to illustrate the performance of our proposed model. The first simulation presented below reports the consistency of the approximate standard errors for the ML estimators of parameters through the EM algorithm with each sample under the stopping criterion in Equation (9), whereas the contents of the second and third simulations are described in the corresponding subsections. In addition, we finish this section of numerical studies with an empirical illustration based on real data.

Here, 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

$$f(y_i|\boldsymbol{\theta}) = \sum_{j=1}^{2} p_j \phi_{\text{SMSN}}(y_i|\mu_{ij} + b\Delta_j, \sigma_j^2, \lambda_j, \boldsymbol{\nu}), \quad i = 1, \ldots, n,$$

where Z_{ij} is a component indicator of Y_i with $P(Z_{ij} = 1) = p_j$, $j = 1, 2$, $x_i^\top = (x_{i1}, x_{i2})$, such that $x_{i1} \sim \mathrm{U}(0, 1)$ and $x_{i2} \sim \mathrm{U}(0, 1)$, for $i = 1, \ldots, n$, and ε_1 and ε_2 follow a distribution as in the assumption given in Equation (3). 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 $p_1 = 0.2$. In addition, we consider the following scenarios (depicted in Figure 2): scenario 1 (well separated components) with $\sigma_1^2 = 0.2$ and $\sigma_2^2 = 0.4$, and scenario 2 (poorly separated components) with $\sigma_1^2 = 2$ and $\sigma_2^2 = 2$. For each combination of parameters, we generated 1000 Monte Carlo samples of size $n = 1000$ from the FM-SMSN-LR models, under four different situations: FM-SN-LR, FM-ST-LR ($\nu = 3$), FM-SSL-LR ($\nu = 3$) and FM-SCN-LR ($\boldsymbol{\nu}^\top = (0.1, 0.1)$). The average values and standard deviations (MC SD) of the estimators across the 1000 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 the Subsection 3.3. Moreover, we compute coverage probability of each parameter (COV), which is defined by $\mathrm{COV}(\widehat{\theta}) = (1/m)\sum_{j=1}^{m} I(\theta \in [\widehat{\theta}_{\mathrm{L}}, \widehat{\theta}_{\mathrm{U}}])$, where I is the indicator function such that θ lies in the interval $[\widehat{\theta}_{\mathrm{L}}, \widehat{\theta}_{\mathrm{U}}]$, with $\widehat{\theta}_{\mathrm{L}}$ and $\widehat{\theta}_{\mathrm{U}}$ being estimated lower and upper bounds of the 95% CI, respectively. The results are presented in Table 1. Note that under both scenarios (well and poorly separated components), the results suggest that the proposed FM-SMSN-LR model produces satisfactory estimates.

It can bee seen 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

Table 1. Simulation study I: mean and MC SD are the respective estimated means and standard deviations from fitting a FM-SMSN-LR model based on 1000 samples. IM SE is the average value of the approximate standard error obtained through the information-based method. COV is the coverage probability. True values of parameters are in parentheses.

Parameter		Scenario 1: $\sigma_1^2 = 0.2$, $\sigma_2^2 = 0.4$				Scenario 2: $\sigma_1^2 = \sigma_2^2 = 2$			
		SN	ST($\nu = 3$)	SCN ($\nu = 0.1$)	SSL($\nu = 3$)	SN	ST($\nu = 3$)	SCN ($\nu = 0.1$)	SSL($\nu = 3$)
$\beta_0(-1)$	Mean	-0.9971	-1.0038	-0.9953	-0.9989	-1.0119	-1.0070	-0.9965	-1.0413
	IM SE	0.0602	0.0859	0.0777	0.0883	0.1928	0.3345	0.2369	0.3238
	MC SD	0.0698	0.0755	0.0713	0.0770	0.0925	0.1214	0.1324	0.1284
	COV	90.6%	96.7%	96.6%	96.0%	99.4%	95.7%	91.8%	95.8%
$\beta_1(-4)$	Mean	-4.0002	-3.9985	-3.9996	-3.9947	-3.9949	-3.9958	-3.9963	-4.0005
	IM SE	0.0368	0.0418	0.0402	0.0423	0.0889	0.1021	0.0974	0.0985
	MC SD	0.0365	0.0426	0.0403	0.0449	0.0899	0.1076	0.0950	0.1031
	COV	94.7%	94.2%	95.5%	95.0%	95.0%	92.9%	95.4%	93.3%
$\beta_2(-3)$	Mean	-3.0012	-2.9998	-3.0014	-2.9938	-2.9994	-2.9989	-2.9967	-3.0013
	IM SE	0.0374	0.0424	0.0410	0.0432	0.0859	0.1005	0.0975	0.1020
	MC SD	0.0370	0.0442	0.0413	0.0430	0.0836	0.1046	0.0977	0.1109
	COV	95.6%	93.7%	94.0%	96.0%	96.2%	94.4%	94.2%	92.0%
$\mu_1(-4)$	Mean	-4.0026	-3.9945	-4.0040	-4.0166	-4.0295	-3.9806	-4.0899	-3.9924
	IM SE	0.0853	0.0800	0.0894	0.0854	0.1396	0.2782	0.1896	0.2531
	MC SD	0.0691	0.0876	0.0744	0.0859	0.1111	0.3161	0.2483	0.2202
	COV	98.2%	99.8%	98.6%	98.6%	97.3%	92.3%	84.5%	94.8 %
$\mu_2(1)$	Mean	0.9992	1.0012	1.0007	0.9945	0.9990	1.0103	1.0391	0.9955
	IM SE	0.0837	0.0878	0.0862	0.0873	0.0744	0.1098	0.0861	0.0983
	MC SD	0.0630	0.0625	0.0656	0.0625	0.0692	0.1000	0.1060	0.0813
	COV	98.3%	99.7%	98.4%	99.0%	96.7%	96.7%	86.4%	97.7%
σ_1^2	Mean	0.2097	0.2089	0.2084	0.1946	2.0069	2.2009	1.9385	1.9221
	IM SE	0.0680	0.0575	0.0643	0.0543	1.4238	0.9880	0.7385	1.5234
	MC SD	0.0427	0.0639	0.0644	0.0539	0.5626	1.0118	0.8238	0.9698
	COV	88.7%	89.8%	88.9%	89.0%	99.6%	87.3%	83.3%	89.1%
σ_2^2	Mean	0.3991	0.4026	0.3940	0.3988	2.0452	1.9839	1.8290	2.1521
	IM SE	0.0274	0.0385	0.0343	0.0381	0.1978	0.3796	0.1898	0.2758
	MC SD	0.0283	0.0501	0.0423	0.0463	0.1816	0.2642	0.3309	0.3109
	COV	94.0%	85.9%	85.5%	88.0%	95.9%	93.7%	72.5%	89.2%
$\lambda_1(1)$	Mean	1.0916	1.0534	1.0894	0.9679	1.1614	1.0068	0.6175	0.8514
	IM SE	0.7420	0.4956	0.6466	0.4814	1.4279	1.0923	1.2206	2.7316
	MC SD	0.8216	0.4983	0.6441	0.4385	0.4974	0.7792	1.3124	1.1426
	COV	94.3%	96.3%	95.9%	98.0%	99.6%	96.9%	88.4%	92.4%
$\lambda_2(-4)$	Mean	-4.0874	-4.1108	-4.0739	-4.1418	-4.2153	-4.0168	-3.7773	-4.0682
	IM SE	0.5446	0.5969	0.5971	0.6086	0.6299	0.8950	0.6262	0.6219
	MC SD	0.5406	0.6141	0.6007	0.5477	0.5967	0.6555	0.8671	0.6494
	COV	96.8%	95.5%	94.3%	96.0%	96.8%	94.5%	86.8%	93.6%
$p_1(0.2)$	Mean	0.1998	0.2004	0.1999	0.1985	0.1987	0.2033	0.2028	0.2000
	IM SE	0.0126	0.0131	0.0130	0.0131	0.0146	0.2218	0.0159	0.0204
	MC SD	0.0126	0.0125	0.0129	0.0127	0.0138	0.0235	0.0213	0.0191
	COV	95.3%	95.8%	95.0%	94.0%	96.3%	92.9%	87.3%	94.6%
ν	Mean	-	3.0735	0.1070	2.9791	-	3.2216	0.1342	4.4543
$\gamma(0.1)$	Mean	-	-	0.1098	-	-	-	0.1415	-

approximation for the variances of the ML estimates of Equation (10) is reliable. Note also that the coverage probability (COV) for the regression parameters is quite stable for two scenarios, indicating that the proposed asymptotic approximation for the variance estimates of the ML estimates is reliable.

4.2 Asymptotic properties of the EM estimates (simulation study II)

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 $x_i^\top = (x_{i1}, x_{i1})$, such that $x_{i1} \sim \mathrm{U}(0,1)$ and $x_{i2} \sim \mathrm{U}(0,1)$, for $i = 1, \ldots, n$. We choose sample sizes $n = 100$, 250, 500, 1000, 2500 and 5000. The true values of the parameters were taken as $\beta_0 = -1$, $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top = (-4, -3)^\top$, $\mu_1 = -4$, $\mu_2 = 1$, $\sigma_1^2 = 0.2$, $\sigma_2^2 = 0.4$ and $p_1 = 0.2$. For each combination of parameters and sample sizes, we generated 1000 random samples from the FM-SMSN-LR models, under three different situations: FM-SN-LR, FM-ST-LR ($\nu = 3$), FM-SSL-LR ($\nu = 3$) and FM-SCN-LR ($\boldsymbol{\nu}^\top = (0.1, 0.1)$). 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

$$\text{Bias}(\theta_i) = \frac{1}{1000}\sum_{i=1}^{1000}(\theta_i^{(j)} - \theta_i) \ \text{ and } \ \text{RMSE}(\theta_i) = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}(\theta_i^{(j)} - \theta_i)^2},$$

where $\widehat{\theta}_i^{(j)}$ is the estimate of θ_i for the jth sample. The results for β_0, β_1 and β_2 are shown in Figure 3; the results for μ_1, σ_1 and λ_1 are shown in Figure 4; the results for μ_2, σ_2, λ_2 are shown in Figure 5; and the results for p_1 are shown in Figure 6. 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.

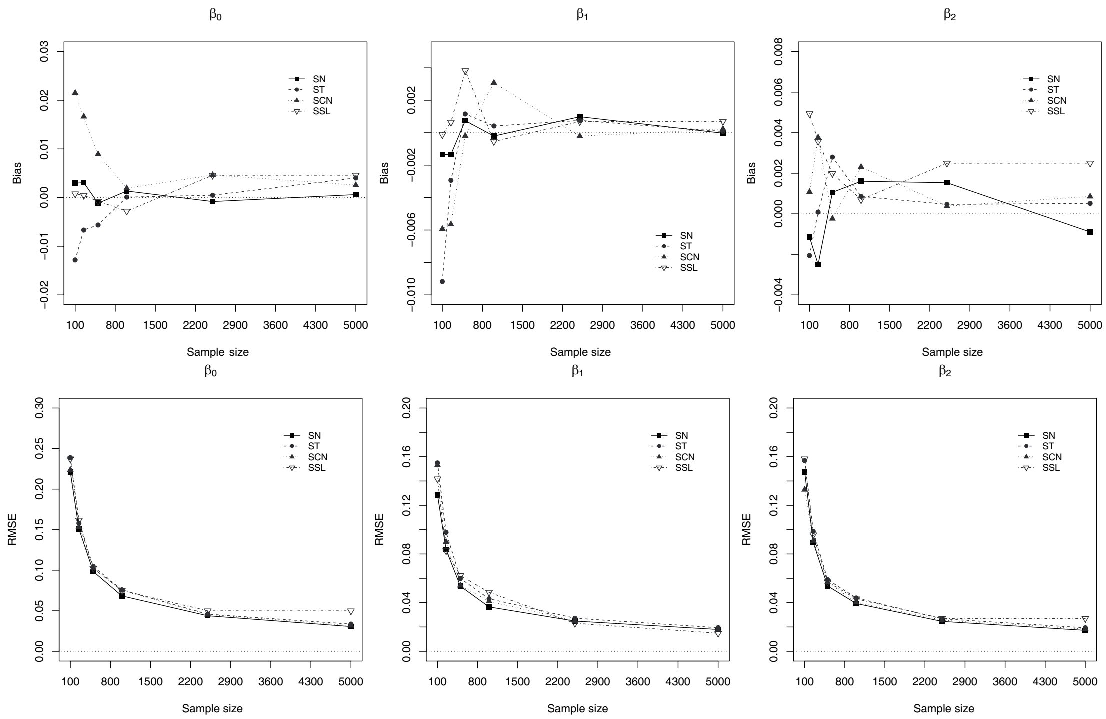


Figure 3. Average bias (1st row) and average RMSE (2nd row) of the estimators of β_0,β_1, β_2 for simulation II.

4.3 Robustness of the EM estimates (simulation study III)

The purpose of this simulation study is to compare the effect of the robustness of the estimates of the FM-SMSN-LR models in the presence of outliers on the response variable. We compare the FM-SN-LR, FM-ST-LR ($\nu = 3$), FM-SSL-LR ($\nu = 3$) and the FM-CN-LR ($(\nu, \gamma) = (0.1, 0.1)$) models. In this scenario, we generated 500 samples of size $n = 500$ of the FM-SMSN-LR model with $f(\varepsilon_i) = \sum_{j=1}^{2} p_j \phi_{\text{SMSN}}(\varepsilon_i|\mu_j + b\Delta_j, \sigma_j^2, \lambda_j, \boldsymbol{\nu})$. The true values of the parameters were taken as $\beta_0 = -1$, $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top = (-4, -3)^\top$, $\mu_1 = -4$, $\mu_2 = 1$, $\sigma_1^2 = 0.2$, $\sigma_2^2 = 0.4$ and $p_1 = 0.2$. To assess how much the EM estimates are influenced by the presence of outliers, we replaced observation y_{150} by $y_{150}(\upsilon) = y_{150} + \upsilon$, with $\upsilon = 1, 2, \ldots, 10$. For each replication, we obtained the parameter estimates with and without outliers, with the three FM-SMSN-LR models.

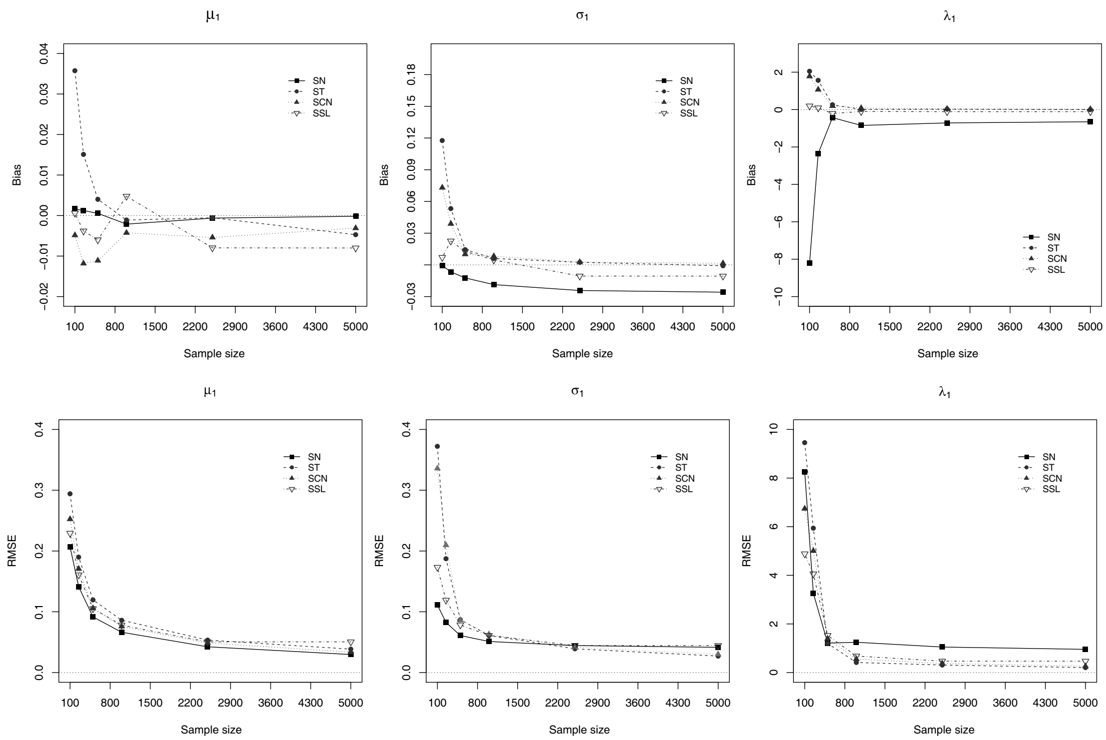


Figure 4. Average bias (1st row) and average RMSE (2nd row) of the estimators of μ_1,σ_1, λ_1 for simulation II.

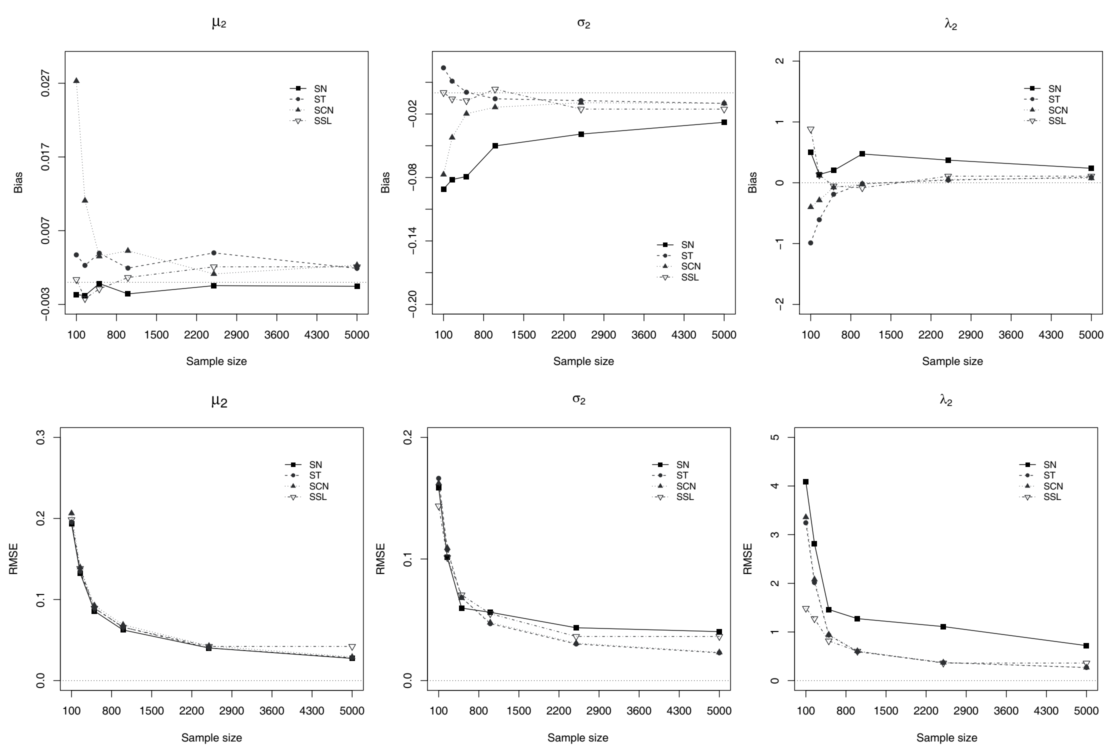


Figure 5. Average bias (1st row) and average RMSE (2nd row) of the estimators of μ_2,σ_2, λ_2 for simulation II.

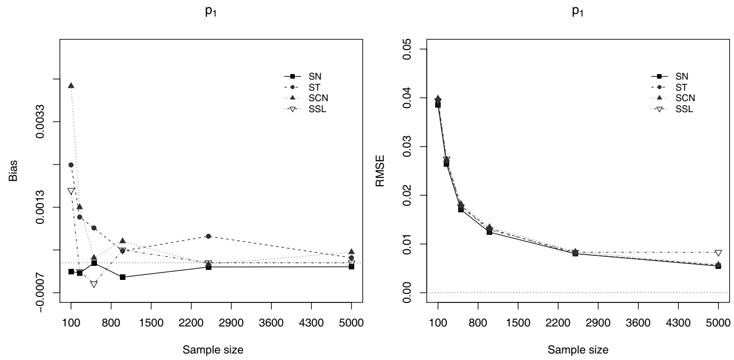


Figure 6. Average bias (1st row) and average RMSE (2nd row) of the estimators of p_1 for simulation II.

We are interested in evaluating the relative change (RC) in the estimates as a function of υ. 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 RC is defined by

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

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

Figure 7 shows the average values of the relative changes undergone by all the parameters. We note that for all parameters, the average RCs suddenly increase under FM-SN-LR model as the υ 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.

4.4 Empirical illustration

Next, the proposed techniques are illustrated with the analysis a real dataset, the one previously analyzed by Cook and Weisberg (1982) in a normal regression setting. The dataset comes from the Australian Institute of Sport (AIS) and consists of measurements of 202 athletes. Here, we focus on percent body fat (Bfat), which is assumed to be explained by the sum of skin folds (ssf) and height in cm (Ht). Thus, we consider the FM-SMSN-LR model given by

$$\mathrm{Bfat}_i = \beta_0 + \beta_1 \mathrm{ssf}_i + \beta_2 \mathrm{Ht}_i + \varepsilon_i, \quad i = 1, \ldots, 202,$$

where ε_i belongs to the FM-SMSN family.

By using the `FMsmsnReg` package (see the appendix), we fit the FM-SMSN-LR models as was described in Section 3. Table 2 compares the fit of various mixture models for $g = 1$ to 5 components, using the model selection criteria discussed in Subsection 3.3. Note from this table that, as expected, the heavy-tailed models perform significantly better than the SN model (and the symmetric counterparts such as the normal and Student-t models), with mixtures of two ($g = 2$) components being significantly better in all cases, except for the normal case (FM-N), where a mixture of $g = 3$ is needed.

Moreover, the 2-component FM-ST-LR model fits the data substantially better. This conclusion also is verified through a hypotheses procedure for testing the number of components in the FM-ST-LR model. As reported by Turner (2000), we can use parametric

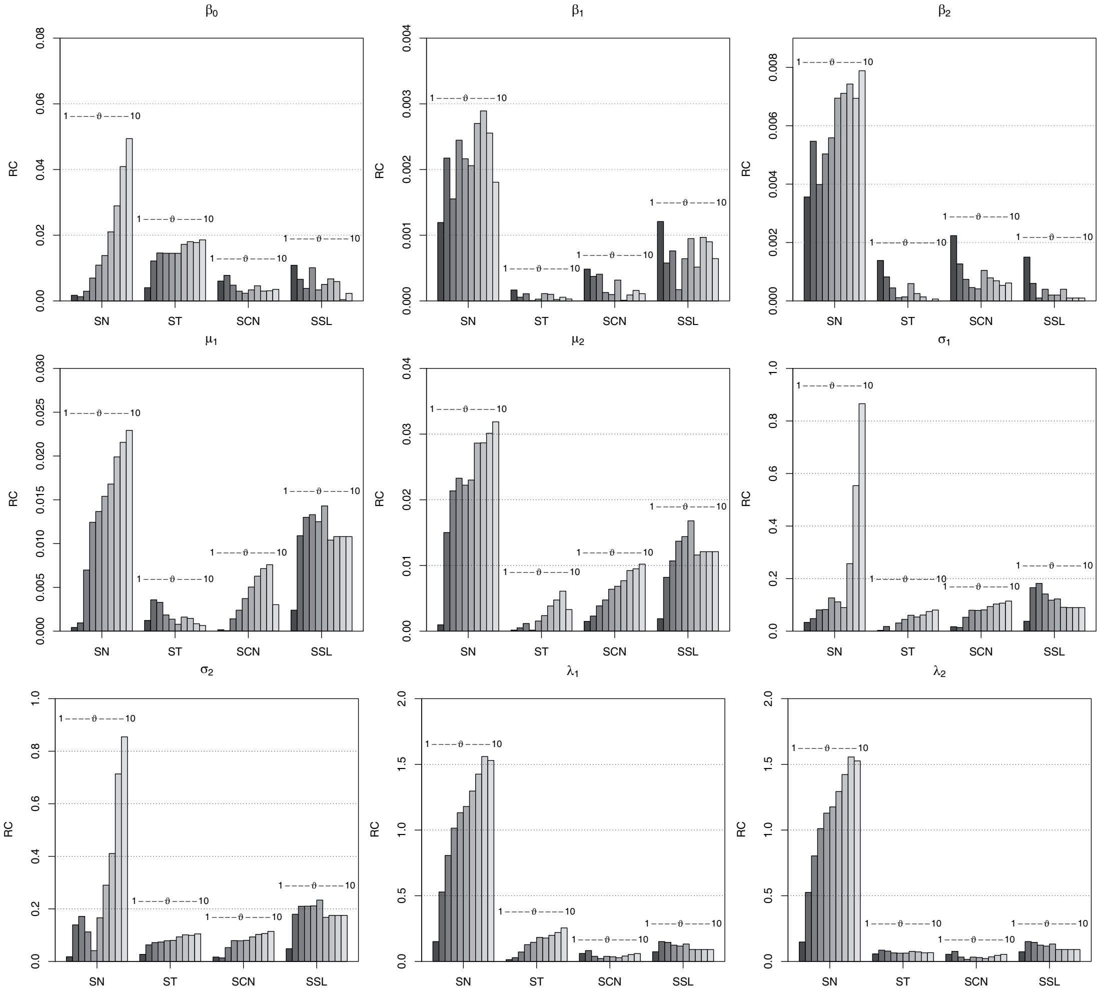


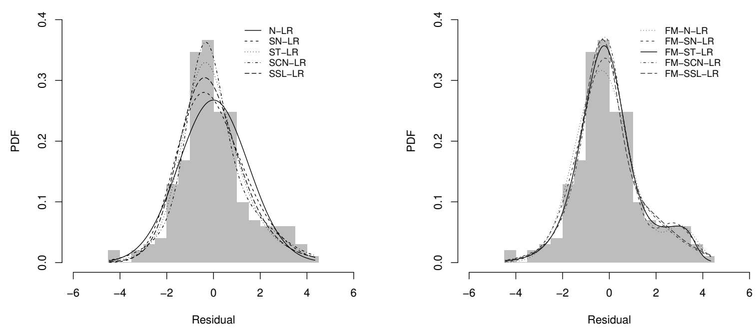
Figure 7. Average RCs of estimates with different perturbations υ for simulation study III.

Table 2. Comparison of maximum log-likelihood, AIC$_c$ and BIC$_A$ for fitted FM-SMSN-LR models using the AIS data. The number of parameters is denoted by m.

Model	g	m	log-lik	AIC$_c$	BIC$_a$
FM-N	1	5	-367.2395	744.7850	745.1792
FM-N	2	8	-359.2902	735.3265	735.7009
FM-N	3	11	-355.2892	733.9679	734.1192
FM-T	1	6	-363.9525	738.2111	738.6053
FM-T	2	9	-358.2494	733.2449	733.6194
FM-T	3	12	-356.3237	736.0369	736.1881
FM-SN	1	6	-363.0346	738.5001	738.9097
FM-SN	2	10	-356.3079	733.7675	734.0164
FM-SN	3	14	-354.1438	738.5336	738.2486
FM-SN	4	18	-353.1388	746.0152	744.7987
FM-SN	5	22	-352.2579	754.1695	751.5973
FM-ST	1	7	-360.7632	736.1038	736.5070
FM-ST	2	11	-353.9696	731.3286	731.4799
FM-ST	3	15	-353.8492	740.2790	739.7994
FM-ST	4	19	-352.3138	746.8034	745.2888
FM-ST	5	23	-351.7865	755.7752	752.7944
FM-SCN	1	8	-357.0375	738.5001	738.9097
FM-SCN	2	12	-353.7235	733.0978	733.1278
FM-SCN	3	16	-354.1656	743.2717	742.5722
FM-SCN	4	20	-352.0380	748.7169	746.8773
FM-SCN	5	24	-352.8184	760.4164	756.9983
FM-SSL	1	7	-362.3246	739.2264	739.6296
FM-SSL	2	11	-354.1580	731.7054	731.8566
FM-SSL	3	15	-354.1941	740.9689	740.4892
FM-SSL	4	19	-352.2586	746.6930	745.1785
FM-SSL	5	23	-352.3504	756.9031	753.9224

Table 3. AIS 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	14.7241	0.0001	14.51593	0.00253	14.6622	0.0025	14.7475	0.0025
β_1	0.1799	0.0012	0.17972	0.00850	0.1805	0.0089	0.1796	0.0091
β_2	-0.0757	0.1302	-0.07536	0.19264	-0.0757	0.1458	-0.0754	0.1513
p_1	0.1543	0.9295	0.15418	1.04192	0.1483	1.0841	0.1514	1.0393
μ_1	2.5504	2.2932	1.93244	4.00942	2.3654	3.8355	2.3891	3.9553
μ_2	-0.4652	1.8546	-0.35226	2.94875	-0.4120	2.5091	-0.4263	2.6266
σ_1^2	0.8483	0.5074	3.80681	1.57056	2.2957	1.6255	2.3158	1.6615
σ_2^2	2.2793	0.4021	1.06550	11.56693	1.1240	7.1021	0.9740	7.0029
λ_1	0.1624	0.8467	-5.70438	0.52991	-3.5415	0.4408	-4.8612	0.3724
λ_2	-2.2318	1.7509	-0.62860	9.52263	-1.0111	7.9389	-1.0144	11.9961
ν	-	-	7.45874	-	0.2270	-	2.3036	-
γ	-	-	-		0.3075	-	-	



(a) (b)

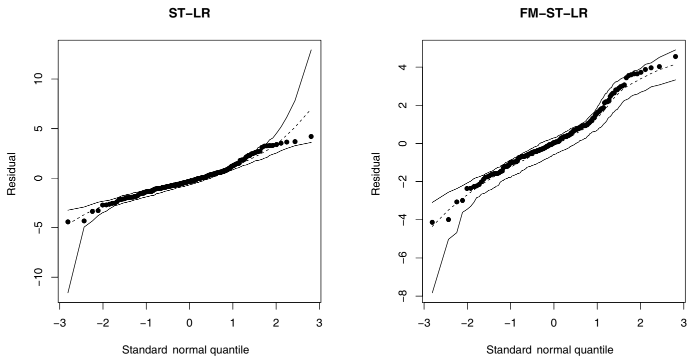
Figure 8. Panels (a) and (b) display the histogram ordinary residuals superimposed on the FM-SMSN-LR residual PDF for $g = 1$ and $g = 2$ components, respectively with AIS dataset.

or semiparametric bootstrap to test hypotheses concerning the number of components in the mixture. Following the method proposed by Turner (2000), we considered 1000 bootstrap statistics for testing $g = 1$ versus $g = 2$, in which case the p-value was 0.027 for the parametric bootstrap. Accordingly, there is strong evidence that at least two components are needed. For testing $g = 2$ versus $g = 3$, the bootstrap p-value was 0.984, so there is no evidence that more than two components are required to model the AIS dataset.

Table 3 presents the ML estimates of the parameters considering the four models with $g = 2$, say FM-SN-LR, FM-ST-LR, FM-SCN-LR and the FM-SSL-LR, along with the corresponding standard errors (SE), obtained via the information-based procedure presented in Subsection 3.3. Notice from Table 3 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 assumption.

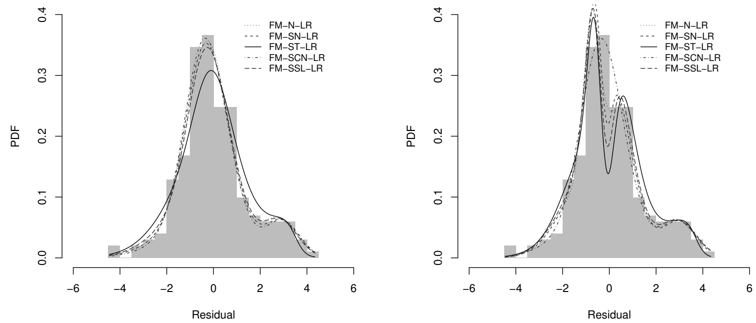
In Figure 8, we plot the histogram of OLS residuals and then display the residual PDFs for the four FM-SMSN-LR models superimposed on a single set of coordinate axes, with $g = 1$ and $g = 2$ components respectively. Additional results related to $g = 3$ and $g = 4$ components are given in Figure 10. Based on this graphical representation, it appears once again that the FM-ST-LR, FT-SCN-LR and the FT-SSL-LR models have quite reasonable and better fit than the FM-SN-LR model with $g = 2$ components.

In order to detect incorrect specification of the error distribution for our best model (FM-ST-LR), we present quantile versus quantile (QQ) plots and simulated envelopes for the residuals $(y - \hat{y})$ in Figure 9. The QQ plots for the other models are given in Figure 11. This figure provides strong evidence that the FM-ST-LR (with $g = 2$ components) yields a better fit to the current data than the ST-LR model (with $g = 1$ component), since there are no observations falling outside the envelope.



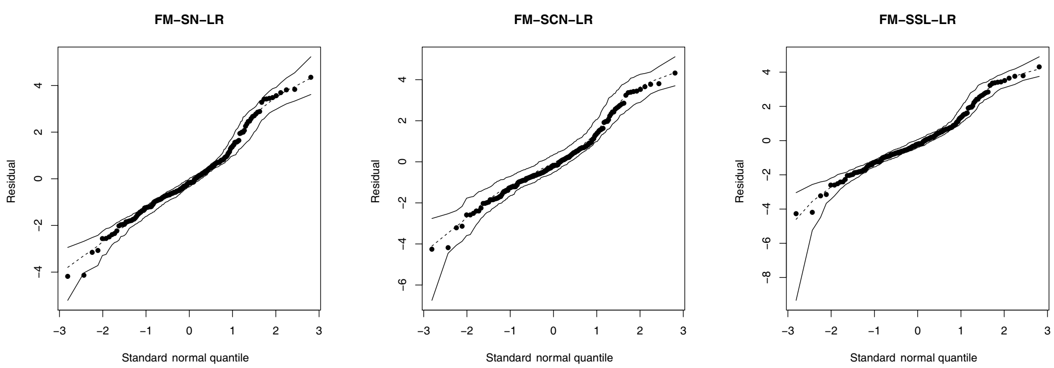
(a) (b)

Figure 9. Panels (a) and (b) display the QQ plots and simulated envelopes for the residual $(y - \widehat{y})$ with for $g = 1$ and $g = 2$ components, respectively with AIS dataset.



(a) (b)

Figure 10. Panels (a) and (b) display the histogram of ordinary residuals with FM-SMSN-LR residual with for $g = 3$ and $g = 4$ components, respectively with AIS dataset.



(a) (b) (c)

Figure 11. Panels (a), (b) and (c) display the QQ plots and simulated envelopes for the residual $(y - \widehat{y})$ for $g = 2$ components based on FM-SN, FM-SCN and FM-SSL distributions, respectively with AIS dataset.

5. Conclusions

In this paper we consider a regression model whose error term follows a finite mixture of 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, simultaneously accommodating 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 of Zeller et al. (2016), 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 PDF estimators, to model the unknown distributional shape of the errors. For this structure we developed a simple EM-type algorithm to perform ML inference of the parameters with closed-form expression at the E-step. The proposed methods are implemented using the `FMsmsnReg` package, 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 a real dataset and several simulation studies.

The proposed methods can be extended to multivariate settings using the multivariate SMSN class of distributions (Cabral et al., 2012), 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).

Appendix: Sample output from the *FMsmsnReg* package

```
------------------------------
Finite Mixture of Scale Mixture Skew Normal Regression Model
------------------------------
Observations = 202
Family = Skew.t

------
Estimates
------
       Estimate      SE
beta0  14.51593 0.00253
beta1   0.17972 0.00850
beta2  -0.07536 0.19264
mu1     1.93244 4.00942
mu2    -0.35226 2.94875
sigma1  3.80681 1.57056
sigma2  1.06550 11.5669
shape1 -5.70438 0.52991
shape2 -0.62860 9.52263
pii1    0.15418 1.04192
nu      7.45874      NA

------------
Model selection criteria
------------
Loglik     AIC     BIC     EDC      ICL
Value -357.030 730.235 766.626 739.502 2916.687

----
Details
----
Convergence reached? = TRUE
EM iterations = 147 / 500
Criteria = 1e-07
Processing time = 27.11465 secs
```

Acknowledgment

The authors thank the Editors and Reviewers for their constructive comments on an earlier version of this manuscript. Luis Benites and Rocío Maehara acknowledges support from CNPq-Brazil. Victor H. Lachos was supported from CNPq-Brazil (Grant 306334/2015-1). Partial support from CAPES is also acknowledged.

References

Andrews, D.F. and Mallows, C.L., 1974. Scale mixtures of normal distributions. Journal of the Royal Statistical Society B, 36, 99-102.

Arellano-Valle, R.B. and Genton, M.G., 2010. Multivariate unified skew-elliptical distributions. Chilean Journal of Statistics, 1, 17-33.

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

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

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

Basford, K., Greenway, D., McLachlan, G., and 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 and Data Analysis, 54, 2926-2941.

Benites, L., Maehara, R., and Lachos, V.H., 2016. FMsmsnReg: Regression Models with Finite Mixtures of Skew Heavy-Tailed Errors. R package version 1.0. http://CRAN.R-project.org/package=FMsmsnReg

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P., and Lindsay, B.G., 1994. The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. Annals of the Institute of Statistical Mathematics, 46, 373-388.

Branco, M.D. and 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., and Prates, M.O., 2012. Multivariate mixture modeling using skew-normal independent distributions. Computational Statistics and Data Analysis, 56, 126-142.

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

Cook, R.D. and Weisberg, S., 1982. Residuals and Influence in Regression. Chapman and Hall/CRC, Boca Raton, FL, US.

Dempster, A., Laird, N., and Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society B, 39, 1-38.

Galea, M., Paula, G.A., and Bolfarine, H., 1997. Local influence in elliptical linear regression models. Journal of the Royal Statistical Society D, 46, 71-79.

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

Garay, A.M., Lachos, V.H., and Abanto-Valle, C.A., 2011. Nonlinear regression models based on scale mixtures of skew-normal distributions. Journal of the Korean Statistical Society, 40, 115-124.

Holzmann, H. and Munk, A., 2006. Identifiability of finite mixture of elliptical distributions. Scandinavian Journal of Statistics, 33, 753-763.

Hurvich, C. and Tsai, C.L., 1989. Regression and time series model selection in small samples. Biometrika, 76, 297-307.

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

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

Lange, K.L. and Sinsheimer, J.S., 1993. Normal/independent distributions and their applications in robust regression. Journal of Computational and Graphical Statistics, 2, 175-198.

Lindsay, B., 1995. Mixture models: Theory, geometry and applications. Institute of Mathematical Statistics, Hayward, CA, US.

Liu, C. and Rubin, D.B., 1994. The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. Biometrika, 80, 267-278.

McNicholas, P., Murphy, T., and McDaid, A.D.F., 2010. Serial and parallel implementations of model-based clustering via parsimonious gaussian mixture models. Computational Statistics and Data Analysis, 54, 711-723.

Otianiano, C., Rathie, P., and Ozelim, L., 2015. On the identifiability of finite mixture of skew-normal and skew-t distributions. Statistics and Probability Letters, 106, 103-108.

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

Prates, M.O., Dey, D.K., and Lachos, V.H., 2012. A dengue fever study in the state of Rio de Janeiro with the use of generalized skew-normal/independent spatial fields. Chilean Journal of Statistics, 3, 143-155.

Quandt, R.E. and Ramsey, J.B., 1978. Estimating mixtures of normal distributions and switching regressions. Journal of the American Statistical Association, 73, 730-738.

R Core Team, 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rao, C.R., 1973. Linear Statistical Inference and its Applications. Wiley, New York.

Sclove, S., 1987. Application of model-selection criteria to some problems in multivariate analysis. Psychometrika, 52, 333-343.

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

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 C, 49, 371-384.

Zeller, C.B., Cabral, C.R., Lachos, V.H., 2016. Robust mixture regression modeling based on scale mixtures of skew-normal distributions. TEST, 25, 375-396.

Information for authors

The editorial board of the Chilean Journal of Statistics (ChJS) is seeking papers, which will be refereed. We encourage the authors to submit a PDF electronic version of the manuscript in a free format to Víctor Leiva, Editor-in-Chief of the ChJS (E-mail: chilean.journal.of.statistics@gmail.com). Submitted manuscripts must be written in English and contain the name and affiliation of each author followed by a leading abstract and keywords. The authors must include a"cover letter" presenting their manuscript and mentioning: "We confirm that this manuscript has been read and approved by all named authors. In addition, we declare that the manuscript is original and it is not being published or submitted for publication elsewhere".

Preparation of accepted manuscripts

Manuscripts accepted in the ChJS must be prepared in Latex using the ChJS format. The Latex template and ChJS class files for preparation of accepted manuscripts are available at http://chjs.mat.utfsm.cl/files/ChJS.zip. Such as its submitted version, manuscripts accepted in the ChJS must be written in English and contain the name and affiliation of each author, followed by a leading abstract and keywords, but now mathematics subject classification (primary and secondary) are required. AMS classification is available at http://www.ams.org/mathscinet/msc/. Sections must be numbered 1, 2, etc., where Section 1 is the introduction part. References must be collected at the end of the manuscript in alphabetical order as in the following examples:

Arellano-Valle, R., (1994). Elliptical Distributions: Properties, Inference and Applications in Regression Models. Unpublished Ph.D. Thesis. Department of Statistics, University of São Paulo, Brazil.

Cook, R.D., (1997). Local influence. In Kotz, S., Read, C.B. and Banks, D.L. (Eds.), Encyclopedia of Statistical Sciences, Update, Vol. 1, Wiley, New York, pp. 380-385.

Rukhin, A.L., (2009). Identities for negative moments of quadratic forms in normal variables. Statistics and Probability Letters 79, 1004-1007.

Stein, M.L., (1999). Statistical Interpolation of Spatial Data: Some Theory for Kriging. Springer, New York.

Tsay, R.S., Pena, D., and Pankratz, A.E., (2000). Outliers in multivariate time series. Biometrika 87, 789-804.

References in the text must be given by the author's name and year of publication, e.g., Gelfand and Smith (1990). In the case of more than two authors, the citation must be written as Tsay et al. (2000).

Copyright