Finite Mixtures of Regression Models

Luis Enrique Benites Sánchez

TESE APRESENTADA
AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA
OBTENÇÃO DO TÍTULO
DE
DOUTOR EM CIÊNCIAS

Programa: Estatística

Orientador: Prof. Dr. Heleno Bolfarine

Coorientador: Prof. Dr. Víctor Hugo Lachos Dávila

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da CAPES e do
CNPq

São Paulo, maio de 2018

Finite Mixtures of Regression Models

Esta é a versão original da tese elaborada pelo
candidato Luis Enrique Benites Sánchez, tal como
submetida à Comissão Julgadora.

Finite Mixtures of Regression Models

Esta versão da tese contém as correções e alterações sugeridas pela Comissão Julgadora durante a defesa da versão original do trabalho, realizada em 6/4/2018. Uma cópia da versão original está disponível no Instituto de Matemática e Estatística da Universidade de São Paulo.

Comissão Julgadora:

- Prof. Dr. Heleno Bolfarine
- Prof. Dr. Víctor Hugo Lachos Dávila
- Prof. Dr. Marcos Oliveira Prates
- Prof. Dr. Celso Rômulo Barbosa Cabral
- Prof. Dr. Mário de Castro Andrade Filho

Agradecimentos

A Jesus, meu Senhor e Salvador, pois reconheço que, se não fosse por Ele, não teria feito nada.

- A minha esposa Rocío, por seu apoio, compreensão e paciência.
- A minha mãe, Delia, que sempre me apoia em tudo e dava força para lutar.
- Aos meus irmãos, Christian, Marcos e Susan, por seu constante apoio.
- A minha avó, Isabel, porque sei que sempre ora por mim.
- Aos meus tios, Luis e Miro, pelos conselhos e apoio.
- A minha tia Isabel, pelo apoio brindado no decorrer de meu desenvolvimento profissional.
- A minha prima Ariana, pelo apoio.
- Ao meu orientador, Professor Heleno Bolfarine, um excelente pesquisador e pessoa.
- Ao meu coorientador, Professor Prof. Víctor Hugo Lachos Dávila, pela amizade, oportunidade, ânimo, motivação e paciência.
- Ao Professor Filidor Vilca pela amizade e aconselhamentos.
- Aos professores do Departamento de Estatística do IME-USP, pelos ensinamentos concedidos.
- À CAPES e ao CNPq, pelo apoio financeiro recebido na elaboração deste trabalho.

Resumo

BENITES, L. E. **Mistura de modelos de regressão**. 2018. 121 f. Tese (Doutorado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2018.

Esta tese composta por três artigos, visa propor extensões das misturas finitas nos modelos de regressão. Aqui vamos considerar uma classe flexível de distribuições tanto univariada como multivariada, que permitem modelar adequadamente dados assimétricos, que apresentam multimodalidade, caldas pesadas e observações atípicas. Esta classe possui casos especiais tais como as distribuições skew-normal, skew-t, skew slash, skew normal contaminada, assim como os casos simétricos. Inicialmente, é proposto um modelo baseado na suposição de que o erros seguem uma mistura finita da distribuição mistura de escala skew-normal (SMSN) ao invés da convencional distribuição normal. Em seguida, temos um modelo de regressão censurado onde consideramos que o erro segue uma mistura finita da distribuição da mistura de escala normal (SMN). E por último, é considerada uma mistura finita de regressão multivariada onde o erro tem uma distribuição SMSN multivariada. Para todos os modelos propostos foram desenvolvidos dois pacotes do software R, que estão exemplificados no apêndice.

Palavras-chave: Modelo de regressão linear; Mistura finita; Mistura da escala da distribuição normal assimétrica multivariada; Dados censurados; Algoritmo EM.

Abstract

BENITES, L. E. **Finite Mixtures of regression models**. 2010. 121 f. Tese (Doutorado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2018.

This dissertation consists of three articles, proposing extensions of finite mixtures in regression models. Here we consider a flexible class of both univariate and multivariate distributions, which allow adequate modeling of asymmetric data that have multimodality, heavy tails and outlying observations. This class has special cases such as skew-normal, skew-t, skew-slash and skew normal contaminated distributions, as well as symmetric cases. Initially, a model is proposed based on the assumption that the errors follow a finite mixture of scale mixture of skew-normal (FM-SMSN) distribution rather than the conventional normal distribution. Next, we have a censored regression model where we consider that the error follows a finite mixture of scale mixture of normal (SMN) distribution. Next, we propose a censored regression model where we consider that the error follows a finite mixture of scale mixture of normal (SMN) distribution. Finally, we consider a finite mixture of multivariate regression where the error has a multivariate SMSN distribution. For all proposed models, two R packages were developed, which are reported in the appendix.

Palavras-chave: Linear regression model; Multivariate scale mixture of skew normal distributions; Censored data, EM algorithm; Finite mixture.

Contents

List of Figures	xi
List of Tables	xiii
1 Introduction	1
1.1 Preliminaries	1
1.1.1 Truncation	2
1.1.2 Censoring	3
1.2 Dissertation organization	3
2 Robust regression models using finite mixtures of skew heavy-tailed distributions	5
2.1 Introduction	5
2.2 Scale mixtures of skew-normal distributions	8
2.3 The SMSN linear regression model	10
2.3.1 Parameter estimation via an EM-type algorithm	10
2.4 The linear regression model with FM–SMSN errors	12
2.4.1 The model	12
2.4.2 Parameter estimation via the ECME algorithm	14
2.4.3 Starting values	16
2.4.4 Model selection	16
2.4.5 Approximated standard errors	17
2.5 Simulation experiments	18
2.5.1 Parameter recovery (simulation study 1)	19
2.5.2 Asymptotic properties of the EM estimates (simulation study 2)	21
2.5.3 Robustness of the EM estimates (simulation study 3)	22
2.6 Application	22
2.7 Conclusions	27
3 Regression modeling of censored data based on mixtures of scale mixtures of normal distributions	29
3.1 Introduction	29
3.2 The model	31
3.2.1 Some characteristics of the SMN-CR model	31
3.2.2 The model	33

3.2.3	Parameter estimation via an EM-type algorithm	33
3.3	The FM-SMN-CR model	36
3.3.1	Identifiability	38
3.3.2	Maximum likelihood estimation via EM-type algorithm	41
3.3.3	Starting values	44
3.3.4	Stopping rule	45
3.3.5	Variance matrix and asymptotic properties	45
3.4	Simulation studies	47
3.4.1	Simulation study 1: parameter recovery	49
3.4.2	The effect of different censoring proportions (simulation study 2)	50
3.4.3	The effect of different sample sizes (simulation study 3)	51
3.5	Application	53
3.6	Conclusions	57
4	Finite mixture of regression models based on multivariate scale mixtures of skew-normal distributions	59
4.1	Introduction	59
4.2	The multivariate scale mixtures of skew-normal distributions	61
4.2.1	Preliminaries	61
4.2.2	Multivariate skew-normal distribution	63
4.2.3	Multivariate skew- t distribution	64
4.2.4	Multivariate skew-slash distribution	64
4.2.5	Multivariate skew-contaminated normal distribution	64
4.3	The linear regression model for multivariate SMSN responses	65
4.3.1	Parameter estimation via the EM algorithm	65
4.4	The FM-SMSN-MR model	68
4.4.1	The model and parameter estimation	68
4.4.2	Starting values	71
4.4.3	Aitken's acceleration stopping criterion	72
4.4.4	Obtaining of standard errors	73
4.5	Simulation study	75
4.5.1	Parameter recovery (simulation study 1)	76
4.5.2	Asymptotic properties of the EM estimates (simulation study 2)	78
4.6	Application	83
4.7	Conclusions	85
5	Conclusion and future studies	87
5.1	Final considerations	87
5.2	Suggestions for future research	88
A	Chapter 2 - Ordinary residuals and QQ-plots	89
B	Chapter 4 - The observed information matrix	91
C	"FMsmnReg" Package manual	93

D “CensMixReg” Package manual	97
Bibliography	99

List of Figures

2.1	AIS dataset. (Panels a and b) displays the histogram with a kernel density estimate superimposed and the boxplot of ordinary residuals, respectively.	7
2.2	Target mixture densities from which the dataset was simulated. (a) Scenario 1 and (b) Scenario 2 of simulation study.	19
2.3	Simulation study 2. Average bias (first row) and average RMSE (second row) of the estimates of $\beta_0, \beta_1, \beta_2$	20
2.4	Simulation study 2. Average bias (first row) and average RMSE (second row) of the estimates of $\mu_1, \sigma_1, \lambda_1$	23
2.5	Simulation study 2. Average bias (first row) and average RMSE (second row) of the estimates of $\mu_2, \sigma_2, \lambda_2$	23
2.6	Simulation study 2. Average bias (first row) and average RMSE (second row) of the estimates of p_1	24
2.7	Simulation study 3. Average relative changes of estimates for different perturbations v	24
2.8	AIS dataset. Histogram of ordinary residuals superimposed for FM-SMSN-LR residual density model for a) $g = 1$ and b) $g = 2$ components.	27
2.9	AIS dataset. QQ-plots and simulated envelopes for the residual $(y - \hat{y})$ with a) $g = 1$ and b) $g = 2$ components, based on the FM-ST-LR model.	28
3.1	Target mixture densities from which the dataset was simulated. (a) Scenario 1 and (b) Scenario 2 of simulation study 1.	48
3.2	Simulation study 3: Average bias (first row) and RMSE (second row) of parameter estimates in the FM-SMN-CR model with 8% censoring level.	53
3.3	Simulation study 3: Average bias (first row) and RMSE (second row) of parameter estimates in the FM-SMN-CR model with 35% censoring level.	53
4.1	Simulation study 1 (asymptotic properties). Bias (first row) and RMSE (second row) of the estimates $\beta_1 = (\beta_{11}, \beta_{21}, \beta_{31})^\top$	79
4.2	Simulation study 1 (asymptotic properties). Bias (first row) and RMSE (second row) of the estimates $\beta_2 = (\beta_{11}, \beta_{22}, \beta_{32})^\top$	80
4.3	Simulation study 1 (asymptotic properties). Bias (first column) and RMSE (second column) of the estimates $\lambda_1 = (\lambda_{11}, \lambda_{21})^\top$ and $\lambda_2 = (\lambda_{12}, \lambda_{22})^\top$	81
4.4	Simulation study 1 (asymptotic properties). Bias (first column) and RMSE (second column) of the estimates Σ_1	82

4.5	Simulation study 1 (Asymptotic properties). Bias (first column) and RMSE (second column) of the estimates Σ_2	82
4.6	hsb2 dataset. Kernel density estimate of the response variable (a) y_1 , (b) y_2 and explanatory variable (c) x_1	83
4.7	hsb2 dataset. Profile log-likelihood of ν for models with $G = 2$ components. (a) ST-MMRM, (b) SSL-MMRM and (c) SCN-MMRM model.	84
A.1	AIS dataset. Histogram of ordinary residuals with FM-SMSN-LR residual density model for a) $g = 3$ and b) $g = 4$ components.	89
A.2	AIS dataset. QQ-plots and simulated envelopes for the residual $(y - \hat{y})$ for $g = 2$ components based on FM-SN (left), FM-SCN (middle) and FM-SSL (right) distributions, respectively.	89

List of Tables

2.1	Simulation Study 1: mean and MC sd Sd are the respective mean estimates 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.	21
2.2	Comparison of log-likelihood maximum, AIC_c and BIC_A for fitted FM-SMSN-LR models using the AIS data. The number of parameters is denoted by m	25
2.3	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. . .	26
3.1	Simulation Study 1 (Scenario 1). Mean, sample standard errors (MC SE), average values (across 500 samples) of the standard errors computed using the information method (MC IM SE) and the percentage coverage of the resulting 95% confidence intervals (COV MC) from the FM-SMN-CR. True values of parameters are in parentheses.	48
3.2	Simulation study 1 (Scenario 2). Mean, sample standard errors (MC SE), average values (across 500 samples) of the standard errors computed using the information method (MC IM SE) and the percentage coverage of the resulting 95% confidence intervals (COV MC) from the FM-SMN-CR. True values of parameters are in parentheses.	50
3.3	Simulation study 2: Relative bias and RMSE of $\hat{\theta}$ for different proportions of censoring.	51
3.4	Simulation study 3: Bias and root relative mean squared errors of ECME estimates	52
3.5	Simulation study 3: Bias and root relative mean squared errors of ECME estimates	54
3.6	ML estimation results for fitting different mixture models to the wage rate data. The model with the smallest BIC value for each family is shown in bold. The best chosen model is indicated by ”*”	55
3.7	Wage rate data. results of the parameter estimation via the EM-type algorithm. .	56
4.1	$k_m = E[U^{-m/2}]$ for different SMSN models.	63
4.2	Simulation study 1: Mean and MC Sd are the respective mean estimates, standard deviations and percentage of coverage (COV) from fitting a ST-MMRM with different sample size settings based on 500 samples. IM SE is the average value of the approximate standard error obtained through the information-based method in Subsection 4.4.4. True values of parameters are in parentheses.	77

4.3	Simulation study 1: Mean and MC Sd are the respective mean estimates, standard deviations and percentage of coverage (COV) from fitting a SCN-MMRM with different sample size settings based on 500 samples. IM SE is the average value of the approximate standard error obtained through the information-based method in Subsection 4.4.4. True values of parameters are in parentheses.	78
4.4	hsb2 dataset. Selection criteria for various SMSN-MMRM models.	84
4.5	hsb2 dataset. Estimation and standard errors (SE) for parameters under the SMSN-MMRM models.	85

Chapter 1

Introduction

1.1 Preliminaries

Finite mixtures of distributions have been widely used to investigate the relationship of variables coming from several homogeneous latent groups that are known, presenting multimodality, asymmetry and heavy tails. These models have been applied in several areas, such as astronomy, biology, image processing, economics, etc. The importance of finite mixtures of distributions can be found in a large number of books published in this area. For example, McNicholas (2017) shows an approach for clustering and classification, Ferraro *et al.* (2017) proposes a restricted method based on data for estimating the mixture of linear regression models, Pyne *et al.* (2016) shows the application of a mixture of models for large volumes of data. Tatarinova and Schumitzky (2015) provide a broad introduction to nonlinear mixture models from a Bayesian perspective. More information is available in McLachlan and Peel (2000) and from a Bayesian perspective, in Frühwirth-Schnatter (2006).

Although most statistical applications are designed to deal with non-unimodal data, in reality the data may show heterogeneity due to asymmetry and multimodality. However, the asymmetry and shape variations, possibly due to intrinsic aspects of the data, can indicate that not all observations come from the same population. In other words, a dataset can be a mixture of observations from various populations.

This dissertation consists of three articles. In the first article we consider a generalization of the work of Bartolucci and Scaccia (2005) that considers a finite mixture of normal distributions, and in our case we consider the SMSN distribution in each component of the finite mixture. This class of distributions is more robust and is useful to fit atypical and asymmetric observations better. An EM algorithm is developed to obtain the parameter estimates. To compute the standard errors, we

consider a method suggested by Basford *et al.* (1997a) where we approximate the inverse of the covariance matrix. We also consider a bootstrap procedure proposed by Turner (2000) to test the number of components in the finite mixture. Finally, we carry out different simulation studies are performed to illustrate the performance of the proposed model.

The second article presents a censored linear regression model where the error term is a finite mixture of the SMN distribution. In this paper, we generalize the works of Arellano-Valle *et al.* (2012), Garay *et al.* (2017), Massuia *et al.* (2015), Karlsson and Laitila (2014) and Caudill (2012). In this way we find that the modeling of the data is very flexible, since it simultaneously accommodates multimodality, heavy tails and asymmetry. To illustrate the proposed model, we use a well-known example involving the participation of married women in the workforce, which is routinely considered in the Tobit model of Greene (2012) or some other alternative models. In this case, the observed response is the wage rate, which is typically considered censored when it is below zero. This means that for working women, only positive wages are recorded and for women who do not work, the wage is zero, see Mroz (1987).

In the last paper, we propose a finite mixture of multivariate regressions whose error follows a multivariate distribution of SMSN. Our proposal is an extension of Zeller *et al.* (2016), who considers a univariate finite mixture of regressions where the error follows a SMSN distribution. In this work we use the EM algorithm to obtain the parameter estimates. Different simulation studies are performed to show the performance of the proposed method. Finally, an application of the model is performed considering the dataset known as “high school and beyond” (hsb2), previously analyzed by Lim *et al.* (2017).

It is important to clarify the difference between the two main causes of incomplete data: censoring and truncation, which will be used throughout our development of the second paper.

1.1.1 Truncation

Truncation happens when some observations exist that are not available in both the response variable and the regressor variables. Therefore, a condition is imposed on the dataset that makes the participation of certain individuals in the study impossible. An example of a truncation to the left is presented in Colosimo and Giolo (2006), where the distribution of the lifetime of the residents of a certain region is studied, observing a sample extracted from the local welfare database. In this case, as can be seen, only those residents who have reached retirement are part of the sample. Other examples of truncation are available in Kalbfleisch and Lawless (1992), Nelson (1990) and Arnold and Press (1989).

1.1.2 Censoring

Censoring occurs when the response variable data in a regression model is not completely available for some units of the sample, but, for these units the data on the regressor variables are known. This can happen for several reasons, for example when there are limitations of the measurement equipment or the experimental design. Thus, a censored observation contains only partial information about an event of interest. An example is presented in Richard (1996), in which the certificate, containing the situation of the student (approved or not) is delivered to all of them, but only the students who are approved are notified of their exact score. So, suppose we want to study the relationship between scores and some other explanatory variables such as parents' schooling, social class, and gender. In this case, the scores are the censored answers to the left because, if y_i denotes the score of the i -th student and that student failed, we will not know the exact value of y_i , we will only know that this value is lower than 40%. Models involving censored responses where the random error term has normal distribution have been studied in several works in the literature, such as Thompson and Nelson (2003), Park *et al.* (2007) and Vaida and Liu (2012). In censored linear regression models with Student- t distribution, we can mention the works of Arellano-Valle *et al.* (2012) and Matos *et al.* (2013).

1.2 Dissertation organization

This dissertation is organized as follows: In Chapter 2, we present the first paper proposing a linear regression model in which the error term follows a finite mixture of SMSN distributions. In Chapter 3, we present a censored regression model where the random error term follows a finite mixture of SMN distributions. In Chapter 4, we present a finite mixture of regression models based on multivariate scale mixtures of skew-normal distributions. Suggestions for future studies and some conclusions are discussed in Chapter 5, and finally in the Appendix we present additional information and some R routines incorporated in the packages **FMsmnReg** and **CensMisReg**.

Chapter 2

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

In this paper we extend the regression model based on the assumption that the error term follows a mixture of normals (Bartolucci and Scaccia, 2005) by consider 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. We developed a simple EM-type algorithm to perform maximum likelihood (ML) inference of the parameters of the proposed model with closed form expressions for both E- and M-steps. Furthermore, the empirical information matrix is derived analytically to account for standard errors and a bootstrap procedure for testing the number of components in the mixture is implemented. The practical utility of the new method is illustrated with the analysis of a real dataset and several simulation studies. The proposed algorithm and methods are implemented in the R package `FMsmsnReg`.

2.1 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 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, 1974b; 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 (FM-LR) models have been recently proposed. For instance, Bartolucci and Scaccia (2005), Soffritti and Galimberti (2011) and Galimberti e 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 (SMSN) distributions (FM-SMSN-LR model) by extending the mixture model based on symmetrical distributions. The classical approach to finite mixture modeling is a research area with several challenging aspects. There are nontrivial issues, like non-identifiability and an unbounded likelihood. In this context, Holzmann *et al.* (2006) have established the identifiability of finite mixtures of elliptical distributions under conditions on the characteristic or density generators. More recently Otiniano *et al.* (2015) have 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 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 (1974b). Besides, 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.

Figure 2.1: AIS dataset. (Panels a and b) displays the histogram with a kernel density estimate superimposed and the boxplot of ordinary residuals, respectively.

More specifically, we extended the linear regression model proposed by Bartolucci and Scaccia (2005), which is defined as

$$Y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim \sum_{j=1}^g p_j \mathcal{N}(\mu_j, \sigma_j^2), \quad i = 1, \dots, n,$$

where the Y_i are the responses, $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^\top$ is a vector explanatory variables values, p_j are positive weights summing to 1, the μ_j satisfy the constraint $\sum_{j=1}^g p_j \mu_j = 0$, by assuming that the distribution of the error term follows a finite mixtures of SMSN distributions, so that the FM-SMSN-LR is defined. It is important to stress that our proposal is different of the one proposed by Zeller *et al.* (2016), where the linear regression is modeling with different regression functions, the so-called *mixture of regressions* or *switching regression*.

In order to motivate our proposed model, we use the AIS dataset available in the R package FMsmnReg. Figure 2.1 (panels a and b) displays the histogram with a kernel density estimate superimposed and the boxplot of ordinary residuals, respectively, obtained by fitting a N-LR model to the AIS data. The plots reveal that the residuals behave somewhat multimodality with evident

presence of outliers, indicating the necessity of consider a robustified structure. Finally, an important question that is addressed in this paper is whether a mixture model ($g \geq 2$) is needed instead a one component model. Thus, we use the parametric bootstrap log-likelihood ratio statistics which was proposed by Turner (2000).

The remainder of the paper is organized as follows. In Section 2.2, 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.

2.2 Scale mixtures of skew-normal distributions

In this section 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 $\lambda \in \mathbb{R}$, denoted by $Z \sim \text{SN}(\mu, \sigma^2, \lambda)$, if its pdf is given by:

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

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 $\text{SMSN}(\mu, \sigma^2, \lambda; H)$, if it has the following stochastic representation:*

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

where $Z \sim \text{SN}(0, \sigma^2, \lambda)$, 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*.

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

$Y|U = u \sim \text{SN}(\mu, u^{-1}\sigma^2, \lambda)$. Thus, the density of Y is given by

$$f(y) = \phi_{\text{SMSN}}(y|\mu, \sigma^2, \lambda, \boldsymbol{\nu}) = 2 \int_0^\infty \phi(y|\mu, u^{-1}\sigma^2) \Phi\left(u^{1/2}\lambda(y - \mu)/\sigma\right) dH(u|\boldsymbol{\nu}). \quad (2.1)$$

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

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

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

where $\Delta = \sigma\delta$, with $\delta = \lambda/\sqrt{1 + \lambda^2}$. Also we have $W_\Phi(\cdot) = \phi(\cdot)/\Phi(\cdot)$, $A = \lambda y_0$, with $y_0 = (y - \mu)/\sigma$, f_0 is the pdf of $Y_0 \sim \text{SMN}(\mu, \sigma^2; H)$, $U_y \stackrel{d}{=} U|Y_0 = y$ and $K_r = E[U^{-r/2}]$, $r = 1, 2, \dots$

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

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

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 $Y \sim \text{ST}(\mu, \sigma^2, \lambda; \nu)$.

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

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

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

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

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

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 The SMSN linear regression model

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

$$Y_i = \mathbf{x}_{ic}^\top \boldsymbol{\beta}_c + \varepsilon_i, \quad i = 1, \dots, n, \quad (2.2)$$

where the Y_i are the responses, $\mathbf{x}_{ic} = (1, x_{i1}, \dots, x_{ip})^\top$ is a vector of explanatory variable values of dimension $(p+1) \times 1$, $\boldsymbol{\beta}_c = (\beta_0, \beta_1, \dots, \beta_p)^\top$ is the regression parameter vector and the random errors $\varepsilon_i \stackrel{iid}{\sim} \text{SMSN}\left(-\sqrt{\frac{2}{\pi}}K_1\Delta, \sigma^2, \lambda; H\right)$, 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

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

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

2.3.1 Parameter estimation via an EM-type algorithm

In this subsection we develop an EM-type algorithm 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 :

$$Y_i|T_i = t_i, U_i = u_i \stackrel{ind}{\sim} N(\mathbf{x}_{ic}^\top \boldsymbol{\beta}_c + \Delta t_i, u_i^{-1} \Gamma), \quad (2.3)$$

$$T_i|U_i = u_i \stackrel{ind}{\sim} \text{TN}(b, u_i^{-1}; (b, \infty)), \quad (2.4)$$

$$U_i \stackrel{iid}{\sim} H(\cdot | \boldsymbol{\nu}), \quad (2.5)$$

where $\text{TN}(r, s; (a, b))$ 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 $\text{TN}(\mu_{T_i} + b, u_i^{-1} M_T^2; (b, \infty))$, with $M_T^2 = \Gamma/(\Delta^2 + \Gamma)$ and $\mu_{T_i} = (\Delta/(\Delta^2 + \Gamma))(y_i - \mathbf{x}_{ic}^\top \boldsymbol{\beta}_c - \Delta b)$. 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 $\hat{\xi} = g(\hat{\boldsymbol{\theta}})$, where $g(\cdot)$ is a generic function of $\boldsymbol{\theta} = (\boldsymbol{\beta}_c^\top, \sigma^2, \lambda, \boldsymbol{\nu}^\top)^\top$. Thus, let $\mathbf{y} = (y_1, \dots, y_n)^\top$, $\mathbf{t} = (t_1, \dots, t_n)^\top$ and $\mathbf{u} = (u_1, \dots, u_n)^\top$. It follows that the complete log-likelihood function associated with $(\mathbf{y}, \mathbf{t}, \mathbf{u})$ is given by:

$$\ell_c(\boldsymbol{\theta} | \mathbf{y}, \mathbf{t}, \mathbf{u}) = c - \frac{n}{2} \log \Gamma - \frac{1}{2\Gamma} \sum_{i=1}^n U_i (y_i - \mathbf{x}_{ic}^\top \boldsymbol{\beta}_c - \Delta T_i)^2 + \sum_{i=1}^n \log(h(u_i; \boldsymbol{\nu})), \quad (2.6)$$

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

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

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

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

where $\hat{\boldsymbol{\theta}}^{(k)}$ is an updated value of $\hat{\boldsymbol{\theta}}$. When the M-step turns out to be analytically intractable, it can be replaced with a sequence of conditional maximization (CM) steps. The resulting proce-

ture is known as the Expectation Conditional Maximization (ECM) algorithm. The Expectation-Conditional Maximization Either (ECME) algorithm (Meng and Rubin, 1993), a faster extension of EM and ECM, is obtained by maximizing the constrained Q-function with some CM-steps that maximize the corresponding constrained current marginal likelihood function, called CML-steps. Next, we describe this EM-type algorithm (ECME) for ML estimation of the parameters of the SMSN-LR model

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

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

$$\begin{aligned}\hat{\boldsymbol{\beta}}_c^{(k+1)} &= \left(\sum_{i=1}^n \hat{u}_i^{(k)} \mathbf{x}_{ic} \mathbf{x}_{ic}^\top \right)^{-1} \left(\sum_{i=1}^n \hat{u}_i^{(k)} y_i \mathbf{x}_{ic} - \hat{\Delta}^{(k)} \mathbf{x}_{ic} \hat{ut}_i^{(k)} \right), \\ \hat{\Delta}^{(k+1)} &= \left(\sum_{i=1}^n (y_i - \mathbf{x}_{ic}^\top \hat{\boldsymbol{\beta}}_c^{(k+1)}) \hat{ut}_i^{(k)} \right) / \sum_{i=1}^n \hat{ut}_i^{2(k)}, \\ \hat{\Gamma}^{(k+1)} &= \frac{1}{n} \sum_{i=1}^n \left[\hat{u}_i^{(k)} \left(y_i - \mathbf{x}_{ic}^\top \hat{\boldsymbol{\beta}}_c^{(k+1)} \right)^2 - 2 \hat{\Delta}^{(k+1)} (y_i - \mathbf{x}_{ic}^\top \hat{\boldsymbol{\beta}}_c^{(k+1)}) \hat{ut}_i^{(k)} + \hat{\Delta}^{2(k+1)} \hat{ut}_i^{2(k)} \right].\end{aligned}$$

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

$$\hat{\boldsymbol{\nu}}^{(k+1)} = \operatorname{argmax}_{\boldsymbol{\nu}} \sum_{i=1}^n \log \left(\phi_{\text{SMSN}}(y_i | \mathbf{x}_{ic}^\top \hat{\boldsymbol{\beta}}_c^{(k+1)} + b \hat{\Delta}^{(k+1)}, (\hat{\sigma}^2)^{(k+1)}, \hat{\lambda}^{(k+1)}, \boldsymbol{\nu}) \right), \quad (2.7)$$

where $\phi_{\text{SMSN}}(\cdot | \boldsymbol{\theta})$ is the SMSN density defined in (2.1). Note that $\hat{\sigma}^{2(k+1)}$ and $\hat{\lambda}^{(k+1)}$ 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 (see Subsection 4.2).

2.4 The linear regression model with FM–SMSN errors

2.4.1 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 (2.2), follows a g -component mixture of SMSN densities given by:

$$f(\epsilon_i) = \sum_{j=1}^g p_j \phi_{\text{SMSN}}(\epsilon_i | \mu_j + b \Delta_j, \sigma_j^2, \lambda_j, \boldsymbol{\nu}_j), \quad p_j \geq 0, \quad \sum_{j=1}^g p_j = 1, \quad i = 1, \dots, n, \quad j = 1, \dots, g, \quad (2.8)$$

where the μ'_j 's satisfy the identifiability constraint $\sum_{j=1}^g p_j \mu_j = 0$ and from Proposition 1, we have that $E(\epsilon_i) = 0$. Let $\mu_{ij} = \mathbf{x}_{ic}^\top \boldsymbol{\beta}_c + \mu_j = \beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta} + \mu_j = \vartheta_j + \mathbf{x}_i^\top \boldsymbol{\beta}$, where $\vartheta_j = \beta_0 + \mu_j$, $\mathbf{x}_i^\top = (x_{i1}, \dots, x_{ip})^\top$ and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$. For a random sample $\mathbf{y} = (y_1, \dots, y_n)^\top$, the g -component mixture of SMSN densities is given by

$$f(y_i|\boldsymbol{\Theta}) = \sum_{j=1}^g p_j \phi_{\text{SMSN}}(y_i|\mu_{ij} + b\Delta_j, \sigma_j^2, \lambda_j, \boldsymbol{\nu}_j), \quad (2.9)$$

where $\mu_{ij} = \mathbf{x}_i^\top \boldsymbol{\beta} + \vartheta_j$, $p_1, \dots, p_g$ are the mixing probabilities and $\boldsymbol{\Theta} = (\boldsymbol{\beta}^\top, (p_1, \dots, p_{g-1})^\top, \vartheta_1, \dots, \vartheta_g, \sigma_1^2, \dots, \sigma_g^2, \lambda_1, \dots, \lambda_g, \nu_1, \dots, \nu_g)^\top$ is the vector with all parameters. Given $\mathbf{y}$, the observed data log-likelihood function is given by

$$\ell(\boldsymbol{\Theta}|\mathbf{y}) = \sum_{i=1}^n \log \left(\sum_{j=1}^g p_j \phi_{\text{SMSN}}(y_i|\mu_{ij} + b\Delta_j, \sigma_j^2, \lambda_j, \boldsymbol{\nu}_j) \right). \quad (2.10)$$

Concerning the parameter $\boldsymbol{\nu}_j$ of the mixing distribution $H(\cdot|\boldsymbol{\nu}_j)$, for $j = 1, \dots, g$, 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 $\boldsymbol{\nu}_1 = \dots = \boldsymbol{\nu}_g = \boldsymbol{\nu}$. 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:

$$P(Z_{ij} = 1) = 1 - P(Z_{ij} = 0) = p_j, \quad \sum_{j=1}^g p_j = 1 \quad \text{and} \quad y_i|Z_{ij} = 1 \sim \text{SMSN}(\mu_{ij} + b\Delta_j, \sigma_j^2, \lambda_j; H(\boldsymbol{\nu})). \quad (2.11)$$

Note that by integrating out $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{ig})^\top$ we obtain the marginal density (2.1) and $\mathbf{Z}_1, \dots, \mathbf{Z}_n$ are independent random vectors, each one having a multinomial distribution with probability function

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

which we denote by $\mathbf{Z}_i \sim M(1; p_1, \dots, p_g)$. These latent vectors appear in the hierarchical representation given next, which is used to build the ECME algorithm. From (2.11) along with Definition

1, the FM-SMSN-LR model can be represented as

$$Y_i|u_i, t_i, Z_{ij} = 1 \stackrel{ind}{\sim} N(\mu_{ij} + \Delta_j t_i, u_i^{-1} \Gamma_j), \quad (2.12)$$

$$T_i|u_i, Z_{ij} = 1 \stackrel{ind}{\sim} \text{TN}(b, u_i^{-1}, (b, \infty)), \quad (2.13)$$

$$U_i|Z_{ij} = 1 \stackrel{ind}{\sim} H(u_i; \boldsymbol{\nu}) \quad (2.14)$$

and

$$\mathbf{Z}_i \stackrel{ind}{\sim} M(1; p_1, \dots, p_g), \quad i = 1, \dots, n, \quad j = 1, \dots, g, \quad (2.15)$$

where $\Gamma_j = (1 - \delta_j^2)\sigma_j^2$, $\Delta_j = \sigma_j\delta_j$ and $\delta_j = \lambda_j/\sqrt{1 + \lambda_j^2}$.

2.4.2 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 (4.15)-(4.16), we have that the complete-data log-likelihood function is

$$\begin{aligned} \ell_c(\boldsymbol{\Theta}|\mathbf{y}, \mathbf{t}, \mathbf{u}, \mathbf{z}) &= c + \sum_{i=1}^n \sum_{j=1}^g Z_{ij} \left\{ \log(p_j) - \frac{1}{2} \log \Gamma_j - \frac{u_i}{2\Gamma_j} (y_i - \mu_{ij} - \Delta_j t_i)^2 \right. \\ &\quad \left. + \log(h(u_i|\boldsymbol{\nu})) + \log[\phi_{\text{TN}}(t_i|b, u_i^{-1}, (b, \infty))] \right\}, \end{aligned} \quad (2.16)$$

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

$$\hat{z}_{ij} = E[Z_{ij}|\hat{\boldsymbol{\Theta}}, y_i], \quad \hat{s}_{1ij} = E[Z_{ij}U_i|\hat{\boldsymbol{\Theta}}, y_i], \quad \hat{s}_{2ij} = E[Z_{ij}U_iT_i|\hat{\boldsymbol{\Theta}}, y_i] \quad \text{and} \quad \hat{s}_{3ij} = E[Z_{ij}U_iT_i^2|\hat{\boldsymbol{\Theta}}, y_i],$$

known properties of conditional expectation, we obtain:

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

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

where

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

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

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

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

In the CML-step we update the estimate of ν by direct maximization of the marginal log-likelihood, circumventing the computation of the conditional expectations $\widehat{s}_{4ij} = E[Z_{ij} \log(h(U_i | \nu)) | \widehat{\Theta}, y_i]$ and $\widehat{s}_{5ij} = E[Z_{ij} \log(\phi_{\text{TN}}(T_i | b, u_i^{-1}, (b, \infty))) | \widehat{\Theta}^{(k)}, y_i]$. Thus, the ECME algorithm for maximum likelihood estimation of Θ is defined as follows:

E-step: Given a current estimate $\widehat{\Theta}^{(k)}$, compute $\widehat{z}_{ij}$, $\widehat{s}_{1ij}$, $\widehat{s}_{2ij}$, $\widehat{s}_{3ij}$, for $i = 1, \dots, n$ and $j = 1, \dots, g$.

CM-steps: Update $\widehat{\Theta}^{(k)}$ by maximizing $Q(\Theta | \widehat{\Theta}^{(k)}) = E[\ell_c(\Theta) | \mathbf{y}, \widehat{\Theta}^{(k)}]$ over Θ , which leads to the following closed-form expressions:

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

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

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

and

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

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

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

Following Bartolucci and Scaccia (2005) we obtain the estimative 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 $j = 1, \dots, g$. 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 adopting the Aitken acceleration (Aitken, 1927) method as the stopping criterion. At iteration k , we first compute the Aitken acceleration factor $c^{(k)} = \frac{\ell^{(k+1)} - \ell^{(k)}}{\ell^{(k)} - \ell^{(k-1)}}$, where following Böhning *et al.* (1994), the asymptotic estimate of the log-likelihood at iteration $k + 1$ is given by

$$\ell_{\infty}^{(k+1)} = \ell^{(k)} + \frac{1}{1 - c^{(k)}} \left[\ell^{(k+1)} - \ell^{(k)} \right]. \quad (2.17)$$

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 $\varepsilon = 10^{-6}$).

2.4.3 Starting values

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 $p_j, \mu_j, \sigma_j^2, \lambda_j$ and $\boldsymbol{\nu}$, $j = 1, \dots, g$, are obtained by fitting the mixture model (4.12) to the OLS residuals (Bartolucci and Scaccia, 2005), which can be provided through the R library `mixsmsn` (Prates *et al.*, 2013).

2.4.4 Model selection

Consider the problem of comparing several FM-SMSN-LR models, with different number of component densities. 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 (Schlove, 1987), denoted by (BIC_a) . These criteria are defined as

$$AIC_c = -2\ell(\widehat{\boldsymbol{\theta}}) + \frac{2n\rho}{n - \rho - 1} \quad \text{and} \quad BIC_a = -2\ell(\widehat{\boldsymbol{\theta}}) + \rho \log \left(\frac{n + 2}{24} \right),$$

where $\ell(\boldsymbol{\theta})$ is the current log-likelihood, ρ is the number of free parameters that have to be estimated in the model, and where n is the sample size.

2.4.5 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 $\hat{\Theta}$ by the inverse of the observed information matrix. Let $\mathbf{I}_o(\Theta) = -\partial^2 \ell(\Theta|\mathbf{y}) / \partial \Theta \partial \Theta^\top$ be the observed information matrix, where $\ell(\Theta|\mathbf{y})$ is the observed log-likelihood function in (2.10). In this work we use the alternative method suggested by Basford *et al.* (1997a) and parameterization, which consists of approximating the inverse of the covariance matrix by:

$$\mathbf{I}_o(\hat{\Theta}) = \sum_{i=1}^n \hat{\mathbf{s}}_i \hat{\mathbf{s}}_i^\top, \quad \text{where} \quad \hat{\mathbf{s}}_i = \left. \frac{\partial}{\partial \Theta} \log \left(\sum_{j=1}^g p_j \phi_{\text{SMSN}}(y_i | \mu_{ij} + b\Delta_j, \sigma_j^2, \lambda_j, \boldsymbol{\nu}) \right) \right|_{\Theta=\hat{\Theta}}, \quad (2.18)$$

where

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

It is important to stress that the standard error of ν , obtained from $\hat{s}_{i,\nu}$, depends heavily on the calculation of conditional expectation $E[\log(U_i) | y_{\text{obs}_i}, \hat{\boldsymbol{\theta}}]$, which relies on computationally intensive Monte Carlo integrations, since we do not have an analytical expression for this expected value. Therefore the expressions for the elements $\hat{s}_{i,\beta}, \hat{s}_{i,p_j}, \hat{s}_{i,\vartheta_j}, \hat{s}_{i,\sigma_j^2}, \hat{s}_{i,\lambda_j}$, $j = 1, \dots, g$, are given as follows:

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

and

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

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

After some algebraic manipulation, we obtain:

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

where the expression $I_{ij}^{\Phi}(w)$ and $I_{ij}^{\phi}(w)$ are defined in Basso *et al.* (2010). Now, returning to our original parameterization, we find the Hessian matrix for the original parameter vector $\Theta^* = (\beta^{\top}, (p_1, \dots, p_{g-1})^{\top}, \beta_0, \mu_1, \dots, \mu_g, \sigma_1^2, \dots, \sigma_g^2, \lambda_1, \dots, \lambda_g)^{\top}$. Using the delta method, we have

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

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

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

where $\boldsymbol{\gamma} = \frac{\partial \boldsymbol{\mu}}{\partial \mathbf{p}^{\top}} = \mathbf{A}(\mathbf{p}^{-1})^{\top}$ is a matrix of dimension $g \times (g-1)$, with $\mathbf{A} = (\mu_g - \mu_1, \dots, \mu_g - \mu_{g-1})^{\top}$ and $\mathbf{p}^{-1} = (1/p_1, \dots, 1/p_{g-1})^{\top}$. The square root of the diagonal elements of the $(p + 4g) \times (p + 4g)$ matrix $\mathbf{J}(\Theta^*|\Theta) \mathbf{I}_o(\hat{\Theta})^{-1} \mathbf{J}(\Theta^*|\Theta)^{\top}$ are the standard errors of parameters Θ^* .

2.5 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 FMsmnReg (Benites *et al.*, 2016). The first simulation study shows

Figure 2.2: Target mixture densities from which the dataset was simulated. (a) Scenario 1 and (b) Scenario 2 of simulation study.

the consistency of the approximate standard errors for the ML estimates of parameters through the EM algorithm to each sample under the stopping criterion in (4.18). 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.

2.5.1 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 $x_{i2} \sim U(0, 1)$, $i = 1, \dots, n$, and ε_1 and ε_2 follow a distribution as in the assumption given in (4.12). 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 (depicted in Figure 2.2): scenario 1 (well separated components): $\sigma_1^2 = 0.2$ and $\sigma_2^2 = 0.4$, and

Figure 2.3: *Simulation study 2. Average bias (first row) and average RMSE (second row) of the estimates of $\beta_0, \beta_1, \beta_2$.*

scenario 2 (poorly separated components): $\sigma_1^2 = 2$ 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 ($\nu^\top = (0.1, 0.1)$). The average values (Mean) and standard deviations (MC Sd) of the estimates 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 Section 2.4.5. Moreover we compute coverage probability of each parameter (COV), which is defined by $\text{COV}(\hat{\theta}) = (1/m) \sum_{j=1}^m I(\theta \in [\hat{\theta}_L, \hat{\theta}_U])$, where I is the indicator function such that θ lies in the interval $[\hat{\theta}_L, \hat{\theta}_U]$, with $\hat{\theta}_L$ and $\hat{\theta}_U$ the estimated lower and upper bounds of the 95% CI, respectively.

The results are presented in Table 3.1. Note that under both scenarios (well and poorly separated components), the results suggest that the proposed FM-SMSN-LR model produced satisfactory estimates.

We also can 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 4.19) 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.

Table 2.1: *Simulation Study 1: mean and MC sd Sd are the respective mean estimates 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	-

2.5.2 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 $x_{i2} \sim U(0,1)$, $i = 1, \dots, n$. We choose sample sizes $n = 100, 250, 500, 1000, 2500$ and 5000 . The true values of the parameters were taken as $\beta_0 = -1$, $\beta = (\beta_1, \beta_2)^\top = (-4, -3)^\top$, $\mu_1 = -4$, $\mu_2 = 1$, $\sigma_1^2 = 0.2$, $\sigma_2^2 = 0.4$ and $p_1 = 0.8$. 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 $\hat{\theta}_i^{(j)}$ is the estimate of θ_i for the j -th sample. The results for β_0 , β_1 and β_2 are shown in Figure 2.3, the results for μ_1 , σ_1 and λ_1 are shown in Figure 2.4, the results for μ_2 , σ_2 , λ_2 are shown in Figure 2.5 and the results for p_1 are shown in Figure 2.6. 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.

2.5.3 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 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. Under this scenario, we generated 500 samples of size $n = 500$ of the FM-SMSN-LR model with black $\varepsilon_i \sim \sum_{j=1}^2 p_j \text{SMSN}(\varepsilon_i | \mu_j + b\Delta_j, \sigma_j^2, \lambda_j, \boldsymbol{\nu}_j)$. 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 $p_1 = 0.8$. To assess how much the EM estimates are influenced by the presence of outliers, we replaced the observation y_{150} by: $y_{150}(v) = y_{150} + v$, with $v = 1, 2, \dots, 10$. For each replication, we obtained the parameter estimates with and without outliers, under the three FM-SMSN-LR models.

2.6 Application

In this Section the proposed techniques are illustrated with the analysis of one data set. As application of the proposed methodology, we consider the dataset previously analyzed by Cook and Weisberg (1982) in a normal regression setting. The data set comes from the Australian Institute of Sport (AIS) and consists of measurements taken on 202 athletes. Here, we focus on percent body fat (Bfat), which is assumed to be explained by the name sum of skin folds (ssf) and Height in cm

Figure 2.4: *Simulation study 2. Average bias (first row) and average RMSE (second row) of the estimates of $\mu_1, \sigma_1, \lambda_1$.*

Figure 2.5: *Simulation study 2. Average bias (first row) and average RMSE (second row) of the estimates of $\mu_2, \sigma_2, \lambda_2$.*

(Ht). Thus, we consider the following FM-SMSN-LR model:

$$Bfat_i = \beta_0 + \beta_1 ssf_i + \beta_2 Ht_i + \varepsilon_i,$$

where ε_i belongs to the FM-SMSN family for $i = 1, \dots, 202$.

Figure 2.6: *Simulation study 2. Average bias (first row) and average RMSE (second row) of the estimates of p_1 .*

Figure 2.7: *Simulation study 3. Average relative changes of estimates for different perturbations v .*

Using the R package `FMsmnReg` (see Appendix A), we fit the FM-SMSN-LR models as was described in Section 2.2. Table 2.2 compares the fit of various mixture models for $g = 1$ to 5 components, using the model selection criteria discussed in Subsection 4.3. Note from this table that, as expected, the heavy-tailed models perform significantly better than the SN model (and the symmetric counterpart such as the normal and Student-t models), with mixtures of two ($g = 2$)

Table 2.2: Comparison of log-likelihood maximum, 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

components significantly better in all cases, except for the normal case (FM-N) where a mixture of $g = 3$ is needed. Moreover, the 2-components FM-ST-LR model fits the data substantially better. This conclusion also is verified through a testing of hypotheses procedure for testing the number of components in the FM-ST-LR model. As reported by Turner (2000) we can use parametric or semiparametric bootstrap for testing hypotheses concerning the number of components in the mixture. Following the methodology proposed by Turner (2000), we considered 1000 bootstrap statistics for testing $G = 1$ versus $G = 2$, in this case the p -value was 0.027 for the parametric bootstrap. Accordingly, there is strong evidence that at least two components is needed. For testing $G = 2$ versus $G = 3$ the bootstrap p -value was 0.984, thus there is no evidence that more than two components is required to model the AIS dataset .

Table 2.3 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 model, along with the correspond-

ing standard errors (SE), obtained via the information-based procedure presented in Subsection 2.4.5.

Table 2.3: 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	-	-	-

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 assumption.

Figure 2.8: AIS dataset. Histogram of ordinary residuals superimposed for FM-SMSN-LR residual density model for a) $g = 1$ and b) $g = 2$ components.

In Figure 2.8, 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, with $g = 1$ and $g = 2$ components respectively. Additional results related to $g = 3$ and $g = 4$ components are given in Appendix B (Figure A.1). 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 QQ-plots and simulated envelopes for the residuals $(y - \hat{y})$ in Figure 2.9. The QQ-plots for the other models are given in Figure A.2 (Appendix B). 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.

2.7 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

Figure 2.9: AIS dataset. QQ-plots and simulated envelopes for the residual $(y - \hat{y})$ with a) $g = 1$ and b) $g = 2$ components, based on the FM-ST-LR model.

important to stress that our proposal is different from that proposed by 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 density estimators, to model the unknown distributional shape of the errors. 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 `FMsmnReg`, 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).

Chapter 3

Regression modeling of censored data based on mixtures of scale mixtures of normal distributions

In the framework of censored regression models, the distribution of the error term can depart significantly from normality, for instance, due to the presence of multimodality, skewness and/or atypical observations. In this paper we propose a novel censored linear regression model where the random errors follow a finite mixture of scale mixtures of normal (SMN) distribution. The SMN is an attractive class of symmetrical heavy-tailed densities that includes the normal, Student-t, slash and the contaminated normal distribution as special cases. This approach allows us to model data with great flexibility, accommodating simultaneously multimodality, heavy tails and skewness depending on the structure of the mixture components. We develop an analytically tractable and efficient EM-type algorithm for iteratively computing the maximum likelihood estimates of the parameters, with standard errors and prediction of the censored values as a by-products. The proposed algorithm has closed-form expressions at the E-step, that rely on formulas for the mean and variance of the truncated SMN distributions. The efficacy of the method is verified through the analysis of simulated and real datasets. The methodology addressed in this paper is implemented in the R package `CensMixReg`.

3.1 Introduction

Regression models with censored dependent variables are applied in many fields, like econometric analysis, geostatistics, clinical trials, medical surveys and engineering studies, among others. For

example, in econometrics, the study of the labor force participation of married women is usually conducted under the censored Tobit model, for instance, Chib (1992). In this case, the observed response is the annual working hours, which is typically considered as left-censored at zero, i.e., working women have strictly positive annual working hours and annual wage incomes (Mroz, 1987). In AIDS research, the viral load measures may be subject to some upper and lower detection limits, below or above which they are not quantifiable. As a result, the viral load responses are either left or right censored depending on the diagnostic assays used (Wu, 2010).

In general, for mathematical tractability it is assumed that the random errors have a normal distribution (Wei and Tanner, 1990). However, it is well known that several phenomena are not always in agreement with this assumption, yielding data with a distribution having heavier tails, skewness or/and 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 transformations yet presents a Gaussian framework.

Many extensions of the classic Gaussian censored regression (N-CR, hereafter) model have been proposed to broaden the applicability of linear regression analysis to situations where the Gaussian error term assumption may be inadequate. For instance, Arellano-Valle *et al.* (2012) advocated the use of the Student-t distribution in the context of truncated regression models. Massuia *et al.* (2015) developed diagnostic measures for censored regression models using the Student-t distribution (t-CR), including the implementation of an EM algorithm for maximum likelihood (ML) estimation. A censored regression model based on the family of SMN distributions (SMN-CR model), as defined by Andrews and Mallows (1974a). The Student-t (T), Pearson type VII (PVII), slash (SL), power exponential (PE), contaminated normal (CN) and, obviously, the normal (N) distributions are included in this class. The SMN-CR models has been recently proposed by Garay *et al.* (2015) to estimate the regression parameters with a simple and efficient EM-type algorithm for iteratively computing ML estimates of the parameters is also presented.

Even when modeling using normal mixtures, overestimation of the number of components (that is, the number of densities in the mixture of the random error) necessary to capture the asymmetric and/or heavy-tailed nature of each subpopulation can occur. Thus in this article we

propose a mixture model for the random errors based on the SMN class of distributions (FM-SMN-CR model) by extending the mixture of normal mixtures proposed by Caudill (2012) and Karlsson and Laitila (2014). Moreover, we consider the case in which each component has different levels of heaviness in the tails of the distribution. More specifically, our objectives are: (i) to propose a mixture censored regression model (and associated likelihood inference) based on the mixtures of SMN distributions, where each component is modeled with a different kurtosis parameter $\boldsymbol{\nu}$, extending the recent works of Arellano-Valle *et al.* (2012), Caudill (2012), Karlsson and Laitila (2014), Massuia *et al.* (2015) and Garay *et al.* (2015); (ii) to implement and evaluate the proposed methods through intensive simulations and the analysis of a real life dataset. (iii) For a researcher interested in our methodology, we implement an R package CensMixReg.

The remainder of the paper is organized as follows. In Section 2, we briefly discuss the truncated SMN distributions and some of their properties. In addition, we present the SMN-CR model proposed by Garay *et al.* (2015) and the related ML estimation. In Section 3, we present the FM-SMN-CR model, including the EM algorithm for ML estimation, and derive the empirical information matrix analytically to obtain the standard errors. In Sections 4 and 5, numerical examples using both simulated and real data are given to illustrate the performance of the proposed method. Finally, some concluding remarks are presented in Section 6.

3.2 The model

3.2.1 Some characteristics of the SMN-CR model

We give a brief introduction of the SMN model. In the following definitions, $N(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2 and $\phi(\cdot|\mu, \sigma^2)$ denotes its probability density function (pdf) and $\phi(\cdot)$ and $\Phi(\cdot)$ denote, respectively, the pdf and the cumulative distribution function (cdf) of the standard normal distribution. Also, $\stackrel{d}{=}$ means “has the same distribution as”. First, we give the definition of the SMN distributions and then we introduce some useful properties. We say that a random variable X has a SMN distribution with location parameter $\mu \in \mathbb{R}$, scale parameter $\sigma^2 \in (0, \infty)$ and degree of freedom $\nu \in (0, \infty)$, denoted by $X \sim \text{SMN}(\mu, \sigma^2, \boldsymbol{\nu})$, if it has the following representation:

$$X \stackrel{d}{=} \mu + U^{-1/2}Z, \quad (3.1)$$

where $Z \sim N(0, \sigma^2)$, U is a positive random variable with cdf $H(\cdot|\boldsymbol{\nu})$ and $\boldsymbol{\nu}$ is a scalar or vector parameter indexing the distribution of U . Also, Z and U are independent random variables. The

pdf of (3.1) is

$$f_{SMN}(x|\mu, \sigma^2, \boldsymbol{\nu}) = (2\pi\sigma^2)^{-1/2} \int_0^\infty u^{1/2} \exp \left[-(x - \mu)^2 / (2u^{-1}\sigma^2) \right] dH(u|\boldsymbol{\nu}). \quad (3.2)$$

Let $X \sim \text{SMN}(\mu, \sigma^2, \boldsymbol{\nu})$ if X has the density of (3.2). A random variable Y has a truncated SMN distribution in the interval (a, b) if $Y \stackrel{d}{=} X|X \in (a, b)$. In this case we write $Y \sim \text{SMN}_\nu(\mu, \sigma^2, \boldsymbol{\nu}; (a, b))$.

It is straightforward to prove that the density of Y is given by

$$f_{\text{TSMN}}(y|\mu, \sigma^2, \boldsymbol{\nu}; (a, b)) = \quad (3.3)$$

$$f_{SMN}(y|\mu, \sigma^2, \boldsymbol{\nu}) \left[F_{SMN} \left(\frac{b - \mu}{\sigma} \right) - F_{SMN} \left(\frac{a - \mu}{\sigma} \right) \right]^{-1}, \quad a < y < b,$$

where $y \in (a, b)$, $f_{\text{TSMN}}(y|\mu, \sigma^2, \boldsymbol{\nu}; (a, b)) = 0$ otherwise, where $F_{SMN}(\cdot)$ denotes the cdf of the standard SMN distribution. The following result is very important for our subsequent exposition. It was provided by Garay *et al.* (2015) and presents the first two moments of the truncated SMN distribution.

Proposition 2. *Let $X \sim \text{SMN}(0, 1, \boldsymbol{\nu})$ with scale factor U and mixture distribution $H(\cdot|\boldsymbol{\nu})$. Then, for $a < b$, $E[U^r X^s | X \in (a, b)]$ for $r \geq 1$ and $s = 0, 1, 2$ is given by:*

$$\begin{aligned} E[U^r | X \in (a, b)] &= \tau(a, b) [E_\Phi(r, b) - E_\Phi(r, a)]; \\ E[U^r X | X \in (a, b)] &= \tau(a, b) \left[E_\phi \left(r - \frac{1}{2}, a \right) - E_\phi \left(r - \frac{1}{2}, b \right) \right]; \\ E[U^r X^2 | X \in (a, b)] &= \tau(a, b) \left[E_\Phi(r - 1, b) - E_\Phi(r - 1, a) + a E_\phi \left(r - \frac{1}{2}, a \right) \right. \\ &\quad \left. - b E_\phi \left(r - \frac{1}{2}, b \right) \right], \end{aligned}$$

where

$$\begin{aligned} \tau(a, b) &= (F_{SMN}(b) - F_{SMN}(a))^{-1}; \\ E_\phi(r, h) &= E \left[U^r \phi \left(h U^{\frac{1}{2}} \right) \right] = \int_0^\infty u^r \phi \left(h u^{\frac{1}{2}} \right) dH(u|\boldsymbol{\nu}); \\ E_\Phi(r, h) &= E \left[U^r \Phi \left(h U^{\frac{1}{2}} \right) \right] = \int_0^\infty u^r \Phi \left(h u^{\frac{1}{2}} \right) dH(u|\boldsymbol{\nu}). \end{aligned}$$

The following result will be useful for the implementation of the EM algorithm (see Section 3.2.3). The proof can be found in Garay *et al.* (2015), Corollary 1.

Corollary 1. *Let $Y \sim \text{SMN}(\mu, \sigma^2, \boldsymbol{\nu})$ with scale factor U and $\mathcal{A} = (a, b)$. Then, for $r \geq 1$,*

$$\begin{aligned} \mathbb{E}[U^r | Y \in \mathcal{A}] &= \mathbb{E}[U^r | X \in \mathcal{A}^*]; \\ \mathbb{E}[U^r Y | Y \in \mathcal{A}] &= \mu \mathbb{E}[U^r | X \in \mathcal{A}^*] + \sigma \mathbb{E}[U^r X | X \in \mathcal{A}^*]; \\ \mathbb{E}[U^r Y^2 | Y \in \mathcal{A}] &= \mu^2 \mathbb{E}[U^r | X \in \mathcal{A}^*] + 2\mu\sigma \mathbb{E}[U^r X | X \in \mathcal{A}^*] \\ &\quad + \sigma^2 \mathbb{E}[U^r X^2 | X \in \mathcal{A}^*], \end{aligned}$$

where $X \sim \text{SMN}(0, 1, \boldsymbol{\nu})$ and $\mathcal{A}^* = (a^*, b^*)$, with $a^* = (a - \mu) / \sigma$ and $b^* = (b - \mu) / \sigma$.

3.2.2 The model

Let us consider first a linear regression model where the responses $Y_1, \dots, Y_n$ are observed and are independent random variables with errors which are independent and identically distributed (iid) according to a SMN distribution. To be more precise, let us write

$$Y_i = \mathbf{x}_{ic}^\top \boldsymbol{\beta}_c + \varepsilon_i, \quad (3.4)$$

$$\varepsilon_i \stackrel{iid}{\sim} \text{SMN}(0, \sigma^2, \boldsymbol{\nu}), \quad (3.5)$$

where $i = 1, \dots, n$, Y_i is the response for subject i , $\boldsymbol{\beta}_c = (\beta_0, \beta_1, \dots, \beta_p)^\top$ is a vector of regression parameters and $\mathbf{x}_{ic}^\top = (1, x_{i1}, \dots, x_{ip})$ is a vector of explanatory variable values. From Equations (3.4)-(3.5), we have that $Y_i \sim \text{SMN}(\mathbf{x}_{ic}^\top \boldsymbol{\beta}_c, \sigma^2, \boldsymbol{\nu})$, for $i = 1, \dots, n$. We call (3.4)-(3.5) the SMN-R model. We are interested in the case where left-censored observations can occur. That is, the observations are of the form

$$Y_{\text{obs}_i} = \begin{cases} \kappa_i & \text{if } Y_i \leq \kappa_i; \\ Y_i & \text{if } Y_i > \kappa_i, \end{cases} \quad (3.6)$$

$i = 1, \dots, n$, for some threshold point κ_i . We have chosen to work with the left censored case, but the results are easily extendable to other censoring types.

3.2.3 Parameter estimation via an EM-type algorithm

In this subsection we developed an Expectation-Maximization (EM) type algorithm (Dempster *et al.*, 1977), a reliable and widely used framework, to obtain maximum likelihood estimates of the parameters of the SMN-CR model. Supposing there are m censored values of the characteristic of interest, then we can partition the observed sample $\mathbf{y}_{\text{obs}}$ into two subsamples of m censored and $n - m$ uncensored values, such that $\mathbf{y}_{\text{obs}} = \{\kappa_1, \dots, \kappa_m, y_{m+1}, \dots, y_n\}$. Then, the log-likelihood

function of the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}_c^\top, \sigma^2, \boldsymbol{\nu})^\top$ is given by

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= \log \left(\prod_{i=1}^n \left[F_{SMN} \left(\frac{\kappa_i - \mathbf{x}_{ic}^\top \boldsymbol{\beta}_c}{\sigma} \right) \right]^{\mathbb{I}_i} \left[f_{SMN}(y_i | \mathbf{x}_{ic}^\top \boldsymbol{\beta}_c, \sigma^2, \boldsymbol{\nu}) \right]^{1-\mathbb{I}_i} \right) \\ &= \sum_{i=1}^m \log \left[F_{SMN} \left(\frac{\kappa_i - \mathbf{x}_{ic}^\top \boldsymbol{\beta}_c}{\sigma} \right) \right] + \sum_{i=m+1}^n \log \left[f_{SMN}(y_i | \mathbf{x}_{ic}^\top \boldsymbol{\beta}_c, \sigma^2, \boldsymbol{\nu}) \right],\end{aligned}\tag{3.7}$$

where $\mathbb{I}_i = 1$ if $\{y_i \leq \kappa_i\}$ and $\mathbb{I}_i = 0$ otherwise. To apply the EM method, we need a representation of the model in terms of missing data. First, observe that, by Equation (3.1), if $Y_i \sim \text{SMN}(\mathbf{x}_{ic}^\top \boldsymbol{\beta}_c, \sigma^2, \boldsymbol{\nu})$ then

$$Y_i | U_i = u_i \sim N(\mathbf{x}_{ic}^\top \boldsymbol{\beta}_c, u_i^{-1} \sigma^2), \quad U_i \sim H(\cdot | \boldsymbol{\nu}).\tag{3.8}$$

This relation is a convenient stochastic representation of the SMN-R model, and will be useful in path E of the algorithm.

In the case of censoring, we can consider the unobserved y_i as a realization of the latent unobservable variable $Y_i \sim \text{SMN}(\mathbf{x}_{ic}^\top \boldsymbol{\beta}_c, \sigma^2, \boldsymbol{\nu})$, $i = 1, \dots, m$. The key to the development of our EM-type algorithm is to consider the augmented dataset $\{y_1, \dots, y_n, u_1, \dots, u_n\}$, that is, we treat the problem as if $(y_1, \dots, y_m)^\top$ were in fact observed. As a consequence, we can use the representation (3.8) to obtain the complete-data log-likelihood, which apart from constants which do not depend on $\boldsymbol{\theta}$, is given by

$$\ell_c(\boldsymbol{\theta}) = -\frac{n}{2} \log \sigma^2 + \frac{1}{2} \sum_{i=1}^n \log u_i - \frac{1}{2\sigma^2} \sum_{i=1}^n u_i (y_i - \mathbf{x}_{ic}^\top \boldsymbol{\beta}_c)^2 + \sum_{i=1}^n \log h(u_i | \boldsymbol{\nu}),$$

where $h(\cdot | \boldsymbol{\nu})$ is density of the random variable U . In what follows, the superscript (k) indicates the estimate of the related parameter at stage k of the algorithm. In path E of the algorithm, we must obtain the so-called Q-function

$$Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(k)}) = E_{\boldsymbol{\theta}^{(k)}}[\ell_c(\boldsymbol{\theta}) | \mathbf{y}_{\text{obs}}],$$

where $E_{\boldsymbol{\theta}^{(k)}}$ means that the expectation is being affected using $\boldsymbol{\theta}^{(k)}$ for $\boldsymbol{\theta}$. Observe that the expression of the Q-function is completely determined by the knowledge of the expectations

$$\mathcal{E}_{si}(\boldsymbol{\theta}^{(k)}) = E_{\boldsymbol{\theta}^{(k)}}[U_i Y_i^s | y_{\text{obs}_i}], \quad s = 0, 1, 2.$$

as well as

$$\mathbb{E}_{\boldsymbol{\theta}^{(k)}}[\log(U_i)|y_{\text{obs}_i}] \quad \text{and} \quad \mathbb{E}_{\boldsymbol{\theta}^{(k)}}[\log h(U_i|\boldsymbol{\nu})|y_{\text{obs}_i}].$$

Thus, dropping unimportant constants, the Q-function can be written in a synthetic form as

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) &= -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n \left[\mathcal{E}_{2i}(\boldsymbol{\theta}^{(k)}) - 2\mathcal{E}_{1i}(\boldsymbol{\theta}^{(k)})\mathbf{x}_{ic}^\top \boldsymbol{\beta}_c + \mathcal{E}_{0i}(\boldsymbol{\theta}^{(k)})(\mathbf{x}_{ic}^\top \boldsymbol{\beta}_c)^2 \right] \\ &+ \frac{1}{2} \sum_{i=1}^n \mathbb{E}_{\boldsymbol{\theta}^{(k)}}[\log(U_i)|y_{\text{obs}_i}] + \sum_{i=1}^n \mathbb{E}_{\boldsymbol{\theta}^{(k)}}[\log h(U_i|\boldsymbol{\nu})|y_{\text{obs}_i}]. \end{aligned}$$

We have adopted this strategy to update the estimate of $\boldsymbol{\nu}$, by direct maximization of the marginal log-likelihood (Expectation/Conditional Maximization Either (ECME) algorithm), circumventing the computation of $\mathbb{E}_{\boldsymbol{\theta}^{(k)}}[\log(U_i)|y_{\text{obs}_i}]$ and $\mathbb{E}_{\boldsymbol{\theta}^{(k)}}[\log h(U_i|\boldsymbol{\nu})|y_{\text{obs}_i}]$. Now, observe that

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}^{(k)}}[U_i Y_i^s | Y_{\text{obs}_i}] &= \mathbb{E}_{\boldsymbol{\theta}^{(k)}} \left\{ \mathbb{E}_{\boldsymbol{\theta}^{(k)}}[U_i Y_i^s | Y_i] | Y_{\text{obs}_i} \right\} \\ &= \mathbb{E}_{\boldsymbol{\theta}^{(k)}} \left\{ Y_i^s \mathbb{E}_{\boldsymbol{\theta}^{(k)}}[U_i | Y_i] | Y_{\text{obs}_i} \right\}. \end{aligned}$$

Defining $d(\boldsymbol{\theta}^{(k)}, y_i) = (y_i - \mathbf{x}_i^\top \boldsymbol{\beta}^{(k)})/\sigma^{(k)}$, we have that *for an uncensored observation i* , $Y_{\text{obs}_i} = y_i$, such that

$$\mathcal{E}_{si}(\boldsymbol{\theta}^{(k)}) = y_i^s \mathbb{E}_{\boldsymbol{\theta}^{(k)}}[U_i | y_i]. \quad (3.9)$$

The value of $\mathbb{E}_{\boldsymbol{\theta}^{(k)}}[U_i | y_i]$ was computed before by Osorio *et al.* (2007). Thus, for example,

- If $Y_i \sim PVII(\mathbf{x}_{ic}^\top \boldsymbol{\beta}_c, \sigma^2, \nu, \delta)$, we have

$$\mathbb{E}_{\boldsymbol{\theta}^{(k)}}[U_i | y_i] = \frac{(\nu + 1)}{\delta + d^2(\boldsymbol{\theta}^{(k)}, y_i)};$$

- If $Y_i \sim SL(\mathbf{x}_{ic}^\top \boldsymbol{\beta}_c, \sigma^2, \nu)$, we have

$$\mathbb{E}_{\boldsymbol{\theta}^{(k)}}[U_i | y_i] = \frac{\Gamma(\nu + 1.5, d^2(\boldsymbol{\theta}^{(k)}, y_i)/2)}{\Gamma(\nu + 0.5, d^2(\boldsymbol{\theta}^{(k)}, y_i)/2)};$$

- If $Y_i \sim CN(\mathbf{x}_{ic}^\top \boldsymbol{\beta}_c, \sigma^2, \nu, \gamma)$, we have

$$\mathbb{E}_{\boldsymbol{\theta}^{(k)}}[U_i | y_i] = \frac{1 - \nu + \nu \gamma^{1.5} e^{0.5(1-\gamma)d^2(\boldsymbol{\theta}^{(k)}, y_i)}}{1 - \nu + \nu \gamma^{0.5} e^{0.5(1-\gamma)d^2(\boldsymbol{\theta}^{(k)}, y_i)}},$$

and, for a censored observation i , we have that $Y_{\text{obs}_i} = \kappa_i$ iff $Y_i \leq \kappa_i$, such that

$$\mathcal{E}_{si}(\boldsymbol{\theta}^{(k)}) = \mathbb{E}_{\boldsymbol{\theta}^{(k)}}[U_i Y_i^s | Y_i \leq \kappa_i], \quad (3.10)$$

which can be easily obtained from Proposition 2.

When the M-step turns out to be analytically intractable, it can be replaced with a sequence of conditional maximization (CM) steps. The resulting procedure is known as the ECM algorithm (Meng and Rubin, 1993). The ECME algorithm (Liu and Rubin, 1994), a faster extension of EM and ECM, is obtained by maximizing the constrained Q-function with some CM-steps that maximize the corresponding constrained current marginal likelihood function, called CML-steps. Next, we describe this ECME algorithm for ML estimation of the parameters of the SMN-CR model.

E-step: Given $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$. For $i = 1, \dots, n$.

- If the observation i is uncensored, then for $s = 0, 1, 2$, compute $\mathcal{E}_{si}(\boldsymbol{\theta}^{(k)})$, as given in (3.9);
- If the observation i is censored, then for $s = 0, 1, 2$, compute $\mathcal{E}_{si}(\boldsymbol{\theta}^{(k)})$ in (3.10) using Proposition 2.

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

$$\boldsymbol{\beta}_c^{(k+1)} = \left(\sum_{i=1}^n \mathcal{E}_{0i}(\boldsymbol{\theta}^{(k)}) \mathbf{x}_{ic} \mathbf{x}_{ic}^\top \right)^{-1} \sum_{i=1}^n \mathbf{x}_{ic} \mathcal{E}_{1i}(\boldsymbol{\theta}^{(k)}), \quad (3.11)$$

$$\begin{aligned} \sigma^{2(k+1)} &= \frac{1}{n} \sum_{i=1}^n \left[\mathcal{E}_{2i}(\boldsymbol{\theta}^{(k)}) - 2\mathcal{E}_{1i}(\boldsymbol{\theta}^{(k)}) \mathbf{x}_{ic}^\top \boldsymbol{\beta}_c^{(k+1)} \right. \\ &\quad \left. + \mathcal{E}_{0i}(\boldsymbol{\theta}^{(k)}) (\mathbf{x}_{ic}^\top \boldsymbol{\beta}_c^{(k+1)})^2 \right]. \end{aligned} \quad (3.12)$$

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

$$\begin{aligned} \nu^{(k+1)} &= \operatorname{argmax}_{\nu} \left\{ \sum_{i=1}^m \log \left[F_{SMN} \left(\frac{\kappa_i - \mathbf{x}_{ic}^\top \boldsymbol{\beta}_c^{(k+1)}}{\sigma^{(k+1)}} \right) \right] \right. \\ &\quad \left. + \sum_{i=m+1}^n \log \left[f_{SMN}(y_i | \mathbf{x}_{ic}^\top \boldsymbol{\beta}_c^{(k+1)}, \sigma^{2(k+1)}, \nu) \right] \right\}. \end{aligned} \quad (3.13)$$

This process is iterated until a suitable convergence rule is satisfied, see the Subsection 3.4.

3.3 The FM-SMN-CR model

Ignoring censoring for the moment, we first consider a more general and framework for the random error ϵ_i of the regression model defined in (3.4)-(3.5), which is assumed to follow a mixture of SMN

distributions. More precisely, we replace assumption (3.5) with

$$\epsilon_i \sim \sum_{j=1}^G p_j \text{SMN}(\mu_j, \sigma_j^2, \boldsymbol{\nu}_j). \quad (3.14)$$

We assume that the number of components G is known and fixed, whereas $p_1, \dots, p_G$ is an unknown vector of mixture weights and are constrained to be positive. The μ_j 's satisfy the next condition:

$$\sum_{j=1}^G p_j \mu_j = 0.$$

Therefore $E(\epsilon_i) = 0$. The mixture regression model defined by (3.4) and (3.14) has the following stochastic representation: let Z_i be a latent class variable such that given $Z_i = j$, then response Y_i has distribution

$$Y_i \sim \text{SMN}(\beta_0 + \mu_j + \mathbf{x}_i^\top \boldsymbol{\beta}, \sigma_j^2, \boldsymbol{\nu}_j), \quad i = 1, \dots, n, \quad j = 1, \dots, G, \quad (3.15)$$

where $P(Z_i = j) = p_j$. We have that the marginal distribution of Y_i is

$$f(y_i | \mathbf{x}_i, \boldsymbol{\theta}) = \sum_{j=1}^G p_j \text{SMN}(y_i | \varphi_j + \mathbf{x}_i^\top \boldsymbol{\beta}, \sigma_j^2, \boldsymbol{\nu}), \quad (3.16)$$

where $\varphi_j = \beta_0 + \mu_j$, μ_j is scalar and $\boldsymbol{\theta} = (\boldsymbol{\gamma}^\top, \boldsymbol{\theta}_1^\top, \dots, \boldsymbol{\theta}_G^\top)^\top$, with $\boldsymbol{\gamma} = (\beta_0, \boldsymbol{\beta}^\top)^\top \in \mathbb{R}^{p+1}$ and $\boldsymbol{\theta}_j = (p_j, \sigma_j^2, \mu_j, \boldsymbol{\nu}_j)^\top$, $\mathbf{x}_i^\top = (x_{i1}, \dots, x_{in})^\top$ and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$. The model (3.16) is the regression model based on a mixture of SMN distributions. Here, we propose an extension of (3.16) with censored responses. Then, given $Z_i = j$, let us assume (3.15) and (3.6). In this case, we have a SMN-CR model for each group j , such that the conditional distribution of $Y_{\text{obs}_i} | Z_i = j$ is

$$g_{ij}(y_i | \mathbf{x}_i, \boldsymbol{\gamma}, \boldsymbol{\theta}_j) = \left[F_{\text{SMN}} \left(\frac{\kappa_i - \varphi_j - \mathbf{x}_i^\top \boldsymbol{\beta}}{\sigma_j} \right) \right]^{\mathbb{I}_i} \left[f_{\text{SMN}}(y_i | \varphi_j + \mathbf{x}_i^\top \boldsymbol{\beta}, \sigma_j^2, \boldsymbol{\nu}) \right]^{1 - \mathbb{I}_i},$$

which means that the marginal distribution of Y_{obs_i} is

$$f(y_i | \mathbf{x}_i, \boldsymbol{\theta}) = \sum_{j=1}^G p_j g_{ij}(y_i | \mathbf{x}_i, \boldsymbol{\gamma}, \boldsymbol{\theta}_j). \quad (3.17)$$

The model defined by (3.17) will be called the FM-SMN-CR model.

3.3.1 Identifiability

Identifiability is an important condition for parameter estimation in the case of mixture distributions. Identifiability of mixtures has been discussed by several authors, among others, Teicher (1967), Yakowitz and Spragins (1968) and Holzmann *et al.* (2006). Moreover, testing hypotheses and classification of random variables can be discussed only if the class of all finite mixtures is identifiable. Following the ideas of Ateya (2012) and Yoichi (2011), to show identifiability for mixtures with censored observations, it is enough to show identifiability for the complete case. Thus in the following we show the identifiability of a mixture of the Student-t, slash and contaminated normal distributions for the complete case. Holzmann *et al.* (2006) proved the following: Let ϕ be the characteristic generators that give rise to families of elliptical densities $\{f_{\alpha,p} : \alpha = (\theta, \mu, \sigma^2) \in \mathbb{R}\}$. Suppose there exists a total ordering ($\leq$) on the $\beta = (a, \theta) \in (0, \infty) \times \theta$ such that $\beta_1 = \beta_2$ implies

$$\lim_{u \rightarrow \infty} \frac{\phi_{\beta_2}(u)}{\phi_{\beta_1}(u)} \rightarrow 0,$$

for corresponding characteristic generators. The finite mixture from the class $\{f_{\alpha,p} : \alpha = (\theta, \mu, \sigma^2) \in \mathbb{R}\}$ of elliptical distributions in $\mathbb{R}$ is identifiable. By using the Holzmann's approach, we prove the following proposition.

Proposition 3. *The class all finite mixture distributions relative to the Student-t, slash and contaminated normal are identifiable.*

Proof. The associated characteristic function of the scale mixture normal distribution is:

$$\phi(t) = \int_0^\infty e^{-i\mu t} \exp\left(-\frac{1}{2}u\sigma^2 t^2\right) h(u|\nu) dt.$$

(i) For the Student-t distribution: The characteristics generator is given by:

$$\phi(u; \nu) = \frac{(\sqrt{\nu u})^{\nu/2}}{2^{\nu/2-1} \Gamma(\nu/2)} K_{\nu/2}(\sqrt{u\nu}), \quad u \geq 0,$$

where $K_{\nu/2}(\cdot)$ denotes the modified Bessel function of the second kind. Then

$$\frac{\phi_{\beta_2}(u)}{\phi_{\beta_1}(u)} = \frac{\phi(a_2 u; \nu_2)}{\phi(a_1 u; \nu_1)}.$$

Using the expression above, we have

$$\begin{aligned}
\frac{\phi_{\beta_2}(u)}{\phi_{\beta_1}(u)} &= \frac{\phi(a_2 u; \nu_2)}{\phi(a_1 u; \nu_1)} \\
&= \frac{(\sqrt{\nu_2 a_2 u})^{\nu_2/2} K_{\nu_2/2}(\sqrt{a_2 u \nu_2}) / (2^{\nu_2/2-1} \Gamma(\nu_2/2))}{(\sqrt{\nu_1 a_1 u})^{\nu_1/2} K_{\nu_1/2}(\sqrt{a_1 u \nu_1}) / (2^{\nu_1/2-1} \Gamma(\nu_1/2))} \\
&= u^{(\nu_2-\nu_1)/2} \left(\frac{K_{\nu_2/2}(\sqrt{a_2 u \nu_2})}{K_{\nu_1/2}(\sqrt{a_1 u \nu_1})} \right) \left[\frac{2^{(\nu_2-\nu_1)/2} \Gamma(\nu_2/2) a_2^{\nu_2/4}}{\Gamma(\nu_1/2) a_1^{\nu_1/4}} \right] \\
&\propto u^{(\nu_2-\nu_1)/2} \left(\frac{K_{\nu_2/2}(\sqrt{a_2 u \nu_2})}{K_{\nu_1/2}(\sqrt{a_1 u \nu_1})} \right). \tag{3.18}
\end{aligned}$$

Using the asymptotic expansions of $K_\nu(z)$ (see Watson (1944), p. 202) we have

$$K_\nu(z) \approx \left(\frac{\pi}{2z} \right)^{1/2} e^{-z} \left[1 + \frac{4\nu^2 - 1}{8z} + \frac{(4\nu^2 - 1)(4\nu^2 - 3^2)}{2!(8z)^2} + \dots \right].$$

Apply limit in (3.18) we have

$$\lim_{u \rightarrow \infty} \frac{\phi_{\beta_2}(u)}{\phi_{\beta_1}(u)} = \lim_{u \rightarrow \infty} u^{(\nu_2-\nu_1)/2} e^{-\sqrt{u}(\sqrt{a_2 \nu_2} - \sqrt{a_1 \nu_1})}. \tag{3.19}$$

Then, if $a_1 \nu_1 < a_2 \nu_2$, or $a_1 \nu_1 = a_2 \nu_2$ and $\nu_1 > \nu_2$ implies that

$$\lim_{u \rightarrow \infty} \frac{\phi_{\beta_2}(u)}{\phi_{\beta_1}(u)} = 0$$

and hence the finite mixture Student- t distribution is identifiable.

(ii) For the slash distribution: The characteristic generator is given by:

$$\phi(u; \mu, \nu, \sigma) = 2\nu \exp(i\mu u) \int_0^1 u^{\nu-1} \exp\left(-\frac{1}{2} \frac{\sigma^2}{u} t^2\right) dt.$$

Using a second-order Taylor series approximation, we have

$$\phi(u; \mu, \nu, \sigma) = 2\nu \exp(i\mu u) \left(\frac{1}{\nu} - \frac{\sigma^2}{\nu-1} \right).$$

Then

$$\begin{aligned} \frac{\phi_{\beta_2}(u)}{\phi_{\beta_1}(u)} &= \frac{\phi(a_2 u; \nu_2)}{\phi(a_1 u; \nu_1)} \\ &= \frac{2\nu_2 \exp(ia_2 \mu_2 u) \left(\frac{1}{\nu_2} - \frac{\sigma_2^2}{\nu_2 - 1} \right)}{2\nu_1 \exp(ia_1 \mu_1 u) \left(\frac{1}{\nu_1} - \frac{\sigma_1^2}{\nu_1 - 1} \right)} \\ &\propto \exp[iu(a_2 \mu_2 - a_1 \mu_1)]. \end{aligned}$$

If $a_2 \mu_2 < a_1 \mu_1$ or $(a_2 = a_1 \text{ and } \mu_2 < \mu_1)$ or $(\mu_2 = \mu_1 \text{ and } a_2 < a_1)$, we have that

$$\lim_{u \rightarrow \infty} \frac{\phi_{\beta_2}(u)}{\phi_{\beta_1}(u)} = 0$$

and hence the finite mixture of slash distribution is identifiable.

(iii) For the contaminated-normal distribution: The characteristics generator is given by:

$$\phi(u; \nu, \gamma) = \exp(i\mu u) [\nu \exp(-\gamma^{-1} \sigma^2 u^2 / 2) + (1 - \nu) \exp(\sigma^2 u^2 / 2)].$$

Then

$$\begin{aligned} \frac{\phi_{\beta_2}(u)}{\phi_{\beta_1}(u)} &= \frac{\phi(a_2 u; \nu_2)}{\phi(a_1 u; \nu_1)} \\ &= \frac{\exp(i\mu_2 u) [\nu_2 \exp(-\gamma_2^{-1} \sigma_2^2 a_2^2 u^2 / 2) + (1 - \nu_2) \exp(-\sigma_2^2 a_2^2 u^2 / 2)]}{\exp(i\mu_1 u) [\nu_1 \exp(-\gamma_1^{-1} \sigma_1^2 a_1^2 u^2 / 2) + (1 - \nu_1) \exp(-\sigma_1^2 a_1^2 u^2 / 2)]} \\ &= \exp(iu(\mu_2 - \mu_1)) \times \\ &\quad \left\{ \frac{\nu_2}{\nu_1 \exp[u^2(a_2^2 \sigma_2^2 \gamma_2^{-1} - a_1^2 \sigma_1^2 \gamma_1^{-1}) / 2] + (1 - \nu_1) \exp[u^2(a_2^2 \sigma_2^2 - \gamma_1^{-1} a_1^2 \sigma_1^2)]} \right. \\ &\quad \left. + \frac{1 - \nu_2}{\nu_1 \exp[u^2(a_2^2 \sigma_2^2 - a_1^2 \sigma_1^2 \gamma_1^{-1}) / 2] + (1 - \nu_1) \exp[u^2(a_2^2 \sigma_2^2 - a_1^2 \sigma_1^2)]} \right\}. \end{aligned}$$

If $(\mu_2 = \mu_1 \text{ or } \mu_2 \neq \mu_1)$ and $[(a_1 \sigma_1 = a_2 \sigma_2 \text{ and } \gamma_2 < \gamma_1) \text{ or } (a_1 \sigma_2 > a_1 \sigma_1 \text{ and } \gamma_1 < \gamma_2)]$, we have that

$$\lim_{u \rightarrow \infty} \frac{\phi_{\beta_2}(u)}{\phi_{\beta_1}(u)} = 0$$

and hence the finite mixture of contaminated-normal distribution is identifiable.

□

3.3.2 Maximum likelihood estimation via EM-type algorithm

In this section, we present the EM-type algorithm for ML estimation of the FM-SMN-CR model. To explore the EM algorithm, we present the FM-SMN-CR in an incomplete-data framework, using the results presented in Section 2.

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

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

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

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

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

$$Y_i | Z_{ij} = 1 \sim SMN(\varphi_j + \mathbf{x}_i^\top \boldsymbol{\beta}, \sigma_j^2, \boldsymbol{\nu}).$$

For the random vector $\mathbf{Z}_i$ we use the notation $\mathbf{Z}_i \sim \text{Multinomial}(1, p_1, \dots, p_G)$. Observe that $Z_{ij} = 1$ iff $Z_i = j$. Thus, from (3.1), the setup defined above can be written hierarchically as

$$Y_i | U_i = u_i, Z_{ij} = 1 \sim N(\mathbf{x}_i^\top \boldsymbol{\beta} + \varphi_j, u_i^{-1} \sigma_j^2), \quad (3.20)$$

$$U_i | Z_{ij} = 1 \sim H(u_i; \boldsymbol{\nu}), \quad (3.21)$$

$$\mathbf{Z}_i \sim \text{Multinomial}(1, p_1, \dots, p_G), \quad (3.22)$$

for $i = 1, \dots, n$, all independent. Then, under the hierarchical representation (3.20)–(3.22), it follows that the complete log-likelihood function, apart from constants that do not depend on $\boldsymbol{\theta}$, is given by $\ell_c(\boldsymbol{\theta}) = \sum_{i=1}^n \ell_{ic}(\boldsymbol{\theta})$, where

$$\begin{aligned} \ell_{ic}(\boldsymbol{\theta}) &= \sum_{j=1}^G z_{ij} \log p_j - \frac{1}{2} \sum_{j=1}^G z_{ij} \log \sigma_j^2 - \frac{1}{2} \sum_{j=1}^G \frac{z_{ij} u_i}{\sigma_j^2} (y_i - \mathbf{x}_i^\top \boldsymbol{\beta} - \varphi_j)^2 \\ &+ \sum_{j=1}^G z_{ij} \log h(u_i | \boldsymbol{\nu}) + \frac{1}{2} \sum_{j=1}^G z_{ij} \log(u_i). \end{aligned} \quad (3.23)$$

As before, the superscript (k) indicates the estimate of the related parameter at stage k of the algorithm. It follows, after some simple algebra, that the conditional expectation of the complete log-likelihood function has the form

$$\begin{aligned}
Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) &= \sum_{i=1}^n \sum_{j=1}^G \mathcal{Z}_{ij}(\boldsymbol{\theta}^{(k)}) \log p_j - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^G \mathcal{Z}_{ij}(\boldsymbol{\theta}^{(k)}) \log \sigma_j^2 - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^G \frac{\mathcal{E}_{2ij}(\boldsymbol{\theta}^{(k)})}{\sigma_j^2} \\
&+ \sum_{i=1}^n \sum_{j=1}^G \frac{\mathcal{E}_{1ij}(\boldsymbol{\theta}^{(k)})}{\sigma_j^2} (\mathbf{x}_i^\top \boldsymbol{\beta} + \varphi_j) - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^G \frac{\mathcal{E}_{0ij}(\boldsymbol{\theta}^{(k)})}{\sigma_j^2} (\mathbf{x}_i^\top \boldsymbol{\beta} + \varphi_j)^2 \\
&+ \sum_{i=1}^n \sum_{j=1}^G E[Z_{ij} \log h(u_i|\boldsymbol{\nu})|\boldsymbol{\theta}^{(k)}, y_{\text{obs}_i}] \\
&+ \sum_{i=1}^n \sum_{j=1}^G E[Z_{ij} \log U_i|\boldsymbol{\theta}^{(k)}, y_{\text{obs}_i}], \tag{3.24}
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{E}_{sij}(\boldsymbol{\theta}^{(k)}) &= E[Z_{ij} U_i Y_i^s | \boldsymbol{\theta}^{(k)}, y_{\text{obs}_i}], \quad s = 0, 1, 2, \\
\mathcal{Z}_{ij}(\boldsymbol{\theta}^{(k)}) &= E[Z_{ij} | \boldsymbol{\theta}^{(k)}, y_{\text{obs}_i}].
\end{aligned}$$

By using known properties of conditional expectation, we obtain

$$\mathcal{Z}_{ij}(\boldsymbol{\theta}^{(k)}) = \frac{p_j^{(k)} g_{ij}(y_i | \mathbf{x}_i, \boldsymbol{\gamma}^{(k)}, \boldsymbol{\theta}_j^{(k)})}{\sum_{j=1}^G p_j^{(k)} g_{ij}(y_i | \mathbf{x}_i, \boldsymbol{\gamma}^{(k)}, \boldsymbol{\theta}_j^{(k)})},$$

and

$$\mathcal{E}_{sij}(\boldsymbol{\theta}^{(k)}) = \mathcal{Z}_{ij}(\boldsymbol{\theta}^{(k)}) E[U_i Y_i^s | \boldsymbol{\theta}^{(k)}, y_{\text{obs}_i}, Z_{ij} = 1],$$

where $s = 0, 1, 2$. The conditional expectations of the form $E[U_i Y_i^s | \boldsymbol{\theta}^{(k)}, y_{\text{obs}_i}, Z_{ij} = 1]$, $s = 0, 1, 2$, can be computed following the same paths that led to equations (3.9) and (3.10) in Section 2. We have that

$$E[U_i Y_i^s | \boldsymbol{\theta}^{(k)}, Y_{\text{obs}_i}, Z_{ij}] = E\left\{Y_i^s E[U_i | \boldsymbol{\theta}^{(k)}, Y_i, Z_{ij}] | \boldsymbol{\theta}^{(k)}, Y_{\text{obs}_i}, Z_{ij}\right\}$$

and $E[U_i | \boldsymbol{\theta}^{(k)}, y_i, Z_{ij} = 1]$ can be easily derived from the result of Osorio *et al.* (2007) and for a censored observation,

$$E[U_i Y_i^s | \boldsymbol{\theta}^{(k)}, Y_i, Z_{ij} = 1] = E[U_i Y_i^s | \boldsymbol{\theta}^{(k)}, Y_i \leq \kappa_i], \tag{3.25}$$

Thus, we have closed form expression for all the quantities involved in path E of the algorithm. Ignore for the moment the restriction that the $p_1, \dots, p_G$ are ordered. Next, we describe the ECME algorithm for maximum likelihood estimation of the parameters of the FM-SMN-CR model.

E-step: Given $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$, compute $\mathcal{E}_{sij}(\boldsymbol{\theta}^{(k)})$, $s = 0, 1, 2$ and $\mathcal{Z}_{ij}(\boldsymbol{\theta}^{(k)})$ for $i = 1, \dots, n$, $j = 1, \dots, G$.

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

$$p_j^{(k+1)} = \frac{\sum_{i=1}^n \mathcal{Z}_{ij}(\boldsymbol{\theta}^{(k)})}{n}, \quad (3.26)$$

$$\begin{aligned} \boldsymbol{\beta}^{(k+1)} &= \left(\sum_{i=1}^n \sum_{j=1}^G \frac{\mathcal{E}_{0ij}(\boldsymbol{\theta}^{(k)})}{\sigma_j^{2(k)}} \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1} \\ &\times \sum_{i=1}^n \sum_{j=1}^G \frac{\mathbf{x}_i}{\sigma_j^{2(k)}} \left(\mathcal{E}_{1ij}(\boldsymbol{\theta}^{(k)}) - \mathcal{E}_{0ij}(\boldsymbol{\theta}^{(k)}) \varphi_j^{(k)} \right), \end{aligned} \quad (3.27)$$

$$\varphi_j^{(k+1)} = \frac{\sum_{i=1}^n \left[\mathcal{E}_{1ij}(\boldsymbol{\theta}^{(k)}) - \mathcal{E}_{0ij}(\boldsymbol{\theta}^{(k)}) \mathbf{x}_i^\top \boldsymbol{\beta}^{(k)} \right]}{\sum_{i=1}^n \mathcal{E}_{0ij}(\boldsymbol{\theta}^{(k)})}, \quad (3.28)$$

$$\sigma_j^{2(k+1)} = \frac{\sum_{i=1}^n \Delta_{ij}^{(k)}}{\sum_{i=1}^n \mathcal{Z}_{ij}(\boldsymbol{\theta}^{(k)})}, \quad j = 1, \dots, G, \quad (3.29)$$

where

$$\begin{aligned} \Delta_{ij}^{(k)} &= \left(\mathcal{E}_{2ij}(\boldsymbol{\theta}^{(k)}) + \mathcal{E}_{0ij}(\boldsymbol{\theta}^{(k)}) \varphi_j^{2(k)} + \mathcal{E}_{0ij}(\boldsymbol{\theta}^{(k)}) (\mathbf{x}_i^\top \boldsymbol{\beta}^{(k)})^2 \right. \\ &\quad - 2\mathcal{E}_{1ij}(\boldsymbol{\theta}^{(k)}) \varphi_j^{(k)} - 2\mathcal{E}_{1ij}(\boldsymbol{\theta}^{(k)}) \mathbf{x}_i^\top \boldsymbol{\beta}^{(k)} \\ &\quad \left. + 2\mathcal{E}_{0ij}(\boldsymbol{\theta}^{(k)}) \varphi_j^{(k)} \mathbf{x}_i^\top \boldsymbol{\beta}^{(k)} \right). \end{aligned}$$

Following Bartolucci and Scaccia (2005), we can also obtain an estimate of β_0 as $\beta_0^{(k)} = \sum_{j=1}^G p_j^{(k)} \varphi_j^{(k)}$,

and an estimate of μ_j as $\mu_j^{(k)} = \varphi_j^{(k)} - \beta_0^{(k)}$, $j = 1, \dots, G$.

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

$$\nu_j^{(k+1)} = \arg \max_{\nu} \sum_{i=1}^n \log \left(\sum_{j=1}^G p_j^{(k+1)} g_{ij}(y_i | \mathbf{x}_i, \nu_j, \beta_0^{(k+1)}, \boldsymbol{\beta}^{(k+1)}, \boldsymbol{\theta}_j^{(k+1)}) \right), \quad (3.30)$$

where $g_{ij}(y_i | \mathbf{x}_i, \gamma, \boldsymbol{\theta}_j)$ as defined in (3.17).

3.3.3 Starting values

It is well known that EM and related algorithms can be sensitive to the initial values chosen, see the discussion in McLachlan and Peel (2000) for example. Note that mixture models can provide a multimodal log-likelihood function. In this sense, the method of maximum likelihood estimation through EM algorithm may not give maximum global solutions if the starting values are far from the real parameter values. Thus, the choice of starting values for the EM algorithm in the mixture context plays a big role in parameter estimation. Note that the number of components G is considered known *a priori* and therefore they are not parameters. In our examples and simulation studies, we consider the following procedure for the FM-SMN-CR model (available in the package `CensMixReg`) (Benites and Lachos, 2015):

- For $\beta_0^{(0)}$ and $\beta^{(0)}$, use the ordinary least-squares (OLS) estimate in the regression model defined in (3.4).
- Partition the residuals into G groups using the K-means clustering algorithm ;
- Compute the proportion of data points belonging to the same cluster j , say $p_j^{(0)}$, $j = 1, \dots, G$. This is the initial value for p_j ;
- For each group j , compute the initial values $\mu_j^{(0)}$, $(\sigma_j^2)^{(0)}$ using the method of moments estimators.
- The starting values for ν_j are taken to be equal to 3 in the Student-t and slash cases and $\nu_j = (0.1, 0.1)$ in the contaminated normal case.

We adopt the Bayesian information criterion (BIC; Schwarz (1978)) to select the number of components in mixture models. Herein, the form of BIC is given by

$$\text{BIC} = -2\ell(\hat{\theta}) + \rho \log n,$$

where $\ell(\hat{\theta})$ is the current log-likelihood and ρ is the number of the parameters in the model. Accordingly, models with small BIC scores are preferred. Recently McNicholas and Murphy (2008) showed the effectiveness of the BIC in selecting the number of components for Gaussian mixture models.

3.3.4 Stopping rule

To assess the convergence of the EM algorithm for the FM-SMN-CR model, we adopt the Aitken acceleration based stopping criterion

$$|\ell^{(k+1)} - \ell_{\infty}^{(k+1)}| < \varepsilon,$$

where $\ell^{(k+1)}$ is the observed log-likelihood evaluated at $\boldsymbol{\theta}^{(k)}$ ignoring the order restriction on the p_j . Assuming convergence to a MLE $\hat{\boldsymbol{\theta}}$, we simply permute the group index j on $\boldsymbol{\theta}$, satisfying the ordering restriction. Also, $\ell_{\infty}^{(k+1)}$ is the asymptotic estimate of the log-likelihood at iteration $k+1$ (McLachlan and Krishnan, 2008, Chap. 4.9) where

$$\ell_{\infty}^{(k+1)} = \ell^{(k)} + \frac{\ell^{(k+1)} - \ell^{(k)}}{1 - c^{(k)}}. \quad (3.31)$$

In the above, $c^{(k)}$ denotes the Aitken acceleration at the k th iteration, given by

$$c^{(k)} = \frac{\ell^{(k+1)} - \ell^{(k)}}{\ell^{(k)} - \ell^{(k-1)}},$$

where ε is the desired tolerance. For numerical analysis in Section 3.4 and 3.5, a default value of $\varepsilon = 10^{-6}$ will be used to terminate the iterations. Note that the above procedure is also applicable for the simple case ($g = 1$) by treating $Z_{ij} = 1$.

3.3.5 Variance matrix and asymptotic properties

A simple way of obtaining the standard errors of ML estimates of mixture model parameters is to approximate the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$ by the inverse of the observed information matrix. 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 (3.17). In this work we use the alternative method suggested by Basford *et al.* (1997a), which consists of approximating the inverse of the covariance matrix by

$$\mathbf{I}_o(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^n \hat{\mathbf{s}}_i \hat{\mathbf{s}}_i^\top, \quad \text{where} \quad \hat{\mathbf{s}}_i = \mathbf{E} \left[\frac{\partial \ell_{ic}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \middle| \mathbf{y}_{\text{obs}}, \boldsymbol{\theta} \right], \quad (3.32)$$

where $\ell_{ic}(\boldsymbol{\theta})$ is as given in (3.23) and

$$\hat{\mathbf{s}}_i = (\hat{s}_{i,\boldsymbol{\beta}}^\top, \hat{s}_{i,p_1}, \dots, \hat{s}_{i,p_{G-1}}, \hat{s}_{i,\sigma_1^2}, \dots, \hat{s}_{i,\sigma_G^2}, \hat{s}_{i,\varphi_1}, \dots, \hat{s}_{i,\varphi_G}, \hat{s}_{i,\nu_1}, \dots, \hat{s}_{i,\nu_G})^\top.$$

Expressions for the elements $\hat{s}_{i,\beta}, \hat{s}_{i,p_j}, \hat{s}_{i,\sigma_j^2}, \hat{s}_{i,\varphi_j}, \hat{s}_{i,\mu_j}$, are given in the following:

$$\begin{aligned}\hat{s}_{i,\beta} &= \sum_{j=1}^G \left\{ \frac{1}{\hat{\sigma}_j^2} \left[\mathcal{E}_{1ij}(\hat{\theta}) \mathbf{x}_i - \mathcal{E}_{0ij}(\hat{\theta}) (\mathbf{x}_i^\top \hat{\beta} + \hat{\varphi}_j) \mathbf{x}_i \right] \right\}, \\ \hat{s}_{i,p_j} &= \frac{\mathcal{Z}_{ij}(\hat{\theta})}{\hat{p}_j} - \frac{\mathcal{Z}_{ig}(\hat{\theta})}{\hat{p}_g}, \\ \hat{s}_{i,\sigma_j^2} &= -\frac{1}{2} \left\{ \frac{1}{\hat{\sigma}_j^4} \left[\mathcal{Z}_{ij}(\hat{\theta}) \hat{\sigma}_j^2 - \mathcal{E}_{2ij}(\hat{\theta}) + 2\mathcal{E}_{1ij}(\hat{\theta}) (\mathbf{x}_i^\top \hat{\beta} + \hat{\varphi}_j) \right. \right. \\ &\quad \left. \left. - \mathcal{E}_{0ij}(\hat{\theta}) (\mathbf{x}_i^\top \hat{\beta} + \hat{\varphi}_j)^2 \right] \right\}, \\ \hat{s}_{i,\varphi_j} &= \frac{1}{\sigma_j^2} \left[\mathcal{E}_{1ij}(\hat{\theta}) - \mathcal{E}_{0ij}(\hat{\theta}) (\mathbf{x}_i^\top \hat{\beta} + \hat{\varphi}_j) \right], \\ \hat{s}_{i,\nu_j} &= E \left[\frac{\partial \ell_{ic}(\theta)}{\partial \nu_j} \middle| \mathbf{y}_{\text{obs}}, \theta \right],\end{aligned}\tag{3.33}$$

where $\psi(\cdot)$ denotes the digamma function. It is important to stress that the standard error of ν , obtained from $\hat{s}_{i,\nu}$, depends heavily on the calculation of conditional expectation $E[\log(U_i)|y_{\text{obs}_i}, \hat{\theta}]$ (see the equation below), which relies on computationally intensive Monte Carlo integrations. For, the Student-t, we have an analytical expression but for slash and contaminated normal distributions we do not have this expected value. Note (3.33) the following expressions:

- For the Student-t distribution: $U \sim \text{Gamma}(\nu/2, \nu/2)$ with $\nu > 0$, then

$$\hat{s}_{i,\nu_j} = \frac{1}{2} \left\{ \mathcal{Z}_{ij}(\hat{\theta}) \left[\log \left(\frac{\hat{\nu}}{2} \right) + 1 - \psi \left(\frac{\hat{\nu}}{2} \right) + E(\log(U_i)|y_{\text{obs}_i}, \hat{\theta}) \right] - \mathcal{E}_{0ij}(\hat{\theta}) \right\},$$

where

$$E(\log(U_i)|y_{\text{obs}_i}, \hat{\theta}) = \begin{cases} \psi \left(\frac{\nu_j}{2} \right) - \log \left(\frac{\nu_j}{2} \right) & \text{if } Y_i \leq \kappa_i; \\ \psi \left(\frac{\nu_j+1}{2} \right) - \log \left(\frac{\nu_j+d_{ij}^2}{2} \right) & \text{if } Y_i > \kappa_i, \end{cases}\tag{3.34}$$

with $d_{ij} = (y_i - \mathbf{x}_i^\top \beta - \varphi_j)/\sigma_j^2$.

- For the slash distribution: $U \sim \text{Beta}(\nu, 1)$ with $\nu \in \mathbb{R}_+$, then

$$\hat{s}_{i,\nu_j} = \frac{1}{\hat{\nu}_j} + E(\log(U_i)|y_{\text{obs}_i}, \hat{\theta}),$$

where $E(\log(U_i)|y_{\text{obs}_i}, \hat{\theta}) = \int_0^\infty \log(u_i) f(u_i|y_{\text{obs}_i}) du_i$, can be approximated by the Monte Carlo integrations.

- For the contaminated normal-distribution: $f(u_i|\nu) = \nu \mathbb{I}_{\{\gamma\}}(u_i) + (1 - \nu) \mathbb{I}_{\{1\}}(u_i)$, $\nu, \gamma \in (0, 1)$,

then

$$\widehat{s}_{i,\nu_j} = \mathcal{Z}_{ij}(\widehat{\boldsymbol{\theta}})E \left[\frac{\partial}{\partial \nu_j} \log h(U_i|\nu_j) \middle| y_{\text{obs}_i}, \widehat{\boldsymbol{\theta}} \right] + \frac{1}{2} \mathcal{Z}_{ij}(\widehat{\boldsymbol{\theta}})E \left[\frac{\partial}{\partial \nu_j} \log(U_i) \middle| y_{\text{obs}_i}, \widehat{\boldsymbol{\theta}} \right]$$

and

$$\widehat{s}_{i,\gamma_j} = \mathcal{Z}_{ij}(\widehat{\boldsymbol{\theta}})E \left[\frac{\partial}{\partial \gamma_j} \log h(U_i|\nu_j) \middle| y_{\text{obs}_i}, \widehat{\boldsymbol{\theta}} \right] + \frac{1}{2} \mathcal{Z}_{ij}(\widehat{\boldsymbol{\theta}})E \left[\frac{\partial}{\partial \gamma_j} \log(U_i) \middle| y_{\text{obs}_i}, \widehat{\boldsymbol{\theta}} \right].$$

Since the expectations above are unknown, we will not consider the standard error for the contaminated normal using the approximation.

Now, returning to our original parameterization, we find the Hessian matrix for the original parameter vector $\boldsymbol{\theta}^* = (\boldsymbol{\beta}^\top, (p_1, \dots, p_{G-1})^\top, \beta_0, \mu_1, \dots, \mu_G, \sigma_1^2, \dots, \sigma_G^2, \nu_1, \dots, \nu_G)^\top$. Using the delta method, we have

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

where $\mathbf{J}(\boldsymbol{\theta}^*|\boldsymbol{\theta})$ is the a matrix of order $(p + 4G) \times (p + 4G - 1)$, defined by

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

where $\boldsymbol{\gamma} = \frac{\partial \boldsymbol{\mu}}{\partial \mathbf{p}^\top} = (\mathbf{p}^{-1})\mathbf{A}^\top$ is a matrix of dimension $G \times (G - 1)$, with $\mathbf{A} = (\mu_G - \mu_1, \dots, \mu_G - \mu_{G-1})^\top$ and $\mathbf{p}^{-1} = (1/p_1, \dots, 1/p_{G-1}, 1/p_G)^\top$. The square roots of the diagonal elements of the $(p + 4G) \times (p + 4G)$ matrix $\mathbf{J}(\boldsymbol{\theta}^*|\boldsymbol{\theta})\mathbf{I}_o(\widehat{\boldsymbol{\theta}})^{-1}\mathbf{J}(\boldsymbol{\theta}^*|\boldsymbol{\theta})^\top$ are the standard errors of parameters $\boldsymbol{\theta}^*$.

3.4 Simulation studies

In this section, we perform simulation studies to evaluate the performance and properties of the ML estimators of the FM-SMN-CR model, considering the normal, Student-t and slash distributions. The estimates are obtained through the ECME algorithm presented in Section 3.3.2. The goals are to evaluate to accuracy of the estimates (Simulation study 1) and evaluate the effect of different censoring levels (Simulation study 2) and sample sizes (Simulation study 3) on the

Table 3.1: *Simulation Study 1 (Scenario 1). Mean, sample standard errors (MC SE), average values (across 500 samples) of the standard errors computed using the information method (MC IM SE) and the percentage coverage of the resulting 95% confidence intervals (COV MC) from the FM-SMN-CR. True values of parameters are in parentheses.*

Parameters	Measure	FM-N-CR				FM-T-CR				FM-SL-CR			
		8%	10%	20%	35%	8%	10%	20%	35%	8%	10%	20%	35%
$\beta_0(1)$	Mean	1.0669	1.0793	1.1635	1.2975	1.0599	1.0807	1.1688	1.3168	1.0662	1.0779	1.1581	1.3001
	MC SE	0.0809	0.0821	0.0844	0.0848	0.0932	0.0913	0.0949	0.0947	0.0897	0.0814	0.0889	0.0903
	MC IM SE	0.1106	0.1130	0.1204	0.0103	0.1339	0.1303	0.1392	0.1428	0.1307	0.1397	0.1322	0.1328
	95% COV	97.2%	96.4%	83.2%	80.2%	97.8%	97.8%	85.6%	81.0%	97.5%	95.8%	89.8%	82.0%
	Mean	-1.0004	-1.0025	-0.9968	-0.9961	-0.9980	-0.9993	-0.9975	-0.9976	-0.9956	-1.0032	-0.9884	-1.0000
$\beta_1(-1)$	MC SE	0.0646	0.0671	0.0740	0.0826	0.0854	0.0862	0.0908	0.0996	0.0784	0.0839	0.0907	0.0878
	MC IM SE	0.0647	0.0664	0.0723	0.0781	0.0781	0.0768	0.0800	0.0807	0.0701	0.0748	0.0801	0.0754
	95% COV	93.8%	94.8%	94.2%	93.8%	94.2%	92.8%	92.2%	89.4%	98.2%	95.2%	82.8%	88.1%
	Mean	4.0004	4.0033	3.9987	4.0037	4.0018	3.9967	3.9969	3.9982	3.9980	4.0050	4.0035	4.0035
	MC SE	0.0662	0.0648	0.0688	0.0817	0.0827	0.0820	0.0935	0.0979	0.0761	0.0820	0.0840	0.0961
$\beta_2(4)$	MC IM SE	0.0645	0.0668	0.0703	0.0773	0.0779	0.0774	0.0799	0.0796	0.0700	0.0783	0.0799	0.0741
	95% COV	94.8%	95.6%	95.4%	93.6%	94.0%	92.6%	90.0%	88.0%	98.7%	94.2%	90.8%	87.8%
	Mean	0.9327	0.9211	0.8358	0.6979	0.9383	0.9223	0.8356	0.6935	0.9345	0.9205	0.8369	0.7019
	MC SE	0.0664	0.0624	0.0621	0.0590	0.0689	0.0662	0.0679	0.0604	0.0631	0.0607	0.0577	0.0546
	MC IM SE	0.0827	0.0827	0.0824	0.0731	0.0894	0.0899	0.0978	0.0531	0.0897	0.0800	0.0778	0.0417
$\mu_1(1)$	95% COV	96.8%	95.8%	97.8%	80.0%	97.8%	97.8%	99.8%	95.6%	97.7%	95.2%	98.8%	94.8%
	Mean	-4.0650	-4.0778	-4.1624	-4.2989	-4.0590	-4.0796	-4.1690	-4.3222	-4.0637	-4.0781	-4.1545	-4.2933
	MC SE	0.0830	0.0806	0.0799	0.0839	0.0929	0.0897	0.0989	0.0975	0.0836	0.0842	0.0819	0.0963
	MC IM SE	0.0721	0.0861	0.0731	0.0627	0.0995	0.0841	0.0839	0.0824	0.0995	0.0841	0.0715	0.0841
	95% COV	96.0%	93.8%	90.5%	81.0%	97.8%	97.8%	83.0%	80.0%	95.7%	98.2%	81.8%	82.4%
$\mu_2(-4)$	Mean	0.2987	0.2998	0.2984	0.2990	0.3003	0.3022	0.2996	0.2981	0.3173	0.3225	0.3424	0.3542
	MC SE	0.0152	0.0157	0.0169	0.0189	0.0266	0.0268	0.0259	0.0286	0.0414	0.0441	0.0282	0.0638
	MC IM SE	0.0158	0.0160	0.0169	0.0188	0.0259	0.0255	0.0245	0.0241	0.0301	0.0510	0.0101	0.0610
	95% COV	96.8%	94.4%	94.8%	94.2%	93.8%	93.6%	93.6%	90.4%	95.8%	94.6%	91.8%	90.8%
	Mean	0.5987	0.5952	0.5998	0.5925	0.6097	0.5952	0.5941	0.5770	0.6752	0.6932	0.6876	0.6880
$\sigma_1^2(0.3)$	MC SE	0.0635	0.0612	0.0746	0.0760	0.1009	0.1022	0.1048	0.1124	0.1275	0.1284	0.1010	0.1512
	MC IM SE	0.0651	0.0655	0.0700	0.0777	0.1257	0.1259	0.1321	0.1410	0.1010	0.1032	0.1010	0.1032
	95% COV	95.8%	94.6%	94.0%	95.2%	97.8%	97.6%	96.8%	94.6%	98.9%	92.7%	98.9%	89.7%
	Mean	-	-	-	-	3.1000	3.1168	3.2822	3.3568	3.1606	3.2008	3.2191	3.3840
	MC SE	-	-	-	-	0.2612	0.2735	0.2648	0.2676	0.2598	0.2790	0.2498	0.2490
$\nu_1(3)$	MC IM SE	-	-	-	-	0.2273	0.2143	0.1755	0.1997	0.2777	0.2550	0.1977	0.1850
	95% COV	-	-	-	-	92.2%	86.6%	84.4%	82.8%	92.9%	90.8%	88.9%	83.8%
	Mean	-	-	-	-	4.2202	4.2114	4.0905	4.2232	4.2064	4.2192	4.2371	4.2510
	MC SE	-	-	-	-	0.4291	0.4375	0.4305	0.4345	0.4710	0.4825	0.4410	0.4125
	MC IM SE	-	-	-	-	0.4294	0.4022	0.3959	0.3755	0.4037	0.4147	0.3737	0.3247
$\nu_2(4)$	95% COV	-	-	-	-	95.8%	92.8%	89.6%	87.0%	91.2%	90.4%	89.9%	88.4%
	Mean	0.8134	0.8157	0.8328	0.8603	0.8122	0.8156	0.8330	0.8617	0.8130	0.8158	0.8323	0.8595
	MC SE	0.0131	0.0124	0.0122	0.0115	0.0136	0.0130	0.0134	0.0119	0.0123	0.0120	0.0111	0.0109
	MC IM SE	0.0286	0.0286	0.0290	0.0295	0.0294	0.0297	0.0217	0.0257	0.0113	0.0212	0.0213	0.0312
	95% COV	91.8%	90.8%	86.8%	85.8%	90.8%	89.8%	88.6%	84.4%	95.7%	92.8%	88.7%	85.8%

Figure 3.1: *Target mixture densities from which the dataset was simulated. (a) Scenario 1 and (b) Scenario 2 of simulation study 1.*

ML estimates. In all simulation studies we consider 500 Monte Carlo replications of the proposed model with three covariates and we set $\beta_0 = 1$, $\beta = (\beta_1, \beta_2)^\top = (-1, 4)^\top$, $\mu_1 = 1$, $\mu_2 = -4$ and

$p_1 = 0.8$. The initial values of the ECME algorithm were obtained considering the strategy given in Subsection 4.4.2.

Our analyses were performed with a 3.40GHz Intel Core i7 processor with 31.9GB of RAM. The computational procedures were implemented using the R software R Core Team (2015), through the package CensMixReg (Benites and Lachos, 2015).

3.4.1 Simulation study 1: parameter recovery

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. The data were artificially generated from the following FM-SMN-CR model with two components:

$$Y_i = \beta_0 + \mu_j + \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad j = 1, 2,$$

where $\varepsilon_i \sim p_1 \text{SMN}(\mu_1, \sigma_1^2, \nu_1) + p_2 \text{SMN}(\mu_2, \sigma_2^2, \nu_2)$ such that $\mathbf{x}_i^\top = (x_{i1}, x_{i2})^\top$ with $x_{ki} \sim U(0, 1)$, $i = 1, \dots, n$, $k = 1, 2$ and several censoring proportion settings (8%, 10%, 20% and 35%).

In this case, we consider Monte Carlo samples of size $n = 1000$. In addition, we consider the following scenarios (depicted in Figure 3.1):

Scenario 1: $\sigma_1^2 = 0.3$, $\sigma_2^2 = 0.6$ (small and different variances) with $\nu_1 = 3$, $\nu_2 = 4$.

Scenario 2: $\sigma_1^2 = 2.5$ and $\sigma_2^2 = 2.5$ (large and equal variances) with $\nu_1 = 3$, $\nu_2 = 3$.

We used the ML estimation via EM algorithm for each sample, using the stopping criterion (4.18). The average values (Mean) and standard deviations (MC Sd) of the estimates across the 500 Monte Carlo samples were computed. We also were computed the average (IM SE) values of the approximate standard deviations of the estimates obtained through the method described in Subsection 4.4.4. The results are presented in Tables 3.1 and 3.2. Note that in both scenarios, the estimates of the regression parameters β_1 and β_2 are less sensitive to the variation in the censoring level. Also ν_1 and ν_2 tend to be overestimated when the censoring level in the two scenarios increases (a little more in the second scenario). In general, in the context of this study the results suggest that the FM-T-CR and FM-SL-CR models produce estimates that are to changes in the censoring level and that behave well even for poorly separated components, except for the regression parameter β_0 and the heavy-tail parameters ν_1 and ν_2 . It also can be seen from this table that, except for the parameters β_0 , ν_1 and ν_2 , the estimation method of the standard errors

Table 3.2: *Simulation study 1 (Scenario 2). Mean, sample standard errors (MC SE), average values (across 500 samples) of the standard errors computed using the information method (MC IM SE) and the percentage coverage of the resulting 95% confidence intervals (COV MC) from the FM-SMN-CR. True values of parameters are in parentheses.*

Parameters	Measure	FM-N-CR				FM-T-CR				FM-SL-CR			
		8%	10%	20%	35%	8%	10%	20%	35%	8%	10%	20%	35%
$\beta_0(1)$	Mean	1.0703	1.0937	1.2008	1.3778	1.0556	1.0752	1.1960	1.4301	1.1040	1.0966	1.2391	1.4239
	MC SE	0.1639	0.1722	0.1752	0.1819	0.2132	0.2271	0.2082	0.2330	0.2036	0.2148	0.2172	0.2022
	MC IM SE	0.2705	0.2801	0.2922	0.2188	0.3602	0.3595	0.3120	0.3178	0.3803	0.3574	0.4803	0.5482
	COV MC	97.4%	96.8%	88.9%	85.8%	97.8%	96.8%	87.6%	85.0%	97.1%	95.0%	88.1%	81.1%
$\beta_1(-1)$	Mean	-0.9812	-0.9902	-0.9899	-0.9819	-0.9955	-0.9881	-0.9870	-1.0222	-1.0111	-0.9832	-1.0072	-0.9999
	MC SE	0.2161	0.2222	0.2244	0.2469	0.2674	0.2768	0.2675	0.2716	0.2698	0.2720	0.2666	0.2745
	MC IM SE	0.2152	0.2172	0.2270	0.2437	0.2358	0.2275	0.2826	0.2385	0.2571	0.2883	0.2558	0.1558
	95% COV	94.4%	95.4%	96.0%	94.0%	91.4%	91.0%	90.4%	90.0%	93.7%	92.8%	91.8%	90.8%
$\beta_2(4)$	Mean	4.0058	4.0034	3.9982	4.0021	4.0114	4.0126	4.0064	4.0184	3.9794	4.0052	3.9969	4.0053
	MC SE	0.2097	0.2048	0.2282	0.2417	0.2471	0.2759	0.2667	0.2925	0.2682	0.2554	0.2827	0.2763
	MC IM SE	0.2166	0.2135	0.2266	0.2437	0.2309	0.2282	0.2797	0.2404	0.2210	0.2802	0.2094	0.1563
	95% COV	96.4%	96.4%	94.2%	93.4%	96.8%	95.6%	94.8%	93.9%	95.8%	95.2%	91.1%	92.1%
$\mu_1(1)$	Mean	0.9489	0.9367	0.8631	0.7453	0.9833	0.9584	0.8657	0.7043	1.0449	1.0217	0.9643	0.8380
	MC SE	0.0792	0.0813	0.0719	0.0748	0.1085	0.1135	0.1081	0.1151	0.1260	0.1291	0.1140	0.1052
	MC IM SE	0.0606	0.0631	0.0636	0.0670	0.0994	0.1099	0.0978	0.1531	0.1094	0.1642	0.2369	0.5369
	95% COV	95.8%	94.8%	92.8%	80.0%	97.8%	96.8%	92.8%	89.6%	97.6%	95.1%	91.6%	85.8%
$\mu_2(-4)$	Mean	-4.0589	-4.0584	-4.1519	-4.3050	-3.9486	-3.9482	-4.0713	-4.2741	-3.6559	-3.6693	-3.6823	-3.8350
	MC SE	0.1917	0.1928	0.2047	0.2175	0.2802	0.3126	0.3712	0.4618	0.5092	0.5268	0.5164	0.5024
	MC IM SE	0.2941	0.2044	0.2844	0.1539	0.2764	0.2461	0.3263	0.3767	0.4764	0.3778	0.3714	0.9714
	95% COV	95.0%	92.8%	90.0%	81.0%	97.0%	95.8%	90.0%	88.0%	96.0%	92.1%	87.3%	87.8%
$\sigma_1^2(2.5)$	Mean	2.4570	2.4595	2.4385	2.3894	2.4084	2.4383	2.4258	2.6301	2.4302	2.7553	2.3098	2.2676
	MC SE	0.1477	0.1554	0.1514	0.1321	0.2732	0.2524	0.2898	0.2964	0.3631	0.3588	0.3659	0.3254
	MC IM SE	0.1625	0.1640	0.1683	0.1561	0.3424	0.3448	0.3013	0.2751	0.2424	0.3051	0.3714	0.2273
	95% COV	97.6%	96.0%	95.0%	94.2%	92.8%	95.8%	92.2%	91.4%	93.8%	96.3%	95.9%	90.0%
$\sigma_2^2(2.5)$	Mean	2.4813	2.5032	2.4787	2.5096	2.7294	2.7138	2.6790	2.4532	2.6959	2.5845	2.7920	2.8047
	MC SE	0.1679	0.1725	0.1725	0.1666	0.7873	0.8360	0.9526	0.9681	0.9685	0.9512	0.9930	0.8675
	MC IM SE	0.2146	0.2193	0.2406	0.2002	0.8081	0.8105	0.8294	0.8399	0.8100	0.8429	0.8174	0.9487
	95% COV	96.8%	95.3%	93.0%	92.4%	97.8%	96.8%	94.8%	93.6%	96.7%	95.7%	94.8%	93.7%
$\nu_1(3)$	Mean	-	-	-	-	3.1369	3.3621	3.4810	3.5128	3.2126	3.6320	3.4259	3.5320
	MC SE	-	-	-	-	0.4265	0.4214	0.4251	0.4469	0.6968	0.4599	0.6925	0.5599
	MC IM SE	-	-	-	-	0.4705	0.2708	0.2467	0.2295	0.4705	0.2799	0.4714	0.3799
	95% COV	-	-	-	-	93.2%	86.6%	81.4%	80.8%	91.2%	87.5%	85.8%	82.3%
$\nu_2(3)$	Mean	-	-	-	-	2.9422	2.9118	2.9809	3.5403	0.3288	3.8635	3.7834	3.8635
	MC SE	-	-	-	-	0.4291	0.4375	0.4305	0.4345	0.5567	0.5900	0.4404	0.6900
	MC IM SE	-	-	-	-	0.4294	0.4022	0.3959	0.3755	0.4294	0.4055	0.5148	0.5055
	95% COV	-	-	-	-	98.8%	97.4%	95.4%	90.0%	95.8%	95.2%	88.8%	85.9%
$p_1(0.8)$	Mean	0.8103	0.8122	0.8276	0.8521	0.8000	0.8038	0.8234	0.8567	0.7746	0.7789	0.7895	0.8181
	MC SE	0.0172	0.0178	0.0163	0.0165	0.0255	0.0279	0.0279	0.0310	0.0408	0.0419	0.0373	0.0327
	MC IM SE	0.0306	0.0306	0.0307	0.0309	0.0246	0.0358	0.0370	0.0455	0.0346	0.0405	0.0714	0.0790
	95% COV	92.8%	91.3%	87.8%	80.8%	91.8%	90.8%	89.6%	88.4%	90.8%	90.1%	87.8%	82.1%

provides relatively close results (IM SE and MC Sd), indicating that the proposed asymptotic approximation for the variances of the ML estimates (Equation (4.19)) is reliable. This conclusion can be also be obtained through the coverage probability of parameters (COV), since in general a confidence interval above 95% coverage is maintained for each parameter. The COV is defined as $CP(\hat{\theta}) = (1/m) \sum_{j=1}^m I(\theta \in [\hat{\theta}_L, \hat{\theta}_U])$, where I is the indicator function such that θ lies in the interval $[\hat{\theta}_L, \hat{\theta}_U]$, with $\hat{\theta}_L$ and $\hat{\theta}_U$ being the estimated lower and upper bounds of the 95% CI, respectively. Note that the coverage of the 95% confidence intervals is very closed to the true confidence level for all parameters when the censoring level is small and decreases when the censoring level increases.

3.4.2 The effect of different censoring proportions (simulation study 2)

Simulation study 2 considers samples of sizes $n = 1000$, generated assuming 8%, 10%, 20% and 35% censoring level and $\nu_1 = 3$, $\nu_2 = 4$. Table 3.3 presents the relative bias and the root-mean-

Table 3.3: *Simulation study 2: Relative bias and RMSE of $\hat{\theta}$ for different proportions of censoring.*

Parameters	Measure	FM-N-CR				FM-T-CR				FM-SL-CR			
		8%	10%	20%	35%	8%	10%	20%	35%	8%	10%	20%	35%
$\beta_0(1)$	Relative bias	0.0569	0.0796	0.1390	0.2780	0.0480	0.0651	0.1365	0.2476	0.0514	0.0702	0.1312	0.2333
	RMSE	0.1052	0.1147	0.1775	0.3171	0.1101	0.1258	0.1932	0.3321	0.1079	0.1220	0.1938	0.3187
$\beta_1(-1)$	Relative bias	-0.0039	-0.0037	0.0043	-0.0011	-0.0004	0.0068	-0.0043	0.0043	-0.0048	0.0015	-0.0111	0.0002
	RMSE	0.0640	0.0717	0.0740	0.0792	0.0817	0.0867	0.0928	0.0981	0.0781	0.0759	0.0835	0.1001
$\beta_2(4)$	Relative bias	0.0019	-0.0062	0.0075	-0.0059	0.0033	-0.0105	0.0030	-0.0021	0.0056	-0.0029	-0.0014	0.0071
	RMSE	0.0723	0.0675	0.0674	0.0798	0.0887	0.0842	0.0912	0.1027	0.0798	0.0830	0.0889	0.0976
$\mu_1(1)$	Relative bias	-0.0660	-0.0705	-0.1442	-0.2769	-0.0489	-0.0586	-0.1345	-0.2411	-0.0526	-0.0665	-0.1231	-0.2328
	RMSE	0.0916	0.1036	0.1712	0.3061	0.0904	0.1069	0.1789	0.3143	0.0879	0.1034	0.1736	0.3041
$\mu_2(-4)$	Relative bias	-0.0557	-0.0778	-0.1464	-0.2681	-0.0519	-0.0655	-0.1363	-0.2440	-0.0570	-0.0623	-0.1163	-0.2296
	RMSE	0.1052	0.1136	0.1801	0.3098	0.1152	0.1253	0.1980	0.3342	0.1077	0.1171	0.1831	0.2997
$\sigma_1^2(0.3)$	Relative bias	-0.0014	-0.0016	-0.0017	-0.0027	0.0015	0.0002	0.0074	0.0076	0.0153	0.0147	0.0285	0.0484
	RMSE	0.0158	0.0154	0.0162	0.0191	0.0300	0.0294	0.0337	0.0359	0.0447	0.0479	0.0697	0.0855
$\sigma_2^2(0.6)$	Relative bias	0.0004	-0.0015	-0.0007	-0.0113	0.0153	0.0197	0.0176	0.0089	0.0477	0.0733	0.0812	0.0603
	RMSE	0.1235	0.1255	0.1274	0.2191	0.1244	0.1268	0.1320	0.1501	0.1539	0.1568	0.1671	0.1781
$\nu_1(3)$	Relative bias	-	-	-	-	0.0910	0.0745	0.1577	0.1703	0.2831	0.0756	0.2347	0.4631
	RMSE	-	-	-	-	0.5247	0.4832	0.6451	0.6887	0.3722	0.3700	0.4086	0.5914
$\nu_2(4)$	Relative bias	-	-	-	-	0.5620	0.3662	0.7687	0.9416	0.5965	0.6485	0.7037	0.8196
	RMSE	-	-	-	-	0.3457	0.9933	0.3635	0.3350	0.4161	0.5194	0.6697	0.7903
$p_1(0.8)$	Relative bias	0.0133	0.0142	0.0288	0.0553	0.0098	0.0119	0.0269	0.0482	0.0107	0.0131	0.0244	0.0464
	RMSE	0.0182	0.0205	0.0342	0.0611	0.0183	0.0214	0.0359	0.0631	0.0176	0.0204	0.0345	0.0604

square error (RMSE) for all censoring levels. For a given parameter θ , they are defined by

$$\text{Bias}(\hat{\theta}) = \frac{1}{500} \sum_{j=1}^{500} (\hat{\theta}^{(j)} - \theta) \quad \text{and} \quad \text{RMSE}(\hat{\theta}) = \sqrt{\frac{1}{500} \sum_{j=1}^{500} (\hat{\theta}^{(j)} - \theta)^2},$$

where $\hat{\theta}^{(j)}$ is the estimate of θ for the j -th sample. The RMSE of the estimators for all censoring proportions. The RMSE and relative bias (hereafter, the analysis is in absolute value) for β_0 , μ_1 , μ_2 , σ_2^2 and p_1 increase as the proportion of censoring increases for FM-N-CR, FM-T-CR and FM-SL-CR models and the other parameters increase too but in one and/or two models. That is an expected result since less information is available if the proportion of censoring is high. The relative bias of β_1 shows opposite behavior being higher if 8% censoring is observed for the FM-N-CR and FM-SL-CR models and, for β_2 , σ_2^2 the three models do not have monotonic behavior. The same behavior is observed for ν_1 . However, the relative bias is very close of zero for all parameters and the RMSE smoothly increases, except for β_0 , ν_1 and ν_2 . Consequently, the standard errors of the estimators are not strongly affected by the proportion of censoring, except for β_0 , ν_1 and ν_2 where a higher increase in the RMSE is observed if a high proportion of censoring is considered.

3.4.3 The effect of different sample sizes (simulation study 3)

Here, the experimental strategy is to generate artificial samples from the FM-N-CR, FM-T-CR and FM-SL-CR models, assuming $\nu_1 = 3$, $\nu_2 = 4$. Tables 3.4 and 3.5 (see also Figures 3.2 and 3.3) present the relative bias and RMSE where we have chosen various settings of censoring proportions (8% and 35%), with samples three sizes (250, 1000 and 5000). When n increases, the RMSE decreases independently of the censoring pattern (a similar pattern was observed in

Table 3.4: *Simulation study 3: Bias and root relative mean squared errors of ECME estimates*

Niv.	Censored	Parameter	FM-NCR		FM-T-CR		FM-SL-CR	
			Relative bias	RMSE	Relative bias	RMSE	Relative bias	RMSE
8%	$\beta_0(1)$	250	0.0607	0.1750	0.0468	0.2035	0.0456	0.1905
		1000	0.0626	0.1045	0.0446	0.1112	0.0497	0.1095
		5000	0.0596	0.0735	0.0533	0.0803	0.0540	0.0769
	$\beta_1(-1)$	250	-0.0102	0.1499	0.0182	0.1697	0.0052	0.1685
		1000	-0.0017	0.0659	0.0018	0.0834	-0.0074	0.0782
		5000	0.0015	0.0311	0.0019	0.0390	-0.0012	0.0363
	$\beta_2(4)$	250	-0.0013	0.1420	-0.0084	0.1781	0.0013	0.1626
		1000	-0.0007	0.0689	0.0032	0.0898	0.0037	0.0820
		5000	-0.0012	0.0318	-0.0037	0.0381	-0.0015	0.0342
	$\mu_1(1)$	250	-0.0548	0.1398	-0.0505	0.1430	-0.0654	0.1480
		1000	-0.0557	0.0898	-0.0478	0.0899	-0.0478	0.0938
		5000	-0.0564	0.0696	-0.0520	0.0738	-0.0524	0.0715
	$\mu_2(-4)$	250	-0.0511	0.1727	-0.0533	0.1998	-0.0422	0.1976
		1000	-0.0608	0.1009	-0.0462	0.1100	-0.0444	0.1067
		5000	-0.0582	0.0729	-0.0522	0.0809	-0.0553	0.0767
	$\sigma_1^2(0.3)$	250	-0.0039	0.0312	0.0094	0.0607	0.0196	0.0750
		1000	-0.0003	0.0140	0.0033	0.0323	0.0082	0.0451
		5000	-9.e-05	0.0070	0.0022	0.0139	0.0064	0.0202
	$\sigma_2^2(0.6)$	250	-0.0155	0.1230	0.0176	0.2422	0.0447	0.2124
		1000	-0.0052	0.0655	0.0189	0.1247	0.0534	0.1570
		5000	-0.0005	0.0302	-0.0015	0.0551	0.0249	0.1260
	$\nu_1(3)$	250	-	-	0.5182	0.1078	1.6629	1.6085
		1000	-	-	0.1382	0.5913	0.6535	0.9549
		5000	-	-	0.0280	0.2291	0.1850	0.5879
	$\nu_2(4)$	250	-	-	0.8772	1.1203	1.6291	1.4748
		1000	-	-	0.7986	0.1269	1.0783	1.0943
		5000	-	-	0.2275	0.0485	0.9793	0.8182
	$p_1(0.8)$	250	0.0114	0.0277	0.0101	0.0284	0.0094	0.0293
		1000	0.0111	0.0178	0.0095	0.0178	0.0098	0.0185
		5000	0.0113	0.0139	0.0104	0.0148	0.0106	0.0143

simulation study 2) of all parameters for the three models with the exception of β_0 , μ_1 , μ_2 and p_1 in the three models. The same monotonic behavior is observed for the relative bias of σ_1^2 , so their ML estimates tend to be unbiased and consistent. For the sample of size $n = 250$ and FM-N-CR, the biases for the estimators of β_0 , β_1 , μ_1 , σ_1^2 and p_1 tend to be more influenced with the proportion of censoring, experiencing a higher increase in their values than that observed for the other parameters. Similar, behavior is observed for β_0 the FM-T-CR model. If samples of size $n = 5000$ are under consideration similar behavior occurs for the estimators of β_0 , σ_2^2 and p_1 for the FM-N-CR, FM-T-CR and FM-SL-CR models. 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 in the FM-T-CR and FM-SL-CR model do provide good asymptotic

properties, as depicted in Figures 3.2 and 3.3.

Figure 3.2: Simulation study 3: Average bias (first row) and RMSE (second row) of parameter estimates in the FM-SMN-CR model with 8% censoring level.

Figure 3.3: Simulation study 3: Average bias (first row) and RMSE (second row) of parameter estimates in the FM-SMN-CR model with 35% censoring level.

3.5 Application

In order to study the performance of our proposed model and algorithm, here analyze a real dataset. We consider the wage rate dataset described in Mroz (1987), where a measure of the wage of 753 married white women, with ages between 30 and 60 years old in 1975, is evaluated. Of 753 women considered in this study, 428 worked at some point during that year, while the remaining did

Table 3.5: *Simulation study 3: Bias and root relative mean squared errors of ECME estimates*

		FM-NCR			FM-T-CR		FM-SL-CR	
Niv.	Censored Parameter		Relative bias	RMSE	Relative bias	RMSE	Relative bias	RMSE
35%	$\beta_0(1)$	250	0.2540	0.3380	0.2260	0.3735	0.2372	0.3637
		1000	0.2684	0.3145	0.2433	0.3293	0.2449	0.3183
		5000	0.2687	0.3030	0.2513	0.3209	0.2404	0.3081
	$\beta_1(-1)$	250	-0.0066	0.1620	0.0215	0.2048	0.0126	0.2028
		1000	-0.0088	0.0831	-0.0081	0.1009	-0.0016	0.0937
		5000	0.0001	0.0401	-0.0033	0.0460	0.0009	0.0399
	$\beta_2(4)$	250	0.0137	0.1578	-0.0037	0.2156	-0.0079	0.1772
		1000	0.0076	0.0819	0.0041	0.1012	-0.0071	0.0986
		5000	0.0020	0.0410	0.0004	0.0452	0.0014	0.0428
	$\mu_1(1)$	250	-0.0741	0.0198	-0.0304	0.0318	-0.2377	0.3253
		1000	-0.0785	0.0073	-0.0427	0.0134	-0.2360	0.3057
		5000	-0.0728	0.0007	-0.0520	0.0102	-0.2388	0.2020
	$\mu_2(-4)$	250	-0.0776	0.0421	-0.0387	0.0870	-0.2288	0.3423
		1000	-0.0680	0.0088	-0.0391	0.0267	-0.2317	0.3096
		5000	-0.0683	0.0978	-0.0515	0.0253	-0.2309	0.1975
	$\sigma_1^2(0.3)$	250	-0.0065	0.0372	0.0145	0.0782	0.0505	0.0997
		1000	-0.0017	0.0181	0.0087	0.0377	0.0446	0.0836
		5000	-0.0002	0.0084	0.0037	0.0169	0.0303	0.0552
	$\sigma_2^2(0.6)$	250	-0.0028	0.1584	-0.0012	0.2886	0.0988	0.2707
		1000	-0.0068	0.0776	0.0255	0.1366	0.0589	0.1676
		5000	0.0037	0.0332	-0.0013	0.0634	0.0692	0.1280
	$\nu_1(3)$	250	-	-	0.0692	0.0902	0.9498	1.2303
		1000	-	-	0.1447	0.0554	0.8916	1.0845
		5000	-	-	-0.0048	0.0230	0.5804	0.9963
	$\nu_2(4)$	250	-	-	0.0321	0.0882	1.0806	1.1601
		1000	-	-	0.0659	0.0787	0.2628	1.0412
		5000	-	-	0.0680	0.0459	0.0832	0.9049
	$p_1(0.8)$	250	0.0038	0.0638	0.0462	0.0664	0.0473	0.0644
		1000	0.0555	0.0613	0.0470	0.0628	0.0471	0.0610
		5000	0.0544	0.0600	0.0488	0.0624	0.0475	0.0602

not work for pay. Thus, we can consider these last ones as left censored with $\kappa_i = 0$, $i = 1, \dots, n..$. The following variables were considered:

- y_i : *Annual working hours*, defined as hours worked in 1975. These observations are considered left censored at zero;
- x_{i1} : wife's age;
- x_{i2} : educational attainment (in years);
- x_{i3} : father's years of schooling;
- x_{i4} : current labor market experience.

The vectors of explanatory variable values are given by $\mathbf{x}_i^\top = (x_{i1}, x_{i2}, x_{i3}, x_{i4})$ for $i = 1, 2, \dots, 753$. This dataset was analyzed by Arellano-Valle *et al.* (2012) and Massuia *et al.* (2015) using a censored regression model with Student-t responses, and by Garay *et al.* (2015) using a censored regression model with scale mixtures of normal distributions. Here, we revisit this dataset to evaluate the performance of the proposed methods considering the FM-N-CR, FM-T-CR and FM-SL-CR models. Thus, we fitted the censored model

$$y_i|Z_i = j \sim SMN(\beta_0 + \mu_j + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}, \sigma_j^2), \quad i = 1, \dots, 753; \quad j = 1, 2,$$

and also its corresponding normal version. For comparison, we fit the FM-SMN-CR model with

Table 3.6: *ML estimation results for fitting different mixture models to the wage rate data. The model with the smallest BIC value for each family is shown in bold. The best chosen model is indicated by "*"*

Model	g	ℓ	m	BIC	Iterations	r
FM-N-CR	1	-493.6209	7	1020.362	7	0.7436
	2	-459.5368	10	991.938	28	0.9309
	3	-441.8554	13	996.320	93	0.9291
	4	-422.8199	16	997.993	52	0.9646
	5	-410.9691	19	1014.036	86	1.0048
	6	-404.2173	22	1040.277	78	0.9730
FM-T-CR	1	-491.3296	8	1022.404	8	0.6244
	2	-443.1301	11	965.734	19	0.6521
	3	-428.3616	14	985.151	29	0.8905
	4	-423.5796	17	1005.353	79	0.9767
	5	-412.3071	20	1021.394	83	1.0047
	6	-406.6891	23	1049.085	64	0.9801
FM-SL-CR	1	-489.3029	8	1018.350	12	0.7735
	2*	-442.2556	11	964.065	21	0.8445
	3	-434.5875	14	985.858	40	0.9221
	4	-422.6101	17	1004.242	62	0.4465
	5	-412.3039	20	1025.148	41	0.9745
	6	-407.0409	23	1052.296	51	0.9641

$g = 1 - 6$. Table 3.6 contains the log-likelihood, the numbers of parameters and one information-based criterion. The last two columns of this table present the required number of EM iterations and the associated rate of convergence, r , which is assessed in practice as

$$r = \lim_{t \rightarrow \infty} \frac{\|\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}\|}{\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t-1)}\|}.$$

A relative tolerance of 10^{-6} for the estimates of all parameters in the model was used as the convergence criterion. The rate of convergence depends on the fraction of missing information and

a greater value of r implies slower convergence; see Meng (1994). We ran the ECME algorithms as described in Section 3.2. We considered $\nu_1 = \nu_2 = 3$ as initial values for degrees of freedom. When g was large, it perhaps stopped too early. In Table 3.6 note that a small value of BIC is associated with a better fitted model. It can be clearly seen that the two-component FM-SL-CR model the best fit, followed by the two-component FM-T-CR model and FM-N-CR model. In this application, we also note that the estimation procedure for fitting the FM-SL-CR model converge is slow for $g \geq 3$.

Table 3.7 compares the ML estimates along with the standard errors for the best fitted FM-SL-CR model with the corresponding values for the other two competing two-component mixture model. It also shows that the estimates of $\beta_0, \dots, \beta_4$ for the FM-N-CR, FM-T-CR and FM-SL-CR models are close. However, the standard errors (SE) of β are smaller than those of the FM-N-CR and FM-T-CR models, indicating that the FM-SL-CR model produces more precise estimates. The estimates of the scale parameters are not comparable since they are on a different scale. Also, notice that the small value of the estimate of ν for the FM-T-CR and FM-SL-CR models indicates a lack of adequacy of the normal assumption. Table 3.6 compares the fit of the two mixture models using the model selection criteria discussed in Subsection 4.19, as well as compares the fit of the SL-CR, T-CR and N-CR models, that is, the models with slash, Student-t and normal errors, respectively. Note that, as expected, the FM-SL-CR model performs significantly better than the FM-N-CR and FM-T-CR models.

Table 3.7: *Wage rate data. results of the parameter estimation via the EM-type algorithm.*

Parameter	FM-N-CR model		FM-T-CR model		FM-SL-CR model	
	Estimates	SE	Estimates	SE	Estimates	SE
β_0	1.4858	0.3397	1.4481	0.2314	1.4482	0.1725
β_1	0.0002	0.0029	0.0010	0.0012	0.0009	0.0005
β_2	-0.0187	0.0105	-0.0196	0.0055	-0.0196	0.0021
β_3	-0.0064	0.0091	-0.0100	0.0042	-0.0091	0.0016
β_4	0.0107	0.0030	0.0143	0.0014	0.0132	0.0005
μ_1	0.2765	0.0389	0.3666	0.1048	0.3638	0.1135
μ_2	-0.9235	0.2561	-0.8185	0.3569	-0.8203	0.1314
σ_1^2	0.3727	0.0219	0.1541	0.0387	0.1102	0.0138
σ_2^2	0.0363	0.0091	0.0463	0.0106	0.0317	0.0066
p	0.7696	0.0352	0.6907	0.0933	0.6928	0.0288
ν	-	-	4.7351	-	1.6596	-

3.6 Conclusions

In this paper, a novel approach to censored linear regression analysis is reported based on the use of finite mixtures of SMN components for the random errors. This approach includes some previously proposed solutions, namely, the classic Tobit linear models in which the error terms are assumed to follow a Gaussian or a Student-t distribution (Arellano-Valle *et al.*, 2012; Garay *et al.*, 2015), or a finite mixture of Gaussian components (Caudill, 2012; Karlsson and Laitila, 2014). In a sense, each of these models is broadened by the proposed approach because our approach provides better estimates of the regression coefficients when the distribution of the error terms is characterized by the presence of multimodality, outlying observations and also skewness, depending on the structure of the mixture components. Furthermore, the experimental results and the analysis of a real datasets provide support for the usefulness and effectiveness of our proposal. A simple and efficient EM-type algorithm was developed, which has closed-form expressions at the E-step and relies on formulas for the mean and variance of the truncated SMN distributions. The proposed EM algorithm was implemented as part of the R package `CensMixReg` and is available for download at the CRAN repository.

Chapter 4

Finite mixture of regression models based on multivariate scale mixtures of skew-normal distributions

The traditional estimation of mixture regression models is based on the assumption of normality (or symmetry) of components and thus is sensitive to outliers, heavy-tailed errors and/or asymmetric errors. In this work we present a proposal to deal with these issues simultaneously by considering finite mixture of regression models with multivariate scale mixtures of skew-normal distributions. This approach allows us to model data with great flexibility, accommodating simultaneously skewness and heavy tails. The main virtue of considering the mixture of regression models under the class of multivariate scale mixtures of skew-normal distributions is that they have a nice hierarchical representation, which allows easy implementation of inferences. We develop a simple EM-type algorithm to perform maximum likelihood inference of the parameters of the proposed model. The observed information matrix is derived analytically to account for standard errors. In order to examine the robustness of this flexible model against outlying observations, some simulation studies are also presented. Finally, a real dataset is analyzed, illustrating the usefulness of the proposed method. The proposed EM-type algorithm and methods are implemented in the R package `FMsmnReg`.

4.1 Introduction

Finite mixture models have been used for many decades, because they provide a flexible tool for modeling data that arise from heterogeneous populations, but the problem is that each individual

can belong to one of several underlying sub-populations, resulting in heterogeneity of the observed outcomes. Finite mixture models have been extensively discussed in the literature and applied in various areas, such as biology, medicine, physics, economics and marketing, among others. The importance of mixture distributions can be noted from the large number of books on the subject. Recently McNicholas (2017) showed an approach to clustering and classification, Ferraro *et al.* (2017, pag. 181-186) proposed a fully data-driven constrained method for estimation of mixtures of linear regression models, Pyne *et al.* (2016, pag. 57-74) showed the application of mixture models to large datasets, Tatarinova and Schumitzky (2015) provided a broad introduction to the important subject of nonlinear mixture models from a Bayesian perspective, and the special issue of *Structural Equation Modeling: A Multidisciplinary Journal* focused on advances in finite mixture modeling (Lubke and Grimm, 2017).

The main feature of the finite mixture of distributions is the ability to capture more specific properties of the observed data, such as multimodality, heavy tails and/or skewness. In the univariate context for example, Zeller *et al.* (2016) proposed the finite mixture of regressions under the class of scale mixture skew-normal distributions in the univariate context Devijver (2015) considered a finite mixture of Gaussian regression models for high-dimensional data, where the number of covariates may be much larger than the sample size, Ciarleglio and Ogden (2016) proposed a finite mixture regression model incorporating functional predictors by taking a wavelet-based approach in which both the functional predictors and the component-specific coefficient functions are represented in terms of an appropriate wavelet basis; and Song *et al.* (2014) proposed a mixture of linear regression models where the error terms are assumed to follow a Laplace distribution. In the multivariate context, Jedidi (1996) proposed an EM algorithm for estimating finite mixture of multivariate regressions and simultaneous equation models with multiple endogenous variables, whereas recently Lim *et al.* (2017) proposed a multivariate t mixture regression model with missing information. The number of components (*i.e.*, the number of densities in the mixture of random errors) necessary to capture the asymmetric and/or heavy tails of each subpopulation can occur. Thus in this paper we propose a robust mixture regression model based on the multivariate scale mixture Skew-normal distributions (SMSN-MMRM) by extending the mixture of univariate regressions proposed by Zeller *et al.* (2016). More specifically, our objectives are (i) to propose a multivariate regression model based on finite mixture of multivariate SMSN, (ii) to implement and evaluate the proposed model computationally; and (iii) to apply these results to analyze of a real-life dataset.

The remainder of this article is organized as follows: In Section 2 we discuss some preliminary

findings related to the multivariate SMSN distributions and some of their important properties. In Section 3 we propose the robust SMSN-MMRM model, including the EM algorithm for maximum likelihood (ML) estimation. Sections 4 and 5 present numerical examples using real and simulated data to show the performance of the proposed method. Finally, some concluding remarks are presented in Section 6.

4.2 The multivariate scale mixtures of skew-normal distributions

4.2.1 Preliminaries

In this section, we review the multivariate SMSN class of distributions and some properties which are shared by all the elements of this class. For this we refer to Lachos *et al.* (2010a) for proofs and additional properties. We begin our exposition by clarifying the concepts involved in the SMSN definition. First, the SMSN distribution was proposed by Branco and Dey (2001), using scale mixtures generated by some skew-normal (SN) families. In fact, as there are several proposed distributions that extend the normal one by the introduction of skewness, there is a need to specify a convenient version of the skew-normal. Second, extensions and unifications of the skew-normal are carefully surveyed in works like Azzalini (2005) and Arellano-Valle and Azzalini (2006). For our purposes, we say that a $p \times 1$ random vector $\mathbf{Y}$ follows a skew-normal distribution with $p \times 1$ location vector $\boldsymbol{\mu}$, $p \times p$ positive definite dispersion matrix $\boldsymbol{\Sigma}$ and $p \times 1$ skewness parameter vector $\boldsymbol{\lambda}$, and we write $\mathbf{Y} \sim \text{SN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$, if its density is given by

$$\phi_{\text{SN}_p}(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}) = 2\phi_p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\Phi(\boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{-1/2}(\mathbf{y} - \boldsymbol{\mu})), \quad (4.1)$$

where $\boldsymbol{\lambda}^\top$ denotes the transpose of $\boldsymbol{\lambda}$, $\boldsymbol{\Sigma}^{-1/2}$ is the root of $\boldsymbol{\Sigma}$, that is, the symmetric matrix that satisfies $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{\Sigma}^{-1}$, $\phi_p(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ stands for the density of the p -variate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ say, and $\Phi(\cdot)$ represents the distribution function of the standard univariate normal distribution. We drop some indexes when there is no possibility of confusion.

The expression (4.1) of the multivariate skew-normal as mixture components allows us to develop an EM-type algorithm for maximum likelihood estimation with full closed form expressions for the equations at the M step, still providing estimates with a good performance when modeling heterogeneous data with a strong pattern of skewness and outliers. Indeed, it has been used as a powerful tool to extend some traditional normal based models, see Cabral *et al.* (2012), Lachos

et al. (2010b) and Lachos *et al.* (2010a), for example.

Finally, having made these remarks, let us define the multivariate SMSN family. In what follows $\mathbf{0}$ denotes the zero vector of appropriate dimension.

Definition 2. We say that the distribution of a random vector $\mathbf{Y}$ belongs to the skew-normal independent (SMSN) family when $\mathbf{Y} = \boldsymbol{\mu} + U^{-1/2}\mathbf{Z}$, where $\boldsymbol{\mu}$ is a p -dimensional (location) vector, $\mathbf{Z} \sim \text{SN}_p(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ and U is a positive random variable, independent of $\mathbf{Z}$, with distribution function $H(\cdot|\boldsymbol{\nu})$ and density $h(\cdot|\boldsymbol{\nu})$.

Here $\boldsymbol{\nu}$ is a scalar or vector parameter indexing the distribution function H . The notation $\mathbf{Y} \sim \text{SMSN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, H)$ will be used. It is straightforward to show that $\mathbf{Y}|U = u \sim \text{SN}_p(\boldsymbol{\mu}, u^{-1}\boldsymbol{\Sigma}, \boldsymbol{\lambda})$, implying that the SMSN distributions are scale mixtures of the skew-normal distributions. Thus, the marginal density of $\mathbf{Y}$ is

$$\phi_{\text{SMSN}_p}(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = 2 \int_0^\infty \phi_p(\mathbf{y}|\boldsymbol{\mu}, u^{-1}\boldsymbol{\Sigma}) \Phi(u^{1/2}\boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{-1/2}(\mathbf{y} - \boldsymbol{\mu})) dH(u|\boldsymbol{\nu}). \quad (4.2)$$

The SMSN class includes asymmetric distributions such as the skew-normal, skew- t , skew-slash and skew-contaminated normal (and their symmetric counterparts). The last three have heavier tails than the skew-normal (and normal) one and can be used for robust inference. Although we can obtain a rich class of distributions by considering a wide range of choices for the mixing distribution $H(\cdot|\boldsymbol{\nu})$, for many of them the resulting distribution does not have a computationally attractive form, so we restrict ourselves to the four above cited cases. These computational restrictions do not affect our theoretical results, and they are valid for the entire SMSN family.

For computational purposes, a useful parameterization is given by

$$\boldsymbol{\Delta} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}, \quad \boldsymbol{\Gamma} = \boldsymbol{\Sigma}^{1/2}(\mathbf{I} - \boldsymbol{\delta}\boldsymbol{\delta}^\top)\boldsymbol{\Sigma}^{1/2} = \boldsymbol{\Sigma} - \boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \quad (4.3)$$

where $\mathbf{I}$ denotes the identity matrix of appropriate dimension. We recover $\boldsymbol{\lambda}$ and $\boldsymbol{\Sigma}$ using the expressions

$$\boldsymbol{\lambda} = \frac{(\boldsymbol{\Gamma} + \boldsymbol{\Delta}\boldsymbol{\Delta}^\top)^{-1/2}\boldsymbol{\Delta}}{[1 - \boldsymbol{\Delta}^\top(\boldsymbol{\Gamma} + \boldsymbol{\Delta}\boldsymbol{\Delta}^\top)^{-1}\boldsymbol{\Delta}]^{1/2}}, \quad \boldsymbol{\Sigma} = \boldsymbol{\Gamma} + \boldsymbol{\Delta}\boldsymbol{\Delta}^\top. \quad (4.4)$$

From Definition 1, it is simple to obtain the mean vector and the covariance matrix of $\mathbf{Y}$, given respectively by

$$E\{\mathbf{Y}\} = \boldsymbol{\mu} - b\boldsymbol{\Delta}, \quad \text{and} \quad \text{Var}\{\mathbf{Y}\} = K_2\boldsymbol{\Sigma} - b^2\boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \quad (4.5)$$

where $b = -\sqrt{\frac{2}{\pi}}K_1$, where $K_m = E\{U^{-m/2}\}$ and are given in Table 1 for different SMSN distri-

butions.

The following propositions are crucial to our method. The proofs can be found in Lachos *et al.* (2010a). Proposition 4 together with parameterization (4.3) facilitates the implementation of the EM-type algorithm, as will be seen later. Also, Proposition 5 is essential to compute some necessary expectations in our algorithm. Here, $\text{TN}(a, b^2, B)$ denotes the truncated normal distribution on the Borel set B , with the indicated parameters before truncation. Also, $\stackrel{d}{=}$ means “has the same distribution as”.

Proposition 4. $\mathbf{Y} \sim \text{SMSN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, H)$ admits the following hierarchical representation:

$$\mathbf{Y}|U = u, T = t \sim \text{N}_p(\boldsymbol{\mu} + \boldsymbol{\Delta}t, u^{-1}\boldsymbol{\Gamma}), \quad T|U = u \sim \text{TN}(0, u^{-1}, (0, \infty)), \quad U \sim H(u|\boldsymbol{\nu}). \quad (4.6)$$

Proposition 5. Let $\mathbf{Y} \sim \text{SMSN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, H)$, $\mathbf{Y}_0 \sim \text{SMN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, H)$. Let U be the common scale factor and define $U_{\mathbf{y}} \stackrel{d}{=} (U|\mathbf{Y}_0 = \mathbf{y})$. Then, for any measurable function $g : \mathbb{R} \rightarrow \mathbb{R}$,

$$E\{g(U)|\mathbf{Y} = \mathbf{y}\} = 2\gamma(\mathbf{y})E\{g(U_{\mathbf{y}})\Phi(U_{\mathbf{y}}^{1/2}\boldsymbol{\lambda}^\top\boldsymbol{\Sigma}^{-1/2}(\mathbf{y} - \boldsymbol{\mu}))\},$$

where $\gamma(\mathbf{y}) = \phi_{\text{SMN}_p}(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, H)/\phi_{\text{SMSN}_p}(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, H)$.

Thus, computation of the integral $E\{g(U)|\mathbf{Y} = \mathbf{y}\}$ can be done using an often simpler form that involves integration with respect to the distribution of $U_{\mathbf{y}}$.

Table 4.1: $k_m = E[U^{-m/2}]$ for different SMSN models.

Model	k_m
SN	1
ST	$\left(\frac{\nu}{2}\right)^{m/2} \frac{\Gamma\left(\frac{\nu-m}{2}\right)}{\Gamma(\nu/2)}$
SSL	$\frac{2\nu}{\nu-m/2}$
SCN	$\frac{\nu_1}{\nu_2^{m/2}} + 1 - \nu_1$

Next, we present the distributions in the SMSN family under consideration in this work.

4.2.2 Multivariate skew-normal distribution

This is the case when $U = 1$ in Definition 2.

4.2.3 Multivariate skew- t distribution

The p -variate skew- t distribution with ν degrees of freedom, $ST_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu)$ say, is derived from definition 2 by taking U to be distributed as $\text{Gamma}(\nu/2, \nu/2)$, $\nu > 0$ – where $\text{Gamma}(a, b)$ denotes the gamma distribution with mean a/b . If a random vector $\mathbf{Y}$ has this distribution, its density is given by

$$\phi_{ST_p}(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu) = 2t_p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)T\left(\sqrt{\frac{\nu+p}{\nu + d_{\boldsymbol{\Sigma}}(\mathbf{y}, \boldsymbol{\mu})}}\boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{-1/2}(\mathbf{y} - \boldsymbol{\mu})|\nu + p\right),$$

where $t_p(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ and $T(\cdot|\nu + p)$ denote, respectively, the density of the p -variate Student- t distribution with mean vector $\boldsymbol{\mu}$, dispersion matrix $\boldsymbol{\Sigma}$ and ν degrees of freedom – namely $t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ – and the distribution function of the standard univariate Student- t distribution with $\nu + p$ degrees of freedom and $d_{\boldsymbol{\Sigma}}(\mathbf{y}, \boldsymbol{\mu}) = (\mathbf{y} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})$ is the squared Mahalanobis distance between $\mathbf{y}$ and $\boldsymbol{\mu}$ with respect to $\boldsymbol{\Sigma}$. A particular case of the skew- t distribution is the skew-Cauchy distribution, when $\nu = 1$. When $\boldsymbol{\lambda} = \mathbf{0}$ we recover the Student- t distribution. Also, when $\nu \uparrow \infty$, we get the skew-normal distribution as the limiting case.

4.2.4 Multivariate skew-slash distribution

The skew-slash distribution, denoted by $SSL_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu)$, arises when U has density given by $h(u|\nu) = \nu u^{\nu-1}$, $u \in (0, 1)$, $\nu > 0$ – that is, a beta distribution with parameters ν and 1. Its density is given by

$$\phi_{SSL_p}(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu) = 2\nu \int_0^1 u^{\nu-1} \phi_p(\mathbf{y}|\boldsymbol{\mu}, u^{-1}\boldsymbol{\Sigma}) \Phi(u^{1/2}\boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{-1/2}(\mathbf{y} - \boldsymbol{\mu})) du. \quad (4.7)$$

and it can be evaluated using the R function `integrate` (R Development Core Team, 2011), for example. The skew-slash distribution reduces to the skew-normal distribution when $\nu \uparrow \infty$ and reduces to the slash distribution when $\boldsymbol{\lambda} = \mathbf{0}$, with notation $\mathbf{Y} \sim SL_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$. See Wang and Genton (2006) for properties and applications of the SSL distribution.

4.2.5 Multivariate skew-contaminated normal distribution

The multivariate skew-contaminated normal distribution arises when the mixing scale factor U is a discrete random variable taking state ν_2 with probability ν_1 and state 1 with probability $1 - \nu_1$. We will write $\mathbf{Y} \sim SCN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \boldsymbol{\nu})$, where $\boldsymbol{\nu} = (\nu_1, \nu_2)^\top$. The associated density is given

by

$$\begin{aligned}\phi_{SCN_p}(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \boldsymbol{\nu}) &= 2 \left\{ \nu_1 \phi_p(\mathbf{y}|\boldsymbol{\mu}, \nu_2^{-1} \boldsymbol{\Sigma}) \Phi(\nu_2^{1/2} \boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{-1/2} (\mathbf{y} - \boldsymbol{\mu})) \right. \\ &\quad \left. + (1 - \nu_1) \phi_p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Phi(\boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{-1/2} (\mathbf{y} - \boldsymbol{\mu})) \right\}.\end{aligned}$$

Parameters ν_1 and ν_2 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 $\nu_2 = 1$ and reduces to the contaminated normal distribution when $\boldsymbol{\lambda} = \mathbf{0}$, with notation $\mathbf{Y} \sim \text{CN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\nu})$.

4.3 The linear regression model for multivariate SMSN responses

The linear regression model based on multivariate SMSN distributions - hereafter SMSN-MLR model - is defined as

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \boldsymbol{\epsilon}_i, \quad i = 1, \dots, n, \quad (4.8)$$

where $\mathbf{y}_i$ is the vector of responses of dimension $p \times 1$, $\mathbf{X}_i$ is a known $p \times q$ covariate matrix, $\boldsymbol{\beta}$ is a $q \times 1$ vector of unknown regression coefficients and the random errors $\boldsymbol{\epsilon}_i \stackrel{iid}{\sim} \text{SMSN}_p(-\sqrt{\frac{2}{\pi}} K_1 \boldsymbol{\Delta}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}; H)$, with $K_r = E\{U^{-r/2}\}$, $r = 1, 2, \dots$, $\boldsymbol{\Delta} = \boldsymbol{\Sigma}^{1/2} \boldsymbol{\delta}$ and $\boldsymbol{\delta} = \boldsymbol{\lambda} / (1 + \boldsymbol{\lambda}^\top \boldsymbol{\lambda})^{1/2}$, which corresponds to the regression model where the error distribution has mean zero and hence the regression parameters are all comparable. From (4.5), we have that

$$E[\mathbf{Y}_i] = \mathbf{X}_i \boldsymbol{\beta}, \quad \text{Var}[\mathbf{Y}_i] = K_2 \boldsymbol{\Sigma} - b^2 \boldsymbol{\Delta} \boldsymbol{\Delta}^\top,$$

where $b = -\sqrt{\frac{2}{\pi}} K_1$ and $\mathbf{Y}_i \sim \text{SMSN}_p(\mathbf{X}_i \boldsymbol{\beta} + b \boldsymbol{\Delta}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}; H)$, for $i = 1, \dots, n$.

4.3.1 Parameter estimation via the EM algorithm

In this subsection, we present an EM-type algorithm (Dempster *et al.*, 1977) for estimating of the parameters of the SMSN-MLR model. Without confusion, in what follows we use $\text{SMSN}_p(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \boldsymbol{\nu})$ to denote a SMSN density with the specified parameters, mixing scale distribution $H(\cdot|\boldsymbol{\nu})$ and density $h(\cdot|\boldsymbol{\nu})$.

To develop the algorithm, consider the “complete data” $(\mathbf{y}_i^\top, t_i, u_i)^\top$, $i = 1, \dots, n$, that is, a random sample from the random vector $(\mathbf{Y}^\top, T, U)^\top$ of Proposition 4. Let $\mathbf{y} = (\mathbf{y}_1^\top, \dots, \mathbf{y}_n^\top)^\top$, $\mathbf{t} = (t_1, \dots, t_n)^\top$ and $\mathbf{u} = (u_1, \dots, u_n)^\top$. The complete-data log-likelihood function of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\lambda}, \boldsymbol{\nu})^\top$,

where $\boldsymbol{\alpha}$ denotes the vector with the elements of the upper triangular matrix of $\boldsymbol{\Sigma}$, is given by

$$\ell_c(\boldsymbol{\theta}|\mathbf{y}, \mathbf{t}, \mathbf{u}) = cte - \frac{n}{2} \log |\boldsymbol{\Gamma}| + \sum_{i=1}^n \log(h(u_i|\boldsymbol{\nu})) - \frac{1}{2} \sum_{i=1}^n u_i (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta} - \boldsymbol{\Delta} t_i) \boldsymbol{\Gamma}^{-1} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta} - \boldsymbol{\Delta} t_i)^\top,$$

where cte is a constant which is independent of the parameter vector $\boldsymbol{\theta}$.

In what follows the superscript (k) indicates the estimate of the related parameter at stage k of the algorithm. The E-step consists of taking the conditional expectation

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = \mathbb{E}_{\boldsymbol{\theta}^{(k)}} \{\ell_c(\boldsymbol{\theta}|\mathbf{y}, \mathbf{t}, \mathbf{u})|\mathbf{y}\},$$

where $\mathbb{E}_{\boldsymbol{\theta}^{(k)}}$ means that the expectation is being effected using $\boldsymbol{\theta}^{(k)}$ for $\boldsymbol{\theta}$. The following conditional expectations are necessary to obtain $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$

$$\begin{aligned} U(\boldsymbol{\theta}^{(k)}, \mathbf{y}_i) &= \mathbb{E}_{\boldsymbol{\theta}^{(k)}} \{U_i|\mathbf{y}_i\}, \quad \xi(\boldsymbol{\theta}^{(k)}, \mathbf{y}_i) = \mathbb{E}_{\boldsymbol{\theta}^{(k)}} \{U_i T_i|\mathbf{y}_i\} \\ \text{and } \omega(\boldsymbol{\theta}^{(k)}, \mathbf{y}_i) &= \mathbb{E}_{\boldsymbol{\theta}^{(k)}} \{U_i T_i^2|\mathbf{y}_i\}. \end{aligned} \tag{4.9}$$

We do not consider the computation of $\mathbb{E}_{\boldsymbol{\theta}^{(k)}} \{\log(h(u_i|\boldsymbol{\nu}))|\mathbf{y}_i\}$ because the extension of the EM algorithm (ECME algorithm) used here does not employ it. $U(\boldsymbol{\theta}^{(k)}, \mathbf{y}_i)$ can be obtained from Proposition 5 with $g(u) = u$. To obtain $\xi(\boldsymbol{\theta}^{(k)}, \mathbf{y}_i)$ and $\omega(\boldsymbol{\theta}^{(k)}, \mathbf{y}_i)$, we first note that, from Proposition 4,

$$\begin{aligned} T_i|\mathbf{y}_i, u_i &\sim \text{TN}(m(\boldsymbol{\theta}, \mathbf{y}_i), u_i^{-1} \mathbf{M}^2(\boldsymbol{\theta}), (0, \infty)), \quad \text{with } \mathbf{M}^2(\boldsymbol{\theta}) = (1 + \boldsymbol{\Delta}^\top \boldsymbol{\Gamma}^{-1} \boldsymbol{\Delta})^{-1} \\ \text{and } m(\boldsymbol{\theta}, \mathbf{y}_i) &= \mathbf{M}(\boldsymbol{\theta}) \boldsymbol{\Delta}^\top \boldsymbol{\Gamma}^{-1} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta} - b \boldsymbol{\Delta})^\top. \end{aligned}$$

Let $W_\Phi(\cdot) = \phi(\cdot)/\Phi(\cdot)$. Recall that a $\text{TN}(c, d^2, (0, \infty))$ distribution has mean $c + d W_\Phi(c/d)$ and variance $d^2 \{1 - W_\Phi(c/d)[W_\Phi(c/d) + c/d]\}$. Also, for any measurable function $g(\cdot)$, we have $\mathbb{E}\{g(T)U|\mathbf{Y}\} = \mathbb{E}\{UE\{g(T)|\mathbf{Y}, U\}|\mathbf{Y}\}$. Combining these results with the relation

$$m(\boldsymbol{\theta}, \mathbf{y}_i)/\mathbf{M}(\boldsymbol{\theta}) = (1 + \boldsymbol{\Delta}^\top \boldsymbol{\Gamma}^{-1} \boldsymbol{\Delta})^{-1/2} \boldsymbol{\Delta}^\top \boldsymbol{\Gamma}^{-1} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta} - b \boldsymbol{\Delta})^\top = \boldsymbol{\lambda}^T \boldsymbol{\Sigma}^{-1/2} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta} - b \boldsymbol{\Delta})^\top \equiv \mathbf{A}(\boldsymbol{\theta}, \mathbf{y}_i),$$

we obtain the following expressions

$$\begin{aligned}\xi(\boldsymbol{\theta}^{(k)}, \mathbf{y}_i) &= U(\boldsymbol{\theta}^{(k)}, \mathbf{y}_i)(m(\boldsymbol{\theta}^{(k)}, \mathbf{y}_i) + b) + M(\boldsymbol{\theta}^{(k)})\tau(\boldsymbol{\theta}^{(k)}, \mathbf{y}_i) \text{ and} \\ \omega(\boldsymbol{\theta}^{(k)}, \mathbf{y}_i) &= U(\boldsymbol{\theta}^{(k)}, \mathbf{y}_i)(m(\boldsymbol{\theta}^{(k)}, \mathbf{y}_i) + b)^2 + (M(\boldsymbol{\theta}^{(k)}))^2 + M(\boldsymbol{\theta}^{(k)})(m(\boldsymbol{\theta}^{(k)}, \mathbf{y}_i) + 2b)\tau(\boldsymbol{\theta}^{(k)}, \mathbf{y}_i),\end{aligned}\tag{4.10}$$

where

$$\tau(\boldsymbol{\theta}^{(k)}, \mathbf{y}_i) = E_{\boldsymbol{\theta}^{(k)}}\{U_i^{1/2}W_{\Phi}(U_i^{1/2}A(\boldsymbol{\theta}^{(k)}, \mathbf{y}_i))|\mathbf{y}_i\}.\tag{4.11}$$

The expectation $\tau(\boldsymbol{\theta}^{(k)}, \mathbf{y}_i)$ can be derived from Proposition 5 with $g(u) = u^{1/2}W_{\Phi}(u^{1/2}A(\boldsymbol{\theta}^{(k)}, \mathbf{y}_i))$.

The M-step consists of maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ with respect to $\boldsymbol{\theta}$. To do this, we use a faster extension of the original EM called *ECME algorithm* (Liu and Rubin, 1994), replacing the M step with a sequence of conditional maximization steps. Specifically, we update $\boldsymbol{\beta}^{(k)}$ by maximizing the partial function $\boldsymbol{\beta} \mapsto Q(\boldsymbol{\beta}, \boldsymbol{\alpha}^{(k)}, \boldsymbol{\lambda}^{(k)}, \boldsymbol{\nu}^{(k)}|\boldsymbol{\theta}^{(k)})$, update $\boldsymbol{\alpha}^{(k)}$ by maximizing $\boldsymbol{\alpha} \mapsto Q(\boldsymbol{\beta}^{(k+1)}, \boldsymbol{\alpha}, \boldsymbol{\lambda}^{(k)}, \boldsymbol{\nu}^{(k)}|\boldsymbol{\theta}^{(k)})$ and so on. The path corresponding to $\boldsymbol{\nu}$ follows a different pattern, and consists of maximizing the *current* log-likelihood function with respect to $\boldsymbol{\nu}$. The algorithm can be summarized as follows

E-step: Given $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$, compute $U_i^{(k)} = U(\boldsymbol{\theta}^{(k)}, \mathbf{y}_i)$, $\xi_i^{(k)} = \xi(\boldsymbol{\theta}^{(k)}, \mathbf{y}_i)$ and $\omega_i^{(k)} = \omega(\boldsymbol{\theta}^{(k)}, \mathbf{y}_i)$, for $i = 1, \dots, n$.

M-step:

1. Update $\boldsymbol{\mu}^{(k)}$, $\boldsymbol{\Gamma}^{(k)}$ and $\boldsymbol{\Delta}^{(k)}$ using the following closed form expressions

$$\begin{aligned}\boldsymbol{\beta}^{(k+1)} &= \left(\sum_{i=1}^n U_i^{(k)} \mathbf{X}_i^{\top} \mathbf{X}_i\right)^{-1} \sum_{i=1}^n \mathbf{X}_i^{\top} (U_i^{(k)} \mathbf{y}_i - \xi_i^{(k)} \boldsymbol{\Delta}^{(k)})^{\top}; \\ \boldsymbol{\Delta}^{(k+1)} &= \sum_{i=1}^n \xi_i^{(k)} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}^{(k+1)})^{\top} / \sum_{i=1}^n \omega_i^{(k)}; \\ \boldsymbol{\Gamma}^{(k+1)} &= \frac{1}{n} \sum_{i=1}^n \left\{ U_i^{(k)} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}^{(k+1)})^{\top} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}^{(k+1)}) - \xi_i^{(k)} \left[\boldsymbol{\Delta}^{(k+1)} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}^{(k+1)}) \right. \right. \\ &\quad \left. \left. + (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}^{(k+1)})^{\top} (\boldsymbol{\Delta}^{(k+1)})^{\top} \right] + \omega_i^{(k)} \boldsymbol{\Delta}^{(k+1)} (\boldsymbol{\Delta}^{(k+1)})^{\top} \right\};\end{aligned}$$

2. Update $\boldsymbol{\nu}^{(k)}$ by maximizing the current marginal log-likelihood function,

$$\boldsymbol{\nu}^{(k+1)} = \arg \max_{\boldsymbol{\nu}} \sum_{i=1}^n \log(\phi_{\text{SMSN}_p}(\mathbf{y}_i | \mathbf{X}_i \boldsymbol{\beta}^{(k+1)} + b \boldsymbol{\Delta}^{(k+1)}, \boldsymbol{\Sigma}^{(k+1)}, \boldsymbol{\lambda}^{(k+1)}, \boldsymbol{\nu})).$$

In stage 2 of the M-step, the updates $\boldsymbol{\Sigma}^{(k+1)}$ and $\boldsymbol{\lambda}^{(k+1)}$ are obtained using expressions (4.4) and

the maximization can be accomplished by using, for instance, the `optim` routine in the R software.

4.4 The FM-SMSN-MR model

4.4.1 The model and parameter estimation

Extending the ideas presented above, we expanded the model presented in (4.8) to consider the mixture multivariate regression where the random errors ϵ_i follow a SMSN distribution as defined in (4.2). Consider a set of independent data vectors $\mathbf{Y}_i$, for $i = 1, \dots, n$, taken from a G -component mixture of SMSN densities given as:

$$f(\mathbf{y}_i|\boldsymbol{\theta}) = \sum_{j=1}^G \pi_j \phi_{SMSP}(\mathbf{y}_i|\mathbf{X}_i\boldsymbol{\beta}_j + b(\boldsymbol{\nu}_j)\boldsymbol{\Delta}_j, \boldsymbol{\Sigma}_j, \boldsymbol{\lambda}_j, \boldsymbol{\nu}_j), \quad \pi_j \geq 0, \quad \sum_{j=1}^G \pi_j = 1, \quad (4.12)$$

where $\mathbf{x}_i$ and $\boldsymbol{\beta}_j$ has dimensions $p \times q$ and $q \times 1$ respectively. Also, we have that $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_G, \pi_1, \dots, \pi_G)^\top$, $b(\boldsymbol{\nu}_j) = b(\boldsymbol{\nu}) = b$, for $j = 1, \dots, G$, this being the component specific vector of parameters, and π_j 's are the mixing probabilities. We assume that the distributions of the mixing scale factors are identical to $H(\cdot|\boldsymbol{\nu})$. The model defined by (4.12) will be called the SMSN-MMRM.

We assume that the number of components G is known and fixed, whereas $\pi_1, \dots, \pi_G$ is an unknown vector of mixture weights and is constrained to be positive. Our interest is to estimate the unknown parameter $\boldsymbol{\theta}$, given $(\mathbf{x}_{1j}, \mathbf{y}_1), \dots, (\mathbf{x}_{nj}, \mathbf{y}_n)$, which is traditionally estimated by the ML estimate:

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} \sum_{i=1}^n \log(f(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\theta})). \quad (4.13)$$

Therefore, for maximizer (4.13) we propose to use an EM-type algorithm (Dempster *et al.*, 1977), because the expression does not have an explicit solution. Identifiability of mixtures has been discussed by several authors, among others Yakowitz and Spragins (1968), and Atienza *et al.* (2006), both of whom provided sufficient conditions to show identifiability under Teicher's definition (Teicher, 1961, 1963, 1967). In the context of univariate or multivariate linear regression mixtures, Hennig (2000) and Grün and Leisch (2008) proved they are identifiable. Holzmann *et al.* (2006) proved the identifiability the finite mixtures of elliptical distributions. No proof exists of the identifiability of the finite mixture class of skew elliptical distributions, but Otiniano *et al.* (2015) proved the identifiability of finite mixture of multivariate skew-normal and skew-t distributions.

Maximum likelihood (ML) estimation of the model parameters may be carried out through the well-known Expectation-Maximization (EM) algorithm (Dempster *et al.*, 1977). This is a general

purpose algorithm for ML estimation in a wide variety of situations best described as incomplete-data problems. We implement the EM-type algorithm for the estimation of the parameters of the SMSN–MMRM model. As usual in the EM framework for mixtures (McLachlan and Peel, 2000, Sec 2.8.2), we introduce a set of latent component indicators $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{iG})^\top$, $i = 1, \dots, n$, where the coordinate j is a binary random variable defined as

$$Z_{ij} = \begin{cases} 1, & \text{if } \mathbf{Y}_i \text{ belongs to group } j, \\ 0, & \text{otherwise} \end{cases} \quad (4.14)$$

and $\sum_{j=1}^G Z_{ij} = 1$. That is, the distribution of $\mathbf{Y}_i | Z_{ij} = j$ is $\text{SMSN}_p(\boldsymbol{\beta}_j, \boldsymbol{\Sigma}_j, \boldsymbol{\lambda}_j, \boldsymbol{\nu})$. The indicators $\mathbf{Z}_1, \dots, \mathbf{Z}_n$ are independent, each one with multinomial distribution,

$$f(\mathbf{z}_i) = \pi_1^{z_{i1}} \pi_2^{z_{i2}} \dots (1 - \pi_1 - \dots - \pi_{G-1})^{z_{iG}}$$

and we write $\mathbf{Z}_i \sim \text{Multinomial}(1; \pi_1, \dots, \pi_G)$.

From Proposition 4 it follows that the hierarchical representation for the SMSN–MMRM model can be written as

$$\begin{aligned} \mathbf{Y}_i | U_i = u_i, T_i = t_i, Z_{ij} = 1 &\sim \text{N}_p(\mathbf{X}_i \boldsymbol{\beta}_j + \boldsymbol{\Delta}_j t_i, u_i^{-1} \boldsymbol{\Gamma}_j), \\ T_i | U_i = u_i, Z_{ij} = 1 &\sim \text{TN}(b, u_i^{-1}) I_{(b, \infty)}, \\ U_i | Z_{ij} = 1 &\sim \text{H}(u_i | \boldsymbol{\nu}), \\ \mathbf{Z}_i &\sim \text{Multinomial}(1; \pi_1, \dots, \pi_g), \end{aligned} \quad (4.15)$$

$i = 1, \dots, n$, $j = 1, \dots, G$, where

$$\boldsymbol{\Delta}_j = \boldsymbol{\Sigma}_j^{1/2} \boldsymbol{\delta}_j, \quad \boldsymbol{\delta}_j = \frac{\boldsymbol{\lambda}_j}{\sqrt{1 + \boldsymbol{\lambda}_j^\top \boldsymbol{\lambda}_j}}, \quad \boldsymbol{\Gamma}_j = \boldsymbol{\Sigma}_j - \boldsymbol{\Delta}_j \boldsymbol{\Delta}_j^\top.$$

If $U_i = 1$ (that is, the skew-normal case), the model is named SN–MMRM, while if $U_i \sim \text{Gamma}(\nu/2, \nu/2)$ (the skew-t case) it is named ST–MMRM, and so on.

We have that, from representation (4.15), the complete log-likelihood function is

$$\begin{aligned} \ell_c(\boldsymbol{\theta} | \mathbf{y}, \mathbf{t}, \mathbf{u}, \mathbf{z}) &= cte + \sum_{i=1}^n \sum_{j=1}^g Z_{ij} \left[\log(\pi_j) - \frac{1}{2} \log |\boldsymbol{\Gamma}_j| \right. \\ &\quad \left. - \frac{u_i}{2} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}_j - \boldsymbol{\Delta}_j t_i)^\top \boldsymbol{\Gamma}_j^{-1} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}_j - \boldsymbol{\Delta}_j t_i) + \log(h(u_i | \boldsymbol{\nu})) \right], \end{aligned}$$

where cte is a constant and $\mathbf{z} = (\mathbf{z}_1^\top, \dots, \mathbf{z}_n^\top)^\top$. The conditional expectations involved in the E-step, that is, the computation of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = E_{\boldsymbol{\theta}^{(k)}}\{\ell_c(\boldsymbol{\theta}|\mathbf{y}, \mathbf{t}, \mathbf{u}, \mathbf{z})|\mathbf{y}\}$, are

$$\begin{aligned} \rho_{ij}^{(k)} &= E_{\boldsymbol{\theta}^{(k)}}\{Z_{ij}|\mathbf{y}_i\}, \quad U_{ij}^{(k)} = E_{\boldsymbol{\theta}^{(k)}}\{Z_{ij}U_i|\mathbf{y}_i\}, \quad \xi_{ij}^{(k)} = E_{\boldsymbol{\theta}^{(k)}}\{Z_{ij}U_iT_i|\mathbf{y}_i\}, \\ \text{and } \omega_{ij}^{(k)} &= E_{\boldsymbol{\theta}^{(k)}}\{Z_{ij}U_iT_i^2|\mathbf{y}_i\}. \end{aligned}$$

The expression for $U_{ij}^{(k)}$ can be obtained using the relation $E\{Z_{ij}U_i|\mathbf{Y}_i\} = E\{Z_{ij}E\{U_i|\mathbf{Y}_i, Z_{ij}\}|\mathbf{Y}_i\}$. For each fixed $\mathbf{y}_i$, the distribution of $g(\mathbf{Y}_i, Z_{ij}) = Z_{ij}E\{U_i|\mathbf{Y}_i, Z_{ij}\}$ given $\mathbf{Y}_i = \mathbf{y}_i$ is binary, assuming the value $g(\mathbf{y}_i, 1) = E\{U_i|\mathbf{Y}_i = \mathbf{y}_i, Z_{ij} = 1\}$ with probability $P(Z_{ij} = 1|\mathbf{Y}_i = \mathbf{y}_i) = \rho_{ij}^{(k)}$ and zero with probability $P(Z_{ij} = 0|\mathbf{Y}_i = \mathbf{y}_i)$. It is straightforward to show that

$$\rho_{ij}^{(k)} = \frac{\pi_j^{(k)} \phi_{\text{SMSN}_p}(\mathbf{y}_i|\boldsymbol{\theta}_j^{(k)})}{\sum_{l=1}^G \pi_l^{(k)} \phi_{\text{SMSN}_p}(\mathbf{y}_i|\boldsymbol{\theta}_l^{(k)})} \quad (4.16)$$

and that the distribution of $U_i|\mathbf{Y}_i = \mathbf{y}_i, Z_{ij} = 1$ is the distribution of $U_i|\mathbf{Y}_0 = \mathbf{y}_i$ when $\mathbf{Y}_0 \sim \text{SMSN}(\boldsymbol{\theta}_j^{(k)})$. Thus, the associated expected value is $U(\boldsymbol{\theta}_j^{(k)}, \mathbf{y}_i)$ given in (4.9). Using similar arguments in order to obtain the other conditional expectations, we have

$$U_{ij}^{(k)} = \rho_{ij}^{(k)} U(\boldsymbol{\theta}_j^{(k)}, \mathbf{y}_i), \quad \xi_{ij}^{(k)} = \rho_{ij}^{(k)} \xi(\boldsymbol{\theta}_j^{(k)}, \mathbf{y}_i) \quad \text{and} \quad \omega_{ij}^{(k)} = \rho_{ij}^{(k)} \omega(\boldsymbol{\theta}_j^{(k)}, \mathbf{y}_i). \quad (4.17)$$

These expressions can be evaluated using their counterparts for the one-component case, given in (4.9) and (4.11).

The ECME algorithm for the estimation of $\boldsymbol{\theta}$ is a direct extension of the algorithm presented in section 4.3.1 for the case with one component. We have adopted the same strategy to obtain $\boldsymbol{\nu}^{(k+1)}$, by direct maximization of the current marginal log-likelihood (ECME algorithm), thus avoiding the process of computing $E_{\boldsymbol{\theta}^{(k)}}\{Z_{ij} \log(h(u_i|\boldsymbol{\nu}))|\mathbf{y}_i\}$. It can be summarized as follows:

E-step: Given $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$, compute $\rho_{ij}^{(k)}$, $U_{ij}^{(k)}$, $\xi_{ij}^{(k)}$ and $\omega_{ij}^{(k)}$, for $i = 1, \dots, n$ and $j = 1, \dots, g$.

M-step:

1. For $j = 1, \dots, g$, update $\pi_j^{(k)}$, $\beta_j^{(k)}$, $\Gamma_j^{(k)}$ and $\Delta_j^{(k)}$ using the following closed form expressions

$$\begin{aligned}
\pi_j^{(k+1)} &= n^{-1} \sum_{i=1}^n \rho_{ij}^{(k)}; \\
\beta_j^{(k+1)} &= \left(\sum_{i=1}^n U_{ij}^{(k)} \mathbf{X}_i^\top \mathbf{X}_i \right)^{-1} \left(\sum_{i=1}^n \mathbf{X}_i^\top (U_{ij}^{(k)} \mathbf{y}_i - \xi_{ij}^{(k)} \Delta_j^{(k)}) \right); \\
\Delta_j^{(k+1)} &= \left[\sum_{i=1}^n \xi_{ij}^{(k)} (\mathbf{y}_i - \mathbf{X}_i \beta_j^{(k+1)})^\top \right] / \sum_{i=1}^n \omega_{ij}^{(k)}; \\
\Gamma_j^{(k+1)} &= \left(\sum_{i=1}^n \rho_{ij}^{(k)} \right)^{-1} \sum_{i=1}^n \left(U_{ij}^{(k)} (\mathbf{y}_i - \mathbf{X}_i \beta_j^{(k+1)})^\top (\mathbf{y}_i - \mathbf{X}_i \beta_j^{(k+1)}) \right. \\
&\quad \left. - \left[\Delta_j^{(k+1)} (\mathbf{y}_i - \mathbf{X}_i \beta_j^{(k+1)}) + (\mathbf{y}_i - \mathbf{X}_i \beta_j^{(k+1)})^\top (\Delta_j^{(k+1)})^\top \right] \xi_{ij}^{(k)} \right. \\
&\quad \left. + \omega_{ij}^{(k)} \Delta_j^{(k+1)} (\Delta_j^{(k+1)})^\top \right);
\end{aligned}$$

2. Update $\nu^{(k)}$ by maximizing the current marginal log-likelihood function, obtaining

$$\nu^{(k+1)} = \arg \max_{\nu} \sum_{i=1}^n \log \left(\sum_{j=1}^g \pi_j \phi_{\text{SMSN}_p}(\mathbf{y}_i | \mathbf{X}_i \beta_j^{(k+1)}, \Sigma_j^{(k+1)}, \lambda_j^{(k+1)}, \nu) \right).$$

A more parsimonious model is achieved by supposing $\Gamma_1 = \dots = \Gamma_G = \Gamma$, which can be seen as an extension of the FM-NOR model with restricted variance-covariance components. In this case, the updates for $\pi_j^{(k)}$, $\beta_j^{(k)}$ and $\Delta_j^{(k)}$ remain the same, and the update for $\Gamma^{(k)}$ is given as

$$\Gamma^{(k+1)} = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^g \rho_{ij}^{(k)} \Gamma_j^{(k+1)}.$$

4.4.2 Starting values

It is well known that EM and related algorithms can be sensitive to the initial values chosen, see the discussion in McLachlan and Peel (2000) for example. Note that mixture models can provide a multimodal log-likelihood function. In this sense, the method of maximum likelihood estimation through EM algorithm may not give maximum global solutions if the starting values are far from the real parameter values. Thus, the choice of starting values for the EM algorithm in the mixture context plays a big role in parameter estimation. Note that the number of components G is considered known *a priori*, so they are not parameters. In our examples and simulation studies, we consider the following procedure for the SMSN-MMRM model (available in the package `FMsmnReg`) (Benites *et al.*, 2016):

- For $\beta_j^{(0)}$, use the ordinary least-squares (OLS) estimate in the regression model defined in (12).
- Partition the residuals into G groups using the K-means clustering algorithm.
- Compute the proportion of data points belonging to the same cluster j , say $\pi_j^{(0)}$, $j = 1, \dots, G$. This is the initial value for π_j .
- For each group j , compute the initial values $(\Sigma_j)^{(0)}$ using the method of moments estimators.
- The starting values for ν_j are taken to be equal to 3 in the Student-t and slash cases and $\nu_j = (0.1, 0.1)$ in the contaminated normal case.

We adopt the Bayesian information criterion (BIC; Schwarz (1978)) to select the number of components in mixture models. Herein, the form of BIC is given by

$$\text{BIC} = -2\ell(\hat{\theta}) + \rho \log n,$$

where $\ell(\hat{\theta})$ is the current log-likelihood and ρ is the number of parameters in the model. Accordingly, models with small BIC scores are preferred. Recently McNicholas and Murphy (2008) showed the effectiveness of the BIC in selecting the number of components for Gaussian mixture models.

4.4.3 Aitken's acceleration stopping criterion

To assess the convergence of the EM algorithm iterations are alternated repeatedly until a suitable stopping criterion is satisfied. We adopt the Aitken acceleration method (Aitken, 1927). Suppose that the algorithm converges at a linear rate. Then it may require many iterations to achieve satisfactory parameter estimates. To determine convergence, Aitken's acceleration estimates the asymptotic maximum of the log-likelihood at each iteration k such that

$$\ell^{(k+1)} - \ell \approx a(\ell^{(k)} - \ell), \quad 0 < a < 1.$$

If the constant a can be estimated by $a^{(k)}$, then the asymptotic estimate of $\ell^{(k+1)}$ can be calculated as

$$\ell_{\infty}^{(k+1)} = \ell^{(k)} + \frac{\ell^{(k+1)} - \ell^{(k)}}{1 - a^{(k)}}, \quad k > 1, \quad (4.18)$$

where $\ell^{(k+1)}$ is the observed log-likelihood evaluated at $\theta^{(k+1)}$ and $a^{(k)} = (\ell^{(k+1)} - \ell^{(k)})/(\ell^{(k)} - \ell^{(k-1)})$, denotes the Aitken acceleration at the k th iteration, which is very close to 1 at convergence.

As recommended by Böhning *et al.* (1994), the algorithm can be considered to achieve convergence if $\ell_{\infty}^{(k+1)} - \ell^{(k+1)} < \varepsilon$ where ε is the desired tolerance. For the numerical analysis in Section 6 and 7, a default value of $\varepsilon = 10^{-6}$ is used to terminate the iterations. Note that the above procedure is also applicable for the simple case ($G = 1$) by treating $Z_{ij} = 1$. See Dias and Wedel (2004) for a discussion in the context of mixtures of univariate normal distributions.

4.4.4 Obtaining of 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 $\hat{\boldsymbol{\theta}}$ by the inverse of the observed information matrix. 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 (4.12).

In this work we use the alternative method suggested by Basford *et al.* (1997b), which consists of in approximating the inverse of the covariance matrix by $\mathbf{I}_o(\hat{\boldsymbol{\theta}}|\mathbf{y})$, where

$$\mathbf{I}_o(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^n \hat{\mathbf{s}}_i \hat{\mathbf{s}}_i^\top, \quad \text{with} \quad \hat{\mathbf{s}}_i = \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\mathbf{y}_i|\boldsymbol{\theta}), \quad (4.19)$$

where $f(\cdot|\boldsymbol{\theta})$ is the SMSN-MMRM density in (4.12) (with $\boldsymbol{\nu}_1 = \dots = \boldsymbol{\nu}_G = \boldsymbol{\nu}$). That is, $\hat{\mathbf{s}}_i$ is the score vector associated to $\mathbf{y}_i$ and

$$\hat{\mathbf{s}}_i = (\hat{s}_{i,\boldsymbol{\beta}_1^\top}, \dots, \hat{s}_{i,\boldsymbol{\beta}_G^\top}, \hat{s}_{i,\boldsymbol{\lambda}_1}, \dots, \hat{s}_{i,\boldsymbol{\lambda}_G}, \hat{s}_{i,\boldsymbol{\alpha}_1}, \dots, \hat{s}_{i,\boldsymbol{\alpha}_G}, \hat{s}_{i,\pi_1}, \dots, \hat{s}_{i,\pi_{G-1}})^\top.$$

Expressions for the element's partial derivatives with respect to the other component specific parameters are denoted in an analogous way. Then, we have

$$\begin{aligned} \hat{s}_{i,\pi_j} &= \frac{S_j(\mathbf{y}_i) - S_G(\mathbf{y}_i)}{f(\mathbf{y}_i|\boldsymbol{\theta})}, \quad i = 1, \dots, n, \quad j = 1, \dots, G-1, \\ \hat{s}_{i,\boldsymbol{\beta}_j} &= \frac{p_j D_{\boldsymbol{\beta}_j}(S_j(\mathbf{y}_i))}{f(\mathbf{y}_i|\boldsymbol{\theta})}, \quad \hat{s}_{i,\boldsymbol{\alpha}_{jk}} = \frac{p_j D_{\boldsymbol{\alpha}_{jk}}(S_j(\mathbf{y}_i))}{f(\mathbf{y}_i|\boldsymbol{\theta})}, \\ \hat{s}_{i,\boldsymbol{\lambda}_j} &= \frac{p_j D_{\boldsymbol{\lambda}_j}(S_j(\mathbf{y}_i))}{f(\mathbf{y}_i|\boldsymbol{\theta})}, \quad \hat{s}_{i,\boldsymbol{\nu}} = \frac{\sum_{j=1}^G p_j D_{\boldsymbol{\nu}}(S_j(\mathbf{y}_i))}{f(\mathbf{y}_i|\boldsymbol{\theta})}, \\ &\quad i = 1, \dots, n, \quad j = 1, \dots, G. \end{aligned}$$

Note that $D_{\boldsymbol{\nu}}(S_j(\mathbf{y}_i))$ must be obtained for each particular case mentioned in Section 2. After some

algebraic manipulation and using the notation

$$\begin{aligned} \mathbf{I}_{ij}^\Phi(w) &= \int_0^\infty u_i^w \exp\left\{-\frac{1}{2}u_i \mathbf{d}_{ij}\right\} \Phi(u_i^{1/2} \mathbf{A}_{ij}) dH(u_i|\boldsymbol{\nu}), \\ \mathbf{I}_{ij}^\phi(w) &= \int_0^\infty u_i^w \exp\left\{-\frac{1}{2}u_i \mathbf{d}_{ij}\right\} \phi(u_i^{1/2} \mathbf{A}_{ij}) dH(u_i|\boldsymbol{\nu}), \end{aligned}$$

where $\mathbf{d}_{ij} = \mathbf{d}_{\boldsymbol{\Sigma}_j}(\mathbf{y}_i, \mathbf{X}_i \boldsymbol{\beta}_j + b(\boldsymbol{\nu}_j) \boldsymbol{\Delta}_j)$ and $\mathbf{A}_{ij} = \boldsymbol{\lambda}_j^\top \boldsymbol{\Sigma}_j^{-1/2}(\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}_j - b(\boldsymbol{\nu}_j) \boldsymbol{\Delta}_j)$, one can show that

$$\begin{aligned} D_{\boldsymbol{\beta}_j}(\mathbf{S}_j(\mathbf{y}_i)) &= \frac{2|\boldsymbol{\Sigma}_j|^{-1/2}}{(2\pi)^{p/2}} \left[\left(\frac{\partial \mathbf{A}_{ij}}{\partial \boldsymbol{\beta}_j} \right) \mathbf{I}_{ij}^\phi \left(\frac{p+1}{2} \right) - \frac{1}{2} \left(\frac{\partial \mathbf{d}_{ij}}{\partial \boldsymbol{\beta}_j} \right) \mathbf{I}_{ij}^\Phi \left(\frac{p}{2} + 1 \right) \right], \\ D_{\boldsymbol{\alpha}_{jk}}(\mathbf{S}_j(\mathbf{y}_i)) &= \frac{2}{(2\pi)^{p/2}} \left[\left(\frac{\partial |\boldsymbol{\Sigma}_j|^{-1/2}}{\partial \boldsymbol{\alpha}_{jk}} \right) \mathbf{I}_{ij}^\Phi \left(\frac{p}{2} \right) - \frac{1}{2} \left(\frac{\partial \mathbf{d}_{ij}}{\partial \boldsymbol{\alpha}_{jk}} \right) |\boldsymbol{\Sigma}_j|^{-1/2} \mathbf{I}_{ij}^\Phi \left(\frac{p}{2} + 1 \right) \right. \\ &\quad \left. + |\boldsymbol{\Sigma}_j|^{-1/2} \left(\frac{\partial \mathbf{A}_{ij}}{\partial \boldsymbol{\alpha}_{jk}} \right) \mathbf{I}_{ij}^\phi \left(\frac{p+1}{2} \right) \right], \\ D_{\boldsymbol{\lambda}_j}(\mathbf{S}_j(\mathbf{y}_i)) &= \frac{2|\boldsymbol{\Sigma}_j|^{-1/2}}{(2\pi)^{p/2}} \left(\frac{\partial \mathbf{A}_{ij}}{\partial \boldsymbol{\lambda}_j} \right) \mathbf{I}_{ij}^\phi \left(\frac{p+1}{2} \right). \end{aligned}$$

Expressions for the derivatives are standard and are not given here. Direct substitution of $H(\cdot|\boldsymbol{\nu})$ in the integrals above yields the following results for each distribution considered in this work, namely:

- *The skew-normal distribution*

$$\mathbf{I}_{ij}^\Phi(w) = \exp\{-\mathbf{d}_{ij}/2\} \Phi(\mathbf{A}_{ij}), \quad \mathbf{I}_{ij}^\phi(w) = \exp\{-\mathbf{d}_{ij}/2\} \phi(\mathbf{A}_{ij}).$$

- *The skew-t distribution*

$$\begin{aligned} \mathbf{I}_{ij}^\Phi(w) &= \frac{2^w \nu^{\nu/2} \Gamma(w + \nu/2)}{\sqrt{2\pi} \Gamma(\nu/2) (\nu + \mathbf{d}_{ij})^{\nu/2+w}} T \left(\frac{\mathbf{A}_{ij}}{(\mathbf{d}_{ij} + \nu)^{1/2}} \sqrt{2w + \nu} |2w + \nu \right), \\ \mathbf{I}_{ij}^\phi(w) &= \frac{2^w \nu^{\nu/2}}{\sqrt{2\pi} \Gamma(\nu/2)} \left(\frac{1}{\mathbf{d}_{ij} + \mathbf{A}_{ij}^2 + \nu} \right)^{\frac{\nu+2w}{2}} \Gamma \left(\frac{\nu + 2w}{2} \right), \end{aligned}$$

$$\begin{aligned} D_\nu(\mathbf{S}_j(\mathbf{y}_i)) &= (2\pi)^{-p/2} |\boldsymbol{\Sigma}_j|^{-1/2} \left(\mathbf{I}_{ij}^\Phi(p/2) (1 + \log(\nu/2) - \psi(\nu/2)) - \mathbf{I}_{ij}^\Phi(1 + p/2) \right. \\ &\quad \left. + \int_0^\infty u_i^{p/2} \log(u_i) \exp\{-u_i \mathbf{d}_{ij}/2\} \times \Phi(u_i^{1/2} \mathbf{A}_{ij}) h(u_i|\nu) du_i \right), \end{aligned}$$

where $\psi(\cdot)$ denotes the digamma function and $h(\cdot|\nu)$ is the gamma density with both param-

eters $\nu/2$.

- *The skew-slash distribution*

$$\begin{aligned} I_{ij}^\Phi(w) &= \frac{2^{2+\nu}\Gamma(w+\nu)}{d_{ij}^{w+\nu}} P_1(w+\nu, d_{ij}/2) E[\Phi(V_{ij}^{1/2}) A_{ij}], \\ I_{ij}^\phi(w) &= \frac{\nu 2^{w+\nu}\Gamma(w+\nu)}{\sqrt{2\pi}(d_{ij} + A_{ij}^2)^{w+\nu}} P_1\left(w+\nu, \frac{d_{ij} + A_{ij}^2}{2}\right), \end{aligned}$$

where $V_{ij} \sim \text{TG}(w+\nu, d_{ij}/2, (0, 1))$ and

$$\begin{aligned} D_\nu(S_j(\mathbf{y}_i)) &= 2(2\pi)^{-p/2} |\Sigma_j|^{-1/2} \left\{ I_{ij}^\Phi(p/2 + \nu - 1) \right. \\ &\quad \left. + \nu \int_0^1 u_i^{\frac{p}{2} + \nu - 1} \log(u_i) \exp\{-u_i d_{ij}/2\} \Phi(u_i^{1/2} A_{ij}) du_i \right\}. \end{aligned}$$

- *The skew-contaminated normal distribution*

$$\begin{aligned} I_{ij}^\Phi(w) &= \sqrt{2\pi} \{ \nu_1 \nu_2^{w-1/2} \phi(d_{ij}|0, 1/\nu_2) \Phi(\nu_2^{1/2} A_{ij}) + (1 - \nu_1) \phi(d_{ir}) \Phi(A_{ij}) \}, \\ I_{ij}^\phi(w) &= \{ \nu_1 \nu_2^{w-1/2} \phi(d_{ij} + A_{ij}^2|0, 1/\nu_2) + (1 - \nu_1) \phi(d_{ij} + A_{ij}^2) \}, \end{aligned}$$

$$\begin{aligned} D_{\nu_1}(S_j(\mathbf{y}_i)) &= 2\{ \phi_p(y_i|\boldsymbol{\mu}_j, \nu_2^{-1}\Sigma_j) \Phi(\nu_2^{1/2} A_{ij}) - \phi_p(y_i|\boldsymbol{\mu}_j, \Sigma_j) \Phi(A_{ij}) \}, \\ D_{\nu_2}(S_j(\mathbf{y}_i)) &= \nu_1 \phi_p(\boldsymbol{\mu}_j, \nu_2^{-1}\Sigma_j) \{ A_{ij} \nu_2^{-1/2} \phi(\nu_2^{1/2} A_{ij}) + \Phi(\nu_2^{1/2} A_{ij}) (p\nu_2^{-1} - d_{ij}) \}. \end{aligned}$$

In the next section we perform simulation experiments which show evidence that, at least asymptotically, this is a reliable method to provide standard error estimates in the SMSN-MMRM context. More accurate estimates could be perhaps obtained by using the bootstrap approach (Efron and Tibshirani, 1986), but this gain is achieved at with the price of an enormous increase of computing time.

4.5 Simulation study

In this section, we perform simulation studies to evaluate the performance and properties of the ML estimators of the SMSN-MMRM, considering the ST-MMRM and SCN-MMRM. The estimates are obtained through the EM algorithm presented in Section 3.2 by analyzing some artificial and real datasets. The goals re to evaluate to accuracy of the estimates and consistency of the approximate standard errors for the ML estimates of parameters (simulation study 1) and

to show that our proposed ECME algorithm estimates do provide good asymptotic properties. In all simulation studies we consider 500 Monte Carlo replications of the following model

$$f(\mathbf{y}_i|\boldsymbol{\theta}) = \pi\phi_{\text{SMSN}_2}(\mathbf{y}_i|\mathbf{X}_i\boldsymbol{\beta}_1 + b(\boldsymbol{\nu})\boldsymbol{\Delta}_1, \boldsymbol{\Sigma}_1, \boldsymbol{\lambda}_1, \boldsymbol{\nu}) + (1 - \pi)\phi_{\text{SMSN}_2}(\mathbf{y}_i|\mathbf{X}_i\boldsymbol{\beta}_2 + b(\boldsymbol{\nu})\boldsymbol{\Delta}_2, \boldsymbol{\Sigma}_2, \boldsymbol{\lambda}_2, \boldsymbol{\nu}), \quad (4.20)$$

where

$$\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{21}, \beta_{31})^\top, \quad \boldsymbol{\beta}_2 = (\beta_{12}, \beta_{22}, \beta_{32})^\top, \quad \boldsymbol{\Sigma}_j = \begin{bmatrix} \sigma_{j.11} & \sigma_{j.12} \\ \sigma_{j.21} & \sigma_{j.22} \end{bmatrix}, \quad \boldsymbol{\lambda}_j = (\lambda_{1j}, \lambda_{2j})^\top, \quad j = 1, 2.$$

with

$$\boldsymbol{\beta}_1 = (-1, -4, -3)^\top, \quad \boldsymbol{\beta}_2 = (3, 7, 9)^\top, \quad \boldsymbol{\Sigma}_1 = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix},$$

$$\boldsymbol{\lambda}_1 = (3, 5)^\top, \quad \boldsymbol{\lambda}_2 = (1, 4)^\top, \quad \pi = 0.3, \quad \nu = 3,$$

finally

$$\mathbf{X}_{i1} = \begin{bmatrix} 1 & X_{i1.11} & X_{i1.12} \end{bmatrix},$$

such that $X_{i1.k1} \sim U(0, 1)$, $X_{i1.k2} \sim U(2, 4)$ and $X_{i2.k1} \sim U(0, 1)$, $i = 1, \dots, n$, $k = 1, 2$. The initial values of the ECME algorithm were obtained considering the strategy given in Subsection 4.4.

Our analysis was performed with a 3.40GHz Intel Core i7 processor with 31.9GB of RAM. The computations were performed using the R package `FMsmnReg` (Benites *et al.*, 2016), available on CRAN.

4.5.1 Parameter recovery (simulation study 1)

In this section, we consider one scenario for the simulation. Our strategy is to generate artificial samples from model defined in (4.20). We used the ML estimation via EM algorithm for each sample, employing the stopping criterion (20). The average values (Mean) and standard deviations (MC Sd) of the estimates across the 500 Monte Carlo samples were computed. Also, we computed the average (IM SE) values of the approximate standard deviations of the estimates obtained through the method described in Subsection 4.4.4. The results are presented in Tables 4.2 and 4.3. The accuracy of the estimates is very satisfactory for ST and SCN. The estimation method of the standard errors provides relatively close results (IM SE and MC Sd), indicating that the proposed asymptotic approximation of the variances of the ML estimates (equation 4.19) is reliable.

This can also be seen through the coverage probability of parameters (COV), since in general a confidence interval above 95% coverage is maintained for each parameter. The COV is defined as $CP(\hat{\theta}) = (1/m) \sum_{j=1}^m I(\theta \in [\hat{\theta}_L, \hat{\theta}_U])$, where I is the indicator function such that θ lies in the interval $[\hat{\theta}_L, \hat{\theta}_U]$, with $\hat{\theta}_L$ and $\hat{\theta}_U$ being the estimated lower and upper bounds of the 95% CI, respectively. The COV is quite stable for the ST and SCN indicating that the proposed asymptotic approximation of the variance estimates of the ML estimates is reliable. Results for the FM-SN had a similar pattern as those shown in Tables 4.2 and 4.3, so we omit them. Because of the computationally expensive nature of the procedure in the FM-SSL case, we did not consider it. The results suggest that the proposed SMSN-MMRM produces satisfactory estimates and consistency for the approximate standard errors.

j	Measure	Parameters									
		β_{1j}	β_{2j}	β_{3j}	$\sigma_{j,11}$	$\sigma_{j,12}$	$\sigma_{j,22}$	λ_{1j}	λ_{2j}	π	ν
$n = 500$											
1	True	(-1)	(-4)	(-3)	(3)	(1)	(3)	(3)	(5)	(0.3)	(3)
	Mean	-0.9975	-3.9922	-2.9975	3.0188	0.9259	3.0385	2.7437	4.7678	0.2988	3.1450
	MC sd	0.3936	0.2391	0.1237	0.6266	0.4189	0.6232	0.7482	1.2527	0.0206	0.3759
	IM SE	1.2477	0.6186	0.3168	0.0722	0.0848	0.0972	0.9231	1.3421	0.0743	0.5080
	COV	98.2%	97.1%	97.6%	93.1%	92.1%	92.1%	95.8%	95.2%	98.6%	98.8%
2	True	(3)	(7)	(9)	(2)	(1)	(2)	(1)	(4)		
	Mean	3.0419	7.0094	8.8994	2.1289	1.102	2.1460	1.1411	4.2776		
	MC sd	0.1201	0.1331	0.2331	0.3205	0.2788	0.3272	1.2527	0.4458		
	IM SE	0.2925	0.1795	0.1518	0.1319	0.1395	0.2429	3.8053	4.2452		
	COV	97.8%	96.1%	96.4%	94.1%	93.9%	94.2%	98.3%	99.7%		
$n = 1000$											
1	True	(-1)	(-4)	(-3)	(3)	(1)	(3)	(3)	(5)	(0.3)	(3)
	Mean	-0.9785	-3.9963	-2.9992	3.0470	0.9994	3.0856	2.9022	4.9463	0.3002	3.1012
	MC sd	0.2915	0.1709	0.0833	0.4646	0.3251	0.4671	0.6082	0.9591	0.0149	0.2480
	IM SE	0.1924	0.1621	0.0521	0.2190	0.1977	0.2970	0.4145	0.9002	0.0101	0.1234
	COV	94.2%	95.1%	93.6%	92.7%	93.1%	92.7%	94.8%	95.1%	95.0%	94.1%
2	True	(3)	(7)	(9)	(2)	(1)	(2)	(1)	(4)		
	Mean	3.0402	6.997	8.9994	2.0947	1.0807	2.1062	1.1335	4.2398		
	MC sd	0.0883	0.0886	0.1886	0.2295	0.2108	0.2407	0.3602	0.6857		
	IM SE	0.0712	0.0698	0.1698	0.0759	0.1024	0.0381	0.1982	0.5687		
	COV	94.3%	93.8%	94.1%	92.1%	94.6%	92.1%	93.5%	94.7%		
$n = 2500$											
1	True	(-1)	(-4)	(-3)	(3)	(1)	(3)	(3)	(5)	(0.3)	(3)
	Mean	-0.9605	-4.0039	-2.9976	3.1493	1.0967	3.1604	3.1009	5.1239	0.3002	3.1011
	MC sd	0.1764	0.1005	0.0524	0.3061	0.2114	0.3059	0.4645	0.6880	0.0092	0.1650
	IM SE	0.1701	0.1186	0.0511	0.1586	0.1913	0.1587	0.3213	0.5992	0.0101	0.1823
	COV	94.8%	95.0%	94.6%	94.7%	94.2%	93.1%	95.5%	95.4%	95.2%	94.6%
2	True	(3)	(7)	(9)	(2)	(1)	(2)	(1)	(4)		
	Mean	3.0407	6.9991	9.0117	2.0934	1.0918	2.1323	1.1248	4.2458		
	MC sd	0.0495	0.0535	0.0635	0.1401	0.1267	0.1488	0.1955	0.3973		
	IM SE	0.0492	0.0551	0.0687	0.0828	0.0624	0.1195	0.1636	1.2986		
	COV	94.8%	94.6%	95.4%	93.2%	94.7%	93.5%	94.1%	97.1%		

Table 4.2: *Simulation study 1: Mean and MC Sd are the respective mean estimates, standard deviations and percentage of coverage (COV) from fitting a ST-MMRM with different sample size settings based on 500 samples. IM SE is the average value of the approximate standard error obtained through the information-based method in Subsection 4.4.4. True values of parameters are in parentheses.*

j	Measure	Parameters										
		β_{1j}	β_{2j}	β_{3j}	$\sigma_{j,11}$	$\sigma_{j,12}$	$\sigma_{j,22}$	λ_{1j}	λ_{2j}	π	ν	γ
$n = 500$												
1	True	(-1)	(-4)	(-3)	(3)	(1)	(3)	(3)	(5)	(0.3)	(0.1)	(0.1)
	Mean	-1.0349	-4.0124	-2.9891	2.8834	0.8492	2.8752	2.583	4.4876	0.2991	0.1034	0.1026
	MC sd	0.3819	0.2241	0.1168	0.5149	0.3739	0.5235	1.2816	0.3743	0.0206	0.0223	0.0192
	IM SE	0.3112	0.2731	0.0910	0.4923	0.2915	0.4287	0.9214	0.3918	0.0259	0.0284	0.0221
	COV	95.4%	96.0%	93.2%	94.2%	93.7%	94.2%	93.4%	96.2%	95.2%	94.8%	96.6%
2	True	(3)	(7)	(9)	(2)	(1)	(2)	(1)	(4)			
	Mean	2.9936	7.0028	8.0028	1.9939	0.9710	1.9375	0.958	3.8084			
	MC sd	0.0823	0.1201	0.1081	0.2418	0.2223	0.2330	0.3743	0.7303			
	IM SE	0.0656	0.0924	0.0810	0.1724	0.1787	0.1321	2.2423	3.0450			
	COV	94.0%	94.1%	94.4%	93.4%	92.9%	93.2%	98.9%	99.5%			
$n = 1000$												
1	True	(-1)	(-4)	(-3)	(3)	(1)	(3)	(3)	(5)	(0.3)	(0.1)	(0.1)
	Mean	-0.9915	-4.0067	-3.0021	2.9607	0.9339	2.9037	2.7316	4.5552	0.2995	0.1009	0.1000
	MC sd	0.2537	0.1601	0.0793	0.3944	0.2647	0.3738	0.5981	0.9485	0.0143	0.0154	0.0133
	IM SE	0.2713	0.1721	0.0510	0.2843	0.1133	0.1287	0.6212	0.9912	0.0153	0.0131	0.0129
	COV	95.8%	96.3%	92.5%	93.9%	93.1%	92.8%	96.4%	97.1%	95.5%	94.9%	95.6%
2	True	(3)	(7)	(9)	(2)	(1)	(2)	(1)	(4)			
	Mean	2.9992	7.0000	9.0104	2.0086	0.9915	1.9570	0.9758	3.8036			
	MC sd	0.0598	0.0870	0.0741	0.1707	0.1587	0.1523	0.2661	0.4950			
	IM SE	0.0512	0.0724	0.0614	0.1513	0.0787	0.1521	0.2423	1.0450			
	COV	95.3%	94.4%	95.4%	93.6%	92.1%	95.5%	93.3%	98.9%			
$n = 2500$												
1	True	(-1)	(-4)	(-3)	(3)	(1)	(3)	(3)	(5)	(0.3)	(0.1)	(0.1)
	Mean	-0.9838	-3.9966	-3.0046	3.0009	0.9718	2.9485	2.7899	4.599	0.2998	0.0999	0.1001
	MC sd	0.1602	0.0972	0.0488	0.2636	0.1771	0.2263	0.4283	0.5986	0.0090	0.0095	0.0077
	IM SE	0.1632	0.0921	0.0421	0.1243	0.1534	0.1087	0.5225	0.6797	0.0092	0.0101	0.0089
	COV	95.2%	94.8%	94.3%	92.4%	94.1%	92.9%	96.3%	95.8%	95.2%	94.8%	96.2%
2	True	(3)	(7)	(9)	(2)	(1)	(2)	(1)	(4)			
	Mean	2.9966	7.0008	9.0010	1.9945	0.9823	1.9611	0.9598	3.8086			
	MC sd	0.0367	0.0519	0.0441	0.1052	0.0957	0.0967	0.1632	0.3164			
	IM SE	0.0233	0.0559	0.0499	0.0924	0.0799	0.0834	0.1413	0.9163			
	COV	94.4%	95.8%	95.2%	94.2%	93.8%	93.5%	93.1%	97.9%			

Table 4.3: *Simulation study 1: Mean and MC Sd are the respective mean estimates, standard deviations and percentage of coverage (COV) from fitting a SCN-MMRM with different sample size settings based on 500 samples. IM SE is the average value of the approximate standard error obtained through the information-based method in Subsection 4.4.4. True values of parameters are in parentheses.*

4.5.2 Asymptotic properties of the EM estimates (simulation study 2)

Here, the experiment is planned to show the asymptotic properties of the EM estimates. We chosen various samples size settings ($n = 100, 250, 500, 1000$ and 2500). The true values of the parameters in this study are the same considered in simulation study 1. For each combination of parameters and sample sizes, we generated 500 random samples from the SMSN-MMRM. In order to analyze asymptotic properties of the EM estimates, we computed the bias and the relative mean square error (RMSE) for each combination of sample size and parameter values. For a given parameter θ , they are defined by

$$\text{Bias}(\hat{\theta}) = \frac{1}{500} \sum_{j=1}^{500} (\hat{\theta}^{(j)} - \theta),$$

$$\text{RMSE}(\hat{\theta}) = \sqrt{\frac{1}{500} \sum_{j=1}^{500} (\hat{\theta}^{(j)} - \theta)^2},$$

where $\hat{\theta}^{(j)}$ is the estimate of θ for the j -th sample. The results for β_1 , β_2 , Σ_1 , Σ_2 , λ_1 , λ_2 and π_1 are shown in the Figures 4.1, 4.2, 4.3 and 4.4 respectively. When n increases, we can see a pattern of convergence to zero of the bias and RMSE. 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 ST-MMRM model do provide good asymptotic properties.

Figure 4.1: *Simulation study 1 (asymptotic properties). Bias (first row) and RMSE (second row) of the estimates $\beta_1 = (\beta_{11}, \beta_{21}, \beta_{31})^\top$.*

Figure 4.2: *Simulation study 1 (asymptotic properties). Bias (first row) and RMSE (second row) of the estimates $\beta_2 = (\beta_{11}, \beta_{22}, \beta_{32})^\top$.*

Figure 4.3: Simulation study 1 (asymptotic properties). Bias (first column) and RMSE (second column) of the estimates $\lambda_1 = (\lambda_{11}, \lambda_{21})^\top$ and $\lambda_2 = (\lambda_{12}, \lambda_{22})^\top$.

Figure 4.4: Simulation study 1 (asymptotic properties). Bias (first column) and RMSE (second column) of the estimates Σ_1 .

Figure 4.5: Simulation study 1 (Asymptotic properties). Bias (first column) and RMSE (second column) of the estimates Σ_2 .

4.6 Application

We illustrate our proposed methods with a dataset obtained from <http://www.ats.ucla.edu/stat/data/hsb2.csv>, which is available, e.g., in the R package **FMsmnReg**. The hsb2 (high school and beyond) dataset contains 200 observations and 11 variables were randomly sampled from the High School and Beyond survey. The survey was conducted among high school seniors by the National Center for Education Statistics. Lim *et al.* (2017) analyzed this dataset using a mixture multivariate regression models framework, considering the multivariate t distribution. Now we revisit this dataset with the aim of expanding the inferential results to the multivariate SMSN family. Following Lim *et al.* (2017), the explanatory variables included in our proposed model are the reading score (x_1) and the bivariate response variables are the science (y_1) and writing (y_2) scores.

Figure 4.6: *hsb2* dataset. Kernel density estimate of the response variable (a) y_1 , (b) y_2 and explanatory variable (c) x_1 .

In Figures 4.6 (a) and (b) we report the kernel density estimate of y_1 and y_2 . They show that the data have patterns of bimodal distributions. Thus, different models in the SMSN-MMRM class are fitted to these data for $G = 1, 2, 3$. Table 4.4 summarizes the ML estimation results, including the log-likelihood maximum and the AIC, BIC and EDC values. It was not possible to apply the SSL-MMRM with three components, because the algorithm did not converge due to an integration problem at step E. Looking at the AIC, BIC and EDC values, we see that models SN-MMRM, ST-MMRM and SCN-MMRM with $G = 2$ components fit the data better than the same models with $G = 3$ components. In addition, it appears that the SN-MMRM presents a better fit than all other models. Thus, we arrive at the following model for the hsb2 dataset:

$f(\mathbf{y}_i|\boldsymbol{\theta}) = \sum_{j=1}^G \pi_j \phi_{\text{SMSN}_2}(\mathbf{y}_i|\mathbf{X}_i\boldsymbol{\beta}_j + b(\boldsymbol{\nu})\boldsymbol{\Delta}_j, \boldsymbol{\Sigma}_j, \boldsymbol{\lambda}_j, \boldsymbol{\nu})$, where

$$\boldsymbol{\beta}_j = (\beta_{1j}, \beta_{2j})^\top, \quad \boldsymbol{\Sigma}_j = \begin{bmatrix} \sigma_{j.11} & \sigma_{j.12} \\ \sigma_{j.21} & \sigma_{j.22} \end{bmatrix}, \quad \boldsymbol{\lambda}_j = (\lambda_{1j}, \lambda_{2j})^\top, \quad j = 1, 2.$$

Figure 4.7 displays plots of the profile log-likelihood of ν in the case of the ST-MMRM, SSL-MMRM and SCN-MMRM models with $G = 2$ components, which were is used to obtain the initial values of the EM algorithm. Table 4.4 presents the ML estimates of the parameters in the SN-MMRM, ST-MMRM, SSL-MMRM and SCN-MMRM models with $G = 2$ components along with their corresponding approximate standard errors (SE). Although the regression parameter estimates (β_1 and β_2) are similarly small in all models, in general, the standard errors of the model SN-MMRM are smaller than those that obtained for the other models. The significance of the parameters of the models can be evaluated through the asymptotic confidence interval, i.e., estimate $\pm 1.96\text{SE}$, which can be constructed with the results given in Table 4.4.

Figure 4.7: *hsb2* dataset. Profile log-likelihood of ν for models with $G = 2$ components. (a) ST-MMRM, (b) SSL-MMRM and (c) SCN-MMRM model.

Criteria	G	Log-likelihood	AIC	BIC	EDC
SN-MMRM	1	-1471.548	2857.097	2880.185	2862.896
ST-MMRM	1	-1403.938	2823.876	2850.262	2830.503
SCN-MMRM	1	-1403.183	2824.367	2854.051	2831.822
SSL-MMRM	1	-1405.920	2827.839	2854.226	2834.467
SN-MMRM	2	-1352.209	2734.419	2783.893	2746.845
ST-MMRM	2	-1453.949	2939.897	2992.671	2953.152
SCN-MMRM	2	-1431.747	2897.494	2953.566	2911.577
SSL-MMRM	2	-1416.850	2865.699	2918.472	2878.954
SN-MMRM	3	-1359.942	2749.883	2825.745	2768.937
ST-MMRM	3	-1460.450	2968.901	3048.061	2988.783
SCN-MMRM	3	-1458.711	2967.423	3049.881	2988.134

Table 4.4: *hsb2* dataset. Selection criteria for various SMSN-MMRM models.

Parameter	SN-MMRM		ST-MMRM		SSL-MMRM		SCN-MMRM	
	Estimation	SE	Estimation	SE	Estimation	SE	Estimation	SE
β_{10}	57.4976	7.2790	38.2468	5.3257	25.2037	7.8124	44.0423	7.6592
β_{11}	0.0205	0.0480	0.3723	0.0781	0.4194	0.0489	-0.0036	0.0447
β_{20}	14.8794	4.3257	14.6674	4.7687	32.7834	6.2134	37.9924	5.9314
β_{21}	0.6581	0.0520	0.6629	0.0809	0.4223	0.0819	0.3724	0.0784
$\sigma_{1.11}$	62.7730	17.4891	53.3434	18.1651	43.0012	20.1926	36.4339	22.4541
$\sigma_{1.12}$	-18.7771	16.6404	-7.8505	16.3452	-4.3019	8.2139	-7.6228	9.9053
$\sigma_{1.22}$	33.6000	27.0269	21.5366	24.4743	58.5891	25.9519	16.9516	21.9519
$\sigma_{2.11}$	45.0943	16.2038	42.5406	18.2794	64.1300	32.3273	27.0153	22.1316
$\sigma_{2.12}$	-8.4328	19.6661	-2.5524	17.3215	19.5815	15.0325	-7.2294	9.0125
$\sigma_{2.22}$	72.2764	31.3593	47.1879	30.1700	46.3510	18.1282	41.8269	22.2974
λ_{11}	3.3948	3.7438	1.8595	5.8467	-2.6606	2.1882	2.3029	3.4713
λ_{12}	-4.0062	3.0312	-1.9167	2.4910	3.0083	3.8201	-2.5546	2.6691
λ_{21}	-0.3250	1.5327	0.0317	1.7404	2.9220	5.1811	-0.4134	4.0838
λ_{22}	0.9137	3.8142	0.3563	2.8373	-3.0096	3.4115	1.1408	2.7888
p_1	0.3700	0.0698	0.4056	0.1296	0.4147	0.1028	0.5366	0.1238
ν	-	-	22.2542	31.5174	1.7949	0.1318	0.8853	0.1134
γ	-	-	-	-	-	-	0.5341	0.2364

Table 4.5: *hsb2* dataset. Estimation and standard errors (SE) for parameters under the SMSN-MMRM models.

4.7 Conclusions

In this paper we consider a robust approach to finite mixture of multivariate regression models based on multivariate scale mixtures of skew-normal distributions. Our proposed model generalizes the recent works of Basso *et al.* (2010), Liu and Lin (2014), Yao *et al.* (2014) and Zeller *et al.* (2016). This robust regression model simultaneously accommodates asymmetry and heavy tails, thus allowing practitioners from different areas to analyze data in an extremely flexible way. An ECME algorithm is developed by exploring the statistical properties of the class considered, which can be easily implemented and coded with existing statistical software such as the R package, FMsmnReg, Benites *et al.* (2016), available on CRAN, 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.

Chapter 5

Conclusion and future studies

5.1 Final considerations

In this dissertation, we present a frequentist perspective of linear models under the class of scale mixtures of skew normal (SMSN) distributions, both univariate and multivariate. We also consider a model for censored data in the class of univariate scale mixture of normal (SMN) distributions, which is the symmetric case of the SMSN family. The SMSN distribution family is characterized by its flexibility to deal with both censored and uncensored data in the presence of multimodality, asymmetry and outlying observations. The SMSN distributions have as special cases the distributions based on the skew-normal, skew Student- t , skew slash and skew normal contaminated models, among others.

In the first proposed model, a linear regression model where the error follows a finite mixture of SMSN distribution class, is denoted by FM-SMSN-LR. This is a generalization of the work presented by Bartolucci and Scaccia (2005). An EM algorithm was developed to calculate maximum likelihood estimates of the FM-SMSN-LR model. This proposal allows modeling the data with great flexibility, accommodating simultaneously multimodality, asymmetry and heavy tails. It should be noted that our proposal differs from the work of Zeller *et al.* (2016), where they consider a finite mixture of regressions.

A censored linear regression model was also proposed where the random errors are based on a finite mixture of the SMN distribution. Our proposal is an extension of the work developed by Arellano-Valle *et al.* (2012) and Garay *et al.* (2017), where the errors follow the normal and Student- t distributions, respectively. It is also an extension of the model proposed by Caudill (2012), which considers a regression model for censored data where the error follows a finite mixture of normal distributions. This new class of distributions allows obtaining better estimates of the regression co-

efficients when the error distribution is characterized by the presence of multimodality, asymmetry and outlying observations, depending on the structure of the mixture of components.

Finally, we consider a finite mixture of multivariate regressions based on scale mixture of skew normal distributions. Our proposal here is an extension of the works of Basso *et al.* (2010), Liu and Lin (2014), Yao *et al.* (2014) and Zeller *et al.* (2016). This robust regression model allows working with multivariate data while accommodating simultaneously asymmetry and heavy tails, which allows modeling data with great flexibility by researchers in many areas.

All the models developed have been included in the packages FMsmnReg and CensMixReg, which are described in the appendix and can be freely installed.

5.2 Suggestions for future research

Several research topics can be derived and/or directed from the results of this work, among which we can mention the following:

1. To extend the estimation results obtained in the model presented in the first paper (FM-SMSN-LR) where the error follows a finite mixture of a multivariate of SMSN distributions, inspired by the works of Soffritti and Galimberti (2011) and Galimberti and Soffritti (2014).
2. To extend the work proposed in the second paper presented in this thesis, considering an asymmetric class, as developed by Basso *et al.* (2010) and Lee and Scott (2012) for censoring models.
3. For all proposed models, it would be interesting to consider models from a Bayesian perspective.
4. The methods proposed in this dissertation can also be applied and extended to the context of random effect models.

Appendix A

Chapter 2 - Ordinary residuals and QQ-plots

Figure A.1: AIS dataset. Histogram of ordinary residuals with FM-SMSN-LR residual density model for a) $g = 3$ and b) $g = 4$ components.

Figure A.2: AIS dataset. QQ-plots and simulated envelopes for the residual $(y - \hat{y})$ for $g = 2$ components based on FM-SN (left), FM-SCN (middle) and FM-SSL (right) distributions, respectively.

Appendix B

Chapter 4 - The observed information matrix

In this appendix we reparameterize $\Sigma = \mathbf{D}^2$ for ease computation of the above derivatives, recalling that $\mathbf{A}_{ij} = \boldsymbol{\lambda}_j^\top \Sigma_j^{-1/2} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}_j - b(\boldsymbol{\nu}_j) \boldsymbol{\Delta}_j)$ and $\mathbf{d}_{ij} = (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}_j - b(\boldsymbol{\nu}_j) \boldsymbol{\Delta}_j)^\top \Sigma_j^{-1} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}_j - b(\boldsymbol{\nu}_j) \boldsymbol{\Delta}_j)$:

- $\mathbf{d}_{ij}$

$$\begin{aligned} \frac{\partial \mathbf{d}_{ij}}{\partial \boldsymbol{\beta}_j} &= -2\mathbf{D}^{-2}(\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}_j - b(\boldsymbol{\nu}_j) \boldsymbol{\Delta}_j) \\ \frac{\partial \mathbf{d}_{ij}}{\partial \boldsymbol{\alpha}_{jk}} &= -(\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}_j - b(\boldsymbol{\nu}_j) \boldsymbol{\Delta}_j)^\top \mathbf{D}_j^{-1} (\dot{D}_{jk} \mathbf{D}_j^{-1} + \mathbf{D}_j^{-1} \dot{D}_{jk}) \mathbf{D}_j^{-1} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}_j - b(\boldsymbol{\nu}_j) \boldsymbol{\Delta}_j) \end{aligned}$$

- $\mathbf{A}_{ij}$

$$\begin{aligned} \frac{\partial \mathbf{A}_{ij}}{\partial \boldsymbol{\beta}_j} &= -\mathbf{X}_i^\top \mathbf{D}_j^{-1} \boldsymbol{\lambda}_j \\ \frac{\partial \mathbf{A}_{ij}}{\partial \boldsymbol{\lambda}_j} &= \mathbf{D}_j^{-1} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}_j - b(\boldsymbol{\nu}_j) \boldsymbol{\Delta}_j) \\ \frac{\partial \mathbf{A}_{ij}}{\partial \boldsymbol{\alpha}_{jk}} &= -\boldsymbol{\lambda}_j \mathbf{D}_j^{-1} \dot{D}_{jk} \mathbf{D}_j^{-1} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}_j - b(\boldsymbol{\nu}_j) \boldsymbol{\Delta}_j) \end{aligned}$$

- $\mathbf{D}_j$

$$\frac{\partial |\mathbf{D}_j^2|^{-1/2}}{\partial \boldsymbol{\alpha}_{jk}} = \frac{-1}{|\mathbf{D}_j|^2} \frac{\partial |\mathbf{D}_j|}{\partial \boldsymbol{\alpha}_{jk}},$$

where $\dot{D}_{jk} = \frac{\partial \mathbf{D}_j}{\partial \boldsymbol{\alpha}_{jk}}$.

Appendix C

“FMsmnsnReg” Package manual

Initially, we will install the package FMsmnsnReg and for this we will use the following code:

```
install.packages("FMsmnsnReg")
```

Then the package can be loaded from the following command:

```
library("FMsmnsnReg")
```

An important comment we can mention is that the package was tested and installed only on machines with Windows operating system. While it is possible for the command set to work in an environment Linux has not been tested on this operating system. After the package has been loaded to access the manual so that you can get more information about the available data set, you must

```
help(package = "FMsmnsnReg")
```

Thus, the estimation methodology proposed here in this thesis becomes available. Even all databases can also be reused for further analysis. For example, to load the data of the second chapter on Australian Institute of Sport (AIS) and soon we will leave the variables available by their names

```
data(ais)
attach(ais)
```

The FM-SMSN-LR can be adjusted using the following code

```
x1    <- cbind(1, SSF, Ht)
y     <- Bfat
```

```
parST <- FMsmnsnReg(y, x1, g=2, get.init = TRUE, criteria = TRUE,
  group = FALSE, family = "Skew.t", error = 10^-4,
  iter.max = 2000, obs.prob= FALSE, kmeans.param = NULL,
  show.converge=FALSE, cp=0.5)
```

The output obtained is as follows

```
-----
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

For more information on the package arguments, you can look at the available CRAN manual.

Appendix D

“CensMixReg” Package manual

To install and load the package we do

```
install.packages("CensMixReg")  
library("CensMixReg")
```

To load the data set and set the response variable and the explanatory variables, let's see the following codes

```
data(wage.rates)  
  
y      <- wage.rates$wage  
x1     <- cbind(1,wage.rates$age,wage.rates$educ,wage.rates$hours/1000)  
cc     <- c(rep(0,428),rep(1,325))
```

Here we will first get the initial values before doing the model adjustment and for this we use:

```
initial    <- initial.values.fm.smn.cr(cc, y,x1,g=2,algorithm="k-means"  
,family="T",lower=1,upper=20,space=1,plotLog = TRUE,searchNU=TRUE,  
printNU=FALSE, saveFigure = FALSE)
```

Now we adjust the model

```
fitT <- fm.smn.cr(cc, y, x1, Abetas = initial$Abetas, medj = initial$medj  
, sigma2 = initial$sigma2, pii = initial$pii, nu=initial$nu, g = 2,  
family = "T", error = 0.00001, iter.max = 500)
```


Bibliography

- Aitken(1927)** A.C. Aitken. On Bernoulli's numerical solution of algebraic equations. *Proceedings of the Royal Society of Edinburgh*, 46:289–305. Cited in page 16, 72
- Andrews and Mallows(1974a)** D. F. Andrews and C. L. Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society, Series B*, 36:99–102. Cited in page 30
- Andrews and Mallows(1974b)** D. F. Andrews and C. L. Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society, Series B*, pages 99–102. Cited in page 6, 7
- Arellano-Valle et al.(2012)** R. Arellano-Valle, L. Castro, G. González-Farías and K. Muñoz-Gajardo. Student-t censored regression model: properties and inference. *Statistical Methods & Applications*, 21:453–473. Cited in page 2, 3, 30, 31, 55, 57
- Arellano-Valle and Azzalini(2006)** R. B. Arellano-Valle and A. Azzalini. On the unification of families of skew-normal distributions. *Scandinavian Journal of Statistics*, 33:561–574. Cited in page 61
- Arnold and Press(1989)** B. C. Arnold and S. J. Press. Bayesian estimation and prediction for Pareto data. *Journal of the American Statistical Association*, 84:1079–1084. Cited in page 2
- Ateya(2012)** S. F. Ateya. Maximum likelihood estimation under a finite mixture of generalized exponential distributions based on censored data. *Statistical Papers*, 55(2):311–325. Cited in page 38
- Atienza et al.(2006)** N. Atienza, J. Garcia-Heras and J. M. Muñoz-Pichardo. A new condition for identifiability of finite mixture distributions. *Metrika*, 63(2):215–221. Cited in page 68
- Azzalini(2005)** A. Azzalini. The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics*, 32:159–188. Cited in page 61
- Azzalini(1985)** A. Azzalini. A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 12:171–178. Cited in page 6, 8
- Azzalini and Genton(2008)** A. Azzalini and M. Genton. Robust likelihood methods based on the skew-t and related distributions. *International Statistical Review*, 76:1490–1507. Cited in page 6

- Bartolucci and Scaccia(2005)** F. Bartolucci and L. Scaccia. The use of mixtures for dealing with non-normal regression errors. *Computational Statistics & Data Analysis*, 48(4): 821–834. Cited in page 1, 5, 6, 7, 16, 43, 87
- Basford et al.(1997a)** K. E. Basford, D. R. Greenway, G. J. McLachlan and D. Peel. Standard errors of fitted component means of normal mixtures. *Computational Statistics*, 12: 1–18. Cited in page 2, 17, 45
- Basford et al.(1997b)** K. E. Basford, D. R. Greenway, G. J. McLachlan and D. Peel. Standard errors of fitted component means of normal mixtures. *Computational Statistics*, 12: 1–17. Cited in page 73
- Basso et al.(2010)** R. M. Basso, V. H. Lachos, C. R. B. Cabral and P. Ghosh. Robust mixture modeling based on scale mixtures of skew-normal distributions. *Computational Statistics & Data Analysis*, 54:2926–2941. Cited in page 9, 10, 18, 85, 88
- Benites and Lachos(2015)** L. Benites and V. H. Lachos. *CensMixReg: Censored Linear Mixture Regression Models*, 2015. URL <http://CRAN.R-project.org/package=CensMixReg>. R package version 0.7. Cited in page 44, 49
- Benites et al.(2016)** L. Benites, R. M. and V. H. Lachos. *FMsmnReg: Regression Models with Finite Mixtures of Skew Heavy-Tailed Errors*, 2016. URL <http://CRAN.R-project.org/package=FMsmnReg>. R package version 1.0. Cited in page 18, 71, 76, 85
- Böhning et al.(1994)** D. Böhning, E. Dietz, R. Schaub, P. Schlattmann and B. G. Lindsay. The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics*, 46(2):373–388. Cited in page 16, 73
- Branco and Dey(2001)** M. D. Branco and D. K. Dey. A general class of multivariate skew-elliptical distributions. *Journal of Multivariate Analysis*, 79:99–113. Cited in page 6, 61
- Cabral et al.(2012)** C. R. B. Cabral, V. H. Lachos and M. O. Prates. Multivariate mixture modeling using skew-normal independent distributions. *Computational Statistics & Data Analysis*, 56:126–142. Cited in page 28, 61
- Caudill(2012)** Steven B Caudill. A partially adaptive estimator for the censored regression model based on a mixture of normal distributions. *Statistical Methods & Applications*, 21:121–137. Cited in page 2, 28, 31, 57, 87
- Chib(1992)** S. Chib. Bayes inference in the Tobit censored regression model. *Journal of Econometrics*, 51:79–99. Cited in page 30
- Ciarleglio and Ogden(2016)** Adam Ciarleglio and R. Todd Ogden. Wavelet-based scalar-on-function finite mixture regression models. *Computational Statistics & Data Analysis*, 93:86 – 96. ISSN 0167-9473. doi: <http://dx.doi.org/10.1016/j.csda.2014.11.017>. URL <http://www.sciencedirect.com/science/article/pii/S0167947314003454>. Cited in page 60
- Colosimo and Giolo(2006)** E. Colosimo and S. Giolo. *Análise de sobrevivência aplicada*. ABE - Projeto Fisher, 1 edição. Cited in page 2

- Cook and Weisberg(1982)** R. D. Cook and S. Weisberg. *Residuals and Influence in Regression*. Chapman & Hall/CRC, Boca Raton, FL. Cited in page 22
- Dempster et al.(1977)** A. Dempster, N. Laird and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38. Cited in page 7, 65, 68
- Devijver(2015)** E. Devijver. Finite mixture regression: A sparse variable selection by model selection for clustering. *Electronic Journal of Statistics*, 9:2642–2674. Cited in page 60
- Dias and Wedel(2004)** J. G. Dias and M. Wedel. An empirical comparison of EM, SEM and MCMC performance for problematic gaussian mixture likelihoods. *Statistics and Computing*, 14:323–332. Cited in page 73
- Efron and Tibshirani(1986)** B. Efron and R. Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1: 54–75. Cited in page 75
- Ferraro et al.(2017)** M. B. Ferraro, P. Giordani, B. Vantaggi, M. Gagolewski, M. A. Gil, P. Grzegorzewski and O. Hryniewicz. *Soft Methods for Data Science*. Advances in Intelligent Systems and Computing 456. Springer International Publishing, 1 edição. Cited in page 1, 60
- Frühwirth-Schnatter(2006)** S. Frühwirth-Schnatter. *Finite mixture and markov switching models*. Springer, New York. Cited in page 1
- Galea et al.(1997)** Manuel Galea, Gilberto A Paula and Heleno Bolfarine. Local influence in elliptical linear regression models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46(1):71–79. Cited in page 6
- Galimberti and Soffritti(2014)** G. Galimberti and G. Soffritti. A multivariate linear regression analysis using finite mixtures of t distributions. *Computational Statistics & Data Analysis*, 71:138–150. Cited in page 28, 88
- Garay et al.(2015)** Aldo M Garay, Victor H Lachos, Heleno Bolfarine and Celso RB Cabral. Linear censored regression models with scale mixtures of normal distributions. *Statistical Papers*. doi: 10.1007/s00362-015-0696-9. Cited in page 30, 31, 32, 55, 57
- Garay et al.(2017)** Aldo M. Garay, Victor H. Lachos, Heleno Bolfarine and Celso R. B. Cabral. Linear censored regression models with scale mixtures of normal distributions. *Statistical Papers*, 58(1):247–278. Cited in page 2
- Greene(2012)** W. Greene. *Econometric Analysis*. Prentice Hall. Cited in page 2
- Grün and Leisch(2008)** Bettina Grün and Friedrich Leisch. *Finite Mixtures of Generalized Linear Regression Models*, pages 205–230. Physica-Verlag HD, Heidelberg. Cited in page 68
- Hennig(2000)** C. Hennig. Identifiability of models for clusterwise linear regression. *Journal of Classification*, 17(2):273–296. Cited in page 68

- Holzmamn et al.(2006)** Hajo Holzmamn, Axel Munk and Tilmann Gneiting. Identifiability of finite mixtures of elliptical distributions. *Scandinavian Journal of Statistics*, 33(4): 753–763. Cited in page 6, 38, 68
- Hurvich and Tsai(1989)** C.M. Hurvich and C.-L Tsai. Regression and time series model selection in small samples. *Biometrika*, 76:297–307. Cited in page 16
- Jedidi(1996)** K. Jedidi. On estimating finite mixture of multivariate regression and simultaneous equation models. *Structural equation modeling*, 3(3):266–289. Cited in page 60
- Kalbfleisch and Lawless(1992)** J. D. Kalbfleisch and J. F. Lawless. Some useful statistical methods for truncated data. *Journal of Quality Technology*, 24:145–152. Cited in page 2
- Karlsson and Laitila(2014)** M. Karlsson and T. Laitila. Finite mixture modeling of censored regression models. *Statistical Papers*, 55:627–642. Cited in page 2, 28, 31, 57
- Lachos et al.(2010a)** V. H. Lachos, P. Ghosh and R. B. Arellano-Valle. Likelihood based inference for skew-normal independent linear mixed models. *Statistica Sinica*, 20. SS-08-045. Cited in page 61, 62, 63
- Lachos et al.(2010b)** V. H. Lachos, F. V. Labra, H. Bolfarine and P. Ghosh. Multivariate measurement error models based on scale mixtures of the skew-normal distribution. *Statistics*, 44:541–556. Cited in page 61
- Lange and Sinsheimer(1993)** K. L. Lange and J. S. Sinsheimer. Normal/independent distributions and their applications in robust regression. *Journal of Computational and Graphical Statistics*, 2:175–198. Cited in page 6
- Lange et al.(1989)** K. L. Lange, R. Little and J. Taylor. Robust statistical modeling using t distribution. *Journal of the American Statistical Association*, 84:881–896. Cited in page 6
- Lee and Scott(2012)** Gyemin Lee and Clayton Scott. Em algorithms for multivariate gaussian mixture models with truncated and censored data. *Computational Statistics & Data Analysis*, 56(9):2816–2829. Cited in page 88
- Lim et al.(2017)** H. K. Lim, N. N. Narisetty and S. Cheon. Robust multivariate mixture regression models with incompleteness. *Journal of Statistical Computation and Simulation*, 87(2): 328–347. Cited in page 2, 60, 83
- Lindsay(1995)** B.G. Lindsay. *Mixture models: Theory, geometry and applications*. Institute of Mathematical Statistics, Hayward, CA. Cited in page 16
- Liu and Rubin(1994)** C. Liu and D. B. Rubin. The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika*, 80:267–278. Cited in page 36, 67
- Liu and Lin(2014)** Min Liu and Tsung-I Lin. A skew-normal mixture regression model. *Educational and Psychological Measurement*, 74(1):139–162. Cited in page 85, 88

- Lubke and Grimm(2017)** Gitta Lubke and Kevin J. Grimm. Novel approaches in mixture modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 24(2):157–158. doi: 10.1080/10705511.2016.1258314. URL <http://dx.doi.org/10.1080/10705511.2016.1258314>. Cited in page 60
- Massuia et al.(2015)** M. B. Massuia, C. R. B. Cabral, L. A. Matos and V. H. Lachos. Influence diagnostics for Student-t censored linear regression models. *Statistics*, 49:1074–1094. Cited in page 30, 31, 55
- Matos et al.(2013)** Larissa A. Matos, Marcos O. Prates, Ming-Hui Chen and Victor H. Lachos. Likelihood-based inference for mixed-effects models with censored response using the multivariate-t distribution. *Statistica Sinica*, 23(3):1323–1345. Cited in page 3
- McLachlan and Krishnan(2008)** G. J. McLachlan and T. Krishnan. *The EM algorithm and extensions*. John Wiley & Sons, New Jersey. Cited in page 45
- McLachlan and Peel(2000)** G. J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, New York. Cited in page 1, 44, 69, 71
- McNicholas(2017)** P. D. McNicholas. *Mixture model-based classification*. Chapman & Hall/CRC, Boca Raton. Cited in page 1, 60
- McNicholas and Murphy(2008)** P. D. McNicholas and T. B. Murphy. Parsimonious Gaussian mixture models. *Statistics and Computing*, 18:285–296. Cited in page 44, 72
- McNicholas et al.(2010)** P.D. McNicholas, T.B. Murphy, A.F. McDaid and Fros. D. Serial and parallel implementations of model-based clustering via parsimonious gaussian mixture models. *Computational Statistics & Data Analysis*, 54(3):711–723. Cited in page 16
- Meng(1994)** X.-L. Meng. On the global and componentswise rates of convergence of the EM algorithm. *Linear Algebra and its Applications*, 199:413–425. Cited in page 56
- Meng and Rubin(1993)** X.-L. Meng and D. B. Rubin. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 81:633–648. Cited in page 12, 36
- Mroz(1987)** T. A. Mroz. The sensitivity of an empirical model of married women’s hours of work to economic and statistical assumptions. *Econometrica: Journal of the Econometric Society*, pages 765–799. Cited in page 2, 30, 53
- Nelson(1990)** W. Nelson. Hazard plotting of left truncated life data. *Journal of Quality Technology*, 22:230–238. Cited in page 2
- Osorio et al.(2007)** F. Osorio, G. A. Paula and M. Galea. Assessment of local influence in elliptical linear models with longitudinal structure. *Computational Statistics and Data Analysis*, 51:4354–4368. Cited in page 35, 42
- Otiniano et al.(2015)** C. E. G. Otiniano, P. N. Rathie and L. C. S. M. Ozelim. On the identifiability of finite mixture of skew-normal and skew-t distributions. *Statistics & Probability Letters*, 106:103–108. Cited in page 6, 68

- Park et al.(2007)** J. Park, M. Genton and S. Ghosh. Censored time series analysis with autoregressive moving average models. *Canadian Journal of Statistics*, 35:151 – 168. Cited in page 3
- Prates et al.(2013)** M. O. Prates, V. H. Lachos and C. R. B. Cabral. mixsmsn: Fitting finite mixture of scale mixture of skew-normal distributions. *Journal of Statistical Software*, 54(12):1–20. Cited in page 16
- Pyne et al.(2016)** S. Pyne, B. L. S. Prakasa Rao and S. B. Rao (eds.). *Big Data Analytics: Methods and Applications*. Springer India, 1 edição. ISBN 978-81-322-3626-9,978-81-322-3628-3. Cited in page 1, 60
- R Core Team(2015)** R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL <http://www.R-project.org/>. Cited in page 18, 49
- R Development Core Team(2011)** R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. URL <http://www.R-project.org>. ISBN 3-900051-07-0. Cited in page 64
- Richard(1996)** B. Richard. *Regression models censored, sample-selected, or truncated data /*. Sage university papers series. Quantitative applications in the social sciences ;. Sage Publications,. Cited in page 3
- Schwarz(1978)** G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6: 461–464. Cited in page 44, 72
- Sclove(1987)** S.L. Sclove. Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, 52(3):333–343. Cited in page 16
- Soffritti and Galimberti(2011)** Gabriele Soffritti and Giuliano Galimberti. Multivariate linear regression with non-normal errors: a solution based on mixture models. *Statistics and Computing*, 21(4):523–536. Cited in page 6, 28, 88
- Song et al.(2014)** Weixing Song, Weixin Yao and Yanru Xing. Robust mixture regression model fitting by laplace distribution. *Computational Statistics & Data Analysis*, 71:128 – 137. ISSN 0167-9473. doi: <http://dx.doi.org/10.1016/j.csda.2013.06.022>. URL <http://www.sciencedirect.com/science/article/pii/S0167947313002442>. Cited in page 60
- Tatarinova and Schumitzky(2015)** T. Tatarinova and A. Schumitzky. *Nonlinear Mixture Models : A Bayesian Approach*. Imperial College Press. Cited in page 1, 60
- Teicher(1961)** H. Teicher. Identifiability of mixtures. *The Annals of Mathematical Statistics*, 33: 244–248. Cited in page 68
- Teicher(1963)** H. Teicher. Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 34:1265–1269. Cited in page 68
- Teicher(1967)** H. Teicher. Identifiability of product measures. *The Annals of Mathematical Statistics*, 38:1300–1302. Cited in page 38, 68

- Thompson and Nelson(2003)** M. L. Thompson and K. P. Nelson. Linear regression with type i interval- and left-censored response data. *Environmental and Ecological Statistics*, 10(2):221–230. Cited in page 3
- Turner(2000)** T. R. Turner. Estimating the propagation rate of a viral infection of potato plants via mixtures of regressions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 49(3):371–384. Cited in page 2, 8, 25
- Vaida and Liu(2012)** F. Vaida and L. Liu. Fast implementation for normal mixed effects models with censored response. *Journal of Computational and Graphical Statistics*, 18:797–817. Cited in page 3
- Wang and Genton(2006)** J. Wang and M. G. Genton. The multivariate skew-slash distribution. *Journal of Statistical Planning and Inference*, 136:209–220. Cited in page 64
- Watson(1944)** G. N. Watson. *A Treatise on the Theory of Bessel Functions 2nd edn.* Cambridge University Press, Cambridge. Cited in page 39
- Wei and Tanner(1990)** G. C. G. Wei and M. A. Tanner. A Monte Carlo implementation of the EM algorithm and the poor man’s data augmentation algorithms. *Journal of the American Statistical Association*, 85:699–704. Cited in page 30
- Wu(2010)** L. Wu. *Mixed Effects Models for Complex Data.* Chapman & Hall/CRC, Boca Raton, FL. Cited in page 30
- Yakowitz and Spragins(1968)** S. J. Yakowitz and J. D. Spragins. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 39:209–214. Cited in page 38, 68
- Yao et al.(2014)** W. Yao, Y. Wei and C. Yu. Robust mixture regression using the t-distribution. *Computational Statistics & Data Analysis*, 71:116–127. Cited in page 85, 88
- Yoichi(2011)** M. Yoichi. Maximum likelihood estimators in finite mixture models with censored data. *Journal of Statistical Planning and Inference*, 141(1):56–64. Cited in page 38
- Zeller et al.(2016)** C. B. Zeller, C. R. B. Cabral and V. H. Lachos. Robust mixture regression modeling based on scale mixtures of skew-normal distributions. *TEST*, 25(2):375–396. Cited in page 2, 7, 28, 60, 85, 87, 88