Robust Regression Modeling for Censored Data based on Mixtures of Student-t Distributions

Víctor Hugo Lachos[a] Luis Benites Sanchez[b] and Celso Rômulo Barbosa Cabral[c*]

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

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

[c]*Departamento de Estatística, Universidade Federal do Amazonas, Brazil*

Abstract

In the framework of censored regression models, the distribution of the error terms departs significantly from normality, for instance, in the presence of heavy tails, skewness and/or atypical observations. In this paper we extend the censored linear regression model with normal errors to the case where the random errors follow a finite mixture of Student-t distributions. This approach allows us to model data with great flexibility, accommodating multimodality, heavy tails and also skewness depending on the structure of the mixture components. We develop an analytically simple and efficient EM-type algorithm for iteratively computing maximum likelihood estimates of the parameters, with standard errors as a by-product. The algorithm has closed-form expressions at the E-step, that rely on formulas for the mean and variance of the truncated Student-t distributions. The efficacy of the method is verified through the analysis of simulated datasets and modeling a censored real dataset first analyzed under normal and Student-t errors. The proposed algorithm and methods are implemented in the R package *CensMixReg()*.

Keywords: Censored regression model, EM-type algorithms, Finite mixture models, Heavy-tails, Tobit model.

1 Introduction

The problem of estimation of regression models where the dependent variable is censored has been studied in different fields, such as econometric analysis and clinical testing, among many others. For example, in econometrics, the study of the labor force participation of married women

*Correspondence to: Departamento de Estatística, ICE, Universidade Federal do Amazonas. Av. Rodrigo Otávio, 6200, Campus Universitário Arthur Virgílio Filho, Coroado I, CEP 69077-000, Manaus, Amazonas, Brazil. E-mail address: celsoromulo@ufam.edu.br (C.R.B. Cabral)

is usually conducted under the censored Tobit model (see, for instance, Chib, 1992). In this case, the observed response is the wage rate, which is typically considered as censored below zero, i.e., for working women, positive values for the wage rates are registered, whereas for non-working women, the observed wage rate is zero (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 reasons, it is assumed that the random errors have a normal distribution (Wei & Tanner, 1990). However, it is well-known that several phenomena are not always in agreement with this assumption, yielding data with a distribution with heavier 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 transformations, yet presents a robust Gaussian framework.

Many extensions of this classic Gaussian censored regression (N-CR) 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 (tCR), including the implementation of an interesting (and simple) EM (expectation-maximization) algorithm for maximum likelihood (ML) estimation. They demonstrated its robustness aspects against outliers through extensive simulations. A CR model based on the scale mixture of normal distributions (Andrews & Mallows, 1974) has been recently proposed by Garay *et al.* (2015) to estimate the regression parameters robustly, where a simple and efficient EM-type algorithm for iteratively computing ML estimates of the parameters is also presented. Moreover, the proposed algorithm was implemented in the R package *SMNCensReg()*. A drawback of these recent proposals is that they are not appropriate when the data present, for instance, multimodality, heavy tails and skewness, simultaneously.

In the context of finite mixture of censored regression (CR) models, Karlsson & Laitila (2014) (see also, Caudill, 2012) illustrated the use of mixtures of normal distributions with a finite number of components (FM-CR model), which can represent a wide variety of density shapes (Marron & Wand, 1992), including skewness and multimodality. This proposition is doubtlessly very flexible, but there can still be problems related to the simultaneous occurrence of skewness, discrepant observations and multimodality. Even when modeling using normal mixtures, overestimation can occur 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. Thus in this article we propose a robust mixture model for the random errors based on the Student-t distribution (FM-tCR model) by extending the mixture of normal mixtures proposed by Caudill (2012) and Karlsson & Laitila (2014). More specifically, our objectives are: (i) to propose a mixture censored regression model (and associated likelihood inference) based on the mixtures of Student-t distribution, extending the recent works of Arellano-Valle *et al.* (2012), Caudill (2012), Karlsson & Laitila (2014), Massuia *et al.* (2015) and Garay *et al.* (2015); (ii) to implement and evaluate the proposed method computationally; and (iii) to apply these results to the analysis of a real life dataset.

The remainder of the paper is organized as follows. In Section 2, we briefly discuss the truncated Student-t distribution and some of its properties. In addition, we present the tCR model proposed by Massuia *et al.* (2015) and the related ML estimation. In Section 3, we present the robust FM-tCR 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.

2 The Student-t censored regression model

2.1 Preliminaries

Before talking about the censored regression model, for the sake of completeness, we give a brief introduction of the truncated Student-t distribution. In the following definitions, $\mathrm{N}(\mu,\sigma^2)$ denotes the normal distribution with mean μ and variance σ^2, $\mathrm{Gamma}(c,d)$ denotes the gamma distribution with mean c/d and variance c/d^2 and $Z\perp U$ denotes independent random variables Z and U. Also, $\stackrel{d}{=}$ means "has the same distribution as". First, we give the classic definition of the Student-t distribution as a scale mixture of the normal distribution.

We say that a random variable X has a Student-t distribution with location parameter $\mu\in\mathbb{R}$, scale parameter $\sigma^2\in(0,\infty)$ and $\nu\in(0,\infty)$ degrees of freedom, denoted by $X\sim\ \mathrm{t}_\nu(\mu,\sigma^2)$, if it has the following representation:

$$X\stackrel{d}{=}\mu+U^{-1/2}Z, \tag{2.1}$$

where $Z\sim\ \mathrm{N}(0,\sigma^2)$, $U\sim\ \mathrm{Gamma}(\nu/2,\nu/2)$ and $Z\perp U$.

Let $X\sim\ \mathrm{t}_\nu(\mu,\sigma^2)$. A random variable Y has a truncated Student-t distribution in the interval (a,b) if $Y\stackrel{d}{=}X|(X\in(a,b))$. In this case we write $Y\sim\ \mathrm{Tt}_\nu(\mu,\sigma^2;(a,b))$. It is straightforward to prove that the density of Y is given by

$$\mathrm{Tt}_\nu(y|\mu,\sigma^2;(a,b))=\ \mathrm{t}_\nu(y|\mu,\sigma^2)\left[\mathscr{T}_\nu\left(\frac{b-\mu}{\sigma}\right)-\mathscr{T}_\nu\left(\frac{a-\mu}{\sigma}\right)\right]^{-1},\quad y\in(a,b),$$

where $\mathrm{t}_\nu(\cdot|\mu,\sigma^2)$ denotes the density of the Student-t distribution and $\mathscr{T}_\nu(\cdot)$ denotes the distribution function of the standard Student-t distribution with ν degrees of freedom (that is, with $\mu=0$

and $\sigma^2 = 1$).

The following result is very important for our subsequent exposition. It was provided by Genç (2013) (see also Kim, 2008) and presents the first two moments of the truncated Student-t distribution. $\Gamma(\cdot)$ denotes the gamma function.

Lemma 1. *If* $Y \sim \mathrm{Tt}_\nu(\mu, \sigma^2; (a,b))$, *then*

$$\begin{aligned}
\mathrm{E}[Y] &= \mu + G(\nu)\left\{(\nu+\alpha^2)^{-(\nu-1)/2} - (\nu+\beta^2)^{-(\nu-1)/2}\right\}\sigma,\ \nu > 1,\\
\mathrm{E}[Y^2] &= \mu^2 + \sigma^2\left\{A(\nu) + G(\nu)\left[\alpha(\nu+\alpha^2)^{-(\nu-1)/2} - \beta(\nu+\beta^2)^{-(\nu-1)/2}\right]\right\}\\
&\quad +2\mu\sigma G(\nu)\left\{(\nu+\alpha^2)^{-(\nu-1)/2} - (\nu+\beta^2)^{-(\nu-1)/2}\right\}, \nu > 2,
\end{aligned}$$

where $A(\nu) = \left(\frac{\nu}{\nu-2}\right)\frac{\mathscr{T}_\nu(\beta^*) - \mathscr{T}_\nu(\alpha^*)}{\mathscr{T}_\nu(\beta) - \mathscr{T}_\nu(\alpha)}$, $G(\nu) = \frac{\Gamma((\nu-1)/2)\nu^{\nu/2}}{2(\mathscr{T}_\nu(\beta) - \mathscr{T}_\nu(\alpha))\Gamma(\nu/2)\Gamma(1/2)}$, $\alpha = \frac{a-\mu}{\sigma}$, $\beta = \frac{b-\mu}{\sigma}$, $\alpha^* = \frac{\alpha}{\sqrt{(\nu-2)/\nu}}$, $\beta^* = \frac{\beta}{\sqrt{(\nu-2)/\nu}}$.

The following result will be useful for the implementation of the EM algorithm (see Section 2.2). The proof can be found in Massuia *et al.* (2015).

Lemma 2. *Let* $Y \sim \mathrm{Tt}_\nu(\mu, \sigma^2; (a,b))$, $d^2(\mu, \sigma^2, Y) = (Y-\mu)^2/\sigma^2$. *Then, for* $k = 0, 1, 2$ *and for* $r = 1, 2$,

$$\begin{aligned}
\mathrm{E}\left[\left(\frac{\nu+1}{\nu + d^2(\mu,\sigma^2,Y)}\right)^r Y^k\right] &= c(\nu,r)\mathrm{E}[X^k]\left[\mathscr{T}_{\nu+2r}\left(\frac{b-\mu}{\sigma^*}\right) - \mathscr{T}_{\nu+2r}\left(\frac{a-\mu}{\sigma^*}\right)\right]\\
&\quad \times\left[\mathscr{T}_\nu\left(\frac{b-\mu}{\sigma}\right) - \mathscr{T}_\nu\left(\frac{a-\mu}{\sigma}\right)\right]^{-1},
\end{aligned}$$

where

$$X \sim \mathrm{Tt}_{\nu+2r}(\mu, \sigma^{*2}; (a,b)), \quad \textit{with} \quad \sigma^{*2} = \frac{\nu}{(\nu+2r)}\sigma^2,$$

and

$$c(\nu,r) = \left(\frac{\nu+1}{\nu}\right)^r \frac{\Gamma((\nu+1)/2)\Gamma((\nu+2r)/2)}{\Gamma(\nu/2)\Gamma((\nu+2r+1)/2)}.$$

Thus, we consider first a linear regression model where the responses are observed with errors which are independent and identically distributed according to a Student-t distribution. To be more precise, let us write

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

where $\varepsilon_i \sim \mathrm{t}_\nu(0, \sigma^2)$, Y_i is the response for subject i, $\boldsymbol{\beta}_c = (\beta_0, \beta_1, \ldots, \beta_p)^\top = (\beta_0, \boldsymbol{\beta}^\top)^\top$ is a vector of regression parameters and $\mathbf{x}_{ic}^\top = (1, x_{i1}, \ldots, x_{ip})$ is a vector of explanatory variable values. By Equation (2.1), we have a that $Y_i \sim \mathrm{t}_\nu(\mathbf{x}_{ic}^\top \boldsymbol{\beta}_c, \sigma^2)$, for $i = 1, \ldots, n$. We call (2.2) the *tR* model.

We are interested in the case where right-censored observations can occur. That is, the observations are of the form

$$Y_{\mathrm{obs}_i} = \begin{cases} \kappa_i & \text{if} \quad Y_i \geq \kappa_i; \\ Y_i & \text{if} \quad Y_i < \kappa_i, \end{cases} \tag{2.3}$$

$i = 1, \ldots, n$, for some threshold point κ_i. We have chosen to work with the right censored case, which is the most common in applications, but the results are easily extendable to other censoring types. Note that when $\kappa_i = 0$, $i = 1, \ldots, n$, the proposed Student-t censored regression *(tCR)* model, defined in (2.2)-(2.3), is reduced to the Tobit model considered by Arellano-Valle *et al.* (2012) in which an interesting EM algorithm is developed to obtain maximum likelihood estimates.

2.2 Parameter estimation via an EM-type algorithm

In what follows in general we use the traditional convention denoting a random variable by an upper case letter and its realization by the corresponding lower case letter. 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 $\mathbf{y}_{\text{obs}} = \{\kappa_1, \ldots, \kappa_m, y_{m+1}, \ldots, y_n\}$. Then, the log-likelihood function of the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}_c^\top, \sigma^2, \nu)^\top$ is given by

$$\begin{aligned}
\ell(\boldsymbol{\theta}|\mathbf{y}_{\text{obs}}) &= \log\left(\prod_{i=1}^{n}\left[\mathscr{T}_\nu\left(\frac{\mathbf{x}_{ic}^\top\boldsymbol{\beta}_c - \kappa_i}{\sigma}\right)\right]^{\mathbb{I}_{\{y_i \geq \kappa_i\}}}\left[\mathrm{t}_\nu(y_i|\mathbf{x}_{ic}^\top\boldsymbol{\beta}_c, \sigma^2)\right]^{\mathbb{I}_{\{y_i < \kappa_i\}}}\right) \\
&= \sum_{i=1}^{m}\log\left[\mathscr{T}_\nu\left(\frac{\mathbf{x}_{ic}^\top\boldsymbol{\beta}_c - \kappa_i}{\sigma}\right)\right] + \sum_{i=m+1}^{n}\log\left[\mathrm{t}_\nu(y_i|\mathbf{x}_{ic}^\top\boldsymbol{\beta}_c, \sigma^2)\right]. \qquad (2.4)
\end{aligned}$$

To estimate the parameters of the *tCR* model, an alternative is to maximize this log-likelihood function directly, a procedure that can be quite cumbersome. Alternatively, our choice is to use the EM algorithm, a classic, reliable, widely used and general framework developed by Dempster *et al.* (1977) to obtain maximum likelihood estimates.

To apply the EM method, we need a representation of the model in terms of missing data. First, observe that, by Equation (2.1), if $Y_i \sim \mathrm{t}_\nu(\mathbf{x}_{ic}^\top\boldsymbol{\beta}_c, \sigma^2)$ then

$$Y_i|U_i = u_i \sim \mathrm{N}(\mathbf{x}_{ic}^\top\boldsymbol{\beta}_c, u_i^{-1}\sigma^2),\ U_i \sim \text{Gamma}(\nu/2, \nu/2). \qquad (2.5)$$

This relation is a convenient stochastic representation of the *tR* 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 $Y_i \sim \mathrm{t}_\nu(\mathbf{x}_{ic}^\top\boldsymbol{\beta}_c, \sigma^2)$, $i = 1, \ldots, m$. The key to the development of our EM-type algorithm is to consider the augmented data $\{\kappa_1, \ldots, \kappa_m, y_{m+1}, \ldots, y_n, u_1, \ldots, u_n\}$, that is, we treat the problem as if $\mathbf{y}_L = (y_1, \ldots, y_m)^\top$ were in fact observed. As a consequence, we can use the representation (2.5) to obtain the complete-data log-likelihood, given as

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

where $\mathbf{u} = (u_1, \ldots, u_n)^\top$ and $h(\cdot|\nu)$ is the Gamma density with both parameters equal to $\nu/2$.

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

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = \mathrm{E}_{\boldsymbol{\theta}^{(k)}}[\ell_c(\boldsymbol{\theta}|\mathbf{Y}_{\text{obs}}, \mathbf{Y}_L, \mathbf{U})|\mathbf{y}_{\text{obs}}],$$

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

$$\mathscr{E}_{si}(\boldsymbol{\theta}^{(k)}) = \mathrm{E}_{\boldsymbol{\theta}^{(k)}}[U_i Y_i^s | y_{\mathrm{obs}_i}],\ s = 0,1,2,$$

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

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) &= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[\mathscr{E}_{2i}(\boldsymbol{\theta}^{(k)}) - 2\mathscr{E}_{1i}(\boldsymbol{\theta}^{(k)})\mathbf{x}_{ic}^{\top}\boldsymbol{\beta}_c + \mathscr{E}_{0i}(\boldsymbol{\theta}^{(k)})(\mathbf{x}_{ic}^{\top}\boldsymbol{\beta}_c)^2\right] \\ &\quad + \sum_{i=1}^{n}\mathrm{E}_{\boldsymbol{\theta}^{(k)}}[\log h(U_i|\nu)|y_{\mathrm{obs}_i}]. \end{aligned} \tag{2.6}$$

From Massuia *et al.* (2015), we have that *for an uncensored observation i*,

$$\mathscr{E}_{si}(\boldsymbol{\theta}^{(k)}) = y_i^s \mathrm{E}_{\boldsymbol{\theta}^{(k)}}[U_i|y_i], \quad \text{with} \quad \mathrm{E}_{\boldsymbol{\theta}^{(k)}}[U_i|y_i] = \frac{\nu+1}{\nu + d^2(\boldsymbol{\theta}^{(k)}, y_i)} \tag{2.7}$$

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

$$\mathscr{E}_{si}(\boldsymbol{\theta}^{(k)}) = \mathrm{E}_{\boldsymbol{\theta}^{(k)}}\left[\frac{(\nu+1)Y_i^s}{\nu + d^2(\boldsymbol{\theta}^{(k)}, Y_i)}\Big|Y_i \geq \kappa_i\right], \tag{2.8}$$

which can be easily obtained from Lemma 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 & Rubin, 1993). The ECME algorithm (Liu & 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 actual marginal likelihood function, called CML-steps. Next, we describe this EM-type algorithm (ECME) for ML estimation of the parameters of the *tCR* model.
E-step: Given $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$. For $i = 1, \ldots, n$.

- If the observation i is uncensored then, for $s = 0,1,2$, compute $\mathscr{E}_{si}(\boldsymbol{\theta}^{(k)})$ given in (2.7);
- If the observation i is censored then, for $s = 0,1,2$, compute $\mathscr{E}_{si}(\boldsymbol{\theta}^{(k)})$ in (2.8) using Lemma 2 with $r = 1$.

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

$$\boldsymbol{\beta}_c^{(k+1)} = \left(\sum_{i=1}^{n}\mathscr{E}_{0i}(\boldsymbol{\theta}^{(k)})\mathbf{x}_{ic}\mathbf{x}_{ic}^{\top}\right)^{-1}\sum_{i=1}^{n}\mathbf{x}_{ic}\mathscr{E}_{1i}(\boldsymbol{\theta}^{(k)}), \tag{2.9}$$

$$\sigma^{2(k+1)} = \frac{1}{n}\sum_{i=1}^{n}\left[\mathscr{E}_{2i}(\boldsymbol{\theta}^{(k)}) - 2\mathscr{E}_{1i}(\boldsymbol{\theta}^{(k)})\mathbf{x}_{ic}^{\top}\boldsymbol{\beta}_c^{(k+1)} + \mathscr{E}_{0i}(\boldsymbol{\theta}^{(k)})(\mathbf{x}_{ic}^{\top}\boldsymbol{\beta}_c^{(k+1)})^2\right], \tag{2.10}$$

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

$$\nu^{(k+1)} = \mathrm{argmax}_{\nu}\left\{\sum_{i=1}^{m}\log\left[\mathscr{T}_{\nu}\left(\frac{\mathbf{x}_{ic}^{\top}\boldsymbol{\beta}_c^{(k+1)} - \kappa_i}{\sigma^{(k+1)}}\right)\right] + \sum_{i=m+1}^{n}\log\left[\mathrm{t}_{\nu}(y_i|\mathbf{x}_{ic}^{\top}\boldsymbol{\beta}_c^{(k+1)}, \sigma^{2(k+1)})\right]\right\}. \tag{2.11}$$

The more efficient CMLstep (2.11) can be easily accomplished by using, for instance, the optim routine in the R software. The algorithm iterates between the E- and M-steps until reaching convergence. This process is iterated until some distance involving two successive evaluations of the actual log-likelihood $\ell(\boldsymbol{\theta})$, like $||\ell(\boldsymbol{\theta}^{(k+1)}) - \ell(\boldsymbol{\theta}^{(k)})||$ or $||\ell(\boldsymbol{\theta}^{(k+1)})/\ell(\boldsymbol{\theta}^{(k)}) - 1||$, is small enough. This algorithm is implemented as part of the R package *CensRegMod ()* (Massuia *et al.*, 2012), which can be downloaded at no charge from the CRAN repository.

3 The FM-tCR model

Ignoring censoring for the moment, we first consider a more general and robust framework for the random error ε_i of the regression model defined in (2.2), which is assumed to follows a mixture of Student-t distributions. More precisely, the model considered is based on assumptions (2.2) and (2.3), with

$$\varepsilon_i \sim \sum_{j=1}^{G} p_j t_{\nu_j}(\mu_j, \sigma_j^2), \tag{3.12}$$

where p_j are weights adding to 1 and the $\mu_j's$ satisfy the *identifiability constraint* $\sum_{j=1}^{G} p_j \mu_j = 0$ and G is the number of groups (also called components in mixture models).

The mixture regression model considered in (2.2), (2.3) and (3.12) is defined as: let Z_i be a latent class variable such that given $Z_i = j$, the response Y_i depends on the p-dimensional predictor $\mathbf{x}_i^\top = (x_{i1}, \ldots, \mathbf{x}_{ip})$ in a linear way

$$Y_i = \beta_0 + \mu_j + \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_{ij},\ \ \varepsilon_{ij} \sim t_{\nu_j}(0, \sigma_j^2),\ \ i = 1, \ldots, n,\ \ j = 1, \ldots, G, \tag{3.13}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$. Concerning the parameter ν_j, $j = 1, \ldots, G$, for computational convenience we assume that $\nu = \nu_1 = \nu_2 = \ldots, = \nu_G$. This strategy works very well in the empirical studies that we have conducted and greatly simplifies the optimization problem.

Now, suppose $P(Z_i = j) = p_j$ and Z_i is independent of $\mathbf{x}_i$. Then, the conditional density of Y_i given $\mathbf{x}_i$, without observing Z_i, is

$$f(y_i|\mathbf{x}_i, \boldsymbol{\theta}) = \sum_{j=1}^{G} p_j t_\nu(y_i|\varphi_j + \mathbf{x}_i^\top \boldsymbol{\beta}, \sigma_j^2), \tag{3.14}$$

where $\varphi_j = \beta_0 + \mu_j$ and $\boldsymbol{\theta} = (\boldsymbol{\gamma}^\top, \boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_G^\top)^\top$, with $\boldsymbol{\gamma} = (\nu, \beta_0, \boldsymbol{\beta}^\top)^\top$ and $\boldsymbol{\theta}_j = (p_j, \sigma_j^2, \mu_j)^\top$. The model (3.14) is the regression model based on the mixture of Student-t distributions, studied, for instance, by Galimberti & Soffritti (2014).

Following Karlsson & Laitila (2014), the mixture model for censored data can be formulated in a similar way to the model defined in (3.14) as:

$$f(y_i|\mathbf{x}_i, \boldsymbol{\theta}) = \sum_{j=1}^{G} p_j g_{ij}(y_i|\mathbf{x}_i, \boldsymbol{\gamma}, \boldsymbol{\theta}_j), \tag{3.15}$$

where

$$g_{ij}(y_i|\mathbf{x}_i,\boldsymbol{\gamma},\boldsymbol{\theta}_j)=\left[\mathscr{T}_\nu\left(\frac{\varphi_j+\mathbf{x}_i^\top\boldsymbol{\beta}-\kappa_i}{\sigma_j}\right)\right]^{\mathbb{I}_{\{y_i\le\kappa_i\}}}\left[\mathrm{t}_\nu(y_i|\varphi_j+\mathbf{x}_i^\top\boldsymbol{\beta},\sigma_j^2)\right]^{\mathbb{I}_{\{y_i>\kappa_i\}}}.$$

The model defined in (3.15) will be called the FM-tCR model.

3.1 Maximum likelihood estimation via EM algorithm

In this section, we present an EM algorithm for the ML estimation of the FM-tCR model defined in (3.15). To explore the EM algorithm, we present the FM-tCR in an incomplete-data framework, using the results presented in Section 2.

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

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

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

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

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

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

For the vector $\mathbf{Z}_i$ we will use the notation $\mathbf{Z}_i\overset{\text{iid}}{\sim}Multinomial(1,p_1,\ldots,p_g)$. Observe that $Z_{ij}=1$ if and only if $Z_i=j$. Thus, from (2.1), the set-up defined above can be written hierarchically as

$$Y_i|U_i=u_i,Z_{ij}=1 \overset{\text{ind}}{\sim} N(\mathbf{x}_i^\top\boldsymbol{\beta}+\varphi_j,u_i^{-1}\sigma_j^2), \tag{3.16}$$

$$U_i|Z_{ij}=1 \overset{\text{ind}}{\sim} Gamma(\nu/2,\nu/2), \tag{3.17}$$

$$\mathbf{Z}_i \overset{\text{iid}}{\sim} Multinomial(1,p_1,\ldots,p_g), \tag{3.18}$$

for $i=1,\ldots,n$, all independent. For censored data, let $\mathbf{y}_{\text{obs}}=\{\kappa_1,\ldots,\kappa_m,y_{m+1},\ldots,y_n\}$, $\mathbf{y}_L=(y_1,\ldots,y_m)^\top$, $\mathbf{u}=(u_1,\ldots,u_n)^\top$, and $\mathbf{z}=(\mathbf{z}_1^\top,\ldots,\mathbf{z}_n^\top)^\top$. Then, under the hierarchical representation (3.16)–(3.17), it follows that the complete log-likelihood function associated with $\mathbf{y}_c=(\mathbf{y}_{obs}^\top,\mathbf{y}_L,\mathbf{u}^\top,\mathbf{z}^\top)^\top$ is

$$\begin{aligned}\ell_c(\boldsymbol{\theta}|\mathbf{y}_c) &= c+\sum_{i=1}^n\sum_{j=1}^G z_{ij}\log p_j-\frac{1}{2}\sum_{i=1}^n\sum_{j=1}^G z_{ij}\log\sigma_j^2-\frac{1}{2}\sum_{i=1}^n\sum_{j=1}^G\frac{z_{ij}u_i}{\sigma_j^2}(y_i-\mathbf{x}_i^\top\boldsymbol{\beta}_j-\varphi_j)^2\\ &\quad+\sum_{i=1}^n\sum_{j=1}^G z_{ij}\log h(u_i|\nu),\end{aligned} \tag{3.19}$$

where c is a constant that is independent of the parameter vector $\boldsymbol{\theta}$.
Letting $\boldsymbol{\theta}^{(k)} = (\boldsymbol{\gamma}^{(k)\top}, \boldsymbol{\theta}_1^{(k)\top}, \ldots, \boldsymbol{\theta}_G^{(k)\top})^\top$, with $\boldsymbol{\gamma}^{(k)} = (\nu^{(k)}, \beta_0^{(k)}, \boldsymbol{\beta}^{(k)\top})^\top$ and $\boldsymbol{\theta}_j = (p_j^{(k)}, \sigma_j^{2(k)}, \mu_j^{(k)})^\top$, $j = 1, \ldots, G$, the estimates of $\boldsymbol{\theta}$ at the k–th iteration.

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

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) &= c + \sum_{i=1}^{n}\sum_{j=1}^{G} \mathscr{Z}_{ij}(\boldsymbol{\theta}^{(k)}) \log p_j - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{G} \mathscr{Z}_{ij}(\boldsymbol{\theta}^{(k)}) \log \sigma_j^2 - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{G} \frac{\mathscr{E}_{2ij}(\boldsymbol{\theta}^{(k)})}{\sigma_j^2} \\ &+ \sum_{i=1}^{n}\sum_{j=1}^{G} \frac{\mathscr{E}_{1ij}(\boldsymbol{\theta}^{(k)})}{\sigma_j^2}(\mathbf{x}_i^\top \boldsymbol{\beta} + \varphi_j) - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{G} \frac{\mathscr{E}_{0ij}(\boldsymbol{\theta}^{(k)})}{\sigma_j^2}(\mathbf{x}_i^\top \boldsymbol{\beta} + \varphi_j)^2, \end{aligned} \tag{3.20}$$

where

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

By using known properties of conditional expectation, we obtain

$$\mathscr{Z}_{ij}(\boldsymbol{\theta}^{(k)}) = \frac{p_j^{(k)} g_{ij}(y_i|\mathbf{x}_i, \boldsymbol{\gamma}^{(k)}, \boldsymbol{\theta}_j^{(k)})}{\sum_{j=1}^{G} p_j^{(k)} g_{ij}(y_i|\mathbf{x}_i, \boldsymbol{\gamma}^{(k)}, \boldsymbol{\theta}_j^{(k)})}, \tag{3.21}$$

$\mathscr{E}_{0ij}(\boldsymbol{\theta}^{(k)}) = \mathscr{Z}_{ij}(\boldsymbol{\theta}^{(k)}) E_{\boldsymbol{\theta}^{(k)}}[U_i | y_{\mathrm{obs}_i}, Z_{ij} = 1]$, $\mathscr{E}_{1ij}(\boldsymbol{\theta}^{(k)}) = \mathscr{Z}_{ij}(\boldsymbol{\theta}^{(k)}) E_{\boldsymbol{\theta}^{(k)}}[U_i Y_i | y_{\mathrm{obs}_i}, Z_{ij} = 1]$ and $\mathscr{E}_{2ij}(\boldsymbol{\theta}^{(k)}) = \mathscr{Z}_{ij}(\boldsymbol{\theta}^{(k)}) E_{\boldsymbol{\theta}^{(k)}}[U_i Y_i^2 | y_{\mathrm{obs}_i}, Z_{ij} = 1]$, where the conditional expectations of the form

$$E_{\boldsymbol{\theta}^{(k)}}[U_i Y_i^s | y_{\mathrm{obs}_i}, Z_{ij} = 1],\ s = 0, 1, 2, \tag{3.22}$$

can be easily derived from equations (2.7) and (2.8) given in Section 2. Thus, we have closed form expression for all the quantities involved in path E of the algorithm. Next, we describe the ECME algorithm for maximum likelihood estimation of the parameters of the FM-tCR model.

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

$$p_j^{(k+1)} = \frac{\sum_{i=1}^{n} \mathscr{Z}_{ij}(\boldsymbol{\theta}^{(k)})}{n}, \tag{3.23}$$

$$\boldsymbol{\beta}^{(k+1)} = \left(\sum_{i=1}^{n}\sum_{j=1}^{G} \frac{\mathscr{E}_{0ij}(\boldsymbol{\theta}^{(k)})}{\sigma_j^{2(k)}} \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1} \sum_{i=1}^{n}\sum_{j=1}^{G} \frac{\mathbf{x}_i}{\sigma_j^{2(k)}} \left(\mathscr{E}_{1ij}(\boldsymbol{\theta}^{(k)}) - \mathscr{E}_{0ij}(\boldsymbol{\theta}^{(k)}) \varphi_j^{(k)} \right), \tag{3.24}$$

$$\varphi_j^{(k+1)} = \frac{\sum_{i=1}^{n} \left[\mathscr{E}_{1ij}(\boldsymbol{\theta}^{(k)}) - \mathscr{E}_{0ij}(\boldsymbol{\theta}^{(k)}) \mathbf{x}_i^\top \boldsymbol{\beta}^{(k)} \right]}{\sum_{i=1}^{n} \mathscr{E}_{0ij}(\boldsymbol{\theta}^{(k)})}, \tag{3.25}$$

$$\sigma^{2}{}_j^{(k+1)} = \frac{\sum_{i=1}^{n} \Delta_{ij}^{(k)}}{\sum_{i=1}^{n} \mathscr{Z}_{ij}(\boldsymbol{\theta}^{(k)})}, \quad j = 1, \ldots, G, \tag{3.26}$$

where

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

Following Bartolucci & Scaccia (2005), we can also obtain an estimative of β_0 as $\beta_0^{(k)} = \sum_{j=1}^G p_j^{(k)}\varphi_j^{(k)}$, and for $j = 1,\ldots G$, $\mu_j^{(k)}$ as $\varphi_j^{(k)} - \beta_0^{(k)}$.

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

$$\nu^{(k+1)} = argmax_\nu \sum_{i=1}^n \log\left(\sum_{j=1}^G p_j^{(k+1)} g_{ij}(y_i|\mathbf{x}_i, \nu, \beta_0^{(k+1)}, \boldsymbol{\beta}^{(k+1)}, \boldsymbol{\theta}_j^{(k+1)})\right), \tag{3.27}$$

where $g_{ij}(y_i|\mathbf{x}_i, \boldsymbol{\gamma}, \boldsymbol{\theta}_j)$ as defined in (3.15).

A more parsimonious model is achieved by supposing $\sigma_1^2 = \ldots = \sigma_G^2 = \sigma^2$, which can be seen as an extension of the FM-NCR model with restricted variance-covariance components. In this case, the updates for $p_j^{(k)}$, $\boldsymbol{\beta}^{(k)}$ and $\varphi_j^{(k)}$ remain the same, and the update for $\sigma^{2(k)}$ is given as

$$\sigma^{2(k+1)} = \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^G \mathscr{Z}_{ij}(\boldsymbol{\theta}^{(k)})\sigma_j^{2(k+1)}.$$

It is well known 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. In our examples and simulation studies, we consider the following procedure for the FM-tCR model:

- For $\boldsymbol{\beta}^{(0)}$, use the ordinary least-square (OLS) estimate in the regression model defined in (2.2).
- Partition the residuals into G groups using the K-means clustering algorithm (Basso *et al.*, 2010);
- Compute the proportion of data points belonging to the same cluster j, say $p_j^{(0)}$, $j = 1,\ldots,G$. 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 value for ν is taken to be 3.

3.2 Provision 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 $\widehat{\boldsymbol{\theta}}$ by the inverse of the observed information matrix. Let $\mathbf{I}_o(\boldsymbol{\theta}|\mathbf{y}) = -\partial^2 \ell(\boldsymbol{\theta}|\mathbf{y})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top$ be the observed information matrix, where $\ell(\boldsymbol{\theta}|\mathbf{y})$ is the observed log-likelihood function as in (3.15). In this work we use the alternative method suggested by Basford *et al.* (1997), which consists of approximating the inverse of the covariance matrix by

$$\mathbf{I}_o(\widehat{\boldsymbol{\theta}}|\mathbf{y}) = \sum_{i=1}^{n} \widehat{\mathbf{s}}_i \widehat{\mathbf{s}}_i^\top, \tag{3.28}$$

where $\widehat{\mathbf{s}}_i = E\left[\partial\left(\ell_c(\boldsymbol{\theta}|\mathbf{y}_c)\right)/\partial\boldsymbol{\theta}\right]\big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$, with $\ell_c(\boldsymbol{\theta}|\mathbf{y}_c)$ as in (3.19) and

$$\widehat{\mathbf{s}}_i = (\widehat{s}_{i,\boldsymbol{\beta}}, \widehat{s}_{i,\beta_0}, \widehat{s}_{i,\sigma_1^2}, \ldots, \widehat{s}_{i,\sigma_G^2}, \widehat{s}_{i,\mu_1}, \ldots, \widehat{s}_{i,\mu_G}, \widehat{s}_{i,p_1}, \ldots, \widehat{s}_{i,p_{G-1}})^\top.$$

Expressions for the elements $\widehat{s}_{i,\boldsymbol{\beta}}, \widehat{s}_{i,\beta_0}, \widehat{s}_{i,\sigma_j^2}, \widehat{s}_{i,\mu_j}, \widehat{s}_{i,p_j}$ are given in the following:

$$\begin{aligned}
\widehat{s}_{i,\boldsymbol{\beta}} &= \sum_{j=1}^{G}\left\{\frac{1}{\widehat{\sigma}_j^2}\left[\mathscr{E}_{1ij}(\widehat{\boldsymbol{\theta}}^{(k)})\mathbf{x}_i - \mathscr{E}_{0ij}(\widehat{\boldsymbol{\theta}}^{(k)})(\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}} + \widehat{\varphi}_j)\mathbf{x}_i\right]\right\},\\
\widehat{s}_{i,\beta_0} &= \sum_{j=1}^{G}\left\{\frac{1}{\widehat{\sigma}_j^2}\left[\mathscr{E}_{1ij}(\widehat{\boldsymbol{\theta}}^{(k)}) - \mathscr{E}_{0ij}(\widehat{\boldsymbol{\theta}}^{(k)})(\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}} + \widehat{\varphi}_j)\right]\right\},\\
\widehat{s}_{i,\sigma_j^2} &= -\frac{1}{2}\left\{\frac{1}{\widehat{\sigma}_j^4}\left[\mathscr{Z}_{ij}(\widehat{\boldsymbol{\theta}}^{(k)})\widehat{\sigma}_j^2 - \mathscr{E}_{2ij}(\widehat{\boldsymbol{\theta}}^{(k)}) + 2\mathscr{E}_{1ij}(\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}} + \widehat{\varphi}_j) - \mathscr{E}_{0ij}(\widehat{\boldsymbol{\theta}}^{(k)})(\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}} + \widehat{\varphi}_j)^2\right]\right\},\\
\widehat{s}_{i,\mu_j} &= \frac{1}{\widehat{\sigma}_j^2}\left[\mathscr{E}_{1ij}(\widehat{\boldsymbol{\theta}}^{(k)}) - \mathscr{E}_{0ij}(\widehat{\boldsymbol{\theta}}^{(k)})(\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}} + \widehat{\varphi}_j)\right],\\
\widehat{s}_{i,p_j} &= \frac{\mathscr{Z}_{ij}(\widehat{\boldsymbol{\theta}}^{(k)})}{\widehat{p}_j} - 1,\\
\widehat{s}_{i,\nu} &= \frac{1}{2}\sum_{j=1}^{G}\left\{\mathscr{Z}_{ij}(\widehat{\boldsymbol{\theta}}^{(k)})\left[\log\left(\frac{\widehat{\nu}}{2}\right) + 1 - \psi\left(\frac{\widehat{\nu}}{2}\right) + E(\log(U_i)|y_{\text{obs}_i}, \widehat{\boldsymbol{\theta}})\right] - \mathscr{E}_{0ij}(\widehat{\boldsymbol{\theta}}^{(k)})\right\},
\end{aligned} \tag{3.29}$$

where $\psi(x)$ represents the digamma function of x. It is important to stress that the SE of ν, obtained from $\widehat{s}_{i,\nu}$, depends heavily on the calculation of $E[\log(U_i)|y_{\text{obs}_i}, \widehat{\boldsymbol{\theta}}]$, of Equation (3.29), which relies on computationally intensive Monte Carlo integrations, since we do not have an analytical expression for this expected value. Thus, in our analysis we focus solely on comparing the SE of $\boldsymbol{\beta}$, σ_j^2 and p_j, with $j = 1, \ldots, G$.

The information-based approximation (3.28) is asymptotically applicable. However, it is less reliable unless the sample size is sufficiently large. It is common practice to perform the parametric bootstrap approach (Efron & Tibshirani, 1986) to obtain more accurate standard error estimates. However, we do not employ the bootstrap approach, since it requires enormous amounts of computing power.

3.3 Model selection

Because there is no universal criterion for mixture model selection, we chose three criteria to compare the FM-tCR and FM-NCR models. The first three are the Akaike information criterion (AIC), the Bayesian information criterion (BIC) and the efficient determination criterion (EDC). Like the more popular AIC and BIC criteria, EDC has the form

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

where $\ell(\boldsymbol{\theta})$ is the actual log-likelihood, ρ is the number of free parameters that have to be estimated in the model and the penalty term c_n is a convenient sequence of positive numbers. Here, we use $c_n = 0.2\sqrt{n}$, a proposal that was considered in Basso *et al.* (2010) and Cabral *et al.* (2012). We have $c_n = 2$ for AIC, $c_n = \log n$ for BIC, where n is the sample size.

Table 1: Simulation Study 1: Mean and standard deviations (SD) for EM estimates based on 500 samples from FM-tCR model. True values of parameters are in parentheses.

		Scenario 1: ($\sigma_1^2 = 0.3$, $\sigma_2^2 = 0.6$)					Scenario 2: ($\sigma_1^2 = 2$, $\sigma_2^2 = 2$)				
Parameter		10%	20%	30%	40%	50%	10%	20%	30%	40%	50%
$\beta_0(1)$	Mean	1.0851	1.1764	1.2630	1.3786	1.4751	1.1107	1.2173	1.3524	1.5059	1.6978
	SD	0.0070	0.0073	0.0079	0.0084	0.0092	0.0358	0.0354	0.0366	0.0380	0.0381
$\beta_1(-1)$	Mean	-0.9933	-1.0037	-1.0031	-1.0106	-1.0115	-1.0133	-1.0058	-0.9992	-1.0000	-0.9961
	SD	0.1132	0.1146	0.1139	0.1154	0.1151	0.2920	0.2908	0.2763	0.2689	0.2500
$\beta_2(4)$	Mean	3.9970	4.0014	4.0066	4.0004	4.0029	3.9953	3.9997	4.0141	3.9923	3.9603
	SD	0.1115	0.1151	0.1128	0.1159	0.1175	0.3014	0.2824	0.2779	0.2774	0.2490
$\mu_1(1)$	Mean	0.9159	0.8303	0.7380	0.6384	0.5448	0.9172	0.8315	0.7136	0.6010	0.4752
	SD	0.0255	0.0227	0.0212	0.0204	0.0200	0.2111	0.1925	0.1878	0.1889	0.1977
$\mu_2(-4)$	Mean	-4.0760	-4.1800	-4.2684	-4.3768	-4.4614	-4.0597	-4.1699	-4.3224	-4.5097	-4.6943
	SD	0.1324	0.1261	0.1153	0.1103	0.1204	0.6524	0.6121	0.5304	0.4795	0.3884
σ_1^2	Mean	0.3021	0.3010	0.3021	0.3002	0.3034	1.9960	2.0178	2.0212	2.0452	2.1193
	SD	0.0863	0.0871	0.0895	0.0916	0.0965	0.2333	0.2229	0.2333	0.2205	0.2191
σ_2^2	Mean	0.6011	0.5941	0.5654	0.5710	0.6377	2.0405	1.9793	1.7786	1.6927	1.4190
	SD	0.1237	0.1211	0.1171	0.1147	0.1174	0.3489	0.3313	0.3167	0.2930	0.2634
$\nu(3)$	Mean	3.1772	3.1259	3.1418	3.1750	3.4809	3.1386	3.1124	3.0011	3.0576	3.0690
	SD	0.8457	0.7577	0.8808	1.0017	1.4816	0.6392	0.7025	0.6674	1.3066	1.0674
$p_1(0.8)$	Mean	0.8164	0.8342	0.8525	0.8726	0.8905	0.8170	0.8377	0.8596	0.8822	0.9046
	SD	0.1125	0.1209	0.1311	0.1447	0.1602	0.1739	0.1843	0.2049	0.2339	0.2830

4 Simulated studies

In this section, we consider three simulation experiments to show the applicability of our proposed model. Our intention is to show that the FM-tCR can do exactly what it is designed for, that is, satisfactorily model CR models that have serious departures from the normal and Student-t assumptions.

4.1 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 ECM algorithm. This is the first step to ensure that the estimation procedure works satisfactorily. We fit the FM-tCR data that were artificially generated from the following FM-CR model with two components:

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

where Z_{ij} is a component indicator of Y_i with $P(Z_{ij} = 1) = p_j$, $j = 1,2$, $\mathbf{x}_i^\top = (x_{i1}, x_{i2})$, such that $x_{i1} \sim U(0,1)$ and $x_{i2} \sim U(0,1)$, $i = 1,\ldots,n$, and ε_1 and ε_2 follow a distribution as in the assumption given in (3.12) and several censoring proportion settings (10%, 20%, 30%, 40% and 50%). We generated 500 Monte Carlo samples of size $n = 500$, with the following parameter values: $\beta_0 = 1$, $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top = (-1,4)^\top$, $\mu_1 = 1$, $\mu_2 = 1$, $p_1 = 0.8$ and $\nu = 3$. In addition, we consider the following scenarios: scenario 1 (small variances and different): $\sigma_1^2 = 0.3$ and $\sigma_2^2 = 0.6$, and scenario 2 (large variances and equal): $\sigma_1^2 = 2$ and $\sigma_2^2 = 2$. The Monte Carlo mean and corresponding standard deviations (SD) of the ML estimates are presented in the Table 1. The EM estimates across all samples were computed using the R package *CensMixReg()*. 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. In general, the results suggest that the proposed FM-tCR model produced satisfactory estimates when the censoring level was small (around 30%) and lost performance when the censoring level increased.

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

Here, the experiment is planned to show the asymptotic properties of the EM estimates. Our strategy is to generate generated artificial samples from the FM-tCR model (3.13), with $x_i^\top = (x_{1i}, x_{2i})$, such that $x_{i1} \sim U(0,1)$ and $x_{i2} \sim U(0,1)$, $i = 1,\ldots,n$. We chose various settings of censoring proportions $p = 10, 20, 30, 40$ and 50% and samples sizes $n = 100, 150, 200, 300, 400, 500, 700, 800, 900$ and 1000. The true values of the regression parameters were taken as $\beta_0 = 1$, $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top = (-1,4)^\top$, $\sigma_1^2 = 1$ and $\sigma_2^2 = 0.5$. For each combination of parameters, sample sizes and censoring levels, we generated 500 random samples from the FM-tCR. In order to analyze asymptotic properties of the EM estimates, we computed the bias and the mean squared error (MSE) for each combination of sample size, censoring level and parameter values. For θ_i, they are given by

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

where $\widehat{\theta}_i^{(j)}$ is the estimate of θ_i for the j-th sample. The results for β_1, β_2 and σ_1^2 are shown in Figure 1. We can see a pattern of convergence to zero of the bias and RMSE when n increases independently of the censoring pattern (a similar pattern was observed for the other 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-tCR model do provide good asymptotic properties.

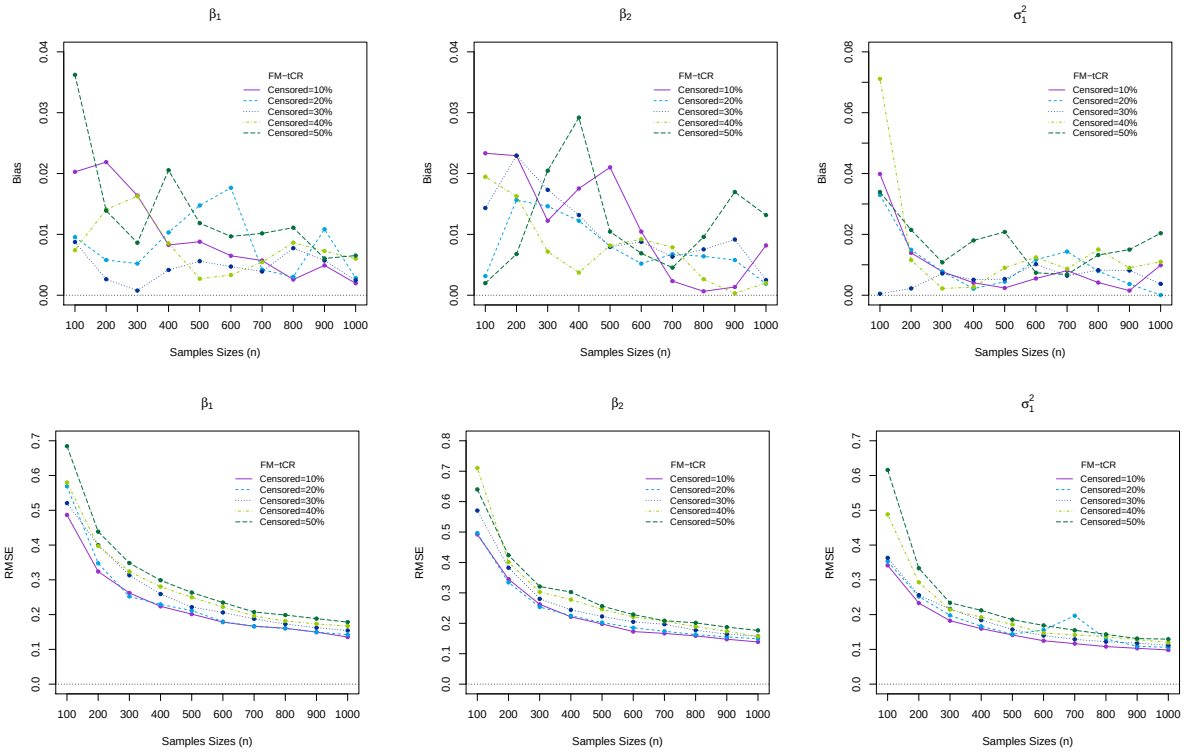

Figure 1: Simulation study 2: Average bias (first row) and RMSE (second row) of parameter estimates under the FM-tCR model.

4.3 Robustness of the EM estimates (simulation study 3)

In this section, we illustrate the ability of our FM-tCR model (compared to the FM-NCR model) to fit data with a mixture structure generated from a different family of distributions. Thus, we generated the error terms from two different mixtures with $G = 2$ components: (a) mixtures of skew normal Birnbaum-Saunders distributions and (b) mixtures of inverse Gaussian distributions.

According to Santana *et al.* (2011), a random variable T follows a skew normal Birnbaum-Saunders distributions (SNBS) if its pdf is given by

$$f(t) = 2\phi(a(t;\alpha,\beta))\Phi(\lambda a(t;\alpha,\beta))A(t;\alpha,\beta), \quad t > 0, \tag{4.30}$$

where $\alpha > 0$, $\beta > 0$ and $\lambda \in \mathbb{R}$ are the shape, scale and skewness parameters respectively. A random variable with pdf as in (4.30) will be denoted by $SNBS(\alpha,\beta,\lambda)$. We can use the R package

bssn() (Maehara & Benites, 2015), to generate random observations from a mixture of SNBS distributions.

On the other hand, according Karlis & Santourian (2008), an inverse Gaussian random variable has pdf given by

$$f(x) = a(\alpha,\beta,\mu,\delta)q\left(\frac{x-\mu}{\delta}\right)^{-1} K_1\left[\delta\alpha q\left(\frac{x-\mu}{\delta}\right)\right] exp(\beta x), \quad x \in \mathbb{R}, \tag{4.31}$$

where $a(\alpha,\beta,\mu,\delta) = \pi^{-1}\alpha exp(\delta\sqrt{\alpha^2-\beta^2}-\beta\mu)$, $q(x) = \sqrt{1+x^2}$ and K_1 is the Bessel function of third order and index 1. Furthermore, α, β, μ and δ are parameters, satisfying $0 \leq |\beta| \leq \alpha$, $\mu \in \mathbb{R}$ and $0 < \delta$. A random variable with pdf as in (4.31) will be denoted by $NIG(\alpha,\beta,\mu,\delta)$.

Table 2: Simulation study 3: Arithmetic averages of the model comparison measures. In parentheses are the percentages in which the respective model was selected for each criterion.

		FM-tCR			FM-NCR		
Model	CR	AIC	BIC	EDC	AIC	BIC	EDC
	10%	2166.078	2204.009	2188.327	2176.248	2209.964	2196.025
		(81.8%)	(67.6%)	(72.6%)	(18.2%)	(32.4%)	(27.4%)
	20%	1956.377	1994.308	1978.626	1967.047	2000.764	1986.824
		(83.8%)	(70.8%)	(75.8%)	(16.2%)	(29.2%)	(24.2%)
$.8NIG(\sqrt{5},-2,1,1)+.2NIG(\sqrt{5},2,2,1)$	30%	1746.918	1784.850	1769.168	1756.237	1789.954	1776.015
		(81.2%)	(66.4%)	(73.6%)	(18.8%)	(33.6%)	(26.4%)
	40%	1531.820	1569.751	1554.069	1540.370	1574.087	1560.147
		(79.6%)	(61%)	(68.4%)	(20.4%)	(39%)	(31.6%)
	50%	1316.714	1354.646	1338.964	1324.907	1358.624	1344.684
		(80%)	(57.2%)	(65.6%)	(20%)	(42.8%)	(34.4%)
		FM-tCR			FM-NCR		
Model	CR	AIC	BIC	EDC	AIC	BIC	EDC
	10%	2659.020	2696.952	2681.270	2717.679	2751.396	2737.456
		(99.4%)	(98.4%)	(99%)	(0.6%)	(1.6%)	(1.0%)
	20%	2406.826	2444.758	2429.075	2457.646	2491.362	2477.423
		(99.6%)	(97.6%)	(98.4%)	(0.4%)	(2.4%)	(1.6%)
$.8SNBS(2.5,1,3)+.2SNBS(0.5,1,4)$	30%	2153.714	2191.645	2175.963	2195.563	2229.280	2215.340
		(98.8%)	(97.6%)	(98.2%)	(1.2%)	(2.4%)	(1.8%)
	40%	1900.141	1938.072	1922.390	1932.415	1966.132	1952.192
		(95.2%)	(92.4%)	(93.6%)	(4.8%)	(7.6%)	(6.4%)
	50%	1635.699	1673.630	1657.948	1660.925	1694.642	1680.702
		(90%)	(84%)	(86.8%)	(10.0%)	(16.0%)	(13.2%)

For each generated dataset, we computed the AIC, BIC and EDC criteria under the FM-NCR and FM-tCR models and their respective values were recorded. Table 2 shows the arithmetic average of these comparison measures, as well as, the percentages in that the Student-t and normal models were chosen. Note that all the measures favored the FM-tCR model. This fact indicates that the FM-tCR model is, in general, more robust to deviations from the model assumptions and fits better than the FM-NCR model when neither is the true generating model. This affirmation can be also observed in Figure 2, where we plot the values of AIC, BIC and EDC for each scenario and model, with 10% censored responses.

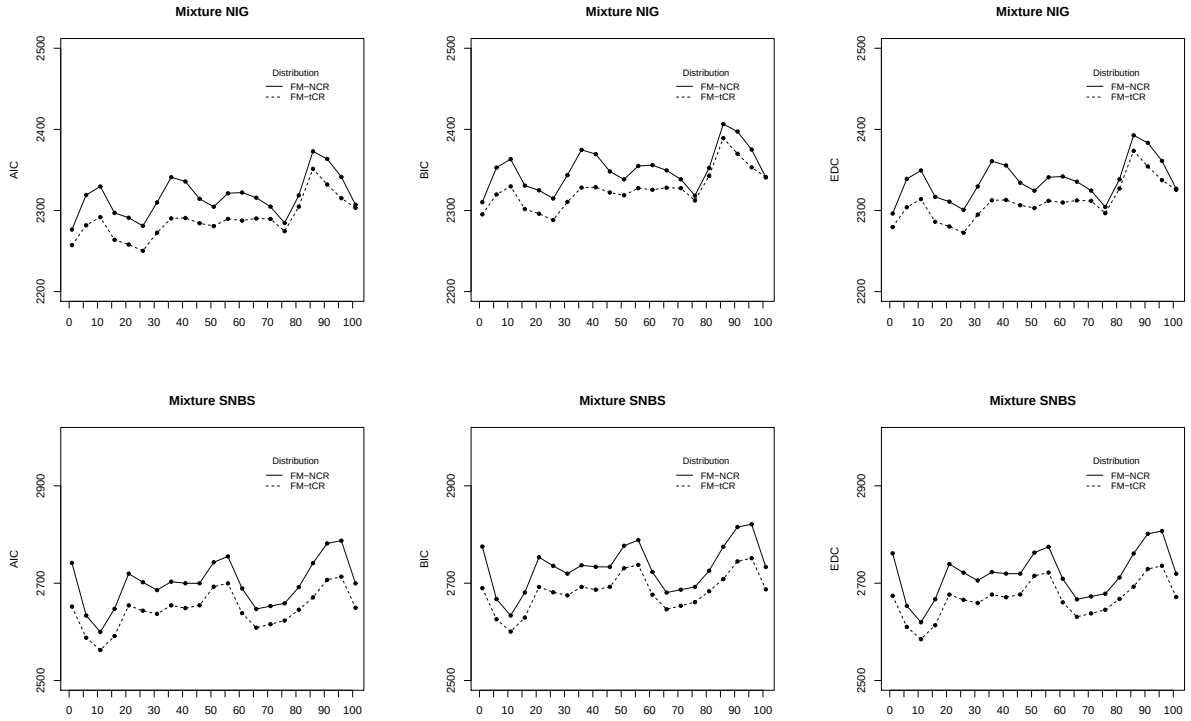

Figure 2: Simulation study 3. Model selection criteria behavior along the 100 generated samples of mixtures of two NIG distributions (first row) and two SNBS distributions(second row).

5 Application

In order to study the performance of our proposed model and algorithm, we analyze a real dataset. The computational procedures were implemented using the R software (R Development Core Team, 2015), through the package *CensMixReg()*. 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 not work for pay. Thus, we can consider these last ones as left censored with $\kappa_i = 0$, $i = 1, \ldots, n$.. The following variables were considered:

- y_i : *wage rates*, defined as the average hourly earnings. If the wage rates are set equal to zero, these wives did not work in 1975. Therefore, these observations are considered left censored at zero;
- x_{i1}: wife's age;
- x_{i2}: educational attainment (in years);
- x_{i3}: husband's hours worked in 1975.

This application is based on left-censoring, which is immediate and follows from (2.3) by reversing the order of y_i and κ_i. Each of the vectors of explanatory variable values is given by $\mathbf{x}_i^\top = (x_{i1}, x_{i2}, x_{i3})$ for $i = 1, 2, \ldots, 753$. 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, more recently by Garay *et al.* (2015), using a censored regression model with scale mixtures of normal distributions. Here, we revisit this dataset in order to evaluate the performance of the proposed methods considering the FM-NCR and FM-tCR models.

The results of the EM algorithm are shown in Table 3. This table shows that the estimates of $\beta_0 - \beta_3$ for the FM-NCR and FM-tCR models are close. However, the standard errors (SE) of $\boldsymbol{\beta}$ are smaller than those of the FM-NCR model, indicating that the FM-tCR model seem to produce more precise estimates. The estimates for the variance components are not comparable since they are on a different scale. Also, notice that the small value of the estimate of ν for the FM-tCR model indicates a lack of adequacy of the normal (FM-NCR) assumption. Table 4 compares the fit of the two mixture models using the model selection criteria discussed in Subsection 3.3. Note that, as expected, the FM-tCR model performs significantly better than the FM-NCR model.

Table 3: Wage rate data: results of the parameter estimation via the EM algorithm.

Parameter	FM-NCR model		FM-tCR model	
	Estimative	SE	Estimative	SE
β_0	6.1596	0.0365	5.6901	0.0369
β_1	-0.0057	0.0155	-0.0096	0.0127
β_2	0.4375	0.0577	0.4180	0.0439
β_3	2.9609	0.5377	2.8746	0.1691
σ_1^2	4.9390	0.6847	3.8023	0.3833
σ_2^2	92.8447	38.2790	8.5156	9.9294
μ_1	0.2056	1.1456	0.1845	0.9159
μ_2	-1.9448	3.5514	-22.2453	2.4183
p	0.9044	0.5542	0.9918	0.6083
ν	-	-	3.1758	-

The robustness of the FM-tCR model can be assessed by considering the influence of a single outlier on the EM estimate of $\boldsymbol{\theta}$. In particular, we can assess how much the EM estimate of $\boldsymbol{\theta}$ is influenced by a change of δ units in a single observation y_i. Replacing y_i by $y_i(\delta) = y_i + \delta$, let $\widehat{\beta}_j(\delta)$ be the EM estimates of β_j after contamination, $j = 0, 1, 2, 3$. We are particularly interested in the relative change $|(\widehat{\beta}_j(\delta) - \widehat{\beta}_j)/\widehat{\beta}_j|$. Figure 3 displays the results of the relative changes of the estimates for different values of δ, under both models, contaminating observation 100 (uncensored) and varying δ between 0 and 10 using 0.5 step size. As expected, the estimates from the FM-tCR model are less affected by variations of δ, especially when δ is large.

Table 4: Wage rate data. Model selection criteria

Criterion	FM-tCR	FM-NCR
log-likelihood	**-1239.529**	-1250.085
AIC	**2499.057**	2518.171
BIC	**2545.298**	2559.787
EDC	**2533.939**	2549.564

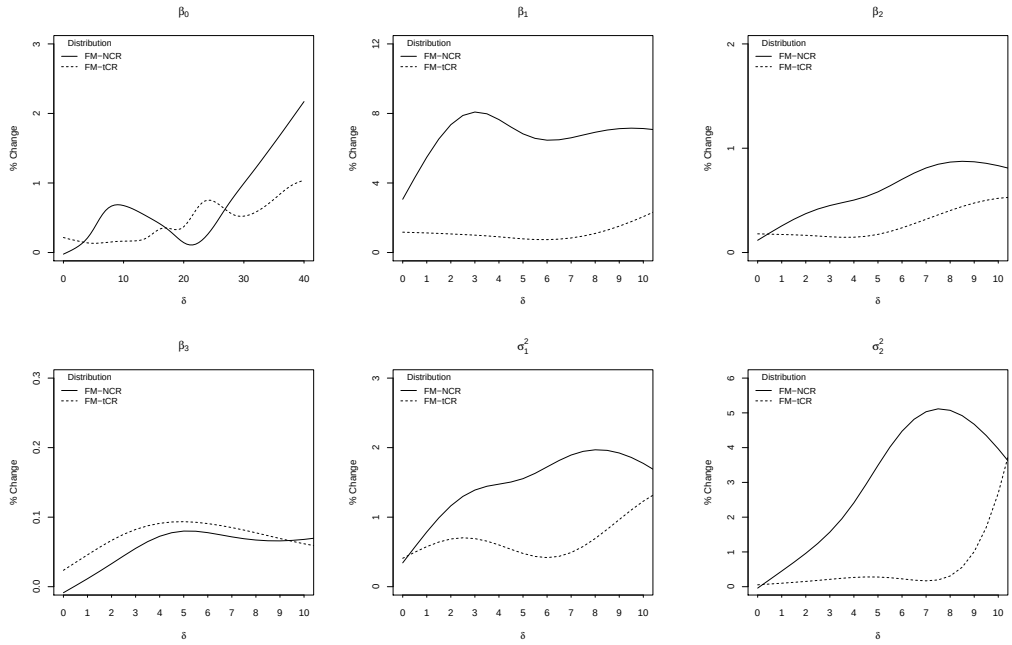

Figure 3: Wage rate data. Relative changes in the maximum-likelihood estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}$ from the FM-NCR (solid line) and FM-tCR (dashed line) models for different contaminations δ.

6 Conclusions

In this paper, a novel approach to censored linear regression analysis has been developed based on the use of finite mixtures of Student-t 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 (Chib, 1992) or a Student-t distribution (Arellano-Valle *et al.*, 2012; Garay *et al.*, 2015) or a finite mixture of Gaussian components (Karlsson & 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 Student-t distributions. The proposed EM algorithm was implemented as part of the R package *CensMixReg()* and is available for download at the CRAN repository.

Recently, Garay *et al.* (2015) considered the problem of censored linear regression models using the normal/independent (NI) distributions. Therefore, it would be a worthwhile task to investigate the applicability of a likelihood-based treatment in the context of finite mixtures of NI censored regression (FM-NICR) models. Other extensions of the current work include, for example, a generalization of FM-tCR linear model to skew-t distribution (Lachos *et al.*, 2010) and multivariate setting.

Acknowledgements Víctor H. Lachos acknowledges support from CNPq-Brazil (Grant 305054/2011-2) and FAPESP-Brazil (Grant 2014/02938-9). Celso R. B. Cabral was supported by CNPq (via BPPesq and Universal Project 2014), and FAPEAM (via Universal Amazonas Project). Luis B. Sánchez was supported by CAPES-Brazil.

References

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

Arellano-Valle, R., Castro, L., González-Farías, G. & Muñoz-Gajardo, K. (2012). Student-t censored regression model: properties and inference. *Statistical Methods & Applications*, **21**, 453–473.

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

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

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

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

Caudill, S. B. (2012). A partially adaptive estimator for the censored regression model based on a mixture of normal distributions. *Statistical Methods & Applications*, **21**(2), 121–137.

Chib, S. (1992). Bayes inference in the Tobit censored regression model. *Journal of Econometrics*, **51**, 79–99.

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

Efron, B. & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, pages 54–75.

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

Garay, A. M., Lachos, V. H., Bolfarine, H. & Cabral, C. R. (2015). Linear censored regression models with scale mixtures of normal distributions. *Statistical Papers*, **(In Press)**.

Genç, A. İ. (2013). Moments of truncated normal/independent distributions. *Statistical Papers*, **54**, 741–764.

Karlis, D. & Santourian, A. (2008). Model-based clustering with non-elliptically contoured distributions. *Statistics and Computing*, **19**, 73–83.

Karlsson, M. & Laitila, T. (2014). Finite mixture modeling of censored regression models. *Statistical Papers*, **55**(3), 627–642.

Kim, H. M. (2008). A note on scale mixtures of skew normal distribution. *Statistics and Probability Letters*, **78**. 1694-1701.

Lachos, V. H., Ghosh, P. & Arellano-Valle, R. B. (2010). Likelihood based inference for skew-normal independent linear mixed models. *Statistica Sinica*, **20**. SS-08-045.

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

Maehara, R. & Benites, L. (2015). bssn: Birnbaum-saunders skew normal. *R package version 0.5*.

Marron, J. S. & Wand, M. P. (1992). Exact mean integrated squared error. *The Annals of Statistics*, pages 712–736.

Massuia, M. B., Matos, L. A. & Lachos, V. H. (2012). *CensRegMod: Fitting Normal and Student-t censored regression model*. R package version 0.0.

Massuia, M. B., Cabral, C. R. B., Matos, L. A. & Lachos, V. H. (2015). Influence diagnostics for Student-t censored linear regression models. *Statistics*, **49**, 1074–1094.

Meng, X. & Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, **81**, 633–648.

Mroz, T. A. (1987). 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.

R Development Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Santana, L., Vilca, F. & Leiva, V. (2011). Influence analysis in skew-Birnbaum Saunders regression models and applications. *Journal of Applied Statistics*, **38**, 1633–1649.

Wei, G. C. G. & Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, **85**, 699–704.

Wu, L. (2010). *Mixed Effects Models for Complex Data*. Chapman & Hall/CRC, Boca Raton, FL.