Mixture of linear experts model for censored data: A novel approach with scale-mixture of normal distributions

Elham Mirfarah[a,*], Mehrdad Naderi[a], Ding-Geng Chen[a]

[a] *Department of Statistics, Faculty of Natural & Agricultural Sciences, University of Pretoria, Pretoria, South Africa*

Abstract

Mixture of linear experts (MoE) model is one of the widespread statistical frameworks for modeling, classification, and clustering of data. Built on the normality assumption of the error terms for mathematical and computational convenience, the classical MoE model has two challenges: 1) it is sensitive to atypical observations and outliers, and 2) it might produce misleading inferential results for censored data. The aim is then to resolve these two challenges, simultaneously, by proposing a robust MoE model for model-based clustering and discriminant censored data with the scale-mixture of normal (SMN) class of distributions for the unobserved error terms. An analytical expectation-maximization (EM) type algorithm is developed in order to obtain the maximum likelihood parameter estimates. Simulation studies are carried out to examine the performance, effectiveness, and robustness of the proposed methodology. Finally, a real dataset is used to illustrate the superiority of the new model.

Keywords: Mixture of linear experts model, Scale-mixture of normal class of distributions, EM-type algorithm, Censored data

1. Introduction

Clusterwise or mixture of regression model (MRM) has recently been considered in statistics for model-based clustering. When the population is heterogeneous and contains several latent source of heterogeneity, the MRM builds several regression models simultaneously, to investigate the relationship between the random phenomena under study. The subjects are then clustered based on the estimated posterior classification probabilities. Upon the normality or non-normality assumption for the mixing components, various MRMs have recently been introduced for modeling heterogeneous data. The classical G-component MRM (DeSarbo and Cron, 1988; Jones and McLachlan, 1992) specifically relies on the assumption that the conditional probability density function (pdf) of the response variable Y given the p-dimension explanatory vector $\boldsymbol{x} = (1, x_1, \ldots, x_{p-1})^\top \in \mathbb{R}^p$ is

$$f(y; \boldsymbol{\Theta}) = \sum_{j=1}^{G} \pi_j \, \phi(y; \boldsymbol{x}^\top \boldsymbol{\beta}_j, \boldsymbol{\sigma}_j^2), \tag{1}$$

where $\phi(\cdot; \mu, \sigma^2)$ stands for the pdf of normal distribution with location and scale parameters μ and σ^2, $\mathcal{N}(\mu, \sigma^2)$, $\boldsymbol{\beta}_j = (\beta_{j0}, \ldots, \beta_{j(p-1)})^\top$ is the jth component regression coefficients vector, and for $\boldsymbol{\theta}_j = (\boldsymbol{\beta}_j, \sigma_j^2)$ the model parameters set is $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G, \pi_1, \ldots, \pi_{G-1}\}$. Bear in mind that the mixing proportion with the constraint $\sum_{j=1}^{G} \pi_j = 1$, is in fact $\pi_j = Pr(Z^* = j)$, where the hidden categorical random variable Z^* indicates from which component each subject is arisen. Recently, the classical MRM (1) has found appealing applications in many fields, such as business, marketing, and biological studies, see Jiang and Tanner (1999); García-Escudero et al. (2010) and Mazza and Punzo (2017) to name a few. It has also been extended to accommodate heavy-tail and/or skew distributed data. In this regard, Liu and Lin (2014) proposed an MRM by replacing $\phi(\cdot)$ in (1) with the pdf of skew-normal (SN) distribution and applied

[*] Corresponding author
Email address: elham_mirfarah@yahoo.com (Elham Mirfarah)

it to the physiological data for illustration purposes. Hu et al. (2017) introduced an MRM by assuming that the components have log-concave densities and developed two EM-type (Dempster et al., 1977) algorithms to obtain the maximum likelihood (ML) parameter estimates. Moreover, Zeller et al. (2016) extended the mixture models based on the scale-mixture of SN (SMSN) class of distributions (Basso et al., 2010) into the regression context.

Built up from the MRM formulation, the MoE model (Jacobs et al., 1991) is perhaps one of the most acknowledged approaches in statistics and machine learning fields. Although the MoE model and MRM share similar structure, they differ in many aspects. In formulation of the MoE model, it is assumed that both mixing proportions and component densities conditionally depend on some input covariates. More precisely, let $Y \in \mathbb{R}$ be the response variable, $\boldsymbol{x} \in \mathbb{R}^p$ and $\boldsymbol{r} = (1, r_1, \ldots, r_{q-1})^\top \in \mathbb{R}^q$ are the vector of explanatory and covariate values corresponding to Y. Instead of considering constant mixing component in model (1), the MoE model assumes that π_j to be modeled as the multinomial logistic function of input $\boldsymbol{r}$, and is known as a gating function. For instance, extending the MRM (1), the pdf of the normal-based MoE (MoE-N) is

$$f(y; \boldsymbol{\Theta}) = \sum_{j=1}^{G} \pi_j(\boldsymbol{r}; \boldsymbol{\tau})\phi(y; \boldsymbol{x}^\top \boldsymbol{\beta}_j, \sigma_j^2), \tag{2}$$

where for the gating parameters $\boldsymbol{\tau} = (\boldsymbol{\tau}_1^\top, \ldots, \boldsymbol{\tau}_{G-1}^\top)^\top$ with $\boldsymbol{\tau}_j = (\tau_{j0}, \ldots, \tau_{j(q-1)})^\top$,

$$\pi_j(\boldsymbol{r}; \boldsymbol{\tau}) = Pr(Z^* = j|\boldsymbol{r}) = \frac{\exp\{\boldsymbol{\tau}_j^\top \boldsymbol{r}\}}{1 + \sum_{l=1}^{G-1} \exp\{\boldsymbol{\tau}_l^\top \boldsymbol{r}\}}, \tag{3}$$

and the model parameters set is $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G, \boldsymbol{\tau}\}$. It should be emphasized that $\boldsymbol{x}$ and $\boldsymbol{r}$ can be exactly or partially identical. Since the introduction of the MoE-N model, considerable amount of contributions have been produced to overcome its potential deficiency in analyzing skew and heavy-tail distributed data. See for instance the works by Nguyen and McLachlan (2016) and Chamroukhi (2016, 2017) on proposing the Laplace, Student-t and skew-t MoE models, respectively.

In many practical situations, such as economic and clinical studies, medical research and epidemiological cancer studies, the data are collected under some imposed detection limits. It might lead to incomplete data with different types of interval, left and/or right-censored responses. In this regard, censored regression model with the normality assumption for the error terms, known as Tobit model, was constructed by Tobin (1958). Since then, the extensions of Tobit model have been introduced by researchers to draw robust inference from censored data. For instance, using the SMN class of distributions for the error terms, Garay et al. (2016, 2017) presented the nonlinear and linear censored regression models to overcome the problem of atypical observations. Mattos et al. (2018) also proposed censored linear regression model with the SMSN class of distributions to accommodate asymmetrically distributed censored datasets. Moreover, mixture of censored regression models based on the Student-t model and on the SMN class of distributions were proposed by Lachos et al. (2019) and Zeller et al. (2019) as a flexible approach for modeling multimodal censored data with fat tails.

Extending the proven proficiency of the MoE model in statistical applications, the main objective of the current paper is to propose an MoE model based on the SMN class of distributions for censored data, hereafter referred as “MoE-SMN-CR model”. Due to the computational complexity, we develop an innovative EM-type algorithm to obtain the ML parameter estimates. The associated variance-covariance matrix of the ML estimators is also approximated by an information-based approach. To illustrate the computational aspects and practical performance of the proposed methodology, a real-data analysis and several simulation studies are presented.

The remainder of the paper is organized as follows. Section 2 briefly reviews the SMN class of distributions. Model formulation and parameter estimation procedure of the MoE-SMN-CR model are presented in Section 3. Five simulation studies are conducted in Section 4 to verify the asymptotic properties of the ML estimates as well as to investigate the performance of the proposed model. The applicability of the proposed method is illustrated in Section 5 by analyzing wage-rates dataset. Finally, we conclude the paper with a discussion and suggestions for future work in Section 6.

2. An overview on the scale-mixture of normal class of distributions

A random variable Y follows an SMN distribution, denoted by $\mathcal{SMN}(\mu, \sigma^2, \boldsymbol{\nu})$, if it is generated by the representation

$$Y = \mu + U^{-1/2} V, \qquad V \perp U, \tag{4}$$

where $V \sim \mathcal{N}(0, \sigma^2)$, U (scale mixture factor) is a positive random variable with the cumulative distribution function (cdf) $H(\cdot; \boldsymbol{\nu})$, and the symbol '$\perp$' indicates independence. Referring to (4), the hierarchical representation of the SMN class of distributions can be written as

$$Y|U = u \sim N(\mu, u^{-1}\sigma^2), \qquad U \sim H(u; \boldsymbol{\nu}). \tag{5}$$

Accordingly, the pdf of random variable Y is obtained by

$$f_{\text{SMN}}(y; \mu, \sigma^2, \boldsymbol{\nu}) = \int_0^\infty \phi(y; \mu, u^{-1}\sigma^2)\, dH(u; \boldsymbol{\nu}), \qquad y \in \mathbb{R}.$$

In what follows, $f_{\text{SMN}}(\cdot; \boldsymbol{\nu})$ and $F_{\text{SMN}}(\cdot; \boldsymbol{\nu})$ will be used to denote the pdf and cdf of the standard SMN distribution ($\mu = 0, \sigma^2 = 1$). With different specifications of the distribution of U, many special cases of the general SMN class of distributions can be obtained. We focus on a few commonly used examples of the SMN class of distributions in this paper:

- Normal (N) distribution: The SMN class of distributions contains the normal model as $U = 1$ with probability one.

- Student-t (T) distribution: If $U \sim Gamma(\nu/2, \nu/2)$, where $Gamma(\alpha,\beta)$ represents the gamma distribution with shape and scale parameters α and β, respectively, the random variable Y then follows the Student-t distribution, $Y \sim \mathcal{T}(\mu, \sigma^2, \nu)$. For $\nu = 1$ the Student-t distribution turns into the Cauchy distribution which has no defined mean and variance.

- Slash (SL) distribution: Let U in (4) follows $Beta(\nu, 1)$, where $Beta(a, b)$ signifies the beta distribution with parameter a and b. Then, Y distributed as a slash model, denoted by $Y \sim \mathcal{SL}(\mu, \sigma^2, \nu)$, with pdf

$$f_{\text{SL}}(y; \mu, \sigma^2, \nu) = \nu \int_0^1 u^{\nu-1} \phi(y; \mu, u^{-1}\sigma^2)\, du, \qquad y \in \mathbb{R}.$$

- Contaminated-normal (CN) distribution: Let U be a discrete random variable with pdf

$$h(u; \nu, \gamma) = \nu \mathbb{I}_\gamma(u) + (1 - \nu)\mathbb{I}_1(u), \qquad \nu, \gamma \in (0, 1),$$

  where $\mathbb{I}_A(\cdot)$ represents the indicator function of the set A. The random variable Y in (4) then follows the contaminated-normal distribution, $Y \sim \mathcal{CN}(\mu, \sigma^2, \nu, \gamma)$, which has the pdf

$$f_{\text{CN}}(y; \mu, \sigma^2, \nu, \gamma) = \nu \phi(y; \mu, \gamma^{-1}\sigma^2) + (1 - \nu)\phi(y; \mu, \sigma^2), \qquad y \in \mathbb{R}.$$

  Note that in the pdf of CN distribution, the parameter ν denotes the proportion of outliers (bad points) and γ is the contamination factor.

More technical details and information of the SMN class of distributions, used for the calculation of some conditional expectations involved in the proposed EM-type algorithm, are provided in the Appendix A with proof in Garay et al. (2017). We will refer to the MoE model of censored data based on the special cases of the SMN class of distributions as MoE-N-CR, MoE-T-CR, MoE-SL-CR and MoE-CN-CR for the normal, Student-t, slash and contaminated-normal cases, respectively.

3. The scale-mixture of normal censored mixture of linear experts model

3.1. Model specification

Extending the classical MoE model with normal distribution in (2), we consider the expert components formulated by the SMN class of distributions. Therefore, the resulting pdf of the response random vector $\boldsymbol{Y} = (Y_1, \dots, Y_n)^\top$, in which the polynomial regression and multinomial logistic model are used for the components and mixing proportions, can be defined as

$$f(y_i; \boldsymbol{\Theta}) = \sum_{j=1}^{G} \pi_j(\boldsymbol{r}_i; \boldsymbol{\tau}) f_{\text{SMN}}(y_i; \boldsymbol{x}_i^\top \boldsymbol{\beta}_j, \sigma_j^2, \boldsymbol{\nu}_j), \qquad i = 1, \dots, n, \tag{6}$$

where $\boldsymbol{x}_i$ and $\boldsymbol{r}_i$ are the vector of explanatory and covariate variables corresponding to Y_i, $\pi_j(\cdot; \boldsymbol{\tau})$ is defined in (3), and for $\boldsymbol{\theta}_j = (\boldsymbol{\beta}_j, \sigma_j^2, \boldsymbol{\nu}_j)$ the model parameters set is $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_G, \boldsymbol{\tau}\}$.

In the MoE-SMN-CR model, we assume that the response vector $\boldsymbol{Y}$ is partially observed. In other words, we suppose some of the response variables are suffering from a type of censoring, that could be interval-, left- or right-censoring. Thus, let the available response variable Y_i be presented as the joint variables (W_i, ρ_i) where W_i represents the uncensored reading $(W_i = Y_{Oi})$ or interval-censoring $(W_i = (c_{i1}, c_{i2})$ for some fixed threshold points $c_{i1}, c_{i2})$ and ρ_i is the censoring indicator: $\rho_i = 1$ if $c_{i1} \le Y_i \le c_{i2}$ and $\rho_i = 0$ if $Y_i = Y_{Oi}$. Note that in this setting if $c_{i1} = -\infty$ (or $c_{i2} = +\infty$) the left-censoring (or right-censoring) is occurred and in case $-\infty \neq c_{i1} < c_{i2} \neq +\infty$ the interval-censored realization is observed. We establish our methodology based on the interval-censoring scheme, however, the left- and right-censoring schemes are also investigated in the simulation and real-data analyses.

The aforementioned setting leads to divide $\boldsymbol{Y}$ to the sets of observed responses and censored cases. Hence, $\boldsymbol{Y}$ can be viewed as the latent variable since it is partially unobserved. Under these assumptions, the log-likelihood function of the MoE-SMN-CR model can be written as

$$\ell(\boldsymbol{\Theta}|\boldsymbol{w}, \boldsymbol{\rho}) = \sum_{i=1}^{n} \log \sum_{j=1}^{G} \pi_j(\boldsymbol{r}_i; \boldsymbol{\tau}) \left[\sigma_j^{-1} f_{\text{SMN}}\left(\frac{y_{Oi} - \boldsymbol{x}_i^\top \boldsymbol{\beta}_j}{\sigma_j}; \boldsymbol{\nu}_j \right) \right]^{1-\rho_i} \left[F_{\text{SMN}}\left(\frac{c_{i2} - \boldsymbol{x}_i^\top \boldsymbol{\beta}_j}{\sigma_j}; \boldsymbol{\nu}_j \right) - F_{\text{SMN}}\left(\frac{c_{i1} - \boldsymbol{x}_i^\top \boldsymbol{\beta}_j}{\sigma_j}; \boldsymbol{\nu}_j \right) \right]^{\rho_i}, \tag{7}$$

where y_{Oi} denotes the realization of Y_{Oi}.

Due to complexity of the log-likelihood (7), there is no analytical solution to obtain the ML estimate of parameters and therefore a numerical search algorithm should be developed. With the embedded hierarchical representation (5), an innovative EM-type algorithm is developed to obtain the ML estimate for the MoE-SMN-CR model.

3.2. EM-based maximum likelihood parameter estimation

Starting from (6) and defining the component label vector $\boldsymbol{Z}_i = (Z_{i1}, \dots, Z_{iG})^\top$ in such a way that the binary latent component-indicators $Z_{ij} = 1$ if and only if $Z_i^* = j$, we have

$$Y_i | Z_{ij} = 1 \sim \mathcal{SMN}(\boldsymbol{x}_i^\top \boldsymbol{\beta}_j, \sigma_j^2, \boldsymbol{\nu}_j), \quad i = 1, \dots, n, \quad j = 1, \dots, G.$$

Now using (5), the hierarchical representation of the MoE-SMN-CR model is

$$\begin{aligned} Y_i | (\boldsymbol{x}_i, U = u_i, Z_{ij} = 1) &\sim \mathcal{N}(\boldsymbol{x}_i^\top \boldsymbol{\beta}_j, u_i^{-1} \sigma_j^2), \\ U_i | Z_{ij} = 1 &\sim H(u_i; \boldsymbol{\nu}_j), \\ Z_i | \boldsymbol{r}_i &\sim \mathcal{M}(1; \pi_1(\boldsymbol{r}_i, \boldsymbol{\tau}), \dots, \pi_G(\boldsymbol{r}_i, \boldsymbol{\tau})), \end{aligned}$$

where $\mathcal{M}(1; \cdot)$ denotes the one trail multinomial distribution. For the realization $\boldsymbol{y} = (y_1, \dots, y_n)^\top$, vector of censoring indicators $\boldsymbol{\rho} = (\rho_1, \dots, \rho_n)^\top$, and hidden values $\boldsymbol{u} = (u_1, \dots, u_n)^\top$ and $\boldsymbol{Z} = (\boldsymbol{Z}_1^\top, \dots, \boldsymbol{Z}_n^\top)^\top$, the log-likelihood function for $\boldsymbol{\Theta}$ associated with complete data $\boldsymbol{y}_c = (\boldsymbol{w}^\top, \boldsymbol{\rho}^\top, \boldsymbol{y}^\top, \boldsymbol{u}^\top, \boldsymbol{Z}^\top)^\top$, is therefore given by

$$\ell_c(\boldsymbol{\Theta}|\boldsymbol{y}_c) \approx \sum_{i=1}^{n} \sum_{j=1}^{G} Z_{ij} \Big\{ \log \pi_j(\boldsymbol{r}_i; \boldsymbol{\tau}) - \frac{1}{2} \log \sigma_j^2 - \frac{u_i}{2\sigma_j^2} (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_j)^2 + \log h(u_i; \boldsymbol{\nu}_j) \Big\}, \tag{8}$$

where $h(\cdot; \boldsymbol{\nu}_j)$ is the pdf of $U_i | Z_{ij} = 1$.

We then develop an expectation conditional maximization either (ECME; Liu and Rubin (1994)) algorithm to estimate parameters of the MoE-SMN-CR model. The ECME algorithm is an extension of expectation conditional maximization (ECM; Meng and Rubin (1993)) that not only inherits its stable properties (e.g. monotone convergence and implementation simplicity) but it can also be faster than ECM. The iterative ECME algorithm replaces some CM-steps of the ECM with the CML-steps that maximize the corresponding constrained log-likelihood function instead. The ECME algorithm for ML estimation of the MoE-SMN-CR model proceeds as follows:

- **Initialization:** Set the number of iteration to $k = 0$ and choose a relative starting point. Due to the multimodal log-likelihood function in the mixture and MoE models, the EM-type algorithm for obtaining parameter estimates might not give the global estimates if the initial points depart too far from the real values. Therefore, the choice of initialization of the EM-based algorithms constitutes an fundamental issue. Nguyen and McLachlan (2016) suggested the starting points for the Laplace MoE model via a modified version of the randomized initial assignment method (McLachlan and Peel, 2000). However, we recommend the following straightforward steps for obtaining the starting points of the MoE-SMN-CR model.

  (i) Partition the sample into G groups using either K-means clustering algorithm (Hartigan and Wong, 1979), k-medoids (Kaufman and Rousseeuw, 1990) or trim-k-means (Cuesta-Albertos et al., 1997) methods.

  (ii) To initialize $\boldsymbol{\tau}$, two strategies can be adopted. As the first and simplest strategy, one can set $\hat{\boldsymbol{\tau}}^{(0)} = \mathbf{0}$. We note that by using this setting, the MoE model reduces to the MRM as a special case. In the second strategy, the information of grouping indices obtained from (i) can be used for initializing $\boldsymbol{\tau}$. Based on the grouping indices, one can fit the generalized linear model to the data and compute $\hat{\boldsymbol{\tau}}^{(0)}$.

  (iii) By utilizing the grouping indices of (i), the least squares method is applied to the jth group to obtain $\hat{\boldsymbol{\beta}}_j^{(0)}$. Moreover, the standard deviation of residuals is used to initialize σ_j^2.

  (iv) Since the normal model belongs to the SMN class of distributions, we adapt $\hat{\boldsymbol{\nu}}_j^{(0)}$ corresponds to an initial assumption near normality. For instance, we set $\hat{\nu}_j^{(0)} = 20$ in the MoE-T-CR and MoE-SL-CR models.

- **E-Step:** At iteration k, the expected value of the complete-data log-likelihood function (8), known as the Q-function, is calculated as

$$Q(\boldsymbol{\Theta}|\hat{\boldsymbol{\Theta}}^{(k)}) = \sum_{i=1}^{n}\sum_{j=1}^{G} \hat{z}_{ij}^{(k)} \left\{ \log \pi_j(\boldsymbol{r}_i;\boldsymbol{\tau}) - \frac{1}{2}\log\sigma_j^2 - \frac{1}{2\sigma_j^2}\left(\widehat{uy^2}_{ij}^{(k)} + (\boldsymbol{x}_i^\top\boldsymbol{\beta}_j)^2\hat{u}_{ij}^{(k)} - 2\widehat{uy}_{ij}^{(k)}\boldsymbol{x}_i^\top\boldsymbol{\beta}_j\right) + \hat{\Upsilon}_{ij}^{(k)} \right\}, \quad (9)$$

  where $\hat{z}_{ij}^{(k)} = E(Z_{ij}|w_i,\rho_i,\hat{\boldsymbol{\theta}}_j^{(k)})$, $\widehat{uy^2}_{ij}^{(k)} = E(U_iY_i^2|w_i,\rho_i,\hat{\boldsymbol{\theta}}_j^{(k)})$, $\hat{u}_{ij}^{(k)} = E(U_i|w_i,\rho_i,\hat{\boldsymbol{\theta}}_j^{(k)})$, $\widehat{uy}_{ij}^{(k)} = E(U_iY_i|w_i,\rho_i,\hat{\boldsymbol{\theta}}_j^{(k)})$, and $\hat{\Upsilon}_{ij}^{(k)} = E(\log h(U_i;\boldsymbol{\nu}_j)|w_i,\rho_i,\hat{\boldsymbol{\theta}}_j^{(k)})$. In what follows, we discuss about the computation of conditional expectations for both uncensored and censored cases.

  (i) For the uncensored observations, we have $\rho_i = 0$, $w_i = y_{Oi}$, and so, $\hat{u}_{ij}^{(k)} = E(U_i|Y = y_{Oi},\hat{\boldsymbol{\theta}}_j^{(k)})$, $\widehat{uy}_{ij}^{(k)} = y_{Oi}\hat{u}_{ij}^{(k)}$, $\widehat{uy^2}_{ij}^{(k)} = y_{Oi}^2\hat{u}_{ij}^{(k)}$,

$$\hat{z}_{ij}^{(k)} = \frac{\pi_j(\boldsymbol{r}_i;\hat{\boldsymbol{\tau}}^{(k)})f_{\text{SMN}}\left(y_{Oi};\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_j^{(k)},\hat{\sigma}_j^{2(k)},\hat{\boldsymbol{\nu}}_j^{(k)}\right)}{\sum_{l=1}^{G}\pi_l(\boldsymbol{r}_i;\hat{\boldsymbol{\tau}}^{(k)})f_{\text{SMN}}\left(y_{Oi};\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_l^{(k)},\hat{\sigma}_l^{2(k)},\hat{\boldsymbol{\nu}}_l^{(k)}\right)}, \qquad \hat{\Upsilon}_{ij}^{(k)} = E(\log h(U_i;\boldsymbol{\nu}_j)|Y = y_{Oi},\hat{\boldsymbol{\theta}}_j^{(k)}).$$

  (ii) For the censored case which is $\rho_i = 1$ and $w_i = (c_{i1}, c_{i2})$, we have

$$\hat{z}_{ij}^{(k)} = E(Z_{ij}|c_{i1}\le Y_i\le c_{i2},\hat{\boldsymbol{\theta}}_j^{(k)}) = \frac{\pi_j(\boldsymbol{r}_i;\hat{\boldsymbol{\tau}}^{(k)})\left[F_{\text{SMN}}\left(\frac{c_{i2}-\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_j^{(k)}}{\hat{\sigma}_j^{(k)}};\hat{\boldsymbol{\nu}}_j^{(k)}\right) - F_{\text{SMN}}\left(\frac{c_{i1}-\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_j^{(k)}}{\hat{\sigma}_j^{(k)}};\hat{\boldsymbol{\nu}}_j^{(k)}\right)\right]}{\sum_{l=1}^{G}\pi_l(\boldsymbol{r}_i;\hat{\boldsymbol{\tau}}^{(k)})\left[F_{\text{SMN}}\left(\frac{c_{i2}-\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_l^{(k)}}{\hat{\sigma}_l^{(k)}};\hat{\boldsymbol{\nu}}_l^{(k)}\right) - F_{\text{SMN}}\left(\frac{c_{i1}-\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_l^{(k)}}{\hat{\sigma}_l^{(k)}};\hat{\boldsymbol{\nu}}_l^{(k)}\right)\right]},$$

$$\hat{u}_{ij}^{(k)} = E(U_i|c_{i1}\le Y_i\le c_{i2},\hat{\boldsymbol{\theta}}_j^{(k)}), \qquad \widehat{uy^2}_{ij}^{(k)} = E(U_iY_i^2|c_{i1}\le Y_i\le c_{i2},\hat{\boldsymbol{\theta}}_j^{(k)}),$$

$$\widehat{uy}_{ij}^{(k)} = E(U_iY_i|c_{i1}\le Y_i\le c_{i2},\hat{\boldsymbol{\theta}}_j^{(k)}), \qquad \hat{\Upsilon}_{ij}^{(k)} = E(\log h(U_i;\boldsymbol{\nu}_j)|c_{i1}\le Y_i\le c_{i2},\hat{\boldsymbol{\theta}}_j^{(k)}).$$

Following Garay et al. (2017), the closed form of the conditional expectations for the particular cases of the SMN class of distributions are provided in Appendix A.

For updating $\hat{\boldsymbol{\Theta}}^{(k)}$, the CM-steps are implemented by maximizing Q-function (9) as follows:

- **CM-step 1:** Calculate $\hat{\boldsymbol{\beta}}_j^{(k)}$ and $\hat{\sigma}_j^{2(k)}$ updates as

$$\hat{\boldsymbol{\beta}}_j^{(k+1)} = \left(\sum_{i=1}^{n} \hat{z}_{ij}^{(k)} \hat{u}_{ij}^{(k)} \boldsymbol{x}_i \boldsymbol{x}_i^{\top} \right)^{-1} \sum_{i=1}^{n} \hat{z}_{ij}^{(k)} \widehat{uy}_{ij}^{(k)} \boldsymbol{x}_i,$$

$$\hat{\sigma}_j^{2(k+1)} = \frac{1}{n_j} \sum_{i=1}^{n} \hat{z}_{ij}^{(k)} \left(\widehat{uy^2}_{ij}^{(k)} - 2\widehat{uy}_{ij}^{(k)} \boldsymbol{x}_i^{\top} \hat{\boldsymbol{\beta}}_j^{(k+1)} + \hat{u}_{ij}^{(k)} \left(\boldsymbol{x}_i^{\top} \hat{\boldsymbol{\beta}}_j^{(k+1)} \right)^2 \right),$$

  where $n_j = \sum_{i=1}^{n} \hat{z}_{ij}^{(k)}$.

- **CM-step 2:** Following Proposition 2 of Nguyen and McLachlan (2016), the update of $\hat{\boldsymbol{\tau}}_j^{(k)}$ can be made as

$$\hat{\boldsymbol{\tau}}_j^{(k+1)} = 4 \left(\sum_{i=1}^{n} \boldsymbol{r}_i \boldsymbol{r}_i^{\top} \right)^{-1} \left(\sum_{i=1}^{n} \left[\hat{z}_{ij}^{(k+1)} - \pi_j(\boldsymbol{r}_i; \hat{\boldsymbol{\tau}}^{(k)}) \right] \boldsymbol{r}_i \right) + \hat{\boldsymbol{\tau}}_j^{(k)}.$$

- **CML-step:** The update of $\hat{\boldsymbol{\nu}}_j^{(k)}$ crucially depends on the conditional expectation $\hat{\Upsilon}_{ij}^{(k)}$ which is quite complicated. However, we can estimate $\boldsymbol{\upsilon} = (\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_G)$ through maximizing the restricted actual log-likelihood function as

$$\begin{aligned} \hat{\boldsymbol{\upsilon}}^{(k+1)} \quad = \quad & \arg\max_{\boldsymbol{\upsilon}} \left\{ \sum_{i=1}^{n} \log \sum_{j=1}^{G} \pi_j(\boldsymbol{r}_i; \hat{\boldsymbol{\tau}}^{(k+1)}) \left[f_{\text{SMN}} \left(\frac{y_{Oi} - \boldsymbol{x}_i^{\top} \hat{\boldsymbol{\beta}}_j^{(k+1)}}{\hat{\sigma}_j^{(k+1)}}; \boldsymbol{\nu}_j \right) / \hat{\sigma}_j^{(k+1)} \right]^{1-\rho_i} \right. \\ & \left. \left[F_{\text{SMN}} \left(\frac{c_{i2} - \boldsymbol{x}_i^{\top} \hat{\boldsymbol{\beta}}_j^{(k+1)}}{\hat{\sigma}_j^{(k+1)}}; \boldsymbol{\nu}_j \right) - F_{\text{SMN}} \left(\frac{c_{i1} - \boldsymbol{x}_i^{\top} \hat{\boldsymbol{\beta}}_j^{(k+1)}}{\hat{\sigma}_j^{(k+1)}}; \boldsymbol{\nu}_j \right) \right]^{\rho_i} \right\}. \end{aligned} \tag{10}$$

  Recommended by Lin et al. (2014) and Zeller et al. (2019), a more parsimonious model can be achieved by assuming the identical mixing component, i.e. $\boldsymbol{\nu}_1 = \boldsymbol{\nu}_2 = \cdots = \boldsymbol{\nu}_G = \boldsymbol{\nu}$. This setting changes the problem of nontrivial high-dimension optimization into the more simple one/two dimensional search. The R function **nlminb**( ) is used to update $\boldsymbol{\upsilon}$ in the numerical parts of the current paper.

The above E- and M-steps are iterated until some convergence criteria are met. We terminate the algorithm when either the maximum number of iterations approaches 1000 or the difference between two consecutive log-likelihood values is less than the per-specified tolerance 10^{-5}.

Remark 1. *To facilitate the estimation of $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_G)$ for the MoE-CN-CR model in the above EM algorithm, one can introduce an extra latent binary variable B_i such that $B_i = 1$ if an observation y_i in group g is a bad point and $B_i = 0$ otherwise. The hierarchical representation of the MoE-CN-CR model can therefore be written as*

$$\begin{aligned} Y_i | (\boldsymbol{x}_i, U = u_i, Z_{ij} = 1, B_i = 1) &\sim \mathcal{N}(\boldsymbol{x}_i^{\top} \boldsymbol{\beta}_j, u_i^{-1} \sigma_j^2), \\ U_i | (Z_{ij} = 1, B_i = 1) &\sim h(u_i; \nu_j, \gamma_j), \\ B_i | (Z_{ij} = 1) &\sim \mathcal{B}(1, \nu_j), \\ Z_i | \boldsymbol{r}_i &\sim \mathcal{M}(1; \pi_1(\boldsymbol{r}_i, \boldsymbol{\tau}), \ldots, \pi_G(\boldsymbol{r}_i, \boldsymbol{\tau})), \end{aligned} \tag{11}$$

where $\mathcal{B}(1, \nu_j)$ denotes the Bernoulli distribution with succeed probability ν_j. Consequently, by computing the Q-function based on (11), the update of $\nu_j^{(k)}$ is

$$\hat{\nu}_j^{(k+1)} = \frac{\sum_{i=1}^{n} \hat{z}_{ij}^{(k)} \hat{b}_{ij}^{(k)}}{\sum_{i=1}^{n} \hat{z}_{ij}^{(k)}},$$

where

$$\hat{b}_{ij}^{(k)} = \begin{cases} \dfrac{\hat{\nu}_j^{(k)}\phi(y_{Oi};\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_j^{(k)},\hat{\gamma}_j^{-1(k)}\hat{\sigma}_j^{2(k)})}{\hat{\nu}_j^{(k)}\phi(y_{Oi};\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_j^{(k)},\hat{\gamma}_j^{-1(k)}\hat{\sigma}_j^{2(k)}) + (1-\hat{\nu}_j^{(k)})\phi(y_{Oi};\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_j^{(k)},\hat{\sigma}_j^{2(k)})}, & \textit{for the uncensoed cases,} \\[2ex] \dfrac{\hat{\nu}_j^{(k)}\Big(\Phi(c_{i2};\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_j^{(k)},\hat{\gamma}_j^{-1(k)}\hat{\sigma}_j^{2(k)}) - \Phi(c_{i1};\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_j^{(k)},\hat{\gamma}_j^{-1(k)}\hat{\sigma}_j^{2(k)})\Big)}{F_{CN}(c_{i2};\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_j^{(k)},\hat{\sigma}_j^{2(k)},\hat{\nu}_j^{(k)},\hat{\gamma}_j^{(k)}) - F_{CN}(c_{i1};\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_j^{(k)},\hat{\sigma}_j^{2(k)},\hat{\nu}_j^{(k)},\hat{\gamma}_j^{(k)})}, & \textit{for the censoed cases.} \end{cases}$$

Since there is no closed-form solution for $\hat{\gamma}_j^{(k+1)}$, it is computed by maximizing the constrained actual observed log-likelihood function (10) *as a function of $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_G)$.*

3.3. Computational and operational aspects

3.3.1. Model selection and performance assessment

In practical model-based clustering, the number of components G is not known and should be estimated from the data. In this regard, one can fit a mixture model for a range of values G and choose the best one based on some model selection criteria. The two commonly used measures, Akaike information criterion (AIC; Akaike (1974)) and Bayesian information criterion (BIC; Schwarz et al. (1978)), are exploited to determine the most plausible value of G. The AIC and BIC are defined as

$$\text{AIC} = 2m - 2\ell_{\max} \qquad \text{and} \qquad \text{BIC} = m \ln n - 2\ell_{\max},$$

where $\ell_{\max}$ is the maximized (observed) log-likelihood, and m the number of free parameters in the model. Although the smallest value of AIC or BIC results in the most favored model, they do not necessarily correspond to optimal clustering. For the sake of classification performance, the misclassification error rate (MCR), Jaccard coefficient index (JCI; Niwattanakul et al. (2013)), Rand index (RI; Rand (1971)) and adjusted Rand index (ARI; Hubert and Arabie (1985)) are used when the true group labels are known. Noted that the lower MCR (close to zero) or a higher RI and JCI (tend to one) indicates a much similarity between the true labels and the cluster labels obtained by the candidate model. An ARI of one also corresponds to perfect agreement, and the expected value of the ARI under random classification is zero. Negative ARI values are possible and indicate classification results that are worse, in some sense, than would be expected by random classification.

3.3.2. Note on computing conditional expectations

As expressed in Appendix A, the conditional expectations of the MoE-SMN-CR sub-models critically depend on the hazard function or the cdf of SMN model. For instance, in the left-censoring scheme, $\widehat{uy}_{ij}^{(k)}$ for the MoE-N-CR model depends on the hazard function of normal distribution as $HF(x) = \phi(x)/\Phi(x)$. The computation of this hazard function for very small values of x (say $x < -35$ as encountered many times in the simulation studies) in R may lead to "NaN". To overcome this issue, Filho and Garay (2017) in the R package "**TSMN**" and Zeller et al. (2019) in the R package "**CensMixReg**" set the denominator to the small machine value (the R command ".Machine$double.xmin" was used). However, this setting may lead to negative value for $\hat{\sigma}^2$ as we found. We recommend to use a remedy for obtaining the exact values of $HF(x)$. In our computation, we have used log-transformation via the following R command

$$HF.x = \exp\Big(\text{dnorm}(x, log = T) - \text{pnorm}(x, log.p = T)\Big).$$

Figure B.8 in the Appendix B highlights the difference of three ways of the HF computation in R. What is observed from Figure B.8 is actually the difference between the computation of $HF(x)$ function for $x < -35$. Similar trick can be applied for the right- and interval-censoring schemes.

3.3.3. Standard error estimates

For estimating the standard error of the ML estimators, we follow Meilijson (1989) to exploit an information-based method for calculating the asymptotic covariance matrix of the ML estimates. Let ℓ_{ci} be the complete-data log-likelihood contributed from the ith observation, viz.

$$\ell_{ci} = \ell_c(\mathbf{\Theta}|\boldsymbol{w}_i^\top, \boldsymbol{\rho}_i^\top, \boldsymbol{y}_i^\top, \boldsymbol{u}_i^\top, \boldsymbol{Z}_i^\top) = \sum_{j=1}^{G} Z_{ij} \left\{ \log \pi_j(\boldsymbol{r}_i; \boldsymbol{\tau}) - \frac{1}{2} \log \sigma_j^2 - \frac{u_i}{2\sigma_j^2}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_j)^2 + \log h(u_i; \boldsymbol{\nu}_j) \right\}.$$

Then, the Fisher information matrix can be approximated by

$$I_o(\hat{\Theta}|\boldsymbol{y}) = \sum_{i=1}^{n} \hat{\boldsymbol{s}}_i \hat{\boldsymbol{s}}_i^\top,$$

where $\hat{s}_i = E\Big(\frac{\partial \ell_{ci}}{\partial \mathbf{\Theta}} \Big| w_i, \rho_i, \hat{\mathbf{\Theta}}\Big)$ is the individual score vector corresponding to the ith observation. The elements of individual score vector $(\hat{s}_{i,\boldsymbol{\tau}_1}^\top, \ldots, \hat{s}_{i,\boldsymbol{\tau}_{G-1}}^\top \hat{s}_{i,\boldsymbol{\beta}_1}^\top, \ldots, \hat{s}_{i,\boldsymbol{\beta}_G}^\top, \hat{s}_{i,\sigma_1^2}, \ldots, \hat{s}_{i,\sigma_G^2})$ have the explicit forms as

$$\begin{aligned}
\hat{s}_{i,\boldsymbol{\tau}_j} &= E\Big(\frac{\partial \ell_{ci}}{\partial \boldsymbol{\tau}_j} \Big| w_i, \rho_i, \hat{\mathbf{\Theta}}\Big) = \Big(\hat{z}_{ij} - \pi_j(\boldsymbol{r}_i; \hat{\boldsymbol{\tau}})\Big) \boldsymbol{r}_i, \\
\hat{s}_{i,\boldsymbol{\beta}_j} &= E\Big(\frac{\partial \ell_{ci}}{\partial \boldsymbol{\beta}_j} \Big| w_i, \rho_i, \hat{\mathbf{\Theta}}\Big) = \frac{\hat{z}_{ij}}{\hat{\sigma}_j^2} \Big(\widehat{uy}_{ij} \boldsymbol{x}_i - \hat{u}_{ij} \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_j \boldsymbol{x}_i\Big), \\
\hat{s}_{i,\sigma_j^2} &= E\Big(\frac{\partial \ell_{ci}}{\partial \sigma_j^2} \Big| w_i, \rho_i, \hat{\mathbf{\Theta}}\Big) = -\frac{\hat{z}_{ij}}{2\hat{\sigma}_j^4} \Big(\hat{\sigma}_j^2 - \widehat{uy^2}_{ij} - \hat{u}_{ij}(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_j)^2 + 2\widehat{uy}_{ij} \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_j\Big).
\end{aligned}$$

As a result, the variance of the ML estimates can be consistently estimated from the diagonal of the inverse of $I_o(\hat{\Theta}|\boldsymbol{y})$ under some regularity conditions. We note that the standard error of $\hat{\nu}$ critically depends on the calculation of $E(\log(U_i)|w_i, \rho_i, \hat{\mathbf{\Theta}})$ which is a computational challenge. It could be mentioned that the inverse of $I_o(\hat{\Theta}|\boldsymbol{y})$ is not always available. One can refer to Yu et al. (2021) to find an innovative interpolation procedure based on the cubic spline interpolation to directly estimate the asymptotic variance-covariance matrix of the ML estimates obtained by the EM algorithm.

4. Monte-Carlo simulation studies

In this section, five Monte-Carlo simulation studies are conducted in order to verify the asymptotic properties of the ML estimates, to assess the fitting and clustering performance of the model, and to check the robustness of the proposed model in dealing with highly peaked and heavily tailed data as well as its sensitivity in presence of outliers.

4.1. Data generation

Note that one of the simplest and straightforward ways for generating interval-censored data is to define the random thresholds as $C_{i1} = Y_i - U_i^{(1)}$ and $C_{i2} = Y_i + U_i^{(2)}$ such that the non-informative condition (1.2) of Gómez et al. (2009) is met. Here the continuous variables $U_i^{(1)}$ and $U_i^{(2)}$ are independently distributed by $\mathcal{U}(0, c)$, where the notation $\mathcal{U}(a, b)$ stands for the uniform distribution on interval (a, b). Recommended by Gómez et al. (2009), a way to go around the non-informative condition is to construct $C_{i1} = \max(Y_i - U_i^{(1)}, Y_i + U_i^{(2)} - c)$ and $C_{i2} = \min(Y_i + U_i^{(2)}, Y_i - U_i^{(1)} + c)$ with $c = 1$. In short, suppose we generate n realizations from model (6), $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$. To have a p% interval-censored data, the following steps are used in our simulation studies.

S_1) Calculate the number of censored samples $\mathcal{NC} = [n \times p] + 1$, where $[a]$ denotes the greatest integer less than or equal to a. Then, generate an index set, $\mathcal{IND}$, as a sample of size $\mathcal{NC}$ from the set $\{1, 2, \cdots, n\}$ without replacement. Use *sample*( ) function in R for this purpose.

S_2) For $i = 1, \ldots, n$, if $i \in \mathcal{IND}$, then

S_{21}) Generate two independent random samples, $u_i^{(1)}$ and $u_i^{(2)}$, from $\mathcal{U}(0, c)$.

S_{22}) Set the thresholds to $c_{i1} = \max(y_i - u_i^{(1)}, y_i + u_i^{(2)} - c), \quad c_{i2} = \min(y_i + u_i^{(2)}, y_i - u_i^{(1)} + c)$.

4.2. Asymptotic properties of the ML estimates

In this section, a simulation study is performed to examine the asymptotic properties of the ML parameter estimates obtained through the ECME algorithm. We simulate 500 Monte-Carlo samples from the special cases of the MoE-SMN-CR model with $G = 2$. The presumed parameters are

$$\boldsymbol{\beta}_1 = (0, -1, -2, -3)^\top, \quad \boldsymbol{\beta}_2 = (-1, 1, 2, 3)^\top, \quad (\sigma_1^2, \sigma_2^2) = (1, 2), \quad \boldsymbol{\tau}_1 = (0.7, 1, 2)^\top,$$

$\nu_1 = \nu_2 = 3$ for the T and SL distributions, and $(\nu_1, \gamma_1) = (\nu_2, \gamma_2) = (0.3, 0.3)$ for the CN model. For each sample size $n = 50, 100, 500, 2000$, we set up $\boldsymbol{x}_i = (1, x_{i1}, x_{i2}, x_{i3})^\top$, such that x_{i1}, x_{i2}, and x_{i3} are generated from $\mathcal{U}(1, 5)$, $\mathcal{U}(-2, 2)$, and $\mathcal{U}(1, 4)$, respectively. Moreover by generating r_{i1} and r_{i2} from $\mathcal{U}(-2, 1)$ and $\mathcal{U}(-1, 1)$, the gating covariate is set to $\boldsymbol{r}_i = (1, r_{i1}, r_{i2})^\top$. By imposing three levels of right-censoring $(7.5\%, 15\%, 30\%)$ on the data, the ECME algorithm described in Section 3.2 is preformed to carry out the ML parameter estimates. To investigate parameter recovery, we compute the bias and the mean squared error (MSE):

$$\text{BIAS}(\hat{\theta}_j) = \frac{1}{500} \sum_{l=1}^{500} (\hat{\theta}_j^{(l)} - \theta_{true}) \qquad \text{and} \qquad \text{MSE}(\hat{\theta}_j) = \frac{1}{500} \sum_{l=1}^{500} (\hat{\theta}_j^{(l)} - \theta_{true})^2,$$

where $\hat{\theta}_j^{(l)}$ denotes the estimate of a specific parameter θ_j at the lth replication.

Figures 1 and 2 display the bias and MSE plots of the parameter estimates of the MoE-N-CR, MoE-T-CR, MoE-SL-CR and MoE-CN-CR models for the censoring levels 7.5% and 30%. To shorten the length of the paper, plots of the 15% censoring level are moved to Appendix C. It can be observed that $\hat{\boldsymbol{\beta}}_j$s have very small (around zero) BIAS for all sample sizes. Moreover, as n increases the MSE of $\hat{\boldsymbol{\beta}}_j$s tend to zero. It is also noticeable that the influence of the censoring on the bias and variability of the σ_j^2 and ν_j estimates increases as the censorship rate increases for all models. However, as can be expected, the bias and variability of σ_j^2 and ν_j tend to decrease toward zero by increasing the sample size. The plots in Figure 2 furthermore show the descending trend in the bias and MSE of the gating function parameter estimates as a function of the sample size. These results indicate that the model parameter estimates via the ECME algorithm are empirically consistent.

4.3. Model selection performance via information criteria

One of the challenges in the MoE models is to choose the optimal number of experts G. In dealing with this challenge, we conduct a simulation study to compare the ability of the proposed MoE-SMN-CR sub-models to select the accurate G. We generate 100 samples of size $n = 500$ from a three-component MoE-SMN-CR model (6), where the mixing variable U is followed by a generalized inverse Gaussian (GIG) distribution with parameter $\vartheta = (\kappa, \chi, \psi)$, denoted by the MoE-SGIG-CR model. Details of the GIG distribution and its new data-generating algorithm can be found in Hrmann and Leydold (2013). It is assumed that the data is left-censored with one of the levels 7.5%, 15% or 30%, $\boldsymbol{x}_i = \boldsymbol{r}_i = (1, x_{i1})^\top$ such that x_{i1} is simulated from $\mathcal{U}(-2, 2)$, $\boldsymbol{\beta}_1 = (-4, 4)^\top$, $\boldsymbol{\beta}_2 = (0, -2)^\top$, $\boldsymbol{\beta}_3 = (0, 4)^\top$, $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 0.1$, $\boldsymbol{\tau}_1 = (0, 13)^\top$, $\boldsymbol{\tau}_2 = (2, 9)^\top$, $\vartheta_1 = (-0.5, 1, 2)$, $\vartheta_2 = (0.5, 1, 2)$, and $\vartheta_3 = (-0.5, 2, 1)$. An example of generated samples with and without censored cases is shown in Figure 3.

In this simulation study, it is assumed that the number of mixture components G is unknown. We therefore fit the MoE-N-CR, MoE-T-CR, MoE-SL-CR and MoE-CN-CR models to the generated data with G ranging from 1 to 5 in each replication. The detailed numerical results including the average values of CPU running time (CPU T. in minute to fit an MoE-SMN-CR model for all $G = 1, \ldots, 5$), AIC and BIC together with the rate of true class identification

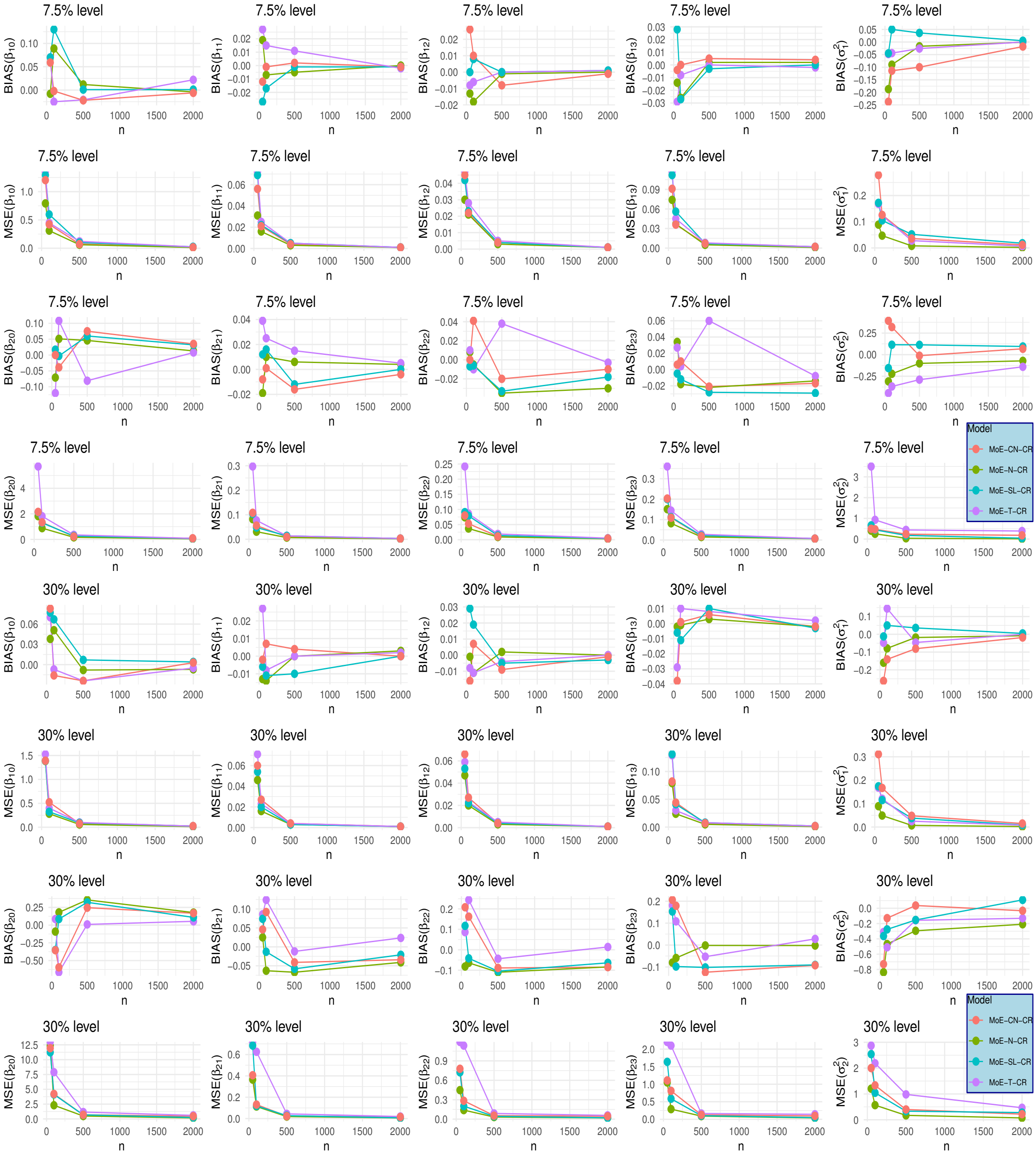

Figure 1: The BIAS and MSE plots of $\boldsymbol{\beta}_j$ and σ_j^2 estimates for the MoE-SMN-CR model (censoring levels 7.5% and 30%).

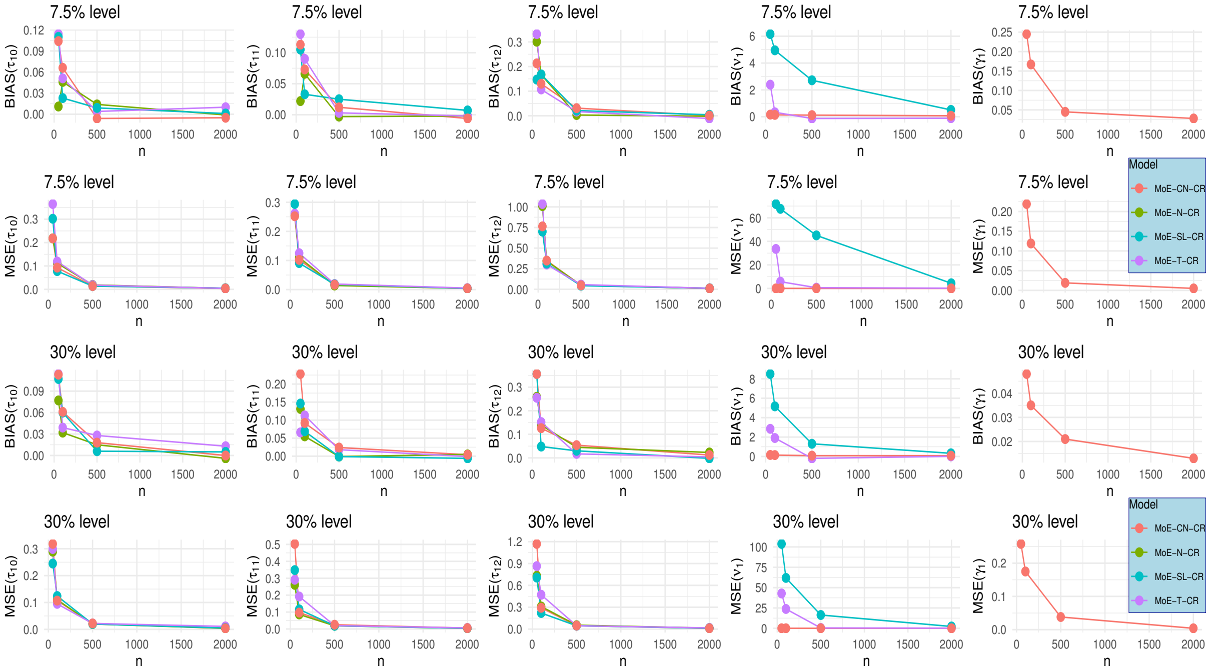

Figure 2: The BIAS and MSE plots of τ_j, ν_j^2 and γ_j estimates for the MoE-SMN-CR model (censoring levels 7.5% and 30%).

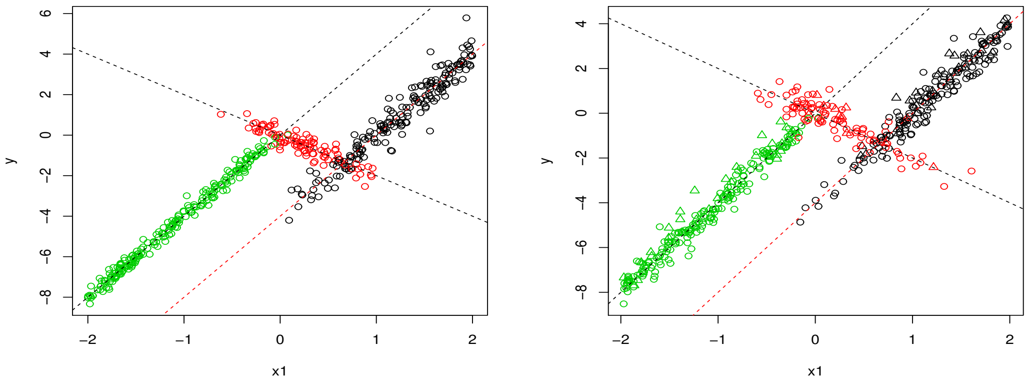

Figure 3: Simulated MoE-SGIG-CR data. Left panel: data without any censored observation. Right panel: data with 15% left-censored observations denoted by △. Dash lines represent the true experts.

(RC; the mean of the number of replications in which the model with $G = 3$ is outperformed) are reported in Table 1. As a rational basis for choosing the most plausible model, Table 1 is also reported the frequencies (in parentheses) supported by the AIC and BIC.

Results depicted in Table 1 suggest that the BIC is more reliable than the AIC for model selection purpose. Based on the RC measure, it can be observed that the MoE-T-CR, MoE-SL-CR and MoE-CN-CR models perform better than the MoE-N-CR model in identifying the number of components since the data are generated from a heavy-tailed distribution. For $G = 3$, the frequencies of plausible model in Table 1 show that the MoE-T-CR and MoE-SL-CR models outperform the other MoE models to fit to the data. In Figure 4, we plot the curve of the estimated experts to a dataset, with 15% censoring level, in which all models suggest $G = 3$ based on the BIC. It could clearly be observed that the MoE-T-CR model fit the data better than the other models.

Table 1: Simulation results, based on 100 replications, for performance comparison of the MoE-SMN-CR sub-models to the generated data from the MoE-SGIG-CR model.

		G = 1		G = 2		G = 3		G = 4		G = 5		RC		
Cens. Level	Model	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	CPU T.
7.5%	MoE-N-CR	1959.914	1972.558	1199.770	1233.487	937.685	**992.475**	918.283	994.146	929.686	1026.622	0.39	0.73	1.277
		(4)	(16)	(3)	(7)	(9)	(16)	(25)	(54)	(30)	(67)			
	MoE-T-CR	1944.642	1961.501	1031.442	1073.588	915.670	**983.104**	882.016	974.737	893.425	1011.434	0.58	0.83	4.016
		(46)	(48)	(84)	(85)	(64)	(60)	(60)	(34)	(50)	(25)			
	MoE-SL-CR	1945.130	1961.988	1048.648	1090.795	895.996	**963.430**	883.740	976.462	903.997	1022.006	0.61	0.84	8.662
		(33)	(34)	(4)	(2)	(20)	(18)	(10)	(9)	(11)	(6)			
	MoE-CN-CR	1946.769	1967.842	1094.680	1145.256	931.721	**1011.798**	910.684	1020.264	913.511	1052.592	0.58	0.77	3.537
		(17)	(2)	(9)	(6)	(7)	(6)	(5)	(3)	(8)	(2)			
15%	MoE-N-CR	1950.180	1962.310	1201.584	1231.051	954.333	**1007.388**	939.155	1011.253	931.957	1027.453	0.35	0.59	1.947
		(8)	(29)	(1)	(4)	(6)	(14)	(22)	(43)	(26)	(54)			
	MoE-T-CR	1936.854	1953.712	1037.712	1079.858	892.430	**959.864**	893.762	986.483	897.911	1015.920	0.59	0.80	5.783
		(46)	(42)	(89)	(89)	(70)	(66)	(61)	(46)	(56)	(38)			
	MoE-SL-CR	1937.009	1953.867	1076.567	1118.713	908.995	**976.428**	905.568	998.289	911.626	1029.635	0.62	0.78	9.208
		(26)	(21)	(6)	(6)	(19)	(17)	(12)	(9)	(10)	(5)			
	MoE-CN-CR	1938.737	1959.810	1136.442	1187.017	962.042	**1042.120**	938.145	1047.725	918.808	1057.890	0.40	0.70	4.267
		(20)	(8)	(4)	(1)	(5)	(3)	(5)	(2)	(8)	(3)			
30%	MoE-N-CR	1947.912	1960.556	1164.255	1197.972	886.857	**941.647**	879.832	955.695	867.216	964.152	0.35	0.78	2.566
		(0)	(6)	(3)	(7)	(8)	(12)	(13)	(34)	(20)	(47)			
	MoE-T-CR	1924.194	1941.052	1018.596	1060.742	823.155	**890.588**	814.159	906.880	810.957	928.966	0.49	0.85	6.124
		(50)	(47)	(87)	(90)	(71)	(69)	(65)	(56)	(61)	(44)			
	MoE-SL-CR	1924.250	1941.109	1044.162	1086.308	838.302	**905.735**	833.820	926.542	840.034	958.043	0.54	0.90	10.016
		(39)	(40)	(2)	(2)	(19)	(17)	(19)	(10)	(12)	(7)			
	MoE-CN-CR	1927.552	1948.625	1089.598	1140.173	876.187	**956.264**	871.465	981.045	866.797	1005.879	0.55	0.85	5.890
		(11)	(7)	(8)	(1)	(2)	(2)	(3)	(0)	(7)	(2)			

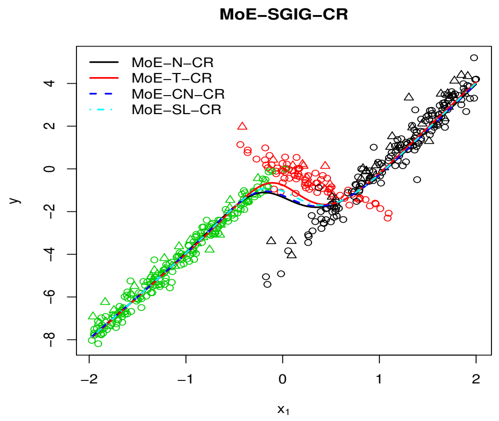

Figure 4: The estimated Experts curves of the special cases of the MoE-SMN-CR model for the MoE-SGIG-CR simulated data with 15% left-censoring.

4.4. Performance in dealing with the highly peaked and thick-tailed data

In this simulation study, we simulate data with $n = 100, 500$ and 2000 observations from a three-component MoE-SMN-CR model via representation (4) under two generating scenarios of U. The first scenario (S1) is conducted by assuming $U^{-1} \sim \mathcal{E}(0.5)$, the exponential distribution with parameter $\lambda = 0.5$, whereas the second one (S2) considers $U \sim \mathcal{BS}(\alpha, 1)$, the Birnbaum-Saunders distribution (Birnbaum and Saunders, 1969) with parameter α and $\beta = 1$. Bear in mind that the former scenario generates data from Laplace distribution which is known as a highly peaked model and the latter scenario provides a heavier tail model than the normal distribution (Naderi et al., 2017, 2019). The Laplace and BS censored MoE models, referred as the MoE-SLap-CR and MoE-SBS-CR, are not considered in this paper since their conditional expectations involved in the ECME algorithm do not exist.

In each replication of 200 trials, the interval-censored data, with level 7.5%, 15% or 30%, are generated from the MoE-SLap-CR and MoE-SBS-CR models with $G = 3$, and the presumed parameter values $\boldsymbol{\theta}_j = (\boldsymbol{\beta}_j, \sigma_j^2, \boldsymbol{\nu}_j)$, $j = 1, 2, 3$, where $\boldsymbol{\beta}_1 = (-2, -1, -2, -3)^\top$, $\boldsymbol{\beta}_2 = (0.5, 1, 2, 3)^\top$, $\boldsymbol{\beta}_3 = (2, 1, 3, 5)^\top$, $(\sigma_1^2, \sigma_2^2, \sigma_3^2) = (1, 3, 5)$, $\boldsymbol{\tau}_1 = (2, 10)^\top$, $\boldsymbol{\tau}_2 = (0.7, 10)^\top$ and $(\alpha_1, \alpha_2, \alpha_3) = (3, 1, 2)$ for the MoE-SBS-CR model. For this purpose, we also set up $\boldsymbol{x}_i =$

Table 2: The average of AIC, BIC, MCR and AIR, over 200 replications, by fitting special cases of the MoE-SMN-CR model to the generated data under S1 scenario.

Model →		MoE-N-CR			MoE-T-CR			MoE-SL-CR			MoE-CN-CR		
n ↓	Measure	7.5%	15%	30%	7.5%	15%	30%	7.5%	15%	30%	7.5%	15%	30%
	AIC	496.030	505.090	514.104	485.298	496.103	497.694	494.472	503.898	506.189	502.770	511.464	514.687
	BIC	545.528	554.588	563.602	542.612	553.417	555.008	551.786	561.212	563.503	567.901	576.594	579.817
100	MCR	0.165	0.192	0.222	0.175	0.185	0.225	0.165	0.178	0.216	0.158	0.175	0.213
	ARI	0.618	0.576	0.499	0.598	0.584	0.497	0.617	0.594	0.506	0.630	0.604	0.517
	JCI	0.632	0.602	0.551	0.619	0.608	0.548	0.630	0.616	0.553	0.641	0.621	0.560
	CPU T.	0.021	0.019	0.007	0.420	0.212	0.114	2.776	2.724	2.351	0.184	0.070	0.026
	AIC	2370.929	2425.217	2618.009	2330.694	2369.694	2531.162	2338.787	2383.382	2554.371	2335.274	2378.907	2545.153
	BIC	2451.006	2505.295	2698.086	2423.415	2462.416	2623.884	2431.509	2476.104	2647.092	2440.640	2484.272	2650.518
500	MCR	0.162	0.167	0.214	0.153	0.163	0.201	0.154	0.158	0.186	0.155	0.157	0.187
	ARI	0.614	0.616	0.572	0.628	0.617	0.577	0.627	0.624	0.598	0.626	0.629	0.596
	JCI	0.630	0.634	0.597	0.642	0.635	0.601	0.642	0.641	0.616	0.639	0.642	0.612
	CPU T.	0.105	0.077	0.096	0.570	0.447	0.496	5.013	5.887	12.063	0.248	0.070	0.058
	AIC	9397.490	9559.804	10548.390	9220.558	9278.635	10085.960	9255.169	9325.645	10221.340	9235.637	9310.615	10157.243
	BIC	9503.907	9666.221	10654.800	9343.778	9401.854	10209.180	9378.389	9448.865	10344.560	9370.060	9445.036	10291.67
2000	MCR	0.154	0.167	0.240	0.147	0.159	0.222	0.147	0.159	0.213	0.146	0.154	0.202
	ARI	0.634	0.614	0.521	0.644	0.621	0.530	0.645	0.623	0.541	0.647	0.623	.556
	JCI	0.645	.0632	0.559	0.654	.0637	0.564	0.656	0.641	0.571	0.656	0.644	0.580
	CPU T.	1.436	1.189	1.279	3.477	2.988	2.855	19.478	18.354	17.189	1.873	1.434	1.363

Table 3: The average of AIC, BIC, MCR and AIR, over 200 replications, by fitting special cases of the MoE-SMN-CR model to the generated data under S2 scenario.

Model →		MoE-N-CR			MoE-T-CR			MoE-SL-CR			MoE-CN-CR		
n ↓	Measure	7.5%	15%	30%	7.5%	15%	30%	7.5%	15%	30%	7.5%	15%	30%
	AIC	522.443	542.212	556.691	498.602	514.384	533.806	508.967	525.134	544.353	502.185	521.979	538.774
	BIC	571.941	591.710	606.189	555.916	571.698	591.121	566.280	582.448	601.666	567.315	587.108	603.903
100	MCR	0.235	0.248	0.263	0.183	0.190	0.209	0.188	0.193	0.211	0.190	0.197	0.214
	ARI	0.498	0.472	0.448	0.589	0.571	0.541	0.584	0.563	0.539	0.581	0.557	0.535
	JCI	0.518	0.501	0.493	0.627	0.599	0.572	0.618	0.586	0.577	0.605	0.572	0.561
	CPU T.	0.032	0.015	0.011	0.505	0.278	0.103	4.544	4.313	2.563	0.168	0.096	0.023
	AIC	2561.848	2564.510	2665.281	2432.404	2432.566	2503.707	2455.665	2464.471	2551.097	2490.907	2502.043	2799.294
	BIC	2641.925	2644.587	2745.358	2525.125	2525.288	2596.428	2548.386	2557.192	2643.818	2596.273	2607.408	2904.659
500	MCR	0.197	0.207	0.217	0.162	0.172	0.187	0.172	0.179	0.190	0.176	0.186	0.201
	ARI	0.588	0.553	0.533	0.636	0.602	0.574	0.624	0.585	0.563	0.618	0.574	0.545
	JCI	0.604	0.583	0.577	0.643	0.622	0.601	0.633	0.612	0.587	0.627	0.596	0.569
	CPU T.	0.172	0.110	0.194	0.850	0.542	0.967	12.179	9.570	14.585	0.284	0.145	0.386
	AIC	10056.300	10424.278	10954.970	9576.709	9829.808	10326.340	9646.046	9956.371	10514.060	9654.003	9976.371	10527.060
	BIC	10162.717	10530.695	11061.380	9699.929	9953.028	10449.560	9769.266	10079.591	10637.280	9794.025	10110.791	10661.420
2000	MCR	0.214	0.223	0.255	0.171	0.174	0.216	0.189	0.178	0.218	0.171	0.181	0.219
	ARI	0.556	0.535	0.513	0.613	0.601	0.569	0.592	0.576	0.533	0.601	0.583	0.550
	JCI	0.578	0.560	0.553	0.644	0.619	0.598	0.607	0.596	0.563	0.626	0.603	0.569
	CPU T.	3.834	2.555	1.446	6.327	4.868	2.219	16.068	14.826	13.642	5.576	4.521	2.079

$(1, x_{i1}, x_{i2}, x_{i3})^\top$, such that x_{i1}, x_{i2}, and x_{i3} are generated from $\mathcal{U}(1,5)$, $\mathcal{U}(0,1)$, and $\mathcal{U}(-2,-1)$, respectively, and $\boldsymbol{r}_i = (1, r_{i1})^\top$ where r_{i1} is generated from $\mathcal{U}(-1,1)$.

We compare the performance of the three-component MoE-N-CR, MoE-T-CR, MoE-SL-CR, and MoE-CN-CR models in terms of model selection indices (AIC and BIC) as well as clustering agreement measures (MCR, JCI, and ARI). Tables 2 and 3 present the average values of AIC, BIC, MCR, JCI, ARI, and CPU running time (in minute), over all 200 replications for the S1 and S2 scenarios of simulation, respectively. Results depicted in these tables reveal that the MoE-T-CR model outperforms the others in terms of AIC and BIC. Although the clustering performance of all models are very closed to each others, as expected from the MoE structure, the MoE-T-CR and MoE-CN-CR models provide a slight improvement in the MCR, JCI and AIR over the MoE-N-CR and MoE-SL-CR models.

4.5. Sensitivity analysis in presence of outliers

This simulation study aims at investigating the robustness of estimating MoE-SMN-CR sub-models in which some outliers are introduced into the simulated data. Each of the three models MoE-SLap-CR, MoE-SBS-CR and MoE-SGIG-CR is considered for data generation. Following Nguyen and McLachlan (2016), we set $\boldsymbol{x}_i = \boldsymbol{r}_i = (1, x_{i1})^\top$ where x_{i1} is generated from $\mathcal{U}(-1,1)$, $\boldsymbol{\beta}_1 = (0,1)^\top$, $\boldsymbol{\beta}_2 = (0,-1)^\top$, $\sigma_1^2 = \sigma_2^2 = 0.01$, $\boldsymbol{\tau}_1 = (0,10)^\top$, $\vartheta_1 = (-0.5, 1, 0.2)$, $\vartheta_2 = (0.5, 1, 0.2)$ for the MoE-SGIG-CR model, and $(\alpha_1, \alpha_2) = (0.5, 1)$ for the MoE-SBS-CR model. We assume

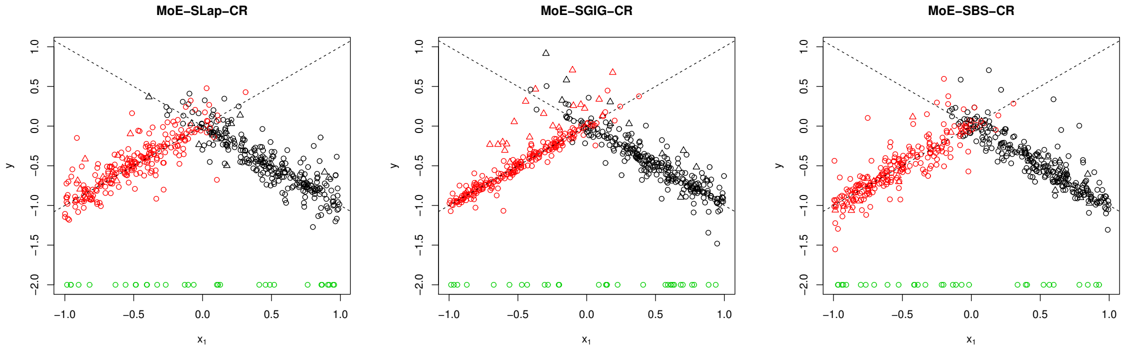

Figure 5: Scatterplots of the simulated data with 7.5% left-censoring (△) generated from the MoE-SLap-CR, MoE-SBS-CR and MoE-SGIG-CR models and containing 6% outliers (green ∘). Dash lines represent the true experts.

Table 4: Simulation results for assessing the robustness of the proposed MoE model to outliers under various censoring levels and outliers percentages.

	Cens. Level →	7.5%				30%			
True model	Fitted model	0%	2%	4%	6%	0%	2%	4%	6%
MoE-SGIG-CR	MoE-N-CR	0.0347	0.0948	0.1418	0.1897	0.1029	0.1626	0.2357	0.2776
	MoE-T-CR	0.0297	0.0855	0.1195	0.1623	0.0698	0.1375	0.1857	0.2068
	MoE-SL-CR	0.0300	0.0856	0.1201	0.1642	0.0733	0.1392	0.1891	0.2161
	MoE-CN-CR	0.0334	0.0865	0.1232	0.1692	0.0928	0.1379	0.2013	0.2481
MoE-SBS-CR	MoE-N-CR	0.0385	0.0979	0.1436	0.1936	0.1061	0.1657	0.2338	0.2788
	MoE-T-CR	0.0342	0.0885	0.1226	0.1654	0.0735	0.1416	0.1918	0.2292
	MoE-SL-CR	0.0344	0.0889	0.1230	0.1662	0.0789	0.1437	0.1932	0.2284
	MoE-CN-CR	0.0375	0.0897	0.1260	0.1729	0.0981	0.1427	0.2076	0.2507
MoE-SLap-CR	MoE-N-CR	0.0451	0.1050	0.1512	0.1978	0.1142	0.1772	0.2401	0.2892
	MoE-T-CR	0.0406	0.0950	0.1287	0.1712	0.0827	0.1519	0.1979	0.2289
	MoE-SL-CR	0.0407	0.0953	0.1290	0.1719	0.0886	0.1548	0.2010	0.2394
	MoE-CN-CR	0.0437	0.0963	0.1329	0.1779	0.1032	0.1539	0.2165	0.2562

left-censoring scheme with levels 7.5% and 30%, and sample size 500. An example of simulated samples with left-censoring level 7.5% from the MoE-SLap-CR, MoE-SBS-CR and MoE-SGIG-CR models is shown in Figure 5. Plots in Figure 5 show that the generated responses are usually greater that -1.6. Apart from the main generated (censored and uncensored) samples, we also add class of outliers with varying probability c ranging from 0% to 6%. To do so, we set $r = x$ where the explanatory variable x is simulated from $\mathcal{U}(-1, 1)$. Moreover, the corresponding response y for all generated x is set to the value -2 (Nguyen and McLachlan, 2016). The green circles in Figures 5 and 6 show the 6% outliers added to the main generated samples. In each trial of 500 replications, the MoE-N-CR, MoE-T-CR, MoE-CN-CR, and MoE-SL-CR models are fitted to the generated data. Figure 6 shows an example fitted MoE curves to the data generated from the MoE-SLap-CR, MoE-SBS-CR and MoE-SGIG-CR models. It can obviously be seen that the heavy-tailed models provide better platforms for describing the data than the MoE-N-CR model.

To assess the impact of the outliers on the parameter estimates and on the quality of the results, in each 500 replication, the mean square error between the true regression mean function and the estimated one is calculated as

$$\text{MSE} = \frac{1}{500}\sum_{i=1}^{500}\Big(E_{\hat{\boldsymbol{\Theta}}}(\boldsymbol{x}_i, \boldsymbol{r}_i) - E_{\boldsymbol{\Theta}_{true}}(\boldsymbol{x}_i, \boldsymbol{r}_i)\Big)^2,$$

where $E_{\boldsymbol{\Theta}}(\boldsymbol{x}_i, \boldsymbol{r}_i) = \sum_{j=1}^{G} \pi_j(\boldsymbol{r}_i; \boldsymbol{\tau})\boldsymbol{x}_i^\top \boldsymbol{\beta}_j$ evaluated at the true and estimated parameters. Table 4 shows, for each of the four MoE models, the average of MSE for various percentage of outliers and censoring levels in the data. First, one can see that the MSE tends towered zero as the level of censoring and percentage of outliers approach zeros for all cases of the MoE-SMN-CR model. Since the three considered scenarios generate fat-tailed data, it can be observed that without outliers (c = 0%) the error of the MoE-N-CR model is greater than those of the other MoE models, reflecting its lack of robustness. Upon inspection of Table 4, one can conclude that by adding outliers to the data, the

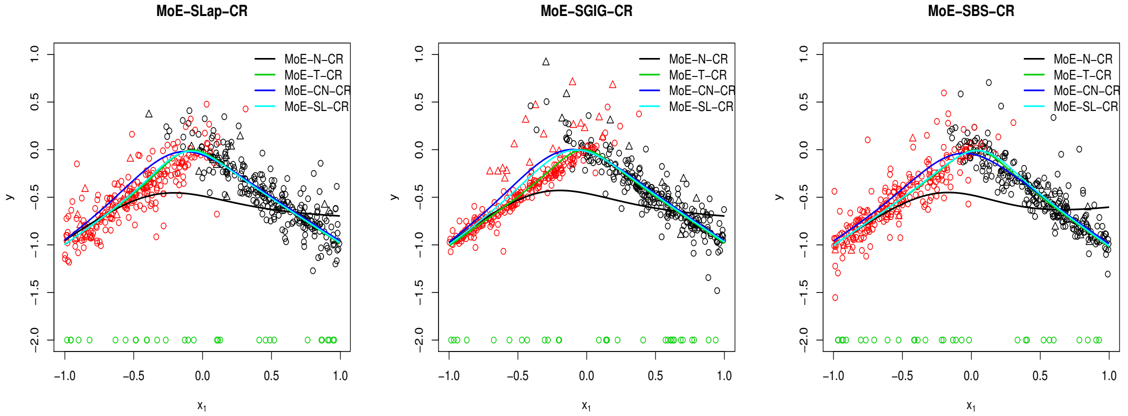

Figure 6: Scatter plots of the artificial data with 7.5% left-censoring (△) generated from the MoE-SLap-CR, MoE-SBS-CR and MoE-SGIG-CR models and containing 6% outliers (green ∘).

MoE-T-CR (and the MoE-SL-CR in the second order) model clearly outperforms others for all situations. It highlights that the MoE-T-CR model is much more robust to outliers under these data generating scenarios.

4.6. *Classification evaluation*

As recommended by the Associate Editor and reviewers, the last simulation investigates the benefits of using the gating function in the proposed MoE-SMN-CR model for the classification purposes. To do so, we compare the clustering performance of our proposed model with the mixture of censored linear regression models based on the SMN class of distributions (MRM-SMN-CR), proposed by (Zeller et al., 2019). Hereafter, we will denote the mixture of censored linear regression models based on the normal, Student-t, slash and contaminated-normal distributions, respectively by MRM-N-CR, MRM-T-CR, MRM-SL-CR and MRM-CN-CR. Following Zeller et al. (2016) and Yang et al. (2020), we generate interval-censored samples of size 500 from a MoE-T-CR model with level 7.5%, 15% or 30% and parameter values

$$\boldsymbol{\beta}_1 = (0,-1,-2,-3)^\top,\ \ \boldsymbol{\beta}_2 = (-1,1,2,3)^\top,\ \ \boldsymbol{\beta}_3 = (0,-2,1,3)^\top,\ \ \boldsymbol{\tau}_1 = (0.7,1,6)^\top,\ \ \boldsymbol{\tau}_2 = (1,0.9,10)^\top$$

$(\sigma_1^2, \sigma_2^2, \sigma_3^2) = (1,2,4)$, and $(\nu_1, \nu_2, \nu_3) = (2,3,5)$. For $\boldsymbol{x}_i = (1, x_{i1}, x_{i2}, x_{i3})^\top$ and $\boldsymbol{r}_i = (1, r_{i1}, r_{i2})^\top$, we generate x_{i1}, x_{i2} and x_{i3} from $\mathcal{U}(1,5)$, $\mathcal{U}(-2,2)$, and $\mathcal{U}(1,4)$, respectively, as well as r_{i1} form $\mathcal{U}(-2,1)$ and r_{i2} from $\mathcal{U}(-1,1)$.

For the sake of clustering comparison, we compute the right number of allocations and MCR of the three-component MoE-SMN-CR and MRM-SMN-CR sub-models for each sample. Table 5 depicts the mean of right allocations (MRA) with its standard deviation (SDRA), the avrage of MCR, and CPU time over 100 replications. It can be observed that the heavy-tailed MoE models have greater MRA and smaller MCR and SDRA which confirms that the MoE-T-CR, MoE-SL-CR and MoE-SL-CR models provide improvement in the right clustering. Moreover, Table 5 reports the percentages that the true MoE-T-CR model is chosen in terms of right allocation in comparison with the other fitted models. As can be expected, the MRA significantly favor true model against the MRM-SMN-CR.

5. Real data analysis

This section considers the wage rates dataset, previously analyzed by Mroz (1987); Caudill (2012) and Karlsson and Laitila (2014), for illustrative purposes of the developed novel MoE-SMN-CR model. This dataset contains 753 observed wage rates (hours of working outside the home) of married white women between the ages of 30 and 60 in 1975, of whom 325 have zero hours working. It means that 43.16% wives did not work in 1975 and can therefore be treated as the left-censored subjects at zero. Recently, Zeller et al. (2019) reanalyzed the wage-rates dataset in order to illustrate the performance of the MRM-SMN-CR. By considering the wife's annual work hours outside home scaled by 1000 as the response variable (y), and the explanatory variables including (x_1) the wife's education in

Table 5: Simulation results, based on 100 replications, for assessing the advantages of using gating function in clustering data when they are generated from the MoE-T-CR model. Percentages that the true model is chosen vs other models are presented in parentheses.

	MRA			SDRA			MCR			CPU T.		
Fitted model	7.5%	15%	30%	7.5%	15%	30%	7.5%	15%	30%	7.5%	15%	30%
MRM-N-CR	410.520 (100)	379.171 (100)	323.700 (100)	19.896	27.371	49.456	0.177	0.242	0.339	0.084	0.060	0.087
MRM-T-CR	417.560 (100)	388.060 (100)	345.000 (100)	13.021	17.756	22.813	0.165	0.227	0.314	0.285	0.184	0.198
MRM-SL-CR	412.770 (100)	383.090 (100)	328.567 (100)	16.028	18.066	24.582	0.174	0.234	0.331	6.915	7.214	11.595
MRM-CN-CR	411.960 (100)	382.350 (100)	329.900 (100)	14.713	21.087	18.746	0.169	0.236	0.333	0.309	0.325	0.362
MoE-N-CR	461.290 (86)	439.790 (89)	407.333 (89)	17.290	25.661	38.462	0.077	0.120	0.182	0.799	0.412	0.178
MoE-T-CR	476.350 (–)	463.890 (–)	432.800 (–)	10.197	15.733	20.671	0.050	0.078	0.134	1.191	0.916	0.651
MoE-SL-CR	469.180 (80)	459.660 (82)	426.767 (78)	10.984	16.404	29.619	0.062	0.085	0.146	10.163	7.269	15.145
MoE-CN-CR	465.637 (83)	447.460 (85)	420.600 (84)	11.043	15.575	29.557	0.069	0.102	0.153	1.269	1.093	0.893

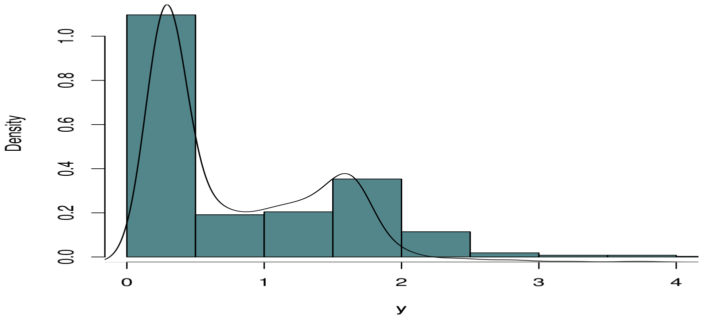

Figure 7: The histogram of the response variable y overlaid with its Kernel density estimate.

years, (x_2) the wife's age, (x_3) the wife's previous labor market experience and (x_4) the wife's previous labor market experience squared, Caudill (2012); Karlsson and Laitila (2014) and Zeller et al. (2019) concluded that a mixture of two-component linear regression censored model provides an appropriate platform for analyzing this dataset. Figure 7 shows the histograms of y overlaid with the estimated kernel density curve. The bimodality of the data and the suitability of the two-component mixture model to fit the data can be observed. It could be mentioned from the histogram that the data are heavily right-tailed distributed.

The mixture modeling allows clustering of the data in terms of the estimated posterior classification probability, $\hat{z}_{ij}$, that a single point belongs to a given group. Although the previous works on the wage-rates dataset focused on the aforementioned explanatory variables and showed that only these variables have significant effects on y, there are eleven measures that could provide more information in investigating the complex relationship of random phenomena under study. One of those variables that we will use for clustering purposes is the living status, labeled as "city", that takes 1 for living in the city and 0 otherwise. Assuming "city" as the group indicator, one can obtain $\hat{z}_{ij}$, and can therefore compute the clustering criteria MCR, RI, ARI and JCI of the MRMs proposed by Zeller et al. (2019). In this regard, the posterior probabilities of the two-component MRM-N-CR, MRM-T-CR, MRM-SL-CR and MRM-CN-CR are computed by fitting them to the considered data. It is observed that all of the models proposed by Zeller et al. (2019) assign data points to one group.

As the advantages of the MoE model, it is possible for the investigator to choose some covariates for the gating function. In analyzing wage-rates data, we consider $\boldsymbol{x} = (1, x_1, x_2, x_3, x_4)^\top$ and $\boldsymbol{r} = (1, r_1, r_2, x_2)^\top$ for gating function, where (r_1) is the unemployment rate in county of residence and (r_2) is the number of kids less than 6 years old in the household. We note that the covariates of the gating function can be the same as $\boldsymbol{x}$, however by considering various combinations of the available explanatory variables, we observe that these three variables provide a better clustering performance. An interesting open issue for future work could be the variable selection problem for both $\boldsymbol{x}$ and $\boldsymbol{r}$ in the MoE models.

By fitting the MoE-N-CR, MoE-T-CR, MoE-SL-CR, and MoE-CN-CR models to this dataset for $G = 1, \ldots, 4$, the two-component MoE model has been selected based on the BIC. It should be noted that our results are not directly comparable with those obtained by Karlsson and Laitila (2014) since they imposed some restrictions on $\boldsymbol{\beta}$ for

Table 6: ML estimates with corresponding approximate standard errors(SE) together with their AIC, BIC, and clustering performance measures.

	MoE-N-CR		MoE-T-CR		MoE-SL-CR		MoE-CN-CR	
Parameter ↓	Estimates	SE	Estimates	SE	Estimates	SE	Estimates	SE
β_{10}	5.5476	0.6362	5.5438	0.6573	5.6223	0.7524	5.4714	0.9077
β_{11}	-0.0554	0.0268	-0.0627	0.0027	-0.0658	0.0287	-0.0607	0.0227
β_{12}	-0.1272	0.0130	-0.1256	0.0014	-0.1227	0.0167	-0.1212	0.0212
β_{13}	0.0653	0.0355	0.0822	0.0050	0.0371	0.0063	0.0485	0.0114
β_{14}	0.0004	0.0002	-0.0003	0.0001	0.0013	0.0029	0.0009	0.0007
β_{20}	1.5064	0.2850	0.7306	0.0675	1.3579	0.4638	1.3405	0.2478
β_{21}	0.0165	0.0025	0.0109	0.0025	0.0259	0.0051	0.0212	0.0028
β_{22}	-0.0592	0.0125	-0.0410	0.0013	-0.0578	0.0109	-0.0560	0.0098
β_{23}	0.2418	0.0205	0.2424	0.0018	0.2426	0.0207	0.2404	0.0128
β_{24}	-0.0047	0.0006	-0.0049	0.0001	-0.0048	0.0007	-0.0047	0.0021
σ_1^2	0.5001	0.0682	0.4365	0.0066	0.3773	0.0836	0.4173	0.1109
σ_2^2	0.7130	0.0568	0.4661	0.0043	0.3120	0.0367	0.4214	0.1291
ν_1	–	–	9.3049	–	9.0866	–	0.0342	–
ν_2	–	–	6.2745	–	1.8225	–	0.1577	–
γ_1	–	–	–	–	–	–	0.2643	–
γ_2	–	–	–	–	–	–	0.2237	–
τ_0	26.7338	5.6234	48.3470	6.3394	14.4513	3.9352	17.4136	4.6459
τ_1	0.2519	0.1023	0.3414	0.2210	0.1845	0.0138	0.1999	0.0851
τ_2	4.1959	1.2368	6.9057	1.7441	0.8211	0.1362	0.7284	0.1246
τ_3	-0.7383	0.2304	-1.2943	0.3588	-0.4177	0.1394	-0.4912	0.1537
AIC	1234.5830		1219.2230		1219.2830		1224.094	
BIC	1308.5680		1302.4570		1302.5160		1316.575	
RI	0.5123		0.5214		0.5323		0.5118	
JCI	0.3676		0.3847		0.4029		0.3713	

estimation. Moreover, it is clear that adding more variables to the model will definitely affect on the likelihood. We therefore can not compare the results of model selection criteria with those reported by Zeller et al. (2019).

Table 6 shows the ML results obtained by fitting the four considered models. The estimates of β_{11} in all MoE-SMN-CR sub-models imply a positive influence of education on the respond variable for the first group which is contrariwise for the second group. Looking at the coefficient estimates of experience, it can be seen that the wives' annual work hours rise as their experience enhanced. However, all models suggest the descending trend, in both group, for the mean of the work hours as a function of age. The estimate of ν_1 in Table 6 is moderately large for the MoE-T-CR, MoE-SL-CR models and quite small for the MoE-CN-CR model. It might support the fact that the best distribution to fit the data is a mixture of normal and a heavy-tailed distributions. The results in Table 6 also reveal that the estimated gating parameters are moderately significant, showing that the considered covariates $\boldsymbol{r}$ have an effect on the analysis. Comparing the estimates of τ_i's for four proposed models, the number of kids less than 6 years old has the highest impact on gating function. Results based on AIC and BIC finally indicate that the MoE-T-CR and MoE-SL-CR models provides an improved fit of the data over the other models. Moreover, by comparing the clustering criteria in Table 6, it turns out that the MoE-SL-CR model yields quite better classification.

6. Conclusions and discussions

This paper proposed a new robust mixture of linear experts model for the censored data based on the scale-mixture of normal class of distributions. This MoE-SMN-CR model extended the classical MoE model which has been demonstrated to solve the two challenges to deal with heavy-tail distributed data and outliers as well as censored data. The newly proposed MoE-SMN-CR model is very extensive which extends the classical MoE model and includes MRM and finite mixture regression model for censored data proposed by Zeller et al. (2019) as special cases. The use of covariates in the gating function is an advantage of the MoE models which might result in better classification of the data. Utilizing the embedded hierarchical structure of the SMN class of distributions, we developed an innovative EM-type algorithm to obtain ML parameter estimates computationally, which is implemented in statistical software R.

Five Monte-Carlo simulation studies were conducted to investigate the performance of the model in applications both for non-linear regression and prediction and for model-based clustering. Results of the simulation studies confirmed that the proposed MoE-SMN-CR model can provide evidence of the robustness to the outliers and atypical observations. Finally, a real-world data analysis demonstrated the applicability and benefit of the proposed approach for practical applications.

As discussed in Section 5, an interesting future direction of the current work is the variable selection for both parts of the regression and gating function. The utility of our current approach can be further extended to the multiple regression on multivariate data rather than simple regression on univariate data, which we are actively exploring. To do so, we refer the reader to the work of Lachos et al. (2017) who proposed an exact ECM algorithm for the mixture of censored multivariate Student-*t* distributions. Another possible extension of the work herein is to consider a full Bayesian approach as a basis of inference and prediction (Peng et al., 1996; Zens, 2019). Recommended by the Associate Editor and the reviewers, one can introduce an MoE model for censored data based on the results of Mattos et al. (2018); Lachos et al. (2020) for handling skew and heavy-tails distributed data, as well as based on the results of Lin et al. (2018); Lin and Wang (2019) for analyzing censored and missing data observations simultaneously.

Acknowledgments

The authors wish to thank the Editor-in-Chief, anonymous Associate Editor, and two referees for their helpful comments and insightful suggestions, helping to improve the first draft of the paper. This work is based upon research supported by the South Africa National Research Foundation and South Africa Medical Research Council (South Africa DST-NRF-SAMRC SARChI Research Chair in Biostatistics, Grant number 114613), as well as by the National Research Foundation of South Africa (Grant Numbers 127727). Opinions expressed and conclusions arrived at are those of the author and are not necessarily to be attributed to the NRF or SAMRC.

Appendix A. Conditional expectations of the special cases of the SMN distributions

Uncensored observations: For the uncensored data y_i, we have $\rho_i = 0$. Therefore, the only necessary conditional expectation $\hat{u}_{ij}^{(k)} = E(U_{ij}|Y = y_i, \hat{\boldsymbol{\theta}}_j^{(k)})$ for the considered models can be computed as follows.

- If $Y \sim \mathcal{N}(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_j^{(k)}, \hat{\sigma}_j^{2(k)})$, in this case, $U = 1$ with probability one, and so $\hat{u}_{ij}^{(k)} = 1$.
- If $Y \sim \mathcal{T}(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_j^{(k)}, \hat{\sigma}_j^{2(k)}, \hat{\nu}_j^{(k)})$, We have
$$\hat{u}_{ij}^{(k)} = \frac{\hat{\nu}_j^{(k)} + 1}{\hat{\nu}_j^{(k)} + \delta\Big(y_i, \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_j^{(k)}, \hat{\sigma}_j^{(k)}\Big)},$$
where $\delta(y, \mu, \sigma) = \Big((y - \mu)/\sigma\Big)^2$.
- If $Y \sim \mathcal{SL}(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_j^{(k)}, \hat{\sigma}_j^{2(k)}, \hat{\nu}_j^{(k)})$, We have
$$\hat{u}_{ij}^{(k)} = 2\left(\delta\Big(y_i, \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_j^{(k)}, \hat{\sigma}_j^{(k)}\Big)\right)^{-1} \frac{\Gamma\left(\hat{\nu}_j^{(k)} + 1.5,\ 0.5\delta\Big(y_i, x_i^\top \hat{\boldsymbol{\beta}}_j^{(k)}, \hat{\sigma}_j^{(k)}\Big)\right)}{\Gamma\left(\hat{\nu}_j^{(k)} + 0.5,\ 0.5\delta\Big(y_i, \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_j^{(k)}, \hat{\sigma}_j^{(k)}\Big)\right)}.$$
- If $Y \sim \mathcal{CN}(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_j^{(k)}, \hat{\sigma}_j^{2(k)}, \hat{\nu}_j^{(k)}, \hat{\gamma}_j^{(k)})$, We have
$$\hat{u}_{ij}^{(k)} = \frac{1 - \hat{\nu}_j^{(k)} + \hat{\nu}_j^{(k)}(\hat{\gamma}_j^{(k)})^{1.5} \exp\left\{0.5(1 - \hat{\gamma}_j^{(k)})\delta\Big(y_i, \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_j^{(k)}, \hat{\sigma}_j^{(k)}\Big)\right\}}{1 - \hat{\nu}_j^{(k)} + \hat{\nu}_j^{(k)}(\hat{\gamma}_j^{(k)})^{0.5} \exp\left\{0.5(1 - \hat{\gamma}_j^{(k)})\delta\Big(y_i, \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_j^{(k)}, \hat{\sigma}_j^{(k)}\Big)\right\}}.$$

Censored cases: In the censored cases, we have $\rho_i = 1$. For the sake of notation, let
$$T_{ij}^{(k)} = \frac{Y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_j^{(k)}}{\hat{\sigma}_j^{(k)}} \sim \text{SMN}(0, 1, \hat{\boldsymbol{\nu}}_j^{(k)}), \qquad \hat{t}_{ij1}^{(k)} = \frac{c_{i1} - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_j^{(k)}}{\hat{\sigma}_j^{(k)}}, \qquad \hat{t}_{ij2}^{(k)} = \frac{c_{i2} - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_j^{(k)}}{\hat{\sigma}_j^{(k)}}.$$

Therefore, the necessary conditional expectations $\hat{u}_{ij}^{(k)} = E(U_{ij}|c_{i1} \leq Y_i \leq c_{i2}, \hat{\boldsymbol{\theta}}_j^{(k)})$, $\widehat{uy}_{ij}^{(k)} = E(U_iY_i|c_{i1} \leq Y_i \leq c_{i2}, \hat{\boldsymbol{\theta}}_j^{(k)})$, and $\widehat{uy^2}_{ij}^{(k)} = E(U_iY_i^2|c_{i1} \leq Y_i \leq c_{i2}, \hat{\boldsymbol{\theta}}_j^{(k)})$ for the considered models can be computed as follows.

$$\hat{u}_{ij}^{(k)} = E\left(U_{ij}|\hat{t}_{ij1}^{(k)} \leq T_{ij}^{(k)} \leq \hat{t}_{ij2}^{(k)}, \hat{\boldsymbol{\theta}}_j^{(k)}\right) = \frac{E_\Phi\left(1, \hat{t}_{ij2}^{(k)}\right) - E_\Phi\left(1, \hat{t}_{ij1}^{(k)}\right)}{F_{SMN}\left(\hat{t}_{ij2}^{(k)}; \hat{\boldsymbol{\nu}}_j^{(k)}\right) - F_{SMN}\left(\hat{t}_{ij1}^{(k)}; \hat{\boldsymbol{\nu}}_j^{(k)}\right)},$$

$$\begin{aligned}
\widehat{uy}_{ij}^{(k)} &= \left(x_i^\top \hat{\boldsymbol{\beta}}_j^{(k)}\right)\widehat{u}_{ij}^{(k)} + \hat{\sigma}_j^{(k)} E\left(U_{ij}T_{ij}\middle|\hat{t}_{ij1}^{(k)} \leq T_{ij}^{(k)} \leq \hat{t}_{ij2}^{(k)}, \hat{\boldsymbol{\theta}}_j^{(k)}\right)\\
&= \left(x_i^\top \hat{\boldsymbol{\beta}}_j^{(k)}\right)\left\{\frac{E_\Phi\left(1, \hat{t}_{ij2}^{(k)}\right) - E_\Phi\left(1, \hat{t}_{ij1}^{(k)}\right)}{F_{SMN}\left(\hat{t}_{ij2}^{(k)}; \hat{\boldsymbol{\nu}}_j^{(k)}\right) - F_{SMN}\left(\hat{t}_{ij1}^{(k)}; \hat{\boldsymbol{\nu}}_j^{(k)}\right)}\right\} + \hat{\sigma}_j^{(k)}\left\{\frac{E_\phi\left(0.5, \hat{t}_{ij1}^{(k)}\right) - E_\phi\left(0.5, \hat{t}_{ij2}^{(k)}\right)}{F_{SMN}\left(\hat{t}_{ij2}^{(k)}; \hat{\boldsymbol{\nu}}_j^{(k)}\right) - F_{SMN}\left(\hat{t}_{ij1}^{(k)}; \hat{\boldsymbol{\nu}}_j^{(k)}\right)}\right\},
\end{aligned}$$

$$\begin{aligned}
\widehat{uy^2}_{ij}^{(k)} &= \left(x_i^\top \hat{\boldsymbol{\beta}}_{\boldsymbol{j}}^{(k)}\right)^2 \widehat{u}_{ij}^{(k)} + 2\left(x_i^\top \hat{\boldsymbol{\beta}}_{\boldsymbol{j}}^{(k)}\right)\hat{\sigma}_j^{(k)}\widehat{uy}_{ij}^{(k)} + \hat{\sigma}_j^{2(k)} E\left(U_{ij}T_{ij}^2\middle|\hat{t}_{ij1}^{(k)} \leq T_{ij}^{(k)} \leq \hat{t}_{ij2}^{(k)}, \hat{\boldsymbol{\theta}}_j^{(k)}\right),\\
&= \left(x_i^\top \hat{\boldsymbol{\beta}}_{\boldsymbol{j}}^{(k)}\right)^2 \widehat{u}_{ij}^{(k)} + 2\left(x_i^\top \hat{\boldsymbol{\beta}}_{\boldsymbol{j}}^{(k)}\right)\hat{\sigma}_j^{(k)}\widehat{uy}_{ij}^{(k)} + \frac{\hat{\sigma}_j^{2(k)}}{F_{SMN}\left(\hat{t}_{ij2}^{(k)}; \hat{\boldsymbol{\nu}}_j^{(k)}\right) - F_{SMN}\left(\hat{t}_{ij1}^{(k)}; \hat{\boldsymbol{\nu}}_j^{(k)}\right)}\\
&\qquad \left(E_\Phi\left(0, \hat{t}_{ij2}^{(k)}\right) - E_\Phi\left(0, \hat{t}_{ij1}^{(k)}\right) + \left(\hat{t}_{ij1}^{(k)}\right)E_\phi\left(0.5, \hat{t}_{ij1}^{(k)}\right) - \left(\hat{t}_{ij2}^{(k)}\right)E_\phi\left(0.5, \hat{t}_{ij2}^{(k)}\right)\right),
\end{aligned}$$

where

$$E_\phi(r, h) = E\left(U^r\phi(h\sqrt{U})\right) \quad \text{and} \quad E_\Phi(r, h) = E\left(U^r\Phi(h\sqrt{U})\right).$$

In the following, the closed forms of $E_\phi(r, h)$ and $E_\Phi(r, h)$ for the special cases of SMN class of distributions are presented.

- For the normal distribution, we have

$$E_\phi(r, h) = \phi(h) \quad \text{and} \quad E_\Phi(r, h) = \Phi(h).$$

- In the case of Student-t distribution, we have

$$\begin{aligned}
E_\phi(r, h) &= \frac{\Gamma\left(\dfrac{\hat{\nu}_j^{(k)} + 2r}{2}\right)}{\sqrt{2\pi}\Gamma(\hat{\nu}_j^{(k)}/2)}\left(\frac{\hat{\nu}_j^{(k)}}{2}\right)^{\frac{\hat{\nu}_j^{(k)}}{2}}\left(\frac{2}{h^2 + \hat{\nu}_j^{(k)}}\right)^{\frac{\hat{\nu}_j^{(k)} + 2r}{2}},\\
E_\Phi(r, h) &= \Gamma\left(\frac{\hat{\nu}_j^{(k)} + 2r}{2}\right)\left(\frac{2}{\hat{\nu}_j^{(k)}}\right)^r F_{PVII}\left(h; \hat{\nu}_j^{(k)} + 2r, \hat{\nu}_j^{(k)}\right)\Big/\Gamma(\frac{\hat{\nu}_j^{(k)}}{2}),
\end{aligned}$$

where $F_{PVII}(\cdot; \nu, \delta)$ denotes the cdf of Pearson type VII distribution.

- For the slash model, we have

$$E_\phi(r, h) = \frac{\hat{\nu}_j^{(k)}}{\sqrt{2\pi}}\left(\frac{2}{h^2}\right)^{\hat{\nu}_j^{(k)} + r}\Gamma(\hat{\nu}_j^{(k)} + r, \frac{h^2}{2}) \quad \text{and} \quad E_\Phi(r, h) = \frac{\hat{\nu}_j^{(k)}}{\hat{\nu}_j^{(k)} + r}F_{SL}\left(h; \hat{\nu}_j^{(k)} + r\right).$$

- For the contaminated-normal distribution, we have

$$\begin{aligned}
E_\phi(r, h) &= \left(\hat{\gamma}_j^{(k)}\right)^r \hat{\nu}_j^{(k)}\phi\left(h\sqrt{\hat{\gamma}_j^{(k)}}\right) + \left(1 - \hat{\nu}_j^{(k)}\right)\phi(h),\\
E_\Phi(r, h) &= \left(\hat{\gamma}_j^{(k)}\right)^r F_{CN}\left(h; \hat{\nu}_j^{(k)}, \hat{\gamma}_j^{(k)}\right) + \left(1 - \left(\hat{\gamma}_j^{(k)}\right)^r\right)\left(1 - \hat{\nu}_j^{(k)}\right)\Phi(h).
\end{aligned}$$

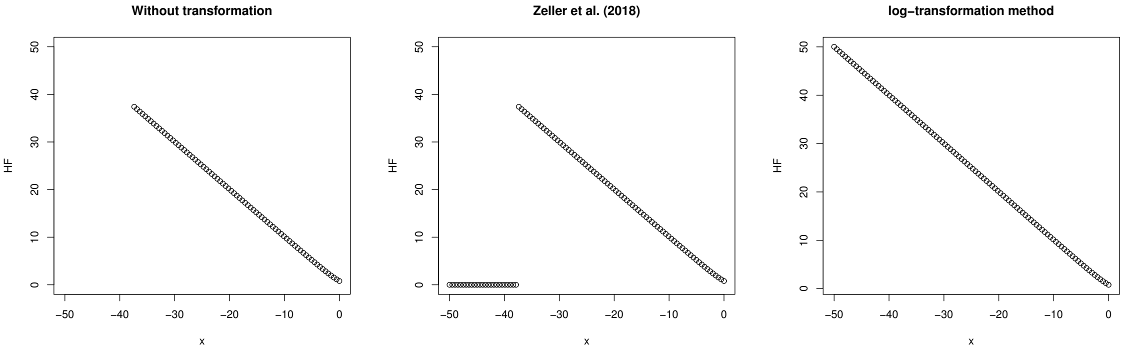

Figure B.8: The normal hazard function plots computed based on three ways in R.

Appendix B. The hazard function plots of the normal distribution

Appendix C. Further plots of the first simulation

References

Akaike, H., 1974. A new look at the statistical model identification, in: Selected Papers of Hirotugu Akaike. Springer, pp. 215–222.

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.

Birnbaum, Z.W., Saunders, S.C., 1969. A new family of life distributions. Journal of Applied Probability 6, 319–327.

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

Chamroukhi, F., 2016. Robust mixture of experts modeling using the *t* distribution. Neural Networks 79, 20–36.

Chamroukhi, F., 2017. Skew *t* mixture of experts. Neurocomputing 266, 390–408.

Cuesta-Albertos, J.A., Gordaliza, A., Matrán, C., et al., 1997. Trimmed *k*-means: An attempt to robustify quantizers. The Annals of Statistics 25, 553–576.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological) 39, 1–22.

DeSarbo, W.S., Cron, W.L., 1988. A maximum likelihood methodology for clusterwise linear regression. Journal of Classification 5, 249–282.

Filho, E.B.D.A., Garay, A.W.M., 2017. TSMN: Truncated Scale Mixtures of Normal Distributions. R package version 1.0.0.

Garay, A.M., Lachos, V.H., Bolfarine, H., Cabral, C.R., 2017. Linear censored regression models with scale mixtures of normal distributions. Statistical Papers 58, 247–278.

Garay, A.M., Lachos, V.H., Lin, T.I., 2016. Nonlinear censored regression models with heavy-tailed distributions. Statistics and Its Interface 9, 281–293.

García-Escudero, L., Gordaliza, A., Mayo-Iscar, A., Martín, R.S., 2010. Robust clusterwise linear regression through trimming. Computational Statistics & Data Analysis 54, 3057–3069.

Gómez, G., Calle, M.L., Oller, R., Langohr, K., 2009. Tutorial on methods for interval-censored data and their implementation in R. Statistical Modelling 9, 259–297.

Hartigan, J.A., Wong, M.A., 1979. Algorithm as 136: A *k*-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics) 28, 100–108.

Hu, H., Yao, W., Wu, Y., 2017. The robust EM-type algorithms for log-concave mixtures of regression models. Computational Statistics & Data Analysis 111, 14–26.

Hubert, L., Arabie, P., 1985. Comparing partitions. Journal of Classification 2, 193–218.

Hrmann, W., Leydold, J., 2013. Generating generalized inverse Gaussian random variates. Statistics and Computing 24, 547–557.

Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E., et al., 1991. Adaptive mixtures of local experts. Neural Computation 3, 79–87.

Jiang, W., Tanner, M.A., 1999. Hierarchical mixtures-of-experts for exponential family regression models: approximation and maximum likelihood estimation. Annals of Statistics , 987–1011.

Jones, P., McLachlan, G., 1992. Fitting finite mixture models in a regression context. Australian Journal of Statistics 34, 233–240.

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

Kaufman, L., Rousseeuw, P.J., 1990. Finding groups in data. John Wiley & Sons, Hoboken, New Jersey.

Lachos, V.H., Cabral, C.R., Prates, M.O., Dey, D.K., 2019. Flexible regression modeling for censored data based on mixtures of student-*t* distributions. Computational Statistics 34, 123–152.

Lachos, V.H., Garay, A.M., Cabral, C.R., et al., 2020. Moments of truncated scale mixtures of skew-normal distributions. Brazilian Journal of Probability and Statistics 34, 478–494.

Lachos, V.H., Moreno, E.J.L., Chen, K., Cabral, C.R.B., 2017. Finite mixture modeling of censored data using the multivariate student-*t* distribution. Journal of Multivariate Analysis 159, 151–167.

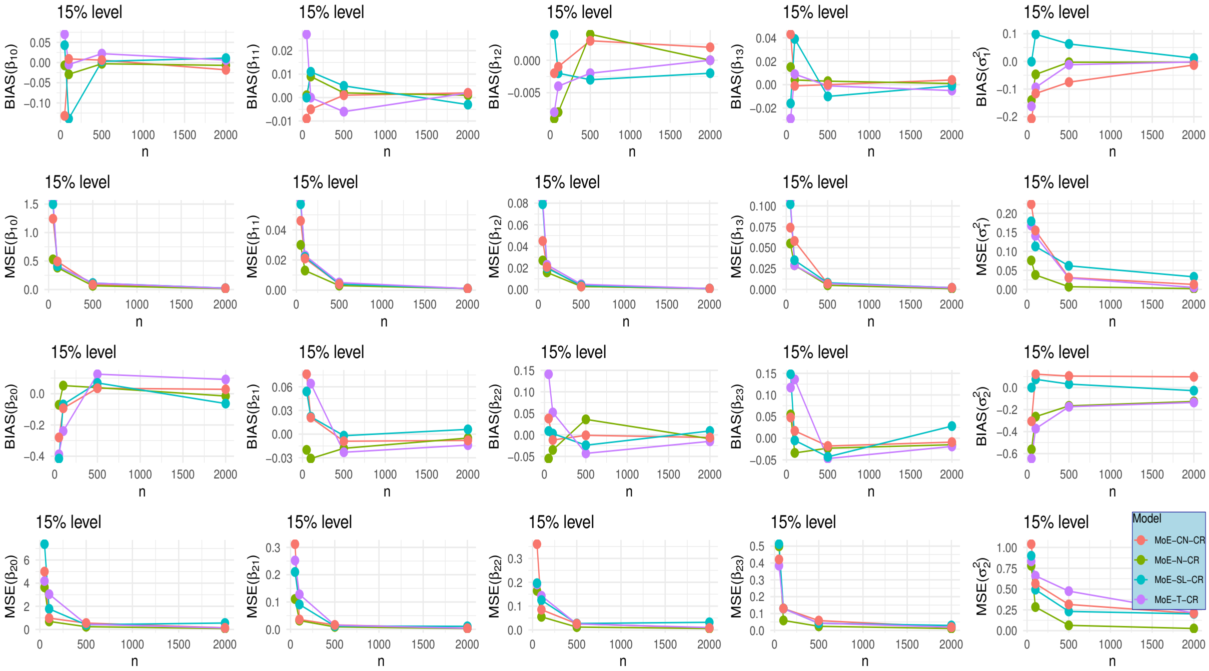

Figure C.9: The BIAS and MSE plots of $\boldsymbol{\beta}_j$ and σ_j^2 estimates for the MoE-SMN-CR model (censoring levels 15%).

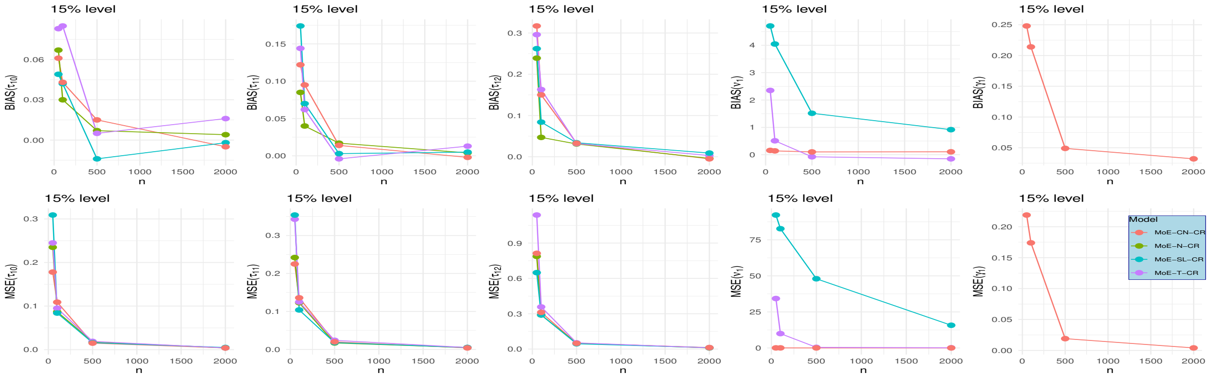

Figure C.10: The BIAS and MSE plots of $\boldsymbol{\beta}_j$ and σ_j^2 estimates for the MoE-SMN-CR model (censoring levels 15%).

Lin, T.I., Ho, H.J., Lee, C.R., 2014. Flexible mixture modelling using the multivariate skew-t-normal distribution. Statistics and Computing 24, 531–546.

Lin, T.I., Lachos, V.H., Wang, W.L., 2018. Multivariate longitudinal data analysis with censored and intermittent missing responses. Statistics in Medicine 37, 2822–2835.

Lin, T.I., Wang, W.L., 2019. Multivariate-t linear mixed models with censored responses, intermittent missing values and heavy tails. Statistical Methods in Medical Research 29, 1288–1304.

Liu, C., Rubin, D.B., 1994. The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. Biometrika 81, 633–648.

Liu, M., Lin, T.I., 2014. A skew-normal mixture regression model. Educational and Psychological Measurement 74, 139–162.

Mattos, T.d.B., Garay, A.M., Lachos, V.H., 2018. Likelihood-based inference for censored linear regression models with scale mixtures of skew-normal distributions. Journal of Applied Statistics 45, 2039–2066.

Mazza, A., Punzo, A., 2017. Mixtures of multivariate contaminated normal regression models. Statistical Papers 61, 787–822.

McLachlan, G., Peel, D., 2000. Finite mixture models. John Wiley & Sons, New York.

Meilijson, I., 1989. A fast improvement to the EM algorithm on its own terms. Journal of the Royal Statistical Society: Series B (Methodological) 51, 127–138.

Meng, X.L., Rubin, D.B., 1993. Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika 80, 267–278.

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 , 765–799.

Naderi, M., Arabpour, A., Lin, T.I., Jamalizadeh, A., 2017. Nonlinear regression models based on the normal mean-variance mixture of Birnbaum-Saunders distribution. Journal of the Korean Statistical Society 46, 476–485.

Naderi, M., Hung, W.L., Lin, T.I., Jamalizadeh, A., 2019. A novel mixture model using the multivariate normal mean-variance mixture of Birnbaum-Saunders distributions and its application to extrasolar planets. Journal of Multivariate Analysis 171, 126–138.

Nguyen, H.D., McLachlan, G.J., 2016. Laplace mixture of linear experts. Computational Statistics & Data Analysis 93, 177–191.

Niwattanakul, S., Singthongchai, J., Naenudorn, E., Wanapu, S., 2013. Using of jaccard coefficient for keywords similarity, in: Proceedings of the international multiconference of engineers and computer scientists, pp. 380–384.

Peng, F., Jacobs, R.A., Tanner, M.A., 1996. Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition. Journal of the American Statistical Association 91, 953–960.

Rand, W.M., 1971. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association 66, 846–850.

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

Tobin, J., 1958. Estimation of relationships for limited dependent variables. Econometrica: Journal of the Econometric Society 26, 24–36.

Yang, Y.C., Lin, T.I., Castro, L.M., Wang, W.L., 2020. Extending finite mixtures of t linear mixed-effects models with concomitant covariates. Computational Statistics & Data Analysis 148, 106961.

Yu, L., Chen, D.G., Liu, J., 2021. Efficient and direct estimation of the variance-covariance matrix in EM algorithm with interpolation method. Journal of Statistical Planning and Inference 211, 119–130.

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

Zeller, C.B., Cabral, C.R.B., Lachos, V.H., Benites, L., 2019. Finite mixture of regression models for censored data based on scale mixtures of normal distributions. Advances in Data Analysis and Classification 13, 89–116.

Zens, G., 2019. Bayesian shrinkage in mixture-of-experts models: identifying robust determinants of class membership. Advances in Data Analysis and Classification 13, 1019–1051.