

---

# FINITE MIXTURE MODELING OF CENSORED AND MISSING DATA USING THE MULTIVARIATE SKEW-NORMAL DISTRIBUTION

---

**Francisco H. C. de Alencar**  
Departamento de Estatística  
Universidade Estadual de Campinas  
Campinas, Brazil  
hildemardealencar@gmail.com

**Christian E. Galarza**  
Departamento de Estadística  
Escuela Superior Politecnica del Litoral  
Guayaquil, Ecuador  
chedgala@espol.edu.ec

**Larissa A. Matos**  
Departamento de Estatística  
Universidade Estadual de Campinas  
Campinas, Brazil  
larissam@unicamp.br

**Victor H. Lachos**  
Department of Statistics  
University of Connecticut  
Storrs CT 06269, U.S.A.  
hlachos@uconn.edu

September 24, 2020

## ABSTRACT

Finite mixture models have been widely used to model and analyze data from a heterogeneous population. Moreover, data of this kind can be missing or subject to some upper and/or lower detection limits because of the restriction of experimental apparatuses. Another complication arises when measures of each population depart significantly from normality, for instance, asymmetric behavior. For such data structures, we propose a robust model for censored and/or missing data based on finite mixtures of multivariate skew-normal distributions. This approach allows us to model data with great flexibility, accommodating multimodality and skewness, simultaneously, depending on the structure of the mixture components. We develop an analytically simple, yet efficient, EM-type algorithm for conducting maximum likelihood estimation of the parameters. The algorithm has closed-form expressions at the E-step that rely on formulas for the mean and variance of the truncated multivariate skew-normal distributions. Furthermore, a general information-based method for approximating the asymptotic covariance matrix of the estimators is also presented. Results obtained from the analysis of both simulated and real datasets are reported to demonstrate the effectiveness of the proposed method. The proposed algorithm and method are implemented in the new R package `CensMFM`.

**Keywords** Censored data · Detection limit · EM-type algorithms · Finite mixture models · Multivariate skew-normal distribution · Truncated distributions.

## 1 Introduction

Modeling based on finite mixture distributions is a rapidly developing area with a wide range of applications. Finite mixture models are now applied in such diverse areas as biology, biometrics, genetics, medicine and marketing, among others. There are various features of finite mixture distributions that make them useful in statistical modeling. For instance, statistical models which are based on finite mixture distributions capture many specific properties of real data such as multimodality, skewness, kurtosis, and unobserved heterogeneity. The importance of mixture distributions can be noted from the large number of books on mixtures, including [31], [14], [29], [20] and [9].

In many research areas, such as environmental pollution and infectious diseases measurements often exhibit complex features such as censored responses and missing values [24, 23]. Moreover, the proportion of censoring in these studies may be substantial, so the use of crude/ad hoc methods, such as substituting a threshold value or some arbitrary

point like a midpoint between zero and cutoff for detection, might lead to biased estimates of the model parameters. Furthermore, multivariate data are commonly seen with simultaneous occurrence of multimodality and skewness and inferential procedures become complicated when the data exhibit these features. The mixture distribution can be used quite effectively to analyze this kind of data. [21] proposed a flexible mixture modeling framework using the multivariate skew-normal distribution, where a feasible EM algorithm is developed for finding the maximum likelihood (ML) estimates. In the context of finite mixtures for correlated censored data, [16] proposed a Gaussian mixture model to flexibly approximate the underlying distribution of the observed data, where an EM algorithm in a multivariate setting was developed to cope with the censored data. More recently, [19] proposed a robust model for censored data based on finite mixtures of multivariate Student-t distributions (FM-MtC model), including the implementation of an exact EM algorithm for ML estimation. This approach allows modeling data with great flexibility, accommodating multimodality, and kurtosis depending on the structure of the mixture components. These methods are undoubtedly very flexible, but the problems related to the simultaneous occurrence of skewness, anomaly observations and multimodality remain. Even when modeling using Student-t mixtures, overestimation of the number of components necessary to capture the asymmetric nature of each subpopulation can occur [11]. So far, to the best of our knowledge there are no studies simultaneously accounting for multivariate censored responses, missing values, heterogeneity and skewness.

In this article, we propose a robust mixture model for censored data based on the multivariate skew-normal distribution so that the FM-MSNC model is defined and a fully likelihood-based approach is carried out, including the implementation of an exact EM-type algorithm for the ML estimation. The interval censoring mechanism of the proposed model allows us to handle missing and censored values simultaneously. We show that the E-step reduces to computing the first two moments of a truncated multivariate skew-normal distribution. The general formulas for these moments were derived efficiently by [15], for which we use the `MomTrunc` package in R. The likelihood function is easily computed as a byproduct of the E-step and is used for monitoring convergence and for model selection. Furthermore, we consider a general information-based method for obtaining the asymptotic covariance matrix of the ML estimate. The method proposed in this paper is implemented in the R package `CensMFM`, which is available for download from the CRAN repository.

The remainder of the paper is organized as follows. In Section 2, we briefly discuss some preliminary results related to the multivariate extended skew-normal (ESN) and related truncated extended skew-normal (TESN) distributions, in addition, to some of their key properties are presented. In section 3, we present the multivariate skew-normal censored (MSNC) model and the related ML estimation. In Section 4, we introduce the robust FM-MSNC model, including the EM algorithm for ML estimation, and derive the empirical information matrix analytically to obtain the standard errors. In Sections 5 and 6, numerical examples using both simulated and real data, respectively, are given to illustrate the performance of the proposed method. Finally, some concluding remarks are presented in Section 7.

## 2 Background

### 2.1 The multivariate skew-normal distribution

In this subsection we present the skew-normal distribution and some of its properties. We say that a  $p \times 1$  random vector  $\mathbf{Y}$  follows a multivariate SN distribution with  $p \times 1$  location vector  $\boldsymbol{\mu}$ ,  $p \times p$  positive definite dispersion matrix  $\boldsymbol{\Sigma}$  and  $p \times 1$  skewness parameter vector  $\boldsymbol{\lambda} \in \mathbb{R}^p$ , and we write  $\mathbf{Y} \sim \text{SN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ , if its pdf is given by

$$SN_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}) = 2\phi_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\Phi_1(\boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{-1/2}(\mathbf{y} - \boldsymbol{\mu})), \quad (1)$$

where  $\Phi_1(\cdot)$  represents the cumulative distribution function (cdf) of the standard univariate normal distribution. If  $\boldsymbol{\lambda} = \mathbf{0}$  then (1) reduces to the symmetric  $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  pdf which is denoted by  $\phi_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ . Except by a straightforward difference in the parameterization considered in (1), this model corresponds to that introduced by [5], whose properties were extensively studied in [4] [see also, 2].

**Proposition 1** *If  $\mathbf{Y} \sim \text{SN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ , then for any  $\mathbf{y} \in \mathbb{R}^p$*

$$F_{\mathbf{Y}}(\mathbf{y}) = P(\mathbf{Y} \leq \mathbf{y}) = 2\Phi_{p+1}(\mathbf{z}^\top, 0)^\top; \mathbf{0}, \boldsymbol{\Omega}), \quad (2)$$

where  $\mathbf{z} = \mathbf{y} - \boldsymbol{\mu}$  and  $\boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Sigma} & -\boldsymbol{\Delta} \\ -\boldsymbol{\Delta}^\top & 1 \end{pmatrix}$ , with  $\boldsymbol{\Delta} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\lambda}/(1 + \boldsymbol{\lambda}^\top \boldsymbol{\lambda})^{1/2}$ .

It is worth mentioning that the multivariate skew-normal distribution is not closed to marginalization and conditioning. Next we present its extended version which has these properties, called the multivariate ESN distribution.

### 2.2 The extended multivariate skew-normal distribution (ESN)

We say that a  $p \times 1$  random vector  $\mathbf{Y}$  follows an ESN distribution with  $p \times 1$  location vector  $\boldsymbol{\mu}$ ,  $p \times p$  positive definite dispersion matrix  $\boldsymbol{\Sigma}$ , a  $p \times 1$  skewness parameter vector  $\boldsymbol{\lambda} \in \mathbb{R}^p$ , and shift parameter  $\tau \in \mathbb{R}$ , denoted by

$\mathbf{Y} \sim \text{ESN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau)$ , if its pdf is given by

$$\text{ESN}_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau) = \xi^{-1} \phi_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \Phi_1(\tau + \boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{-1/2}(\mathbf{y} - \boldsymbol{\mu})), \quad (3)$$

with  $\xi = \Phi_1(\tau/(1 + \boldsymbol{\lambda}^\top \boldsymbol{\lambda})^{1/2})$ . Note that when  $\tau = 0$ , we retrieve the skew-normal distribution defined in (1), that is,  $\text{ESN}_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, 0) \equiv \text{SN}_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ . It is also interesting to note that

$$\text{ESN}_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau) \rightarrow \phi_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}), \text{ as } \tau \rightarrow +\infty.$$

The following propositions are crucial to develop our methods. The proofs are given in [3].

**Proposition 2** Let  $\mathbf{Y} \sim \text{ESN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau)$  and  $\mathbf{Y}$  is partitioned as  $\mathbf{Y} = (\mathbf{Y}_1^\top, \mathbf{Y}_2^\top)^\top$  of dimensions  $p_1$  and  $p_2$  ( $p_1 + p_2 = p$ ), respectively. Let

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}, \quad \boldsymbol{\mu} = (\boldsymbol{\mu}_1^\top, \boldsymbol{\mu}_2^\top)^\top, \quad \boldsymbol{\lambda} = (\boldsymbol{\lambda}_1^\top, \boldsymbol{\lambda}_2^\top)^\top \quad \text{and} \quad \boldsymbol{\varphi} = (\boldsymbol{\varphi}_1^\top, \boldsymbol{\varphi}_2^\top)^\top$$

be the corresponding partitions of  $\boldsymbol{\Sigma}$ ,  $\boldsymbol{\mu}$ ,  $\boldsymbol{\lambda}$  and  $\boldsymbol{\varphi} = \boldsymbol{\Sigma}^{-1/2} \boldsymbol{\lambda}$ . Then,

$$\begin{aligned} \mathbf{Y}_1 &\sim \text{ESN}_{p_1}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}, c_{12} \boldsymbol{\Sigma}_{11}^{1/2} \tilde{\boldsymbol{\varphi}}_1, c_{12} \tau), \\ \mathbf{Y}_2 \mid \mathbf{Y}_1 = \mathbf{y}_1 &\sim \text{ESN}_{p_2}(\boldsymbol{\mu}_{2.1}, \boldsymbol{\Sigma}_{22.1}, \boldsymbol{\Sigma}_{22.1}^{1/2} \boldsymbol{\varphi}_2, \tau_{2.1}) \end{aligned}$$

where  $c_{12} = (1 + \boldsymbol{\varphi}_2^\top \boldsymbol{\Sigma}_{22.1} \boldsymbol{\varphi}_2)^{-1/2}$ ,  $\tilde{\boldsymbol{\varphi}}_1 = \boldsymbol{\varphi}_1 + \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12} \boldsymbol{\varphi}_2$ ,  $\boldsymbol{\Sigma}_{22.1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12}$ ,  $\boldsymbol{\mu}_{2.1} = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} (\mathbf{y}_1 - \boldsymbol{\mu}_1)$  and  $\tau_{2.1} = \tau + \tilde{\boldsymbol{\varphi}}_1^\top (\mathbf{y}_1 - \boldsymbol{\mu}_1)$ .

**Proposition 3** If  $\mathbf{Y} \sim \text{ESN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau)$ , then for any  $\mathbf{y} \in \mathbb{R}^p$

$$F_{\mathbf{Y}}(\mathbf{y}) = P(\mathbf{Y} \leq \mathbf{y}) = \frac{\Phi_{p+1}(\mathbf{z}^\top, \tilde{\tau}; \mathbf{0}, \boldsymbol{\Omega})}{\Phi_1(\tilde{\tau})}, \quad (4)$$

with  $\mathbf{z}$  and  $\boldsymbol{\Omega}$  as defined in Proposition 1, and  $\tilde{\tau} = \tau/(1 + \boldsymbol{\lambda}^\top \boldsymbol{\lambda})^{1/2}$ .

Hereafter, for  $\mathbf{Y} \sim \text{ESN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau)$ , we will denote its cdf as  $F_{\mathbf{Y}}(\mathbf{y}) \equiv \tilde{\Phi}_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau)$  for simplicity.

Let  $\mathbb{A}$  be a Borel set in  $\mathbb{R}^p$ . We say that the random vector  $\mathbf{Y}$  has a truncated extended skew-normal distribution on  $\mathbb{A}$  when  $\mathbf{Y}$  has the same distribution as  $\mathbf{Y} \mid (\mathbf{Y} \in \mathbb{A})$ . In this case, the pdf of  $\mathbf{Y}$  is given by

$$f(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\nu}; \mathbb{A}) = \frac{\text{ESN}_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau)}{P(\mathbf{Y} \in \mathbb{A})} \mathbf{1}_{\mathbb{A}}(\mathbf{y}),$$

where  $\mathbf{1}_{\mathbb{A}}$  is the indicator function of  $\mathbb{A}$ . We use the notation  $\mathbf{Y} \sim \text{TESN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau; \mathbb{A})$ . If  $\mathbb{A}$  has the form

$$\begin{aligned} \mathbb{A} &= \{(x_1, \dots, x_p) \in \mathbb{R}^p : a_1 \leq x_1 \leq b_1, \dots, a_p \leq x_p \leq b_p\} \\ &= \{\mathbf{x} \in \mathbb{R}^p : \mathbf{a} \leq \mathbf{x} \leq \mathbf{b}\}, \end{aligned} \quad (5)$$

then we use the notation  $\{\mathbf{Y} \in \mathbb{A}\} = \{\mathbf{a} \leq \mathbf{Y} \leq \mathbf{b}\}$ , where  $\mathbf{a} = (a_1, \dots, a_p)^\top$  and  $\mathbf{b} = (b_1, \dots, b_p)^\top$ . Here, we say that the distribution of  $\mathbf{Y}$  is doubly truncated. Analogously we define  $\{\mathbf{Y} \geq \mathbf{a}\}$  and  $\{\mathbf{Y} \leq \mathbf{b}\}$ . Thus, we say that the distribution of  $\mathbf{Y}$  is truncated from below and truncated from above, respectively. For convenience, we also use the notation  $\mathbf{Y} \sim \text{TESN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau; [\mathbf{a}, \mathbf{b}])$ . In particular, we denote  $\mathbf{W}$  to follow a truncated  $p$ -variate normal distribution on  $[\mathbf{a}, \mathbf{b}]$  as  $\mathbf{W} \sim \text{TN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}; [\mathbf{a}, \mathbf{b}])$ .

For the general doubly truncated case, we define the normalizing constant  $\mathcal{L}_p(\mathbf{a}, \mathbf{b}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau) = P(\mathbf{a} \leq \mathbf{Y} \leq \mathbf{b})$  as

$$\mathcal{L}_p(\mathbf{a}, \mathbf{b}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau) = \int_{\mathbf{a}}^{\mathbf{b}} \text{ESN}_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau) d\mathbf{y}.$$

When all  $\boldsymbol{\lambda}$  and  $\tau$  are equal to zero, we have a normal integral  $\mathcal{L}_p(\mathbf{a}, \mathbf{b}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{0}, 0) = L_p(\mathbf{a}, \mathbf{b}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \int_{\mathbf{a}}^{\mathbf{b}} \phi_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{y}$ . Note that we use calligraphic style  $\mathcal{L}_p$  when we work with the skewed extended version and Roman style  $L_p$  for the symmetric case.

The following properties of the truncated multivariate extended skew-normal distributions are useful for implementation of the EM-algorithm. The proofs are given in [15].

**Proposition 4** Let  $\mathbf{Y} \sim \text{TESN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau; [\mathbf{a}, \mathbf{b}])$ . For any measurable function  $g(\cdot)$ , we have that

$$\mathbb{E} \left[ g(\mathbf{Y}) \frac{\phi_1(\tau + \boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{-1/2}(\mathbf{Y} - \boldsymbol{\mu}))}{\Phi_1(\tau + \boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{-1/2}(\mathbf{Y} - \boldsymbol{\mu}))} \right] = \eta \frac{L_p(\mathbf{a}, \mathbf{b}; \boldsymbol{\mu} - \boldsymbol{\mu}_b, \boldsymbol{\Gamma})}{\mathcal{L}_p(\mathbf{a}, \mathbf{b}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau)} \mathbb{E}[g(\mathbf{W})], \quad (6)$$

with  $\eta = \phi_1(\tau; 0, 1 + \boldsymbol{\lambda}^\top \boldsymbol{\lambda})/\xi$ ,  $\boldsymbol{\mu}_b = \tilde{\tau} \boldsymbol{\Delta}$ ,  $\boldsymbol{\Gamma} = \boldsymbol{\Sigma} - \boldsymbol{\Delta} \boldsymbol{\Delta}^\top$  and  $\mathbf{W} \sim \text{TN}_p(\boldsymbol{\mu} - \boldsymbol{\mu}_b, \boldsymbol{\Gamma}; [\mathbf{a}, \mathbf{b}])$ .

**Proposition 5** Let  $\mathbf{Y} \sim ESN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau; [\mathbf{a}, \mathbf{b}])$ , where  $\mathbf{Y}$  is partitioned as  $\mathbf{Y} = (\mathbf{Y}_1^\top, \mathbf{Y}_2^\top)^\top$  of dimensions  $p_1$  and  $p_2$  ( $p_1 + p_2 = p$ ), with corresponding partitions of  $\mathbf{a}$ ,  $\mathbf{b}$ ,  $\boldsymbol{\mu}$ ,  $\boldsymbol{\Sigma}$ ,  $\boldsymbol{\lambda}$  and  $\varphi$ . Then, for any measurable function  $g(\cdot)$ , we have that

$$\mathbb{E}_{\mathbf{Y}_2} \left[ g(\mathbf{Y}_2) \frac{\phi_1(\tau + \boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{-1/2}(\mathbf{Y} - \boldsymbol{\mu}))}{\Phi_1(\tau + \boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{-1/2}(\mathbf{Y} - \boldsymbol{\mu}))} \middle| \mathbf{Y}_1 \right] = \frac{\eta_{2.1} L_{2.1}}{\mathcal{L}_{2.1}} \mathbb{E}[g(\mathbf{W}_2)], \quad (7)$$

where  $L_{2.1} = L_{p_2}(\mathbf{a}_2, \mathbf{b}_2; \boldsymbol{\mu}_{2.1} - \boldsymbol{\mu}_{b2.1}, \boldsymbol{\Gamma}_{22.1})$ ,  $\mathcal{L}_{2.1} = \mathcal{L}_{p_2}(\mathbf{a}_2, \mathbf{b}_2; \boldsymbol{\mu}_{2.1}, \boldsymbol{\Sigma}_{22.1}, \boldsymbol{\lambda}_{2.1}, \tau_{2.1})$  and  $\mathbf{W}_2 \sim TN_p(\boldsymbol{\mu}_{2.1} - \boldsymbol{\mu}_{b2.1}, \boldsymbol{\Gamma}_{22.1}, [\mathbf{a}_2, \mathbf{b}_2])$  with  $\boldsymbol{\lambda}_{2.1} = \boldsymbol{\Sigma}_{22.1}^{1/2} \boldsymbol{\varphi}_2$ ,  $\boldsymbol{\mu}_{2.1}$ ,  $\boldsymbol{\Sigma}_{22.1}$ , and  $\tau_{2.1}$  as in proposition 2, and  $\eta_{2.1}$ ,  $\boldsymbol{\mu}_{b2.1}$  and  $\boldsymbol{\Gamma}_{22.1}$  can be computed as expressions  $\eta$ ,  $\boldsymbol{\mu}_b$  and  $\boldsymbol{\Gamma}$  in proposition 4 but using the new set of parameters  $\boldsymbol{\mu}_{2.1}$ ,  $\boldsymbol{\Sigma}_{22.1}$ ,  $\boldsymbol{\lambda}_{2.1}$  and  $\tau_{2.1}$  (instead of  $\boldsymbol{\mu}$ ,  $\boldsymbol{\Sigma}$ ,  $\boldsymbol{\lambda}$  and  $\tau$ ).

Observe that Propositions 4 and 5 depend on formulas for  $g(\mathbf{Y})$ , where  $\mathbf{Y} \sim TESN(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau; [\mathbf{a}, \mathbf{b}])$ . Closed form expressions for these expectations were obtained recently by [15], for which the `meanvarTMD()` function of the `R MomTrunc` library can be used.

### 3 Multivariate skew-normal model for censored and missing responses

Now we present the robust multivariate skew-normal model for censored data. So, we write

$$\mathbf{Y}_1, \dots, \mathbf{Y}_n \sim SN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}), \quad (8)$$

where for each  $i \in \{1, \dots, n\}$ ,  $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{ip})^\top$  is a  $p \times 1$  vector of responses for sample unit  $i$ ,  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)^\top$  is the location vector and the dispersion matrix  $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\alpha})$  depends on an unknown and reduced parameter vector  $\boldsymbol{\alpha}$  and skewness parameter  $\boldsymbol{\lambda}$ . We assume that  $\mathbf{Y}_1, \dots, \mathbf{Y}_n$  are independent and identically distributed. We consider a similar approach to that proposed by [19] to model the censored responses. Thus, the observed data for the  $i$ th subject are given by  $(\mathbf{V}_i, \mathbf{C}_i)$ , where each element of  $\mathbf{V}_i = (V_{i1}, \dots, V_{ip})^\top$  represents either the vector of uncensored observations ( $V_{ik} = V_{0i}$ ) or the interval censoring level ( $V_{ik} \in [V_{1ik}, V_{2ik}]$ ), and  $\mathbf{C}_i = (C_{i1}, \dots, C_{ip})^\top$  is the vector of censoring indicators, satisfying

$$C_{ik} = \begin{cases} 1 & \text{if } V_{1ik} \leq Y_{ik} \leq V_{2ik}; \\ 0 & \text{if } Y_{ik} = V_{0i}. \end{cases} \quad (9)$$

for all  $i \in \{1, \dots, n\}$  and  $k \in \{1, \dots, p\}$ , i.e.,  $C_{ik} = 1$  if  $Y_{ik}$  is located within a specific interval. In this case, (8) and (9) define the multivariate skew-normal interval censored model (hereafter, the MSNC model). Missing observations can be handled by considering  $V_{1ik} = -\infty$  and  $V_{2ik} = +\infty$ .

#### 3.1 The likelihood function

Let  $\mathbf{y} = (\mathbf{y}_1^\top, \dots, \mathbf{y}_n^\top)^\top$ , where  $\mathbf{y}_i = (y_{i1}, \dots, y_{ip})^\top$  is a realization of  $\mathbf{Y}_i \sim SN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ . To obtain the likelihood function of the MSNC model, we first treat the observed and censored components of  $\mathbf{y}_i$ , separately, i.e.,  $\mathbf{y}_i = (\mathbf{y}_i^o, \mathbf{y}_i^c)^\top$ , where  $C_{ik} = 0$  for all elements in the  $p_i^o$ -dimensional vector  $\mathbf{y}_i^o$ , and  $C_{ik} = 1$  for all elements in the  $p_i^c$ -dimensional vector  $\mathbf{y}_i^c$ . Accordingly, we write  $\mathbf{V}_i = \text{vec}(\mathbf{V}_i^o, \mathbf{V}_i^c)$ , where  $\mathbf{V}_i^c = (\mathbf{V}_{1i}^c, \mathbf{V}_{2i}^c)$  with

$$\begin{aligned} \boldsymbol{\mu}_i &= (\boldsymbol{\mu}_i^o, \boldsymbol{\mu}_i^c)^\top, \quad \boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\alpha}) = \begin{pmatrix} \boldsymbol{\Sigma}_i^{oo} & \boldsymbol{\Sigma}_i^{oc} \\ \boldsymbol{\Sigma}_i^{co} & \boldsymbol{\Sigma}_i^{cc} \end{pmatrix}, \\ \boldsymbol{\lambda}_i &= (\boldsymbol{\lambda}_i^o, \boldsymbol{\lambda}_i^c)^\top \quad \text{and} \quad \boldsymbol{\varphi}_i = (\boldsymbol{\varphi}_i^o, \boldsymbol{\varphi}_i^c)^\top. \end{aligned} \quad (10)$$

Then, using Proposition 2, we have that  $\mathbf{Y}_i^o \sim SN_{p_i^o}(\boldsymbol{\mu}_i^o, \boldsymbol{\Sigma}_i^{oo}, c_i^{oc} \boldsymbol{\Sigma}_i^{oo1/2} \tilde{\boldsymbol{\varphi}}_i^o)$  and  $\mathbf{Y}_i^c \mid \mathbf{Y}_i^o = \mathbf{y}_i^o \sim ESN_{p_i^c}(\boldsymbol{\mu}_i^{co}, \boldsymbol{\Sigma}_i^{cc.o}, \boldsymbol{\Sigma}_i^{cc.o1/2} \boldsymbol{\varphi}_i^c, \tau_i^{co})$ , where

$$\boldsymbol{\mu}_i^{co} = \boldsymbol{\mu}_i^c + \boldsymbol{\Sigma}_i^{co} \boldsymbol{\Sigma}_i^{oo-1} (\mathbf{y}_i^o - \boldsymbol{\mu}_i^o), \quad \boldsymbol{\Sigma}_i^{cc.o} = \boldsymbol{\Sigma}_i^{cc} - \boldsymbol{\Sigma}_i^{co} (\boldsymbol{\Sigma}_i^{oo})^{-1} \boldsymbol{\Sigma}_i^{oc}, \quad (11)$$

$$\tilde{\boldsymbol{\varphi}}_i^o = \boldsymbol{\varphi}_i^o + \boldsymbol{\Sigma}_i^{oo-1} \boldsymbol{\Sigma}_i^{oc} \boldsymbol{\varphi}_i^c, \quad c_i^{oc} = (1 + \boldsymbol{\varphi}_i^{c\top} \boldsymbol{\Sigma}_i^{cc.o} \boldsymbol{\varphi}_i^c)^{-1/2} \quad \text{and} \quad \tau_i^{co} = \tilde{\boldsymbol{\varphi}}_i^{o\top} (\mathbf{y}_i^o - \boldsymbol{\mu}_i^o). \quad (12)$$

Let  $\mathbf{V} = \text{vec}(\mathbf{V}_1, \dots, \mathbf{V}_n)$  and  $\mathbf{C} = \text{vec}(\mathbf{C}_1, \dots, \mathbf{C}_n)$  denote the observed data. Therefore, the log-likelihood function of  $\boldsymbol{\theta} = (\boldsymbol{\mu}^\top, \boldsymbol{\alpha}^\top, \boldsymbol{\lambda}^\top)^\top$ , given the observed data  $(\mathbf{V}, \mathbf{C})$  is

$$\ell(\boldsymbol{\theta} \mid \mathbf{V}, \mathbf{C}) = \sum_{i=1}^n \ln L_i, \quad (13)$$

where  $L_i$  represents the likelihood function of  $\theta$  for the  $i$ th sample, given by

$$\begin{aligned} L_i &\equiv L_i(\theta | \mathbf{V}_i, \mathbf{C}_i) = f(\mathbf{V}_i | \mathbf{C}_i, \theta) = f(\mathbf{V}_{1i}^c \leq \mathbf{y}_i^c \leq \mathbf{V}_{2i}^c | \mathbf{y}_i^o, \theta) f(\mathbf{y}_i^o | \theta) \\ &= \mathcal{L}_{p_i^c}(\mathbf{V}_{1i}^c, \mathbf{V}_{2i}^c; \boldsymbol{\mu}_i^{co}, \boldsymbol{\Sigma}_i^{cc,o}, \boldsymbol{\Sigma}_i^{cc,o1/2} \boldsymbol{\varphi}_i^c, \tau_i^{co}) \\ &\quad \times SN_{p_i^o}(\mathbf{y}_i^o; \boldsymbol{\mu}_i^o, \boldsymbol{\Sigma}_i^{oo}, c_i^{oo} \boldsymbol{\Sigma}_i^{oo1/2} \hat{\boldsymbol{\varphi}}_i^o). \end{aligned}$$

### 3.2 Parameter estimation via the EM algorithm

We now describe how to carry out ML estimation for the MSNC model. The EM algorithm, originally proposed by [13], is a very popular iterative optimization strategy and commonly used to obtain ML estimates for incomplete-data problems. This algorithm has many attractive features, such as numerical stability, simplicity of implementation and quite reasonable memory requirement [27].

By the essential property of a multivariate SN distribution, we can write

$$\mathbf{Y}_i | (T_i = t_i) \sim N_p(\boldsymbol{\mu} + \boldsymbol{\Delta} t_i, \boldsymbol{\Gamma}) \text{ and } T_i \sim \text{HN}(0, 1), \quad (14)$$

with HN referring to a half normal distribution and with  $\boldsymbol{\Delta}$  and  $\boldsymbol{\Gamma}$  as defined in the previous section. The complete-data log-likelihood function of an equivalent set of parameters  $\boldsymbol{\theta} = (\boldsymbol{\mu}^\top, \boldsymbol{\Delta}^\top, \boldsymbol{\alpha}_\Gamma^\top)^\top$ , where  $\boldsymbol{\alpha}_\Gamma = \text{vech}(\boldsymbol{\Gamma})$ , is given by  $\ell_c(\boldsymbol{\theta}) = \sum_{i=1}^n \ell_{ic}(\boldsymbol{\theta})$ , where the individual complete-data log-likelihood is

$$\ell_{ic}(\boldsymbol{\theta}) = -\frac{1}{2} \{ \ln |\boldsymbol{\Gamma}| + (\mathbf{y}_i - \boldsymbol{\mu} - \boldsymbol{\Delta} t_i)^\top \boldsymbol{\Gamma}^{-1} (\mathbf{y}_i - \boldsymbol{\mu} - \boldsymbol{\Delta} t_i) \} + c,$$

with  $c$  being a constant that does not depend on  $\boldsymbol{\theta}$ . Subsequently, the EM algorithm for the MSNC model can be summarized as follows:

**E-step:** Given the current estimate  $\hat{\boldsymbol{\theta}}^{(k)} = (\hat{\boldsymbol{\mu}}^{(k)}, \hat{\boldsymbol{\Delta}}^{(k)}, \hat{\boldsymbol{\alpha}}_\Gamma^{(k)})$  at the  $k$ th step of the algorithm, the E-step provides the conditional expectation of the complete data log-likelihood function

$$Q(\boldsymbol{\theta} | \hat{\boldsymbol{\theta}}^{(k)}) = \mathbb{E}[\ell_c(\boldsymbol{\theta}) | \mathbf{V}, \mathbf{C}, \hat{\boldsymbol{\theta}}^{(k)}] = \sum_{i=1}^n Q_i(\boldsymbol{\theta} | \hat{\boldsymbol{\theta}}^{(k)}),$$

where

$$\begin{aligned} Q_i(\boldsymbol{\theta} | \hat{\boldsymbol{\theta}}^{(k)}) &\propto -\frac{1}{2} \ln |\hat{\boldsymbol{\Gamma}}^{(k)}| - \frac{1}{2} \text{tr} \left[ \left\{ \hat{\mathbf{y}}_i^{2(k)} + \hat{\boldsymbol{\mu}}^{(k)} \hat{\boldsymbol{\mu}}^{(k)\top} + \hat{t}_i^{2(k)} \hat{\boldsymbol{\Delta}}^{(k)} \hat{\boldsymbol{\Delta}}^{(k)\top} \right. \right. \\ &\quad - \hat{\boldsymbol{\mu}}^{(k)} \hat{\mathbf{y}}_i^{(k)\top} - \hat{\mathbf{y}}_i^{(k)} \hat{\boldsymbol{\mu}}^{(k)\top} - \hat{t}_i^{(k)} \hat{\boldsymbol{\Delta}}^{(k)\top} - \hat{\boldsymbol{\Delta}}^{(k)} \hat{t}_i^{(k)\top} \\ &\quad \left. \left. + \hat{t}_i^{(k)} \hat{\boldsymbol{\Delta}}^{(k)} \hat{\boldsymbol{\mu}}^{(k)\top} + \hat{t}_i^{(k)} \hat{\boldsymbol{\mu}}^{(k)} \hat{\boldsymbol{\Delta}}^{(k)\top} \right\} \hat{\boldsymbol{\Gamma}}^{-1(k)} \right], \end{aligned}$$

with  $\hat{\mathbf{y}}_i^{r(k)} = \mathbb{E}_{T_i, \mathbf{Y}_i}[\mathbf{Y}_i^r | \mathbf{V}_i, \mathbf{C}_i, \hat{\boldsymbol{\theta}}^{(k)}]$ ,  $\hat{t}_i^{r(k)} = \mathbb{E}_{T_i, \mathbf{Y}_i}[T_i^r | \mathbf{V}_i, \mathbf{C}_i, \hat{\boldsymbol{\theta}}^{(k)}]$  (for  $r = \{0, 1, 2\}$ , with  $\mathbf{Y}_i^0 = 1$ ,  $\mathbf{Y}_i^1 = \mathbf{Y}_i$  and  $\mathbf{Y}_i^2 = \mathbf{Y}_i \mathbf{Y}_i^\top$ ) and  $\hat{t}_i^{(k)} = \mathbb{E}_{T_i, \mathbf{Y}_i}[T_i \mathbf{Y}_i | \mathbf{V}_i, \mathbf{C}_i, \hat{\boldsymbol{\theta}}^{(k)}]$ . Then, we can use Propositions 4 and 5 to obtain closed form expressions for these conditional expectations as follows:

1. If the  $i$ th subject has only non-censored components, then

$$\begin{aligned} \hat{\mathbf{y}}_i^{r(k)} &= \mathbb{E}_{\mathbf{Y}_i}[\mathbf{Y}_i^r | \mathbf{V}_i, \mathbf{C}_i, \hat{\boldsymbol{\theta}}^{(k)}] = \mathbf{y}_i^r, \\ \hat{t}_i^{r(k)} &= \mathbb{E}_{T_i, \mathbf{Y}_i}[T_i^r | \mathbf{V}_i, \mathbf{C}_i, \hat{\boldsymbol{\theta}}^{(k)}] = \mathbb{E}_{T_i}[T_i^r | \mathbf{Y}_i, \hat{\boldsymbol{\theta}}^{(k)}], \\ \hat{t}_i^{(k)} &= \mathbb{E}_{T_i, \mathbf{Y}_i}[T_i \mathbf{Y}_i | \mathbf{V}_i, \mathbf{C}_i, \hat{\boldsymbol{\theta}}^{(k)}] = \mathbf{y}_i \mathbb{E}_{T_i}[T_i | \mathbf{Y}_i, \hat{\boldsymbol{\theta}}^{(k)}], \end{aligned}$$

with  $\mathbf{y}_i^0 = 1$ ,  $\mathbf{y}_i^1 = \mathbf{y}_i$  and  $\mathbf{y}_i^2 = \mathbf{y}_i \mathbf{y}_i^\top$  and  $\mathbb{E}_{T_i}[T_i^r | \mathbf{Y}_i, \hat{\boldsymbol{\theta}}^{(k)}] = \mathbb{E}_{T_i}[T_i^r | \mathbf{Y}_i]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(k)}}$  for  $r = \{1, 2\}$ . These last conditional expectations can be obtained directly from the results given in [11].

2. If the  $i$ th subject has only censored components, from Proposition 4 we have

$$\begin{aligned} \hat{\mathbf{y}}_i^{r(k)} &= \mathbb{E}_{\mathbf{Y}_i}[\mathbf{Y}_i^r | \mathbf{V}_i, \mathbf{C}_i, \hat{\boldsymbol{\theta}}^{(k)}] = \hat{\mathbf{w}}_i^{r(k)}, \\ \hat{t}_i^{(k)} &= M^2(\hat{\boldsymbol{\theta}}^{(k)}) \hat{\boldsymbol{\Delta}}^{(k)\top} \hat{\boldsymbol{\Gamma}}^{-1(k)} (\hat{\mathbf{w}}_i^{(k)} - \hat{\boldsymbol{\mu}}^{(k)}) + \hat{\gamma}_i^{(k)} M(\hat{\boldsymbol{\theta}}^{(k)}), \\ \hat{t}_i^{2(k)} &= M^4(\hat{\boldsymbol{\theta}}^{(k)}) \hat{\boldsymbol{\Delta}}^{(k)\top} \hat{\boldsymbol{\Gamma}}^{-1(k)} (\hat{\mathbf{w}}_i^{2(k)} - 2\hat{\mathbf{w}}_i^{(k)} \hat{\boldsymbol{\mu}}^{(k)\top} + \hat{\boldsymbol{\mu}}^{(k)} \hat{\boldsymbol{\mu}}^{(k)\top}) \hat{\boldsymbol{\Gamma}}^{-1(k)} \hat{\boldsymbol{\Delta}}^{(k)} \\ &\quad + M^2(\hat{\boldsymbol{\theta}}^{(k)}) + \hat{\gamma}_i^{(k)} M^3(\hat{\boldsymbol{\theta}}^{(k)}) \hat{\boldsymbol{\Delta}}^{(k)\top} \hat{\boldsymbol{\Gamma}}^{-1(k)} (\hat{\mathbf{w}}_{0i}^{(k)} - \hat{\boldsymbol{\mu}}^{(k)}), \\ \hat{t}_i^{(k)} &= M^2(\hat{\boldsymbol{\theta}}^{(k)}) (\hat{\mathbf{w}}_i^{2(k)} - \hat{\mathbf{w}}_i^{(k)} \hat{\boldsymbol{\mu}}^{(k)\top}) \hat{\boldsymbol{\Gamma}}^{-1(k)} \hat{\boldsymbol{\Delta}}^{(k)} + \hat{\gamma}_i^{(k)} M(\hat{\boldsymbol{\theta}}^{(k)}) \hat{\mathbf{w}}_{0i}^{(k)}, \end{aligned}$$

where

$$\begin{aligned} M^2(\boldsymbol{\theta}) &= (1 + \boldsymbol{\Delta}^\top \boldsymbol{\Gamma}^{-1} \boldsymbol{\Delta})^{-1}, \quad \widehat{\mathbf{w}}_i^{(k)} = \mathbb{E}[\mathbf{W}_i | \widehat{\boldsymbol{\theta}}^{(k)}], \\ \widehat{\mathbf{w}}_i^{2(k)} &= \mathbb{E}[\mathbf{W}_i \mathbf{W}_i^\top | \widehat{\boldsymbol{\theta}}^{(k)}] \quad \text{and} \quad \widehat{\mathbf{w}}_{0i}^{(k)} = \mathbb{E}[\mathbf{W}_{0i} | \widehat{\boldsymbol{\theta}}^{(k)}], \end{aligned}$$

with  $\mathbf{W}_i \sim \text{TSN}_p(\widehat{\boldsymbol{\mu}}^{(k)}, \widehat{\boldsymbol{\Sigma}}^{(k)}, \widehat{\boldsymbol{\lambda}}^{(k)}, [\mathbf{v}_{1i}, \mathbf{v}_{2i}])$ ,  $\mathbf{W}_{0i} \sim \text{TN}_p(\widehat{\boldsymbol{\mu}}^{(k)}, \widehat{\boldsymbol{\Gamma}}^{(k)}, [\mathbf{v}_{1i}, \mathbf{v}_{2i}])$  and

$$\widehat{\gamma}_i^{(k)} = \frac{1}{\sqrt{\frac{\pi}{2}(1 + \widehat{\boldsymbol{\lambda}}^{(k)\top} \widehat{\boldsymbol{\lambda}}^{(k)})}} \frac{L_p(\mathbf{v}_{1i}, \mathbf{v}_{2i}; \widehat{\boldsymbol{\mu}}^{(k)}, \widehat{\boldsymbol{\Gamma}}^{(k)})}{\mathcal{L}_p(\mathbf{v}_{1i}, \mathbf{v}_{2i}; \widehat{\boldsymbol{\mu}}^{(k)}, \widehat{\boldsymbol{\Sigma}}^{(k)}, \widehat{\boldsymbol{\lambda}}^{(k)}, 0)}.$$

3. If the  $i$ th subject has both censored and uncensored components and given that  $(\mathbf{Y}_i | \mathbf{V}_i, \mathbf{C}_i)$ ,  $(\mathbf{Y}_i | \mathbf{V}_i, \mathbf{C}_i, \mathbf{Y}_i^o)$ , and  $(\mathbf{Y}_i^c | \mathbf{V}_i, \mathbf{C}_i, \mathbf{Y}_i^o)$  are equivalent processes, we have from Proposition 5 that

$$\begin{aligned} \widehat{\mathbf{y}}_i^{(k)} &= \mathbb{E}(\mathbf{Y}_i | \mathbf{Y}_i^o, \mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}^{(k)}) = \text{vec}(\mathbf{y}_i^o, \widehat{\mathbf{w}}_i^{c(k)}), \\ \widehat{\mathbf{y}}_i^{2(k)} &= \mathbb{E}(\mathbf{Y}_i \mathbf{Y}_i^\top | \mathbf{Y}_i^o, \mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}^{(k)}) = \begin{pmatrix} \mathbf{y}_i^o \mathbf{y}_i^o{}^\top & \mathbf{y}_i^o \widehat{\mathbf{w}}_i^{c(k)\top} \\ \widehat{\mathbf{w}}_i^{c(k)} \mathbf{y}_i^o{}^\top & \widehat{\mathbf{w}}_i^{2c(k)} \end{pmatrix}, \\ \widehat{\mathbf{y}}_{0i}^{(k)} &= \text{vec}(\mathbf{y}_i^o, \widehat{\mathbf{w}}_{0i}^{c(k)}), \\ \widehat{t}_i^{(k)} &= M^2(\widehat{\boldsymbol{\theta}}^{(k)}) \widehat{\boldsymbol{\Delta}}^{(k)\top} \widehat{\boldsymbol{\Gamma}}^{-1(k)} (\widehat{\mathbf{y}}_i^{(k)} - \widehat{\boldsymbol{\mu}}^{(k)}) + \widehat{\gamma}_i^{(k)} M(\widehat{\boldsymbol{\theta}}^{(k)}), \\ \widehat{t}_i^{2(k)} &= M^4(\widehat{\boldsymbol{\theta}}^{(k)}) \widehat{\boldsymbol{\Delta}}^{(k)\top} \widehat{\boldsymbol{\Gamma}}^{-1(k)} (\widehat{\mathbf{y}}_i^{2(k)} - 2\widehat{\mathbf{y}}_i^{(k)} \widehat{\boldsymbol{\mu}}^{(k)\top} + \widehat{\boldsymbol{\mu}}^{(k)} \widehat{\boldsymbol{\mu}}^{(k)\top}) \widehat{\boldsymbol{\Gamma}}^{-1(k)} \widehat{\boldsymbol{\Delta}}^{(k)} \\ &\quad + M^2(\widehat{\boldsymbol{\theta}}^{(k)}) + \widehat{\gamma}_i^{(k)} M^3(\widehat{\boldsymbol{\theta}}^{(k)}) \widehat{\boldsymbol{\Delta}}^{(k)\top} \widehat{\boldsymbol{\Gamma}}^{-1(k)} (\widehat{\mathbf{y}}_{0i}^{(k)} - \widehat{\boldsymbol{\mu}}^{(k)}), \\ \widehat{t\mathbf{y}}_i^{(k)} &= M^2(\widehat{\boldsymbol{\theta}}^{(k)}) (\widehat{\mathbf{y}}_i^{2(k)} - \widehat{\mathbf{y}}_i^{(k)} \widehat{\boldsymbol{\mu}}^{(k)\top}) \widehat{\boldsymbol{\Gamma}}^{-1(k)} \widehat{\boldsymbol{\Delta}}^{(k)} + \widehat{\gamma}_i^{(k)} M(\widehat{\boldsymbol{\theta}}^{(k)}) \widehat{\mathbf{y}}_{0i}^{(k)}, \end{aligned}$$

where

$$\widehat{\mathbf{w}}_i^{c(k)} = \mathbb{E}[\mathbf{W}_i^c | \widehat{\boldsymbol{\theta}}^{(k)}], \quad \widehat{\mathbf{w}}_i^{2c(k)} = \mathbb{E}[\mathbf{W}_i^c \mathbf{W}_i^{c\top} | \widehat{\boldsymbol{\theta}}^{(k)}] \quad \text{and} \quad \widehat{\mathbf{w}}_{0i}^{c(k)} = \mathbb{E}[\mathbf{W}_{0i}^c | \widehat{\boldsymbol{\theta}}^{(k)}],$$

with  $\mathbf{W}_i^c \sim \text{TESN}_{p_i^c}(\widehat{\boldsymbol{\mu}}_i^{co(k)}, \widehat{\boldsymbol{\Sigma}}_i^{cc.o(k)}, \widehat{\boldsymbol{\lambda}}_i^{co(k)}, \widehat{\tau}_i^{co(k)}, [\mathbf{v}_{1i}^c, \mathbf{v}_{2i}^c])$ ,  $\mathbf{W}_{0i}^c \sim \text{TN}_{p_i}(\widehat{\mathbf{m}}_i^{co(k)}, \widehat{\boldsymbol{\Gamma}}_i^{cc.o(k)}, [\mathbf{v}_{1i}^c, \mathbf{v}_{2i}^c])$  and

$$\widehat{\gamma}_i^{(k)} = \frac{\eta_i^{co} L_p(\mathbf{v}_{1i}^c, \mathbf{v}_{2i}^c; \widehat{\mathbf{m}}_i^{co(k)}, \widehat{\boldsymbol{\Gamma}}_i^{cc.o(k)})}{\mathcal{L}_p(\mathbf{v}_{1i}^c, \mathbf{v}_{2i}^c; \widehat{\boldsymbol{\mu}}_i^{co(k)}, \widehat{\boldsymbol{\Sigma}}_i^{cc.o(k)}, \widehat{\boldsymbol{\lambda}}_i^{co(k)}, \widehat{\tau}_i^{co(k)})},$$

where  $\lambda_i^{co} = \boldsymbol{\Sigma}_i^{cc.o^{1/2}} \boldsymbol{\varphi}_i^c$ ,  $\mathbf{m}_i^{co} = \boldsymbol{\mu}_i^{co} - \boldsymbol{\mu}_{bi}^{co}$ , and  $\eta_i^{co}$ ,  $\boldsymbol{\mu}_{bi}^{co}$  and  $\boldsymbol{\Gamma}_i^{cc.o}$  can be computed as expressions  $\eta$ ,  $\boldsymbol{\mu}_b$  and  $\boldsymbol{\Gamma}$  in Proposition 4 but using the new set of parameters  $\boldsymbol{\mu}_i^{co}$ ,  $\boldsymbol{\Sigma}_i^{cc.o}$ ,  $\boldsymbol{\lambda}_i^{co}$  and  $\tau_i^{co}$  (instead of  $\boldsymbol{\mu}$ ,  $\boldsymbol{\Sigma}$ ,  $\boldsymbol{\lambda}$  and  $\tau$ ).

To compute  $\mathbb{E}[\mathbf{W}_{0i}]$ ,  $\mathbb{E}[\mathbf{W}_i]$  and  $\mathbb{E}[\mathbf{W}_i \mathbf{W}_i^\top]$  in items 2 and 3, we use the R library `MomTrunc`.

**M-step:** Conditionally maximizing  $Q(\boldsymbol{\theta} | \widehat{\boldsymbol{\theta}}^{(k)}) = \sum_{i=1}^n Q_i(\boldsymbol{\theta} | \widehat{\boldsymbol{\theta}}^{(k)})$  with respect to each entry of  $\boldsymbol{\theta}$ , we update the estimate  $\widehat{\boldsymbol{\theta}}^{(k+1)} = (\widehat{\boldsymbol{\mu}}^{(k+1)}, \widehat{\boldsymbol{\Delta}}^{(k+1)}, \widehat{\boldsymbol{\alpha}}_\Gamma^{(k+1)})$  by

$$\widehat{\boldsymbol{\mu}}^{(k+1)} = \frac{1}{n} \sum_{i=1}^n \left\{ \widehat{\mathbf{y}}_i^{(k)} - \widehat{t}_i^{(k)} \widehat{\boldsymbol{\Delta}}^{(k)} \right\}, \quad (15)$$

$$\widehat{\boldsymbol{\Delta}}^{(k+1)} = \left\{ \sum_{i=1}^n \widehat{t}_i^{2(k)} \right\}^{-1} \sum_{i=1}^n \left\{ \widehat{t\mathbf{y}}_i^{(k)} - \widehat{t}_i^{(k)} \widehat{\boldsymbol{\mu}}^{(k+1)} \right\}, \quad (16)$$

$$\begin{aligned} \widehat{\boldsymbol{\Gamma}}^{(k+1)} &= \frac{1}{n} \sum_{i=1}^n \left\{ \widehat{\mathbf{y}}_i^{2(k)} + \widehat{\boldsymbol{\mu}}^{(k)} \widehat{\boldsymbol{\mu}}^{(k)\top} + \widehat{t}_i^{2(k)} \widehat{\boldsymbol{\Delta}}^{(k)} \widehat{\boldsymbol{\Delta}}^{(k)\top} - \widehat{\boldsymbol{\mu}}^{(k)} \widehat{\mathbf{y}}_i^{(k)\top} - \widehat{\mathbf{y}}_i^{(k)} \widehat{\boldsymbol{\mu}}^{(k)\top} \right. \\ &\quad \left. - \widehat{t\mathbf{y}}_i^{(k)} \widehat{\boldsymbol{\Delta}}^{(k)\top} - \widehat{\boldsymbol{\Delta}}^{(k)} \widehat{t\mathbf{y}}_i^{(k)\top} + \widehat{t}_i^{(k)} \widehat{\boldsymbol{\Delta}}^{(k)} \widehat{\boldsymbol{\mu}}^{(k)\top} + \widehat{t}_i^{(k)} \widehat{\boldsymbol{\mu}}^{(k)} \widehat{\boldsymbol{\Delta}}^{(k)\top} \right\}. \end{aligned} \quad (17)$$

The algorithm is iterated until a suitable convergence rule is satisfied. In the later analysis, the algorithm is terminated when the relative distance between two successive evaluations of the log-likelihood defined in (13) is less than a tolerance, i.e.,  $|\ell(\widehat{\boldsymbol{\theta}}^{(k+1)} | \mathbf{V}, \mathbf{C}) / \ell(\widehat{\boldsymbol{\theta}}^{(k)} | \mathbf{V}, \mathbf{C}) - 1| < \epsilon$ , for example,  $\epsilon = 10^{-6}$ . Once converged, we can recover  $\widehat{\boldsymbol{\lambda}}$  and  $\widehat{\boldsymbol{\Sigma}}$  using the expressions

$$\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Gamma}} + \widehat{\boldsymbol{\Delta}} \widehat{\boldsymbol{\Delta}}^\top \quad \text{and} \quad \widehat{\boldsymbol{\lambda}} = \frac{\widehat{\boldsymbol{\Sigma}}^{-1/2} \widehat{\boldsymbol{\Delta}}}{(1 - \widehat{\boldsymbol{\Delta}}^\top \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{\Delta}})^{1/2}}.$$

It is important to stress that, from Eqs. (15)-(17), the E-step reduces to the computation of  $\widehat{\mathbf{y}}_i^2, \widehat{\mathbf{y}}_i, \widehat{t}_i, \widehat{t}_i^2$  and  $\widehat{t\mathbf{y}}_i$ , for which we have implementable expressions. As pointed out for an anonymous referee, since missing values are treated as interval censored data, the computation burden relies heavily on the dimension of censored vector for evaluating the expectations of TESN and TN random vectors. In next subsection, we briefly discuss how to circumvent this problem, such that missing values do not represent neither a mathematical or computational burden.

### 3.3 Efficient computation of expectations

In the event that there are missing values, we can partition the censored vector as  $\mathbf{Y}_{cens} = (\mathbf{Y}_c^\top, \mathbf{Y}_m^\top)^\top$ , that is, as missing and (truly) censored, in order to avoid unnecessary calculation of integrals for obtaining its expectation. Considering the partition above such that  $\dim(\mathbf{Y}_c) = p_c^c, \dim(\mathbf{Y}_m) = p_m^c$ , where  $p_c^c + p_m^c = p^c$ , it follows that

$$\mathbb{E}[\mathbf{Y}_{cens}|\mathbf{Y}_{obs}] = \mathbb{E} \left[ \begin{array}{c} \mathbb{E}[\mathbf{Y}_m|\mathbf{Y}_c, \mathbf{Y}_{obs}] \\ \mathbf{Y}_c|\mathbf{Y}_{obs} \end{array} \right] \quad (18)$$

and  $\text{var}[\mathbf{Y}_{cens}|\mathbf{Y}_{obs}]$  is given by

$$\left[ \begin{array}{cc} \mathbb{E}[\text{var}[\mathbf{Y}_m|\mathbf{Y}_c, \mathbf{Y}_{obs}]] + \text{var}[\mathbb{E}[\mathbf{Y}_m|\mathbf{Y}_c, \mathbf{Y}_{obs}]] & \text{cov}[\mathbb{E}[\mathbf{Y}_m|\mathbf{Y}_c, \mathbf{Y}_{obs}], \mathbf{Y}_c|\mathbf{Y}_{obs}] \\ \text{cov}[\mathbf{Y}_c|\mathbf{Y}_{obs}, \mathbb{E}[\mathbf{Y}_m|\mathbf{Y}_c, \mathbf{Y}_{obs}]] & \text{var}[\mathbf{Y}_c|\mathbf{Y}_{obs}] \end{array} \right]. \quad (19)$$

By noting that  $\mathbf{Y}_m = (\mathbf{V} = (-\infty, \infty), \mathbf{C} = \mathbf{1})$ , we have that  $\mathbf{Y}_m|\mathbf{Y}_c, \mathbf{Y}_{obs}$  is a non-truncated partition following a ESN distribution which moments have closed forms. Then, the computation of the first two moments of  $\mathbf{Y}_{cens}|\mathbf{Y}_{obs}$  can be calculated using Eqs. (18) and (19), these last only depending on the computation of the truncated moments of  $\mathbf{Y}_c|\mathbf{Y}_{obs}$ , these are  $\mathbb{E}[\mathbf{Y}_c|\mathbf{Y}_{obs}]$  and  $\text{var}[\mathbf{Y}_c|\mathbf{Y}_{obs}]$ . As can be seen, we can use the latter equations to treat missing data as censored in a neat manner, where the truncated moments are computed only over the  $p_c^c$ -variate partition, avoiding some unnecessary integrals and saving a significant computational effort.

**Remark 1** *In general, TESN distributions are not closed under marginalization but conditioning. For instance,  $\mathbf{Y}_m|\mathbf{Y}_{obs}$  does not follow a TESN distribution but its conditional distribution  $\mathbf{Y}_m|\mathbf{Y}_c, \mathbf{Y}_{obs}$  does. Furthermore, since  $\mathbf{V} = (-\infty, \infty)$  for missing observations, we have that  $\mathbf{Y}_m|\mathbf{Y}_c, \mathbf{Y}_{obs}$  is a (conditionally) non-truncated partition, following a ESN distribution. For this particular case,  $\mathbf{Y}_c|\mathbf{Y}_{obs}$  follow a TESN distribution due to the aforementioned condition.*

## 4 The FM-MSNC model

Ignoring censoring for the moment, we consider a more general and robust framework for the multivariate response variable  $\mathbf{Y}_i$  of the model defined in (8), which is assumed to follow a mixture of multivariate skew normal distributions:

$$\mathbf{Y}_i \sim \sum_{j=1}^G \pi_j SN_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \boldsymbol{\lambda}_j), \quad (20)$$

where  $\pi_j$  are weights adding to 1 and  $G$  is the number of groups, also called components in mixture models. The mixture model considered in (20) can also be by letting  $Z_{ij}$  be a latent class variable, such that

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

Thus, given  $Z_{ij} = 1$ , the response  $\mathbf{Y}_i$  follows a multivariate skew-normal distribution

$$\mathbf{Y}_i \sim SN_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \boldsymbol{\lambda}_j), \quad i \in \{1, \dots, n\}, \quad j \in \{1, \dots, G\}. \quad (21)$$

Now, suppose  $\Pr(Z_{ij} = 1) = \pi_j$ . Then the density of  $\mathbf{Y}_i$ , without observing  $Z_{ij}$ , is given by

$$f(\mathbf{y}_i | \boldsymbol{\theta}) = \sum_{j=1}^G \pi_j SN_p(\mathbf{y}_i; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \boldsymbol{\lambda}_j), \quad (22)$$

where  $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \dots, \boldsymbol{\theta}_G^\top)^\top$ , with  $\boldsymbol{\theta}_j = (\pi_j, \boldsymbol{\mu}_j^\top, \boldsymbol{\Sigma}_j, \boldsymbol{\lambda}_j)^\top$ .

We treat the observed and censored components of  $\mathbf{Y}_i$ , separately, i.e.  $\mathbf{y}_i = (\mathbf{y}_i^o{}^\top, \mathbf{y}_i^c{}^\top)^\top$ , with respective partitioned parameters as in (10). Following [19], we define the mixture model for censored data as a mixture of the MSNC

models given in (13), viz.

$$f(\mathbf{V}_i | \mathbf{C}_i, \boldsymbol{\theta}) = \sum_{j=1}^G \pi_j f_{ij}(\mathbf{V}_i | \mathbf{C}_i, \boldsymbol{\theta}), \quad (23)$$

with

$$\begin{aligned} f_{ij}(\mathbf{V}_i | \mathbf{C}_i, \boldsymbol{\theta}) &= \mathcal{L}_{p_i^c}(\mathbf{V}_{1i}^c, \mathbf{V}_{2i}^c; \boldsymbol{\mu}_i^{co}, \boldsymbol{\Sigma}_i^{cc.o}, \boldsymbol{\Sigma}_i^{cc.o^{1/2}} \boldsymbol{\varphi}_i^c, \tau_i^{co}) \\ &\times SN_{p_i^o}(\mathbf{y}_i^o; \boldsymbol{\mu}_i^o, \boldsymbol{\Sigma}_i^{oo}, c_i^{oo} \boldsymbol{\Sigma}_i^{oo^{1/2}} \tilde{\boldsymbol{\varphi}}_i^o), \end{aligned}$$

where, for each component  $j$ , the arguments are defined as (11) and (12), respectively. The model defined in (23) will be called the FM-MSNC model. Thus, the log-likelihood function given the observed data  $(\mathbf{V}, \mathbf{C})$  is given by

$$\ell(\boldsymbol{\theta} | \mathbf{V}, \mathbf{C}) = \sum_{i=1}^n \ln\{f(\mathbf{V}_i | \mathbf{C}_i, \boldsymbol{\theta})\}.$$

#### 4.1 Maximum likelihood estimation via the EM algorithm

In this section, we present an EM algorithm for the ML estimation of the FM-MSNC model. To do so, we present the FM-MSNC model in an incomplete-data framework, using the results presented in Section 3. We recall that the likelihood associated with finite mixtures of skew-normal distributions may be unbounded, as shown by [11]. Using a straightforward extension of their argument, it can be shown that the likelihood may be unbounded in the FM-MSNC case as well. Despite this, following [32] (p. 41), we shall henceforth refer to the solution provided by the EM algorithm as the ML estimate even in situations where it may not globally maximize the likelihood.

Using the stochastic representation of the skew-normal distribution given in (14), it follows that the complete data log-likelihood function is  $\ell_c(\boldsymbol{\theta}) = \sum_{i=1}^n \ell_{ic}(\boldsymbol{\theta})$ , where, for each  $i \in \{1, \dots, n\}$ ,

$$\begin{aligned} \ell_{ic}(\boldsymbol{\theta}) &= c + \sum_{j=1}^G z_{ij} \ln \pi_j - \frac{1}{2} \sum_{j=1}^G z_{ij} \ln (|\boldsymbol{\Gamma}_j|) \\ &- \frac{1}{2} \sum_{j=1}^G z_{ij} (\mathbf{y}_i - \boldsymbol{\mu}_j - \boldsymbol{\Delta}_j t_i)^\top \boldsymbol{\Gamma}_j^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_j - \boldsymbol{\Delta}_j t_i), \end{aligned} \quad (24)$$

with  $c$  being a constant which is independent of the parameter vector  $\boldsymbol{\theta}$ .

For each  $j \in \{1, \dots, G\}$ , let  $\hat{\boldsymbol{\theta}}_j^{(k)} = (\hat{\pi}_j^{(k)}, \hat{\boldsymbol{\mu}}_j^{(k)}, \hat{\boldsymbol{\Sigma}}_j^{(k)}, \hat{\boldsymbol{\lambda}}_j^{(k)})^\top$ , and let  $\hat{\boldsymbol{\theta}}^{(k)} = (\hat{\boldsymbol{\theta}}_1^{(k)\top}, \dots,$

$\hat{\boldsymbol{\theta}}_G^{(k)\top})^\top$  be the estimate 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

$$\begin{aligned} Q(\boldsymbol{\theta} | \hat{\boldsymbol{\theta}}^{(k)}) &\propto \sum_{i=1}^n \sum_{j=1}^G \mathcal{Z}_{ij}(\hat{\boldsymbol{\theta}}^{(k)}) \ln \pi_j - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^G \mathcal{Z}_{ij}(\boldsymbol{\theta}^{(k)}) \ln (|\hat{\boldsymbol{\Gamma}}_j^{(k)}|) \\ &- \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^G \text{tr} \left[ \hat{\boldsymbol{\Gamma}}_j^{-1(k)} \left\{ \mathcal{E}_{2ij}(\hat{\boldsymbol{\theta}}^{(k)}) - \hat{\boldsymbol{\mu}}_j^{(k)} \mathcal{E}_{1ij}(\hat{\boldsymbol{\theta}}^{(k)}) - \mathcal{E}_{1ij}(\hat{\boldsymbol{\theta}}^{(k)}) \hat{\boldsymbol{\mu}}_j^{(k)\top} \right. \right. \\ &- \mathcal{E}_{3ij}(\hat{\boldsymbol{\theta}}^{(k)}) \hat{\boldsymbol{\Delta}}_j^{(k)\top} - \hat{\boldsymbol{\Delta}}_j^{(k)} \mathcal{E}_{3ij}(\hat{\boldsymbol{\theta}}^{(k)}) + \mathcal{Z}_{ij}(\hat{\boldsymbol{\theta}}^{(k)}) \hat{\boldsymbol{\mu}}_j^{(k)} \hat{\boldsymbol{\mu}}_j^{(k)\top} \\ &\left. \left. + \mathcal{E}_{4ij}(\hat{\boldsymbol{\theta}}^{(k)}) \hat{\boldsymbol{\Delta}}_j^{(k)} \hat{\boldsymbol{\Delta}}_j^{(k)\top} + \mathcal{E}_{5ij}(\hat{\boldsymbol{\theta}}^{(k)}) \hat{\boldsymbol{\Delta}}_j^{(k)} \hat{\boldsymbol{\mu}}_j^{(k)\top} + \mathcal{E}_{5ij}(\hat{\boldsymbol{\theta}}^{(k)}) \hat{\boldsymbol{\mu}}_j^{(k)} \hat{\boldsymbol{\Delta}}_j^{(k)\top} \right\} \right], \end{aligned}$$

where

$$\begin{aligned} \mathcal{E}_{1ij}(\hat{\boldsymbol{\theta}}^{(k)}) &= \mathbb{E}(Z_{ij} \mathbf{Y}_i | \mathbf{V}_i, \mathbf{C}_i, \hat{\boldsymbol{\theta}}^{(k)}), \quad \mathcal{E}_{2ij}(\hat{\boldsymbol{\theta}}^{(k)}) = \mathbb{E}(Z_{ij} \mathbf{Y}_i \mathbf{Y}_i^\top | \mathbf{V}_i, \mathbf{C}_i, \hat{\boldsymbol{\theta}}^{(k)}), \\ \mathcal{E}_{3ij}(\hat{\boldsymbol{\theta}}^{(k)}) &= \mathbb{E}(Z_{ij} T_i \mathbf{Y}_i | \mathbf{V}_i, \mathbf{C}_i, \hat{\boldsymbol{\theta}}^{(k)}), \quad \mathcal{E}_{4ij}(\hat{\boldsymbol{\theta}}^{(k)}) = \mathbb{E}(Z_{ij} T_i^2 | \mathbf{V}_i, \mathbf{C}_i, \hat{\boldsymbol{\theta}}^{(k)}), \\ \mathcal{E}_{5ij}(\hat{\boldsymbol{\theta}}^{(k)}) &= \mathbb{E}(Z_{ij} T_i | \mathbf{V}_i, \mathbf{C}_i, \hat{\boldsymbol{\theta}}^{(k)}) \quad \text{and} \quad \mathcal{Z}_{ij}(\hat{\boldsymbol{\theta}}^{(k)}) = \mathbb{E}(Z_{ij} | \mathbf{V}_i, \mathbf{C}_i, \hat{\boldsymbol{\theta}}^{(k)}). \end{aligned}$$



By using known properties of conditional expectation, we obtain

$$\mathcal{Z}_{ij}(\hat{\boldsymbol{\theta}}^{(k)}) = \frac{\hat{\pi}_j^{(k)} f_{ij}(\mathbf{V}_i | \mathbf{C}_i, \hat{\boldsymbol{\theta}}_j^{(k)})}{\sum_{j=1}^G \hat{\pi}_j^{(k)} f_{ij}(\mathbf{V}_i | \mathbf{C}_i, \hat{\boldsymbol{\theta}}_j^{(k)})}, \quad (25)$$

$$\begin{aligned} \mathcal{E}_{1ij}(\hat{\boldsymbol{\theta}}^{(k)}) &= \mathcal{Z}_{ij}(\hat{\boldsymbol{\theta}}^{(k)}) \mathbb{E}(\mathbf{Y}_i | \mathbf{V}_i, \mathbf{C}_i, \hat{\boldsymbol{\theta}}^{(k)}, Z_{ij} = 1) \\ \mathcal{E}_{2ij}(\hat{\boldsymbol{\theta}}^{(k)}) &= \mathcal{Z}_{ij}(\hat{\boldsymbol{\theta}}^{(k)}) \mathbb{E}(\mathbf{Y}_i \mathbf{Y}_i^\top | \mathbf{V}_i, \mathbf{C}_i, \hat{\boldsymbol{\theta}}^{(k)}, Z_{ij} = 1), \\ \mathcal{E}_{3ij}(\hat{\boldsymbol{\theta}}^{(k)}) &= \mathcal{Z}_{ij}(\hat{\boldsymbol{\theta}}^{(k)}) \mathbb{E}(T_i \mathbf{Y}_i | \mathbf{V}_i, \mathbf{C}_i, \hat{\boldsymbol{\theta}}^{(k)}, Z_{ij} = 1), \\ \mathcal{E}_{4ij}(\hat{\boldsymbol{\theta}}^{(k)}) &= \mathcal{Z}_{ij}(\hat{\boldsymbol{\theta}}^{(k)}) \mathbb{E}(T_i^2 | \mathbf{V}_i, \mathbf{C}_i, \hat{\boldsymbol{\theta}}^{(k)}, Z_{ij} = 1) \end{aligned} \quad (26)$$

and

$$\mathcal{E}_{5ij}(\hat{\boldsymbol{\theta}}^{(k)}) = \mathcal{Z}_{ij}(\hat{\boldsymbol{\theta}}^{(k)}) \mathbb{E}(T_i | \mathbf{V}_i, \mathbf{C}_i, \hat{\boldsymbol{\theta}}^{(k)}, Z_{ij} = 1).$$

The conditional expectations  $\mathbb{E}(\mathbf{Y}_i | \mathbf{V}_i, \mathbf{C}_i, \hat{\boldsymbol{\theta}}^{(k)}, Z_{ij} = 1)$ ,  $\mathbb{E}(\mathbf{Y}_i \mathbf{Y}_i^\top | \mathbf{V}_i, \mathbf{C}_i, \hat{\boldsymbol{\theta}}^{(k)}, Z_{ij} = 1)$ ,  $\mathbb{E}(T_i \mathbf{Y}_i | \mathbf{V}_i, \mathbf{C}_i, \hat{\boldsymbol{\theta}}^{(k)}, Z_{ij} = 1)$  and  $\mathbb{E}(T_i^2 | \mathbf{V}_i, \mathbf{C}_i, \hat{\boldsymbol{\theta}}^{(k)}, Z_{ij} = 1)$  and  $\mathbb{E}(T_i | \mathbf{V}_i, \mathbf{C}_i, \hat{\boldsymbol{\theta}}^{(k)}, Z_{ij} = 1)$  can be directly obtained from expressions  $\hat{\mathbf{y}}_i$ ,  $\hat{\mathbf{y}}_i^2$ ,  $\hat{t}_i \mathbf{y}_i$ ,  $\hat{t}_i^2$  and  $\hat{t}_i$ , respectively, given in Subsection 3.2. Thus, we have closed form expressions for all the quantities involved in the E-step of the algorithm. Next, we describe the EM algorithm for maximum likelihood estimation of the parameters in the FM-MSNC model.

**E-step:** Given  $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^{(k)}$ , compute  $\mathcal{E}_{sij}(\hat{\boldsymbol{\theta}}^{(k)})$  for all  $s \in \{1, 2, 3, 4, 5\}$  and  $\mathcal{Z}_{ij}(\hat{\boldsymbol{\theta}}^{(k)})$  for all  $i \in \{1, \dots, n\}$ ,  $j \in \{1, \dots, G\}$ .

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

$$\begin{aligned} \hat{\pi}_j^{(k+1)} &= \frac{1}{n} \sum_{i=1}^n \mathcal{Z}_{ij}(\hat{\boldsymbol{\theta}}^{(k)}), \\ \hat{\boldsymbol{\mu}}_j^{(k+1)} &= \left\{ \sum_{i=1}^n \mathcal{Z}_{ij}(\hat{\boldsymbol{\theta}}^{(k)}) \right\}^{-1} \sum_{i=1}^n \{ \mathcal{E}_{1ij}(\hat{\boldsymbol{\theta}}^{(k)}) - \mathcal{E}_{5ij}(\hat{\boldsymbol{\theta}}^{(k)}) \hat{\boldsymbol{\Delta}}_j^{(k)} \} \\ \hat{\boldsymbol{\Delta}}_j^{(k+1)} &= \left\{ \sum_{i=1}^n \mathcal{E}_{4ij}(\hat{\boldsymbol{\theta}}^{(k)}) \right\}^{-1} \sum_{i=1}^n \{ \mathcal{E}_{3ij}(\hat{\boldsymbol{\theta}}^{(k)}) - \mathcal{E}_{5ij}(\hat{\boldsymbol{\theta}}^{(k)}) \hat{\boldsymbol{\mu}}_j^{(k+1)} \} \\ \hat{\boldsymbol{\Gamma}}_j^{(k+1)} &= \left\{ \sum_{i=1}^n \mathcal{Z}_{ij}(\hat{\boldsymbol{\theta}}^{(k)}) \right\}^{-1} \sum_{i=1}^n \left\{ \mathcal{E}_{2ij}(\hat{\boldsymbol{\theta}}^{(k)}) - \hat{\boldsymbol{\mu}}_j^{(k)} \mathcal{E}_{1ij}(\hat{\boldsymbol{\theta}}^{(k)}) - \mathcal{E}_{1ij}(\hat{\boldsymbol{\theta}}^{(k)}) \hat{\boldsymbol{\mu}}_j^{(k)\top} \right. \\ &\quad - \mathcal{E}_{3ij}(\hat{\boldsymbol{\theta}}^{(k)}) \hat{\boldsymbol{\Delta}}_j^{(k)\top} - \hat{\boldsymbol{\Delta}}_j^{(k)} \mathcal{E}_{3ij}(\hat{\boldsymbol{\theta}}^{(k)\top}) + \mathcal{Z}_{ij}(\hat{\boldsymbol{\theta}}^{(k)}) \hat{\boldsymbol{\mu}}_j^{(k)} \hat{\boldsymbol{\mu}}_j^{(k)\top} \\ &\quad \left. + \mathcal{E}_{4ij}(\hat{\boldsymbol{\theta}}^{(k)}) \hat{\boldsymbol{\Delta}}_j^{(k)} \hat{\boldsymbol{\Delta}}_j^{(k)\top} + \mathcal{E}_{5ij}(\hat{\boldsymbol{\theta}}^{(k)}) \hat{\boldsymbol{\Delta}}_j^{(k)} \hat{\boldsymbol{\mu}}_j^{(k)\top} + \mathcal{E}_{5ij}(\hat{\boldsymbol{\theta}}^{(k)}) \hat{\boldsymbol{\mu}}_j^{(k)} \hat{\boldsymbol{\Delta}}_j^{(k)\top} \right\}, \end{aligned}$$

for all  $j \in \{1, \dots, G\}$ .

It is well known that mixture models can provide a multimodal log-likelihood function. In this sense, the method of maximum likelihood estimation through the EM algorithm may not give 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-MSNC model:

- (i) Partition the data (censoring levels replacing the censored observations) into  $G$  groups using the K-means clustering algorithm [11].
- (ii) Compute the proportion of data points belonging to the same cluster  $j$ , say  $\pi_j^{(0)}$ ,  $j \in \{1, \dots, G\}$ . This gives the initial value for  $\pi_j$ .
- (iii) For each group  $j$ , compute the initial values  $\boldsymbol{\mu}_j^{(0)}$ ,  $\boldsymbol{\Sigma}_j^{(0)}$ ,  $\boldsymbol{\lambda}_j^{(0)}$  using the R package `mixsmsn` [33].

## 4.2 Model selection

Because there is no universal criterion for mixture model selection, we chose three criteria to compare the models considered in this work, namely, the Akaike information criterion (AIC) [1], Bayesian information criterion (BIC) [35] and efficient determination criterion (EDC) [6]. Like the AIC and BIC, EDC has the form  $-2\ell(\hat{\boldsymbol{\theta}}) + \rho c_n$ , where  $\ell(\boldsymbol{\theta})$  is the actual log-likelihood,  $\rho$  is the number of free parameters that has 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 [8] and [11]. Note that  $c_n$  constant is given by  $c_n = 2$  for AIC and  $c_n = \log n$  for BIC, with  $n$  being the sample size.

## 4.3 Provision of standard errors

In this section, we describe how to obtain the standard errors of the ML estimates for the FM-MSNC model. We follow the information-based method exploited by [7] to compute the asymptotic covariance of the ML estimates. The empirical information matrix, according to [30]'s formula, is defined as

$$\mathbf{I}_e(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^n s(\mathbf{y}_i|\boldsymbol{\theta})s^\top(\mathbf{y}_i|\boldsymbol{\theta}) - \frac{1}{n}S(\mathbf{y}_i|\boldsymbol{\theta})S^\top(\mathbf{y}_i|\boldsymbol{\theta}), \quad (27)$$

where  $S(\mathbf{y}_i|\boldsymbol{\theta}) = \sum_{i=1}^N s(\mathbf{y}_i|\boldsymbol{\theta})$  and  $s(\mathbf{y}_i|\boldsymbol{\theta})$  is the empirical score function for the  $i$ th subject. It is noted from the result of [26] that the individual score can be determined as

$$s(\mathbf{y}_i|\boldsymbol{\theta}) = \mathbf{E} \left( \frac{\partial \ell_i(\boldsymbol{\theta}|\mathbf{y}_c)}{\partial \boldsymbol{\theta}} \middle| \mathbf{V}_i, \mathbf{C}_i, \boldsymbol{\theta} \right). \quad (28)$$

Using the ML estimates  $\hat{\boldsymbol{\theta}}$  in  $s(\mathbf{y}_i|\boldsymbol{\theta})$ , leads to  $S(\mathbf{y}_i|\hat{\boldsymbol{\theta}}) = 0$ , so from (27) we have that

$$\mathbf{I}_e(\hat{\boldsymbol{\theta}}|\mathbf{y}) = \sum_{i=1}^n \hat{\mathbf{s}}_i \hat{\mathbf{s}}_i^\top, \quad (29)$$

where  $\hat{\mathbf{s}}_i$  is an individual score vector given by  $\hat{\mathbf{s}}_i = (\hat{s}_{i,\boldsymbol{\mu}_1}, \dots, \hat{s}_{i,\boldsymbol{\mu}_G}, \hat{s}_{i,\boldsymbol{\sigma}^2_1}, \dots, \hat{s}_{i,\boldsymbol{\sigma}^2_G}, \hat{s}_{i,\boldsymbol{\lambda}_1}, \dots, \hat{s}_{i,\boldsymbol{\lambda}_G}, \hat{s}_{i,\pi_1}, \dots, \hat{s}_{i,\pi_{G-1}})^\top$ , where  $\boldsymbol{\sigma}_j^2$  is a vector with  $p(p+1)/2$  distinct elements of  $\boldsymbol{\Sigma}_j$ .

First we reparameterize  $\boldsymbol{\Sigma}_j = \mathbf{F}_j^2$  for ease of computation and theoretical derivation, where  $\mathbf{F}_j$  is the square root of  $\boldsymbol{\Sigma}_j$  containing  $p(p+1)/2$  distinct elements.

Now we have that  $\hat{\mathbf{s}}_i = (\hat{s}_{i,\boldsymbol{\mu}_1}, \dots, \hat{s}_{i,\boldsymbol{\mu}_G}, \hat{s}_{i,\boldsymbol{\alpha}_1}, \dots, \hat{s}_{i,\boldsymbol{\alpha}_G}, \hat{s}_{i,\boldsymbol{\lambda}_1}, \dots, \hat{s}_{i,\boldsymbol{\lambda}_G}, \hat{s}_{i,\pi_1}, \dots, \hat{s}_{i,\pi_{G-1}})^\top$ . So, the expressions for the elements of  $\hat{\mathbf{s}}_i$  are given by:

$$\begin{aligned} \hat{s}_{i,\pi_j} &= \frac{\mathcal{Z}_{ij}(\hat{\boldsymbol{\theta}})}{\hat{\pi}_j} - \frac{\mathcal{Z}_{ij}(\hat{\boldsymbol{\theta}})}{\hat{\pi}_G}, \\ \hat{s}_{i,\boldsymbol{\mu}_j} &= (\hat{s}_{i,\boldsymbol{\mu}_{j1}}, \dots, \hat{s}_{i,\boldsymbol{\mu}_{jp}}) = \hat{\Gamma}_j^{-1} \left( \mathcal{E}_{1ij}(\hat{\boldsymbol{\theta}}) - \mathcal{Z}_{ij}(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{\mu}}_j - \mathcal{E}_{5ij}(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{\Delta}}_j \right), \\ \hat{s}_{i,\boldsymbol{\alpha}_j} &= (\hat{s}_{i,\boldsymbol{\alpha}_{j,11}}, \dots, \hat{s}_{i,\boldsymbol{\alpha}_{j,pp}}) = -\frac{\mathcal{Z}_{ij}(\hat{\boldsymbol{\theta}})}{2} \text{tr} \left( \hat{\Gamma}_j^{-1} A_j(\hat{\boldsymbol{\theta}}) \right) - \frac{1}{2} \left\{ \mathcal{E}_{1ij}(\hat{\boldsymbol{\theta}})^\top \hat{\Gamma}_j^{-1} A_j(\hat{\boldsymbol{\theta}}) \hat{\Gamma}_j^{-1} \hat{\boldsymbol{\mu}}_j \right. \\ &\quad + \hat{\boldsymbol{\mu}}_j^\top \hat{\Gamma}_j^{-1} A_j(\hat{\boldsymbol{\theta}}) \hat{\Gamma}_j^{-1} \mathcal{E}_{1ij}(\hat{\boldsymbol{\theta}}) - \mathcal{E}_{3ij}(\hat{\boldsymbol{\theta}})^\top \left( \hat{\Gamma}_j^{-1} \dot{\mathbf{F}}_j(r) \hat{\boldsymbol{\delta}}_j - \hat{\Gamma}_j^{-1} A_j(\hat{\boldsymbol{\theta}}) \hat{\Gamma}_j^{-1} \hat{\boldsymbol{\Delta}}_j \right) \\ &\quad - \text{tr} \left( \mathcal{E}_{2ij}(\hat{\boldsymbol{\theta}}) \hat{\Gamma}_j^{-1} A_j(\hat{\boldsymbol{\theta}}) \hat{\Gamma}_j^{-1} \right) - \left( \hat{\boldsymbol{\delta}}_j^\top \dot{\mathbf{F}}_j(r) \hat{\Gamma}_j^{-1} - \hat{\boldsymbol{\Delta}}_j^\top \hat{\Gamma}_j^{-1} A_j(\hat{\boldsymbol{\theta}}) \hat{\Gamma}_j^{-1} \hat{\boldsymbol{\Delta}}_j \right) \mathcal{E}_{3ij}(\hat{\boldsymbol{\theta}}) \\ &\quad - \hat{\boldsymbol{\mu}}_j^\top \hat{\Gamma}_j^{-1} A_j(\hat{\boldsymbol{\theta}}) \hat{\Gamma}_j^{-1} \hat{\boldsymbol{\mu}}_j + \mathcal{E}_{5ij}(\hat{\boldsymbol{\theta}}) \hat{\boldsymbol{\mu}}_j^\top \left( \hat{\Gamma}_j^{-1} \dot{\mathbf{F}}_j(r) \hat{\boldsymbol{\delta}}_j - \hat{\Gamma}_j^{-1} A_j(\hat{\boldsymbol{\theta}}) \hat{\Gamma}_j^{-1} \hat{\boldsymbol{\Delta}}_j \right) \\ &\quad + \mathcal{E}_{4ij}(\hat{\boldsymbol{\theta}}) \left( \hat{\boldsymbol{\delta}}_j^\top \dot{\mathbf{F}}_j(r) \hat{\Gamma}_j^{-1} \hat{\boldsymbol{\Delta}}_j - \hat{\boldsymbol{\Delta}}_j^\top \hat{\Gamma}_j^{-1} A_j(\hat{\boldsymbol{\theta}}) \hat{\Gamma}_j^{-1} \hat{\boldsymbol{\Delta}}_j + \hat{\boldsymbol{\Delta}}_j^\top \hat{\Gamma}_j^{-1} \dot{\mathbf{F}}_j(r) \hat{\boldsymbol{\delta}}_j \right) \\ &\quad \left. + \left( \hat{\boldsymbol{\delta}}_j^\top \dot{\mathbf{F}}_j(r) \hat{\Gamma}_j^{-1} - \hat{\boldsymbol{\Delta}}_j^\top \hat{\Gamma}_j^{-1} A_j(\hat{\boldsymbol{\theta}}) \hat{\Gamma}_j^{-1} \right) \hat{\boldsymbol{\mu}}_j \mathcal{E}_{5ij}(\hat{\boldsymbol{\theta}}) \right\}, \end{aligned}$$

$$\begin{aligned}
\widehat{s}_{i,\lambda_j} &= (\widehat{s}_{i,\lambda_{j1}}, \dots, \widehat{s}_{i,\lambda_{jp}}) = \frac{Z_{ij}(\widehat{\theta})}{2} \text{tr} \left( \widehat{\Gamma}_j^{-1} B_j(\widehat{\theta}) \right) - \frac{1}{2} \left\{ \text{tr} \left( \mathcal{E}_{2ij}(\widehat{\theta}) \widehat{\Gamma}_j^{-1} B_j(\widehat{\theta}) \widehat{\Gamma}_j^{-1} \right) \right. \\
&- \mathcal{E}_{1ij}(\widehat{\theta})^\top \widehat{\Gamma}_j^{-1} B_j(\widehat{\theta}) \widehat{\Gamma}_j^{-1} \widehat{\mu}_j + \widehat{\mu}_j^\top \widehat{\Gamma}_j^{-1} B_j(\widehat{\theta}) \widehat{\Gamma}_j^{-1} \widehat{\mu}_j - \widehat{\mu}_j^\top \widehat{\Gamma}_j^{-1} B_j(\widehat{\theta}) \widehat{\Gamma}_j^{-1} \mathcal{E}_{1ij}(\widehat{\theta}) \\
&- \mathcal{E}_{3ij}(\widehat{\theta})^\top \left( \widehat{\Gamma}_j^{-1} B_j(\widehat{\theta}) \widehat{\Gamma}_j^{-1} \widehat{\Delta}_j + \widehat{\Gamma}_j^{-1} b_j(\widehat{\theta}) \right) \\
&- \left( \widehat{\Delta}_j^\top \widehat{\Gamma}_j^{-1} B_j(\widehat{\theta}) \widehat{\Gamma}_j^{-1} + b_j(\widehat{\theta})^\top \widehat{\Gamma}_j^{-1} \right) \mathcal{E}_{3ij}(\widehat{\theta}) \\
&+ \mathcal{E}_{5ij}(\widehat{\theta}) \widehat{\mu}_j^\top \left( \widehat{\Gamma}_j^{-1} B_j(\widehat{\theta}) \widehat{\Gamma}_j^{-1} \widehat{\Delta}_j + \widehat{\Gamma}_j^{-1} b_j(\widehat{\theta}) \right) \\
&+ \left( \widehat{\Delta}_j^\top \widehat{\Gamma}_j^{-1} B_j(\widehat{\theta}) \widehat{\Gamma}_j^{-1} + b_j(\widehat{\theta})^\top \widehat{\Gamma}_j^{-1} \right) \widehat{\mu}_j \mathcal{E}_{5ij}(\widehat{\theta}) \\
&\left. + \mathcal{E}_{4ij}(\widehat{\theta}) \left( b_j(\widehat{\theta})^\top \widehat{\Gamma}_j^{-1} \widehat{\Delta}_j + \widehat{\Delta}_j^\top \widehat{\Gamma}_j^{-1} B_j(\widehat{\theta}) \widehat{\Gamma}_j^{-1} \widehat{\Delta}_j + \widehat{\Delta}_j^\top \widehat{\Gamma}_j^{-1} b_j(\widehat{\theta}) \right) \right\},
\end{aligned}$$

where

$$\begin{aligned}
A_j(\widehat{\theta}) &= \left( \dot{\mathbf{F}}_j(r) (\mathbf{I} - \widehat{\delta}_j \widehat{\delta}_j^\top) \widehat{\mathbf{F}}_j + \widehat{\mathbf{F}}_j (\mathbf{I} - \widehat{\delta}_j \widehat{\delta}_j^\top) \dot{\mathbf{F}}_j(r) \right), \\
B_j(\widehat{\theta}) &= \widehat{\mathbf{F}}_j \left( \frac{\dot{\mathbf{R}}_j(r) (1 + \widehat{\lambda}_j^\top \widehat{\lambda}_j) - 2\lambda_{jr} \widehat{\lambda}_j \widehat{\lambda}_j^\top}{(1 + \widehat{\lambda}_j^\top \widehat{\lambda}_j)^2} \right) \widehat{\mathbf{F}}_j, \\
b_j(\widehat{\theta}) &= \widehat{\mathbf{F}}_j \left( \frac{\dot{\lambda}_j(r) (1 + \widehat{\lambda}_j^\top \widehat{\lambda}_j) - \lambda_{jr} \widehat{\lambda}_j}{(1 + \widehat{\lambda}_j^\top \widehat{\lambda}_j)^{3/2}} \right),
\end{aligned}$$

$$\dot{\mathbf{F}}_j(r) = \frac{\partial \mathbf{F}_j}{\partial \sigma_{jr}^2} \Big|_{\sigma^2 = \widehat{\sigma}^2}, \quad \dot{\mathbf{R}}_j(r) = \frac{\partial \lambda_j \lambda_j^\top}{\partial \lambda_{jr}} \Big|_{\lambda = \widehat{\lambda}}, \quad \text{and} \quad \dot{\lambda}_j(r) = \frac{\partial \lambda_j}{\partial \lambda_{jr}} \Big|_{\lambda = \widehat{\lambda}}, \quad \text{with } r = 1, 2, \dots, p.$$

## 5 Simulation studies

In order to study the performance of our proposed method, we present five simulation studies. The first and second study investigates whether we can estimate the true parameter values and their respective standard errors accurately by using the proposed EM algorithm and approximated empirical information matrix, respectively involving censoring and missing data. The third one investigates the number of mixture components by comparing the FM-MSNC with two groups and FM-MNC with various groups. The fourth study investigates the ability of the FM-MSNC model to cluster observations. Finally, the last one shows the asymptotic behavior of the EM estimates for the proposed model. The computations were done using the R package *CensMFM*.

### 5.1 Performance of the ML Estimates over censoring data

This simulation study is designed to verify if we can estimate the true parameter values of the FM-MSNC model accurately when we have censoring data by using the proposed EM algorithm. We simulated several datasets considering mixtures with two components from model (23) with two left-censoring proportion settings (5% and 30%), taken in each mixture component, and different samples sizes  $n \in \{500, 1000, 2000\}$ . For each combination, we generated 500 Monte Carlo (MC) samples. Summary statistics of the estimates across the 500 MC samples were computed, such as the mean estimate (MC mean), the empirical standard error (MC Sd), and the mean of the approximate standard errors of the estimates, obtained through the method described in Section 4.3 (IM SE).

We consider small and different variances with the following parameter setup:

$$0.65 SN_2 \left( \left[ \begin{array}{c} -3 \\ -4 \end{array} \right], \left[ \begin{array}{cc} 3 & 1 \\ 1 & 4.5 \end{array} \right], \left[ \begin{array}{c} -2 \\ 2 \end{array} \right] \right) + 0.35 SN_2 \left( \left[ \begin{array}{c} 2 \\ 2 \end{array} \right], \left[ \begin{array}{cc} 2 & 1 \\ 1 & 3.5 \end{array} \right], \left[ \begin{array}{c} -3 \\ 4 \end{array} \right] \right).$$

Figure 1 shows the simulated data from the FM-MSNC model with their respective density contours for the skew-normal (top panel) and normal (bottom panel) distributions and the allocations in each group for samples sizes 500, 1000 and 2000 with left-censoring proportion of 5%. The black points represent the first component and the red triangles represent the second component of the mixture. One can note that the contour lines of the skew-normal distribution are more appropriate to represent the shape assumed by the generated data.

The results are presented in Table 1. This table shows that, regardless the sample size, the Monte Carlo mean of the parameter estimates deviates further the true values as the censoring level increases, i.e., the parameter estimates are

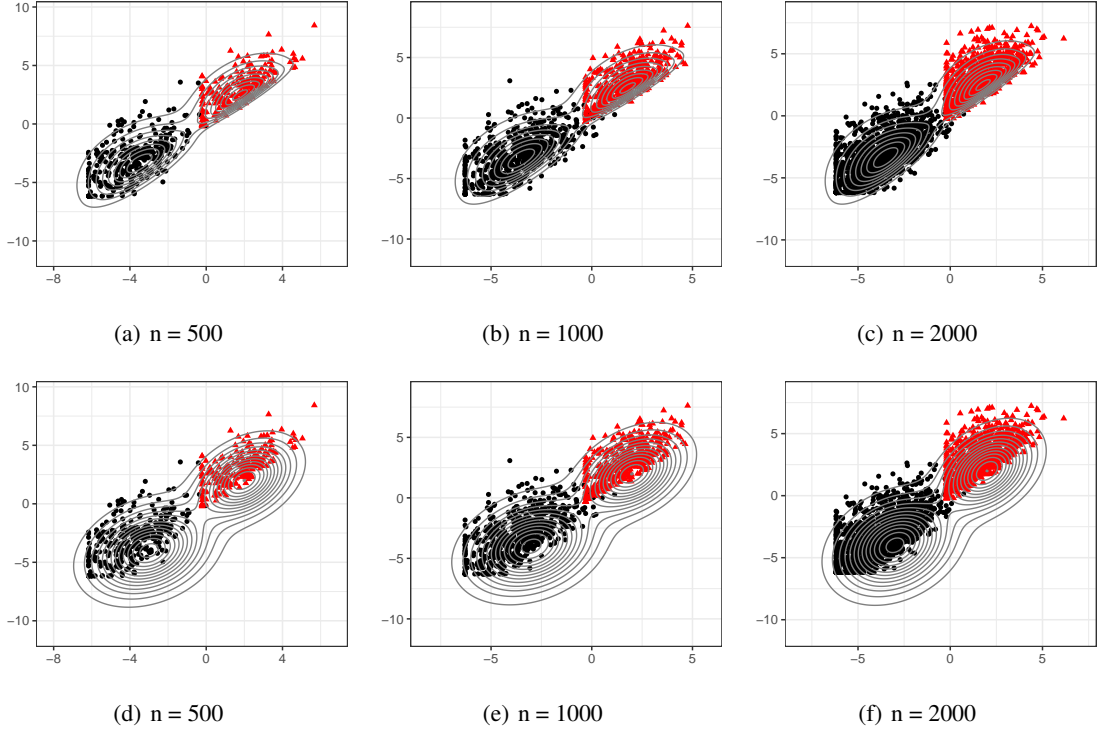


Figure 1: Simulated data: Performance of the ML Estimates over censoring data. Scatter plot for some simulated data from FM-MSNC model with the respective density contours, skew-normal (top panel) and normal (bottom panel) with 5% censoring level.

affected by the censoring level. In particular, the estimates of  $\mu_1$  and  $\mu_2$  appear to be less affected by increasing the censoring level than the other parameters. Furthermore, the estimates of the standard errors, i.e., MC Sd and IM SE, provide relatively close results, which may indicate that the asymptotic approach proposed for the standard errors of the ML estimates is reliable.

Table 1: Simulated data: Performance of the ML Estimates over censoring data. Parameter estimates based on 500 simulated samples. Monte Carlo (MC) mean, MC Sd are the respective mean estimates and standard deviations. IM SE is the average value of the approximate standard error obtained through the information-based method.

| Censoring  | $j$ | Measure | Parameter  |            |                 |                 |                 |                |                |        |
|------------|-----|---------|------------|------------|-----------------|-----------------|-----------------|----------------|----------------|--------|
|            |     |         | $\mu_{j1}$ | $\mu_{j2}$ | $\alpha_{j,11}$ | $\alpha_{j,12}$ | $\alpha_{j,22}$ | $\lambda_{j1}$ | $\lambda_{j2}$ | $\pi$  |
| $n = 500$  |     |         |            |            |                 |                 |                 |                |                |        |
| 5%         | 1   | True    | (-3)       | (-4)       | (1.7121)        | (0.2620)        | (2.1051)        | (-2)           | (2)            | (0.65) |
|            |     | MC mean | -3.1797    | -4.1235    | 1.6262          | 0.2923          | 2.2499          | -1.7236        | 2.1533         | 0.6607 |
|            |     | MC Sd   | 0.3101     | 0.3705     | 0.1510          | 0.1138          | 0.1906          | 0.5262         | 0.5739         | 0.0274 |
|            | 2   | IM SE   | 0.2844     | 0.3406     | 0.1361          | 0.0924          | 0.1944          | 0.5397         | 0.6202         | 0.0236 |
|            |     | True    | (2)        | (2)        | (1.3798)        | (0.3101)        | (1.8449)        | (-3)           | (4)            |        |
|            |     | MC mean | 2.0196     | 2.0954     | 1.3152          | 0.3177          | 1.7883          | -2.7361        | 3.7058         |        |
| 30%        | 1   | MC Sd   | 0.2381     | 0.2808     | 0.1431          | 0.0928          | 0.1595          | 1.0896         | 1.3303         |        |
|            |     | IM SE   | 0.2677     | 0.2922     | 0.1391          | 0.0981          | 0.1816          | 1.2411         | 1.5434         |        |
|            |     | True    | (-3)       | (-4)       | (1.7121)        | (0.2620)        | (2.1051)        | (-2)           | (2)            | (0.65) |
|            | 2   | MC mean | -3.3139    | -4.1708    | 1.5445          | 0.3483          | 2.2909          | -1.3744        | 2.0529         | 0.7030 |
|            |     | MC Sd   | 0.3360     | 0.4545     | 0.1580          | 0.1413          | 0.2700          | 0.5082         | 0.7388         | 0.0238 |
|            |     | IM SE   | 0.4719     | 0.5176     | 0.2080          | 0.1396          | 0.2762          | 0.7304         | 0.8737         | 0.0238 |
| $n = 1000$ |     |         |            |            |                 |                 |                 |                |                |        |
| 5%         | 1   | True    | (2)        | (2)        | (1.3798)        | (0.3101)        | (1.8449)        | (-3)           | (4)            |        |
|            |     | MC mean | 2.2651     | 2.5165     | 1.3373          | 0.2533          | 1.6079          | -2.3831        | 3.1945         |        |
|            |     | MC Sd   | 0.5131     | 0.5107     | 0.2661          | 0.1682          | 0.2207          | 1.5120         | 1.4642         |        |
|            | 2   | IM SE   | 0.3213     | 0.3134     | 0.1929          | 0.1202          | 0.1726          | 1.4296         | 1.4575         |        |
|            |     | True    | (-3)       | (-4)       | (1.7121)        | (0.2620)        | (2.1051)        | (-2)           | (2)            | (0.65) |
|            |     | MC mean | -3.1812    | -4.1450    | 1.6219          | 0.2804          | 2.2407          | -1.6988        | 2.1169         | 0.6594 |
| 30%        | 1   | MC Sd   | 0.1876     | 0.2248     | 0.0911          | 0.0615          | 0.1354          | 0.3611         | 0.4228         | 0.0165 |
|            |     | IM SE   | 0.1988     | 0.2351     | 0.0962          | 0.0607          | 0.1369          | 0.3660         | 0.4177         | 0.0167 |
|            |     | True    | (2)        | (2)        | (1.3798)        | (0.3101)        | (1.8449)        | (-3)           | (4)            |        |
|            | 2   | MC mean | 2.2651     | 2.5165     | 1.3373          | 0.2533          | 1.6079          | -2.3831        | 3.1945         |        |
|            |     | MC Sd   | 0.5131     | 0.5107     | 0.2661          | 0.1682          | 0.2207          | 1.5120         | 1.4642         |        |
|            |     | IM SE   | 0.3213     | 0.3134     | 0.1929          | 0.1202          | 0.1726          | 1.4296         | 1.4575         |        |

Continued on next page

Table 1 – Continued from previous page

| Censored | j  | Measure | Parameter  |            |                 |                 |                 |                |                |        |        |
|----------|----|---------|------------|------------|-----------------|-----------------|-----------------|----------------|----------------|--------|--------|
|          |    |         | $\mu_{j1}$ | $\mu_{j2}$ | $\alpha_{j,11}$ | $\alpha_{j,12}$ | $\alpha_{j,22}$ | $\lambda_{j1}$ | $\lambda_{j2}$ | $\pi$  |        |
|          |    |         | $n = 2000$ |            |                 |                 |                 |                |                |        |        |
| 30%      | 1  | True    | (-3)       | (-4)       | (1.7121)        | (0.2620)        | (2.1051)        | (-2)           | (2)            | (0.65) |        |
|          |    | MC mean | -3.3168    | -4.2133    | 1.5402          | 0.3276          | 2.2758          | -1.3758        | 2.0334         | 0.7021 |        |
|          |    | MC Sd   | 0.2029     | 0.2736     | 0.0978          | 0.0776          | 0.1758          | 0.3513         | 0.5272         | 0.0162 |        |
|          | 1  | IM SE   | 0.3071     | 0.3289     | 0.1374          | 0.0878          | 0.1840          | 0.4791         | 0.5681         | 0.0160 |        |
|          |    | True    | (2)        | (2)        | (1.3798)        | (0.3101)        | (1.8449)        | (-3)           | (4)            |        |        |
|          |    | MC mean | 2.3043     | 2.5243     | 1.3347          | 0.2208          | 1.5994          | -2.3618        | 3.0762         |        |        |
|          | 2  | MC Sd   | 0.3216     | 0.2956     | 0.1899          | 0.1000          | 0.1513          | 1.0722         | 0.8913         |        |        |
|          |    | IM SE   | 0.2144     | 0.2025     | 0.1295          | 0.0743          | 0.1172          | 0.8982         | 0.8670         |        |        |
|          |    | True    | (-3)       | (-4)       | (1.7121)        | (0.2620)        | (2.1051)        | (-2)           | (2)            | (0.65) |        |
|          | 5% | 1       | MC mean    | -3.2080    | -4.1723         | 1.6098          | 0.2779          | 2.2581         | -1.6429        | 2.1188 | 0.6600 |
|          |    |         | MC Sd      | 0.1577     | 0.1728          | 0.0678          | 0.0525          | 0.1010         | 0.2864         | 0.3225 | 0.0118 |
|          |    |         | IM SE      | 0.1411     | 0.1681          | 0.0672          | 0.0429          | 0.0978         | 0.2515         | 0.2926 | 0.0118 |
| 2        |    | True    | (2)        | (2)        | (1.3798)        | (0.3101)        | (1.8449)        | (-3)           | (4)            |        |        |
|          |    | MC mean | 2.0480     | 2.1143     | 1.3204          | 0.3059          | 1.7857          | -2.5958        | 3.4666         |        |        |
|          |    | MC Sd   | 0.1283     | 0.1995     | 0.0642          | 0.0553          | 0.0845          | 0.5176         | 0.6986         |        |        |
| 2        |    | IM SE   | 0.2882     | 0.3822     | 0.0739          | 0.0501          | 0.0888          | 0.6434         | 0.8032         |        |        |
|          |    | True    | (-3)       | (-4)       | (1.7121)        | (0.2620)        | (2.1051)        | (-2)           | (2)            | (0.65) |        |
|          |    | MC mean | -3.3559    | -4.2548    | 1.4775          | 0.3046          | 2.1378          | -1.3237        | 2.0660         | 0.7036 |        |
| 30%      |    | 1       | MC Sd      | 0.1785     | 0.2527          | 0.1310          | 0.0859          | 0.3274         | 0.2934         | 0.4413 | 0.0109 |
|          |    |         | IM SE      | 0.2138     | 0.2299          | 0.0921          | 0.0641          | 0.1320         | 0.3289         | 0.4023 | 0.0113 |
|          |    |         | True       | (2)        | (2)             | (1.3798)        | (0.3101)        | (1.8449)       | (-3)           | (4)    |        |
|          | 2  | MC mean | 2.2991     | 2.5248     | 1.3155          | 0.2141          | 1.6032          | -2.2326        | 2.9887         |        |        |
|          |    | MC Sd   | 0.2436     | 0.2913     | 0.1381          | 0.0824          | 0.1245          | 0.7816         | 0.7760         |        |        |
|          |    | IM SE   | 0.1601     | 0.1569     | 0.0922          | 0.0542          | 0.0847          | 0.5911         | 0.5862         |        |        |

## 5.2 Performance of the ML Estimates over missing data

In order to evaluate the performance of FM-MSNC model for dealing with partially incomplete data, a simulation study was conducted. Various ways of using models for imputation are described in [25], among them, one of the most relevant is the missing completely at random (MCAR). We simulated several datasets considering mixtures with two components from model (23) with two missing data proportion settings (5% and 20%), taken in each mixture component, and different samples sizes  $n \in (500, 700, 900)$ . For each combination, we generated 500 Monte Carlo (MC) samples. Summary statistics of the estimates across the 500 MC samples were computed, such as the mean estimate (MC mean), the empirical standard error (MC Sd), and the mean of the approximate standard errors of the estimates, obtained through the method described in Section 4.3 (IM SE). We consider small and different variances with the following parameter as in the simulation about asymptotic properties in 5.5.

Table 2 shows the results for this simulation. The results obtained are similar to those of simulation 5.1 and the same conclusions can be drawn. Additionally, we note that  $\lambda$  estimates appear to be more strongly affected as we increase the proportion of missing data in the sample.

Table 2: Simulated data: Performance of the ML Estimates over missing data. Parameter estimates based on 500 simulated samples. Monte Carlo (MC) mean, MC Sd are the respective mean estimates and standard deviations. IM SE is the average value of the approximate standard error obtained through the information-based method.

| Missing | j   | Measure | Parameter  |            |                 |                 |                 |                |                |        |        |
|---------|-----|---------|------------|------------|-----------------|-----------------|-----------------|----------------|----------------|--------|--------|
|         |     |         | $\mu_{j1}$ | $\mu_{j2}$ | $\alpha_{j,11}$ | $\alpha_{j,12}$ | $\alpha_{j,22}$ | $\lambda_{j1}$ | $\lambda_{j2}$ | $\pi$  |        |
|         |     |         | $n = 500$  |            |                 |                 |                 |                |                |        |        |
| 5%      | 1   | True    | (-5)       | (-4)       | (1.7121)        | (0.2620)        | (2.1051)        | (-2)           | (3)            | (0.65) |        |
|         |     | MC mean | -5.2150    | -3.3411    | 1.6844          | 0.4594          | 1.9442          | -1.0396        | 1.4896         | 0.6495 |        |
|         |     | MC Sd   | 0.5773     | 0.8843     | 0.1075          | 0.2005          | 0.2071          | 1.0020         | 1.5677         | 0.0219 |        |
|         | 1   | IM SE   | 0.8618     | 0.9443     | 0.2666          | 0.1876          | 0.2603          | 0.9352         | 1.0664         | 0.0220 |        |
|         |     | True    | (2)        | (3)        | (1.3798)        | (0.3101)        | (1.8449)        | (-2)           | (3)            |        |        |
|         |     | MC mean | 1.7699     | 3.4694     | 1.3709          | 0.4857          | 1.7308          | -0.9892        | 1.5876         |        |        |
|         | 2   | MC Sd   | 0.5614     | 0.7703     | 0.1231          | 0.1904          | 0.1922          | 1.1875         | 1.6737         |        |        |
|         |     | IM SE   | 0.8408     | 0.9545     | 0.2759          | 0.2176          | 0.3247          | 1.2244         | 1.3864         |        |        |
|         |     | True    | (-5)       | (-4)       | (1.7121)        | (0.2620)        | (2.1051)        | (-2)           | (3)            | (0.65) |        |
|         | 20% | 1       | MC mean    | -5.2797    | -3.2378         | 1.6751          | 0.4903          | 1.9116         | -0.8182        | 1.1958 | 0.6495 |
|         |     |         | MC Sd      | 0.6079     | 0.9007          | 0.1120          | 0.1908          | 0.2215         | 0.9494         | 1.5060 | 0.0228 |
|         |     |         | IM SE      | 1.0618     | 1.1557          | 0.3267          | 0.2297          | 0.3216         | 1.1195         | 1.2620 | 0.0227 |
| 1       |     | True    | (2)        | (3)        | (1.3798)        | (0.3101)        | (1.8449)        | (-2)           | (3)            |        |        |

Continued on next page

Table 2 – Continued from previous page

| Missing   | j | Measure | Parameter  |            |                 |                 |                 |                |                |        |
|-----------|---|---------|------------|------------|-----------------|-----------------|-----------------|----------------|----------------|--------|
|           |   |         | $\mu_{j1}$ | $\mu_{j2}$ | $\alpha_{j,11}$ | $\alpha_{j,12}$ | $\alpha_{j,22}$ | $\lambda_{j1}$ | $\lambda_{j2}$ | $\pi$  |
|           |   | MC mean | 1.6992     | 3.5252     | 1.3680          | 0.5125          | 1.7031          | -0.7323        | 1.3092         |        |
|           |   | MC Sd   | 0.5987     | 0.7648     | 0.1359          | 0.1924          | 0.2047          | 1.1693         | 1.5322         |        |
|           |   | IM SE   | 1.2860     | 1.4228     | 0.3530          | 0.2721          | 0.4106          | 1.6672         | 1.7740         |        |
| $n = 700$ |   |         |            |            |                 |                 |                 |                |                |        |
| 5%        | 1 | True    | (-5)       | (-4)       | (1.7121)        | (0.2620)        | (2.1051)        | (-2)           | (3)            | (0.65) |
|           |   | MC mean | -5.1275    | -3.2608    | 1.6902          | 0.4589          | 1.9299          | -1.0984        | 1.4147         | 0.6493 |
|           |   | MC Sd   | 0.5118     | 0.9054     | 0.0931          | 0.2000          | 0.1867          | 0.9515         | 1.5788         | 0.0188 |
|           | 2 | IM SE   | 0.6699     | 0.7212     | 0.2189          | 0.1539          | 0.2158          | 0.7571         | 0.8440         | 0.0186 |
|           |   | True    | (2)        | (3)        | (1.3798)        | (0.3101)        | (1.8449)        | (-2)           | (3)            |        |
|           |   | MC mean | 1.8256     | 3.4568     | 1.3733          | 0.4711          | 1.7356          | -1.1133        | 1.6816         |        |
| 20%       | 1 | MC Sd   | 0.4959     | 0.7484     | 0.1053          | 0.1725          | 0.1741          | 1.0682         | 1.5869         |        |
|           |   | IM SE   | 0.6973     | 0.8268     | 0.2228          | 0.1766          | 0.2621          | 1.0000         | 1.1773         |        |
|           |   | True    | (-5)       | (-4)       | (1.7121)        | (0.2620)        | (2.1051)        | (-2)           | (3)            | (0.65) |
| 5%        | 1 | MC mean | -5.1948    | -3.2079    | 1.6799          | 0.4817          | 1.8994          | -0.9175        | 1.1938         | 0.6488 |
|           |   | MC Sd   | 0.5467     | 0.8905     | 0.0937          | 0.1907          | 0.1882          | 0.9160         | 1.4792         | 0.0192 |
|           |   | IM SE   | 0.8447     | 0.8977     | 0.2686          | 0.1866          | 0.2680          | 0.9194         | 1.0006         | 0.0192 |
|           | 2 | True    | (2)        | (3)        | (1.3798)        | (0.3101)        | (1.8449)        | (-2)           | (3)            |        |
|           |   | MC mean | 1.7787     | 3.5349     | 1.3666          | 0.4987          | 1.7052          | -0.8822        | 1.3254         |        |
|           |   | MC Sd   | 0.5295     | 0.7466     | 0.1172          | 0.1741          | 0.1832          | 1.0146         | 1.4570         |        |
| 20%       | 1 | IM SE   | 0.9450     | 1.0692     | 0.2765          | 0.2118          | 0.3275          | 1.2557         | 1.3766         |        |
|           |   | True    | (-5)       | (-4)       | (1.7121)        | (0.2620)        | (2.1051)        | (-2)           | (3)            | (0.65) |
|           |   | MC mean | -5.1377    | -3.4011    | 1.6847          | 0.4243          | 1.9552          | -1.1829        | 1.6348         | 0.6488 |
| 5%        | 1 | MC Sd   | 0.4678     | 0.8574     | 0.0794          | 0.1869          | 0.1790          | 0.8858         | 1.4904         | 0.0169 |
|           |   | IM SE   | 0.4769     | 0.4857     | 0.1699          | 0.1183          | 0.1781          | 0.5865         | 0.6471         | 0.0163 |
|           |   | True    | (2)        | (3)        | (1.3798)        | (0.3101)        | (1.8449)        | (-2)           | (3)            |        |
|           | 2 | MC mean | 1.8644     | 3.3518     | 1.3744          | 0.4342          | 1.7513          | -1.2648        | 1.9203         |        |
|           |   | MC Sd   | 0.4578     | 0.6741     | 0.0891          | 0.1616          | 0.1540          | 0.9946         | 1.4206         |        |
|           |   | IM SE   | 0.5043     | 0.5860     | 0.1656          | 0.1297          | 0.1999          | 0.7796         | 0.9222         |        |
| 20%       | 1 | True    | (-5)       | (-4)       | (1.7121)        | (0.2620)        | (2.1051)        | (-2)           | (3)            | (0.65) |
|           |   | MC mean | -5.2002    | -3.3434    | 1.6760          | 0.4495          | 1.9244          | -0.9997        | 1.4014         | 0.6485 |
|           |   | MC Sd   | 0.5123     | 0.8387     | 0.0839          | 0.1816          | 0.1866          | 0.8815         | 1.4234         | 0.0175 |
|           | 2 | IM SE   | 0.6459     | 0.6892     | 0.2138          | 0.1502          | 0.2198          | 0.7351         | 0.8213         | 0.0168 |
|           |   | True    | (2)        | (3)        | (1.3798)        | (0.3101)        | (1.8449)        | (-2)           | (3)            |        |
|           |   | MC mean | 1.7946     | 3.4393     | 1.3699          | 0.4709          | 1.7147          | -0.9902        | 1.5475         |        |
| 20%       | 1 | MC Sd   | 0.5079     | 0.6797     | 0.0986          | 0.1640          | 0.1650          | 1.0069         | 1.3840         |        |
|           |   | IM SE   | 0.7262     | 0.7971     | 0.2178          | 0.1622          | 0.2598          | 1.0105         | 1.0992         |        |
|           |   | True    | (-5)       | (-4)       | (1.7121)        | (0.2620)        | (2.1051)        | (-2)           | (3)            | (0.65) |

To exemplify the predictive accuracies on the imputation of missing values, we compare the FM-MSNC with the traditional randomization-based mean imputation (MI) predictor [25], known as a common heuristic by filling in a single value for each missing value with the observed sample mean of the associated attribute. As a measure of precision, we use the mean absolute error (MAE) and the mean absolute relative error (MARE). They are defined as

$$MAE = \frac{1}{m} \sum_{i=1}^n \sum_{j=1}^g |y_{ij} - \hat{y}_{ij}| \quad \text{and} \quad MARE = \frac{1}{m} \sum_{i=1}^n \sum_{j=1}^g \left| \frac{y_{ij} - \hat{y}_{ij}}{y_{ij}} \right|, \quad (30)$$

where  $m$  is the number of missing entries,  $y_{ij}$  is the actual value and  $\hat{y}_{ij}$  is the respective predictive value. The MAE and MARE measures, for both FM-MSNC and MI method, are listed in Table 3. We can see that the FM-MSNC predictor exhibits considerable promising accuracy in the prediction of missing values when compared with those of MI imputations for all cases.

Table 3: Simulated data: Performance of the ML Estimates over missing data. Average prediction accuracies for the both imputation methods FM-MSNC and mean imputation (MI) with varying sample size,  $n \in (500, 700, 900)$ , and proportions of missing values 5%, 10% and 20%.

| Imputation method | Missing rate(%) | MAE    |        |        | MARE   |        |        |
|-------------------|-----------------|--------|--------|--------|--------|--------|--------|
|                   |                 | 500    | 700    | 900    | 500    | 700    | 900    |
| FM-MSNC           | 5%              | 1.9009 | 1.8556 | 1.8444 | 0.8642 | 0.8747 | 0.8252 |
|                   | 10%             | 2.0681 | 2.0243 | 2.0025 | 0.9883 | 1.0274 | 0.9678 |
|                   | 20%             | 2.4074 | 2.3693 | 2.3605 | 1.2264 | 1.2386 | 1.2345 |
| MI                | 5%              | 3.2550 | 3.0297 | 2.7708 | 1.6897 | 1.8542 | 1.5337 |
|                   | 10%             | 3.2594 | 3.0498 | 2.7717 | 1.7402 | 1.8841 | 1.5501 |
|                   | 20%             | 3.2567 | 3.0563 | 2.7733 | 1.9881 | 1.8875 | 1.6624 |

### 5.3 Number of mixture components

In this section, we compare the ability of some classic model selection criteria discussed in Subsection 4.2 to select the appropriate model. One may argue that an arbitrary multivariate density can always be approximated by a finite mixture of normal multivariate distributions, see [31, Chapter 1], for example. Thus, an interesting comparison can be made if we consider a sample from a two-component FM-MSNC(2) and use some model choice criteria to compare this model with the FM-MNC and several components under different censoring levels. Here we consider 100 samples of size 500 from a two-component FM-MSNC(2) model with left censoring levels at 5%, 10% or 20%, and parameter values set at

$$0.65 SN_2 \left( \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \begin{bmatrix} 1.5 & 0 \\ 0 & 1.5 \end{bmatrix}, \begin{bmatrix} -5 \\ 10 \end{bmatrix} \right) + 0.35 SN_2 \left( \begin{bmatrix} -2 \\ -1 \end{bmatrix}, \begin{bmatrix} 1.5 & 0 \\ 0 & 1.5 \end{bmatrix}, \begin{bmatrix} -5 \\ 10 \end{bmatrix} \right).$$

The results are presented in Table 4, under different censoring levels, where it can be seen that all criteria favor the true model, that is, the FM-MSNC(2) model instead the FM-MNC model with two, three and four components, as expected. This is evidence that these measures are capable of detecting departures from normality. It is important to emphasize that the FM-MNC models with three and four components have 17 and 23 parameters respectively, while the FM-MSNC(2) model has 11 parameters.

Table 4: Simulated data: Number of mixture components. Percentage when the FM-MSNC model with two components is preferred over the other adjusted FM-MNC models.

|          |     | Censoring |     |     | 5% |     |     | 10% |     |     | 20% |     |   |
|----------|-----|-----------|-----|-----|----|-----|-----|-----|-----|-----|-----|-----|---|
|          |     | Group     |     |     | 2  | 3   | 4   | 2   | 3   | 4   | 2   | 3   | 4 |
| Criteria | AIC | 100       | 100 | 100 | 95 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |   |
|          | BIC | 100       | 100 | 100 | 93 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |   |
|          | EDC | 100       | 100 | 100 | 93 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |   |

As pointed out for an anonymous referee, we can see in Table 4 that the case with 10% censoring and two components is the only case where the correct model was not preferred by model selection criteria for a small number of instances. According to Figure 2, the preferred model in these (atypical) cases was the FM-MNC(2), however the differences in the criteria values related to a FM-MSNC(2) are close to zero. We believe that the amount of data sets generated in the simulation (100 data sets) may not be sufficient and a more intensive simulation study would be required. However, due to the computational burden of the simulation, it would be too time consuming to move over to bigger simulations, say, 1000 data sets.

### 5.4 Clustering

Mixture models in general can be used for two main purposes: 1. estimation, and 2. model-based clustering [28]. In this section, we investigate the ability of the FM-MSNC model to cluster observations, that is, to allocate them into groups of observations that are similar in some sense. We know that each data point belongs to  $g$  heterogeneous populations, but we don't know how to discriminate between them. Fitting the data with mixture models allows clustering the data in terms of the estimated posterior probability that a single point belongs to a given group. For this purpose, we follow the method proposed by [38], to assess the quality of the clustering of each mixture model using an index measure called correct classification rate (CCR), which is based on the posterior assigned to each subject. For the investigation of the clustering ability of the FM-MSNC model, we simulated 500 MC samples considering mixtures with two components from model (22), with sample size  $n \in (100, 200, 300)$ , without censoring and left-censoring proportion settings (5%, 10%, 20%) taken in each mixture component, and parameter values set at

$$0.7 SN_2 \left( \begin{bmatrix} 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 3 & 1 \\ 1 & 4 \end{bmatrix}, \begin{bmatrix} 2 \\ 4 \end{bmatrix} \right) + 0.3 SN_2 \left( \begin{bmatrix} 5 \\ 7 \end{bmatrix}, \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \begin{bmatrix} 3 \\ 5 \end{bmatrix} \right).$$

To fit the data we used the models FM-MSNC and FM-MNC and for each model we obtain the estimate of the posterior probability that an observation  $y_i$  belongs to the  $j$ th component of the mixture,  $\widehat{Z}_{ij}$ . So, if  $\max_j \widehat{Z}_{ij}$  occurs in component  $j$ , then  $y_i$  is classified into group  $j$ . For the  $m$ th sample of the MC, we computed the correct classification rate, denoted by  $CCR_m$ , then obtained the average of the correct classification rate (ACCR) of  $CCR_m$ . Table 5 shows the ACCR values. From this table it is possible to observe that the model produces a high correct classification rate in both fitted models. We see that the rate decreases when the censoring proportion increases, this decrease is stronger for  $n = 100$ . Looking at the  $n$  samples, keeping the censoring proportion fixed, the rate increased when the sample size increased.

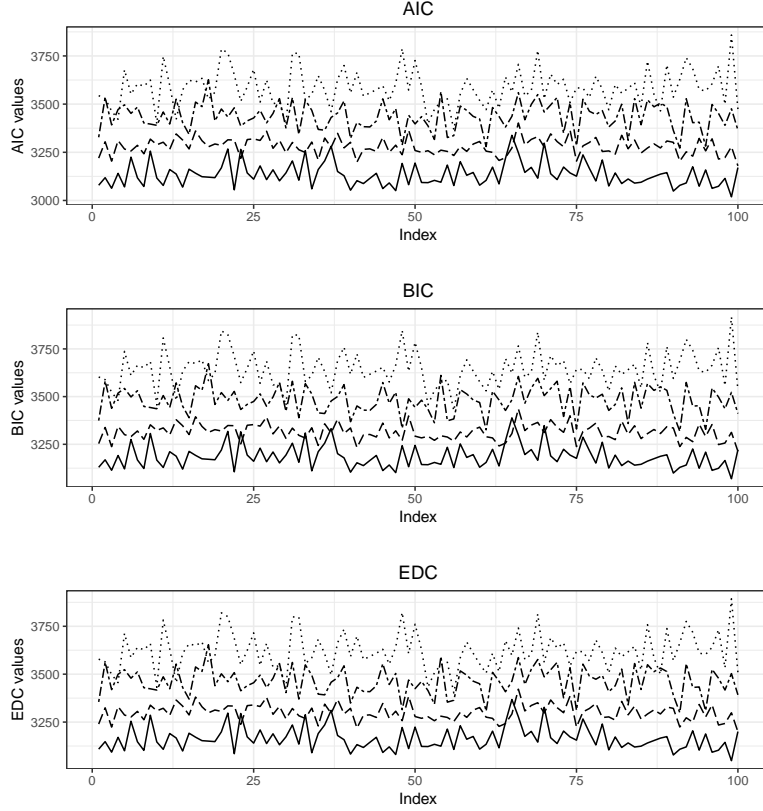


Figure 2: Simulated data: Number of mixture components. AIC, BIC and EDC values for 100 samples and left-censoring level 10%. Solid line: FM-MSNC, long dashed: FM-MNC(2), dot dashed: FM-MNC(3) and dotted: FM-MNC(4).

Table 5: Simulated data: Clustering. ACCR for fitted models FM-MSNC and FM-MNC for the simulated.

| n   | FM-MSNC |        |        |        | FM-MNC |        |        |        |
|-----|---------|--------|--------|--------|--------|--------|--------|--------|
|     | 0%      | 5%     | 10%    | 20%    | 0%     | 5%     | 10%    | 20%    |
| 100 | 0.9685  | 0.9619 | 0.9558 | 0.945  | 0.8801 | 0.8809 | 0.8772 | 0.8735 |
| 200 | 0.9729  | 0.9661 | 0.9599 | 0.9508 | 0.9206 | 0.9191 | 0.916  | 0.9016 |
| 300 | 0.9733  | 0.9661 | 0.9608 | 0.9545 | 0.9248 | 0.9229 | 0.9233 | 0.9155 |

Figure 3 shows the allocations in each group for sample size  $n = 200$  and left-censoring proportion of 0%,5%,10% and 20%, where the groups are represented by black and red points. The first line of graphics (a - d) contains the scatter plot of the generated real data, where the black circles represent an observation erroneously classified as belonging to the black group. The second line of graphics (e - h) contains the scatter plot of the fitted FM-MSNC model, where the black circles represent an observation erroneously classified as belonging to the black group. The last line of graphics (i - l) contains the scatter plot of the fitted FM-MNC model, where the red circles represent an observation erroneously classified as belonging to the red group.

### 5.5 Asymptotic properties

In this simulation study, we analyze the absolute bias and the mean square error (mse) of the estimates obtained from the FM-MSNC model through the proposed EM algorithm. The idea of this simulation is to provide empirical evidence about the consistency of the ML estimates. These measures are defined by

$$bias(\theta_i) = \frac{1}{M} \sum_{m=1}^M |\hat{\theta}_i^{(m)} - \theta_i| \quad \text{and} \quad mse(\theta_i) = \frac{1}{M} \sum_{m=1}^M (\hat{\theta}_i^{(m)} - \theta_i)^2, \quad (31)$$

where  $M$  is the number of MC samples, and  $\hat{\theta}_i^{(m)}$  is the estimated ML of the parameter  $\theta_i$  for the  $m$ th sample. Four different sample sizes ( $n = 300, 600, 900, 1200$ ) are considered. For each sample size, we generated 500 Monte Carlo



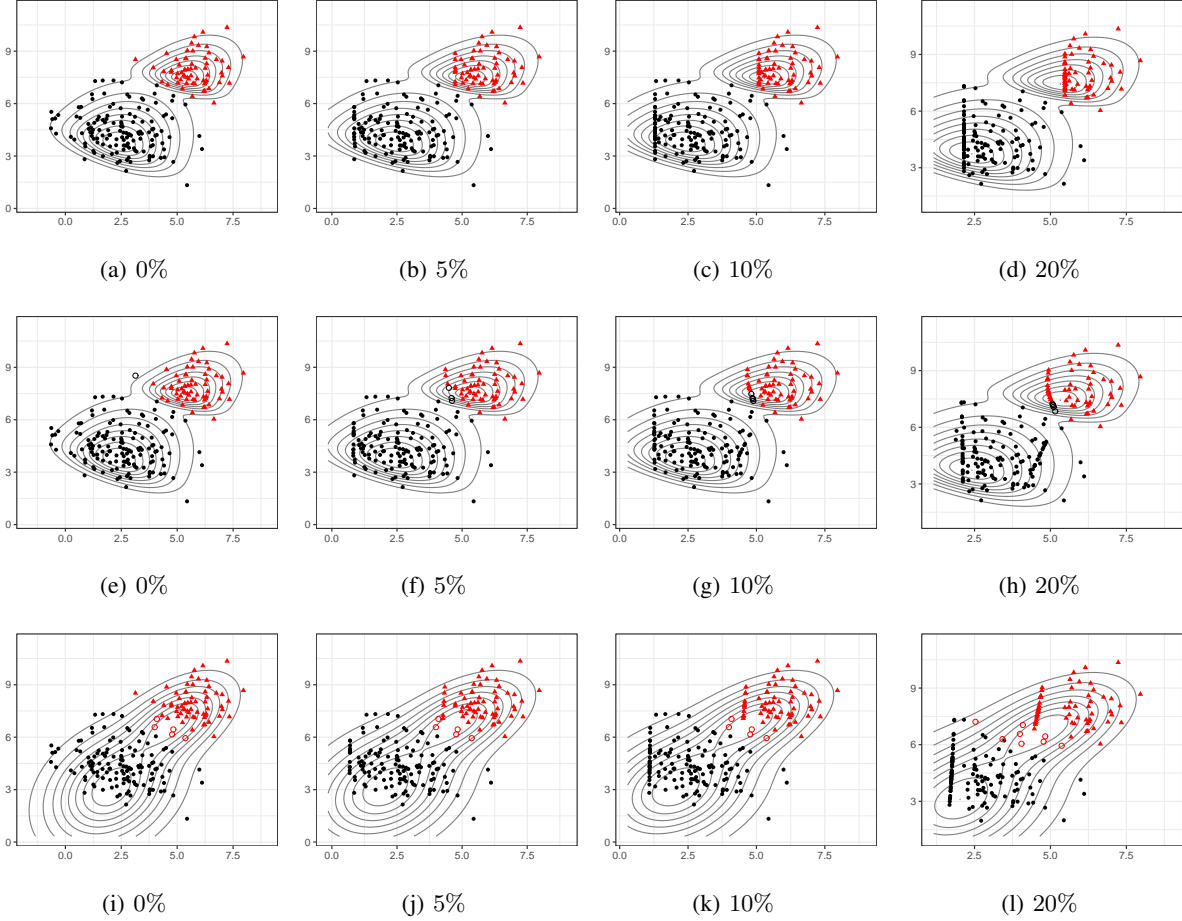


Figure 3: Simulated data: Clustering. Scatter plots for some simulated data from FM-MSNC model with  $n = 200$  and the respective density contours (first line). Clustering scatter plots from fitted skew-normal (second line) and normal (last line) with multiples censoring level.

samples with 5%, 10%, 15% censoring proportion. Using the EM algorithm, the absolute bias and mean squared error for each parameter over the 500 datasets were computed. The parameter setup is as follows

$$0.65 SN_2 \left( \begin{bmatrix} -5 \\ -4 \end{bmatrix}, \begin{bmatrix} 3 & 1 \\ 1 & 4.5 \end{bmatrix}, \begin{bmatrix} -2 \\ 3 \end{bmatrix} \right) + 0.35 SN_2 \left( \begin{bmatrix} 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 2 & 1 \\ 1 & 3.5 \end{bmatrix}, \begin{bmatrix} -2 \\ 3 \end{bmatrix} \right).$$

The results of the estimates of  $\mu_1, \mathbf{F}_1, \lambda_1, \mu_2, \mathbf{F}_2, \lambda_2$  and  $\pi$  are shown in Figures 4, 5, and 6. As a general rule, we can say that the *bias* and *mse* tend to approach zero when the sample size increases, indicating that the estimates based on the proposed EM algorithm, under the FM-MSNC model, provide good asymptotic properties.

## 6 Application

To illustrate the performance of our proposed model and algorithm, we consider a dataset of trace metal concentrations collected by the Virginia Department of Environmental Quality (VDEQ) that was previously analyzed by [16] and [19] using the normal and Student-t distribution, respectively.

This dataset consists of  $p = 5$  concentration levels of dissolved trace metals in independently selected  $n = 184$  freshwater streams across the Commonwealth of Virginia. The five attributes are trace metals: copper (Cu), lead (Pb), zinc (Zn), calcium (Ca) and magnesium (Mg). The Cu, Pb, and Zn concentrations are reported in  $\mu\text{g/L}$  of water, whereas Ca and Mg concentrations are reported in  $\text{mg/L}$  of water. Since the measurements were taken at different times, the presence of multiple limits of detection values are possible for each trace metal [34]. The limits of detection

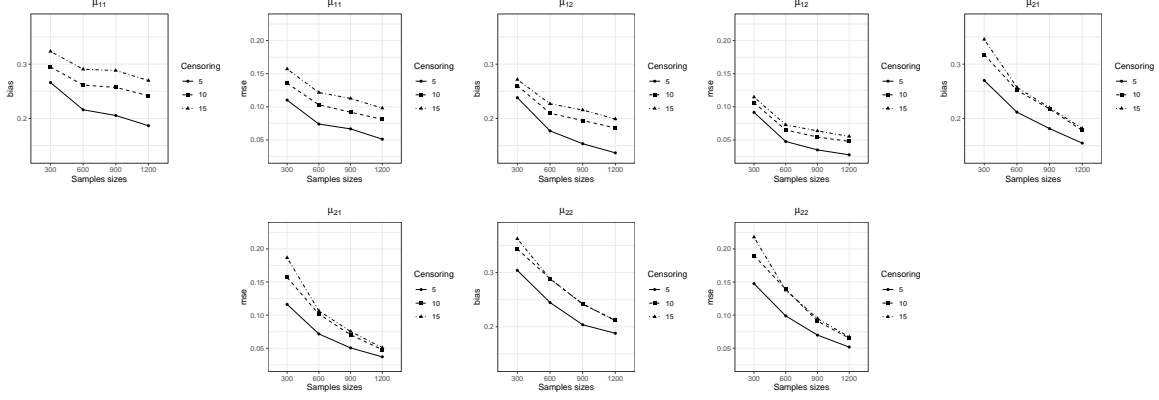


Figure 4: Simulated data: Asymptotic properties. bias and mse of  $\mu_1$  and  $\mu_2$  estimate in the FM-MSNC model with different censoring levels: 5% (solid line), 10% (dashed line), 15% (dot-dashed line).

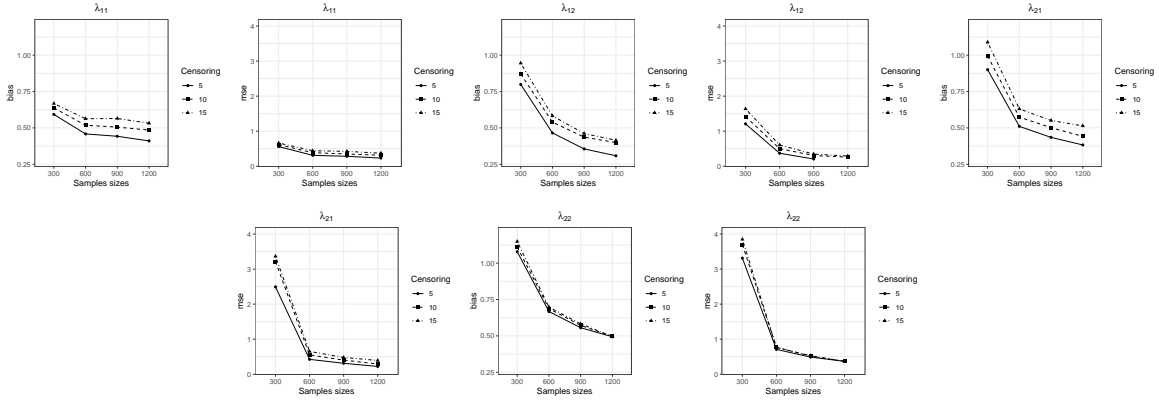


Figure 5: Simulated data: Asymptotic properties. bias and mse of  $\lambda_1$  and  $\lambda_2$  estimate in the FM-MSNC model with different censoring levels: 5% (solid line), 10% (dashed line), 15% (dot-dashed line).

for Cu and Pb are both the  $0.1 \mu\text{g/L}$ ,  $1.0 \mu\text{g/L}$  for Zn, while Ca and Mg have limits of  $0.5 \text{ mg/L}$  and  $1.0 \text{ mg/L}$ , respectively.

The percentage of left-censored values of 2.7% for (Ca), 4.9% for (Cu), 9.8% for (Mg) are small in comparison to 78.3% for (Pb) and 38.6% for (Zn). Also note that 17.9% of the streams had 0 non-detected trace metals, 39.1% had 1, 37.0% had 2, 3.8% had 3, 1.1% had 4 and 1.1% had 5. As the concentration levels are strictly positive measures, to guarantee this, we consider an interval-censoring analysis by setting all lower limits of detection equal to 0 for all trace metals. Also, due to the different scales for each trace metal, we standardize the dataset to have zero mean and variance equal to one as in [37]. The work mentioned before considered this dataset to be left censored without taking into account the possibility of predicting negative concentration levels for the trace metals. For instance, note that Pb censored concentrations take values in the small interval  $[0, 0.1]$ . Thus, after transforming the data, the new limits of detection are  $-0.8776$  (Cu),  $-0.3124$  (Pb),  $-0.4719$  (Zn),  $-0.7894$  (Ca),  $-0.6289$  (Mg). Figure 7 shows the histogram for each original trace metal with the detection limits and all of them together. It can be seen that most of the distributions associated with the variables have two or more modes and are right skewed. For this reason we propose to fit a FM-MSNC model.

We fit the data with 1, 2 and 3 components considering the FM-MSNC, FM-MtC and FM-MNC models, for the FM-MtC model we consider fixed degrees of freedom, as described in [19]. The number of groups of the model is chosen according to the information criteria as shown in Table 6. It can be seen that according to all model selection criteria the FM-MSNC model with three components fits the data best. We considered the variance-covariance ( $\Gamma$ ) to be equal in order to reduce the number of parameters to be estimated (parsimonious model).

The ML estimates of the parameters were obtained using the EM algorithm described in Subsection 4.1. The results are shown in Table 7. As can be seen, 46.99% of the freshwater streams belong to Cluster 1, Cluster 2 contains around

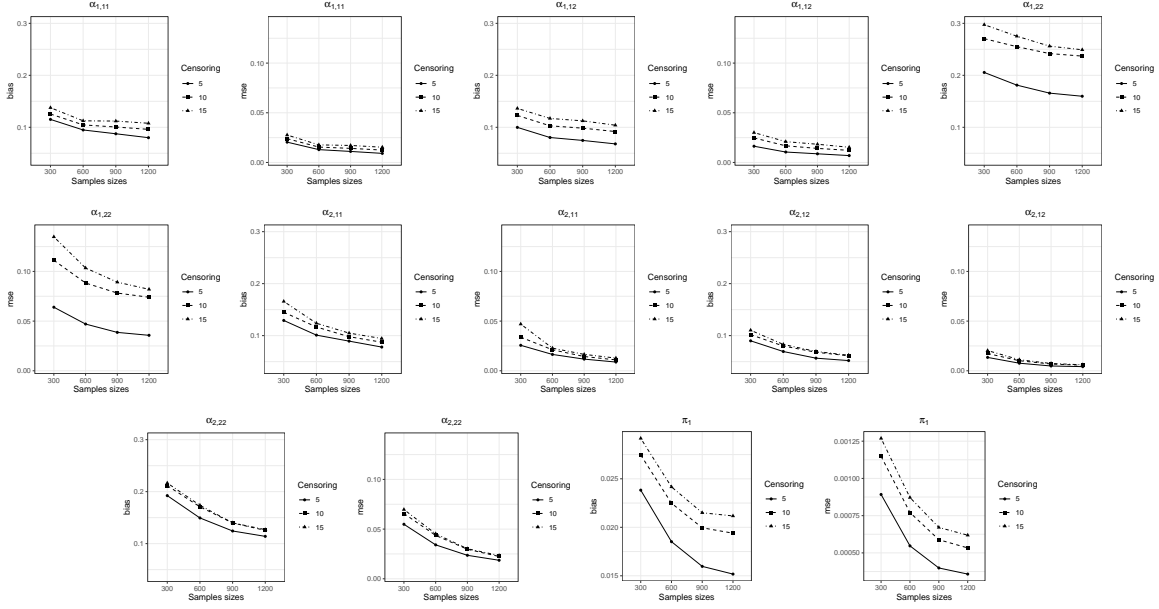


Figure 6: Simulated data: Asymptotic properties. bias and mse of  $\mathbf{F}_1 = \Sigma_1^{1/2}$ ,  $\mathbf{F}_2 = \Sigma_2^{1/2}$  and  $\pi$  estimates in the FM-MSNC model with different censoring levels: 5% (solid line), 10% (dashed line), 15% (dot-dashed line).

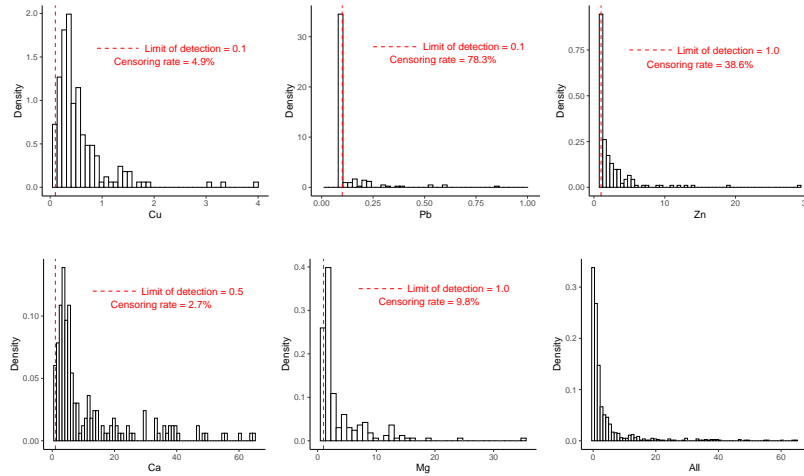


Figure 7: VDEQ data. Histograms of all the original five attributes of Virginia trace metal concentration data and all attributes together. The red line means the censoring limit detection for each concentration.

34.40% of them and the remaining 15.78% belong to Cluster 3. Table 6 shows that the best FM-MNC model has two components, and in this 96.73% of freshwater streams belong to Cluster 1 and the remaining 3.27% are in Cluster 2. For the FM-MtC model, the best model also has two components and three degrees of freedom. In this case, 85.19% of freshwater streams belong to Cluster 1 and the remaining 14.81% belong to Cluster 2.

In Figure 8, we fit the data using the FM-MSNC with three components. The scatter plots of the observations  $y_i$  ( $i = 1, \dots, 184$ ) for each pair of trace metals reveal that it is difficult to classify freshwater streams by visualization because these observations almost mix together.

## 7 Conclusions

In this paper, a novel approach to analyze multiply censored and missing data is presented based on the use of finite mixtures of multivariate skew-normal distributions. This approach generalizes several previously proposed solutions

Table 6: VDEQ data. Model selection criteria for various FM-MSNC and FM-MNC models. Values in bold correspond to the best model according to the criteria.

| Criteria       | FM-MSNC     |             |                  | FM-MNC     |            |              |
|----------------|-------------|-------------|------------------|------------|------------|--------------|
|                | $G = 1$     | $G = 2$     | $G = 3$          | $G = 1$    | $G = 2$    | $G = 3$      |
| Log-likelihood | -1269.302   | -910.3387   | <b>-697.6815</b> | -1351.596  | -1268.848  | -1210.626    |
| AIC            | 2588.604    | 1892.677    | <b>1489.363</b>  | 2743.192   | 2589.695   | 2485.253     |
| BIC            | 2668.977    | 2008.415    | <b>1640.465</b>  | 2807.491   | 2673.284   | 2588.13      |
| EDC            | 2606.427    | 1918.343    | <b>1522.871</b>  | 2757.451   | 2608.232   | 2508.066     |
| Time           | 1.7023 min. | 4.0229 min. | 9.2552 min.      | 1.054 sec. | 7.729 sec. | 26.0381 sec. |

| Criteria       | FM-MtC      |              |             |              |              |           |
|----------------|-------------|--------------|-------------|--------------|--------------|-----------|
|                | $\nu = 3$   |              |             | $\nu = 4$    |              |           |
|                | $G = 1$     | $G = 2$      | $G = 3$     | $G = 1$      | $G = 2$      | $G = 3$   |
| Log-likelihood | -1040.276   | -1018.943    | -1074.852   | -1061.702    | -1036.393    | -1072.487 |
| AIC            | 2120.553    | 2089.887     | 2213.705    | 2163.404     | 2124.786     | 2208.974  |
| BIC            | 2184.852    | 2173.475     | 2316.583    | 2227.702     | 2208.375     | 2311.852  |
| EDC            | 2134.812    | 2108.423     | 2236.519    | 2177.662     | 2143.323     | 2231.788  |
| Time           | 16.672 sec. | 35.9952 sec. | 2.8734 min. | 13.1202 sec. | 29.9673 sec. | 2.68 min. |

Table 7: VDEQ data. ML estimates of parameters from fitting the FM-MSNC model with 3 components to the Virginia trace metal concentration data.

| Parameter                       | Estimate   |
|---------------------------------|--|
| $(\pi_1, \pi_2, \pi_3)$         | (0.4699, 0.3440, 0.1861)   |
| $\mu_1$                         | (-0.3789, -0.5344, -0.722, -0.5744, -0.4833)   |
| $\mu_2$                         | (-0.8189, -0.6838, -0.4485, 0.805, 0.7948)   |
| $\mu_3$                         | (1.2681, -0.463, -0.503, -0.3272, -0.305)  |
| $\lambda_1$                     | (0.4867, -1.4692, 13.2902, -0.4112, -0.0062)   |
| $\lambda_2$                     | (28.3006, -1.012, 6.0129, 4.4658, -1.6659)   |
| $\lambda_3$                     | (-0.5228, 11.9182, 12.3316, -1.3583, 1.2298)   |
| $\mathbf{F}_1 = \Sigma_1^{1/2}$ | $\begin{bmatrix} 0.3328 & 0.1229 & 0.0419 & 0.0338 & 0.0573 \\ & 0.3733 & 0.0060 & -0.0105 & 0.0038 \\ & & 0.5540 & -0.0174 & -0.0061 \\ & & & 0.0952 & 0.0554 \\ & & & & 0.1208 \end{bmatrix}$    |
| $\mathbf{F}_2 = \Sigma_2^{1/2}$ | $\begin{bmatrix} 1.0582 & -0.035 & 0.1628 & 0.1981 & 0.0710 \\ & 0.102 & -0.0074 & 0.0025 & 0.0111 \\ & & 0.4745 & 0.0157 & 0.1260 \\ & & & 1.0244 & 0.3920 \\ & & & & 1.2392 \end{bmatrix}$       |
| $\mathbf{F}_3 = \Sigma_3^{1/2}$ | $\begin{bmatrix} 1.4992 & -0.1232 & -0.2405 & 0.1341 & 0.0950 \\ & 2.2952 & 0.8411 & -0.0459 & -0.0915 \\ & & 2.1716 & -0.1003 & -0.0633 \\ & & & 0.2824 & 0.1472 \\ & & & & 0.2640 \end{bmatrix}$ |

for censored data, such as, the finite mixture of Gaussian components [17, 12, 16] and the finite mixture of Student-t components [19], which are also restricted to a left or right censored problem. 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 vector and covariance matrix of the multivariate truncated skew-normal distribution, for which the R `MomTrunc` library is used [15]. The proposed EM algorithm was implemented as part of the R package `CensMFM` and is available for download at the CRAN repository. The experimental results and the analysis of a real dataset provide support for the usefulness and effectiveness of our proposal.

The method proposed in this paper can be extended to other types of mixture distributions, for example, the multivariate scale mixtures of skew-normal distributions [11] or generalized hyperbolic mixtures [10]. It is also of interest to develop an effective Markov chain Monte Carlo algorithm for the FM-MSNC models in a fully Bayesian treatment. Finally, the proposed methods can also be easily applied to other substantial areas in which the data being analyzed have censored and/or missing observations, for instance, factor analysis models [36] and linear mixed models [22, 18].

## References

- [1] Akaike H (1974) A new look at the statistical model identification. *IEEE Trans Autom Cont* 19:716–723

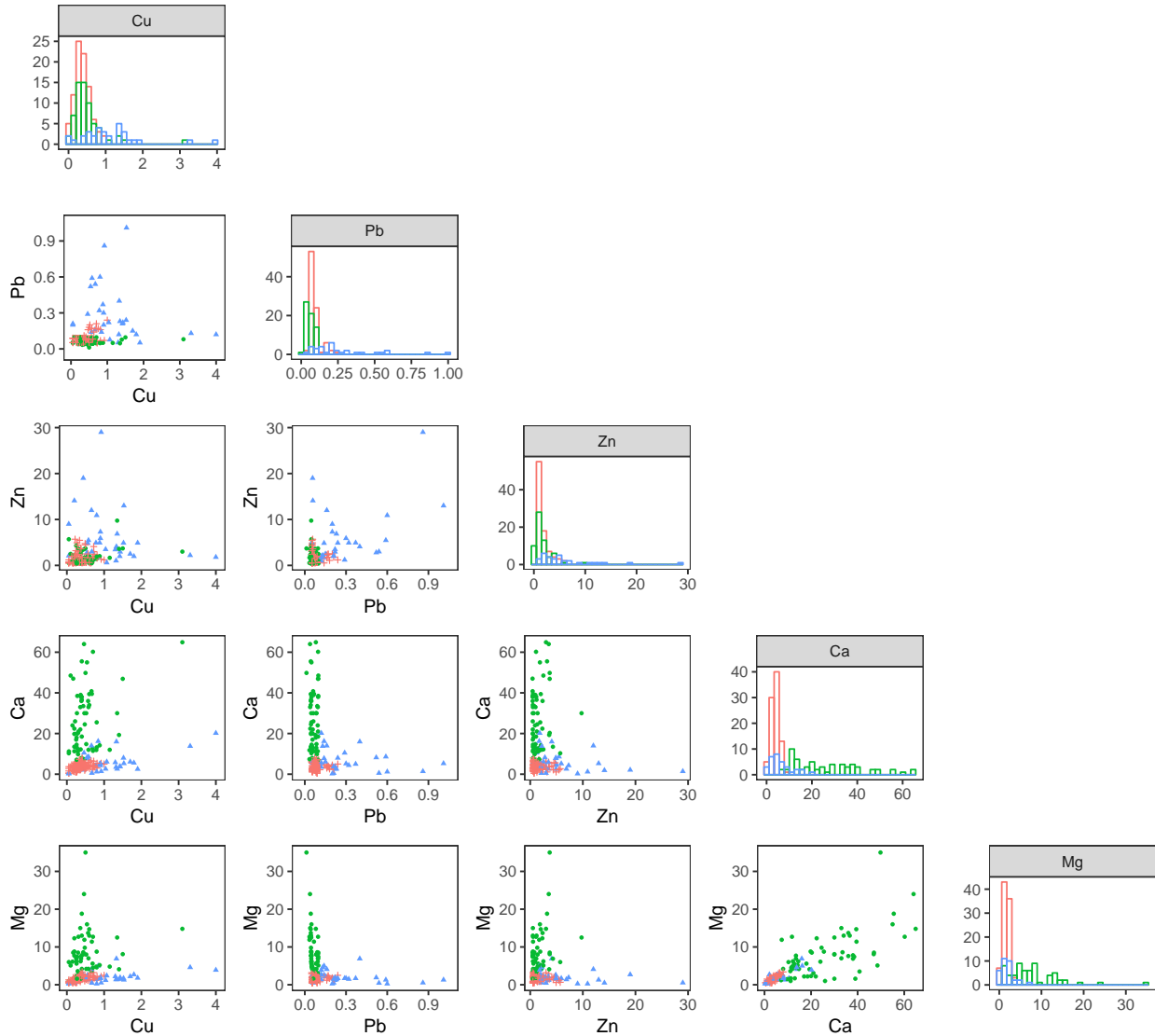


Figure 8: Scatter plots and histograms of the fitted values of  $y_j$  for each predicted clusters of the VDEQ data. Lower diagonal entries: ● cluster 1; ▲ cluster 2; + cluster 3.

- [2] Arellano-Valle RB, Genton MG (2005) On fundamental skew distributions. *Journal of Multivariate Analysis* 96:93–116
- [3] Arellano-Valle RB, Genton MG (2010) Multivariate extended skew-t distributions and related families. *Metron* LXVIII:201–234
- [4] Azzalini A, Capitanio A (1999) Statistical applications of the multivariate skew-normal distribution. *Journal of the Royal Statistical Society, Series B* 61:579–602
- [5] Azzalini A, Dalla-Valle A (1996) The multivariate skew-normal distribution. *Biometrika* 83(4):715–726
- [6] Bai Z, Krishnaiah P, Zhao L (1989) On rates of convergence of efficient detection criteria in signal processing with white noise. *Inform Theory IEEE Trans* 35:380–388
- [7] Basford K, Greenway D, McLachlan G, Peel D (1997) Standard errors of fitted component means of normal mixtures. *Computational Statistics* 12:1–18
- [8] Basso RM, Lachos VH, Cabral CRB, Ghosh P (2010) Robust mixture modeling based on scale mixtures of skew-normal distributions. *Computational Statistics & Data Analysis* 54(12):2926–2941

- [9] Bouveyron C, Celeux G, Murphy T, Raftery A (2019) *Model-Based Clustering and Classification for Data Science: With Applications in R*. Cambridge University Press
- [10] Browne RP, McNicholas PD (2015) A mixture of generalized hyperbolic distributions. *Canadian Journal of Statistics* 43(2):176–198
- [11] Cabral CRB, Lachos VH, Prates MO (2012) Multivariate mixture modeling using skew-normal independent distributions. *Computational Statistics & Data Analysis* 56:126–142
- [12] Caudill SB (2012) A partially adaptive estimator for the censored regression model based on a mixture of normal distributions. *Statistical Methods & Applications* 21:121–137
- [13] 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
- [14] Frühwirth-Schnatter S (2006) *Finite mixture and Markov switching models*. Springer Science & Business Media
- [15] Galarza CE, Kan R, Lachos VH (2020) MomTrunc: Moments of Folded and Doubly Truncated Multivariate Distributions. R Package Version 5.87 URL <http://CRAN.R-project.org/package=MomTrunc>
- [16] He J (2013) Mixture model based multivariate statistical analysis of multiply censored environmental data. *Advances in Water Resources* 59:15–24
- [17] Karlsson M, Laitila T (2014) Finite mixture modeling of censored regression models. *Statistical Papers* 55(3):627–642
- [18] Lachos VH, Bandyopadhyay D, Dey DK (2011) Linear and nonlinear mixed-effects models for censored HIV viral loads using normal/independent distributions. *Biometrics* 67:1594–1604
- [19] Lachos VH, Moreno E JL, Chen K, Cabral CRB (2017) Finite mixture modeling of censored data using the multivariate Student-t distribution. *Journal of Multivariate Analysis* 159:151–167
- [20] Lachos VH, Cabral CRB, Zeller CB (2018) *Finite Mixture of Skewed Distributions*. Springer
- [21] Lin TI (2009) Maximum likelihood estimation for multivariate skew normal mixture models. *Journal of Multivariate Analysis* 100(2):257–265
- [22] Lin TI, Ho HJ, Chen CL (2009) Analysis of multivariate skew normal models with incomplete data. *Journal of Multivariate Analysis* 100(19):2337–2351
- [23] Lin TI, Wang WL (2019) Multivariate-t linear mixed models with censored responses, intermittent missing values and heavy tails. *Statistical Methods in Medical Research* p 0962280219857103
- [24] Lin TI, Lachos VH, Wang WL (2018) Multivariate longitudinal data analysis with censored and intermittent missing responses. *Statistics in Medicine* 37(19):2822–2835
- [25] Little RJ, Rubin DB (2002) *Statistical analysis with missing data*, vol 793. John Wiley & Sons
- [26] Louis TA (1982) Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society, Series B* 44:226–233
- [27] McLachlan GJ, Krishnan T (2008) *The EM Algorithm and Extensions*, 2nd edn. Wiley
- [28] McLachlan GJ, Peel D (2000) *Finite Mixture Models*. Wiley, New York
- [29] McNicholas PD (2016) *Mixture model-based classification*. Chapman and Hall/CRC
- [30] Meilijson I (1989) A fast improvement to the em algorithm on its own terms. *Journal of the Royal Statistical Society, Series B (Methodological)* 51(1):127–138
- [31] Peel D, McLachlan GJ (2000) *Finite Mixture Models*. John Wiley & Sons
- [32] Peel D, McLachlan GJ (2000) Robust mixture modelling using the t distribution. *Statistics and Computing* 10(4):339–348
- [33] Prates MO, Lachos VH, Cabral C (2013) mixsmsn: Fitting finite mixture of scale mixture of skew-normal distributions. *Journal of Statistical Software* 54(12):1–20
- [34] Richmond V (2003) *The quality of virginia non-tidal streams: First year report*
- [35] Schwarz G (1978) Estimating the dimension of a model. *The Annals of Statistics* 6:461–464
- [36] Wang WL, Liu M, Lin TI (2017) Robust skew-t factor analysis models for handling missing data. *Statistical Methods & Applications* 26(4):649–672
- [37] Wang WL, Castro LM, Lachos VH, Lin TI (2019) Model-based clustering of censored data via mixtures of factor analyzers. *Computational Statistics & Data Analysis* 140:104–121
- [38] Zeller CB, Cabral CR, Lachos VH (2016) Robust mixture regression modeling based on scale mixtures of skew-normal distributions. *Test* 25(2):375–396

This figure "test.png" is available in "png" format from:

<http://arxiv.org/ps/2009.10826v1>

