Mathematics Interdisciplinary Research **9** (4) (2024) 385 − 411

Original Scientific Paper

Flexible Parsimonious Mixture of Skew Factor Analysis Based on Normal Mean–Variance Birnbaum-Saunders

Farzane Hashemi, *Jalal Askari** and *Saeed Darijani*

Abstract

The purpose of this paper is to extend the mixture factor analyzers (MFA) model to handle missing and heavy-tailed data. In this model, the distribution of factors loading and errors arise from the multivariate normal mean-variance mixture of the Birnbaum-Saunders (NMVBS) distribution. By using the structures covariance matrix, we introduce parsimonious MFA based on NMVBS distribution. An Expectation Maximization (EM)-type algorithm is developed for parameter estimation. Simulations study and real data sets represent the efficiency and performance of the proposed model.

Keywords: Normal mean-variance distribution, EM-type algorithm, Factor analysis, Heavy-tail, Strongly leptokurtic.

2020 Mathematics Subject Classification: 62H12, 62H25.

How to cite this article
F. Hashemi, J. Askari and S. Darijani, Flexible parsimonious mixture of skew factor analysis based on normal mean variance Birnbaum-Saunders, *Math. Interdisc. Res.* **9** (4) (2024) 385-411.

1. Introduction

Model-based clustering is a useful method to find proper groups for datasets with unlabeled observations. For this issue, the finite mixture model is used to fit

*Corresponding author (E-mail: askari@kashanu.ac.ir)
Academic Editor: Mohammad Amini
Received 21 February 2024, Accepted 16 August 2024
DOI: 10.22052/MIR.2024.254416.1459

the observation and specified the data classification. The multivariate normal mixture models for clustering are discussed in [1]. Originally, a finite mixture model based on multivariate normal distribution is sensitive to skewed and heavy-tailed datasets. Andrews and McNicholas [2] considered multivariate t-distribution replacing normal distribution for handling heavy-tail data. Afterward, Browne and McNicholas [3] proposed a finite mixture model based on generalized hyperbolic distribution for heavy-tail and skewed data. Extensive details for model-based clustering are proposed by [4].

When data have high dimensions, the model-based clustering may produce bias and inaccurate inferences for estimating parameters and group datasets. This bias and inaccurate inferences are more estimates of the component covariance than other parameters. To overcome this problem, MFA was proposed by Ghahramani and Hinton [5] to reduce the number of model-free parameters. Although the initial MFA model [5] assumed the component factors and errors of the multivariate normal distribution as a mixture component, the interest in skewed and heavily-tailed distributions has increasingly been considered a robust framework for data analysis with incomplete data. One can refer to [6–10] a few recently published contributions.

The problem of multimodal data is ubiquitous in many scientific fields and should be addressed appropriately before engaging learning algorithms. To extend the MFA model, four issues are always of interest: (i) theoretical convenience and mathematical properties, (ii) computational tractability such as easy calibration of data and speed of the importation, (iii) the flexibility and robustness in analyzing strongly skewed distributions with outliers, and (iv) reduction in the number of parameters that must be estimated. In this regard, the family of skew-normal [11] distributions is one of the baseline platforms that play a subnational role in constructing efficient MFA. For instance, the mixture of skew-normal factor analysis models was proposed in [12]. Lin et al. [7] introduced the mixture of skew-t factor analysis models as an extension of the mixture of skew-normal factor analysis models for handling heavy-tailed asymmetric datasets. Wei et al. [8] discussed the issue of MFA as well as another class of skew distribution by considering the class of normal mean-variance distributions. MFA has a covariance component with $\boldsymbol{\Sigma}_i = \boldsymbol{B}_i\boldsymbol{B}_i^\top + \boldsymbol{D}_i$ structure where the factor loading matrix $\boldsymbol{B}_i$ is a $p \times q$ matrix and the noise matrix $\boldsymbol{D}_i$ is diagonal. By introducing MFA, McNicholas and Murphy [13] considered eight structures of covariance matrix by constrained or unconstrained $\boldsymbol{B}_i$ and $\boldsymbol{D}_i$. In addition, Murray et al. [6, 14] introduced a parsimonious mixture factor analysis based on the skew-t distribution as an extension of the parsimonious MFA model based on the class of normal mean-variance distributions for handling heavy-tailed asymmetric datasets.

Recently, Hashemi et al. [15] extended the original factor analysis to formulate a new class of skewed models as an alternative to the baseline factor analysis model. Calling the normal mean-variance Birnbaum-Saunders factor analysis (NMVBSFA) model, Hashemi et al. [15] showed that the NMVBSFA not only inherits mathematical and computational aspects of the skew-normal factor analysis

and skew-t factor analysis model but also it can provide a flexible platform for statistical analysis. Hashemi et al. [16] used the mean-mixture normal distribution for proposing a new factor analyzer model. Through a simulation study, they showed that some special cases of mean-mixture normal factor analysis outperform the skew-t factor analysis of large amounts of degrees of freedom. Numerical results of [16] confirmed that the calibration of mean-mixture normal distribution of the data is much faster than some existing models. Therefore, the clustering of multivariate data with high dimensions is one of the main reasons for the introduction of the finite mixture NMVBSFA model. Another of the main reasons is to reduce the free parameters model by introducing 8 parsimonious structures based on McNicholas and Murphy [13].

In this contribution, we advocate the use of NMVBS distribution as the component factors and errors in MFA in an additional eight structures for the covariance matrix. Referring to the properties of the NMVBS distributions and NMVBSFA models discussed in the next section, it is demonstrated that the new mixture factor analysis based on the NMVBSFA model is attractive as it can simultaneously model the skewness and heavy tails. The model parameters are estimated based on the development of an EM-type [17] algorithm using a hierarchical representation. One of the serious concerns related to multivariate mixture models is the potential over-parameterization which is well documented in the work of McNicholas and Murphy [13] on parsimonious Gaussian and t-mixtures models. Following McNicholas and Murphy [13], we will present the ways to address this deficiency by offering eigen-decomposition of the component covariance matrices. Simulation studies for assessing the performance and computational tractability of the proposed methodology are carried out. Finally, real-world data analyses are presented to support the usefulness of the new model.

The layout of this paper is as follows: In Section 2, we recall the NMVBSFA models. Section 3 addresses the estimation and computational framework of the NMVBS mixture. For parameter estimation, we extend an EM-type algorithm of the Expectation Conditional Maximization (ECM), represented by [18]. Meanwhile, we compute the standard errors for the NMVBSFA mixture. In Sections 5 and 6, the proposed methodology is demonstrated with extensive application of a real data set and simulation studies.

2. Statistical models

This section reviews the formulation and some properties of the NMVBS distribution to describe the motivation and parameter of estimation. NMVBS distribution followed by a p-dimensional random vector $\boldsymbol{Y} \in \mathbb{R}^p$, denoted by $\boldsymbol{Y} \sim NMVBS_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \alpha)$, with location, scale and shape parameters $\boldsymbol{\mu} \in \mathbb{R}^p$, $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ and $\boldsymbol{\lambda} \in \mathbb{R}^p$, respectively, can be depicted as:

$$\boldsymbol{Y} \overset{d}{=} \boldsymbol{\mu} + W\boldsymbol{\lambda} + \sqrt{W}\boldsymbol{X}, \tag{1}$$

in which $\boldsymbol{X}$ depicts a p-dimensional normal random vector with $\mathrm{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ representation, W a positive random variable, independent of $\boldsymbol{X}$, with Birnbaum-Saunders distribution [19] with $BS(\alpha, \beta = 1)$ representation and $\stackrel{d}{=}$ shows the identically distribution. The probability density function (pdf) of W is given by:

$$f(w; \alpha, \beta) = A(w, \alpha, \beta)\phi(a(w, \alpha, \beta)), \qquad w > 0,\ \alpha > 0,\ \beta > 0,$$

where $\phi(\cdot)$ is the pdf of the standard normal distribution, $a(w, \alpha, \beta) = (\sqrt{w/\beta} - \sqrt{\beta/w})/\alpha$, and $A(w, \alpha, \beta)$ is the derivative of $a(w, \alpha, \beta)$ concerning w. Specially, the $NMVBS_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \alpha)$ distributed random vector $\boldsymbol{Y}$ in (1) indicated hierarchically as:

$$\boldsymbol{Y}|W = w \sim \mathrm{N}_p(\boldsymbol{\mu} + w\boldsymbol{\lambda}, w\boldsymbol{\Sigma}) \quad \text{and} \quad W \sim BS(\alpha, 1). \tag{2}$$

The pdf of $\boldsymbol{Y}$ is according to

$$\begin{aligned} & f_{NMVBS}(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\Sigma}, \alpha) \\ = \ & \frac{1}{2} f_{GH}\left(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\Sigma}, \frac{1}{2}, \frac{1}{\alpha^2}, \frac{1}{\alpha^2}\right) + \frac{1}{2} f_{GH}\left(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\Sigma}, -\frac{1}{2}, \frac{1}{\alpha^2}, \frac{1}{\alpha^2}\right), \end{aligned} \tag{3}$$

where f_{GH} is the pdf of GH distribution that represent in Appendix A [20]. It is clear that

$$E(\boldsymbol{Y}) = \boldsymbol{\mu} + E(W)\boldsymbol{\lambda} \qquad \text{and} \qquad Cov(\boldsymbol{Y}) = E(W)\boldsymbol{\Sigma} + Var(W)\boldsymbol{\lambda}\boldsymbol{\lambda}^\top.$$

Proposition 2.1. *Let $\boldsymbol{Y}$ and W be the random vector and variable distributed by $NMVBS_{\ p}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \alpha)$ and $BS(\alpha, 1)$, respectively. Then, for any $\boldsymbol{y} \in \mathbb{R}^p$, the pdf of W given $\boldsymbol{Y} = \boldsymbol{y}$ is represented by*

$$\begin{aligned} f(w \mid \boldsymbol{y}) \ = \ & \pi(\boldsymbol{y}) f_{GIG}\left(w; \frac{1-p}{2}, \chi(\boldsymbol{y}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha), \psi(\boldsymbol{\lambda}, \boldsymbol{\Sigma}, \alpha)\right) \\ & + (1 - \pi(\boldsymbol{y})) f_{GIG}\left(w; \frac{-1-p}{2}, \chi(\boldsymbol{y}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha), \psi(\boldsymbol{\lambda}, \boldsymbol{\Sigma}, \alpha)\right), \end{aligned} \tag{4}$$

where $f_{GIG}(.)$ is the pdf of generalized inverse Gaussian (GIG) distribution that represent in Appendix A,

$$\begin{aligned} & \chi(\boldsymbol{y}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha) = (\boldsymbol{y} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) + \alpha^{-2}, \\ & \psi(\boldsymbol{\lambda}, \boldsymbol{\Sigma}, \alpha) = \boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda} + \alpha^{-2}, \\ & \pi(\boldsymbol{y}) = \frac{f_{GH_p}\left(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\Sigma}, 0.5, \alpha^{-2}, \alpha^{-2}\right)}{f_{GH_p}\left(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\Sigma}, 0.5, \alpha^{-2}, \alpha^{-2}\right) + f_{GH_p}\left(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\Sigma}, -0.5, \alpha^{-2}, \alpha^{-2}\right)}. \end{aligned} \tag{5}$$

Furthermore, for $r = \pm 1, \pm 2, \ldots$

$$E(W^r \mid \boldsymbol{y}) = \left(\frac{\chi(\boldsymbol{y}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha)}{\psi(\boldsymbol{\lambda}, \boldsymbol{\Sigma}, \alpha)}\right)^{r/2} \left\{ \pi(\boldsymbol{y}) \frac{K_{(1-p)/2+r}\left(\sqrt{\psi(\boldsymbol{\lambda}, \boldsymbol{\Sigma}, \alpha)\chi(\boldsymbol{y}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha)}\right)}{K_{(1-p)/2}\left(\sqrt{\psi(\boldsymbol{\lambda}, \boldsymbol{\Sigma}, \alpha)\chi(\boldsymbol{y}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha)}\right)} \right.$$
$$\left. + (1 - \pi(\boldsymbol{y})) \frac{K_{-(1+p)/2+r}\left(\sqrt{\psi(\boldsymbol{\lambda}, \boldsymbol{\Sigma}, \alpha)\chi(\boldsymbol{y}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha)}\right)}{K_{-(1+p)/2}\left(\sqrt{\psi(\boldsymbol{\lambda}, \boldsymbol{\Sigma}, \alpha)\chi(\boldsymbol{y}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha)}\right)} \right\}. \tag{6}$$

where $K_\kappa(\cdot)$ *denotes the modified Bessel function of the third kind of order* κ*.*

Proof. See Hahsemi et al. [15]. □

3. The mixture of NMVBSFA model

3.1 The formulation of the model

Consider $\boldsymbol{Y}_j$ is a p-variate random vector of the jth individual for $j = 1, \ldots, n$. In the finite mixture modeling, it is assumed that observations are categorized into g classes. Therefore, the set of unobservable allocation vectors $\boldsymbol{Z}_j = \left(Z_{1j}, \ldots, Z_{gj}\right)^\top$ for $j = 1, \ldots, n$ is usually defined for the sake of investigating statistical properties and computational purposes. Here, the elements of $\boldsymbol{Z}_j$ are the binary outcomes where $Z_{ij} = 1$ if the jth observation belongs to the ith component and otherwise $Z_{ij} = 0$. It follows the point that $\boldsymbol{Z}_j$ has a one-trail multinational distribution with probabilities $(\pi_1, \ldots, \pi_g)$, denoted by $\mathcal{M}(1; \pi_1, \ldots, \pi_g)$. In the mixture of NMVBSFA, we assume that each observation $\boldsymbol{Y}_j$ can be shown as:

$$\boldsymbol{Y}_j = \boldsymbol{\mu}_i + \boldsymbol{B}_i \boldsymbol{U}_{ij} + \boldsymbol{\varepsilon}_{ij} \quad \text{with probability } \pi_i \ (i = 1, \ldots, g), \tag{7}$$

where

$$\begin{bmatrix} \boldsymbol{U}_{ij} \\ \boldsymbol{\varepsilon}_{ij} \end{bmatrix} | Z_{ij} = 1 \sim NMVBS_{q+p}\left(\begin{bmatrix} -a_{\alpha_i}\boldsymbol{\Lambda}_i^{-1/2}\boldsymbol{\lambda}_i \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Lambda}_i^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D}_i \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Lambda}_i^{-1/2}\boldsymbol{\lambda}_i \\ \boldsymbol{0} \end{bmatrix}, \alpha_i\right), \tag{8}$$

$\boldsymbol{D}_i$ is a positive diagonal matrix and $\pi_1 + \ldots + \pi_g = 1$. Bear in mind that, by considering $a_{\alpha_i} = E(W_j | Z_{ij} = 1)$ and $b_{\alpha_i} = \mathrm{Var}(W_j | Z_{ij} = 1)$, $\boldsymbol{\Lambda}_i = a_{\alpha_i}\boldsymbol{I}_q + b_{\alpha_i}\boldsymbol{\lambda}_i\boldsymbol{\lambda}_i^\top$, the mixture of NMVBSFA model (8) performs $E(\boldsymbol{U}_{ij}) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{U}_{ij}) = \boldsymbol{I}_q$. By McNicholas and Murphy [13], we fulfill an eight-member parsimonious mixture of NMVBSFA (PM-NMVBSFA) similar to eight parsimonious mixture of Gaussian factor analyzer (Table 1). By taking into account two representations (2) and (8), the hierarchical representation of the postulated factor analyzer can be written as:

$$\begin{aligned} &\boldsymbol{Y}_j \mid W_j = w_j, Z_{ij} = 1 \sim N_p(\boldsymbol{\mu}_i - a_{\boldsymbol{\alpha}_i}\boldsymbol{B}_i\boldsymbol{\Lambda}_i^{-1/2}\boldsymbol{\lambda}_i + w_j\boldsymbol{B}_i\boldsymbol{\Lambda}_i^{-1/2}\boldsymbol{\lambda}_i, w_j\boldsymbol{\Sigma}_i), \\ &W_j | Z_{ij} = 1 \sim BS(\alpha_i, 1), \quad \text{with probability } \pi_i. \end{aligned} \tag{9}$$

Table 1: Nomination and covariance structures for the parsimonious mixture of Gaussian factor analyzer [13].

Model	$\boldsymbol{B}_i = \boldsymbol{B}$	$\boldsymbol{D}_i = \boldsymbol{D}$	$\boldsymbol{D}_i = d_i \boldsymbol{I}_p$	number of parameters
CCC	Constrained	Constrained	Constrained	$[pq - q(q-1)/2] + 1$
CCU	Constrained	Constrained	Unconstrained	$[pq - q(q-1)/2] + p$
CUC	Constrained	Unconstrained	Constrained	$[pq - q(q-1)/2] + g$
CUU	Constrained	Unconstrained	Unconstrained	$[pq - q(q-1)/2] + gp$
UCC	Unconstrained	Constrained	Constrained	$g[pq - q(q-1)/2] + 1$
UCU	Unconstrained	Constrained	Unconstrained	$g[pq - q(q-1)/2] + p$
UUC	Unconstrained	Unconstrained	Constrained	$g[pq - q(q-1)/2] + g$
UUU	Unconstrained	Unconstrained	Unconstrained	$g[pq - q(q-1)/2] + gp$

Thus, the PM-NMVBSFA model has the pdf of Y_j:

$$f(\boldsymbol{y}_j; \boldsymbol{\Theta}) = \sum_{i=1}^{g} \pi_i f_{NMVBS}(\boldsymbol{y}_j, \boldsymbol{\mu}_i - a_{\alpha_i}\boldsymbol{\eta}_i, \boldsymbol{\Sigma}_i, \boldsymbol{\eta}_i, \alpha_i), \tag{10}$$

where $\boldsymbol{\Sigma}_i = \boldsymbol{B}_i \boldsymbol{\Lambda}_i^{-1} \boldsymbol{B}_i^{\top} + \boldsymbol{D}_i$, $\boldsymbol{\eta}_i = \boldsymbol{B}_i \boldsymbol{\Lambda}_i^{-1/2} \boldsymbol{\lambda}_i$, and for $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{g-1})$ and $\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \boldsymbol{\lambda}_i, \boldsymbol{B}_i, \boldsymbol{D}_i, \alpha_i)$, the parameter set of the model is $\boldsymbol{\Theta} = (\boldsymbol{\pi}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g)$.

Alternatively, using incomplete data, the PM-NMVBSFA can be presented as:

$$\begin{aligned} \boldsymbol{Y}_j \mid Z_{ij} = 1 &\sim NMVBS_p(\boldsymbol{\mu}_i - a_{\alpha_i}\boldsymbol{\eta}_i, \boldsymbol{\Sigma}_i, \boldsymbol{\eta}_i, \alpha_i), \\ \boldsymbol{Z}_j &\sim \mathcal{M}(1; \pi_1, \ldots, \pi_g). \end{aligned} \tag{11}$$

To facilitate mathematical and computational procedures, the scaling transformations

$$\tilde{\boldsymbol{B}}_i \triangleq \boldsymbol{B}_i \boldsymbol{\Lambda}_i^{-1/2} \quad \text{and} \quad \tilde{\boldsymbol{U}}_{ij} \triangleq \boldsymbol{\Lambda}_i^{1/2} \boldsymbol{U}_{ij}, \tag{12}$$

are considered on the matrices of factor loadings $\boldsymbol{B}_i$ and the common factors $\boldsymbol{U}_{ij}$. Therefore, the high-level representation of the PM-NMVBSFA model can alternatively be written as:

$$\begin{aligned} \boldsymbol{Y}_j \mid (\tilde{\boldsymbol{u}}_{ij}, w_j, Z_{ij} = 1) &\sim N_p(\boldsymbol{\mu}_i + \tilde{\boldsymbol{B}}_i \tilde{\boldsymbol{u}}_{ij}, w_j \boldsymbol{D}_i), \\ \tilde{\boldsymbol{U}}_{ij} \mid (w_j, Z_{ij} = 1) &\sim N_q\big((w_j - a_{\alpha_i})\boldsymbol{\lambda}_i, w_j \boldsymbol{I}_q\big), \\ W_j \mid Z_{ij} = 1 &\sim BS(\alpha_i, 1), \\ \boldsymbol{Z}_j &\sim \mathcal{M}(1; \pi_1, \ldots, \pi_g). \end{aligned} \tag{13}$$

Proposition 3.1. *Under conceptualized representation (13), we have:*

i) The distribution of common factors $\tilde{\boldsymbol{U}}_{ij}$ given $(\boldsymbol{y}_j, w_j, Z_{ij} = 1)$ is $N_q(\boldsymbol{q}_{ij}, w_j \boldsymbol{C}_i)$, where

$$\begin{aligned} &\boldsymbol{q}_{ij} = \boldsymbol{C}_i \left\{ \boldsymbol{\xi}_{ij} + \boldsymbol{\lambda}_i (w_j - a_{\alpha_i}) \right\}, \quad \boldsymbol{\xi}_{ij} = \tilde{\boldsymbol{B}}_i^{\top} \boldsymbol{D}_i^{-1} (\boldsymbol{y}_j - \boldsymbol{\mu}_i), \\ &\boldsymbol{C}_i = (\boldsymbol{I}_q + \tilde{\boldsymbol{B}}_i^{\top} \boldsymbol{D}_i^{-1} \tilde{\boldsymbol{B}}_i)^{-1}. \end{aligned}$$

ii) The conditional distribution of W_j *given* $(\boldsymbol{y}_j, Z_{ij} = 1)$ *is obtained as:*

$$f(w_j \mid \boldsymbol{y}_j, Z_{ij} = 1) = \tau_{ij} f_{GIG}\left(w_j; \frac{1-p}{2}, \chi_{ij}, \psi_i\right) + (1-\tau_{ij}) f_{GIG}\left(w_j; \frac{-1-p}{2}, \chi_{ij}, \psi_i\right), \tag{14}$$

where

$$\chi_{ij} = (\boldsymbol{y}_j - \boldsymbol{\mu}_i + a_{\alpha_i}\boldsymbol{\eta}_i)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_j - \boldsymbol{\mu}_i + a_{\alpha_i}\boldsymbol{\eta}_i) + \alpha_i^{-2}, \quad \psi_i = \boldsymbol{\eta}_i^\top \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\eta}_i + \alpha_i^{-2},$$

$$\tau_{ij} = \frac{f_{GH_p}\left(\boldsymbol{y}_j; \boldsymbol{\mu}_i - a_{\alpha_i}\boldsymbol{\eta}_i, \boldsymbol{\eta}_i, \boldsymbol{\Sigma}_i, 0.5, \alpha_i^{-2}, \alpha_i^{-2}\right)}{f_{GH_p}\left(\boldsymbol{y}_j; \boldsymbol{\mu}_i - a_{\alpha_i}\boldsymbol{\eta}_i, \boldsymbol{\eta}_i, \boldsymbol{\Sigma}_i, 0.5, \alpha_i^{-2}, \alpha_i^{-2}\right) + f_{GH_p}\left(\boldsymbol{y}_j; \boldsymbol{\mu}_i - a_{\alpha_i}\boldsymbol{\eta}_i, \boldsymbol{\eta}_i, \boldsymbol{\Sigma}_i, -0.5, \alpha_i^{-2}, \alpha_i^{-2}\right)},$$

iii) Consequently forms (i) *and* (ii)*, we have*

$$E(W_j^r \mid \boldsymbol{y}_j) = \left(\frac{\chi_{ij}}{\psi_i}\right)^{r/2} \left\{ \frac{K_{(1-p)/2+r}\left(\sqrt{\psi_i\chi_{ij}}\right)}{K_{(1-p)/2}\left(\sqrt{\psi_i\chi_{ij}}\right)} + \frac{K_{-(1+p)/2+r}\left(\sqrt{\psi_i\chi_{ij}}\right)}{K_{-(1+p)/2}\left(\sqrt{\psi_i\chi_{ij}}\right)} \right\}, \; r = \pm 1,$$

$$E(\tilde{\boldsymbol{U}}_{ij} \mid \boldsymbol{y}_j) = \boldsymbol{C}_i \left\{ \boldsymbol{\xi}_{ij} + \boldsymbol{\lambda}_i \left(E(W_j \mid \boldsymbol{y}_j) - a_{\alpha_i} \right) \right\},$$

$$E(W_j^{-1}\tilde{\boldsymbol{U}}_{ij} \mid \boldsymbol{y}_j) = \boldsymbol{C}_i \left\{ \boldsymbol{\xi}_{ij} E(W_j^{-1} \mid \boldsymbol{y}_j) + \boldsymbol{\lambda}_i \big(1 - a_{\alpha_i} E(W_j^{-1} \mid \boldsymbol{y}_j)\big) \right\},$$

and

$$E(W_j^{-1}\tilde{\boldsymbol{U}}_{ij}\tilde{\boldsymbol{U}}_{ij}^\top \mid \boldsymbol{y}_j) = \Big\{ E(W_j^{-1}\tilde{\boldsymbol{U}}_{ij} \mid \boldsymbol{y}_j)\boldsymbol{\xi}_{ij}^\top + \big[E(\tilde{\boldsymbol{U}}_{ij} \mid \boldsymbol{y}_j) - a_{\alpha_i} E(W_j^{-1}\tilde{\boldsymbol{U}}_{ij} \mid \boldsymbol{y}_j)\big]\boldsymbol{\lambda}_i^\top + \boldsymbol{I}_q \Big\} \boldsymbol{C}_i. \tag{15}$$

Proof. See Hahsemi et al. [15]. □

3.2 Identifiability issue

Before developing an EM-type algorithm for the ML parameter estimation of (7), it is important to examine the model identifiability. A finite mixture model is identifiable if two sets of parameters cannot yield the same mixture distribution, that is, $f(\boldsymbol{y}; \boldsymbol{\Theta}_1) = f(\boldsymbol{y}; \boldsymbol{\Theta}_2)$ for all $\boldsymbol{y} \in \mathbb{R}^p$ implying that $\boldsymbol{\Theta}_1 = \boldsymbol{\Theta}_2$. Naderi et al. [21] proved the identifiability of finite mixtures of NMVBS distributions based on earlier work by Browne and McNicholas [3]. By using Naderi et al. [21], we provide a sufficient condition for the identifiability of the PM-NMVBSFA (7) in Theorem 3.2 for the most general UUU structure. Based on Hashemi et al. [15], we add

the restriction that $\boldsymbol{B}_i^\top \boldsymbol{D}_i^{-1}\boldsymbol{B}_i$ is a diagonal matrix with its elements arranged in descending order of magnitude. A sufficient condition for the identifiability of class $\mathcal{C}$ is established where

$$\mathcal{C} = \Big\{ f(\boldsymbol{y};\boldsymbol{\Theta}) : f(\boldsymbol{y};\boldsymbol{\Theta}) = \sum_{i=1}^{g} \pi_i f_{NMVBS}(\boldsymbol{y}_j, \boldsymbol{\mu}_i - a_{\alpha_i}\boldsymbol{\eta}_i, \boldsymbol{\Sigma}_i, \boldsymbol{\eta}_i, \alpha_i),$$
$$\text{with } \pi_i > 0; \sum_{i=1}^{g} \pi_i = 1,\ ||(\boldsymbol{\mu}_r - a_{\alpha_r}\boldsymbol{\eta}_r) - (\boldsymbol{\mu}_l - a_{\alpha_l}\boldsymbol{\eta}_l)||, \text{for} r \neq l \in$$
$$\boldsymbol{B}_i^\top \boldsymbol{D}_i^{-1}\boldsymbol{B}_i \text{is a diagonal matrix},\ \boldsymbol{\mu}_l \neq \boldsymbol{\mu}_r, \boldsymbol{\lambda}_l \neq \boldsymbol{\lambda}_r, \boldsymbol{B}_l \neq \boldsymbol{B}_r, \boldsymbol{D}_l \neq \boldsymbol{D}_r, \alpha_l \neq \alpha_r \Big\}.$$

Theorem 3.2. *Suppose that pdf of the PM-NMVBSFA for the general UUU structure is parameterized in a different way as:*

$$f(\boldsymbol{y};\boldsymbol{\Theta}) = \sum_{r=1}^{g} \pi_r f_{NMVBS}(\boldsymbol{y}, \boldsymbol{\mu}_r - a_{\alpha_i}\boldsymbol{\eta}_r, \boldsymbol{\Sigma}_r, \boldsymbol{\eta}_r, \alpha_r),\ \textit{and}$$
$$f(\boldsymbol{y};\tilde{\boldsymbol{\Theta}}) = \sum_{l=1}^{\tilde{g}} \pi_l f_{NMVBS}(\boldsymbol{y}, \tilde{\boldsymbol{\mu}}_l - a_{\tilde{\alpha}_l}\tilde{\boldsymbol{\eta}}_l, \tilde{\boldsymbol{\Sigma}}_l, \tilde{\boldsymbol{\eta}}_l, \tilde{\alpha}_l).$$

For the considered PM-NMVBSFA model in (7), if condition $\mathcal{C}$ is hold, then the equality $f(\boldsymbol{y};\boldsymbol{\Theta}) = f(\boldsymbol{y};\tilde{\boldsymbol{\Theta}})$ implies that $g = \tilde{g}$, and there are $r, l \in \{1, 2, \ldots, g\}$ such that $\boldsymbol{\mu}_l = \tilde{\boldsymbol{\mu}}_r, \boldsymbol{\lambda}_l = \tilde{\boldsymbol{\lambda}}_r, \boldsymbol{B}_l = \tilde{\boldsymbol{B}}_r, \boldsymbol{D}_l = \tilde{\boldsymbol{D}}_r$, and $\alpha_l = \tilde{\alpha}_r$.

Proof. We first note that the finite mixture of our considered special cases of the NMVBS distributions is identifiable, see, for instance, Naderi et al. [21] for the identifiability issue of the finite mixture of the NMVBS distributions. Now, if conditions class $\mathcal{C}$ are fulfilled and the proof can be obtained directly from the proof of Theorem 1 Dang et al. [22]. □

3.3 Parameter estimation via the ECM algorithm

In this section, the parameter estimation of the PM-NMVBSFA model is carried out via an expectation conditional maximization (ECM) algorithm. Meng and Rubin [18] considered the ECM algorithm as an extension of the EM algorithm [17]. Several properties of this algorithm include stable features, implementation simplicity, and monotone convergence.

To simplify notation, we denote the complete data by $\boldsymbol{y}_c = (\boldsymbol{y}, \tilde{\boldsymbol{U}}, \boldsymbol{W}, \boldsymbol{Z})$, where $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$, $\tilde{\boldsymbol{U}} = (\tilde{\boldsymbol{U}}_1, \ldots, \tilde{\boldsymbol{U}}_n)^\top$, $\boldsymbol{W} = (W_1, \ldots, W_n)$ and $\boldsymbol{Z} = (\boldsymbol{z}_1^\top, \ldots, \boldsymbol{z}_n^\top)^\top$. From (13), the log-likelihood function of $\boldsymbol{\Theta}$ for $\boldsymbol{y}_c$, without con-

sidering constant terms is

$$\ell_c(\boldsymbol{\Theta};\boldsymbol{y}_c)=\sum_{i=1}^{g}\sum_{j=1}^{n}z_{ij}\Big\{\log\pi_i-\log\alpha_i-\frac{(W_j-1)^2}{2\alpha_i^2W_j}-\frac{1}{2}\log|\boldsymbol{D}_i|-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{D}_i^{-1}\boldsymbol{\Upsilon}_{ij}\right)$$
$$-\frac{1}{2}\Big((W_j-2a_{\alpha_i}+W_j^{-1}a_{\alpha_i}^2)\boldsymbol{\lambda}_i\boldsymbol{\lambda}_i^\top-2\boldsymbol{\lambda}_i(\tilde{\boldsymbol{U}}_{ij}-a_{\alpha_i}W_j^{-1}\tilde{\boldsymbol{U}}_{ij})^\top+W_j^{-1}\tilde{\boldsymbol{U}}_{ij}\tilde{\boldsymbol{U}}_{ij}^\top\Big)\Big\}, \tag{16}$$

where $\boldsymbol{\Upsilon}_{ij}=W_j^{-1}\big(\boldsymbol{y}_j-\boldsymbol{\mu}_i-\tilde{\boldsymbol{B}}_i\tilde{\boldsymbol{U}}_{ij}\big)\big(\boldsymbol{y}_j-\boldsymbol{\mu}_i-\tilde{\boldsymbol{B}}_i\tilde{\boldsymbol{U}}_{ij}\big)^\top$.

After that, as conditional expectations, we have the following:

$$\begin{aligned}
\hat{z}_{ij}^{(k)}&=E(Z_{ij}\mid\boldsymbol{y}_j;\hat{\boldsymbol{\Theta}}^{(k)}),\quad \hat{w}_{ij}^{(k)}=E(W_j\mid Z_{ij}=1,\boldsymbol{y}_j,\hat{\boldsymbol{\Theta}}^{(k)}),\\
\hat{\boldsymbol{\zeta}}_{0ij}^{(k)}&=E(\tilde{\boldsymbol{U}}_{ij}\mid Z_{ij}=1,\boldsymbol{y}_j,\hat{\boldsymbol{\Theta}}^{(k)}),\quad \hat{\boldsymbol{\zeta}}_{1ij}^{(k)}=E(W_i^{-1}\tilde{\boldsymbol{U}}_{ij}\mid Z_{ij}=1,\boldsymbol{y}_j,\hat{\boldsymbol{\Theta}}^{(k)}),\\
\hat{\boldsymbol{\Omega}}_{ij}^{(k)}&=E(W_i^{-1}\tilde{\boldsymbol{U}}_{ij}\tilde{\boldsymbol{U}}_{ij}^\top\mid Z_{ij}=1,\boldsymbol{y}_j,\hat{\boldsymbol{\Theta}}^{(k)}),\quad \hat{t}_{ij}^{(k)}=E(W_j^{-1}\mid Z_{ij}=1,\boldsymbol{y}_j,\hat{\boldsymbol{\Theta}}^{(k)}).
\end{aligned} \tag{17}$$

Now, the ML parameter estimation of the PM-NMVBSFA model via the ECM algorithm proceeds as follows:

- E-step: At the kth iteration, the so-called Q-function defined as the expected value of complete data log-likelihood (16) concerning the conditional distribution of $\mathcal{S}$ given the observed data $\boldsymbol{y}_j$ evaluated at $\Theta=\hat{\Theta}^{(k)}$, is

$$\begin{aligned}
Q(\boldsymbol{\theta}\mid\hat{\boldsymbol{\theta}}^{(k)})=&\sum_{i=1}^{g}\sum_{j=1}^{n}\hat{z}_{ij}^{(k)}\Big\{\log\pi_i-\log\alpha_i-\frac{1}{2\alpha_i^2}(\hat{w}_{ij}^{(k)}-2+\hat{t}_{ij}^{(k)})-\frac{1}{2}\log|\boldsymbol{D}_i|\\
&-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{D}_i^{-1}\boldsymbol{\Upsilon}_{ij}^{(k)}\right)-\frac{1}{2}\Big((\hat{w}_{ij}^{(k)}-2a_{\alpha_i}+a_{\alpha_i}^2\hat{t}_{ij}^{(k)})\boldsymbol{\lambda}_i\boldsymbol{\lambda}_i^\top\\
&-2\boldsymbol{\lambda}_i^\top(\hat{\boldsymbol{\zeta}}_{0ij}^{(k)}-a_{\alpha_i}\hat{\boldsymbol{\zeta}}_{1ij}^{(k)})+\hat{\boldsymbol{\Omega}}_{ij}^{(k)}\Big)\Big\},
\end{aligned} \tag{18}$$

  where

$$\begin{aligned}
\boldsymbol{\Upsilon}_{ij}^{(k)}=&\hat{t}_{ij}^{(k)}(\boldsymbol{y}_j-\boldsymbol{\mu}_i)(\boldsymbol{y}_j-\boldsymbol{\mu}_i)^\top-\tilde{\boldsymbol{B}}_i\hat{\boldsymbol{\zeta}}_{1ij}^{(k)}(\boldsymbol{y}_j-\boldsymbol{\mu}_i)^\top\\
&-(\boldsymbol{y}_j-\boldsymbol{\mu}_i)\hat{\boldsymbol{\zeta}}_{1ij}^{(k)}\tilde{\boldsymbol{B}}_i^\top+\tilde{\boldsymbol{B}}_i\hat{\boldsymbol{\Omega}}_j^{(k)}\tilde{\boldsymbol{B}}_i^\top,
\end{aligned} \tag{19}$$

- CM-step 1: The CM-steps obtained by maximizing (18) concerning $\boldsymbol{\mu}_i,\boldsymbol{\lambda}_i,\boldsymbol{B}_i$

and $\boldsymbol{D}_i$ proceed as follows:

$$\hat{\pi}_i = \frac{n_i}{n}, \qquad \hat{\boldsymbol{\mu}}_i^{(k+1)} = \frac{\sum_{j=1}^{n} \hat{z}_{ij}^{(k)} (\hat{t}_{ij}^{(k)} \boldsymbol{y}_j - \hat{\tilde{\boldsymbol{B}}}_i^{(k)} \hat{\boldsymbol{\zeta}}_{1ij}^{(k)})}{\sum_{j=1}^{n} \hat{z}_{ij}^{(k)} \hat{t}_{ij}^{(k)}},$$

$$\hat{\boldsymbol{\lambda}}_i^{(k+1)} = \frac{\sum_{j=1}^{n} \hat{z}_{ij}^{(k)} \left(\hat{\tilde{\boldsymbol{\zeta}}}_{0ij}^{(k)} - a_{\hat{\alpha}_i^{(k)}} \hat{\tilde{\boldsymbol{\zeta}}}_{1ij}^{(k)} \right)}{\sum_{j=1}^{n} \hat{z}_{ij}^{(k)} \left(\hat{w}_{ij}^{(k)} - 2a_{\hat{\alpha}_i^{(k)}} + a_{\hat{\alpha}_i^{(k)}}^2 \hat{t}_{ij}^{(k)} \right)},$$

$$\hat{\tilde{\boldsymbol{B}}}_i^{(k+1)} = \left(\sum_{j=1}^{n} \hat{z}_{ij}^{(k)} (\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_i^{(k+1)}) \hat{\tilde{\boldsymbol{\zeta}}}_{1ij}^{(k)\top} \right) \left(\sum_{j=1}^{n} \hat{z}_{ij}^{(k)} \hat{\boldsymbol{\Omega}}_{ij}^{(k)} \right)^{-1},$$

$$\hat{\boldsymbol{D}}_i^{(k+1)} = \frac{1}{\hat{n}_i^{(k)}} \mathrm{Diag} \left(\sum_{j=1}^{n} \hat{z}_{ij}^{(k)} \hat{\boldsymbol{\Upsilon}}_{ij}^{(k+1/2)} \right),$$

where $\hat{\boldsymbol{\Upsilon}}_{ij}^{(k+1/2)}$ is obtained by substituting $\hat{\boldsymbol{\mu}}_i^{(k+1)}$ and $\hat{\tilde{\boldsymbol{B}}}_i^{(k+1)}$ into (19).

Then, the factor loading can be calculated by $\hat{\boldsymbol{B}}_i^{(k+1)} = \hat{\tilde{\boldsymbol{B}}}_i^{(k+1)} \hat{\boldsymbol{\Lambda}}^{(k+1)}$ where $\hat{\boldsymbol{\Lambda}}^{(k+1)} = a_{\hat{\alpha}_i^{(k+1)}} \boldsymbol{I}_q + b_{\hat{\alpha}_i^{(k+1)}} \hat{\boldsymbol{\lambda}}^{(k+1)} \hat{\boldsymbol{\lambda}}^{(k+1)\top}$ that $\hat{\alpha_i}^{(k)}$ evaluated in the next CM step. Also, we have the following estimates for different cases of covariance structures:

- If $\boldsymbol{B}_i = \boldsymbol{B}$, then

$$\hat{\tilde{\boldsymbol{B}}}^{(k+1)} = \left(\sum_{i=1}^{g} \sum_{j=1}^{n} \hat{z}_{ij}^{(k)} (\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_i^{(k+1)}) \hat{\tilde{\boldsymbol{\zeta}}}_{1ij}^{(k)} \right) \left(\sum_{i=1}^{g} \sum_{j=1}^{n} \hat{z}_{ij}^{(k)} \hat{\boldsymbol{\Omega}}_{ij}^{(k)} \right)^{-1}.$$

- If $\boldsymbol{D}_i = d_i \boldsymbol{I}_p$, then

$$\hat{d}_i^{(k+1)} = \frac{1}{pn_i} \mathrm{tr} \left(\sum_{j=1}^{n} \hat{z}_{ij}^{(k)} \hat{\boldsymbol{\Upsilon}}_{ij}^{(k+1/2)} \right).$$

- If $\boldsymbol{D}_i = \boldsymbol{D}$, then

$$\hat{\boldsymbol{D}}^{(k+1)} = \frac{1}{n} \mathrm{diag} \left(\sum_{i=1}^{g} \sum_{j=1}^{n} \hat{z}_{ij}^{(k)} \hat{\boldsymbol{\Upsilon}}_{ij}^{(k+1/2)} \right).$$

- If $\boldsymbol{D}_i = d \boldsymbol{I}_p$, then

$$\hat{d}^{(k+1)} = \frac{1}{pn} \mathrm{tr} \left(\sum_{i=1}^{g} \sum_{j=1}^{n} \hat{z}_{ij}^{(k)} \hat{\boldsymbol{\Upsilon}}_{ij}^{(k+1/2)} \right).$$

- CM-step 2: The update of α_i can be gained by maximizing $Q(\boldsymbol{\Theta} \mid \hat{\boldsymbol{\Theta}}^{(k)})$ or equivalently by solving the root of the following equation:

$$\sum_{j=1}^{n} \hat{z}_{ij}^{(k)} \Big\{ -\frac{1}{\alpha_i} + \frac{1}{\alpha_i^3}(\hat{w}_{ij}^{(k)} - 2 + \hat{t}_{ij}^{(k)}) \\ -\frac{1}{2}\Big(2\alpha_i a_{\alpha_i}\hat{t}_{ij}^{(k)} - 2\alpha_i)\boldsymbol{\lambda}_i\boldsymbol{\lambda}_i^{\top} + 2\alpha_i\boldsymbol{\lambda}_i^{\top}\hat{\boldsymbol{\zeta}}_{1ij}^{(k)}\Big)\Big\} = 0.$$

  Moreover, when α_is are assumed to be the same, say $\alpha_i = \alpha$ with $i = 1, \ldots, g$, the estimator of α is evaluated using functions such as `uniroot` in R programming by

$$\sum_{i=1}^{g}\sum_{j=1}^{n} \hat{z}_{ij}^{(k)} \Big\{ -\frac{1}{\alpha} + \frac{1}{\alpha^3}(\hat{w}_{ij}^{(k)} - 2 + \hat{t}_{ij}^{(k)}) \\ -\frac{1}{2}\Big((2\alpha a_{\alpha_i}\hat{t}_{ij}^{(k)} - 2\alpha)\boldsymbol{\lambda}_i\boldsymbol{\lambda}_i^{\top} + 2\alpha\boldsymbol{\lambda}_i^{\top}\hat{\boldsymbol{\zeta}}_{1ij}^{(k)}\Big)\Big\} = 0.$$

For calculating and predicting factor scores and missing information, denote the ML estimates by $\hat{\boldsymbol{\Theta}} = (\hat{\pi}_1, \cdots, \hat{\pi}_{g-1}, \hat{\boldsymbol{\theta}}_1, \cdots, \hat{\boldsymbol{\theta}}_g)$ where $\hat{\boldsymbol{\theta}}_i = (\hat{\mu}_i, \hat{\boldsymbol{\lambda}}_i, \hat{\boldsymbol{B}}_i, \hat{\boldsymbol{D}}_i, \hat{\alpha}_i)$. The estimator of factor scores can be evaluated as follows:

$$\hat{\boldsymbol{u}}_{ij} = E(\boldsymbol{U}_{ij} \mid \boldsymbol{y}_j, Z_{ij} = 1, \hat{\boldsymbol{\Theta}}) = \hat{\boldsymbol{\Lambda}}_i^{-1/2} E(\tilde{\boldsymbol{U}}_{ij} \mid \boldsymbol{y}_j, Z_{ij} = 1, \hat{\boldsymbol{\Theta}}), \tag{20}$$

where $\hat{\boldsymbol{\Lambda}}_i$ and $E(\tilde{\boldsymbol{U}}_{ij} \mid \boldsymbol{y}_j, Z_{ij} = 1, \hat{\boldsymbol{\Theta}})$ are calculated using $\boldsymbol{\Lambda}_i = b_{\alpha_i}\boldsymbol{I}_q + b_{\alpha_i}\boldsymbol{\lambda}_i\boldsymbol{\lambda}_i^{\top}$ and Proposition 3.1, respectively, evaluated at $\hat{\boldsymbol{\Theta}}$. Therefore, we have

$$\hat{\boldsymbol{u}}_j = \sum_{i=1}^{g} \hat{z}_{ij}\hat{\boldsymbol{u}}_{ij}, \quad j = 1, \ldots, n. \tag{21}$$

3.4 Notes on implementation

Like any other EM-type algorithm, if the ECM algorithm is given good parameter estimates, convergence may be speed up or made easier. When the raw data contains missing values, after filling in the missing values of the kth variable with the mean of the corresponding column regardless of the missing values, for π_i initial estimate, we first divide the datasets into g groups using the k-means function in R. Thus, the estimate of π_i parameters is the number of group members i divided by the number of sample (n). Initial value of $\hat{\boldsymbol{\mu}}_i^{(0)}$, $\hat{\boldsymbol{B}}_i^{(0)}$, $\hat{\boldsymbol{D}}_i^{(0)}$ and $\hat{\alpha}_i^{(0)}$ are described in Hashemi et al. [15] datasets including the ith label.

The Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) [23] are measures to select the number of classes and factors. It is calculated as:

$$\text{AIC} = -2\ell_{\max} + 2m, \quad \text{BIC} = -2\ell_{\max} + m\log n,$$

where m is the number of free parameters, and $\ell_{\max}$ is the maximized log-likelihood value.

Models with fewer AIC and BIC values are generally better fitted. To measure the ability of clustering agreement, we employ the misclassification rate (MCR), the correct classification rate (CCR) and adjusted Rand index (ARI, [24]). The model with the highest CCR and ARI and lower MCR score is considered to provide the most reliable classification accuracy.

4. Standard errors estimation

The hierarchical form of UUU model in PM-NMVBSFA regardless of scaling transformation obtained by

$$\begin{aligned}
\boldsymbol{Y}_j \mid \left(\boldsymbol{U}_{ij}, W_j, Z_{ij}=1\right) &\sim N_p(\boldsymbol{\mu}_i + \boldsymbol{B}_i\boldsymbol{U}_{ij}, W_j\boldsymbol{D}_i),\\
\boldsymbol{U}_{ij} \mid (W_j, Z_{ij}=1) &\sim N_q\left((W_j - a_{\alpha_i})\boldsymbol{\Lambda}_i^{-1/2}\boldsymbol{\lambda}_i, W_j\boldsymbol{\Lambda}_i^{-1}\right),\\
W_j \mid Z_{ij}=1 &\sim BS(\alpha_i, 1),\\
\boldsymbol{Z}_j &\sim \mathcal{M}(1;\pi_1,\ldots,\pi_g).
\end{aligned}$$

In this model, the information-based technique is exploited to compute the estimates of the parameter's standard errors. Following Meilijson [25], the Fisher information matrix can be approximated by

$$I_o(\hat{\boldsymbol{\Theta}} \mid \boldsymbol{y}) = \sum_{j=1}^{n} \hat{\boldsymbol{s}}_j \hat{\boldsymbol{s}}_j^{\top}, \tag{22}$$

where

$$\hat{\boldsymbol{s}}_j = E\left(\frac{\ell_c(\boldsymbol{\Theta} \mid \boldsymbol{y}_{cj})}{\partial\boldsymbol{\Theta}} \mid \boldsymbol{y}_j, \hat{\boldsymbol{\Theta}}\right), \tag{23}$$

is the individual score vector corresponding to $\boldsymbol{y}_j$. Moreover,

$$\begin{aligned}
\ell_c(\boldsymbol{\Theta} \mid \boldsymbol{y}_{cj}) \;=\; & \sum_{i=1}^{g} Z_{ij}\Bigg\{ -\log\alpha_i - \frac{1}{2\alpha_i^2}\left(W_j + W_j^{-1} - 2\right) - \frac{1}{2}\log|\boldsymbol{D}_i| + \frac{1}{2}\log|\boldsymbol{\Lambda}_i| \\
& -\frac{1}{2}\mathrm{tr}\left(W_j^{-1}\boldsymbol{D}_i^{-1}(\boldsymbol{y}_j - \boldsymbol{\mu}_i - \boldsymbol{B_i}\boldsymbol{U}_{ij})(\boldsymbol{y}_j - \boldsymbol{\mu}_i - \boldsymbol{B_i}\boldsymbol{U}_{ij})^{\top}\right) \\
& -\frac{1}{2}b_{\alpha}W_j^{-1}\boldsymbol{\lambda}_i^{\top}\boldsymbol{U}_{ij}\boldsymbol{U}_{ij}^{\top}\boldsymbol{\lambda}_i - \frac{1}{2}\left(W_j + W_j^{-1}a_{\alpha_i}^2 - 2a_{\alpha_i}\right)\boldsymbol{\lambda}_i^{\top}\boldsymbol{\lambda}_i \\
& 22 - \frac{1}{2}a_{\alpha_i}W_j^{-1}\boldsymbol{U}_{ij}^{\top}\boldsymbol{U}_{ij} + \boldsymbol{\lambda}_i^{\top}\boldsymbol{\Lambda}_{\boldsymbol{i}}^{1/2}(\boldsymbol{U}_{ij} - a_{\alpha_i}W_j^{-1}\boldsymbol{U}_{ij})\Bigg\},
\end{aligned}$$

denotes the individual complete-data log-likelihood contributed by $\boldsymbol{y}_{cj} = (\boldsymbol{y}_j, \boldsymbol{U}_{ij}, W_j)$.

Proposition 4.1. *The square-root matrix of* $\boldsymbol{\Lambda}_i = a_{\alpha_i}\boldsymbol{I}_q + b_{\alpha_i}\boldsymbol{\lambda}_i\boldsymbol{\lambda}_i^\top = a_{\alpha_i}(\boldsymbol{I}_q + c_\alpha\boldsymbol{\lambda}_i\boldsymbol{\lambda}_i^\top)$ *is*

$$\boldsymbol{\Lambda}_i^{1/2} = \sqrt{a_{\alpha_i}}\boldsymbol{V}\boldsymbol{\Delta}_i^{1/2}\boldsymbol{V}^{-1},$$

where $c_{\alpha_i} = b_{\alpha_i}/a_{\alpha_i}$ *and*

$$\boldsymbol{V}_i = [\mathbf{v}_{i1}\ \mathbf{v}_{i2}\cdots\mathbf{v}_{iq}],$$

with $\mathbf{v}_{id} = (-\lambda_{i,q-d+1}/\lambda_{i,1},\ \underbrace{0\ldots0}_{q-d-1\ times}\ 1\ \underbrace{0\ldots0}_{d-1\ times})^\top$ *for* $d = 1,\ldots,q-1$, $\mathbf{v}_q = (\lambda_{i,1}/\lambda_{i,q}, \lambda_{i,2}/\lambda_{i,q},\ldots,\lambda_{i,q-1}/\lambda_{i,q}, 1)^\top$ *being eigenvectors of* $\boldsymbol{\Lambda}_i$, *and* $\boldsymbol{\lambda}_i = (\lambda_{i,1},\ldots,\lambda_{i,q})^\top$. *Moreover,* $\boldsymbol{\Delta}_i$ *is a diagonal matrix which has the eigenvalues of* $\boldsymbol{\Lambda}_i$ *along the main diagonal. Specifically,*

$$\boldsymbol{\Delta}_i = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & 1 & 0 \\ 0 & \cdots & 0 & 1 + c_{\alpha_i}\|\boldsymbol{\lambda}_i\|^2 \end{pmatrix}.$$

Proof. The proof is straightforward and hence is omitted. □

Let $\boldsymbol{b}_i = \mathrm{vec}(\boldsymbol{B}_i)$ denote a $pq \times 1$ vector by stacking the column vectors of $\boldsymbol{B}_i$ and $\boldsymbol{d}_i = \mathrm{Diag}(\boldsymbol{D}_i)$ a $p \times 1$ vector including entries on the main diagonal of $\boldsymbol{D}_i$. Using differentiation of the standard matrix, the individual score vector (23) includes the following elements:

$$\hat{\boldsymbol{s}}_{j,\pi_i} = E\left(\frac{\ell_c(\boldsymbol{\Theta} \mid \boldsymbol{y}_{cj})}{\partial\pi_i} \mid \boldsymbol{y}_j, \hat{\boldsymbol{\Theta}}\right) = \frac{\hat{z}_{rj}}{\hat{\pi}_r} - \frac{\hat{z}_{gj}}{\hat{\pi}_g} \quad (r = 1,\ldots,g-1),$$

$$\hat{\boldsymbol{s}}_{j,\boldsymbol{\mu}_i} = E\left(\frac{\ell_c(\boldsymbol{\Theta} \mid \boldsymbol{y}_{cj})}{\partial\boldsymbol{\mu}_i} \mid \boldsymbol{y}_j, \hat{\boldsymbol{\Theta}}\right) = \hat{z}_{ij}\Big\{\hat{\boldsymbol{D}}_i^{-1}\Big\{\hat{t}_{ij}(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_i) - \hat{\boldsymbol{B}}_i\hat{\boldsymbol{\zeta}}_{1ij}\Big\}\Big\},$$

$$\hat{\boldsymbol{s}}_{j,\boldsymbol{b}_i} = E\left(\frac{\ell_c(\boldsymbol{\Theta} \mid \boldsymbol{y}_{cj})}{\partial\boldsymbol{b}_i} \mid \boldsymbol{y}_j, \hat{\boldsymbol{\Theta}}\right) = \hat{z}_{ij}\Big\{\mathrm{vec}\left(\hat{\boldsymbol{D}}_i{}^{-1}\Big\{(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_i)\hat{\boldsymbol{\zeta}}_{1ij} - \hat{\boldsymbol{B}}_i\hat{\boldsymbol{\Omega}}_{ij}\Big\}\right)\Big\},$$

$$\hat{\boldsymbol{s}}_{j,\boldsymbol{d}_i} = E\left(\frac{\ell_c(\boldsymbol{\Theta} \mid \boldsymbol{y}_{cj})}{\partial\boldsymbol{d}} \mid \boldsymbol{y}_j, \hat{\boldsymbol{\Theta}}\right) = \hat{z}_{ij}\Big\{\mathrm{Diag}\left(-\frac{1}{2}\Big\{\hat{\boldsymbol{D}}_i{}^{-1} - \hat{\boldsymbol{D}}_i{}^{-1}\hat{\boldsymbol{\Upsilon}}_j\hat{\boldsymbol{D}}_i{}^{-1}\Big\}\right)\Big\},$$

$$\hat{\boldsymbol{s}}_{j,\boldsymbol{\lambda}_i} = E\left(\frac{\ell_c(\boldsymbol{\Theta} \mid \boldsymbol{y}_{cj})}{\partial\boldsymbol{\lambda}_i} \mid \boldsymbol{y}_j, \hat{\boldsymbol{\Theta}}\right) = \hat{z}_{ij}\Big\{\frac{a_{\hat{\alpha}_i}^{q-1}b_{\hat{\alpha}_i}}{|\hat{\boldsymbol{\Lambda}}_i|}\hat{\boldsymbol{\lambda}}_i - b_{\hat{\alpha}_i}\hat{\boldsymbol{\Omega}}_{ij}\hat{\boldsymbol{\lambda}}_i$$
$$-\left(\hat{w}_{ij} + \hat{t}_{ij}a_{\hat{\alpha}_i}^2 - 2a_{\hat{\alpha}_i}\right)\hat{\boldsymbol{\lambda}}_i + \hat{\boldsymbol{\Lambda}}_i^{\frac{1}{2}}(\hat{\boldsymbol{\zeta}}_{0ij} - a_{\hat{\alpha}_i}\hat{\boldsymbol{\zeta}}_{1ij}) + \hat{\boldsymbol{\delta}}_{ij}\Big\},$$

$$\hat{\boldsymbol{s}}_{j,\alpha_i} = E\left(\frac{\ell_c(\boldsymbol{\Theta} \mid \boldsymbol{y}_{cj})}{\partial \alpha_i} \mid \boldsymbol{y}_j, \hat{\boldsymbol{\Theta}}\right) = \hat{z}_{ij}\Big\{ -\frac{1}{\hat{\alpha}_i} + \frac{1}{\hat{\alpha}_i^3}(\hat{w}_j + \hat{t}_{ij} - 2)$$
$$+ \frac{\hat{K}}{2|\hat{\boldsymbol{\Lambda}}_i|} - a'_{\hat{\alpha}_i}\hat{\boldsymbol{\lambda}}_i^\top \hat{\boldsymbol{\Lambda}}_i^{1/2}\hat{\boldsymbol{\zeta}}_{1ij} - (a'_{\hat{\alpha}_i}a_{\hat{\alpha}_i}\hat{t}_{ij} - a'_{\hat{\alpha}_i})\hat{\boldsymbol{\lambda}}_i^\top\hat{\boldsymbol{\lambda}}_i - \frac{1}{2}b'_{\hat{\alpha}_i}\hat{\boldsymbol{\lambda}}_i^\top\hat{\boldsymbol{\Omega}}_{ij}\hat{\boldsymbol{\lambda}}_i$$
$$- \frac{1}{2}a'_{\hat{\alpha}_i}\hat{\omega}_{ij} + \hat{\boldsymbol{\lambda}}_i^\top\left[\frac{\partial \boldsymbol{\Lambda}_i^{1/2}}{\partial \alpha_i}\right]_{\boldsymbol{\lambda}_i=\hat{\boldsymbol{\lambda}}_i,\alpha_i=\hat{\alpha}_i}(\hat{\boldsymbol{\zeta}}_{0ij} - a_{\hat{\alpha}_i}\hat{\boldsymbol{\zeta}}_{1ij})\Big\},$$

where $\hat{w}_{ij} = E(W_j \mid \boldsymbol{y}_j, \hat{\boldsymbol{\Theta}}, Z_{ij} = 1)$, $\hat{t}_{ij} = E(W_j^{-1} \mid \boldsymbol{y}_j, \hat{\boldsymbol{\Theta}}, Z_{ij} = 1)$, $a'_{\hat{\alpha}_i} = \hat{\alpha}_i$, $b'_{\hat{\alpha}_i} = 2\hat{\alpha}_i + 5\hat{\alpha}_i^3$, $|\hat{\boldsymbol{\Lambda}}_i| = a_{\hat{\alpha}_i}^q + a_{\hat{\alpha}_i}^{q-1}b_{\hat{\alpha}_i}\hat{\boldsymbol{\lambda}}_i^\top\hat{\boldsymbol{\lambda}}_i$,

$\hat{\boldsymbol{\zeta}}_{0ij} = E(\boldsymbol{U}_{ij} \mid \boldsymbol{y}_j, \hat{\boldsymbol{\Theta}}, Z_{ij} = 1) = \hat{\boldsymbol{\Lambda}}_i^{-1/2}E(\tilde{\boldsymbol{U}}_{ij} \mid \boldsymbol{y}_j, \hat{\boldsymbol{\Theta}}, Z_{ij} = 1)$,
$\hat{\boldsymbol{\zeta}}_{1ij} = E(W_j^{-1}\boldsymbol{U}_{ij} \mid \boldsymbol{y}_j, \hat{\boldsymbol{\Theta}}, Z_{ij} = 1) = \hat{\boldsymbol{\Lambda}}_i^{-1/2}E(W_j^{-1}\tilde{\boldsymbol{U}}_{ij} \mid \boldsymbol{y}_j, \hat{\boldsymbol{\Theta}}, Z_{ij} = 1)$,
$\hat{\boldsymbol{\Omega}}_{ij} = E(W_j^{-1}\boldsymbol{U}_{ij}\boldsymbol{U}_{ij}^\top \mid \boldsymbol{y}_j, \hat{\boldsymbol{\Theta}}, Z_{ij} = 1) = \hat{\boldsymbol{\Lambda}}_i^{-1/2}E(W_j^{-1}\tilde{\boldsymbol{U}}_{ij}\tilde{\boldsymbol{U}}_{ij}^\top \mid \boldsymbol{y}_j, \hat{\boldsymbol{\Theta}}, Z_{ij} = 1)\hat{\boldsymbol{\Lambda}}_i^{-1/2}$,
$\hat{\omega}_{ij} = E(W_j^{-1}\boldsymbol{U}_{ij}^\top\boldsymbol{U}_{ij} \mid \boldsymbol{y}_j, \hat{\boldsymbol{\Theta}}, Z_{ij} = 1) = \mathrm{tr}(\hat{\boldsymbol{\Omega}}_j)$,
$\hat{K}_i = \left[\partial|\boldsymbol{\Lambda}_i|/\partial\alpha_i\right]_{\boldsymbol{\lambda}_i=\hat{\boldsymbol{\lambda}}_i,\alpha_i=\hat{\alpha}_i} = qa'_{\hat{\alpha}_i}a_{\hat{\alpha}_i}^{q-1} + \left\{a_{\hat{\alpha}_i}^{q-1}b'_{\hat{\alpha}_i} + (q-1)a'_{\hat{\alpha}_i}a_{\hat{\alpha}_i}^{q-2}b_{\hat{\alpha}_i}\right\}\hat{\boldsymbol{\lambda}}_i^\top\hat{\boldsymbol{\lambda}}_i$,

and $\hat{\boldsymbol{\delta}}_{ij} = (\hat{\delta}_{1ij}, \ldots, \hat{\delta}_{qij})^\top$ with

$$\hat{\delta}_{dij} = \hat{\boldsymbol{\lambda}}_i^\top\left[\frac{\partial \boldsymbol{\Lambda}_i^{1/2}}{\partial \lambda_d}\right]_{\boldsymbol{\lambda}_i=\hat{\boldsymbol{\lambda}}_i,\alpha_i=\hat{\alpha}_i}(\hat{\boldsymbol{\zeta}}_{0ij} - a_{\hat{\alpha}_i}\hat{\boldsymbol{\zeta}}_{1ij}), \quad \text{for} \quad d = 1, \ldots, q.$$

To compute $\partial\boldsymbol{\Lambda}_i^{1/2}/\partial\alpha_i$ and $\partial\boldsymbol{\Lambda}_i^{1/2}/\partial\lambda_d$ for $d = 1, \ldots, q$, Proposition 4.1 and the chain rule can be applied to gain

$$\frac{\partial \boldsymbol{\Lambda}_i^{1/2}}{\partial \alpha_i} = \frac{\partial\left[\sqrt{a_{\alpha_i}}\left(I_q + c_{\alpha_i}\boldsymbol{\lambda}_i\boldsymbol{\lambda}_i^\top\right)^{1/2}\right]}{\partial \alpha_i} = \frac{\alpha_i}{2a_{\alpha_i}}\boldsymbol{\Lambda}_i^{1/2} + \sqrt{a_{\alpha_i}}\boldsymbol{V}_i\frac{\boldsymbol{\Delta}_i^{1/2}}{\partial \alpha_i}\boldsymbol{V}_i^{-1},$$

and

$$\frac{\partial \boldsymbol{\Lambda}_i^{1/2}}{\partial \lambda_d} = \sqrt{a_{\alpha_i}}\left[\frac{\partial \boldsymbol{V}_i}{\partial \lambda_i}\boldsymbol{\Delta}_i^{1/2}\boldsymbol{V}_i^{-1} + \boldsymbol{V}_i\frac{\boldsymbol{\Delta}_i^{1/2}}{\partial \lambda_d}\boldsymbol{V}_i^{-1} - \boldsymbol{V}_i\boldsymbol{\Delta}_i^{1/2}\boldsymbol{V}_i^{-1}\frac{\partial \boldsymbol{V}_i}{\partial \lambda_d}\boldsymbol{V}_i^{-1}\right].$$

As a result, the standard errors of $\hat{\boldsymbol{\Theta}}$ is obtained as the diagonal elements of the square roots (22).

5. Simulations

5.1 Asymptotic properties of the ML estimations

In this experiment, we obtain the performance of the proposed ML estimations based on the ECM algorithm to regain the true parameters for PM-NMVBSFA.

We generate $M = 500$ artificial samples from model (7) with $q = 1$, $g = 2$ and structure UUU under different sample size $n = 100, 200, 500, 1000, 2000$. The common true parameters of these models include

$$\pi_1 = 0.4, \quad \pi_2 = 0.6, \quad \alpha_1 = 0.5, \ \alpha_2 = 1, \ \lambda_1 = -1, \ \lambda_2 = 2$$
$$\boldsymbol{\mu}_1 = (2, 3, 5, 3, 1)^\top, \ \boldsymbol{\mu}_2 = (12, 13, 15, 13, 11)^\top, \ \boldsymbol{B}_1 = (4, 5, 2, 4, 6)^\top,$$
$$\boldsymbol{B}_2 = (3, 5, 3, 4, 7)^\top, \ \boldsymbol{D}_1 = \text{diag}(1, 2, 3, 4, 5)^\top, \ \boldsymbol{D}_2 = \text{diag}(2, 3, 4, 5, 6)^\top.$$

The ECM algorithm is used to fit each simulated dataset under the UUU structure. We computed the average values and the corresponding standard deviations of the ML estimates across all Monte Carlo samples.

The relative absolute bias (RBias) and the root mean squared error (RMSE) were computed for the estimation accuracy measurement:

$$\text{RBias} = \frac{1}{M}\sum_{i=1}^{M}\left|\frac{\hat{\theta}_i - \theta_{true}}{\hat{\theta}_i}\right| \ \text{ and } \ \text{RMSE} = \sqrt{\frac{\sum_{i=1}^{M}(\hat{\theta}_i - \bar{\hat{\theta}})^2}{M}},$$

where $\hat{\theta}_i$ denotes the ML estimate of a specific parameter at the ith replication and θ_{true} is its true value.

In addition, investigating the standard error's estimation consistency is interest. So, by using the observed information matrix (IMSE), we measured the sample standard deviation of parameters (STD) and the average standard errors:

$$\begin{aligned} \text{STD}(\hat{\theta}_i) &= \sqrt{\frac{1}{M-1}\left[\sum_{r=1}^{M}\left(\hat{\theta}_i^{(r)}\right)^2 - \frac{1}{M}\left(\sum_{r=1}^{M}\hat{\theta}_i^{(r)}\right)^2\right]} \ \text{ and } \ \text{IMSE}(\hat{\theta}_i) \quad (24) \\ &= \frac{1}{M}\sum_{r=1}^{M}\text{SE}(\hat{\theta}_i^{(r)}), \end{aligned}$$

where $\text{SE}(\hat{\theta}_i^{(r)})$ denotes the asymptotic standard errors of $\hat{\theta}_i$ at the rth replication. We examine the accuracies of STD estimators with IMSE as well as n increase for the above factor model using discrepancy measures: sum of absolute deviation of STD with IMSE computed by

$$\text{SAD}(\hat{\theta}_i) = |\text{STD}(\hat{\theta}_i) - \text{IMSE}(\hat{\theta}_i)|$$

where $\text{STD}(\hat{\theta}_i)$ is the standard deviation of $\hat{\theta}_i$ and $\text{IMSE}(\hat{\theta}_i)$ is average standard errors using the observed information matrix for parameter $\hat{\theta}_i$ with sample size n.

Figure 1 shows the means of RBias and RMSE for every parameter in UUU structure of the PM-NMVBSFA model. It should be mentioned that the assumed parameter differs for each model. Based on Figure 1, the simulation experiment's

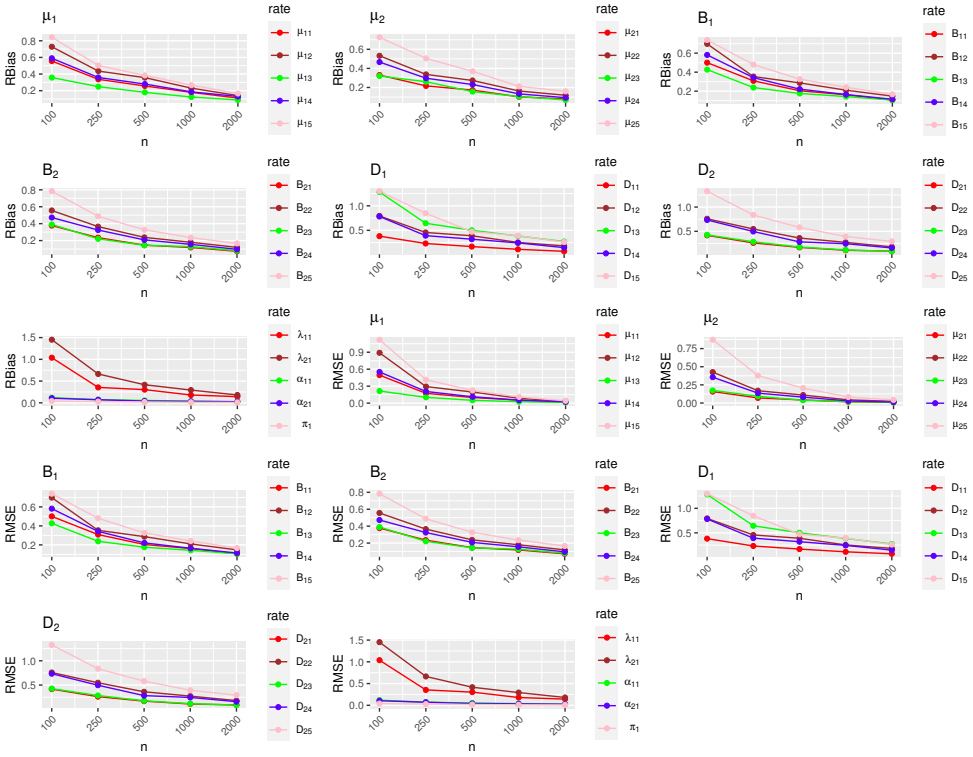

Figure 1: Average RBias and RMSE of parameter estimates in the considered EM-type as a function of sample size.

results show the effectiveness of the proposed ECM algorithm in parameter recovery of all the considered sub-models. Figure 2, depicts that the empirical standard errors (STD) differences and the theoretic results of standard errors (IMSE) tend to get zero as well, revealing the reliability of the proposed asymptotic approximation through (22) to measure the ML estimates standard errors.

5.2 The performance of the factor models via heavy-tailed data

In this simulation study, we examine the performance of the proposed model based on heavy-tailed data in terms of clustering quality. These generated data sets are simulated from a two-component normal inverse Gaussian factor model (NIGFA).

$$\boldsymbol{Y}_j = \boldsymbol{\mu}_i + \boldsymbol{B}_i \boldsymbol{U}_{ij} + \boldsymbol{\varepsilon}_{ij} \quad \text{with probability } \pi_i \ (i = 1, \ldots, g), \tag{25}$$

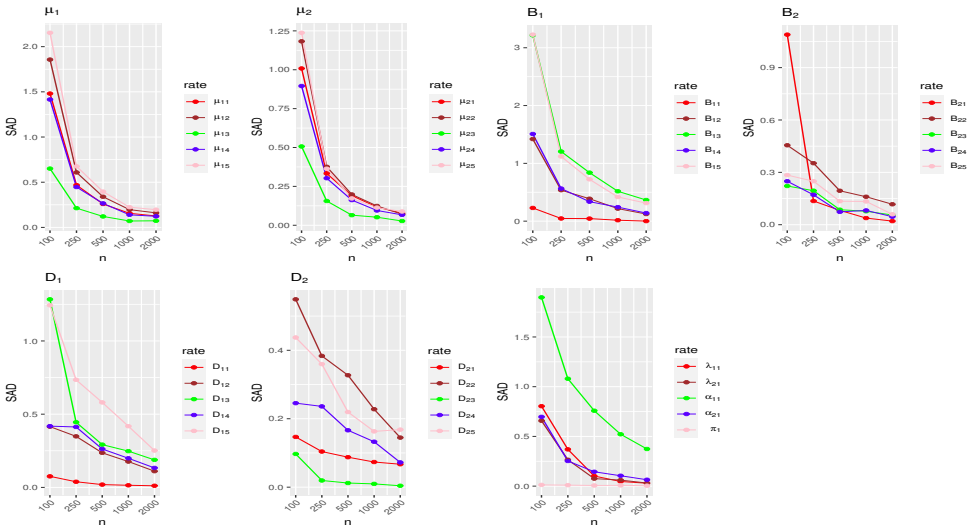

Figure 2: Value of SAD Criterion for parameter estimates in the considered EM-type as a function of sample size.

where

$$\begin{bmatrix} \boldsymbol{U}_{ij} \\ \boldsymbol{\varepsilon}_{ij} \end{bmatrix} \sim NIG_{q+p}\left(\begin{bmatrix} -a_{\theta_i}\boldsymbol{\Lambda}_i^{-1/2}\boldsymbol{\lambda}_i \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Lambda}_i^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D}_i \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Lambda}_i^{-1/2}\boldsymbol{\lambda}_i \\ \boldsymbol{0} \end{bmatrix}, \chi_i, \psi_i\right), \tag{26}$$

where $\boldsymbol{\Lambda}_i = a_{\boldsymbol{\theta}_i}\boldsymbol{I}_q + b_{\boldsymbol{\theta}_i}\boldsymbol{\lambda}_i\boldsymbol{\lambda}_i^\top$, $a_{\boldsymbol{\theta}_i} = \sqrt{\chi_i/\psi_i}$, $b_{\boldsymbol{\theta}_i} = \chi_i/\psi_i \frac{K_{1.5}(\sqrt{\chi_i\psi_i})}{K_{-0.5}(\sqrt{\chi_i\psi_i})} - a^2_{\boldsymbol{\theta}_i}$, and $NIG_p(\boldsymbol{\mu}_i, \boldsymbol{\lambda}_i, \boldsymbol{\Sigma}_i, \chi_i, \psi_i)$ denotes the p-variate NIG distribution, obtained as a special case of GH distribution with $\kappa = -0.5$. The four parsimonious mixtures of NIGFA (PM-NIGFA) models generated in these simulations are: CUU, UCU, UUC and UUU. The true parameters model for any structures are as follows:

$$\pi_1 = 0.35,\ \ \pi_2 = 0.65,\ \ \chi_1 = 4,\ \ \chi_2 = 6,\ \ \psi_1 = 11,\ \ \psi_2 = 1,\ \ \boldsymbol{\lambda}_1 = (1,2)^\top,$$
$$\boldsymbol{\lambda}_2 = (5,2)^\top,\ \ \boldsymbol{\mu}_1 = (10,11,12,13,14)^\top,\ \ \boldsymbol{\mu}_2 = (18,16,16,17,17)^\top,$$

and

- for CUU:

$$\boldsymbol{B}_1 = \boldsymbol{B}_2 = \begin{pmatrix} 3 & 3 & 3 & 4 & 5 \\ 2 & 4 & 6 & 0 & 0 \end{pmatrix}^\top,\ \boldsymbol{D}_1 = \text{diag}(1,2,3,4,5)^\top,$$
$$\boldsymbol{D}_2 = \text{diag}(2,3,4,5,6)^\top,$$

- for UCU:

$$\boldsymbol{B}_1 = \begin{pmatrix} 3 & 3 & 3 & 4 & 5 \\ 2 & 4 & 6 & 0 & 0 \end{pmatrix}^\top, \ \boldsymbol{B}_2 = \begin{pmatrix} 2 & 2 & 2 & 3 & 4 \\ 1 & 3 & 5 & 2 & 2 \end{pmatrix}^\top,$$
$$\boldsymbol{D}_1 = \boldsymbol{D}_2 = \text{diag}(1,2,3,4,5)^\top,$$

- for UUC:

$$\boldsymbol{B}_1 = \begin{pmatrix} 3 & 3 & 3 & 4 & 5 \\ 2 & 4 & 6 & 0 & 0 \end{pmatrix}^\top, \ \boldsymbol{B}_2 = \begin{pmatrix} 2 & 2 & 2 & 3 & 4 \\ 1 & 3 & 5 & 2 & 2 \end{pmatrix}^\top,$$
$$\boldsymbol{D}_1 = \text{diag}(1,1,1,1,1)^\top, \ \boldsymbol{D}_2 = \text{diag}(2,2,2,2,2)^\top,$$

- for UUU:

$$\boldsymbol{B}_1 = \begin{pmatrix} 3 & 3 & 3 & 4 & 5 \\ 2 & 4 & 6 & 0 & 0 \end{pmatrix}^\top, \ \boldsymbol{B}_2 = \begin{pmatrix} 2 & 2 & 2 & 3 & 4 \\ 1 & 3 & 5 & 2 & 2 \end{pmatrix}^\top,$$
$$\boldsymbol{D}_1 = \text{diag}(1,2,3,4,5)^\top, \ \boldsymbol{D}_2 = \text{diag}(2,3,4,5,6)^\top.$$

We generate $n = 400$, PM-NIGFA models with the above true parameters for four structures. For each of the 100 replications, we fitted a full structure parsimonious mixture of factor analysis (PM-FA), the parsimonious mixture of t factor analysis (PM-tFA), the parsimonious mixture of skew-normal factor analysis (PM-SNFA), parsimonious mixture of skew-t factor analysis (PM-STFA) and PM-NMVBSFA with $g = \{1,2,3,4\}$ and $q = 2$. For every structure and model, the number of selected components based on the lowest BIC is $g = 2$. Figure 3 shows a heat map of the average of BIC values for every structure and model based on 100 replications. The comparison of the average of model's BIC values, shows that the BIC is lower for the PM-NMVBSFA in every structure generated based on PM-NIGFA. Once again, the ARI of all structures in five models is given in heat map Figure 4 and shows that the model with the lower BIC also has better clustering performance.

6. Analysis of the Sonar data

Sonar, navigation and distance detection by sound, is a technology that can identify other vessels or ships by using underwater sound emission. Active sonar works by generating sound pulses (known as pings), and then listening for the return pulse. To determine the distance from the target, one can measure the time between receiving and sending the pulse. To measure the direction and alignment of the target, multiple hydrophones can be used, and then the time of receiving the pulse by each of these hydrophones is measured by comparing these times, the direction and alignment of the target can be easily determined. Sonar performance

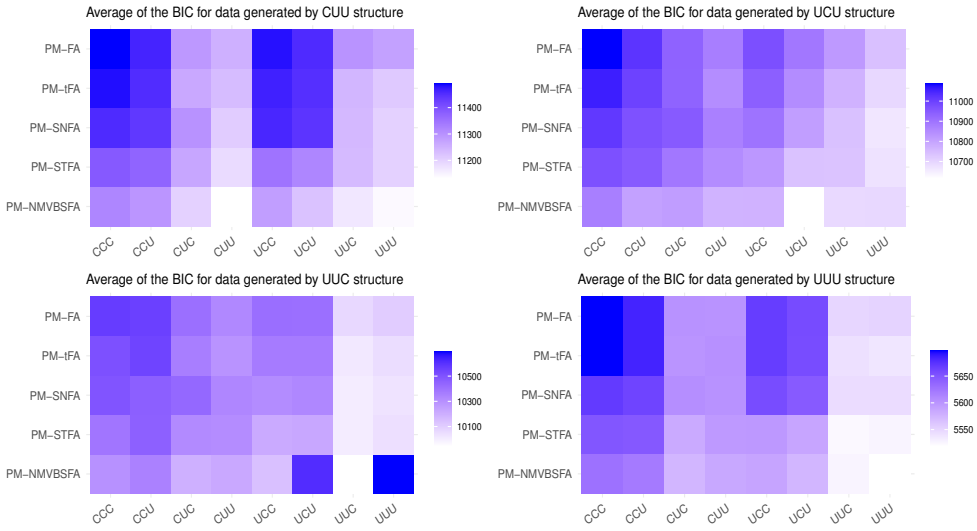

Figure 3: A heat map representation of the average BIC value for all eight structures in each model.

depends on the speed of sound. The speed of sound in freshwater is slower than the speed of sound in seawater. In all waters, the speed of sound depends on the density of the water. Density depends on parameters such as temperature, water salts (usually water salinity) and pressure.

The performance of active sonar is similar to radar. A sound pulse is sent, then the sound waves start moving in all directions. When these waves hit the ground, the incident waves are reflected in all directions and some of the reflected signals reach the active sonar sensor. These reflected signals enable sonar technicians to identify parameters such as signal frequency, signal energy, depth, water temperature, and as a result, the location of the target.

To illustrate our methodology, we consider the Sonar datasets studied by Gorman and Sejnowski [26]. This dataset is available at the UCI Machine Learning Repository *(https://www. kaggle.com/code/martinhewing/uci-sonar-dataset)*. There are $n = 208$ inventories with $p = 60$ continuous attributes between 0 and 1, denoted by $X_1, \ldots, X_{60}$, that contain measurements on 111 bouncing sonar signals off a metal cylinder and 97 bouncing signals off rocks subsets. Figure 5 shows range of skewness and kurtosis for 60 attributes of Sonar data. The results depicted in Figure 5 show that for the considered data most of the attributes are moderately skewed.

To study the performance of the considered model, we implement PM-N, PM-T, PM-SNFA and PM-STFA to compare with the PM-NMVBSFA model with the number of latent factors q ranged from 2 to 7 for $g = 1, \ldots, 5$ to fit the Sonar

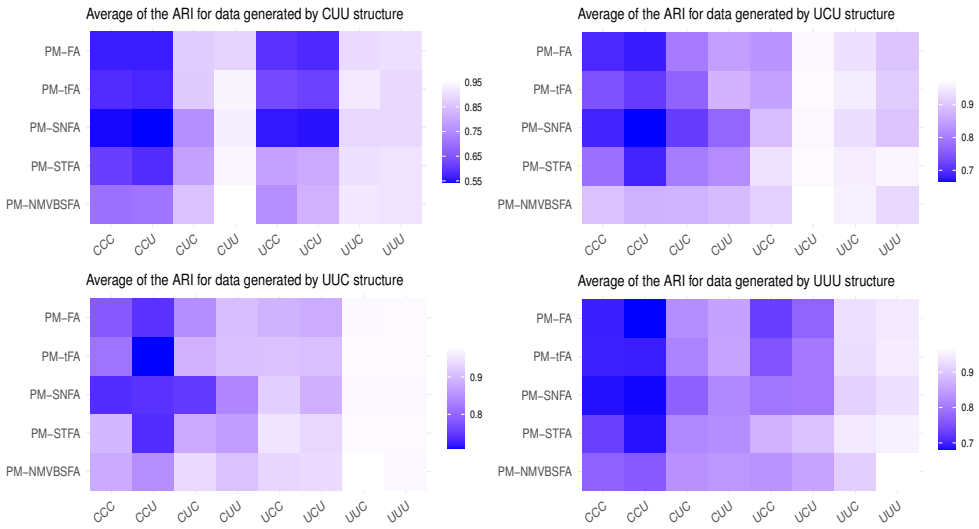

Figure 4: A heat map representation of the average ARI value for all eight structures in each model.

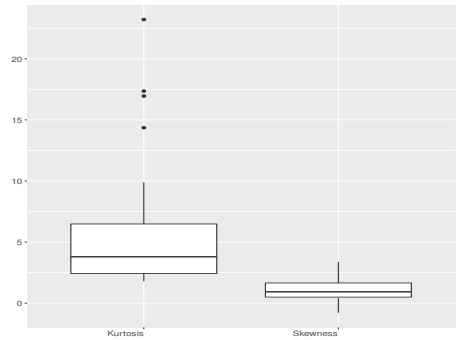

Figure 5: Range of skewness and kurtosis 60 attributes Sonar data.

data datasets. For every model, the number of selected components based on the lowest BIC was $g = 2$. Figure 6 shows the minimum BIC for the eight structures for all models based on each pair (g, q). The BIC of all eight structures in every model was compared to the sonar data for $g = 2$ and $q = 2, \ldots, 7$ in the heat map Figure 7. The model with the lowest BIC (-35307.75) was selected; this was the CUU structure of the PM-NMVBSFA model with $g = 2$ and $q = 7$. Table 2 shows 10 first attribute parameters for the 2 clusters as well as the standard errors for the selected model, we do not show other attributes for the sake of space.

Table 2: Summary of some parameters for the sonar data as well as its standard errors.

	$\boldsymbol{\mu}_1$		$\boldsymbol{B}_1$														$\boldsymbol{D}_1$	
	Par.	SE	Par.	SE	Par.	SE	Par.	SE	Par.	SE	Par.	SE	Par.	SE	Par.	SE	Par.	SE
$i=1$	-0.336	0.030	0.136	-0.084	0.060	-0.172	-0.346	-0.009	0.163	0.019	0.037	0.013	0.014	0.003	0.009	0.022	0.292	0.089
	-0.397	0.036	0.204	-0.048	0.060	-0.215	-0.466	-0.029	0.175	0.021	0.032	0.021	0.019	0.004	0.012	0.015	0.240	0.082
	-0.341	0.036	0.143	0.048	0.020	-0.210	-0.493	0.020	0.198	0.009	0.027	0.021	0.019	0.006	0.009	0.004	0.180	0.092
	-0.353	0.033	0.133	0.036	0.028	-0.108	-0.360	0.014	0.239	0.001	0.026	0.014	0.010	0.004	0.010	0.010	0.199	0.094
	-0.294	0.024	-0.048	-0.049	-0.095	-0.140	-0.314	0.004	0.167	0.010	0.025	0.022	0.006	0.002	0.002	0.011	0.299	0.101
	-0.366	0.025	-0.133	-0.204	0.101	-0.089	-0.282	0.161	0.079	0.004	0.049	0.009	0.003	0.004	0.007	0.001	0.336	0.099
	-0.395	0.024	-0.173	-0.226	0.166	-0.166	-0.237	0.206	0.062	0.002	0.061	0.012	0.007	0.013	0.000	0.004	0.317	0.087
	-0.360	0.027	-0.116	-0.307	0.248	-0.237	-0.324	0.228	0.165	0.017	0.083	0.005	0.009	0.026	0.014	0.017	0.155	0.049
	-0.240	0.009	-0.228	-0.542	0.247	-0.194	-0.240	0.267	0.125	0.021	0.100	0.005	0.001	0.027	0.018	0.032	0.120	0.024
	-0.141	0.007	-0.293	-0.531	0.210	-0.251	-0.428	0.213	0.174	0.019	0.102	0.004	0.003	0.027	0.017	0.031	0.069	0.023

$\pi_1 = 0.549(\text{SE} = 0.023),\ \alpha_1 = 0.667(\text{SE} = 0.142),\ \lambda_1 = (1.259, -0.764, -0.458, 0.926, -1.781, -0.787, 1.168)^\top$
SE for $\boldsymbol{\lambda}_1 = (0.428, 0.253, 0.196, 0.336, 0.569, 0.303, 0.311)$

	$\boldsymbol{\mu}_2$		$\boldsymbol{B}_2$														$\boldsymbol{D}_2$	
	Par.	SE	Par.	SE	Par.	SE	Par.	SE	Par.	SE	Par.	SE	Par.	SE	Par.	SE	Par.	SE
$i=2$	0.365	0.045	0.298	-0.190	0.242	-0.251	-0.370	0.042	0.248	0.019	0.037	0.013	0.014	0.003	0.009	0.022	0.754	0.024
	0.378	0.052	0.397	-0.173	0.282	-0.312	-0.487	0.032	0.276	0.021	0.032	0.021	0.019	0.004	0.012	0.015	0.754	0.019
	0.343	0.050	0.291	-0.047	0.205	-0.290	-0.497	0.070	0.275	0.009	0.027	0.021	0.019	0.006	0.009	0.004	0.830	0.013
	0.394	0.047	0.265	-0.050	0.187	-0.176	-0.368	0.057	0.309	0.001	0.026	0.014	0.010	0.004	0.010	0.010	0.814	0.016
	0.417	0.034	-0.012	-0.073	-0.023	-0.169	-0.290	0.023	0.186	0.010	0.025	0.022	0.006	0.002	0.002	0.011	0.996	0.129
	0.481	0.036	-0.038	-0.269	0.208	-0.135	-0.299	0.192	0.131	0.004	0.049	0.009	0.003	0.004	0.007	0.001	0.976	0.129
	0.456	0.036	-0.055	-0.306	0.282	-0.219	-0.275	0.241	0.126	0.002	0.061	0.012	0.007	0.013	0.000	0.004	0.758	0.225
	0.445	0.043	0.102	-0.454	0.461	-0.333	-0.392	0.290	0.283	0.017	0.083	0.005	0.009	0.026	0.014	0.017	0.429	0.014
	0.286	0.029	-0.014	-0.689	0.449	-0.285	-0.314	0.328	0.241	0.021	0.100	0.005	0.001	0.027	0.018	0.032	0.157	0.007
	0.166	0.027	-0.089	-0.671	0.423	-0.347	-0.479	0.275	0.284	0.019	0.102	0.004	0.003	0.027	0.017	0.031	0.204	0.006

$\pi_2 = 0.451(\text{SE} = 0.023),\ \alpha_2 = 0.701(\text{SE} = 0.145),\ \lambda_2 = (1.260, -1.824, 2.094, -1.502, -0.482, 0.322, 1.672)^\top$
SE for $\boldsymbol{\lambda}_1 = (0.523, 0.489, 0.608, 0.419, 0.108, 0.079, 0.725)$

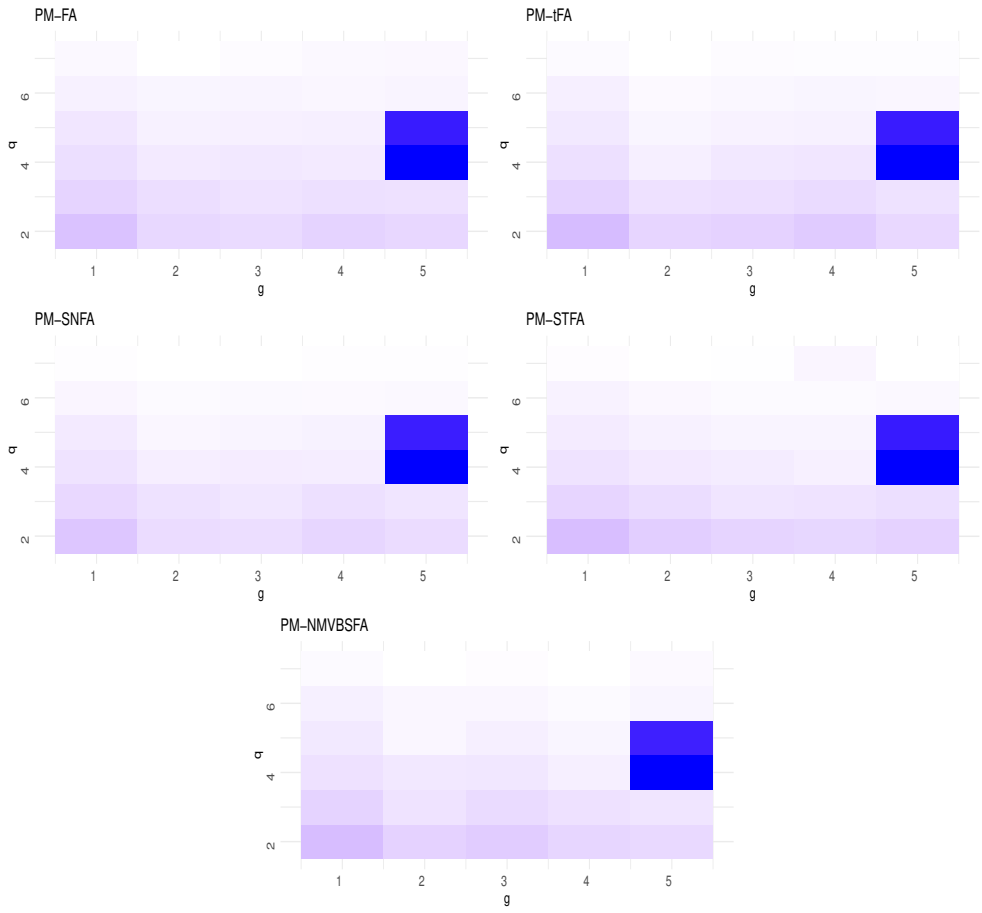

Figure 6: A heat map representation of the minimum BIC value for each value of (g, q) in sonar data sets, where the minimum is taken over the eight models.

The results of clustering selected structures based on Figure 7 are compared in Table 3. The clustering performance of PM-NMVBSFA (ARI = 0.477) is better than that of PM-STFA (ARI = 0.425). According to Table 3 and Figure 7, the CUU structure of PM-NMVBSFA is the best model to fit and the best clustering accuracy for this dataset.

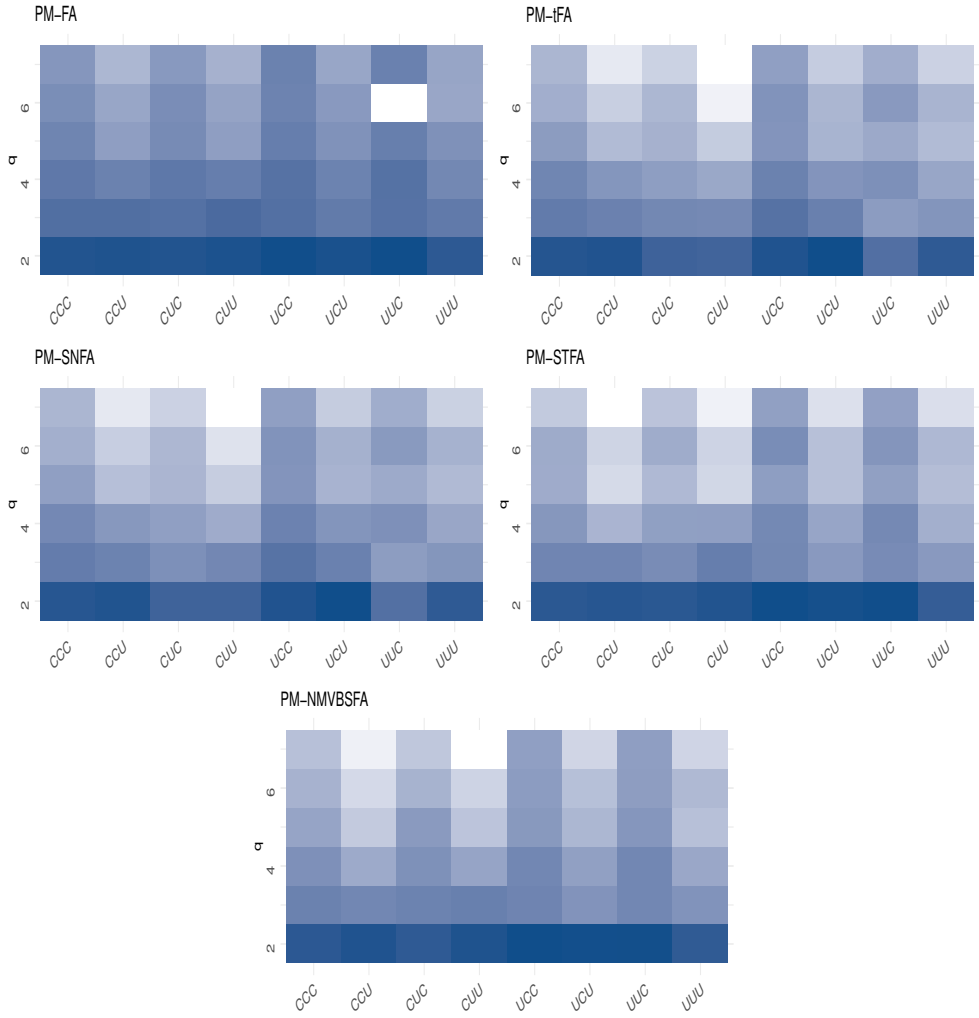

Figure 7: A heat map representation of the BIC value for each models with $g = 2$ for sonar data sets.

Table 3: Performance estimation for factor models fitted to the sonar data.

	PM-FA		PM-tFA		PM-SNFA		PM-STFA		PM-NMVBSFA	
Selected model	$q=6$	UUC	$q=7$	CUU	$q=7$	CUU	$q=7$	CCU	$q=7$	UUC
	1	2	1	2	1	2	1	2	1	2
Rock	15	82	11	86	85	12	88	9	91	6
Metal	38	73	31	80	33	78	29	82	27	84

7. Concluding remarks

A computationally feasible ECM algorithm is developed to estimate PM-NMVBSFA model parameters under eight structures covariance matrix to accommodate asymmetric shapes and heavy tails data. The performance of the proposed finite mixture models has been investigated by model-based clustering using two Monte Carlo simulation studies and a real data set. Numerical results reveal that the PM-NMVBSFA model outperforms other competing models based on model fitting and outright clustering when data contain and exhibit non-normal features such as multimodality, asymmetry, and heavy-tailed data.

There are a few issues as well as possible modifications related to the proposed methodology that deserve further attention. As has been indicated in these models, its PM-NMVBSFA under a missing-information framework as an extended tool to accommodate incomplete data can be challenged. We have recently focused our work on these subjects and it is expected to present the findings in future papers.

Conflicts of Interest. The authors declare that they have no conflicts of interest regarding the publication of this article.

References

[1] G. Celeux and G. Govaert, Gaussian parsimonious clustering models, *Pattern Recognit.* **28** (1995) 781−793, https://doi.org/10.1016/0031-3203(94)00125-6.

[2] J. L. Andrews and P. D. McNicholas, Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions, *Stat. Comput.* **22** (2012) 1021 − 1029, https://doi.org/10.1007/s11222-011-9272-x.

[3] R. P. Browne and P. D. McNicholas, A mixture of generalized hyperbolic distributions, *Canad. J. Statist.* **43** (2015) 176 − 198, https://doi.org/10.1002/cjs.11246.

[4] P. D. McNicholas, *Mixture Model-Based Classification*, Taylor and Francis Group, 2020.

[5] Z. Ghahramani and G. E. Hinton, *The EM Algorithm for Mixtures of Factor Analyzers*, In: Technical Report CRG-TR-96-1. University of Toronto, 1996.

[6] P. M. Murray, R. P. Browne and P. D. McNicholas, Mixtures of skew-t factor analyzers, *Comput. Statist. Data Anal.* **77** (2014) 326 − 335, https://doi.org/10.1016/j.csda.2014.03.012.

[7] T.-I. Lin, W.-L. Wang, G. J. McLachlan and S. X. Lee, Robust mixtures of factor analysis models using the restricted multivariate skew-t distribution, *Stat. Model.* **18** (2018) 50 − 72, https://doi.org/10.1177/1471082X17718119.

[8] Y. Wei, Y. Tang and P. D. McNicholas, Flexible high-dimensional unsupervised learning with missing data, *IEEE Trans. Pattern Anal. Mach. Intell.* **42** (2018) 610 − 621, https://doi.org/10.1109/TPAMI.2018.2885760.

[9] S. X. Lee, T.-I. Lin and G. J. McLachlan, Mixtures of factor analyzers with scale mixtures of fundamental skew normal distributions, *Adv. Data Anal. Classif.* **15** (2021) 481 − 512, https://doi.org/10.1007/s11634-020-00420-9.

[10] W.-L. Wang and T.-I. Lin, Model-based clustering via mixtures of unrestricted skew normal factor analyzers with complete and incomplete data, *Stat. Methods Appl.* **32** (2023) 787 − 817, https://doi.org/10.1007/s10260-022-00674-x.

[11] A. Azzalini, A class of distributions which includes the normal ones, *Scand. J. Statist.* **12** (1985) 171 − 178.

[12] T.-I. Lin, G. J. McLachlan and S. X. Lee, Extending mixtures of factor models using the restricted multivariate skew-normal distribution, *J. Multivariate Anal.* **143** (2016) 398 − 413, https://doi.org/10.1016/j.jmva.2015.09.025.

[13] P. D. McNicholas and T. B. Murphy, Parsimonious gaussian mixture models, *Stat. Comput.* **18** (2008) 285−296, https://doi.org/10.1007/s11222-008-9056-0.

[14] P. M. Murray, P. D. McNicholas and R. P. Browne, A mixture of common skew-t factor analysers, *Stat.* **3** (2014) 68 − 82, https://doi.org/10.1002/sta4.43.

[15] F. Hashemi, M. Naderi, A. Jamalizadeh and T.-I. Lin, A skew factor analysis model based on the normal mean–variance mixture of Birnbaum-Saunders distribution, *J. Appl. Stat.* **47** (2020) 3007 − 3029, https://doi.org/10.1080/02664763.2019.1709054.

[16] F. Hashemi, M. Naderi, A. Jamalizadeh and A. Bekker, A flexible factor analysis based on the class of mean-mixture of normal distributions, *Comput. Statist. Data Anal.* **157** (2021) #107162, https://doi.org/10.1016/j.csda.2020.107162.

[17] A. P. Dempster, N. M. Laird and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *J. Roy. Statist. Soc. Ser. B* **39** (1977) 1 − 22, https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.

[18] X.-L. Meng and D. B. Rubin, Maximum likelihood estimation via the ECM algorithm: a general framework, *Biometrika* **80** (1993) 267 − 278, https://doi.org/10.1093/biomet/80.2.267.

[19] Z. W. Birnbaum and S. C. Saunders, A new family of life distributions, *J. Appl. Prob.* **6** (1969) 319 − 327, https://doi.org/10.2307/3212003.

[20] O. Barndorff-Nielsen and C. Halgreen, Infinite divisibility of the hyperbolic and generalized inverse gaussian distributions, *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **38** (1977) 309−311, https://doi.org/10.1007/BF00533162.

[21] M. Naderi, W. L. Hung, T. I. Lin and A. Jamalizadeh, A novel mixture model using the multivariate normal mean–variance mixture of Birnbaum–Saunders distributions and its application to extrasolar planets, *J. Multivariate Anal.* **171** (2019) 126 − 138, https://doi.org/10.1016/j.jmva.2018.11.015.

[22] U. J. Dang, A. Punzo, P. D. McNicholas, S. Ingrassia and R. P. Browne, Multivariate response and parsimony for Gaussian cluster-weighted models, *J. Classif.* **34** (2017) 4 − 34, https://doi.org/10.1007/s00357-017-9221-2.

[23] G. Schwarz, Estimating the dimension of a model, *Ann. Statist.* **6** (1978) 461 − 464.

[24] L. Hubert and P. Arabie, Comparing partitions, *J. Classif.* **2** (1985) 193−218, https://doi.org/10.1007/BF01908075.

[25] I. Meilijson, A fast improvement to the EM algorithm on its own terms, *J. Roy. Statist. Soc. Ser. B* **51** (1989) 127−138, https://doi.org/10.1111/j.2517-6161.1989.tb01754.x.

[26] R. P. Gorman and T. J. Sejnowski, Analysis of hidden units in a layered network trained to classify sonar targets, *Neur. Net.* **1** (1988) 75 − 89, https://doi.org/10.1016/0893-6080(88)90023-8.

Appendix A. The GH and the generalized inverse Gaussian

The GH distribution of [20] showed its applicability in modeling skewed data, despite its complicated form. Its probability density function (pdf) has the form

$$f_{GH_p}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\lambda},\boldsymbol{\Sigma},\kappa,\chi,\psi) = C\frac{K_{\kappa-\frac{p}{2}}\left(\sqrt{\left(\psi+\boldsymbol{\lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda}\right)\left(\chi+(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)}\right)}{\left(\sqrt{\left(\psi+\boldsymbol{\lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda}\right)\left(\chi+(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)}\right)^{\frac{p}{2}-\kappa}} \times \exp\left\{(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda}\right\}, \quad \boldsymbol{x}\in\mathbb{R}^p, \tag{27}$$

$\boldsymbol{\mu},\ \boldsymbol{\lambda} \in \mathbb{R}^p$, $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ space of $p \times p$ positive definite matrices, where

$$C = (\psi/\chi)^{\kappa/2}(\psi + \boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda})^{p/2-\kappa}/(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}K_\kappa(\sqrt{\psi\chi}),$$

is the normalizing constant. We shall write $\boldsymbol{X} \sim GH_p(\boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\Sigma}, \kappa, \chi, \psi)$ to indicate that the random vector $\boldsymbol{X}$ has the pdf (27). Specifically, W follows the generalized inverse Gaussian (GIG) distribution with pdf

$$f_{GIG}(w; \kappa, \chi, \psi) = \left(\frac{\psi}{\chi}\right)^{\kappa/2} \frac{w^{\kappa-1}}{2K_\kappa(\sqrt{\psi\chi})} \exp\left\{\frac{-1}{2}\left(w^{-1}\chi + w\psi\right)\right\}, \quad w > 0.$$

Farzane Hashemi
Department of Statistics,
University of Kashan,
Kashan, I. R. Iran
e-mail: farzane.hashemi1367@yahoo.com

Jalal Askari
Department of Applied Mathematics,
University of Kashan,
Kashan, I. R. Iran
e-mail: askari@kashanu.ac.ir

Saeed Darijani
Farhangian University Of Kerman,
Kerman, I. R. Iran
e-mail: saeed_darijani@yahoo.com