Journal of Statistical Modelling: Theory and Applications
Vol. 4, No. 2, 2023, pp. 35-56
Yazd University Press 2023

JSMTA

Research Paper

Mixtures of the normal mean-variance of Lindley factor analysis model with missing data

Maryam Darijani[1], Hojatollah Zakerzadeh[1*], Farzane Hashemi[2]
[1]Department of Statistics, Yazd University ,Yazd, Iran
[2]Department of Statistics, University of Kashan, Kashan

Received: Feburary 02, 2024/ Revised: May 17, 2024/ Accepted: May 30, 2024

Abstract: For a heterogeneous community comprised of multiple sub-communities, the model-based clustering method stands out as a suitable recommendation. Moreover, incomplete data collection and information loss may occur due to a variety of causes. In this paper, our focus is on investigating the mixture of factor analysis model in the presence of missing data. Here, the latent factors and errors within each sub-cluster exhibit non-normal characteristics and adhere to the normal mean-variance mixture of Lindley distribution. This model is termed the mixture of normal mean-variance mixture of Lindley factor analysis. To estimate model parameters and generate a single imputation of potential missing values under the missing with random mechanism, we introduce a generalized expectation-maximization algorithm. The number of factors and mixture components are determined by the evaluation criteria of the model. The proposed model's advantage is validated through a real dataset and simulation studies, demonstrating its superior performance compared to existing models.

Keywords: Asymmetry; ECME algorithm; Factor analysis model; Incomplete data; Mixture model; Model-based clustering.
Mathematics Subject Classification (2010): 62H25.

1 Introduction

Spearman (1904) originally introduced the factor analysis (FA) model. When applying the FA model, a reduced number of unobserved variables, referred to as "factors," are expressed as linear combinations of the observed variables. Ghahramani and Hinton

*Corresponding author:hzaker@yazd.ac.ir

(1997) applied the mixture of factor analysis (MFA) model for high-dimensional data clustering. The MFA model has found widespread application in various fields, including medicine (Wall et al., 2012), bioinformatics (McLachlan et al., 2003), behavioral sciences (De Roover et al., 2022), and pattern recognition (Gaarenstroom et al., 1977).

Consider the following the Gaussian MFA model

$$\boldsymbol{Y}_j = \boldsymbol{\mu}_i + \boldsymbol{B}_i \boldsymbol{U}_{ij} + \boldsymbol{\varepsilon}_{ij}, \quad \begin{bmatrix} \boldsymbol{U}_{ij} \\ \boldsymbol{\varepsilon}_{ij} \end{bmatrix} \sim N_{q+p}\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D}_i \end{bmatrix} \right), \; j = 1, \ldots, n, \; i = 1, \ldots, g, \tag{1}$$

where $\boldsymbol{\mu}_i$ is a location vector p-dimensional, $\boldsymbol{B}_i \in \mathbb{R}^{p \times q}$ is the matrix of "factor loadings", $\boldsymbol{U}_{ij}$ is the vector of "factor scores" with $(q < p)$ dimension, $\boldsymbol{\varepsilon}_{ij} \in \mathbb{R}^p$ defines the error vector called "specific factors", and, $\boldsymbol{D}_i$ is a diagonal matrix. For additional information and uses, see Lawley and Maxwell (1962), Joreskog et al. (1979) and Basilevsky (2009). The validity of statistical inference can be impacted by non-normal data because the Gaussian MFA model is not robust to asymmetry, heavy tails, and missing values. Interest in the MFA model based on skewed distributions has grown to address this shortcoming (Lee et al., 2018a,b, 2021). McLachlan et al. (2007), using the multivariate t-distribution introduced the mixture of t-factor analyzers (MTFA) model to increase the validity of the MFA model when the data have heavier tail clusters than the normal distribution. Then, the MFA model was extended based on restricted multivariate skew-normal distribution (rMSN) (Azzalini, 2005) for the component latent factors which is introduced with the title, "mixtures of skew-normal factor analyzers" (MSNFA) (Lin et al., 2016; Maleki and Wraith, 2019), and it was demonstrated that this model is more applicable than MTFA. Also, Murray et al. (2014) developed latent factors and errors in the MFA model based on the skew-t distribution (Azzalini and Capitanio, 2003) which is called mixtures of skew-t factor analyzers (MSTFA). To see the unrestricted and restricted versions of the MSTFA model, refer to Murray et al. (2013) and Lin et al. (2015).

Another family of asymmetric distributions is the generalized hyperbolic (GH) distribution family (Barndorff-Nielsen, 1977), encompassing the skew-Laplace (Arslan, 2010), skew-t, variance-gamma (Fischer et al., 2023), and normal inverse Gaussian (NIG) (Göncü and Yang, 2016) distributions. Certain GH distributions are formulated through the normal mean-variance (NMV) mixture family, where the generalized inverse Gaussian (GIG) distribution serves as the mixing random variable. Consider the following linear random representation

$$\boldsymbol{X} = \boldsymbol{\mu} + W\boldsymbol{\lambda} + \sqrt{W}\boldsymbol{Z}, \tag{2}$$

where $\boldsymbol{Z}$ and W are independent random variables such that $\boldsymbol{Z}$ is a multivariate normal distribution with mean $\boldsymbol{0}$ and variance matrix $\boldsymbol{\Sigma}$, and W is a non-negative random variable with the GIG distribution. Also, $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ are vector of shape and location parameters in R^p, respectively. Then, $\boldsymbol{X}$ follows a p-variate random variable with GH distribution. The normal mean-variance mixture of Lindley (NMVL) distribution, introduced by Naderi et al. (2018), arises when the random variable W in (2) follows the Lindley distribution (Ghitany et al., 2008). In recent years, skewed MFA models based on GH distributions have gained prominence in the literature due to their characteristics, such as heavier tails, and their application to financial and economic data (Murray et al., 2014; Tortora et al., 2016; McNicholas et al., 2017; Hashemi et

al., 2020). Recognizing the substantial impact of missing data on statistical inferences, Wei et al. (2018) extended the generalized hyperbolic factor analysis (GHFA) proposed by Tortora et al. (2016) in the presence of missing data. Additionally, Lin et al. (2018) delved into the study of the restricted multivariate skew-t factor analysis model to enhance the interpretation of relationships between variables in the context of missing data. For further details, refer to Wang (2013, 2015) and Wang et al. (2017).

In this study, we investigate the MFA model with missing data, where the latent factors and errors in each sub-cluster follow the NMVL distribution. The advantage of employing the NMVL distribution lies in its reduced number of input parameters compared to linear combinations of the GH distribution. Additionally, this distribution exhibits a broader range of skewness and kurtosis than competing distributions, and the mixture of normal mean-variance mixture of Lindley factor analysis (MNMVLFA) model performs well in the presence of missing data. Considering the missing at random (MAR) assumption (Rubin, 1976; Little and Rubin, 2019), parameter estimation is conducted through the development of the expectation-maximization-type (EM) algorithm (Dempster et al., 1977), specifically the expectation conditional maximization either (ECME) method (Liu and Rubin, 1994). This method facilitates parameter estimation by effectively handling missing data. To leverage information from the available data, we also provide a forecasting mechanism to generate reasonably estimated values. To enhance the accuracy of calculations, observed and missing vectors are determined using two indicator matrices. To assess the applicability of this method, we examined a real dataset and designed simulation studies.

The structure of this article is as follows. In section 2, we list the MNMVLFA model's symbols and fundamental characteristics. We also look at the issue of the model not being recognized. The MNMVLFA model with missing data is presented in section 3, along with the estimation technique and several of the model's useful features, including initialization, convergence assessment, and performance comparison criteria. Simulation studies are presented in section 4 to assess the current approach used in this study. We assess the benefit of the suggested approach using a real data medical in section 5.

2 Preliminaries

In this section, we offer some required topics and prepare the ground for introducing the MNMVLFA model. As a first step, we must define the notation used in this paper. Let $f_{GH_p}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\lambda},\boldsymbol{\Sigma},\kappa,\chi,\psi)$ be the probability density function (pdf) of a p-dimensional GH distribution introduced by Barndorff-Nielsen (1977) with parameter $\boldsymbol{\mu},\ \boldsymbol{\lambda}\in\mathbb{R}^p$, $\boldsymbol{\Sigma}\in\mathbb{R}^{p\times p}$, $\kappa\in\mathbb{R}$, $\chi,\ \psi>0$ given by

$$f_{GH_p}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\lambda},\boldsymbol{\Sigma},\kappa,\chi,\psi)=C\ \frac{K_{\kappa-p/2}\left(\sqrt{t}\right)}{t^{p-2\kappa}}\exp\left((\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda}\right),\ \ \boldsymbol{x}\in\mathbb{R}^p, \tag{3}$$

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 and $t=\left(\psi+\boldsymbol{\lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda}\right)\left(\chi+(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$. Also $K_\kappa(\cdot)$ interprets the modified Bessel function of the third kind of order κ.

A p-variate NMVL distribution with location vector $\boldsymbol{\mu}$, scale covariance matrix $\boldsymbol{\Sigma}$, skewness vector $\boldsymbol{\lambda}$ and shape parameter α, denoted by $\boldsymbol{X}\sim\ \mathrm{NMVL}_p(\boldsymbol{\mu},\boldsymbol{\lambda},\boldsymbol{\Sigma},\alpha)$ has

the pdf

$$f_{NMVL}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\lambda},\boldsymbol{\Sigma},\alpha) = \frac{\alpha}{1+\alpha} f_{GH_p}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\lambda},\boldsymbol{\Sigma},1,0,2\alpha) + \frac{1}{1+\alpha} f_{GH_p}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\lambda},\boldsymbol{\Sigma},2,0,2\alpha). \quad (4)$$

A two-level hierarchical demonstration from (4) is

$$\boldsymbol{X} \mid W = w \sim \mathrm{N}_p(\boldsymbol{\mu}+\boldsymbol{\lambda}w, w\boldsymbol{\Sigma}), \qquad W \sim \mathrm{Lindley}(\alpha).$$

The pdf of Lindley (α) is expressed as a linear combination of GIG distribution as

$$f_{Lindley}(w;\alpha) = \frac{\alpha}{1+\alpha} f_{GIG}(w;1,0,2\alpha) + \frac{1}{1+\alpha} f_{GIG}(w;2,0,2\alpha),$$

where $f_{GIG}(.)$ is the pdf of GIG distribution (Good, 1953) as

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

The GIG distribution is unimodal. Also, if $\chi = 0$ and $\kappa > 0$, the distribution of the GIG becomes to gamma distribution with parameters κ and $\frac{\psi}{2}$, and if $\psi = 0$ and $\kappa < 0$, the distribution of the GIG becomes to inverse gamma distribution with parameters $-\kappa$ and $\frac{\chi}{2}$.

$\boldsymbol{Y}$ is a p-dimensional random vector of FA model. The NMVLFA model can be written in a matrix-vector form as

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{U} + \boldsymbol{\varepsilon}, \quad (5)$$

along with the assumption

$$\begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{\varepsilon}\end{bmatrix} \sim \mathrm{NMVL}_{q+p}\left(\begin{bmatrix}-a_\alpha\boldsymbol{\Lambda}^{-1/2}\boldsymbol{\lambda}\\ \boldsymbol{0}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Lambda}^{-1} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{D}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{\lambda}\\ \boldsymbol{0}\end{bmatrix}, \alpha\right), \quad (6)$$

that $\boldsymbol{\mu}$, $\boldsymbol{B}$, $\boldsymbol{U}$, $\boldsymbol{\varepsilon}$ and diagonal matrix $\boldsymbol{D} = \mathrm{diag}(\sigma_1^2,\ldots,\sigma_p^2)$, (all arrivals are strictly positive) that called "uniquenesses", were defined in (1). Also

$$W \sim Lindley(\alpha), \quad a_\alpha = E(W) = \frac{\alpha+2}{\alpha(\alpha+1)}, \quad \text{and} \quad b_\alpha = \mathrm{Var}(W) = \frac{\alpha^2+4\alpha+2}{\alpha^2(\alpha+1)^2}.$$

The marginal distribution $\boldsymbol{Y}$ from (5) and (6) is

$$\boldsymbol{Y} \sim \mathrm{NMVL}_p(\boldsymbol{\mu} - a_\alpha\boldsymbol{\delta}, \boldsymbol{\Sigma}, \boldsymbol{\delta}, \alpha), \quad (7)$$

where $\boldsymbol{\delta} = \boldsymbol{B}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{\lambda}^\top$ is a p-dimensional vector of reparameterized shape parameters, and $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{\Lambda}^{-1}\boldsymbol{B}^\top + \boldsymbol{D}$ with $\boldsymbol{\Lambda} = a_\alpha\boldsymbol{I}_q + b_\alpha\boldsymbol{\lambda}\boldsymbol{\lambda}^\top$ to verify the orthogonality of factor loadings. Subsequently, the mean vector and covariance matrix of $\boldsymbol{Y}$ are

$$E(\boldsymbol{Y}) = \boldsymbol{\mu} \quad \text{and} \quad cov(\boldsymbol{Y}) = \boldsymbol{B}\boldsymbol{B}^\top + \frac{\alpha+2}{\alpha(\alpha+1)}\boldsymbol{D}. \quad (8)$$

The rotational invariance of the factor loadings $\boldsymbol{B}$ and the skewness parameter $\boldsymbol{\lambda}$ contribute to the identifiability problem of the NMVLFA model. Assume that an orthogonal matrix of order q is $\boldsymbol{S}$. The marginal distribution in (7) is invariant if $\boldsymbol{B}$ and $\boldsymbol{\lambda}$ are exchanged by $\boldsymbol{BS}$ and $\boldsymbol{S}^\top\boldsymbol{\lambda}$. Furthermore, the covariance structure (8) remains unchanged as a result of these orthogonal modifications. We add $\frac{q(q-1)}{2}$ identifiability limits to the factor loading matrix as a result of these modifications. Numerous approaches have been suggested to address this issue; see to Lopes and West (2004); Bai and Li (2012).

3 Examining aspects of the model

3.1 Incomplete MNMVLFA model

Vector $\boldsymbol{Y} = (\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n)$ is a p-dimensional feature vector with n independent components originating from a heterogeneous community with g subclasses. In a finite mixture model, to determine the density of the observed component $\boldsymbol{Y}_j$, we define an unobserved membership indicator vector $\boldsymbol{Z}_j = (Z_{1j}, \ldots, Z_{gj})$. Therefore, if $\boldsymbol{Y}_j$ belongs to class i, it is assigned the value of one, and the rest are assigned the value of zero.

$$Z_{ij} = \begin{cases} 1 & if\ \boldsymbol{Y}_j \text{ belongs to class } i \\ 0 & \text{otherwise.} \end{cases}$$

Mixing proportions are determined by $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)$ so that $Pr(Z_{ij} = 1) = \pi_i$. It is obvious that

$$\boldsymbol{Z}_j \sim \mathcal{M}(1; \pi_1, \ldots, \pi_g), \qquad \sum_{i=1}^{g} \pi_i = 1,$$

that is the multinomial distribution with pdf

$$f(Z_j; \boldsymbol{\pi}) \propto \pi_1^{z_{1j}} \pi_2^{z_{2j}} \ldots (1 - \pi_1 \ldots - \pi_{g-1})^{z_{gj}}, \qquad \sum_{i=1}^{g} z_{ij} = 1.$$

The MNMVLFA model consists of g sub-models of (6) with mixture ratio $\boldsymbol{\pi}$. In a special case, each $\boldsymbol{Y}_j$ is

$$\boldsymbol{Y}_j = \boldsymbol{\mu}_i + \boldsymbol{B}_i\boldsymbol{U}_{ij} + \boldsymbol{\varepsilon}_{ij} \quad \text{with probability } \pi_i \quad (i = 1, \ldots, g), \tag{9}$$

with the assumption

$$\begin{bmatrix} \boldsymbol{U}_{ij} \\ \boldsymbol{\varepsilon}_{ij} \end{bmatrix} \sim \text{NMVL}_{q+p}\left(\begin{bmatrix} -a_{\alpha_i}\boldsymbol{\Lambda}_i^{-1/2}\boldsymbol{\lambda}_i \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Lambda}_i^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D}_i \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Lambda}_i^{-1/2}\boldsymbol{\lambda}_i \\ \boldsymbol{0} \end{bmatrix}, \alpha_i \right), \tag{10}$$

in which a_{α_i} is a_α when α replaced with α_i and $\boldsymbol{\Lambda}_i = a_{\alpha_i}\boldsymbol{I}_q + b_{\alpha_i}\boldsymbol{\lambda}_i\boldsymbol{\lambda}_i^\top$. According to the model (9), the marginal density of $\boldsymbol{Y}_j$ is

$$f(\boldsymbol{y}_j) = \sum_{i=1}^{g} \pi_i \psi(\boldsymbol{y}_j; \boldsymbol{\theta}_i), \tag{11}$$

in which $\psi(\boldsymbol{y}_j;\boldsymbol{\theta}_i)$ is the NMVL density quoted in (4), $\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \boldsymbol{B}_i, \boldsymbol{D}_i, \boldsymbol{\lambda}_i, \alpha_i)$ shows the ith mixture component of unknown parameters and $\boldsymbol{\Theta} = (\pi_1, \ldots, \pi_{g-1}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g)$ demonstrates the whole unknown parameters. We propose situations where missing values occur for reasons beyond our control. To solve this problem, we formulate this model with missing data in such a way that, $\boldsymbol{Y}_j(p \times 1)$ is partitioned into two subvectors so that $\boldsymbol{y}_j{}^o(p_j{}^o \times 1)$ includes the observed portion of $\boldsymbol{Y}_j$ and $\boldsymbol{y}_j{}^m((p-p_j{}^o)\times 1)$ includes the residual inputs, namely the missing portion of $\boldsymbol{Y}_j$. To simplify calculations, we define two permutation matrices $\boldsymbol{O}_j \in \mathbb{R}^{(p_j^o \times p)}$ and $\boldsymbol{M}_j \in \mathbb{R}^{((p-p_j^o)\times p)}$ so that $\boldsymbol{Y}_j^o = \boldsymbol{O}_j\boldsymbol{Y}_j$ and $\boldsymbol{Y}_j^m = \boldsymbol{M}_j\boldsymbol{Y}_j$. From (11), we have

$$f(\boldsymbol{y}_j^o;\boldsymbol{\Theta}) = \sum_{i=1}^{g} \pi_i \psi(\boldsymbol{y}_j^o;\boldsymbol{\theta}_i).$$

It is easy to see $\boldsymbol{Y}_j = \boldsymbol{O}_j^\top \boldsymbol{Y}_j^o + \boldsymbol{M}_j^\top \boldsymbol{Y}_j^m$ and $\boldsymbol{O}_j\boldsymbol{O}_j^\top + \boldsymbol{M}_j\boldsymbol{M}_j^\top = \boldsymbol{I}_p$. In order to achieve closed-form for the estimators, we use the procedure of Hashemi et al. (2020) by introducing invariant transformations

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

To evaluate the conditional expectation in the E-step for the computational algorithm interpreted in the subsection 3.2, the following theorem is useful.

Theorem 3.1. *From* (7) *and* (9)*, we have*
a. The marginal distribution of the $\boldsymbol{Y}_j^o$'th observation is

$$\boldsymbol{Y}_j^o \mid (Z_{ij} = 1) \sim MNMVL_{p_j^o}(\boldsymbol{\mu}_{ij}^o - a_{\alpha_i}\boldsymbol{\delta}_{ij}^o, \boldsymbol{\Sigma}_{ij}^{oo}, \boldsymbol{\delta}_{ij}^o, \alpha_i). \tag{13}$$

b. The conditional distribution of $\boldsymbol{Y}_j^o$ given w_j is

$$\boldsymbol{Y}_j^o \mid (w_j, Z_{ij} = 1) \sim N_{p_j^o}(\boldsymbol{\mu}_{ij}^o - a_{\alpha_i}\boldsymbol{\delta}_{ij}^o + w_j\boldsymbol{\delta}_{ij}^o, w_j\boldsymbol{\Sigma}_{ij}^{oo}),$$

where $\boldsymbol{\mu}_{ij}^o = \boldsymbol{O}_j\boldsymbol{\mu}_i$, $\boldsymbol{\delta}_{ij}^o = \boldsymbol{O}_j\boldsymbol{\delta}_i$, $\boldsymbol{\delta}_i = \boldsymbol{B}_i\boldsymbol{\Lambda}_i^{-1/2}\boldsymbol{\lambda}_i$ and $\boldsymbol{\Sigma}_{ij}^{oo} = \boldsymbol{O}_j\boldsymbol{\Sigma}_i\boldsymbol{O}_j^\top$.
c. The MNMVLFA model formulated in (9) *has the following hierarchical presentation*

$$\begin{aligned}
\boldsymbol{Y}_j^o \mid (\tilde{\boldsymbol{u}}_{ij}, w_j, Z_{ij} = 1) &\sim N_{p_j^o}(\boldsymbol{\mu}_{ij}^o + \tilde{\boldsymbol{B}}_{ij}^o\tilde{\boldsymbol{u}}_{ij}, w_j\boldsymbol{D}_{ij}^{oo}),\\
\boldsymbol{Y}_j^m \mid (\boldsymbol{Y}_j^o, \tilde{\boldsymbol{u}}_{ij}, w_j, Z_{ij} = 1) &\sim N_{p-p_j^o}(\boldsymbol{\varphi}_{ij}^{m.o}, w_j\boldsymbol{D}_{ij}^{mm.o}),\\
\tilde{\boldsymbol{U}}_{ij} \mid (w_j, Z_{ij} = 1) &\sim N_q((w_j - a_{\alpha_i})\boldsymbol{\lambda}_i, w_j\boldsymbol{I}_q),\\
W_j \mid (Z_{ij} = 1) &\sim \mathit{Lindley}(\alpha_i),\\
\boldsymbol{Z}_j &\sim \mathcal{M}(1;\pi_1,\ldots,\pi_g),
\end{aligned}$$

where $\tilde{\boldsymbol{B}}_{ij}^o = \boldsymbol{O}_j\tilde{\boldsymbol{B}}_i$, $\boldsymbol{\varphi}_{ij}^{m.o} = \boldsymbol{M}_j\left[\boldsymbol{\mu}_i + \tilde{\boldsymbol{B}}_i\tilde{\boldsymbol{U}}_{ij} + \boldsymbol{D}_i\boldsymbol{C}_{ij}^{oo}(\boldsymbol{y}_j - \boldsymbol{\mu}_i - \tilde{\boldsymbol{B}}_i\tilde{\boldsymbol{U}}_{ij})\right]$, $\boldsymbol{D}_{ij}^{oo} = \boldsymbol{O}_j\boldsymbol{D}_i\boldsymbol{O}_j^\top$, $\boldsymbol{D}_{ij}^{mm.o} = \boldsymbol{M}_j\left(\boldsymbol{I}_p - \boldsymbol{D}_i\boldsymbol{C}_{ij}^{oo}\right)\boldsymbol{D}_i\boldsymbol{M}_j^\top$ and, $\boldsymbol{C}_{ij}^{oo} = \boldsymbol{O}_j^\top\left(\boldsymbol{O}_j\boldsymbol{D}_i\boldsymbol{O}_j^\top\right)^{-1}\boldsymbol{O}_j$.
d. Conditional distributions are obtained as

$$\tilde{\boldsymbol{U}}_{ij} \mid (\boldsymbol{y}_j^o, w_j, Z_{ij} = 1) \sim N_q(\boldsymbol{q}_{ij}^o, w_j\boldsymbol{R}_{ij}^{oo}),$$

$$
\begin{aligned}
f(W_j \mid \boldsymbol{y}_j^o, (Z_{ij}=1)) &= \pi_j^o f_{GIG}\left(w_j; 1-\frac{p_j^o}{2}, \chi_{ij}^o, \psi_{ij}^o\right) \\
&\quad +(1-\pi_j^o) f_{GIG}\left(w_j; 2-\frac{p_j^o}{2}, \chi_{ij}^o, \psi_{ij}^o\right), \\
\boldsymbol{Z}_j \mid \boldsymbol{y}_j^o &\sim \mathcal{M}(1; \tilde{\pi}_{1j}, \ldots, \tilde{\pi}_{gj})), \qquad (14)
\end{aligned}
$$

where

$$
\begin{aligned}
\chi_{ij}^o &= (\boldsymbol{y}_j^o - \boldsymbol{\mu}_{ij}^o + a_{\alpha_i}\boldsymbol{\delta}_{ij}^o)^\top \boldsymbol{\Sigma}_{ij}^{oo\,-1}(\boldsymbol{y}_j^o - \boldsymbol{\mu}_{ij}^o + a_{\alpha_i}\boldsymbol{\delta}_{ij}^o), \\
\boldsymbol{\Sigma}_{ij}^{oo} &= \boldsymbol{O}_j \boldsymbol{\Sigma}_i \boldsymbol{O}_j^\top, \\
\psi_{ij}^o &= \boldsymbol{\delta}_{ij}^{o^\top} \boldsymbol{\Sigma}_{ij}^{oo} \boldsymbol{\delta}_{ij}^o + 2\alpha_i, \\
\pi_j^o &= \frac{\alpha_i f_{GH_p}\left(\boldsymbol{y}_j^o; \boldsymbol{\mu}_{ij}^o - a_{\alpha_i}\boldsymbol{\delta}_{ij}^o, \boldsymbol{\delta}_{ij}^o, \boldsymbol{\Sigma}_{ij}^{oo}, 1, 0, 2\alpha_i\right)}{f_{\text{MNMVL}}\left(\boldsymbol{y}_j^o; \boldsymbol{\mu}_{ij}^o - a_{\alpha_i}\boldsymbol{\delta}_{ij}^o, \boldsymbol{\Sigma}_{ij}^{oo}, \boldsymbol{\delta}_{ij}^o, \alpha_i\right)},
\end{aligned}
$$

and $\boldsymbol{q}_{ij}^o = \boldsymbol{R}_{ij}^{oo}\left\{\boldsymbol{b}_{ij}^o + \boldsymbol{\lambda}_i(w_j - a_{\alpha_i})\right\}$, $\boldsymbol{b}_{ij}^o = \tilde{\boldsymbol{B}}_i^\top \boldsymbol{C}_{ij}^{oo}(\boldsymbol{y}_j - \boldsymbol{\mu}_i)$,$\boldsymbol{R}_{ij}^{oo} = (\boldsymbol{I}_q + \tilde{\boldsymbol{B}}_i^\top \boldsymbol{C}_{ij}^{oo}\tilde{\boldsymbol{B}}_i)^{-1}$. *The posterior probability of the* j*th observed feature vector* $\boldsymbol{y}_j^o$ *belongs to component* i *of the mixture is*

$$
\tilde{\pi}_{ij} = E(Z_{ij} \mid \boldsymbol{y}_j^o) = \frac{\pi_i f(\boldsymbol{y}_j^o, \boldsymbol{\theta}_i)}{f(\boldsymbol{y}_j^o, \boldsymbol{\Theta})}.
$$

e. conditional expectations are for $r = \pm 1$

$$
\begin{aligned}
E(W_j^r \mid \boldsymbol{y}_j^o) &= \left(\frac{\chi_{ij}^o}{\psi_{ij}^o}\right)^{r/2} \left\{ \pi_j^o \frac{K_{(1-\frac{p_j^o}{2})+r}\left(\sqrt{\psi_{ij}^o \chi_{ij}^o}\right)}{K_{(1-\frac{p_j^o}{2})}\left(\sqrt{\psi_{ij}^o \chi_{ij}^o}\right)} \right. \\
&\quad \left. +(1-\pi_j^o) \frac{K_{(2-\frac{p_j^o}{2})+r}\left(\sqrt{\psi_{ij}^o \chi_{ij}^o}\right)}{K_{(2-\frac{p_j^o}{2})}\left(\sqrt{\psi_{ij}^o \chi_{ij}^o}\right)} \right\}, \qquad (15) \\
E(\tilde{U}_{ij} \mid \boldsymbol{y}_j^o) &= \boldsymbol{R}_{ij}^{oo}\left\{\boldsymbol{b}_{ij}^o + \boldsymbol{\lambda}_i\left(E(W_j \mid \boldsymbol{y}_j^o) - a_{\alpha_i}\right)\right\} \qquad (16) \\
E(W_j^{-1}\tilde{\boldsymbol{U}}_j \mid \boldsymbol{y}_j^o) &= \boldsymbol{R}_j^{oo}\left\{\boldsymbol{b}_j^o E(W_j^{-1} \mid \boldsymbol{y}_j^o) + \boldsymbol{\lambda}\big(1 - a_{\alpha_i} E(W_j^{-1} \mid \boldsymbol{y}_j^o)\big)\right\}, \qquad (17) \\
E(W_j^{-1}\tilde{\boldsymbol{U}}_{ij}\tilde{\boldsymbol{U}}_{ij}^\top \mid \boldsymbol{y}_j^o) &= \left\{ E(W_j^{-1}\tilde{\boldsymbol{U}}_{ij} \mid \boldsymbol{y}_j^o)\boldsymbol{b}_{ij}^{o\top} \right. \\
&\quad \left. +\big[E(\tilde{\boldsymbol{U}}_{ij} \mid \boldsymbol{y}_j^o) - a_{\alpha_i} E(W_j^{-1}\tilde{\boldsymbol{U}}_{ij} \mid \boldsymbol{y}_j^o)\big]\boldsymbol{\lambda}_i^\top + \boldsymbol{I}_q \right\} \boldsymbol{R}_{ij}^{oo}. \qquad (18)
\end{aligned}
$$

Proof. a. Similar to the relationship described in equation (7), and considering the observed portion of the vector $\boldsymbol{Y}_j$, this can be readily proved.

b. Using the relations (2), (9), (10) and, transformations of relation (12) and these clear points, that $\tilde{\boldsymbol{U}}_{ij}$ has the random representation of relation (2) with location, scale and skewness parameters defined in relation (10), and also $\boldsymbol{\varepsilon}_{ij} = \sqrt{W}\boldsymbol{Z}^*$, which $\boldsymbol{Z}^* \sim N_p(\boldsymbol{0}, \boldsymbol{D}_i)$ is proved.

c. Similar to part (b), and with the assistance of permutation matrices, the partitioning of the matrix $\boldsymbol{Y}_j$, along with the properties of the conditional normal distribution, is demonstrated.

d. It can be proved by performing simple algebraic operations from parts (b) and (c) and using Bayes theorem.
e. It is proved using part (d) and similar to what is given in Hashemi et al. (2020). □

3.2 Estimation of model parameters using ECME algorithm

When there is missing data or a hidden variable in a statistical model, the expectation-maximization (EM) algorithm is an iterative method for computing maximum likelihood (ML) estimates. The EM algorithm's features include monotone convergence and ease of use. Despite these desirable features, the maximization step of the EM algorithm is nearly impossible for the MNMVLFA model. As a result, we use the ECME approach, an extension of the expectation-conditional maximization (ECM) technique Meng and Rubin (1993), to estimate the parameters of the MNMVLFA model. The ECME algorithm shows a faster convergence speed than EM and ECM algorithms while maintaining stability and integrity. The ECME algorithm replaces certain CM-steps of the ECM algorithm with CML-steps, leading to the maximization of the restricted true likelihood function. To browse a symbol of variables, we consider the allocation indicator vector $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$, the latent factor vector $\tilde{\boldsymbol{U}} = (\tilde{\boldsymbol{U}}_1, \ldots, \tilde{\boldsymbol{U}}_n)$, and the vector of mixture variables $\boldsymbol{W} = (W_1, \ldots, W_n)$ simultaneously with the vector of missing values $\boldsymbol{y}^m = (\boldsymbol{y}_1^m, \ldots, \boldsymbol{y}_n^m)$ as the missing data. In the part (c) of theorem 3.1 , the log-likelihood function of $\boldsymbol{\Theta}$ for the complete data, consists of the observed data $\boldsymbol{y}^o = (\boldsymbol{y}_1^o, \ldots, \boldsymbol{y}_n^o)$ and the missing data $\boldsymbol{y}_c = (\boldsymbol{Z}, \tilde{\boldsymbol{U}}, \boldsymbol{W}, \boldsymbol{y}^m)$, by removing the independent values of the parameter is

$$\begin{aligned}
\ell_c(\boldsymbol{\Theta} \mid \boldsymbol{y}_c) = & \sum_{j=1}^{n}\sum_{i=1}^{g} Z_{ij}[\log \pi_i - \frac{1}{2}\log|\boldsymbol{D}_i| + \log f_{Lindley}(W_j;\alpha_i)] \\
& -\frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{g} W_j^{-1}\big(\boldsymbol{y}_j - \boldsymbol{\mu}_i - \tilde{\boldsymbol{B}}_i\tilde{\boldsymbol{U}}_{ij}\big)^\top \boldsymbol{D}_i^{-1}\big(\boldsymbol{y}_j - \boldsymbol{\mu}_i - \tilde{\boldsymbol{B}}_i\tilde{\boldsymbol{U}}_{ij}\big) \\
& -\frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{g}\Big(\big(W_j - 2a_{\alpha_i} + W_j^{-1}a_{\alpha_i}^2\big)\boldsymbol{\lambda}_i\boldsymbol{\lambda}_i^\top - 2\boldsymbol{\lambda}_i(\tilde{\boldsymbol{U}}_{ij} - a_{\alpha_i}W_j^{-1}\tilde{\boldsymbol{U}}_{ij})^\top\Big).
\end{aligned} \tag{19}$$

In the kth repetition of the E-step, we calculate the expected value of $\ell_c(\boldsymbol{\Theta} \mid \boldsymbol{y}_c)$ given the observed value $\boldsymbol{y}_j^o$ and the estimate of the current parameters $\hat{\boldsymbol{\Theta}}^{(k)}$. The Q function is then computed as

$$Q(\boldsymbol{\Theta} \mid \hat{\boldsymbol{\Theta}}^{(k)}) = E\left(\ell_c(\boldsymbol{\Theta} \mid \boldsymbol{y}_c) \mid \boldsymbol{y}_j^o, \hat{\boldsymbol{\Theta}}^{(k)}\right), \tag{20}$$

in which $\hat{\boldsymbol{\Theta}}^{(k)} = (\hat{\pi}_1^{(k)}, \ldots, \hat{\pi}_{g-1}^{(k)}, \hat{\boldsymbol{\theta}}_1^{(k)}, \ldots, \hat{\boldsymbol{\theta}}_g^{(k)})$ and $\hat{\boldsymbol{\theta}}_i^{(k)} = (\hat{\boldsymbol{\mu}}_i^{(k)}, \hat{\tilde{\boldsymbol{B}}}_i^{(k)}, \hat{\boldsymbol{D}}_i^{(k)}, \hat{\boldsymbol{\lambda}}_i^{(k)}, \hat{\alpha}_i^{(k)})$. In order to appraise (20), it is necessary to obtain conditional expectations.

$$\begin{aligned}
\hat{Z}_{ij}^{(k)} &= E(Z_{ij} \mid \boldsymbol{y}_j^o, \hat{\boldsymbol{\Theta}}^{(k)}), & \hat{w}_{ij}^{(k)} &= E(W_j \mid \boldsymbol{y}_j^o, \hat{\boldsymbol{\Theta}}^{(k)}), \\
\hat{t}_{ij}^{(k)} &= E(W_j^{-1} \mid \boldsymbol{y}_j^o, \hat{\boldsymbol{\Theta}}^{(k)}), & \hat{\boldsymbol{\zeta}}_{0ij}^{(k)} &= E(\tilde{\boldsymbol{U}}_{ij} \mid \boldsymbol{y}_j^o, \hat{\boldsymbol{\Theta}}^{(k)}),
\end{aligned}$$

$$\hat{\boldsymbol{\zeta}}_{1ij}^{(k)} = E(W_j^{-1}\tilde{\boldsymbol{U}}_{ij} \mid \boldsymbol{y}_j^o, \hat{\boldsymbol{\Theta}}^{(k)}), \quad \hat{\boldsymbol{\Omega}}_{ij}^{(k)} = E(W_j^{-1}\tilde{\boldsymbol{U}}_{ij}\tilde{\boldsymbol{U}}_{ij}^\top \mid \boldsymbol{y}_j^o, \hat{\boldsymbol{\Theta}}^{(k)}), \tag{21}$$

which can be easily obtained by equations in the (15)-(18) stated in theorem 3.1, all of elements in $\boldsymbol{\Theta}$ changed with $\hat{\boldsymbol{\Theta}}^{(k)}$.

In the $(k+1)$th repetition of the CM-steps, the updated formulas for the MNMVLFA model parameters are given in the following steps:
CM-step 1: compute $\hat{\pi}_i^{(k+1)} = \hat{n}_i^{(k)}/n$, in which $\hat{n}_i^{(k)} = \sum_{j=1}^n \hat{Z}_{ij}^{(k)}$.
CM-step 2: compute

$$\hat{\boldsymbol{\mu}}_i^{(k+1)} = \frac{\sum_{j=1}^n \hat{Z}_{ij}^{(k)}\hat{t}_{ij}^{(k)}\hat{\boldsymbol{q}}_{ij}^{(k)} - \hat{\boldsymbol{D}}_i^{(k)}\sum_{j=1}^n \hat{Z}_{ij}^{(k)}\hat{\boldsymbol{C}}_{ij}^{oo^{(k)}}\hat{\tilde{\boldsymbol{B}}}_i^{(k)}\hat{\boldsymbol{\zeta}}_{1ij}^{(k)}}{\sum_{j=1}^n \hat{Z}_{ij}^{(k)}\hat{t}_{ij}^{(k)}},$$

in which $\hat{\boldsymbol{q}}_{ij}^{(k)} = \hat{\boldsymbol{\mu}}_i^{(k)} + \hat{\boldsymbol{D}}_i^{(k)}\hat{\boldsymbol{C}}_{ij}^{oo^{(k)}}(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_i^{(k)})$.
CM-step 3: By placing $\boldsymbol{\mu}_i = \hat{\boldsymbol{\mu}}_i^{(k+1)}$, and maximizing the function (20) relative to $\tilde{\boldsymbol{B}}_i$, the updated $\hat{\tilde{\boldsymbol{B}}}_i^{(k)}$ is

$$\hat{\tilde{\boldsymbol{B}}}_i^{(k+1)} = \left(\sum_{j=1}^n \hat{Z}_{ij}^{(k)}\left[\hat{\boldsymbol{E}}_{ij}^{oo^{(k)}}\hat{\boldsymbol{\Omega}}_{ij}^{(k)} + (\hat{\boldsymbol{q}}_{ij}^{(k)} - \hat{\boldsymbol{\mu}}_i^{(k+1)})\hat{\boldsymbol{\zeta}}_{1ij}^{(k)\top}\right]\right)\left(\sum_{j=1}^n \hat{Z}_{ij}^{(k)}\hat{\boldsymbol{\Omega}}_{ij}^{(k)}\right)^{-1},$$

in which $\hat{\boldsymbol{E}}_{ij}^{oo^{(k)}} = \left(\boldsymbol{I}_p - \hat{\boldsymbol{D}}_i^{(k)}\hat{\boldsymbol{C}}_{ij}^{oo^{(k)}}\right)\hat{\tilde{\boldsymbol{B}}}_i^{(k)}$.
CM-step 4: By placing $\boldsymbol{\mu}_i = \hat{\boldsymbol{\mu}}_i^{(k+1)}$ and $\tilde{\boldsymbol{B}}_i = \hat{\tilde{\boldsymbol{B}}}_i^{(k+1)}$,and maximizing the function (20) relative to $\tilde{\boldsymbol{D}}_i$, the updated $\hat{\boldsymbol{D}}_i^{(k)}$ is obtained

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

in which

$$\begin{aligned}\hat{\boldsymbol{\Upsilon}}_{ij}^{(k+1)} &= \hat{t}_{ij}^{(k)}(\hat{\boldsymbol{q}}_{ij}^{(k)} - \hat{\boldsymbol{\mu}}_i^{(k+1)})(\hat{\boldsymbol{q}}_{ij}^{(k)} - \hat{\boldsymbol{\mu}}_i^{(k+1)})^\top + \left(\boldsymbol{I}_p - \hat{\boldsymbol{D}}_i^{(k)}\hat{\boldsymbol{C}}_{ij}^{oo^{(k)}}\right)\hat{\boldsymbol{D}}_i^{(k)} \\ &+ \left(\hat{\boldsymbol{E}}_{ij}^{oo^{(k)}} - \hat{\tilde{\boldsymbol{B}}}_i^{(k+1)}\right)\hat{\boldsymbol{\Omega}}_{ij}^{(k)}\left(\hat{\boldsymbol{E}}_{ij}^{oo^{(k)}} - \hat{\tilde{\boldsymbol{B}}}_i^{(k+1)}\right) \\ &+ \left(\hat{\boldsymbol{q}}_{ij}^{oo(k)} - \hat{\boldsymbol{\mu}}_i^{(k+1)}\right)\hat{\tilde{\boldsymbol{\zeta}}}_{1ij}^{(k)\top}\left(\hat{\boldsymbol{E}}_{ij}^{oo^{(k)}} - \hat{\tilde{\boldsymbol{B}}}_i^{(k+1)}\right)^\top \\ &+ \left(\hat{\boldsymbol{E}}_{ij}^{oo^{(k)}} - \hat{\tilde{\boldsymbol{B}}}_i^{(k+1)}\right)\hat{\tilde{\boldsymbol{\zeta}}}_{1ij}^{(k)}\left(\hat{\boldsymbol{q}}_{ij}^{oo(k)} - \hat{\boldsymbol{\mu}}_i^{(k+1)}\right)^\top.\end{aligned}$$

CM-step 5: compute

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

in which $a_{\hat{\alpha}_i^{(k)}} = \frac{\hat{\alpha}_i^{(k)}+2}{\hat{\alpha}_i^{(k)}(\hat{\alpha}_i^{(k)}+1)}$.

We employ 'CML-step' to maximize the restricted true log-likelihood function as there is no closed-form for estimating the α parameter.
CML-step 6:

$$\hat{\alpha}_i^{(k+1)} = \arg\max_{\alpha_i} \sum_{j=1}^{n} \log f_{\text{MNMVL}} \left(\boldsymbol{y}_j^o; \boldsymbol{\mu}_{ij}^{o(k+1)} - a_{\alpha_i} \boldsymbol{\delta}_{ij}^{o(k+1)}, \boldsymbol{\Sigma}_{ij}^{oo(k+1)}, \boldsymbol{\delta}_{ij}^{o(k+1)}, \alpha_i \right),$$

in which $\hat{\boldsymbol{\mu}}_{ij}^{o(k+1)}$, $\hat{\boldsymbol{\delta}}_{ij}^{o(k+1)}$ and $\hat{\boldsymbol{\Sigma}}_{ij}^{oo(k+1)}$ are estimates of $\boldsymbol{\mu}_{ij}^o$, $\boldsymbol{\delta}_{ij}^o$ and $\boldsymbol{\Sigma}_{ij}^{oo}$, respectively.

In the ECME algorithm, E-step and CM/CML-steps continue until the reliable convergence criterion is estimated. For example, the loop created in step $(k+1)$ stops when either $\| \hat{\boldsymbol{\Theta}}^{(k+1)} - \hat{\boldsymbol{\Theta}}^{(k)} \|$ or $|\ell(\hat{\boldsymbol{\Theta}}^{(k+1)}) - \ell(\hat{\boldsymbol{\Theta}}^{(k)})|$ becomes less than a user-defined tolerance. This means that the algorithm continues until we reach convergence in the value of parameters or likelihood function. In this perusal, if the maximum number of repetitions achieves $k_{\max} = 10,000$, the algorithm terminates or when the absolute difference between the log-likelihood value and its asymptotic estimate is less than $\epsilon = 10^{-5}$. Upon convergence, the resulting ML estimates are indicated by $\hat{\boldsymbol{\theta}} = (\hat{\pi}_i, \hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{B}}_i, \hat{\boldsymbol{D}}_i, \hat{\boldsymbol{\lambda}}_i, \hat{\alpha}_i)$, in which $\hat{\boldsymbol{B}}_i = \hat{\tilde{\boldsymbol{B}}}_i \hat{\boldsymbol{\Lambda}}_i^{1/2}$, and $\hat{\boldsymbol{\Lambda}}_i$ is $\hat{\boldsymbol{\Lambda}}_i = a_{\hat{\alpha}_i} \boldsymbol{I}_q + b_{\hat{\alpha}_i} \hat{\boldsymbol{\lambda}}_i \hat{\boldsymbol{\lambda}}_i^\top$.

3.3 Forecast of factor scores and missing values

Predicting the factor scores is helpful for additional analysis after the suggested model's parameters have been estimated. For example, one may seek to use factor information for data reconstruction into a lower-dimensional subspace, or one may be interested in knowing how factor scores differ between groups. By applying the law of repeated expectations and utilizing (14), we arrive at

$$\hat{\boldsymbol{u}}_{ij} = E(\boldsymbol{U}_{ij} \mid \boldsymbol{y}_j^o, \hat{\boldsymbol{\Theta}}) = \hat{\boldsymbol{\Lambda}}_i^{-1/2} \hat{\boldsymbol{R}}_{ij}^{oo} \left\{ \hat{\boldsymbol{b}}_{ij}^o + \hat{\boldsymbol{\lambda}}_i \left(E(W_j \mid \boldsymbol{y}_j^o, \hat{\boldsymbol{\Theta}}) - a_{\hat{\alpha}_i} \right) \right\}, \qquad (22)$$

in which $E(W_j \mid \boldsymbol{y}_j^o, \hat{\boldsymbol{\Theta}})$ is given by (15). accordingly, the estimated factor scores in accordance with $\boldsymbol{y}_j^0$ can be obtained as

$$\hat{\boldsymbol{u}}_j = \sum_{j=1}^{n} \hat{Z}_{ij} \hat{\boldsymbol{u}}_{ij}.$$

In addition, substituting acceptable values instead of missing values is a necessary step for constructing a full dataset to employ standard statistical methods. Another advantage of the ML technique is that it assigns a value for any missing data, with the title "single imputation." A minimum mean squared conditional predictor for $\boldsymbol{y}_j^m$ using the ECME algorithm is

$$\hat{\boldsymbol{y}}_j^m = E(\boldsymbol{Y}_j^m \mid \boldsymbol{y}_j^o, \hat{\boldsymbol{\Theta}}) = \boldsymbol{M}_j \sum_{j=1}^{n} \hat{Z}_{ij} \left[\hat{\boldsymbol{\mu}}_i + \hat{\boldsymbol{B}}_i \hat{\boldsymbol{u}}_{ij} + \hat{\boldsymbol{D}}_i \hat{\boldsymbol{C}}_{ij}^{oo} (\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_i - \hat{\boldsymbol{B}}_i \hat{\boldsymbol{u}}_{ij}) \right]. \qquad (23)$$

While the predictor (23) fails to capture the uncertainty surrounding the predictions of the unknown missing values, addressing this issue can be achieved through a more universally functional multiple imputation (MI) procedure Schafer (1997). Significantly,

MI operates under the assumption that the frequentist criterion (or the method of ML) has been met, and subsequently generates single imputations multiple times for replicated samples obtained through MCMC or bootstrap techniques.

3.4 Evaluation criteria

The log-likelihood function of a finite mixture model can lead to multiple states due to its complexity. Consequently, the EM-based algorithm may face challenges in achieving convergence, and even when convergence is reached, it may not yield a comprehensive and dependable solution. To address this challenge effectively, it is advisable to test various initial values and subsequently select the one that yields the highest likelihood. This can be easily accomplished through multiple random initializations and the extraction of initial parameter values based on different initial partitions (e.g., randomly assigning each point to a cluster) or employing the k-means clustering method Hartigan and Wong (1979). This involves determining the initial $\boldsymbol{\pi}_i$ based on the sample proportion of cluster labels and considering the initial $\boldsymbol{\mu}_i$ as the sample mean vectors for each cluster. Subsequently, the common FA model is fitted to each clustered sample, and the resulting estimates are used as the initial values for $\boldsymbol{B}_i$ and $\boldsymbol{D}_i$. Additionally, the initial skewness parameters $\boldsymbol{\lambda}_i$ are obtained through parameter estimation in the NMVL distribution. Finally, the initial value of parameter α_i is considered a fixed value, such as $\hat{\alpha}_i^{(0)} = 1$. Model-based clustering using the MFA method and its various types typically involves unknown quantities, including the number of components g and the number of factors q. In such cases, the Bayesian information criterion (BIC) Schwarz (1978) proves useful for selecting g and q, particularly for determining g Keribin (2000). AIC and BIC are

$$-2\ell(\hat{\boldsymbol{\Theta}}) + mc_n,$$

here, $\ell(\hat{\boldsymbol{\Theta}})$ denotes the maximum log-likelihood value, m represents the number of free parameters to be estimated within the model , and c_n is 2 for the Akaike information criterion(AIC) and $\log(n)$ for BIC. Biernacki et al. (2000) noted that BIC might not be the optimal method for determining the number of components in a model-based clustering approach. As an alternative, they introduced an integrated completed likelihood (ICL) criterion designed to select g in a way that facilitates proper partitioning of the data. ICL is defined as follows

$$ICL = BIC + 2ENT(\hat{\boldsymbol{z}}),$$

in which, $ENT((\hat{\boldsymbol{z}})) = -\sum_{i=1}^{g}\sum_{j=1}^{n}\hat{z}_{ij}\log\hat{z}_{ij}$ represents the entropy used for assessing the overlap of clusters. Models with smaller BIC or ICL values indicate a better fit. In cases where there is a discrepancy between the two criteria, the ICL criterion is preferred due to its imposition of a higher penalty on more complex models. To verify the convergence of the ECME algorithm in the MNMVLFA model, we measure the absolute difference between the log-likelihood value and its asymptotic estimate using the Aitken acceleration method Aitken (1925)

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

in which $\ell_{\infty}^{(k+1)}$ McLachlan and Krishnan (2007) represents the extreme value of the log-likelihood function in the $(k+1)$th repetition. In comparing the classification performance of different model-based classifiers, we adopt the adjusted Rand index (ARI) and the correct classification rate (CCR) with higher values meant for good classification results. The ARI criterion, proposed by Hubert and Arabie (1985), extends the Rand index (RI) Rand (1971), which confirms the agreement between two separate partitions of the same data. Typically, ARI takes values in the range of (0,1), but in exceptional cases, when the level of agreement is weak, it may assume a negative value. The CCR is determined for each permutation of cluster labels and the reported value is the largest value among all permutations.

4 Simulation study

In the following, we present simulation studies to scrutinize various aspects of the model and computation procedures. The initial simulation study assesses the fitting and clustering efficacy of the MNMVLFA model in comparison to other models.

4.1 Evaluation of fitting and clustering efficiency of the proposed model

For simulation, synthetic data are generated with $n = 500$ observations from a mixture of normal inverse Gaussian factor analysis (MNIGFA) distribution that $g = 2$ is considered. In per repetition of $M = 100$ experiment, a random sample of size n=500 is extracted from the MNIGFA distribution $\boldsymbol{Y}_j = \boldsymbol{\mu}_i + \boldsymbol{B}_i\boldsymbol{U}_{ij} + \boldsymbol{\epsilon}_{ij}$, where

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

where $\boldsymbol{\Lambda}_i = a_{\theta_i}\boldsymbol{I}_q + b_{\theta_i}\boldsymbol{\lambda}_i\boldsymbol{\lambda}_i^{\top}$, $a_{\theta_i} = \sqrt{\chi_i/\psi_i}$, $b_{\theta_i} = \chi_i/\psi_i\frac{K_{1.5}(\sqrt{\chi_i\psi_i})}{K_{-0.5}(\sqrt{\chi_i\psi_i})} - a_{\theta_i}^2$, and $NIG_p(\boldsymbol{\mu}_i, \boldsymbol{\lambda}_i, \boldsymbol{\Sigma}_i, \chi_i, \psi_i)$ defines the p-variate NIG distribution, that this distribution is a particular state of GH distribution with $\kappa = -0.5$. In order to generate non-normal mixtures, the NIG distribution is considered because it creates the right amount of asymmetry and leptokurtosis. The hypothetical two-component parameters of MNIGFA with $q = 2$ are expressed as $\pi_1 = \frac{1}{3}$, $\pi_2 = \frac{2}{3}$, $\psi_1 = 11$, $\psi_2 = 5$, $\chi_1 = 4$, $\chi_2 = 8$,

$$\boldsymbol{\mu}_1 = (0,0,0,0,0)^{\top}, \quad \boldsymbol{\mu}_2 = (10,10,10,10,10)^{\top},$$

$$\boldsymbol{D}_1 = \text{diag}\{1,2,3,4,3\}, \quad \boldsymbol{D}_2 = \text{diag}\{1,2,3,5,4\}, \quad \boldsymbol{\lambda}_1 = (1,9)^{\top}, \quad \boldsymbol{\lambda}_2 = (2,8)^{\top},$$

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

The mentioned initial values were deliberately chosen to ensure that the generated data have high skewness, heavy-tails and strong separation between the two classes. We compare the model presented in this article (MNMVLFA) with MFA, MSNFA, and MSTFA in terms of performance. In this comparison, the MNIGFA model is not considered because, as expected, the real model has the best performance. Also, we consider two levels of missingness, 10% and 20%. To enhance the robustness of

clustering settlement, we employ the ARI, which corrects the inefficiency of the RI resulting from a fortuitous settlement. The model with the highest ARI provides the most accurate classification. Lower AIC and BIC values indicate a better fit for the model.

In Table 1, we present a summary of the fitting results of 500 simulated samples, along with the ARI values used to evaluate the clustering. AIC or BIC best aligns with the MNMVLFA model in the table, resulting in improved classification accuracy (ARI = 0.843). According to the numerical results, the proposed MNMVLFA model provides superior density estimation and improved clustering compared to its alternatives.

Table 1: Performance of varied skew mixture models fitted to 500 simulated MNIGFA datasets.

missing rate	Model	$\ell_{\max}$	m	AIC	BIC	ICL	ARI	CCR
10	MFA	–3515.52	39	7109.04	7253.48	7272.77	0.70	0.85
	MSNFA	–3503.39	43	7092.77	7252.04	7272.96	0.77	0.91
	MSTFA	–3502.87	45	7095.75	7262.42	7285.33	0.73	0.90
	MNMVLFA	–3484.81	45	7059.62	7226.29	7248.74	0.84	0.96
20	MFA	–3192	39	6461.99	6606.44	6632.13	0.63	0.76
	MSNFA	–3181.33	43	6448.67	6607.93	6635.75	0.75	0.90
	MSTFA	–3181.5	45	6453.01	6619.67	6645.94	0.71	0.88
	MNMVLFA	–3168.31	45	6426.63	6593.3	6633.88	0.82	0.95

4.2 Finite-sample attributes of EM-type estimators

Using a three-component MNMVLFA model, we produce 500 three-vary Monte Carlo samples that show the true parameter values in Table 2. We accept sample sizes $n = 250,\ 500$ and 1000. For simulated datasets, the MNMVLFA model with $g = 3$, $p = 3$, and $q = 1$ is fitted and the results of the estimated parameters are gathered. In this simulation, we consider the rate of missing above 30%. In Table 2, the mean values and standard deviations (Std.) of the EM estimates amongst 500 experiments are brought. To search for the accuracy of the estimate, we also calculate the absolute bias (AB) and the mean squared error (MSE)

$$\text{AB} = \frac{1}{500}\sum_{r=1}^{n}|\hat{\theta}^{(r)} - \theta_{\text{true}}| \quad \text{and} \quad \text{MSE} = \frac{1}{500}\sum_{r=1}^{n}(\hat{\theta}^{(r)} - \theta_{\text{true}}\ )^2,$$

where $\hat{\theta}^{(r)}$ specifies the parameter estimate for a particular element in $\boldsymbol{\Theta}$ obtained from the rth iteration. A close inspection of Table 2 provides evidence that all estimated parameters are within the true range and the standard deviations are reasonable. Furthermore, both AB and MSE values decrease with increasing n, experimentally confirming the accuracy of the ML estimates calculated with the EM algorithm.

Table 2: Mean and standard deviations (Std.) for EM-type estimates–500 samples from the MNMVLFA model for 30% missing.

Component		μ_{i1}	μ_{i2}	μ_{i3}	λ_{i1}	b_{i1}	b_{i2}	b_{i3}	d_{i1}	d_{i2}	d_{i3}	α	π
1	True	0	0	0	4	4	5	2	1	1	0.9	1	0.3
2	Parameter	5	5	5	-2	1	3	5	0.5	0.8	1	1.2	0.3
3		-2	-2	-2	-1	2	2	4	0.2	0.3	0.7	0.6	0.4
$n = 250$													
	Mean	0.058	0.057	0.095	4.054	3.940	4.922	2.319	1.162	1.400	0.795	1.290	0.323
1	Std.	0.531	0.446	0.596	0.504	0.466	0.581	0.255	0.186	0.248	0.246	0.261	0.059
	AB	0.396	0.344	0.438	0.381	0.347	0.430	0.215	0.144	0.193	0.185	0.215	0.042
	MSE	0.282	0.201	0.363	0.256	0.220	0.342	0.071	0.036	0.061	0.060	0.080	0.003
	Mean	4.962	4.931	4.949	-1.978	1.015	3.104	4.718	0.419	0.713	1.185	1.495	0.345
2	Std.	0.260	0.402	0.315	0.384	0.572	0.427	0.323	0.348	0.322	0.645	0.459	0.059
	AB	0.189	0.286	0.246	0.261	0.379	0.298	0.257	0.288	0.270	0.508	0.374	0.049
	MSE	0.069	0.165	0.101	0.146	0.325	0.182	0.110	0.127	0.110	0.441	0.221	0.004
	Mean	-1.990	-1.908	-2.040	-0.970	1.966	1.950	3.919	0.329	0.283	0.801	0.640	
3	Std.	0.151	0.162	0.220	0.233	0.239	0.442	0.377	0.292	0.310	0.403	0.401	
	AB	0.118	0.133	0.175	0.174	0.188	0.322	0.308	0.245	0.254	0.351	0.354	
	MSE	0.023	0.026	0.050	0.055	0.058	0.197	0.148	0.090	0.103	0.184	0.185	
$n = 500$													
	Mean	0.032	0.017	0.067	3.986	3.990	4.944	2.175	1.104	1.364	0.835	1.138	0.312
1	Std.	0.315	0.338	0.386	0.296	0.354	0.392	0.162	0.125	0.189	0.166	0.164	0.037
	AB	0.237	0.234	0.282	0.221	0.255	0.284	0.132	0.094	0.140	0.124	0.123	0.024
	MSE	0.100	0.114	0.153	0.088	0.125	0.156	0.027	0.016	0.037	0.027	0.031	0.001
	Mean	4.941	4.914	4.945	-2.041	1.064	3.039	4.750	0.449	0.765	1.040	1.234	0.338
2	Std.	0.178	0.260	0.228	0.249	0.360	0.278	0.220	0.262	0.230	0.391	0.306	0.037
	AB	0.142	0.197	0.168	0.178	0.258	0.207	0.162	0.193	0.172	0.304	0.232	0.013
	MSE	0.035	0.075	0.055	0.064	0.133	0.078	0.051	0.071	0.054	0.164	0.098	0.001
	Mean	-1.981	-1.985	-2.037	-0.990	1.983	1.966	3.938	0.176	0.271	0.778	0.634	
3	Std.	0.089	0.102	0.140	0.156	0.158	0.294	0.230	0.179	0.198	0.247	0.263	
	AB	0.073	0.083	0.112	0.118	0.125	0.222	0.194	0.147	0.160	0.199	0.211	
	MSE	0.008	0.011	0.021	0.024	0.025	0.087	0.056	0.033	0.040	0.066	0.073	

Continued.

Component		μ_{i1}	μ_{i2}	μ_{i3}	λ_{i1}	b_{i1}	b_{i2}	b_{i3}	d_{i1}	d_{i2}	d_{i3}	α	π
1	True	0	0	0	4	4	5	2	1	1	0.9	1	0.3
2	Parameter	5	5	5	-2	1	3	5	0.5	0.8	1	1.2	0.3
3		-2	-2	-2	-1	2	2	4	0.2	0.3	0.7	0.6	0.4
$n = 1000$													
	Mean	0.017	0.012	0.062	3.989	3.993	4.953	2.104	1.084	1.276	0.920	1.104	0.309
1	Std.	0.196	0.279	0.265	0.176	0.259	0.251	0.117	0.092	0.130	0.133	0.147	0.025
	AB	0.146	0.193	0.213	0.144	0.198	0.227	0.088	0.069	0.107	0.096	0.119	0.018
	MSE	0.039	0.073	0.094	0.036	0.074	0.103	0.014	0.009	0.021	0.018	0.029	0.001
	Mean	4.944	4.934	4.954	-2.034	1.052	3.033	4.754	0.457	0.766	1.058	1.229	0.331
2	Std.	0.096	0.213	0.178	0.174	0.257	0.193	0.152	0.179	0.140	0.292	0.203	0.026
	AB	0.098	0.147	0.113	0.115	0.166	0.129	0.118	0.131	0.107	0.226	0.149	0.009
	MSE	0.020	0.045	0.025	0.031	0.068	0.038	0.025	0.034	0.020	0.093	0.043	0.001
	Mean	-1.992	-1.990	-2.013	-0.966	1.985	1.976	3.946	0.174	0.278	0.776	0.639	
3	Std.	0.076	0.098	0.114	0.119	0.117	0.228	0.143	0.128	0.136	0.203	0.197	
	AB	0.053	0.047	0.062	0.092	0.093	0.166	0.128	0.108	0.114	0.176	0.169	
	MSE	0.005	0.004	0.007	0.015	0.015	0.057	0.026	0.018	0.020	0.046	0.042	

5 Hepatitis disease data

An initial study of 155 patients with severe and hazardous hepatitis was conducted Diaconis and Efron (1983). The data comprise both continuous and discrete values. Nineteen variables, including age, sex, and outcomes of modulus biochemical measurements, were measured for each patient. The dataset encompasses $p = 6$ numeric properties measured on $n = 155$ patients, with one binary variable for dichotomous classification. Specifically, 85 samples exhibit Yes for HISTOLOGY, while 75 samples indicate No for HISTOLOGY. Furthermore, most properties exhibit weak to strong asymmetry and mild to extremely heavy tails, indicating that both subcommunities deviate significantly from normal distributions. The presence of missing values varies widely across features, with some properties containing none, while others have 67 missing values. We applied the MFA, MSNFA, MSTFA, and MNMVLFA models to fit this dataset, with the range of q taking values from 1 to $q_{\max} = 3$.

We consider the value of g as 2, which corresponds to two memberships in the manifest group. Table 3 summarizes ML results, containing the number of parameters, the maximum log-likelihood value, and The values of ICL and BIC criteria. To compare clustering performance in the mentioned models, the classification settlement obtained by CCR is also presented in the final column of Table 3. According to the Table, the MNMVLFA with $q = 2$ is the best model to fit this dataset (BIC=5671.659 and ICL=5701.320) and the best classification accuracy (CCR = 0.731) for this dataset. In competition with the other three classical methods, the MNMVLFA model is more flexible.

Table 3: Performance of four mixtures of factor models fitted to the Hepatitis data.

Model	q	$\ell_{\max}$	m	BIC	ICL	CCR
MFA	1	-2830.280	37	5847.168	5884.611	0.587
	2	-2728.446	47	5693.933	5724.100	0.609
	3	-2726.010	55	5729.409	5756.471	0.619
MSNFA	1	-2827.425	39	5851.544	5889.669	0.574
	2	-2710.141	51	5677.498	5712.231	0.626
	3	-2709.085	61	5725.819	5756.877	0.619
MSTFA	1	-2821.121	41	5849.022	5890.463	0.574
	2	-2711.103	53	5689.508	5728.961	0.587
	3	-2706.151	63	5730.038	5765.494	0.619
MNMVLFA	1	-2803.295	41	5813.370	5852.910	0.675
	2	-2702.179	53	5671.659	5701.320	0.731
	3	-2697.481	63	5712.698	5739.421	0.692

We are interested in investigating how missing values can be imputed to the four MFA models. A suitable specification of latent factors is generally thought to produce more accurate imputations of missing data. Figure 1 displays the paired dispersion diagram of the missing values predicted by using (23) for all models. Our results show that either the MFA and MSNFA or MSTFA models provide the same attributed values, while the MNMVLFA model prepares different results.

In addition, we are eager to determine if the estimated factor scores are influenced by the assumptions underlying the factor distribution. Figure 2 compares the scatter plots of the factor scores obtained from the fitted models. The greater the dispersion, the larger the difference between the two models. In Figure 2, the factor scores obtained in two MFA and MSNFA models show relatively high dispersion compared to our

proposed model (MNMVLFA). However, the difference between the MNMVLFA and MSTFA models is relatively small due to the amount of dispersion. Therefore, in Figure 3, these two models are compared more accurately. From the available results, it is evident that fitting the MNMVLFA model on real data yields more reliable results compared to the MFA, MSNFA, and MSTFA models.

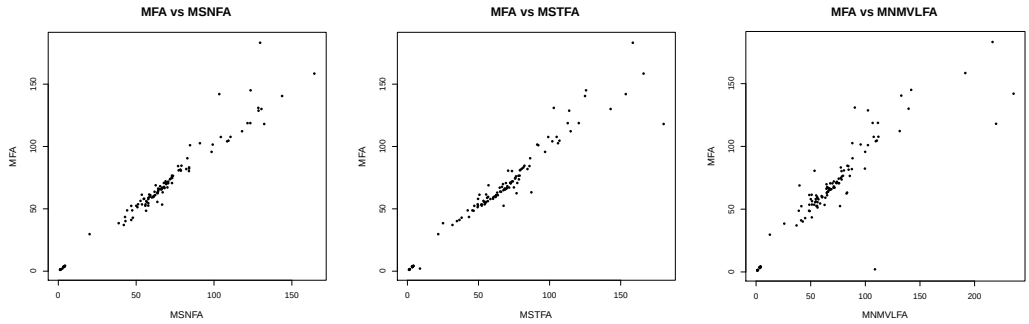

Figure 1: Scatter plots of imputed missing values using the MFA of all models for the Hepatitis data.

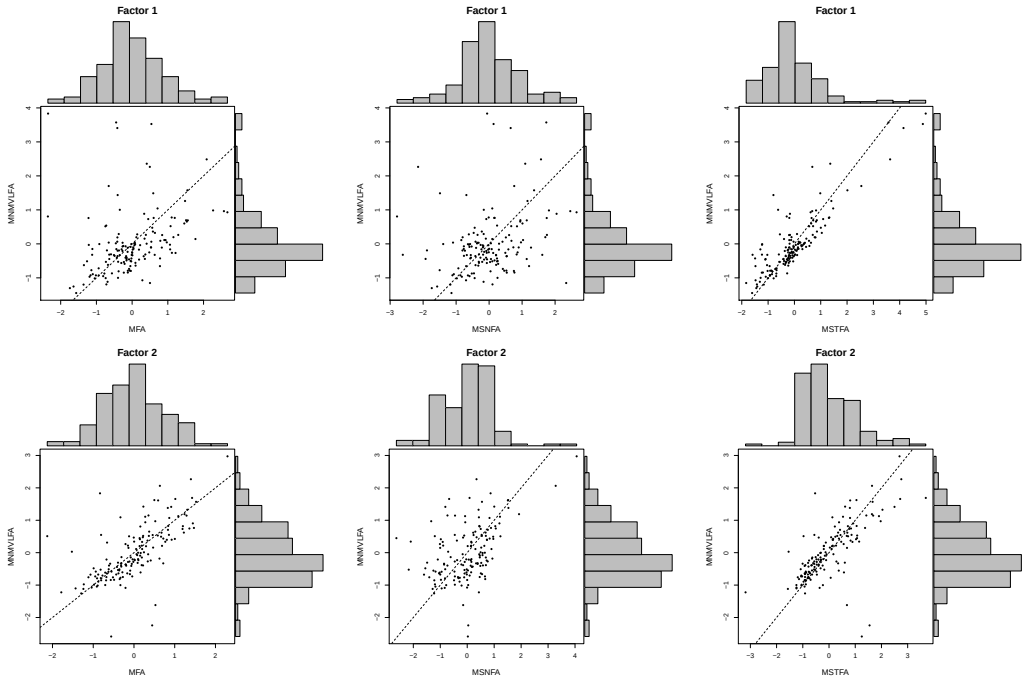

Figure 2: Scatter-histogram plots of factor scores obtained from the fitted models.

Figure 3 shows the fitted lines of MSTFA model against the MNMVLFA model acquired by marginalization of the fitted distributions superimposed on the scatter points for seven pairs of variables. Here the emphasis is on AGE against other variables,

in which the missing values are attributed by (23). In Figure 3, the scatter of the real data is more consistent with the MNMVLFA model. Also, some outlier values are not included in the contour of the MSTFA model, while the proposed model (MNMVLFA) also covers outliers. As a result, the lines of the MNMVLFA model have a better fit with the scatter of points compared to the MSTFA model.

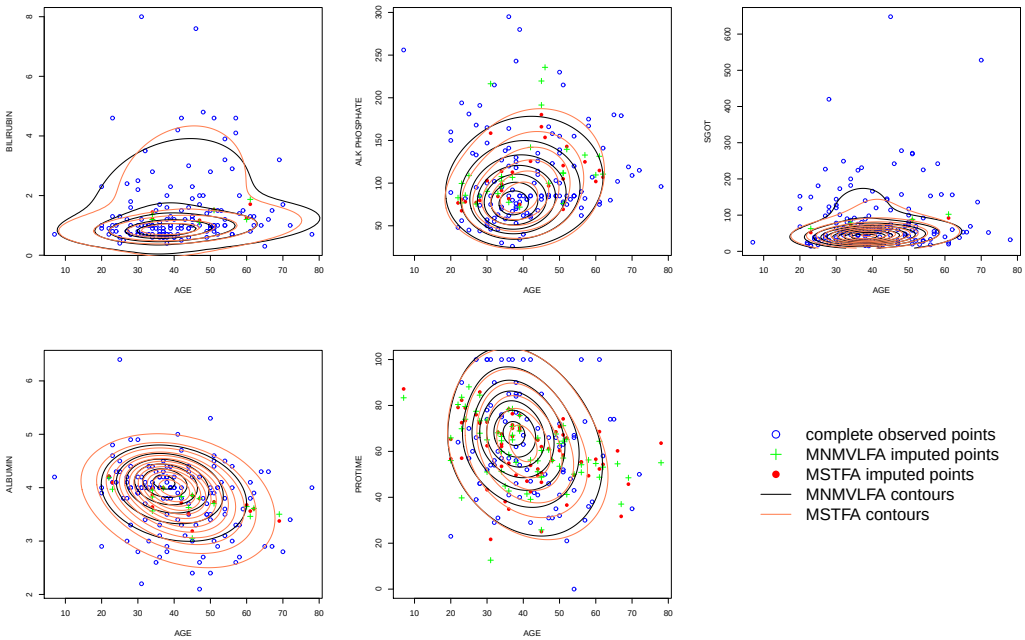

Figure 3: Scatterplots of fitted MSTFA and MNMVLFA lines for fourteen pairs of variables from the hepatitis data.

Conclusion

In this study, we developed the MFA model based on the NMVL distribution Naderi et al. (2018) to analyze heavy-tailed and asymmetric data in the presence of missing data. Initially, we review some preliminary results, focusing on the representation of the NMVL distribution. After that, using the development of EM-type algorithms, the ECME algorithm was employed to obtain parameter estimates for the MNMVLFA model in the presence of missing values. Then, two simulation studies and a real data example were designed to confirm the superiority of the proposed model. In the first simulation, the MNMVLFA model demonstrates superior density estimation and improved clustering compared to other models for data with high skewness and kurtosis with missing data. The second simulation verifies the accuracy of the ML estimates calculated using the ECME algorithm by examining the asymptotic properties of the estimators. In the real data example, the appropriateness of the proposed model fit was checked on the data in comparison with the MFA, NSNFA, and MSTFA models,

which demonstrates the superiority of the model. For future work, we plan to explore machine learning techniques to estimate the number of components and dimensions of factor loadings. We also intend to present a whole class of MNMVFA models in the presence of missing data, which include asymmetric distributions introduced by Darijani et al. (2024) and the normal mean-variance mixture of Birnbaum-Saunders factor analysis (NMVBSFA) model introduced by Hashemi et al. (2020).

References

Aitken, A. (1925). On Bernoulli's numerical solution of algebraic equations. *Proceedings of the Royal Society of Edinburgh*, **46**:289–305.

Arslan, O. (2010). An alternative multivariate skew Laplace distribution: Properties and estimation. *Statistical Papers*, **51**(4):865–887.

Azzalini, A. (2005). The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics*, **32**(2):159–188.

Azzalini, A. and Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **65**(2):367–389.

Bai, J. and Li, K. (2012). Statistical analysis of factor models of high dimension. *The Annals of Statistics*, **40**(1):436–465.

Barndorff-Nielsen, O. (1977). Exponentially decreasing distributions for the logarithm of particle size. *Proceedings of the Royal Society of London A Mathematical and Physical Sciences*, **353**(1674):401–419.

Basilevsky, A.T. (2009). *Statistical Factor Analysis and Related Methods: Theory and Applications*. John Wiley & Sons.

Biernacki, C., Celeux, G. and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(7):719–725.

Darijani, M., Zakerzadeh, H. and Jafari, A.A. (2024). Introducing a family of distributions by using the class of normal mean-variance mixture. *Journal of Statistical Theory and Practice*, **18**(1):17.

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

De Roover, K., Vermunt, J.K. and Ceulemans, E. (2022). Mixture multigroup factor analysis for unraveling factor loading noninvariance across many groups. *Psychological Methods*, **27**(3):281–306.

Diaconis, P. and Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American*, **248**(5):116–131.

Fischer, A., Gaunt, R.E. and Sarantsev, A. (2023). The variance-gamma distribution: A review. *arXiv preprint arXiv:230305615*.

Gaarenstroom, P.D., Perone, S.P. and Moyers, J.L. (1977). Application of pattern recognition and factor analysis for characterization of atmospheric particulate composition in southwest desert atmosphere. *Environmental Science & Technology*, **11**(8):795–800.

Ghahramani, Z. and Hinton, G.E. (1997). *The EM Algorithm for Mixtures of Factor Analyzers*. Vol. 60, Technical Report CRG-TR-96-1. University of Toronto.

Ghitany, M.E., Atieh, B. and Nadarajah, S. (2008). Lindley distribution and its application. *Mathematics and Computers in Simulation*, **78**(4):493–506.

Göncü, A. and Yang, H. (2016). Variance-gamma and normal-inverse Gaussian models: Goodness-of-fit to Chinese high-frequency index returns. *The North American Journal of Economics and Finance*, **36**:279–292.

Good, I.J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, **40**(3/4):237–264.

Hartigan, J.A. and Wong, M.A. (1979). Algorithm AS 136: A K-means clustering algorithm. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **28**(1):100–108.

Hashemi, F., Naderi, M., Jamalizadeh, A. and Lin, T. (2020). A skew factor analysis model based on the normal mean-variance mixture of Birnbaum-Saunders distribution. *Journal of Applied Statistics*, **47**(16):3007–3029.

Hubert, L. and Arabie, P.(1985). Comparing partitions. *Journal of Classification*, **2**(1):193–218.

Joreskog, K.G., Sorbom, D. and Magidson, J. (1979). *Advances in Factor Analysis and Structural Equation Models*. Abt books.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā: The Indian Journal of Statistics, Series A*, **62**:49–66.

Lawley, D.N. and Maxwell, A.E. (1962). Factor analysis as a statistical method. *Journal of the Royal Statistical Society, Series D (The Statistician)*, **12**(3):209–229.

Lee, S.X. and McLachlan, G.J. (2018). On formulations of skew factor models: Skew errors versus skew factors. *arXiv preprint arXiv:1810.04842*.

Lee, S.X., Lin, T.I. and McLachlan, G.J. (2018). Mixtures of factor analyzers with fundamental skew symmetric distributions. *arXiv preprint arXiv:1802.02467*.

Lee, S.X., Lin, T.I. and McLachlan, G.J. (2021). Mixtures of factor analyzers with scale mixtures of fundamental skew normal distributions. *Advances in Data Analysis and Classification*, **15**(2):481–512.

Lin, T.I., McLachlan, G.J. and Lee, S.X. (2016). Extending mixtures of factor models using the restricted multivariate skew-normal distribution. *Journal of Multivariate Analysis*, **143**:398–413.

Lin, T.I., Wang, W.L., McLachlan, G.J. and Lee, S.X. (2018). Robust mixtures of factor analysis models using the restricted multivariate skew-t distribution. *Statistical Modelling*, **18**(1):50–72.

Lin, T.I., Wu, P.H., McLachlan, G.J. and Lee, S.X. (2015). A robust factor analysis model using the restricted skew-t distribution. *Test*, **24**(3):510–531.

Little, R.J. and Rubin, D.B. (2019). *Statistical Analysis with Missing Data*. John Wiley & Sons.

Liu, C. and Rubin, D.B. (1994). The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika*, **81**(4):633–648.

Lopes, H.F. and West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica*, **14**(1):41–67.

Maleki, M. and Wraith, D. (2019). Mixtures of multivariate restricted skew-normal factor analyzer models in a Bayesian framework. *Computational Statistics*, **34**(3):1039–1053.

McLachlan, G.J. and Krishnan, T. (2007). *The EM algorithm and extensions, 2nd edition*. John Wiley & Sons.

McLachlan, G.J., Bean, R.W. and Jones L.B.T. (2007). Extension of the mixture of factor analyzers model to incorporate the multivariate t-distribution. *Computational Statistics & Data Analysis*, **51**(11):5327–5338.

McLachlan, G.J., Peel, D. and Bean, R.W. (2003). Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics & Data Analysis*, **41**(3-4):379–388.

McNicholas, S.M., McNicholas, P.D. and Browne, R.P. (2017). A mixture of variance-gamma factor analyzers. *Big and Complex Data Analysis: Methodologies and Applications*, 369–385.

Meng, X. and Rubin, D.B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, **80**(2):267–278.

Murray, P.M., Browne, R.P. and McNicholas, P.D. (2013). Mixtures of unrestricted skew-t factor analyzers. *arXiv preprint arXiv:1310.6224*.

Murray, P.M., Browne, R.P. and McNicholas, P.D. (2014). Mixtures of skew-t factor analyzers. *Computational Statistics & Data Analysis*, **77**:326–335.

Murray, P.M., McNicholas, P.D. and Browne, R.P. (2014). A mixture of common skew-t factor analysers. *Stat*, **3**(1):68–82.

Naderi, M., Arabpour, A. and Jamalizadeh, A. (2018). Multivariate normal mean-variance mixture distribution based on Lindley distribution. *Communications in Statistics-Simulation and Computation*, **47**(4):1179–1192.

Rand, W.M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66**(336):846–850.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, **63**(3):581–592.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. CRC press.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**(2):461–464.

Spearman, C. (1904). The proof and measurement of association between two things. *The American journal of psychology*, **3/4**(1987):441–471.

Tortora, C., McNicholas, P.D. and Browne, R.P. (2016). A mixture of generalized hyperbolic factor analyzers. *Advances in Data Analysis and Classification*, **10**(4):423–440.

Wall, M.M., Guo, J. and Amemiya, Y. (2012). Mixture factor analysis for approximating a nonnormally distributed continuous latent factor with continuous and dichotomous observed variables. *Multivariate Behavioral Research*, **47**(2):276–313.

Wang, W.L. (2013). Mixtures of common factor analyzers for high-dimensional data with missing information. *Journal of Multivariate Analysis*, **117**:120–133.

Wang, W.L. (2015). Mixtures of common t-factor analyzers for modeling high-dimensional data with missing values. *Computational Statistics & Data Analysis*, **83**:223–235.

Wang, W.L., Liu, M. and Lin, T.I. (2017). Robust skew-t factor analysis models for handling missing data. *Statistical Methods & Applications*, **26**:649–672.

Wei, Y., Tang, Y. and McNicholas, P.D. (2018). Flexible high-dimensional unsupervised learning with missing data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **42**(3):610–621.