Noname manuscript No.
(will be inserted by the editor)

Flexible factor model for handling missing data in supervised learning

Andriette Bekker · Farzane Hashemi · Mohammad Arashi

Received: date / Accepted: date

Abstract This paper presents an extension of the factor analysis model based on the normal mean-variance mixture of the Birnbaum-Sanders in the presence of nonresponses and missing data. This model can be used as a powerful tool to model non-normal features observed from data such as strongly skewed and heavy-tail noises. Missing data may occur due to operator error or incomplete data capturing therefore cannot be ignored in factor analysis modeling. We implement an EM-type algorithm for maximum likelihood estimation and propose single imputation of possible missing values under a missing at random mechanism. The potential and applicability of our proposed method are illustrated through analysing both simulated and real datasets.

Keywords Automobile dataset · Asymmetry · ECME algorithm · Factor analysis model · Heavy tails · Incomplete data · Liver disorders dataset

A. Bekker
Department of Statistics, Faculty of Natural & Agricultural Sciences, University of Pretoria, Pretoria, South Africa

F. Hashemi
Department of Statistics, Faculty of Natural & Agricultural Sciences, University of Pretoria, Pretoria, South Africa
Department of Statistics, Faculty of Mathematical Sciences, University of Kashan, Kashan, Iran

M. Arashi (Corresponding author)
Department of Statistics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhah, Iran
Department of Statistics, Faculty of Natural & Agricultural Sciences, University of Pretoria, Pretoria, South Africa
E-mail: arashi@um.ac.ir

1 Introduction

Consider the following Gaussian factor analysis model

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

where $\boldsymbol{B} \in \mathbb{R}^{p\times q}$ stands for the matrix of factor loadings, with $q < p$, $\boldsymbol{U}_j$'s are the latent variables called *common factors*, $\boldsymbol{\varepsilon}_j \in \mathbb{R}^p$ denote the model errors called *specific factors*, and $\boldsymbol{D} = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$. We refer to [3] for more details and applications.

The Gaussian factor analysis model lacks robustness againts asymmetry, heavy tails and missing values and this may strongly affect statistical inference validity. McLachlan et al. [18] considered a robust extension by adopting the multivariate t distribution for both the errors and factors. When there is asymmetry in multivariate data, Liu and Lin [17] suggested the multivariate skew normal factor model to compensate for missing data and to obtain valid interpretation among variables relationships. To overcome the restriction of skew normal factor analysis for heavy tails data, Wang et al. [28] presented a robust skew factor analysis model based on a restricted version of the multivariate skew-t (see [12]); called skew-t factor analysis model with missing data.

Several different skew factor analysis models based on the family of generalized hyperbolic distribution have been proposed in the literature, for example, [20,21,26]. Recently, Wei et al. [29] developed generalized hyperbolic factor analysis introduced by Tortora et al. [26] in the presence of missing values. Specifically, a p-variate random variable $\boldsymbol{X}$ is said to have generalized hyperbolic distribution if it can be generated through the linear stochastic representation

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

where $\boldsymbol{Z} \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ is independent of W with generalized inverse Gaussian distribution [6]. Pourmousa et al. [22] represented the normal mean-variance Birnbaum-Sanders distribution when W in (2) has the Birnbaum-Saunders distribution, denoted by $W \sim BS(\alpha, 1)$. Very recently, Hashemi et al. [7] introduced a skew extension of the classical factor analysis model based on the multivariate normal mean-variance Birnbaum-Sanders distribution for modeling multivariate data with abnormality and heavy-tailed behavior.

The aim of this paper is to develop a flexible factor analysis model for handling skewed and heavy-tailed data in the presence missing values. The missing values might be present in the outlying data which can create seriously biased estimates and subsequently leads to distorted inference. We specifically propose a model called the normal mean-variance Birnbaum-Sanders factor analysis (NMVBSFA), for missing information as an alternative. For parameter estimation, we develop an EM-type algorithm namely expectation conditional maximization either (ECME) represented by Liu and Rubin [16]. Throughout

this paper, data's missingness is assumed to be MAR with an ignorable mechanism [23,25,15]. The missingness is unrelated to the missing values. In this setup, parameters that govern the data model and the missing-data mechanism are distinct, and the likelihood inference can ignore the missingness mechanism. When the MAR condition holds, the full data log-likelihood and the observed data log-likelihood will give identical inferences for interest parameters. Our proposed procedure is also valid if the mechanism is missing completely at random since it is a particular case of MAR. Refer to Liu and Lin [17] and Lin et al. [14] for more details. To ease the computational burden, two indicator matrices that determine the observed and missing locations of each observation separately are introduced. This methodology was used to analyze a real dataset and simulation study.

The rest of this paper is organized as follows. In section 2, we establish the notation and briefly review some preliminaries of the NMVBSFA model. In section 3, we formulate the NMVBSFA model under an incomplete-data specification and present the development of ECME algorithm for parameter estimation and missing-data imputation. The methodologies are illustrated in section 4 with a real data example. We assume that the missing mechanism is the MAR to validate our proposed procedure for the real data analysis. We conduct two simulation studies in section 5 to examine the validity of the model and finite-sample properties of the ML estimators. Some concluding remarks are given in section 6.

2 Preliminaries

In this section, we provide some preliminaries and pave the road for our factor analysis model. First, we need to define the notation that will be used throughout the 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 generalized hyperbolic distribution introduced by Barndor-Nielsen and Halgreen [2] 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

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

where $K_\kappa(\cdot)$ denotes the modified Bessel function of the third kind of order κ, and $C=(\psi/\chi)^{\kappa/2}(\psi+\boldsymbol{\lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda})^{p/2-\kappa}/(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}K_\kappa(\sqrt{\psi\chi})$ is the normalizing constant. A p-variate normal mean-variance Birnbaum-Sanders distribution with location vector $\boldsymbol{\mu}$, scale covariance matrix $\boldsymbol{\Sigma}$, skewness vector $\boldsymbol{\lambda}$ and shape parameter α, denoted by $\boldsymbol{X}\sim NMVBS_p(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda},\alpha)$ has the

pdf

$$f_{NMVBS}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\lambda},\boldsymbol{\Sigma},\alpha) = \frac{1}{2} f_{GH_p}\left(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\lambda},\boldsymbol{\Sigma},\frac{1}{2},\frac{1}{\alpha^2},\frac{1}{\alpha^2}\right) + \frac{1}{2} f_{GH_p}\left(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\lambda},\boldsymbol{\Sigma},-\frac{1}{2},\frac{1}{\alpha^2},\frac{1}{\alpha^2}\right). \quad (4)$$

A two-level hierarchical representation of (4) is

$$\boldsymbol{X} \mid W = w \sim N_p(\boldsymbol{\mu} + w\boldsymbol{\lambda}, w\boldsymbol{\Sigma}),\ W \sim BS(\alpha, 1). \quad (5)$$

The pdf of BS$(\alpha, 1)$ then can be expressed [4] as

$$f_{BS}(w;\alpha,\beta=1) = \frac{1}{2} f_{GIG}\left(w;\frac{1}{2},\frac{1}{\alpha^2},\frac{1}{\alpha^2}\right) + \frac{1}{2} f_{GIG}\left(w;-\frac{1}{2},\frac{1}{\alpha^2},\frac{1}{\alpha^2}\right), \quad (6)$$

where $f_{GIG}(.)$ is the pdf of generalized inverse Gaussian (GIG) distribution (see [6]) with pdf

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

The NMVBSFA model is then given by

$$\boldsymbol{Y}_j = \boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{U}_j + \boldsymbol{\varepsilon}_j, \quad j = 1, \ldots, n, \quad (7)$$

along with the assumption of

$$\begin{bmatrix}\boldsymbol{U}_j \\ \boldsymbol{\varepsilon}_j\end{bmatrix} \sim NMVBS_{q+p}\left(\begin{bmatrix}-a_\alpha \boldsymbol{\Lambda}^{-1/2}\boldsymbol{\lambda} \\ \boldsymbol{0}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Lambda}^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{\lambda} \\ \boldsymbol{0}\end{bmatrix}, \alpha\right), \quad (8)$$

where $\boldsymbol{\mu}$ is a p-dimensional location vector, $\boldsymbol{B} \in \mathbb{R}^{p\times q}$ is factor loadings, $\boldsymbol{U}_j$ is a q-dimensional vector ($q < p$) of latent variables called *common factors*, $\boldsymbol{\varepsilon}_j \in \mathbb{R}^p$ is a p-dimensional vector of errors called *specific factors*, $\boldsymbol{D} = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ and $\boldsymbol{\Lambda} = a_\alpha \boldsymbol{I}_q + b_\alpha \boldsymbol{\lambda}\boldsymbol{\lambda}^\top$, such that,

$$W_j \sim BS(\alpha, 1),\ a_\alpha = E(W_j) = 1{+}0.5\alpha^2 \text{ and } b_\alpha = \text{Var}(W_j) = \alpha^2\left(1 + \frac{5}{4}\alpha^2\right).$$

According to (5), the proposed NMVBSFA model formulated by (7) and (8) allow the following two-level hierarchical representation

$$\begin{aligned}\boldsymbol{Y}_j \mid (W_j = w_j) &\sim N_p(\boldsymbol{\mu} - a_\alpha \boldsymbol{B}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{\lambda} + w_j \boldsymbol{B}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{\lambda}, w_j\boldsymbol{\Sigma}), \\ W_j &\sim BS(\alpha, 1). \end{aligned} \quad (9)$$

Combining the pdfs $f(\boldsymbol{y}_j \mid w_j)$ and $f(w_j)$ in (9) and integrating out the weight variable w_j yields the marginal distribution of $\boldsymbol{Y}_j$, given by

$$\boldsymbol{Y}_j \sim NMVBS_p(\boldsymbol{\mu} - a_\alpha\boldsymbol{\eta}, \boldsymbol{\Sigma}, \boldsymbol{\eta}, \alpha), \quad (10)$$

where $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{\Lambda}^{-1}\boldsymbol{B}^\top + \boldsymbol{D}$ and $\boldsymbol{\eta} = \boldsymbol{B}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{\lambda}$ is a p-dimensional reparameterized skewness parameter vector.

3 The NMVBSFA Model with Missing Information

3.1 Model formulation

For this section we refine the NMVBSFA model to accommodate incomplete data. To derive the estimating equations allowing for missing values, $\boldsymbol{Y}_j$ is partitioned into the observed part $\boldsymbol{Y}_j^o$ with dimensional p_j^o and the missing part $\boldsymbol{Y}_j^m$ with dimensional $(p - p_j^o)$. To simplify calculations, two auxiliary permutation matrices $\boldsymbol{O}_j \in \mathbb{R}^{(p_j^o \times p)}$ and $\boldsymbol{M}_j \in \mathbb{R}^{((p-p_j^o)\times p)}$ are introduced such that $\boldsymbol{Y}_j^o = \boldsymbol{O}_j \boldsymbol{Y}_j \in \mathbb{R}^{(p_j^o \times 1)}$ and $\boldsymbol{Y}_j^m = \boldsymbol{M}_j \boldsymbol{Y}_j \in \mathbb{R}^{((p-p_j^o)\times 1)}$ for $j = 1, \ldots, n$. 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 $\boldsymbol{O}_j^\top \boldsymbol{O}_j + \boldsymbol{M}_j^\top \boldsymbol{M}_j = \boldsymbol{I}_p$. To obtain closed-form expressions for the estimators, we follow the strategy of [7] by proposing the invariant transformations:

$$\tilde{\boldsymbol{B}}_i \stackrel{\Delta}{=} \boldsymbol{B}\boldsymbol{\Lambda}^{-1/2} \quad \text{and} \quad \tilde{\boldsymbol{U}}_j \stackrel{\Delta}{=} \boldsymbol{\Lambda}^{1/2}\boldsymbol{U}_j. \tag{11}$$

The following proposition is useful for evaluating the required conditional expectation in the E-step for the computational algorithm described in the next section.

Proposition 1 *From* (7) *and* (9) *of NMVBSFA, we have that:*

(a) The marginal distribution of the observed component $\boldsymbol{Y}_j^o$ *is*

$$\boldsymbol{Y}_j^o \sim NMVBS_{p_j^o}(\boldsymbol{\mu}_j^o - a_\alpha \boldsymbol{\eta}_j^o, \boldsymbol{\Sigma}_j^{oo}, \boldsymbol{\eta}_j^o, \alpha). \tag{12}$$

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

$$\boldsymbol{Y}_j^o \mid w_j \sim N_{p_j^o}(\boldsymbol{\mu}_j^o - a_\alpha \boldsymbol{\eta}_j^o + w_j \boldsymbol{\eta}_j^o, w_j \boldsymbol{\Sigma}_j^{oo}),$$

where $\boldsymbol{\mu}_j^o = \boldsymbol{O}_j \boldsymbol{\mu}$, $\boldsymbol{\eta}_j^o = \boldsymbol{O}_j \boldsymbol{\eta}$ *and* $\boldsymbol{\Sigma}_j^{oo} = \boldsymbol{O}_j \Sigma \boldsymbol{O}_j^\top$.

(c) The conditional distribution of $\boldsymbol{Y}_j^m$ *given* $\boldsymbol{y}_j^o$, $\tilde{\boldsymbol{u}}_j$ *and* w_j *is*

$$\boldsymbol{Y}_j^m \mid (\boldsymbol{y}_j^o, \tilde{\boldsymbol{u}}_j, w_j) \sim N_{p-p_j^o}(\boldsymbol{\varphi}_j^{m.o}, W_j \boldsymbol{D}_j^{mm.o})$$

where

$$\begin{aligned}
\boldsymbol{\varphi}_j^{m.o} &= \boldsymbol{M}_j \left[\boldsymbol{\mu} + \tilde{\boldsymbol{B}}\tilde{\boldsymbol{u}}_j + \boldsymbol{D}\boldsymbol{C}_j^{oo}(\boldsymbol{y}_j - \boldsymbol{\mu} - \tilde{\boldsymbol{B}}\tilde{\boldsymbol{u}}_j)\right], \\
\boldsymbol{D}_j^{mm.o} &= \boldsymbol{M}_j \left(\boldsymbol{I}_p - \boldsymbol{D}\boldsymbol{C}_{ij}^{oo}\right) \boldsymbol{D}\boldsymbol{M}_j^\top, \\
\boldsymbol{C}_j^{oo} &= \boldsymbol{O}_j^\top \left(\boldsymbol{O}_j \boldsymbol{D} \boldsymbol{O}_j^\top\right)^{-1} \boldsymbol{O}_j.
\end{aligned}$$

(d)

$$\begin{aligned}
f(w_j \mid \boldsymbol{y}_j^o) = \; & \pi_j^o f_{GIG}\left(w_j; \frac{1-p_j^o}{2}, \chi_j^o, \psi_j^o\right) \\
& + (1-\pi_j^o) f_{GIG}\left(w_j; \frac{-1-p_j^o}{2}, \chi_j^o, \psi_j^o\right),
\end{aligned} \tag{13}$$

where

$$\chi_j^o = (\boldsymbol{y}_j - \boldsymbol{\mu} + a_\alpha \boldsymbol{\eta})^\top \boldsymbol{S}_j^{oo} (\boldsymbol{y}_j - \boldsymbol{\mu} + a_\alpha \boldsymbol{\eta}) + \alpha^{-2}, \quad \boldsymbol{S}_j^{oo} = \boldsymbol{O}_j \boldsymbol{\Sigma}_j^{oo} \boldsymbol{O}_j^\top,$$
$$\psi_j^o = \boldsymbol{\eta} \boldsymbol{S}_j^{oo} \boldsymbol{\eta} + \alpha^{-2}, \quad \pi_j^o = \frac{f_{GH_p}\left(\boldsymbol{y}_j^o; \boldsymbol{\mu}_j^o - a_\alpha \boldsymbol{\eta}_j^o, \boldsymbol{\eta}_j^o, \boldsymbol{\Sigma}_j^{oo}, 0.5, \alpha^{-2}, \alpha^{-2}\right)}{2 f_{\text{NMVBS}}\left(\boldsymbol{y}_j^o; \boldsymbol{\mu}_j^o - a_\alpha \boldsymbol{\eta}_j^o, \boldsymbol{\Sigma}_j^{oo}, \boldsymbol{\eta}_j^o, \alpha\right)}$$
. (14)

(e) $\boldsymbol{Y}_j^o \mid (\tilde{\boldsymbol{u}}_j, w_j) \sim N_{p_j^o}(\boldsymbol{\mu}_j^o + \tilde{\boldsymbol{B}}_j^o \tilde{\boldsymbol{u}}_j, w_j \boldsymbol{D}_j^{oo})$ *where* $\boldsymbol{D}_j^{oo} = \boldsymbol{O}_j \boldsymbol{D} \boldsymbol{O}_j^\top$.
(f) $\tilde{\boldsymbol{U}}_j \mid (\boldsymbol{y}_j^o, w_j) \sim N_q(\boldsymbol{q}_j^o, w_j \boldsymbol{R}_j^{oo})$ *where* $\boldsymbol{q}_j^o = \boldsymbol{R}_j^{oo}\left\{\boldsymbol{b}_j^o + \boldsymbol{\lambda}(w_j - a_\alpha)\right\}$, $\boldsymbol{b}_j^o = \tilde{\boldsymbol{B}}^\top \boldsymbol{C}_j^{oo}(\boldsymbol{y}_j - \boldsymbol{\mu})$ *and* $\boldsymbol{R}_j^{oo} = (\boldsymbol{I}_q + \tilde{\boldsymbol{B}}^\top \boldsymbol{C}_j^{oo} \tilde{\boldsymbol{B}})^{-1}$.
(g) Based on part (d) and (f), we have the following conditional expectations:

$$E(W_j^r \mid \boldsymbol{y}_j^o) = \left(\frac{\chi_j^o}{\psi_j^o}\right)^{r/2} \left\{ \frac{K_{(1-p_j^o)/2+r}\left(\sqrt{\psi_j^o \chi_j^o}\right)}{K_{(1-p_j^o)/2}\left(\sqrt{\psi_j^o \chi_j^o}\right)} + \frac{K_{-(1+p_j^o)/2+r}\left(\sqrt{\psi_j^o \chi_j^o}\right)}{K_{-(1+p_j^o)/2}\left(\sqrt{\psi_j^o \chi_j^o}\right)} \right\},$$
$r = \pm 1$ (15)

$$E(\tilde{\boldsymbol{U}}_{ij} \mid \boldsymbol{y}_j^o) = \boldsymbol{R}_j^{oo}\left\{\boldsymbol{b}_j^o + \boldsymbol{\lambda}\left(E(W_j \mid \boldsymbol{y}_j^o) - a_\alpha\right)\right\} \quad (16)$$

$$E(W_j^{-1}\tilde{\boldsymbol{U}}_j \mid \boldsymbol{y}_j^o) = \boldsymbol{R}_j^{oo}\left\{\boldsymbol{b}_j^o E(W_j^{-1} \mid \boldsymbol{y}_j^o) + \boldsymbol{\lambda}\big(1 - a_\alpha E(W_j^{-1} \mid \boldsymbol{y}_j^o)\big)\right\}. \quad (17)$$

and

$$E(W_j^{-1}\tilde{\boldsymbol{U}}_j\tilde{\boldsymbol{U}}_j^\top \mid \boldsymbol{y}_j^o) = \Big\{ E(W_j^{-1}\tilde{\boldsymbol{U}}_j \mid \boldsymbol{y}_j^o)\boldsymbol{b}_j^{o\top} + \left[E(\tilde{\boldsymbol{U}}_j \mid \boldsymbol{y}_j^o) - a_\alpha E(W_j^{-1}\tilde{\boldsymbol{U}}_j \mid \boldsymbol{y}_j^o)\right]\boldsymbol{\lambda}^\top + \boldsymbol{I}_q \Big\} \boldsymbol{R}_j^{oo}. \quad (18)$$

Proof See Appendix.

The NMVBSFA model with missing data suffers from the identifiability problem concerning the factor loading matrix $\boldsymbol{B}$. One way to overcome this problem proposed by Lawley [10] is to add the restriction such that $\boldsymbol{B}^\top \boldsymbol{D}^{-1} \boldsymbol{B}$ is a diagonal matrix with its diagonal elements arranged in decreasing order. The second method used here is to constrain $\boldsymbol{B}$ so that its upper-right triangle is zero and its diagonals are strictly positive [5]. In both approaches, $q(q-1)/2$ constraints are imposed on $\boldsymbol{B}$. However, the loading elements' identifiability is not necessarily a concern in practice when carrying out the EM-based ML estimation [24] because there is a consistent sequence of roots to the likelihood equation. Notice that the number of factors fulfills the constraint $(p-q)^2 \geq (p+q)$, as suggested by Lawley and Maxwell [11]. Interested readers may also refer to Liu and Lin [17] and Wang et al. [28] for more discussion.

3.2 Parameter estimation via the ECME algorithm

In this section, we estimate parameters of the NMVBSFA model by developing an ECME algorithm, which is an extension to the ECM algorithm [19]. The key feature of the ECME is that it replaces some CM-steps of ECM with CML-steps that maximize the corresponding constrained actual likelihood function instead. To simplify the notation, we denote the complete data by $\boldsymbol{y}_c = (\boldsymbol{y}^o, \boldsymbol{y}^m, \tilde{\boldsymbol{U}}, \boldsymbol{W})$, where $\boldsymbol{y}^o = (\boldsymbol{y}_1^o, \ldots, \boldsymbol{y}_n^o)$ are the observed parts of the experimental data. In contrast, $\boldsymbol{y}^m = (\boldsymbol{y}_1^m, \ldots, \boldsymbol{y}_n^m)$, $\tilde{\boldsymbol{U}} = (\tilde{\boldsymbol{U}}_1, \ldots, \tilde{\boldsymbol{U}}_n)$ and $\boldsymbol{W} = (W_1, \ldots, W_n)$ are viewed as the hidden data in the EM framework. The complete-data log-likelihood function of $\boldsymbol{\theta} = (\boldsymbol{\mu}, \tilde{\boldsymbol{B}}, \boldsymbol{D}, \boldsymbol{\lambda}, \alpha)$, can be written as

$$\begin{aligned}\ell_c(\boldsymbol{\theta} \mid \boldsymbol{y}_c) = &- n\log\alpha - \sum_{j=1}^{n} -\frac{(W_j-1)^2}{2\alpha^2 W_j} - \frac{n}{2}\log|\boldsymbol{D}| \\ &- \frac{1}{2}\sum_{j=1}^{n} W_j^{-1}\big(\boldsymbol{y}_j - \boldsymbol{\mu} - \tilde{\boldsymbol{B}}\tilde{\boldsymbol{U}}_j\big)^\top \boldsymbol{D}^{-1}\big(\boldsymbol{y}_j - \boldsymbol{\mu} - \tilde{\boldsymbol{B}}\tilde{\boldsymbol{U}}_j\big) \\ &- \frac{1}{2}\sum_{j=1}^{n}\Big(\big(W_j - 2a_\alpha + W_j^{-1}a_\alpha^2\big)\boldsymbol{\lambda}\boldsymbol{\lambda}^\top - 2\boldsymbol{\lambda}(\tilde{\boldsymbol{U}}_j - a_\alpha W_j^{-1}\tilde{\boldsymbol{U}}_j)^\top\Big).\end{aligned} \tag{19}$$

On the k^{th} iteration of the E-step, we compute the expected value of the $\ell_c(\boldsymbol{\theta} \mid \boldsymbol{y}_c)$ conditional on the observed data $\boldsymbol{y}_j^o$ and current parameter estimates $\hat{\boldsymbol{\theta}}^{(k)}$, called the Q function:

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

where $\hat{\boldsymbol{\theta}}^{(k)} = (\hat{\boldsymbol{\mu}}^{(k)}, \hat{\tilde{\boldsymbol{B}}}^{(k)}, \hat{\boldsymbol{D}}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k)}, \hat{\alpha}^{(k)})$. To evaluate (20), we need to calculate the following conditional expectations:

$$\begin{aligned}&\hat{w}_j^{(k)} = E(W_j \mid \boldsymbol{y}_j^o, \hat{\boldsymbol{\theta}}^{(k)}), \quad \hat{t}_j^{(k)} = E(W_j^{-1} \mid \boldsymbol{y}_j^o, \hat{\boldsymbol{\theta}}^{(k)}), \quad \hat{\boldsymbol{\zeta}}_{0j}^{(k)} = E(\tilde{\boldsymbol{U}}_j \mid \boldsymbol{y}_j^o, \hat{\boldsymbol{\theta}}^{(k)}), \\ &\hat{\boldsymbol{\zeta}}_{1j}^{(k)} = E(W_j^{-1}\tilde{\boldsymbol{U}}_j \mid \boldsymbol{y}_j^o, \hat{\boldsymbol{\theta}}^{(k)}), \quad \hat{\boldsymbol{\Omega}}_j^{(k)} = E(W_j^{-1}\tilde{\boldsymbol{U}}_j\tilde{\boldsymbol{U}}_j^\top \mid \boldsymbol{y}_j^o, \hat{\boldsymbol{\theta}}^{(k)}),\end{aligned} \tag{21}$$

which are obtainable directly by Eqs. (15)-(18) given in Proposition 1 with all the elements in $\boldsymbol{\theta}$ replaced by $\hat{\boldsymbol{\theta}}^{(k)}$.

On the $(k+1)^{\text{th}}$ iteration of the CM-steps, the updated formula for model parameters are summarized below.

CM-step 1: Calculate

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

where $\hat{\boldsymbol{q}}_j^{(k)} = \hat{\boldsymbol{\mu}}^{(k)} + \hat{\boldsymbol{D}}^{(k)}\hat{\boldsymbol{C}}_j^{oo^{(k)}}(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}^{(k)})$.

CM-step 2: Given $\boldsymbol{\mu} = \hat{\boldsymbol{\mu}}^{(k+1)}$, update $\hat{\tilde{\boldsymbol{B}}}^{(k)}$ by maximizing (20) over $\tilde{\boldsymbol{B}}$, which gives

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

where $\hat{\boldsymbol{E}}_j^{oo^{(k)}} = \left(\boldsymbol{I}_p - \hat{\boldsymbol{D}}^{(k)}\hat{\boldsymbol{C}}_j^{oo^{(k)}}\right)\hat{\tilde{\boldsymbol{B}}}^{(k)}$.

CM-step 3: Given $\boldsymbol{\mu} = \hat{\boldsymbol{\mu}}^{(k+1)}$ and $\tilde{\boldsymbol{B}} = \hat{\tilde{\boldsymbol{B}}}^{(k)}$, update $\hat{\boldsymbol{D}}^{(k)}$ by maximizing (20) over $\boldsymbol{D}$

$$\hat{\boldsymbol{D}}^{(k+1)} = \frac{1}{n}\text{Diag}\left(\sum_{j=1}^{n}\hat{\boldsymbol{\Upsilon}}_j^{(k+1)}\right),$$

where

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

CM-step 4: Calculate

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

where $a_{\hat{\alpha}^{(k)}} = 1 + 0.5\hat{\alpha}^{(k)2}$.

Since there is no closed-form solution for updating α, we adopt the so-called 'CML-step' to maximize the restricted actual log-likelihood function. That is,

CML-step 5:

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

where $\hat{\boldsymbol{\mu}}^{o^{(k+1)}}$, $\hat{\boldsymbol{\eta}}_j^{o^{(k+1)}}$ and $\hat{\boldsymbol{\Sigma}}_j^{oo^{(k+1)}}$ are $\boldsymbol{\mu}_j^o$, $\boldsymbol{\eta}_j^o$ and $\boldsymbol{\Sigma}_j^{oo}$ in (12), respectively, evaluated at the current estimates at the start of the $(k+1)^{\text{th}}$ iteration. The above optimization procedure can be easily done by utilizing the built-in R function `optim` in range $(0, 100)$.

To monitor the convergence, by using the likelihood increasing property of the ECM algorithm, the default stopping rule is $\ell(\hat{\boldsymbol{\Theta}}^{(k)}; \boldsymbol{y}) - \ell(\hat{\boldsymbol{\Theta}}^{(k-1)}; \boldsymbol{y}) < \epsilon$ where $\ell(\hat{\boldsymbol{\Theta}}^{(k)}; \boldsymbol{y})$ is the log-likelihood evaluated for the parameter estimation at iteration (k) and ϵ is the user-specified tolerance. In our analysis, the algorithm is terminated if the maximum number of iterations $k_{\max} = 5000$ is reached or when the relative difference between two successive log-likelihood values is less than $\epsilon = 10^{-6}$. Upon convergence, the resulting maximum likelihood estimates are denoted by $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{B}}, \hat{\boldsymbol{D}}, \hat{\boldsymbol{\lambda}}, \hat{\alpha})$, where $\hat{\boldsymbol{B}} = \hat{\tilde{\boldsymbol{B}}}\hat{\boldsymbol{\Lambda}}^{1/2}$, and $\hat{\boldsymbol{\Lambda}}$ is $\boldsymbol{\Lambda} = a_\alpha \boldsymbol{I}_q + b_\alpha \boldsymbol{\lambda}\boldsymbol{\lambda}^\top$ with $\boldsymbol{\lambda}$ and α replaced by their estimates.

3.3 Prediction of factor scores and missing values

In addition to estimate the model parameters, factor scores can be predicted and used in subsequent analyses. For instance, researchers may want to use the factor information for data reconstruction in lower-dimensional subspaces [13]. The conditional predictor of unobserved factor scores corresponding to $\boldsymbol{y}_j^o$ is

$$\hat{\boldsymbol{u}}_j = E(\boldsymbol{U}_j \mid \boldsymbol{y}_j^o, \hat{\boldsymbol{\theta}}) = \hat{\boldsymbol{\Lambda}}^{-1/2} E(\tilde{\boldsymbol{U}}_j \mid \boldsymbol{y}_j^o, \hat{\boldsymbol{\theta}}), \tag{22}$$

where $E(\tilde{\boldsymbol{U}}_j \mid \boldsymbol{y}_j^o, \hat{\boldsymbol{\theta}})$ is given by (16) evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$.

Furthermore, filling in the missing data with plausible values is an important task for creating a complete dataset in order to apply standard statistical methods. The maximum likelihood approach provides a simple way of imputing one value for each missing datum, referred to as single imputation. As a by-product of our ECME algorithm, conditional prediction of $\boldsymbol{y}_j^m$ can be obtained via

$$\hat{\boldsymbol{y}}_j^m = E(\boldsymbol{Y}_j^m \mid \boldsymbol{y}_j^o, \hat{\boldsymbol{\theta}}) = \boldsymbol{M}_j \left(\hat{\boldsymbol{\mu}} + \hat{\boldsymbol{B}}\hat{\boldsymbol{u}}_j + \hat{\boldsymbol{D}}\hat{\boldsymbol{C}}_j^{oo}(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}} - \hat{\boldsymbol{B}}\hat{\boldsymbol{u}}_j) \right). \tag{23}$$

3.4 Practical implementation issues

Good initial values of parameter estimates for the ECME algorithm can speed up or enable the convergence. In particular, when the number of latent factors is over-specified or the proportion of missing values is too large, converging at the boundary of the parameter space might occur due to a poor choice of initial value $\hat{\boldsymbol{\theta}}^{(0)}$. When the raw data contains missing values, we first fill in the missing values of the kth variable, say y_{jk}^m, by the sample mean of observed values of corresponding variable $\bar{y}_{\cdot k} = \sum_{j=1}^{n_k^o} y_{jk}^o / n_k^o$, where n_k^o is the number of observations in the kth variable, for $k = 1, \ldots, p$. A strategy for specifying $\hat{\boldsymbol{\mu}}^{(0)}$, $\hat{\boldsymbol{B}}^{(0)}$, $\hat{\boldsymbol{D}}^{(0)}$, $\hat{\boldsymbol{\lambda}}^{(0)}$ and $\hat{\alpha}^{(0)}$ are described in [7].

To determine the most plausible value of q, we adopt the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) taking the form of

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

Table 1 Acronyms used for all factor analysis models under comparison.

FA	Gaussian factor analysis	Liu and Lin [17] for $\boldsymbol{\lambda} = \mathbf{0}$
SNFA	Skew normal factor analysis	Liu and Lin [17]
STFA	Skew-t factor analysis	Wang et al. [28]
GHFA	Generalized hyperbolic factor analysis	Wei et al. [29]
NMVBSFA	Normal mean-variance Birnbaum-Sanders factor analysis	Proposed in this paper

such that ℓ_{max} is the maximized log-likelihood, and m is the number of free parameters in the considered model. The smaller the AIC and BIC values, the most parsimonious model is favored over more complex ones.

4 Automobile Data Analysis

In the subsequent sections, we compare the performance of some factor analysis models. For convenience, we use the abbreviations in Table 1. Here, we consider the automobile data studied by Kibler et al. [9]. This dataset, available at the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml). There are $n = 205$ inventories with $p = 15$ continuous attributes, denoted by $X_1, \ldots, X_{15}$, summarized in Table 2. In addition, most attributes exhibit mild to strong asymmetry and light to extremely heavy tails, indicating that the subpopulation deviate substantially from normality. There are 45 observations what have at least one missing value. Without considering the missing data, using the generalized Shapiro-Wilk test for multivariate normality proposed by Villasenor Alva and Estrada [27], the p-value corresponding to the test statistic 0.912 is smaller than 2e-16, confirming a serious departure from normality of the data. To compare the performance of the proposed model, we implement

Table 2 An overview of the automobile data.

Attributes	Description	Number of missing values	Sample mean	Sample sd	Sample skewness	Sample kurtosis
x_1	normalized-losses (Nor-Los)	41	122.0	35.4	0.8	3.5
x_2	wheel-base (Whe)	0	98.8	6.0	1.0	4.0
x_3	length (Len)	0	174.0	12.3	0.2	2.9
x_4	width (Wid)	0	65.9	2.1	0.9	3.7
x_5	height (Hei)	0	53.7	2.4	0.1	2.5
x_6	curb-weight (Cur)	0	2555.6	520.7	0.7	2.9
x_7	engine-size (Eng)	0	126.9	41.6	1.9	8.1
x_8	bore (Bor)	4	3.3	0.3	0.1	2.2
x_9	stroke (Str)	4	3.3	0.3	-0.7	5.0
x_{10}	compression-ratio (Com-Rat)	0	10.1	4.0	2.6	8.1
x_{11}	horsepower (Hor)	2	104.3	39.7	1.4	5.5
x_{12}	peak-rpm (Peak-rpm)	2	5125.4	479.3	0.1	3.0
x_{13}	city-mpg (City-mpg)	0	25.2	6.5	0.7	3.5
x_{14}	highway-mpg (Hig-mpg)	0	30.8	6.9	0.5	3.4
x_{15}	price (Price)	4	13207.1	7947.1	1.8	6.1

several factor analyzers to compare with the NMVBSFA model with q ranging from 2 to $q_{\max} = 10$ to fit this dataset.

Table 3 lists the maximum likelihood results, including the maximized log-likelihood values and the number of parameters together with the BIC

Table 3 Estimation performance for factor models in Table 1 fitted to the automobile data.

q	FA					SNFA				
		Original data		Standardized data			Original data		Standardized data	
	m	$\ell_{\max}$	BIC	$\ell_{\max}$	BIC	m	$\ell_{\max}$	BIC	$\ell_{\max}$	BIC
2	59	-10977.83	22269.72	-2745.97	5806.03	61	-10945.39	22215.49	-2712.25	5749.21
3	72	-10794.46	21972.17	-2562.61	5508.48	75	-10743.42	21886.08	-2511.58	5422.38
4	84	-10756.16	21959.45	-2524.31	5495.76	88	-10703.15	21874.73	-2471.30	5411.04
5	95	-10730.97	21967.63	-2499.12	5503.93	100	-10677.19	21886.68	-2445.34	5422.99
6	105	-10709.45	21977.82	-2474.68	5508.27	111	-10616.59	21824.04	-2409.65	5410.16
7	114	-10716.11	22039.03	-2454.90	5516.62	121	-10722.51	22089.10	-2385.93	5415.95
8	122	-10711.59	22072.59	-2442.37	5534.16	130	-10679.06	22050.10	-2355.50	5402.99
9	129	-10697.30	22081.26	-2433.71	5554.09	138	-10666.91	22068.40	-2355.43	5445.45
10	135	-10693.41	22105.42	-2433.24	5585.09	145	-10666.28	22104.40	-2354.95	5481.74
q	STFA					GHFA				
		Original data		Standardized data			Original data		Standardized data	
	m	$\ell_{\max}$	BIC	$\ell_{\max}$	BIC	m	$\ell_{\max}$	BIC	$\ell_{\max}$	BIC
2	62	-10779.64	21889.31	-2547.80	5425.62	76	-10795.99	21927.34	-2564.15	5463.65
3	76	-10632.69	21669.94	-2400.85	5206.24	89	-10624.78	21659.44	-2392.93	5195.74
4	89	-10589.90	21653.54	-2358.05	5189.85	101	-10585.78	21650.63	-2353.93	5186.93
5	101	-10526.19	21590.00	-2294.34	5126.31	112	-10500.20	21543.35	-2268.35	5079.66
6	112	-10502.20	21600.58	-2270.08	5136.35	122	-10474.93	21551.36	-2243.08	5087.67
7	122	-10601.24	21851.88	-2277.32	5204.05	131	-10584.10	21822.94	-2217.55	5089.83
8	131	-10605.09	21907.50	-2231.66	5160.64	139	-10589.27	21881.19	-2205.15	5112.95
9	139	-10585.97	21911.85	-2223.35	5186.60	146	-10569.57	21884.37	-2187.60	5120.43
10	146	-10595.22	21967.61	-2229.96	5237.08	152	-10574.88	21932.25	-2176.11	5134.72
q	NMVBSFA									
		Original data		Standardized data						
	m	$\ell_{\max}$	BIC	$\ell_{\max}$	BIC					
2	62	-10778.25	21886.53	-2546.41	5422.85					
3	76	-10621.88	21648.32	-2390.04	5184.62					
4	89	-10582.51	21638.77	-2350.66	5175.07					
5	101	-10501.17	21539.97	-2269.32	5076.27					
6	112	-10467.73	**21531.65**	-2235.77	**5067.72**					
7	122	-10541.64	21732.69	-2216.39	5082.19					
8	131	-10543.52	21784.35	-2195.88	5089.08					
9	139	-10525.12	21790.14	-2184.81	5109.53					
10	146	-10518.64	21814.44	-2171.06	5119.28					

values compared to the other models; listed in Table 1. From this table, the NMVBSFA with $q = 6$ provides the best fitting performance (for original data BIC=21531.65, for standard data BIC=5067.72) for this dataset. This example demonstrates that the NMVBSFA provides improvements over the other methods because it provides more flexibility in capturing non-normal features. Figure 1 graphs the factor-5 scores estimated by using (22) and we observe that the estimated factor scores obtained by the NMVBSFA model appear to be more suitable as its skewness and kurtosis are larger compared to that of the other two models.

Figure 2 depicts the pairwise scatter plots of the 57 missing values predicted by using (23) for the model in Table 1. We find that either FA and SNFA or STFA models offer very similar imputed values, while the GHFA and NMVBSFA models provide rather different imputations. Therefore, the NMVBSFA and GHFA models have similar performance for the analysis of missing values. Thus, we are interested in whether the estimated factor scores are affected by the assumed underlying factor distribution. Figure 3 depicts the scatter plots of factor scores estimated using the NMVBSFA against GHFA according to equation (22) and marginal histograms on the plot's horizontal and vertical axes. Comparing the spread of estimated factor scores, the two models appear to provide somewhat different results. The factor scores' marginal histograms look skewed and have long tails, indicating that the NMVBSFA model may be more appropriate for capturing the true latent factors.

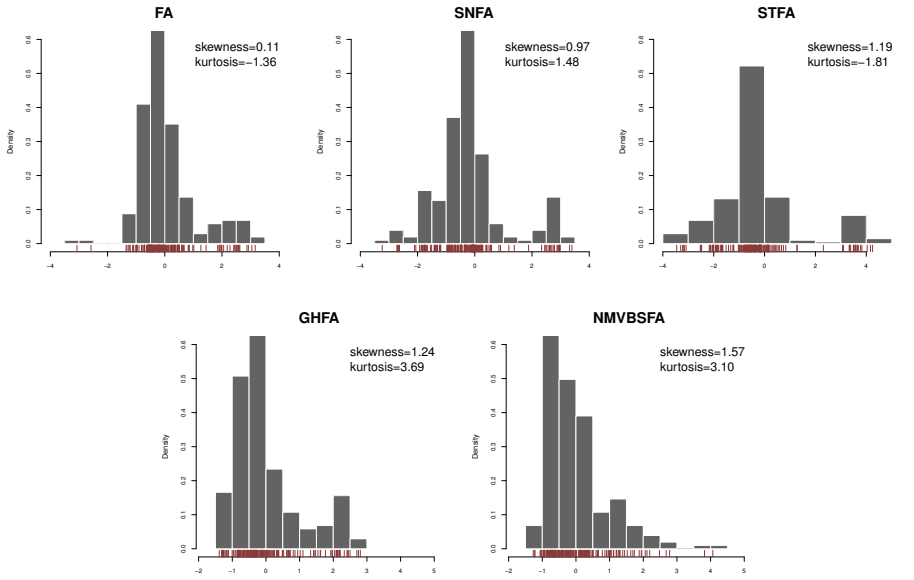

Fig. 1 Histograms of the estimated 5nd factor scores obtained of factor analysis models

Figures 4 and 5 display the fitted contours of SNFA and STFA models versus the NMVBSFA model obtained by marginalization of the fitted distributions superimposed on the scatter points for seven pairs of variables, mainly focusing on Price (15^{th} attribute) versus the remaining attributes, where the missing values are imputed by using (23). As can be seen in Figures 4 and 5, the contours of NMVBSFA appear to match the scattering shape more ideally than SNFA or STFA.

5 Monte Carlo simulations

5.1 Simulation 1

The first simulation experiment aims at examining the finite sample behaviour of ML estimators and their standard errors. We generate $M = 500$ samples from the model (7) with $q = 1$ with different sample sizes n =100, 500 and 1000. The true parameters values are specified as $\boldsymbol{\mu} = (12, 13, 15, 13, 11)^\top$, $\boldsymbol{B} = (4, 5, 2, 4, 6)^\top$, $\boldsymbol{D} = \text{diag}(1, 2, 3, 4, 5)$, $\lambda = 2$, and $\alpha = 0.5$ throughout this experiment. Each simulated dataset was fitted under the true model via the ECME algorithm. The average values of the maximum likelihood estimates across all samples were computed. To proceed with the experimental study, synthetic missing values are introduced to the simulated data under MAR mechanism. In the MAR experiment, missing items are obtained by deleting

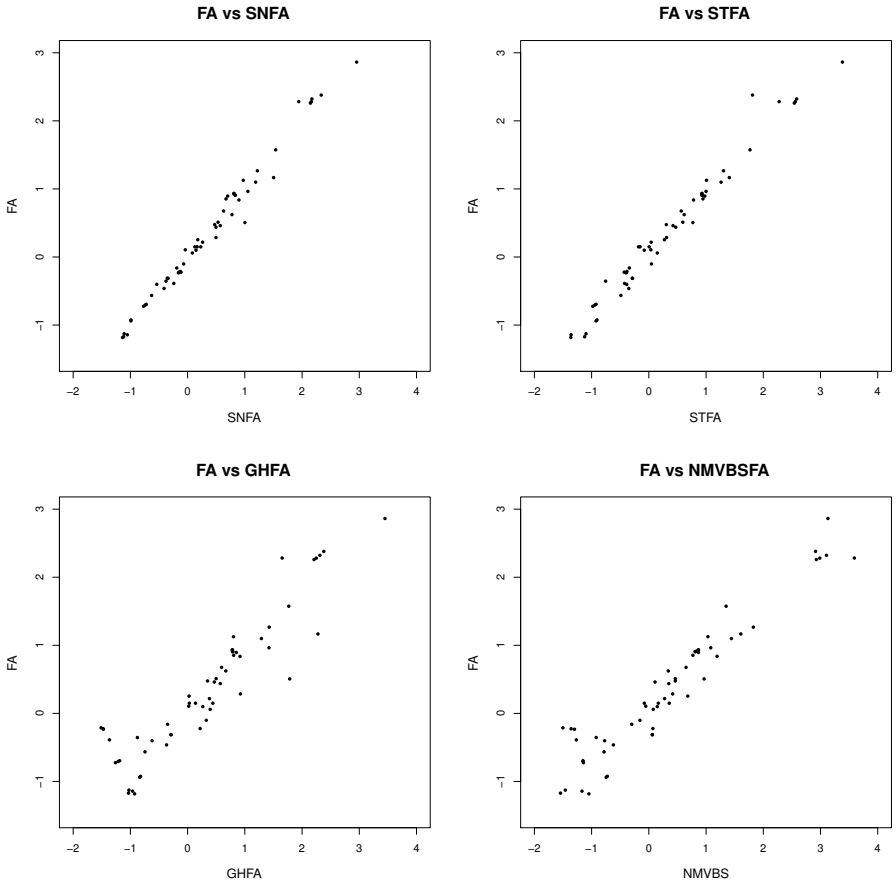

Fig. 2 Scatter plots of imputed missing values using the factor analysis models Table 1 for the automobile data

at random under low (r = 10%), moderate (r = 20%) and relatively high (r = 30%) rates of missingness.

To investigate the estimation accuracies, we compute the absolute relative bias (RB) and root mean squared error (RMSE) respectively given by

$$\mathrm{RB}(\hat{\theta}_i) = \frac{1}{M}\sum_{r=1}^{M}\left|\frac{\hat{\theta}_i^{(r)} - \theta_i}{\theta_i}\right| \quad \text{and} \quad \mathrm{RMSE}(\hat{\theta}_i) = \sqrt{\frac{1}{M}\sum_{r=1}^{M}\left(\hat{\theta}_i^{(r)} - \theta_i\right)^2},$$

where $\hat{\theta}_i^{(r)}$ is the estimate of a specific parameter θ_i at the rth replication. The results are given in Table 4. From Table 4, the RB and RMSE values for all parameters are getting smaller when the sample size increases. This shows the maximum likelihood estimates obtained by the proposed estimating procedure are close to the real values when we increase the sample size. Also, the RB

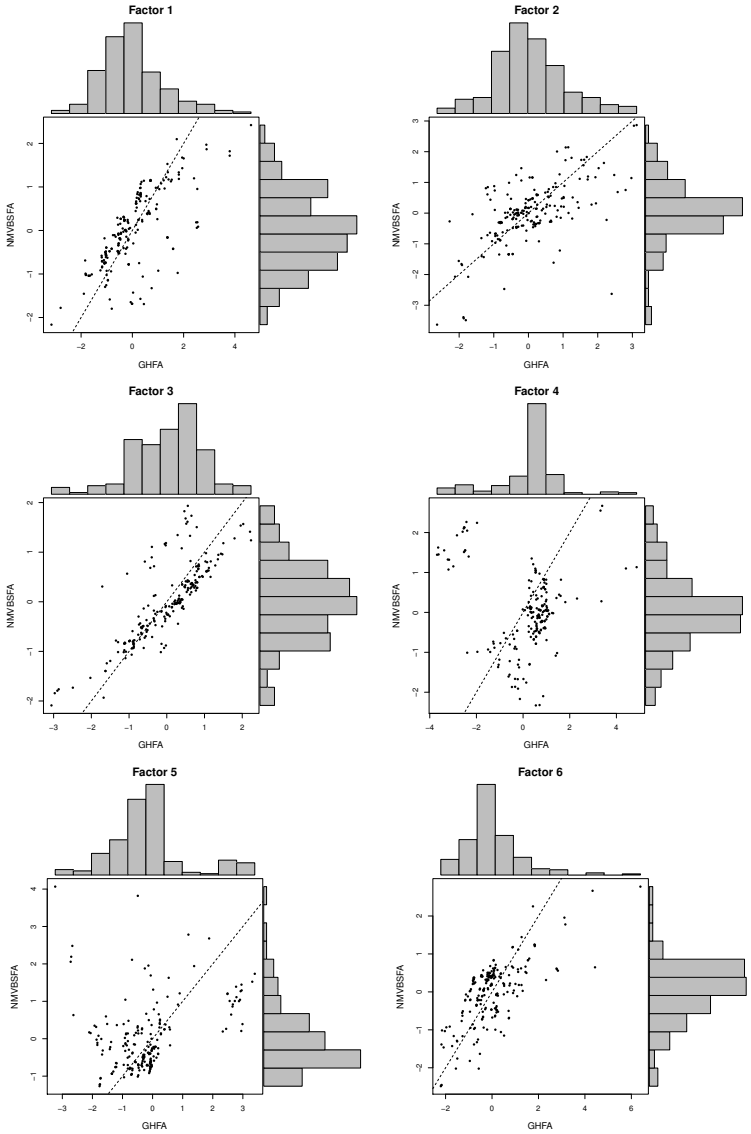

Fig. 3 Scatter-histogram plots of factor scores obtained from the fitted GHFA and NMVB-SFA models

and RMSE values for all parameters are getting bigger when the missingness rate rises.

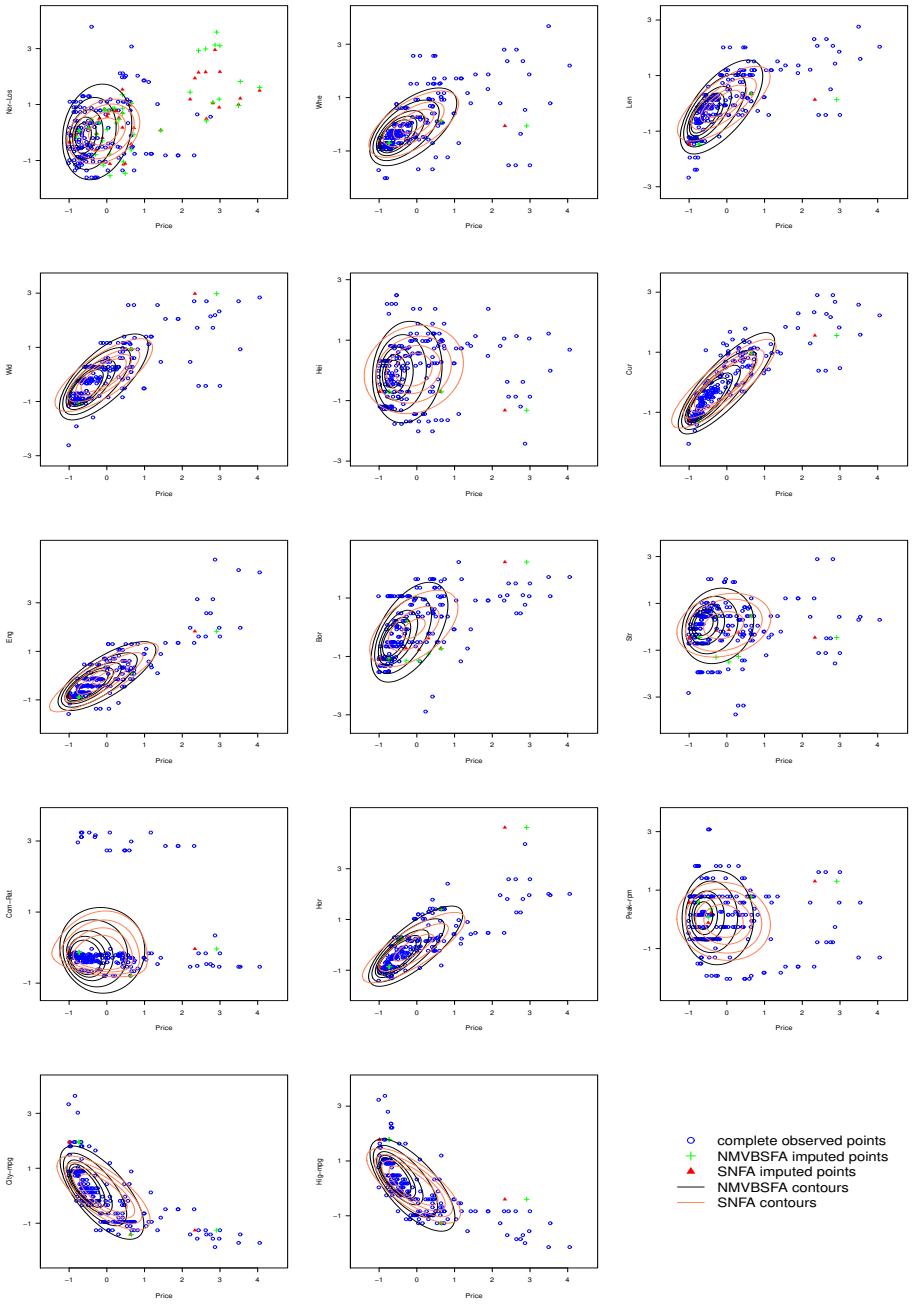

Fig. 4 Scatter plots superimposed on fitted SNFA and NMVBSFA contours for Fourteen pairs of variables of the automobile data

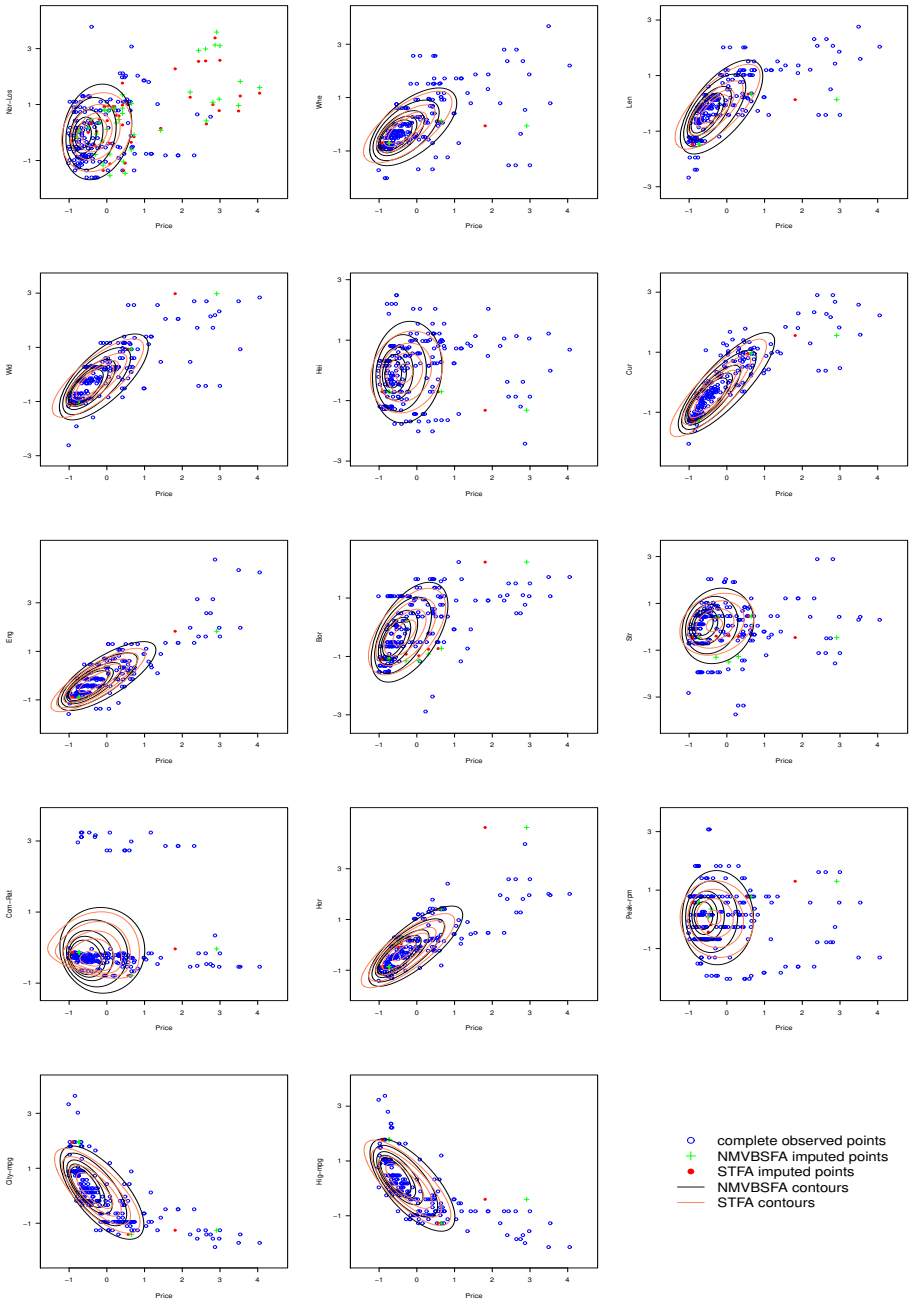

Fig. 5 Scatter plots superimposed on fitted STFA and NMVBSFA contours for fourteen pairs of variables of the automobile data

Table 4 Simulation results for assessing the asymptotic properties of parameter estimates.

n	r	Measure	$\mu_1(12)$	$\mu_2(13)$	$\mu_3(15)$	$\mu_4(13)$	$\mu_5(11)$	$b_1(4)$	$b_2(5)$	$b_3(2)$	$b_4(4)$	$b_5(6)$	$d_{11}(1)$	$d_{22}(2)$	$d_{33}(3)$	$d_{44}(4)$	$d_{55}(5)$	$\lambda_1(2)$	$\alpha(0.5)$
100	10%	RB	0.0255	0.0291	0.0132	0.0259	0.0466	0.0724	0.0747	0.1011	0.0848	0.0774	0.4792	0.3699	0.1591	0.2210	0.2877	0.9712	0.3466
		RMSE	0.1515	0.2484	0.0618	0.1876	0.4093	0.1344	0.2128	0.0691	0.1838	0.3539	0.3614	0.8357	0.3601	1.3323	3.3782	7.9787	0.0432
	20%	RB	0.0266	0.0299	0.0132	0.0266	0.0468	0.0727	0.0769	0.1079	0.0865	0.0780	0.5053	0.3791	0.1671	0.2445	0.3111	0.9537	0.3539
		RMSE	0.1609	0.2497	0.0625	0.2013	0.4182	0.1350	0.2176	0.0772	0.1893	0.3669	0.3909	0.9431	0.3883	1.5906	3.7778	8.3840	0.0447
	30%	RB	0.0275	0.0312	0.0141	0.0279	0.0475	0.0733	0.0793	0.1159	0.0886	0.0797	0.5185	0.3845	0.1732	0.2512	0.3087	1.1015	0.3620
		RMSE	0.1705	0.2531	0.0631	0.2182	0.4293	0.1368	0.2189	0.0843	0.1904	0.3669	0.4006	0.9791	0.4260	1.6082	3.9278	12.0196	0.0458
500	10%	RB	0.0127	0.0151	0.0060	0.0132	0.0213	0.0408	0.0434	0.0601	0.0398	0.0457	0.3854	0.3301	0.1020	0.1555	0.2209	0.6695	0.2141
		RMSE	0.0351	0.0596	0.0115	0.0458	0.0903	0.0413	0.0665	0.0207	0.0397	0.1122	0.1851	0.5191	0.1688	0.5451	1.5637	1.8806	0.0156
	20%	RB	0.0127	0.0155	0.0061	0.0131	0.0219	0.0424	0.0428	0.0602	0.0419	0.0472	0.3901	0.3367	0.1096	0.1580	0.2216	0.6791	0.2190
		RMSE	0.0368	0.0615	0.0125	0.0467	0.0925	0.0435	0.0684	0.0208	0.0438	0.1154	0.1983	0.5321	0.1701	0.5785	1.6624	1.9452	0.0158
	30%	RB	0.0133	0.0157	0.0064	0.0136	0.0227	0.0440	0.0419	0.0604	0.0430	0.0477	0.3946	0.3383	0.1122	0.1649	0.2301	0.6885	0.2196
		RMSE	0.0380	0.0623	0.0138	0.0483	0.0954	0.0464	0.0705	0.0219	0.0455	0.1202	0.2000	0.5453	0.1761	0.5963	1.8771	1.9686	0.0157
1000	10%	RB	0.0090	0.0103	0.0051	0.0086	0.0160	0.0330	0.0302	0.0420	0.0371	0.0340	0.4292	0.3316	0.0950	0.1406	0.2121	0.7192	0.1657
		RMSE	0.0175	0.0292	0.0093	0.0192	0.0454	0.0257	0.0334	0.0109	0.0320	0.0620	0.2060	0.4937	0.1279	0.4141	1.2218	2.1019	0.0090
	20%	RB	0.0091	0.0103	0.0054	0.0090	0.0166	0.0333	0.0302	0.0422	0.0373	0.0349	0.4294	0.3377	0.0993	0.1377	0.2182	0.7214	0.1779
		RMSE	0.0177	0.0298	0.0101	0.0199	0.0460	0.0255	0.0332	0.0109	0.0319	0.0630	0.2060	0.5063	0.1287	0.3992	1.2435	2.1086	0.0093
	30%	RB	0.0092	0.0105	0.0056	0.0095	0.0166	0.0333	0.0320	0.0426	0.0373	0.0366	0.4298	0.3386	0.1008	0.1463	0.2177	0.7269	0.1833
		RMSE	0.0179	0.0302	0.0101	0.0227	0.0478	0.0257	0.0368	0.0114	0.0319	0.0687	0.2121	0.5131	0.1291	0.4453	1.2656	2.1873	0.0098

5.2 Liver disorders dataset

In this simulation study, we used the liver disorders dataset which is available at the UCI machine learning repository (https://archive.ics.uci.edu/ml). Data include the measurements of $p = 6$ attributes with two groups and we focus a subset of $n = 200$ observations from the second selector field. Hashemi et al. [7] demonstrated that the best model is the NMVBSFA model with $q = 2$ because it has the smallest BIC. As we also want to know the performance of the SNFA, STFA, GHFA and NMVBSFA in accommodating missing values, simulated random missing values were introduced into the data. In this experimental study, a random sample was deleted from the original data to model low (r = 10%) and high (r = 30%) rate of missing, respectively. For each considered factor analysis model, we also report the AIC and BIC score as a measure of model fits for comparison and the predictive accuracy on the imputation of missing values, assessed by the mean squared prediction errors (MSPE):

$$\text{MSPE} = \frac{1}{n^*}\sum_{j=1}^{n}(\boldsymbol{y}_j^m - \hat{\boldsymbol{y}}_j^m)^\top(\boldsymbol{y}_j^m - \hat{\boldsymbol{y}}_j^m), \tag{24}$$

where $n^* = \sum_{j=1}^{n}(p - p_j^o)$ is the number of total missing values. A smaller value of MSPE indicates a more accurate prediction of missing values.

Figure 6 shows the boxplots of AIC, BIC and MSPE values as a function of q ranging from 1 to 3 for each model. We found that NMVBS with $q = 2$ is still the best model no matter which selection criterion is used.

6 Conclusion

This paper presents computationally tractable techniques for the NMVBSFA model to accommodate data with missing values, asymmetric features and heavy-tailed noises simultaneously. The proposed ECME algorithm, incorporating two auxiliary indicator matrices, presents a certain degree of convenience and flexibility in its implementation. The experimental studies have highlighted the capability of the NMVBSFA model as a promising tool for robust modeling of data in the presence of missing values. The superiority of

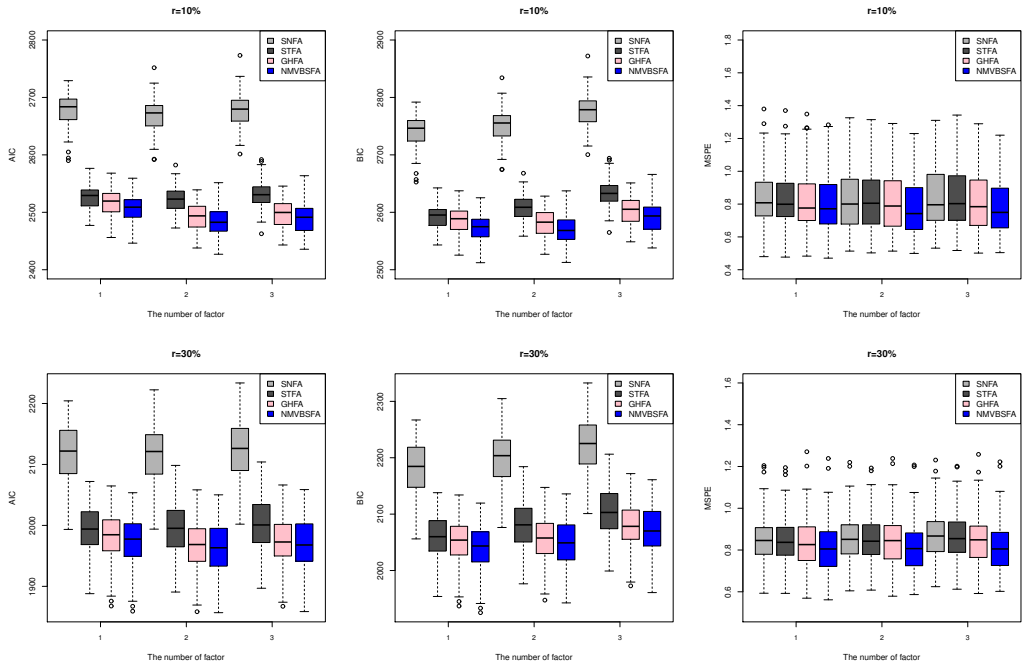

Fig. 6 The boxplots of AIC, BIC and MSPE values over 100 replicates for the AIS data with synthetic 10% and 30% missing values.

the NMVBSFA over its nested models (FA, SNFA and STFA) lies in the fact that the non-negligible effects of skewness and outliers can be simultaneously taken into account.

Concerning the future work, a natural generalization is to broaden the current method in a finite mixture representation for model-based clustering and classification based on [14]. It is of considerable interest to compare fuzzy method for clustering and classification [8] to examine the sensitivity of missing data mechanism in the NMVBSFA models.

Acknowledgment

The authors would like to sincerely thank the Editor, anonymous Associate Editor, and two reviewers for their constructive and insightful comments, which significantly improved the presentation. This work was based on research supported by the National Research Foundation, South African (SRUG190308422768 grant No. 120839 and IFR170227223754 grant No. 109214.), the South African NRF SARChI Research chair in Computational and Methodological Statistics (UID: 71199). The research of the corresponding author is supported by a grant from Ferdowsi University of Mashhad (N.2/54034). We would like to

thank Prof. McNicholas (Department of Mathematics and Statistics, McMaster University, Hamilton, ON, Canada) for his valuable comments on the earlier version of this paper.

Appendix A. Proof of Proposition 1

(a) Let $\boldsymbol{Y}_j$ $(j = 1, \ldots, n)$, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\lambda}$ be a partitioned as

$$\boldsymbol{Y}_j = \begin{bmatrix} \boldsymbol{Y}_j^o \\ \boldsymbol{Y}_j^m \end{bmatrix} = \begin{bmatrix} \boldsymbol{O}_j \boldsymbol{Y}_j \\ \boldsymbol{M}_j \boldsymbol{Y}_j \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_j^o \\ \boldsymbol{\mu}_j^m \end{bmatrix} = \begin{bmatrix} \boldsymbol{O}_j \boldsymbol{\mu} \\ \boldsymbol{M}_j \boldsymbol{\mu} \end{bmatrix}, \quad \boldsymbol{\lambda} = \begin{bmatrix} \boldsymbol{\lambda}_j^o \\ \boldsymbol{\lambda}_j^m \end{bmatrix} = \begin{bmatrix} \boldsymbol{O}_j \boldsymbol{\lambda} \\ \boldsymbol{M}_j \boldsymbol{\lambda} \end{bmatrix},$$
$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_j^{oo} & \boldsymbol{\Sigma}_j^{om} \\ \boldsymbol{\Sigma}_j^{mo} & \boldsymbol{\Sigma}_j^{mm} \end{bmatrix} = \begin{bmatrix} \boldsymbol{O}_j \boldsymbol{\Sigma} \boldsymbol{O}_j^\top & \boldsymbol{O}_j \boldsymbol{\Sigma} \boldsymbol{M}_j^\top \\ \boldsymbol{M}_j \boldsymbol{\Sigma} \boldsymbol{O}_j^\top & \boldsymbol{M}_j \boldsymbol{\Sigma} \boldsymbol{M}_j^\top \end{bmatrix}.$$

Based on [22], the marginal distribution of the observed component $\boldsymbol{Y}_j^o$ is

$$\boldsymbol{Y}_j^o \sim \text{NMVBS}_{p_j^o}(\boldsymbol{\mu}_j^o - a_\alpha \boldsymbol{\eta}_j^o, \boldsymbol{\Sigma}_j^{oo}, \boldsymbol{\eta}_j^o, \alpha).$$

(b) Using part (a) and (9), proof of part (b) is omitted.

(c) We have $\boldsymbol{Y}_j = \begin{bmatrix} \boldsymbol{Y}_j^o \\ \boldsymbol{Y}_j^m \end{bmatrix} \mid (\tilde{\boldsymbol{u}}_j, w_j) \sim N_p(\boldsymbol{\mu} + \tilde{\boldsymbol{B}}\tilde{\boldsymbol{u}}_j, w_j \boldsymbol{D})$. By Theorem 2.5.1 of [1], we can show that

$$\begin{aligned} \boldsymbol{\varphi}_j^{m.o} =& E(\boldsymbol{Y}_j^m \mid \boldsymbol{Y}_j^o, \tilde{\boldsymbol{u}}_j, w_j) = \boldsymbol{M}_j \boldsymbol{\mu} \\ &+ \boldsymbol{M}_j \tilde{\boldsymbol{B}} \tilde{\boldsymbol{u}}_j + \boldsymbol{M}_j \boldsymbol{D} \boldsymbol{O}_j^\top (\boldsymbol{O}_j \boldsymbol{D} \boldsymbol{O}_j^\top)^{-1} (\boldsymbol{O}_j \boldsymbol{Y}_j - \boldsymbol{O}_j \boldsymbol{\mu} - \boldsymbol{O}_j \tilde{\boldsymbol{B}} \tilde{\boldsymbol{u}}_j) \\ =& \boldsymbol{M}_j \left[\boldsymbol{\mu} + \tilde{\boldsymbol{B}} \tilde{\boldsymbol{u}}_j + \boldsymbol{D} \boldsymbol{O}_j^\top (\boldsymbol{O}_j \boldsymbol{D} \boldsymbol{O}_j^\top)^{-1} \boldsymbol{O}_j (\boldsymbol{Y}_j - \boldsymbol{\mu} - \tilde{\boldsymbol{B}} \tilde{\boldsymbol{u}}_j) \right], \end{aligned}$$

and

$$\begin{aligned} \boldsymbol{D}_j^{mm.o} =& cov(\boldsymbol{Y}_j^m \mid \boldsymbol{Y}_j^o, \tilde{\boldsymbol{u}}_j, w_j) = \boldsymbol{M}_j \boldsymbol{D} \boldsymbol{M}_j^\top - \boldsymbol{M}_j \boldsymbol{D} \boldsymbol{O}_j^\top (\boldsymbol{O}_j \boldsymbol{D} \boldsymbol{O}_j^\top)^{-1} \boldsymbol{O}_j \boldsymbol{D} \boldsymbol{M}_j^\top \\ =& \boldsymbol{M}_j \left(\boldsymbol{I}_p - \boldsymbol{D} \boldsymbol{O}_j^\top (\boldsymbol{O}_j \boldsymbol{D} \boldsymbol{O}_j^\top)^{-1} \boldsymbol{O}_j \right) \boldsymbol{D} \boldsymbol{M}_j^\top. \end{aligned}$$

Thus, $\boldsymbol{Y}_j^m \mid (\boldsymbol{Y}_j^o, \tilde{\boldsymbol{u}}_j, w_j) \sim N_{p-p_j^o}(\boldsymbol{\varphi}_j^{m.o}, W_j \boldsymbol{D}_j^{mm.o})$.

The proofs of part (d), (e) and part (f) are straightforward omitted based on part (a) and [7].

(g) It follows from part (d) that the distribution of $W_j \mid \boldsymbol{y}_j^o$ has a mixture of two GIG distributions. Thus, $E(W_j^r \mid \boldsymbol{y}_j^o) = \pi_j^o E(V_{1j}^r) + (1 - \pi_j^o) E(V_{2j}^r)$, where $V_{1j} \sim GIG\left(\frac{1-p_j^o}{2}, \chi_j^o, \psi_j^o\right)$ and $V_{2j} \sim GIG\left(-\frac{1+p_j^o}{2}, \chi_j^o, \psi_j^o\right)$. Therefore,

$$\begin{aligned} E(W_j^r \mid \boldsymbol{y}_j^o) =& \pi_j^o \left(\frac{\chi_j^o}{\psi_j^o} \right)^{r/2} \frac{K_{(1-p^o)/2+r}\left(\sqrt{\psi_j^o \chi_j^o}\right)}{K_{(1-p^o)/2}\left(\sqrt{\psi_j^o \chi_j^o}\right)} \\ &+ (1 - \pi_j^o) \left(\frac{\chi_j^o}{\psi_j^o} \right)^{r/2} \frac{K_{-(1+p^o)/2+r}\left(\sqrt{\psi_j^o \chi_j^o}\right)}{K_{-(1+p^o)/2}\left(\sqrt{\psi_j^o \chi_j^o}\right)}, \quad r = \pm 1. \end{aligned}$$

From part (f), we can see that $E(\tilde{\boldsymbol{U}}_j \mid \boldsymbol{y}_j^o, W_j) = \boldsymbol{R}_j^{oo}\left\{\boldsymbol{b}_j^o + \boldsymbol{\lambda}(W_j - a_\alpha)\right\}$. Applying the law of iterative expectations, we get

$$\begin{aligned} E(\tilde{\boldsymbol{U}}_j \mid \boldsymbol{y}_j^o) =& E\left\{E(\tilde{\boldsymbol{U}}_j \mid \boldsymbol{y}_j^o, W_j) \mid \boldsymbol{y}_j^o\right\} = E\left\{\boldsymbol{R}_j^{oo}\left\{\boldsymbol{b}_j^o + \boldsymbol{\lambda}(W_j - a_\alpha)\right\} \mid \boldsymbol{y}_j^o\right\} \\ =& \boldsymbol{R}_j^{oo}\left\{\boldsymbol{b}_j^o + \boldsymbol{\lambda}(E(W_j \mid \boldsymbol{y}_j^o) - a_\alpha)\right\}. \end{aligned}$$

Using the law of iterative expectations and part (f), it is easy to verify that

$$\begin{aligned} E(W_j^{-1}\tilde{\boldsymbol{U}}_j \mid \boldsymbol{y}_j^o) =& E\left\{W_j^{-1}E(\tilde{\boldsymbol{U}}_j \mid \boldsymbol{y}_j^o, W_j) \mid \boldsymbol{y}_j^o\right\} \\ =& E\left\{W_j^{-1}\left[\boldsymbol{R}_j^{oo}\left\{\boldsymbol{b}_j^o + \boldsymbol{\lambda}(W_j - a_\alpha)\right\}\right] \mid \boldsymbol{y}_j\right\} \\ =& \boldsymbol{R}_j^{oo}\left\{\boldsymbol{b}_j^o E(W_j^{-1} \mid \boldsymbol{y}_j^o) + \boldsymbol{\lambda}\left[1 - a_\alpha E(W_j^{-1} \mid \boldsymbol{y}_j^o)\right]\right\}. \end{aligned}$$

Using the law of iterative expectations, we obtain

$$\begin{aligned} E(W_j^{-1}\tilde{\boldsymbol{U}}_j\tilde{\boldsymbol{U}}_j^\top \mid \boldsymbol{y}_j^o) =& E(W_j^{-1}E(\tilde{\boldsymbol{U}}_j\tilde{\boldsymbol{U}}_j^\top \mid \boldsymbol{y}_j^o, W_j) \mid \boldsymbol{y}_j^o) \\ =& E\left\{W_j^{-1}\left[E(\tilde{\boldsymbol{U}}_j \mid \boldsymbol{y}_j^o, W_j)E(\tilde{\boldsymbol{U}}_j \mid \boldsymbol{y}_j^o, W_j)^\top + W_j\boldsymbol{R}_j^{oo}\right] \mid \boldsymbol{y}_j\right\} \\ =& E\left\{W_j^{-1}\left[E(\tilde{\boldsymbol{U}}_j \mid \boldsymbol{y}_j^o, W_j)\left(\boldsymbol{R}_j^{oo}\left\{\boldsymbol{b}_j^o + \boldsymbol{\lambda}(W_j - a_\alpha)\right\}\right)^\top + W_j R_j^{oo}\right] \mid \boldsymbol{y}_j^o\right\} \\ =& \left\{E(W_j^{-1}\tilde{\boldsymbol{U}}_j \mid \boldsymbol{y}_j^o)\boldsymbol{b}_j^{o\top} + \left[E(\tilde{\boldsymbol{U}}_j \mid \boldsymbol{y}_j^o) - a_\alpha E(W_j^{-1}\tilde{\boldsymbol{U}}_j \mid \boldsymbol{y}_j^o)\right]\boldsymbol{\lambda}^\top + \boldsymbol{I}_q\right\}\boldsymbol{R}_j^{oo}. \end{aligned}$$

References

1. Anderson, T.W.: An Introduction to Multivariate Statistical Analysis (Wiley Series in Probability and Statistics), 3 edn. (2003)
2. Barndorff-Nielsen, O., Halgreen, C.: Infinite divisibility of the hyperbolic and generalized inverse gaussian distributions. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete **38**(4), 309–311 (1977)
3. Basilevsky, A.T.: Statistical factor analysis and related methods: theory and applications, vol. 418. John Wiley & Sons (2009)
4. Desmond, A.F.: On the relationship between two fatigue-life models. IEEE Transactions on Reliability **35**(2), 167–169 (1986)
5. Fokoué, E., Titterington, D.: Mixtures of factor analysers. bayesian estimation and inference by stochastic simulation. Machine Learning **50**(1), 73–94 (2003)
6. Good, I.J.: The population frequencies of species and the estimation of population parameters. Biometrika **40**(3-4), 237–264 (1953)
7. Hashemi, F., Naderi, M., Jamalizadeh, A., Lin, T.I.: A skew factor analysis model based on the normal mean–variance mixture of Birnbaum–Saunders distribution. Journal of Applied Statistics **47**(16), 3007–3029 (2020)
8. Hashemi, F., Naderi, M., Mashinchi, M.: Clustering right-skewed data stream via Birnbaum–Saunders mixture models: A flexible approach based on fuzzy clustering algorithm. Applied Soft Computing **82**, 105539 (2019). DOI 10.1016/j.asoc.2019.105539
9. Kibler, D., Aha, D.W., Albert, M.K.: Instance-based prediction of real-valued attributes. Computational Intelligence **5**(2), 51–57 (1989)
10. Lawley, D.N.: The estimation of factor loadings by the method of maximum likelihood. Proceedings of the Royal Society of Edinburgh **60**(1), 64–82 (1940)
11. Lawley, D.N., Maxwell, A.E.: Factor analysis as a statistical method. Journal of the Royal Statistical Society: Series D (The Statistician) **12**(3), 209–229 (1962)

12. Lee, S.X., Mclachlan, G.J.: On mixtures of skew normal and skew t-distributions. Advances in Data Analysis and Classification **7**(3), 241–266 (2013)
13. Lin, T.I., Ho, H.J., Lee, C.R.: Flexible mixture modelling using the multivariate skew-t-normal distribution. Statistics and Computing **24**(4), 531–546 (2014)
14. Lin, T.I., Wang, W.L., McLachlan, G.J., Lee, S.X.: Robust mixtures of factor analysis models using the restricted multivariate skew-t distribution. Statistical Modelling **18**(1), 50–72 (2018)
15. Little, R., Rubin, D.: Statistical analysis with missing data. John Wiley (2002)
16. Liu, C., Rubin, D.B.: The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. Biometrika **81**(4), 633–648 (1994)
17. Liu, M., Lin, T.: Skew-normal factor analysis models with incomplete data. Journal of Applied Statistics **42**(4), 789–805 (2015)
18. McLachlan, G.J., Bean, R., Jones, L.B.T.: Extension of the mixture of factor analyzers model to incorporate the multivariate t-distribution. Computational Statistics & Data Analysis **51**(11), 5327–5338 (2007)
19. Meng, X.L., Rubin, D.B.: Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika **80**(2), 267–278 (1993)
20. Murray, P.M., Browne, R.P., McNicholas, P.D.: Mixtures of skew-t factor analyzers. Computational Statistics & Data Analysis **77**, 326–335 (2014a)
21. Murray, P.M., McNicholas, P.D., Browne, R.P.: A mixture of common skew-t factor analysers. Stat **3**(1), 68–82 (2014b)
22. Pourmousa, R., Jamalizadeh, A., Rezapour, M.: Multivariate normal mean–variance mixture distribution based on Birnbaum–Saunders distribution. Journal of Statistical Computation and Simulation **85**(13), 2736–2749 (2015)
23. Rubin, D.B.: Inference and missing data. Biometrika **63**(3), 581–592 (1976)
24. Rubin, D.B., Thayer, D.T.: Em algorithms for ml factor analysis. Psychometrika **47**(1), 69–76 (1982)
25. Schafer, J.L.: Analysis of incomplete multivariate data. CRC press (1997)
26. Tortora, C., McNicholas, P.D., Browne, R.P.: A mixture of generalized hyperbolic factor analyzers. Advances in Data Analysis and Classification **10**(4), 423–440 (2015). DOI 10.1007/s11634-015-0204-z
27. Villasenor Alva, J.A., Estrada, E.G.: A generalization of shapiro–wilk's test for multivariate normality. Communications in StatisticsTheory and Methods **38**(11), 1870–1883 (2009)
28. Wang, W.L., Liu, M., Lin, T.I.: Robust skew-t factor analysis models for handling missing data. Statistical Methods & Applications **26**(4), 649–672 (2017)
29. Wei, Y., Tang, Y., McNicholas, P.D.: Flexible high-dimensional unsupervised learning with missing data. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(3), 610–621 (2020)