*Statistical Modelling* 2018; **18(1)**: 50–72

# Robust mixtures of factor analysis models using the restricted multivariate skew-*t* distribution

**Tsung-I Lin[1,2], Wan-Lun Wang[3], Geoffrey J. McLachlan[4] and Sharon X. Lee[4]**
[1]Institute of Statistics, National Chung Hsing University, Taichung 402, Taiwan.
[2]Department of Public Health, China Medical University, Taichung 404, Taiwan.
[3]Department of Statistics, Feng Chia University, Taichung 407, Taiwan.
[4]Department of Mathematics, University of Queensland, St Lucia, 4072, Australia.

**Abstract:** This article introduces a robust extension of the mixture of factor analysis models based on the restricted multivariate skew-*t* distribution, called mixtures of skew-*t* factor analysis (MSTFA) model. This model can be viewed as a powerful tool for model-based clustering of high-dimensional data where observations in each cluster exhibit non-normal features such as heavy-tailed noises and extreme skewness. Missing values may be frequently present due to the incomplete collection of data. A computationally feasible EM-type algorithm is developed to carry out maximum likelihood estimation and create single imputation of possible missing values under a missing at random mechanism. The numbers of factors and mixture components are determined via penalized likelihood criteria. The utility of our proposed methodology is illustrated through analysing both simulated and real datasets. Numerical results are shown to perform favourably compared to existing approaches.

**Key words:** model-based clustering, factor analysis, heavy tails, missing values, rMST distribution, robustness

**Received** October 2016; **revised** May 2017; **accepted** June 2017

## 1 Introduction

Mixtures of factor analyzers (MFA; Ghahramani and Hinton, 1997) have been increasingly adopted to cluster high-dimensional data as it can use fewer parameters to model the component covariance matrices through their parsimonious factor-analytic representations. The MFA approach is widely applied in the fields of pattern recognition (Ueda et al., 2000), bioinformatics (McLachlan et al., 2002, 2003) and the social and psychological sciences (Wall et al., 2012), among many others. However, this model may result in unavoidable misleading inferential results since the component factors and errors are assumed to follow the multivariate normal distributions, which can be seriously affected by outliers. To enhance the robustness of MFA for data with clusters having longer than normal tails, McLachlan et al. (2007)

Address for correspondence: Tsung-I Lin, Institute of Statistics, National Chung Hsing University, Taichung 402, Taiwan.
E-mail: tilin@nchu.edu.tw

 10.1177/1471082X17718119

considered using the multivariate *t*-distribution to establish a robust extension known as mixtures of *t*-factor analyzers (MtFA).

To accommodate the observed data exhibiting asymmetric characteristics, Lin et al. (2016) extended the MFA model by incorporating the restricted multivariate skew-normal (rMSN) distribution for the component latent factors called mixtures of skew-normal factor analyzers (MSNFA). The rMSN distribution was originally introduced by Bolfarine et al. (2007) and later discussed by Lee and McLachlan (2013). The adjective term 'restricted' is referred to the restriction on the random vector of latent skewing variables such that the skewness is regulated by a single latent skewing variable. The rMSN distribution is equivalent to the classical version introduced by Azzalini and Dalla Valle (1996) after reparameterization.

Over the past decade, there has been increasing interest in adopting more flexible parametric families such as the multivariate skew-*t* distribution of Azzalini and Capitanio (2003), the multivariate skew-elliptical distribution of Sahu et al. (2003) and the multivariate skew-*t*-normal distribution of Lin et al. (2014) to accommodate substantial deviations of normality for improving the robustness against fat tails and extreme skewness. The restricted multivariate skew-*t* (rMST) distribution studied by Pyne et al. (2009) is a variant of the multivariate skew-*t* distribution of Sahu et al. (2003) and a special case of the canonical fundamental skew-*t* (CFUST) distribution introduced by Arellano-Valle and Genton (2005) and studied recently by Lee and McLachlan (2016). In the single-component version of mixtures of skew factor analyzers, models, Lin et al. (2015) proposed the skew-*t* factor analysis (STFA) model in which the latent factors and errors are assumed to jointly follow an rMST distribution to handle heavy-tailed noises and skewness in the data. Until recently, there have been other proposals of mixtures of skew factor analyzers using the generalized hyperbolic distribution (Tortora et al., 2016) and the generalized hyperbolic skew-*t* (GHST) distribution (Murray et al., 2014a), where the latter is henceforth referred to as mixtures of generalized hyperbolic skew-*t* factor analyzers (MGHSTFA). Note that the GHST distribution (Barndorff-Nielsen and Shephard, 2001) becomes a normal and not a skew-normal distribution (Lee and Poon, 2011) in the limit as the degrees of freedom (df) tend to infinity.

In this article, we propose a mixture of skew-*t* factor analysis (MSTFA) model constructed by adopting a joint rMST distribution for the component factors and errors. Missing values frequently occur in the data collection process for a variety of reasons. Under the missing at random (MAR) assumption (Rubin, 1976, 1987), an expectation conditional maximization either (ECME) algorithm (Liu and Rubin, 1994) is developed for computing the maximum likelihood (ML) estimates for the MSTFA model with incomplete data. To enhance the computational efficiency, two missingness indicator matrices are incorporated into the procedure to identify the locations of observed and missing values. Therefore, our proposed MSTFA model would be more widely applicable than the MGHSTFA adopted by Murray et al. (2014b), as it includes the MSNFA model as a limiting case and allows the handling of missing values. In order to utilize the available data information, we also offer a predictor for producing reasonably estimated values for the absent data.

The rest of the article unfolds as follows. In Section 2, we establish the notation and outline some preliminaries of the STFA model along with its rotational non-identifiability issue. In Section 3, we introduce the MSTFA model under an incomplete-data specification, present the development of the estimation algorithm and describe some practical aspects related to the initialization, convergence assessment and performance evaluation. The methodology developed in this article is illustrated in Section 4 with three real data examples. Section 5 gives some concluding remarks and recommends some directions for future research.

## 2 Notation and background

At the beginning, we establish some conventions and the notation that will be used throughout this article. Let $\phi_p(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the probability density function (pdf) of a $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ $p$-dimensional random vector and $\Phi(\cdot)$ the cumulative distribution function (cdf) of the univariate standard normal distribution. In other notation, we let $t_p(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ be the pdf of $t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$, a $p$-dimensional multivariate $t$-distribution with location $\boldsymbol{\mu}$, scale covariance matrix $\boldsymbol{\Sigma}$ and df $\nu$; $T(\cdot; \nu)$ the cdf of the Student's $t$-distribution with df $\nu$; $TN(\mu, \sigma^2; (a, b))$ the truncated normal distribution for $N(\mu, \sigma^2)$ lying within the truncated interval $(a, b)$; $\boldsymbol{M}^{1/2}$ the square root of a positive definite (symmetric) matrix $\boldsymbol{M}$; $\mathbf{1}_p$ a $p \times 1$ vector of ones; $\boldsymbol{I}_p$ the identity matrix of size $p$; **Diag**$(\cdot)$ a diagonal matrix created by extracting the main diagonal elements of a square matrix or the diagonalization; vec$(\cdot)$ an operator that vectorizes a matrix by stacking its columns vertically.

A $p$-dimensional random vector $\boldsymbol{X}$ is said to follow the rMST distribution with location $\boldsymbol{\xi} \in \mathbb{R}^p$, scale covariance matrix $\boldsymbol{\Sigma}$, shape parameter $\boldsymbol{\lambda} \in \mathbb{R}^p$ and df $\nu \in (0, \infty)$, denoted by $rSt_p(\boldsymbol{\xi}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu)$, if it has the pdf given by

$$f(\boldsymbol{x}; \boldsymbol{\xi}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu) = 2t_p(\boldsymbol{x}; \boldsymbol{\xi}, \boldsymbol{\Omega}, \nu) T\left(\frac{q}{\sigma}\sqrt{\frac{p+\nu}{U+\nu}}; p+\nu\right), \tag{2.1}$$

where $\boldsymbol{\Omega} = \boldsymbol{\Sigma} + \boldsymbol{\lambda}\boldsymbol{\lambda}^\top$, $q = \boldsymbol{\lambda}^\top \boldsymbol{\Omega}^{-1}(\boldsymbol{x} - \boldsymbol{\xi})$, $\sigma = (1 - \boldsymbol{\lambda}^\top \boldsymbol{\Omega}^{-1}\boldsymbol{\lambda})^{-1/2}$ and $U = (\boldsymbol{x} - \boldsymbol{\xi})^\top \boldsymbol{\Omega}^{-1}(\boldsymbol{x} - \boldsymbol{\xi})$ denotes the Mahalanobis squared distance between $\boldsymbol{x}$ and $\boldsymbol{\xi}$ with respect to $\boldsymbol{\Omega}$.

Let $\boldsymbol{Y}$ be a $p$-dimensional random vector. The STFA model considered by Lin et al. (2015) can be written in a matrix–vector form as

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{U} + \boldsymbol{\varepsilon} \text{ with}$$

$$\begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{\varepsilon} \end{bmatrix} \sim rSt_{q+p}\left(\begin{bmatrix} -a_\nu \boldsymbol{\Lambda}^{-1/2}\boldsymbol{\lambda} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Lambda}^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{D} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Lambda}^{-1/2}\boldsymbol{\lambda} \\ \mathbf{0} \end{bmatrix}, \nu\right), \tag{2.2}$$

where $\boldsymbol{\mu}$ is a $p$-dimensional location vector, $\boldsymbol{B}$ is a $p \times q$ matrix of factor loadings, $\boldsymbol{U}$ is the $q$-dimensional ($q < p$) vector of unobserved factors, $\boldsymbol{\varepsilon}$ is a $p$-dimensional vector of error terms and $\boldsymbol{D}$ is a diagonal matrix whose components entries are strictly positive called 'uniquenesses'. The marginal distribution $\boldsymbol{Y}$ arising from model (2.2) is

$$\boldsymbol{Y} \sim rST_p(\boldsymbol{\mu} - a_\nu \boldsymbol{\alpha}, \boldsymbol{\Delta}, \boldsymbol{\alpha}, \nu), \tag{2.3}$$

where $\boldsymbol{\alpha} = \boldsymbol{B}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{\lambda}$ is a $p$-dimensional vector of reparameterized shape parameters, and $\boldsymbol{\Delta} = \boldsymbol{B}\boldsymbol{\Lambda}^{-1}\boldsymbol{B}^\top + \boldsymbol{D}$ with $\boldsymbol{\Lambda} = \boldsymbol{I}_q + \left(1 - a_\nu^2(\nu-2)/\nu\right)\boldsymbol{\lambda}\boldsymbol{\lambda}^{\mathrm{T}}$ for satisfying the orthogonality of factor loadings. To have a zero mean vector for $\boldsymbol{U}$, we can find that

$$a_\nu = \sqrt{\frac{\nu}{\pi}}\frac{\Gamma\left((\nu-1)/2\right)}{\Gamma\left(\nu/2\right)}. \tag{2.4}$$

Consequently, the mean vector and covariance matrix of $\boldsymbol{Y}$ are

$$E(\boldsymbol{Y}) = \boldsymbol{\mu} \text{ and } \operatorname{cov}(\boldsymbol{Y}) = \frac{\nu}{\nu-2}(\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}} + \boldsymbol{D}), \tag{2.5}$$

from which the results are the same as the $t$-factor analysis model regardless of the values of $\boldsymbol{\lambda}$.

For the STFA model, there is an identifiability issue associated with the rotational invariance of the factor matrix $\boldsymbol{B}$ and skewness parameter $\boldsymbol{\lambda}$. Let $\boldsymbol{R}$ be any orthogonal matrix of order $q$. It is easy to verify that the marginal distribution in (2.3) is invariant when $\boldsymbol{B}$ and $\boldsymbol{\lambda}$ are replaced by $\boldsymbol{B}\boldsymbol{R}$ and $\boldsymbol{R}^\top\boldsymbol{\lambda}$, respectively. Moreover, such an orthogonal transformation will make the covariance structure in (2.5) unaltered. This results in $q(q-1)/2$ identifiability constraints being generally applied to the factor loading matrix. Many methods have been proposed for solving the rotational identifiability indeterminacy, see, for example, Lopes and West (2004) and Bai and Li (2012).

## 3 Methods

### 3.1 The MSTFA model with missing information

Let $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$ be $n$ independent $p$-dimensional feature vectors arising from a heterogeneous population consisting of $g$ relatively homogeneous classes. In the finite mixture framework, to designate which component density has generated the observation $\boldsymbol{Y}_j$, it is convenient to introduce an unobserved membership indicator vector $\boldsymbol{Z}_j = (Z_{1j}, \ldots, Z_{gj})$ in which one of the $Z_{ij}$'s is one and the others are zero, defined as

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

Define the mixing proportions by $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)$ such that $\Pr(Z_{ij} = 1) = \pi_i$. It follows that

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

which is the multinomial distribution with pdf

$$f(\boldsymbol{z}_j; \boldsymbol{\pi}) = \pi_1^{z_{1j}} \pi_2^{z_{2j}} \cdots (1 - \pi_1 - \cdots - \pi_{g-1})^{z_{gj}}, \quad \sum_{i=1}^{g} z_{ij} = 1.$$

The MSTFA model is constituted by a mixture of $g$ submodels of (2.2) with mixture proportions $\boldsymbol{\pi}$. Specifically, each $\boldsymbol{Y}_j$ can be formulated as

$$\boldsymbol{Y}_j = \boldsymbol{\mu}_i + \boldsymbol{B}_i \boldsymbol{U}_{ij} + \boldsymbol{\varepsilon}_{ij}, \quad \text{with probability } \pi_i \ \ (i = 1, \ldots, g) \tag{3.1}$$

along with the assumption of

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

where $a_{\nu_i}$ is $a_\nu$ in (2.4) with $\nu$ replaced with $\nu_i$ and $\boldsymbol{\Lambda}_i = \boldsymbol{I}_q + \left(1 - \frac{\nu_i - 2}{\nu_i} a_{\nu_i}^2\right) \boldsymbol{\lambda}_i \boldsymbol{\lambda}_i^\top$.

Based on model (3.1), the marginal density of $\boldsymbol{Y}_j$ takes the form

$$f(\boldsymbol{y}_j) = \sum_{i=1}^{g} \pi_i \psi(\boldsymbol{y}_j; \boldsymbol{\theta}_i), \tag{3.2}$$

where $\psi(\boldsymbol{y}_j; \boldsymbol{\theta}_i)$ is the rMST density defined in (2.1), $\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \boldsymbol{B}_i, \boldsymbol{D}_i, \boldsymbol{\lambda}_i, \nu_i)$ is composed of the unknown parameters of the $i$th mixture component and $\boldsymbol{\Theta} = (\pi_1, \ldots, \pi_{g-1}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g)$ represents the entire unknown parameters.

We consider the situation when missing values occur due to nonresponse for uncontrolled reasons. To formulate the MSTFA model with missing information, we partition $\boldsymbol{Y}_j$ $(p \times 1)$ into two subvectors such that $\boldsymbol{y}_j^{\text{o}}$ $(p_j^{\text{o}} \times 1)$ contains the observed part of $\boldsymbol{Y}_j$ and $\boldsymbol{y}_j^{\text{m}}$ $((p - p_j^{\text{o}}) \times 1)$ contains the remaining entries, namely the missing part of $\boldsymbol{Y}_j$. To facilitate further computational developments, we additionally define two permutation matrices $\boldsymbol{O}_j$ $(p_j^{\text{o}} \times p)$ and $\boldsymbol{M}_j$ $((p - p_j^{\text{o}}) \times p)$ such that $\boldsymbol{Y}_j^{\text{o}} = \boldsymbol{O}_j \boldsymbol{Y}_j$ and $\boldsymbol{Y}_j^{\text{m}} = \boldsymbol{M}_j \boldsymbol{Y}_j$. From (3.2), integrating out $\boldsymbol{y}_j^{\text{m}}$ gives

$$f(\boldsymbol{y}_j^{\text{o}}; \boldsymbol{\Theta}) = \sum_{i=1}^{g} \pi_i \psi(\boldsymbol{y}_j^{\text{o}}; \boldsymbol{\theta}_i), \tag{3.3}$$

where

$$\psi(\boldsymbol{y}_j^{\mathrm{o}};\boldsymbol{\theta}_i)=2t_{p_j^{\mathrm{o}}}(\boldsymbol{y}_j^{\mathrm{o}};\boldsymbol{\mu}_{ij}^{\mathrm{o}}-a_{\nu_i}\boldsymbol{\alpha}_{ij}^{\mathrm{o}},\boldsymbol{\Omega}_{ij}^{\mathrm{oo}},\nu_i)T\left(A_{ij}^{\mathrm{o}}\sqrt{\frac{\nu_i+p_j^{\mathrm{o}}}{\nu_i+G_{ij}^{\mathrm{o}}}};\nu_i+p_j^{\mathrm{o}}\right) \tag{3.4}$$

is the rMST density for component $i$ with

$$\begin{aligned}
&\boldsymbol{\mu}_{ij}^{\mathrm{o}}=\boldsymbol{O}_j\boldsymbol{\mu}_i,\quad \boldsymbol{\alpha}_{ij}^{\mathrm{o}}=\boldsymbol{O}_j\boldsymbol{\alpha}_i,\quad \boldsymbol{\Omega}_{ij}^{\mathrm{oo}}=\boldsymbol{O}_j\boldsymbol{\Omega}_i\boldsymbol{O}_j^{\mathrm{T}},\quad S_{ij}^{\mathrm{oo}}=\boldsymbol{O}_j^{\mathrm{T}}\boldsymbol{\Omega}_{ij}^{\mathrm{oo}^{-1}}\boldsymbol{O}_j,\\
&h_{ij}^{\mathrm{o}}=\boldsymbol{\alpha}_i^{\top}S_{ij}^{\mathrm{oo}}(\boldsymbol{y}_j-\boldsymbol{\mu}_i+a_{\nu_i}\boldsymbol{\alpha}_i),\quad \sigma_{ij}^{\mathrm{o}}=\sqrt{1-\boldsymbol{\alpha}_i^{\mathrm{T}}S_{ij}^{\mathrm{oo}}\boldsymbol{\alpha}_i},\\
&A_{ij}^{\mathrm{o}}=h_{ij}^{\mathrm{o}}/\sigma_{ij}^{\mathrm{o}},\quad G_{ij}^{\mathrm{o}}=(\boldsymbol{y}_j-\boldsymbol{\mu}_i+a_{\nu_i}\boldsymbol{\alpha}_i)^{\mathrm{T}}S_{ij}^{\mathrm{oo}}(\boldsymbol{y}_j-\boldsymbol{\mu}_i+a_{\nu_i}\boldsymbol{\alpha}_i),
\end{aligned} \tag{3.5}$$

where $\boldsymbol{\alpha}_i=\boldsymbol{B}_i\boldsymbol{\Lambda}_i^{-1/2}\boldsymbol{\lambda}_i$, $\boldsymbol{\Omega}_i=\boldsymbol{\Sigma}_i+\boldsymbol{\alpha}_i\boldsymbol{\alpha}_i^{\mathrm{T}}$ and $\boldsymbol{\Sigma}_i=\boldsymbol{B}_i\boldsymbol{\Lambda}_i^{-1}\boldsymbol{B}_i^{\mathrm{T}}+\boldsymbol{D}_i$.

Alternatively, the MSTFA model with incomplete data can be represented as

$$\begin{aligned}
\boldsymbol{Y}_j^{\mathrm{o}}\mid(Z_{ij}=1)&\sim rSt_{p_j^{\mathrm{o}}}(\boldsymbol{\mu}_{ij}^{\mathrm{o}}-a_{\nu_i}\boldsymbol{\alpha}_{ij}^{\mathrm{o}},\boldsymbol{\Sigma}_{ij}^{\mathrm{oo}},\boldsymbol{\alpha}_{ij}^{\mathrm{o}},\nu_j),\\
\boldsymbol{Z}_j&\sim\mathcal{M}(1;\pi_1,\ldots,\pi_g),
\end{aligned} \tag{3.6}$$

where

$$\boldsymbol{\Sigma}_{ij}^{\mathrm{oo}}=\boldsymbol{O}_j\boldsymbol{\Sigma}_i\boldsymbol{O}_j^{\mathrm{T}}. \tag{3.7}$$

To obtain closed-form expressions for estimators, we follow the strategy used in Lin et al. (2016) by making the invariant transformation:

$$\tilde{\boldsymbol{B}}_i\stackrel{\triangle}{=}\boldsymbol{B}_i\boldsymbol{\Lambda}_i^{-1/2}\ \text{ and }\ \tilde{\boldsymbol{U}}_{ij}\stackrel{\triangle}{=}\boldsymbol{\Lambda}_i^{1/2}\boldsymbol{U}_{ij}.$$

The following proposition presents a high-level hierarchical representation of model (3.6) that greatly facilitates parameter estimation in a complete-data framework.

*Proposition 1. The MSTFA model formulated in (3.1) admits a convenient hierarchal representation:*

$$\begin{aligned}
\boldsymbol{Y}_j^{\mathrm{o}}|(\tilde{\boldsymbol{u}}_{ij},\nu_j,w_j,Z_{ij}=1)&\sim N_{p_j^{\mathrm{o}}}(\boldsymbol{\mu}_{ij}^{\mathrm{o}}+\tilde{\boldsymbol{B}}_{ij}^{\mathrm{o}}\tilde{\boldsymbol{u}}_{ij},w_j^{-1}\boldsymbol{D}_{ij}^{\mathrm{oo}}),\\
\boldsymbol{Y}_j^{\mathrm{m}}|(\boldsymbol{Y}_j^{\mathrm{o}},\tilde{\boldsymbol{u}}_{ij},\nu_j,w_j,Z_{ij}=1)&\sim N_{p-p_j^{\mathrm{o}}}(\boldsymbol{\varphi}_{ij}^{\mathrm{m}\cdot\mathrm{o}},w_j^{-1}\boldsymbol{D}_{ij}^{\mathrm{mm}\cdot\mathrm{o}}),\\
\tilde{\boldsymbol{U}}_{ij}|(\nu_j,w_j,Z_{ij}=1)&\sim N_q\big((\nu_j-a_{\nu_i})\boldsymbol{\lambda}_i,w_j^{-1}\boldsymbol{I}_q\big),\\
V_j|(w_j,Z_{ij}=1)&\sim TN\big(0,w_j^{-1};(0,\infty)\big),\\
W_j|(Z_{ij}=1)&\sim\mathrm{Gamma}(\nu_i/2,\nu_i/2),\\
\boldsymbol{Z}_j&\sim\mathcal{M}(1;\pi_1,\ldots,\pi_g),
\end{aligned} \tag{3.8}$$

where $\tilde{\boldsymbol{B}}_{ij}^{\mathrm{o}} = \boldsymbol{O}_j\tilde{\boldsymbol{B}}_i$, $\boldsymbol{\varphi}_j^{\mathrm{m\cdot o}} = \boldsymbol{M}_j\big[\boldsymbol{\mu}_i + \tilde{\boldsymbol{B}}_i\tilde{\boldsymbol{u}}_{ij} + \boldsymbol{D}_i\boldsymbol{C}_{ij}^{\mathrm{oo}}(\boldsymbol{y}_j - \boldsymbol{\mu}_i - \tilde{\boldsymbol{B}}_i\tilde{\boldsymbol{u}}_{ij})\big]$, $\boldsymbol{D}_{ij}^{\mathrm{oo}} = \boldsymbol{O}_j\boldsymbol{D}_i\boldsymbol{O}_j^{\mathrm{T}}$, $\boldsymbol{D}_{ij}^{\mathrm{mm\cdot o}} = \boldsymbol{M}_j(\boldsymbol{I}_p - \boldsymbol{D}_i\boldsymbol{C}_{ij}^{\mathrm{oo}})\boldsymbol{D}_i\boldsymbol{M}_j^{\mathrm{T}}$ *and* $\boldsymbol{C}_{ij}^{\mathrm{oo}} = \boldsymbol{O}_j^{\mathrm{T}}(\boldsymbol{O}_j\boldsymbol{D}_i\boldsymbol{O}_j^{\mathrm{T}})^{-1}\boldsymbol{O}_j$.

By Proposition 1 in conjunction with Bayes' theorem, we have the following conditional distributions:

$$\begin{aligned}
\tilde{\boldsymbol{U}}_{ij} \mid (\boldsymbol{y}_j^{\mathrm{o}}, \nu_j, w_j, Z_{ij} = 1) &\sim N_q(\boldsymbol{q}_{ij}^{\mathrm{o}}, w_j^{-1}\boldsymbol{R}_{ij}^{\mathrm{oo}}),\\
V_j \mid (\boldsymbol{y}_j^{\mathrm{o}}, w_j, Z_{ij} = 1) &\sim TN(h_{ij}^{\mathrm{o}}, w_j^{-1}\sigma_{ij}^{\mathrm{o}2}; (0, \infty)),\\
f(W_j; \boldsymbol{y}_j^{\mathrm{o}}, Z_{ij} = 1) &= \frac{\Phi(\sqrt{w_j}A_{ij}^{\mathrm{o}})}{T(A_{ij}^{\mathrm{o}}\sqrt{\frac{\nu_i + p_j^{\mathrm{o}}}{\nu_i + G_{ij}^{\mathrm{o}}}}; \nu_i + p_j^{\mathrm{o}})} f_G\left(w_j; \frac{\nu_i + p_j^{\mathrm{o}}}{2}, \frac{\nu_i + G_{ij}^{\mathrm{o}}}{2}\right),\\
\boldsymbol{Z}_j \mid \boldsymbol{y}_j^{\mathrm{o}} &\sim \mathcal{M}(1; \tilde{\pi}_{1j}, \ldots, \tilde{\pi}_{gj}),
\end{aligned} \tag{3.9}$$

where $f_G(w; \alpha, \beta) \propto w^{\alpha-1}e^{-\beta w}$ denotes the pdf of Gamma$(\alpha, \beta)$, $\boldsymbol{q}_{ij}^{\mathrm{o}} = \boldsymbol{R}_{ij}^{\mathrm{oo}}\{\boldsymbol{b}_{ij}^{\mathrm{o}} + \boldsymbol{\lambda}(\nu_j - a_{\nu_i})\}$, $\boldsymbol{R}_{ij}^{\mathrm{oo}} = (\boldsymbol{I}_q + \tilde{\boldsymbol{B}}_i^{\mathrm{T}}\boldsymbol{C}_{ij}^{\mathrm{oo}}\tilde{\boldsymbol{B}}_i)^{-1}$, $\boldsymbol{b}_{ij}^{\mathrm{o}} = \tilde{\boldsymbol{B}}_i^{\mathrm{T}}\boldsymbol{C}_{ij}^{\mathrm{oo}}(\boldsymbol{y}_j - \boldsymbol{\mu}_i)$ and

$$\tilde{\pi}_{ij} = E(Z_{ij} \mid \boldsymbol{y}_j^{\mathrm{o}}) = \frac{\pi_i\psi(\boldsymbol{y}_j^{\mathrm{o}}; \boldsymbol{\theta}_i)}{f(\boldsymbol{y}_j^{\mathrm{o}}; \boldsymbol{\Theta})} \tag{3.10}$$

is the posterior probability that the $j$th observed feature vector $\boldsymbol{y}_j^{\mathrm{o}}$ belongs to component $i$ of the mixture.

## 3.2 Parameter estimation via the ECME algorithm

The EM algorithm (Dempster et al., 1977) is a popular iterative procedure for computing ML estimates when there are missing values in the data or latent variables in the model. Two attractive properties of EM are simplicity of implementation and monotone convergence. Despite these desirable features, however, the implementation of EM is not straightforward for the MSTFA model because the M-step is intractable. To circumvent this difficulty, we exploit the ECME algorithm (Liu and Rubin, 1994) with the complicated M-step of EM replaced by several computationally simple conditional maximization (CM) steps that maximize either the constrained $Q$ function, called the CMQ-step, as in the ECM algorithm of Meng and Rubin (1993), or the corresponding constrained actual likelihood function, called the CML-step.

To simplify notation, we treat the allocation indicator vector $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$, the latent factor vector $\tilde{\boldsymbol{U}} = (\tilde{\boldsymbol{U}}_1, \ldots, \tilde{\boldsymbol{U}}_n)$, the vector of hidden variables $\boldsymbol{V} = (V_1, \ldots, V_n)$ and the vector of scale variables $\boldsymbol{W} = (W_1, \ldots, W_n)$ together with the vector of missing values $\boldsymbol{y}^{\mathrm{m}} = (\boldsymbol{y}_1^{\mathrm{m}}, \ldots, \boldsymbol{y}_m^{\mathrm{m}})$ as the missing data. According to hierarchy (3.8), the log-likelihood function of $\boldsymbol{\Theta}$ for the complete data, comprising the observed data $\boldsymbol{y}^{\mathrm{o}} = (\boldsymbol{y}_1^{\mathrm{o}}, \ldots, \boldsymbol{y}_n^{\mathrm{o}})$ and the missing data $(\boldsymbol{Z}, \tilde{\boldsymbol{U}}, \boldsymbol{V}, \boldsymbol{W}, \boldsymbol{y}^{\mathrm{m}})$ apart from additive

constants, is

$$\begin{aligned}
&\ell_c(\boldsymbol{\Theta}; \boldsymbol{y}^{\mathrm{o}}, \boldsymbol{Z}, \tilde{\boldsymbol{U}}, \boldsymbol{V}, \boldsymbol{W}, \boldsymbol{y}^{\mathrm{m}}) \\
&= \sum_{j=1}^{n}\sum_{i=1}^{g} Z_{ij}\Big\{ \log \pi_i - \frac{1}{2}\log|\boldsymbol{D}_i| + \frac{\nu_i}{2}(\log W_j - W_j) + \frac{\nu_i}{2}\log\left(\frac{\nu_i}{2}\right) - \log\Gamma\left(\frac{\nu_i}{2}\right) \\
&\qquad - \frac{W_j}{2}\big[(V_j - a_{\nu_i})^2 \boldsymbol{\lambda}_i^{\mathrm{T}}\boldsymbol{\lambda}_i - 2(V_j - a_{\nu_i})\boldsymbol{\lambda}_i^{\mathrm{T}}\tilde{\boldsymbol{U}}_{ij}\big] - \frac{1}{2}\mathrm{tr}(\boldsymbol{D}_i^{-1}\boldsymbol{\Upsilon}_{ij})\Big\}, \qquad (3.11)
\end{aligned}$$

where

$$\boldsymbol{\Upsilon}_{ij} = W_j(\boldsymbol{y}_j - \boldsymbol{\mu}_i - \tilde{\boldsymbol{B}}_i\tilde{\boldsymbol{U}}_{ij})(\boldsymbol{y}_j - \boldsymbol{\mu}_i - \tilde{\boldsymbol{B}}_i\tilde{\boldsymbol{U}}_{ij})^{\mathrm{T}}. \qquad (3.12)$$

In what follows, we define $c_{ij}^{\mathrm{o}}(r) = [(\nu_i + p_j^{\mathrm{o}} + r)/(G_{ij}^{\mathrm{o}} + \nu_i)]^{1/2}$, where $r = -2, 0, 2$. We establish the following proposition, which is useful for evaluating the conditional expectation of (3.11), known as the $Q$-function.

*Proposition 2. Given the specifications of (3.8) and (3.9), we have the following conditional expectations (the symbol* $| \cdots$ *represents conditioning on* $\boldsymbol{y}_j^{\mathrm{o}}$ *and* $Z_{ij} = 1$*):*

$$E(W_j \mid \cdots) = \{c_{ij}^{\mathrm{o}}(0)\}^2 \frac{T(A_{ij}^{\mathrm{o}} c_{ij}^{\mathrm{o}}(2); \nu_i + p_j^{o} + 2)}{T(A_{ij}^{\mathrm{o}} c_{ij}^{\mathrm{o}}(0); \nu_i + p_j^{\mathrm{o}})}, \qquad (3.13)$$

$$\begin{aligned}
E(\log W_j \mid \cdots) &= E(W_j \mid \cdots) - \log\left(\frac{\nu_i + G_{ij}^{\mathrm{o}}}{2}\right) - \frac{\nu_i + p_j^{\mathrm{o}}}{\nu_i + G_{ij}^{\mathrm{o}}} + \mathrm{DG}\left(\frac{\nu_i + p_j^{\mathrm{o}}}{2}\right) \\
&\quad + \frac{T(A_{ij}^{\mathrm{o}} c_{ij}^{\mathrm{o}}(0); \nu_i + p_j^{\mathrm{o}})}{\int_{-\infty}^{A_{ij}^{\mathrm{o}}} t_1\left(x; 0, \frac{\nu_i + G_{ij}^{\mathrm{o}}}{\nu_i + p_j^{\mathrm{o}}}, \nu_i + p_j^{\mathrm{o}}\right) f_{\nu_i}(x)dx}, \qquad (3.14)
\end{aligned}$$

$$E(W_j V_j \mid \cdots) = h_{ij}^{\mathrm{o}} E(W_j \mid \cdots) + \sigma_{ij}^{\mathrm{o}} c_{ij}^{\mathrm{o}}(0) \frac{t(A_{ij}^{\mathrm{o}} c_{ij}^{\mathrm{o}}(0); \nu_i + p_j^{\mathrm{o}})}{T(A_{ij}^{\mathrm{o}} c_{ij}^{\mathrm{o}}(0); \nu_i + p_j^{\mathrm{o}})}, \qquad (3.15)$$

$$E(W_j V_j^2 \mid \cdots) = \sigma_{ij}^{\mathrm{o}^2} + h_{ij}^{\mathrm{o}} E(W_j V_j \mid \cdots), \qquad (3.16)$$

$$\begin{aligned}
E(W_j \tilde{\boldsymbol{U}}_{ij} \mid \cdots) &= \boldsymbol{R}_{ij}^{\mathrm{oo}}\Big\{\boldsymbol{b}_{ij}^{\mathrm{o}} E(W_j \mid \cdots) + \boldsymbol{\lambda}_i\Big[E(W_j V_j \mid \cdots) \\
&\qquad - a_{\nu_i} E(W_j \mid \cdots)\Big]\Big\}, \qquad (3.17)
\end{aligned}$$

$$\begin{aligned}
E(W_j V_j \tilde{\boldsymbol{U}}_{ij} \mid \cdots) &= \boldsymbol{R}_{ij}^{\mathrm{oo}}\Big\{\boldsymbol{b}_{ij}^{\mathrm{o}} E(W_j V_j \mid \cdots) + \boldsymbol{\lambda}_i\Big[E(W_j V_j^2 \mid \cdots) \\
&\qquad - a_{\nu_i} E(W_j V_j \mid \cdots)\Big]\Big\} \qquad (3.18)
\end{aligned}$$

*and*

$$E(W_j\tilde{\boldsymbol{U}}_{ij}\tilde{\boldsymbol{U}}_{ij}^{\top} \mid \cdots) = \Big\{\Big[E(W_jV_j\tilde{\boldsymbol{U}}_{ij} \mid \cdots) - a_{\nu_i}E(W_j\tilde{\boldsymbol{U}}_{ij} \mid \cdots)\Big]\boldsymbol{\lambda}_i^{\top} + E(W_j\tilde{\boldsymbol{U}}_{ij} \mid \cdots)\boldsymbol{b}_{ij}^{\mathrm{o}\top} + \boldsymbol{I}_q\Big\}\boldsymbol{R}_{ij}^{\mathrm{oo}}, \tag{3.19}$$

where $\mathrm{DG(x)} = d\log\Gamma(x)/dx$ is the digamma function and

$$f_{\nu_i}(x) = \mathrm{DG}\left(\frac{p_j^{\mathrm{o}}+\nu_i+1}{2}\right) - \mathrm{DG}\left(\frac{p_j^{\mathrm{o}}+\nu_i}{2}\right) - \frac{1}{\pi(\nu_i+G_{ij})} - \log\left(1+\frac{x^2}{G_{ij}+\nu_i}\right) + \frac{(\nu_i+p_j^{\mathrm{o}}+1)x^2}{(G_{ij}+\nu)(x^2+G_{ij}+\nu)}.$$

To evaluate the $Q$-function, we require the following conditional expectations:

$$\begin{aligned}
&\hat{\tau}_{ij}^{(k)} = E(Z_{ij} \mid \boldsymbol{y}_j^{\mathrm{o}}, \hat{\boldsymbol{\Theta}}^{(k)}), \quad \hat{w}_{ij}^{(k)} = E(W_j \mid \boldsymbol{y}_j^{\mathrm{o}}, Z_{ij}=1, \hat{\boldsymbol{\Theta}}^{(k)}),\\
&\hat{\kappa}_{ij}^{(k)} = E(\log W_j \mid \boldsymbol{y}_j^{\mathrm{o}}, Z_{ij}=1, \hat{\boldsymbol{\Theta}}^{(k)}), \quad \hat{s}_{1ij}^{(k)} = E(W_jV_j \mid \boldsymbol{y}_j^{\mathrm{o}}, Z_{ij}=1, \hat{\boldsymbol{\Theta}}^{(k)}),\\
&\hat{s}_{2ij}^{(k)} = E(W_jV_j^2 \mid \boldsymbol{y}_j^{\mathrm{o}}, Z_{ij}=1, \hat{\boldsymbol{\Theta}}^{(k)}), \quad \hat{\boldsymbol{\Psi}}_{ij}^{(k)} = E(W_j\tilde{\boldsymbol{U}}_{ij}\tilde{\boldsymbol{U}}_{ij}^{\top} \mid \boldsymbol{y}_j^{\mathrm{o}}, Z_{ij}=1, \hat{\boldsymbol{\Theta}}^{(k)}),\\
&\hat{\boldsymbol{\eta}}_{ij}^{(k)} = E(W_j\tilde{\boldsymbol{U}}_{ij} \mid \boldsymbol{y}_j^{\mathrm{o}}, Z_{ij}=1, \hat{\boldsymbol{\Theta}}^{(k)}), \quad \hat{\boldsymbol{\xi}}_{ij}^{(k)} = E(W_jV_j\tilde{\boldsymbol{U}}_{ij} \mid \boldsymbol{y}_j^{\mathrm{o}}, Z_{ij}=1, \hat{\boldsymbol{\Theta}}^{(k)}).
\end{aligned}$$

Let $\hat{\tau}_{ij}$ be the estimate of the posterior probability of the $j$th observation belonging to the $i$th component of the mixture model. We let $\hat{\tau}_{ij}$ be the value of $\hat{\tau}_{ij}^{(k)}$ for $\hat{\boldsymbol{\Theta}}^{(k)} = \hat{\boldsymbol{\Theta}}$.

In the E-step, we calculate

$$\begin{aligned}
&Q(\boldsymbol{\Theta}; \hat{\boldsymbol{\Theta}}^{(k)}) \qquad (3.20)\\
&= \sum_{i=1}^{g}\sum_{j=1}^{n}\hat{\tau}_{ij}^{(k)}\Big\{\log\pi_i - \frac{1}{2}\log|\boldsymbol{D}_i| + \frac{\nu_i}{2}(\hat{\kappa}_{ij}^{(k)} - \hat{w}_{ij}^{(k)}) + \frac{\nu_i}{2}\log\left(\frac{\nu_i}{2}\right) - \log\Gamma\left(\frac{\nu_i}{2}\right)\\
&\quad -\frac{1}{2}\big[(\hat{s}_{2ij}^{(k)} - 2a_{\nu_i}\hat{s}_{1ij}^{(k)} + a_{\nu_i}^2\hat{w}_{ij}^{(k)})\boldsymbol{\lambda}_i\boldsymbol{\lambda}_i^{\top} - 2\boldsymbol{\lambda}_i^{\top}(\hat{\boldsymbol{\xi}}_{ij}^{(k)} - a_{\nu_i}\hat{\boldsymbol{\eta}}_{ij}^{(k)}) - \mathrm{tr}(\boldsymbol{D}_i^{-1}\boldsymbol{\Upsilon}_{ij}^{(k)})\big]\Big\},
\end{aligned}$$

with $\boldsymbol{\Upsilon}_{ij}^{(k)} = E(\boldsymbol{\Upsilon}_{ij} \mid \boldsymbol{y}_j^{\mathrm{o}}, Z_{ij}=1, \hat{\boldsymbol{\Theta}}^{(k)})$, where $\boldsymbol{\Upsilon}_{ij}$ is given by (3.12).

For updating $\hat{\boldsymbol{\Theta}}^{(k)}$, the CM-steps obtained by maximizing (3.20) proceed as follows:

CM-step 1: Calculate $\hat{\pi}^{(k+1)} = \hat{n}_i^{(k)}/n$, where $\hat{n}_i^{(k)} = \sum_{j=1}^{n} \hat{\tau}_{ij}^{(k)}$.

CM-step 2: Calculate

$$\hat{\boldsymbol{\mu}}_i^{(k+1)} = \frac{\sum_{j=1}^{n} \hat{\tau}_{ij}^{(k)} \hat{w}_{ij}^{(k)} \hat{\boldsymbol{q}}_{ij}^{(k+1/2)} - \hat{\boldsymbol{D}}_i^{(k)} \sum_{j=1}^{n} \hat{\tau}_{ij}^{(k)} \hat{\boldsymbol{C}}_{ij}^{\mathrm{oo}^{(k)}} \hat{\tilde{\boldsymbol{B}}}_i^{(k)} \hat{\boldsymbol{\eta}}_{ij}^{(k)}}{\sum_{j=1}^{n} \hat{\tau}_{ij}^{(k)} \hat{w}_{ij}^{(k)}},$$

where $\hat{\boldsymbol{q}}_{ij}^{(k+1/2)} = \hat{\boldsymbol{\mu}}_i^{(k+1)} + \hat{\boldsymbol{D}}_i^{(k)} \hat{\boldsymbol{C}}_{ij}^{\mathrm{oo}^{(k)}} (\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_i^{(k+1)})$.

CM-step 3: Calculate

$$\hat{\tilde{\boldsymbol{B}}}_i^{(k+1)} = \Big( \sum_{j=1}^{n} \hat{\tau}_{ij}^{(k)} \big[ \hat{\boldsymbol{E}}_{ij}^{\mathrm{oo}(k)} \hat{\boldsymbol{\Psi}}_{ij}^{(k)} + (\hat{\boldsymbol{q}}_{ij}^{(k+1/2)} - \hat{\boldsymbol{\mu}}_i^{(k+1)}) \hat{\boldsymbol{\eta}}_{ij}^{(k)^{\mathrm{T}}} \big] \Big) \Big( \sum_{j=1}^{n} \hat{\tau}_{ij}^{(k)} \hat{\boldsymbol{\Psi}}_{ij}^{(k)} \Big)^{-1},$$

where $\hat{\boldsymbol{E}}_{ij}^{\mathrm{oo}(k)} = (\boldsymbol{I}_p - \hat{\boldsymbol{D}}_i^{(k)} \hat{\boldsymbol{C}}_{ij}^{\mathrm{oo}^{(k)}}) \hat{\tilde{\boldsymbol{B}}}_i^{(k)}$.

CM-step 4: Calculate

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

where

$$\begin{aligned} \hat{\boldsymbol{\Upsilon}}_j^{(k+1/2)} = {} & \hat{w}_{ij}^{(k)} (\hat{\boldsymbol{q}}_{ij}^{(k+1/2)} - \hat{\boldsymbol{\mu}}_i^{(k+1)}) (\hat{\boldsymbol{q}}_{ij}^{(k+1/2)} - \hat{\boldsymbol{\mu}}_i^{(k+1)})^{\mathrm{T}} + (\boldsymbol{I}_p - \hat{\boldsymbol{D}}_i^{(k+1)} \hat{\boldsymbol{C}}_{ij}^{\mathrm{oo}(k)}) \hat{\boldsymbol{D}}_i^{(k+1)} \\ & + \big( \hat{\boldsymbol{E}}_{ij}^{\mathrm{oo}(k)} - \hat{\tilde{\boldsymbol{B}}}_i^{(k+1)} \big) \hat{\boldsymbol{\Psi}}_{ij}^{(k)} \big( \hat{\boldsymbol{E}}_{ij}^{\mathrm{oo(k)}} - \hat{\tilde{\boldsymbol{B}}}_i^{(k+1)} \big)^{\mathrm{T}} \\ & + \big( \hat{\boldsymbol{q}}_{ij}^{(k)} - \hat{\boldsymbol{\mu}}_i^{(k+1)} \big) \hat{\boldsymbol{\eta}}_{ij}^{(k)\mathrm{T}} \big( \hat{\boldsymbol{E}}_j^{\mathrm{oo(k)}} - \hat{\tilde{\boldsymbol{B}}}_i^{(k+1)} \big)^{\mathrm{T}} \\ & + \big( \hat{\boldsymbol{E}}_{ij}^{\mathrm{oo}(k)} - \hat{\tilde{\boldsymbol{B}}}_i^{(k+1)} \big) \hat{\boldsymbol{\eta}}_{ij}^{(k)} \big( \hat{\boldsymbol{q}}_{ij}^{\mathrm{oo}(k)} - \hat{\boldsymbol{\mu}}_i^{(k+1)} \big)^{\mathrm{T}}. \end{aligned}$$

CM-step 5: Calculate

$$\hat{\boldsymbol{\lambda}}_i^{(k+1)} = \frac{\sum_{j=1}^{n} \hat{\tau}_{ij}^{(k)} (\hat{\boldsymbol{\zeta}}_{ij}^{(k)} - a_{\hat{\nu}_i^{(k)}} \hat{\boldsymbol{\eta}}_{ij}^{(k)})}{\sum_{j=1}^{n} \hat{\tau}_{ij}^{(k)} (\hat{s}_{2ij}^{(k)} - 2a_{\hat{\nu}_i^{(k)}} \hat{s}_{1ij}^{(k)} + a_{\hat{\nu}_i^{(k)}}^2 \hat{w}_{ij}^{(k)})}.$$

CML-step 6: In light of (3.6), when the dfs are assumed to be unequal, the updated estimate of $\nu_i$ is obtained as

$$\hat{\nu}_i^{(k+1)} = \underset{\nu_i}{\operatorname{argmax}} \sum_{j=1}^{n} \hat{\tau}_{ij}^{(k)} \psi(\boldsymbol{y}_j^{\mathrm{o}}; \hat{\boldsymbol{\mu}}_{ij}^{\mathrm{o}(k+1)}, \hat{\boldsymbol{\Sigma}}_{ij}^{\mathrm{oo}(k+1)}, \hat{\boldsymbol{\alpha}}_{ij}^{\mathrm{o}(k+1)}, \nu_i), \quad i = 1, \ldots, g,$$

where $\psi(\boldsymbol{y}_j^{\mathrm{o}}; \cdots)$ is the rMST pdf as defined in (3.4). In the case where the dfs are constrained to be identical, say $\nu_1 = \cdots = \nu_g = \nu$, we update $\nu$ by maximizing the constrained actual observed log-likelihood function, namely

$$\hat{\nu}^{(k+1)} = \underset{\nu}{\operatorname{argmax}} \sum_{j=1}^{n} \log \left\{ \sum_{i=1}^{g} \hat{\pi}_i^{(k+1)} \psi(\boldsymbol{y}_j^{\mathrm{o}}; \hat{\boldsymbol{\mu}}_{ij}^{\mathrm{o}(k+1)}, \hat{\boldsymbol{\Sigma}}_{ij}^{\mathrm{oo}(k+1)}, \hat{\boldsymbol{\alpha}}_{ij}^{\mathrm{o}(k+1)}, \nu) \right\},$$

where $\hat{\boldsymbol{\mu}}_{ij}^{\mathrm{o}(k+1)}$, $\hat{\boldsymbol{\alpha}}_{ij}^{\mathrm{o}(k+1)}$ and $\hat{\boldsymbol{\Sigma}}_{ij}^{\mathrm{oo}(k+1)}$ are $\boldsymbol{\mu}_{ij}^{\mathrm{o}}$, $\boldsymbol{\alpha}_{ij}^{\mathrm{o}}$ and $\boldsymbol{\Sigma}_{ij}^{\mathrm{oo}}$ in (3.5) and (3.7), respectively, evaluated at the current estimates at the start of the $(k+1)$th iteration.

In the aforementioned ECME procedure, the E-step and CM/CML-steps are repeated alternately until a reliable convergence criterion is satisfied, as detailed in the next section. Having reached the convergence, we denote the ML estimate by $\hat{\boldsymbol{\Theta}} = \{\hat{\pi}_i, \hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{B}}_i, \hat{\boldsymbol{D}}_i, \hat{\boldsymbol{\lambda}}_i, \hat{\nu}_i\}_{i=1}^{g}$, where $\hat{\boldsymbol{B}}_i = \hat{\tilde{\boldsymbol{B}}}_i \hat{\boldsymbol{\Lambda}}_i^{1/2}$ and $\hat{\boldsymbol{\Lambda}}_i = \boldsymbol{I}_q + \left(1 - \frac{\hat{\nu}_i - 2}{\hat{\nu}_i} a_{\hat{\nu}_i}^2\right) \hat{\boldsymbol{\lambda}}_i \hat{\boldsymbol{\lambda}}_i^{\mathrm{T}}$. Consequently, component memberships can be determined based on the maximum a posteriori (MAP) classification rule (McLachlan and Peel, 2000). More precisely, each partially observed vector $\boldsymbol{y}_j^{\mathrm{o}}$ is assigned to the component to which it has the highest estimated posterior probability $\hat{\tau}_{ij}$ of belonging.

## 3.3 Prediction of factor scores and missing values

In addition to estimating the model parameters, factor scores can be predicted and used in subsequent analyses. For instance, researchers may want to investigate how factor scores differ between groups, or to use the factor information for data reconstruction in lower-dimensional subspaces (Lin et al., 2014). From (3.9), using the law of iterated expectations, it can be verified that

$$\hat{\boldsymbol{u}}_{ij} = E(\boldsymbol{U}_{ij} \mid \boldsymbol{y}_j^{\mathrm{o}}, \hat{\boldsymbol{\Theta}}) = \hat{\boldsymbol{\Lambda}}_i^{-1/2} \hat{\boldsymbol{R}}_{ij}^{\mathrm{oo}} \left\{ \hat{\boldsymbol{b}}_{ij}^{\mathrm{o}} + \hat{\boldsymbol{\lambda}}_i \left[ E(V_j \mid \boldsymbol{y}_j^{\mathrm{o}}, \hat{\boldsymbol{\Theta}}) - a_{\hat{\nu}_i} \right] \right\}, \tag{3.21}$$

where

$$\begin{aligned} &E(V_j \mid \boldsymbol{y}_j^{\mathrm{o}}, \hat{\boldsymbol{\Theta}}) \\ &= \hat{h}_{ij}^{\mathrm{o}} + \hat{\sigma}_{ij}^{\mathrm{o}} \frac{(\hat{\nu}_i + \hat{G}_{ij}^{\mathrm{o}}) \Gamma((\hat{\nu}_i + p_j^{\mathrm{o}} - 2)/2) \hat{c}_{ij}^{\mathrm{o}}(-2)}{2\Gamma((\hat{\nu}_i + p_j^{\mathrm{o}})/2)} \frac{t(\hat{A}_{ij}^{\mathrm{o}} \hat{c}_{ij}^{\mathrm{o}}(-2); \hat{\nu}_i + p_j^{\mathrm{o}} - 2)}{T(\hat{A}_{ij}^{\mathrm{o}} \hat{c}_{ij}^{\mathrm{o}}(0); \hat{\nu}_i + p_j^{\mathrm{o}})}. \end{aligned}$$

Therefore, the estimated factor scores corresponding to $\boldsymbol{y}_j^{\mathrm{o}}$ can be calculated as

$$\hat{\boldsymbol{u}}_j = \sum_{j=1}^{n} \hat{\tau}_{ij} \hat{\boldsymbol{u}}_{ij}. \tag{3.22}$$

Furthermore, filling in the missing data with plausible values is an important task for creating a completed dataset so that standard statistical methods can be applied. The ML 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, a minimum mean squared *conditional predictor* for $\boldsymbol{Y}_j^{\mathrm{m}}$ is given by

$$\hat{\boldsymbol{y}}_j^{\mathrm{m}} = E(\boldsymbol{Y}_j^{\mathrm{m}} \mid \boldsymbol{y}_j^{\mathrm{o}}, \hat{\boldsymbol{\Theta}}) = \boldsymbol{M}_j \sum_{j=1}^{n} \hat{\tau}_{ij} (\hat{\boldsymbol{\mu}}_i + \hat{\boldsymbol{B}}_i \hat{\boldsymbol{u}}_{ij} + \hat{\boldsymbol{D}}_i \hat{\boldsymbol{C}}_{ij}^{\mathrm{oo}} (\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_i - \hat{\boldsymbol{B}}_i \hat{\boldsymbol{u}}_{ij})). \tag{3.23}$$

Although the predictor (3.23) does not reflect the uncertainty about the predictions of the unknown missing values, the problem can be overcome by a more broadly applicable *multiple imputation* (MI) method (Schafer, 1997; Little and Rubin, 2002). Notably, MI assumes that the frequentist criterion (or the ML approach) has been satisfied and then proceeds drawing single imputation separately several times for replicated samples obtained from the MCMC or bootstrap techniques.

## 3.4 Practical implementation issues

Because the log-likelihood function of a finite mixture model is usually complicated and might have multiple modes, the EM-based algorithm does not guarantee convergence to the global maximizer in the case where the likelihood function is bounded or to the appropriate local maximizer if the likelihood function is unbounded. One simple way to alleviate this problem is to try many different initial values and choose the one providing the highest likelihood. This can be easily performed by extracting initial parameter values according to different initial partitions obtained by multiple times of *random starts* (randomly assigning each point to a cluster) or $k$-means clustering (Hartigan and Wong, 1979). As such, the initial $\pi_i$'s are given by the sample proportions of cluster labels, and the initial $\boldsymbol{\mu}_i$'s are taken to be the sample mean vectors corresponding to each cluster. Then, we fit the ordinary factor analysis model to each clustered sample and set the resulting estimates as the initial values of $\boldsymbol{B}_i$ and $\boldsymbol{D}_i$. The initial skewness parameters $\boldsymbol{\lambda}_i$'s are initialized based on the sample skewness of the estimated factor scores. Finally, we take a small fixed value for the initial df, say $\hat{\nu}^{(0)} = 5$.

In model-based clustering via the MFA approach and its variants, the number of components $g$ and the number of factors $q$ are typically unknown. The Bayesian information criterion (BIC; Schwarz, 1978) is commonly used as an aid to choose $g$

and $q$ consistently, especially for $g$ (Keribin, 2000). The form of BIC is

$$\text{BIC} = m \log n - 2\ell_{\max}(\hat{\boldsymbol{\Theta}}),$$

where $\ell_{\max}(\hat{\boldsymbol{\Theta}})$ is the maximized log-likelihood value and $m$ is the number of free parameters to be estimated. Biernacki et al. (2000) pointed out that BIC may not be a good way of assessing the number of components in a model-based clustering perspective. Instead, they proposed an integrated completed likelihood (ICL) criterion whose purpose aims at selecting a suitable $g$ that leads to a sensible partitioning of the data. The ICL is defined as

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

where $\text{ENT}(\hat{(\tau)}) = -\sum_{i=1}^{g}\sum_{j=1}^{n} \hat{\tau}_{ij} \log \hat{\tau}_{ij}$ is the entropy used to measure the overlap of clusters. Models with smaller values of BIC or ICL imply a more adequate fit. In our study, the model with the lowest BIC or ICL value is considered to be the 'best'. If there is an inconsistency selected by two criteria, we will favour the model picked by ICL, as it places a higher penalty on more complex models.

To precisely monitor the convergence of the ECME algorithm for the MSTFA model, we adopt an Aitken acceleration method (Aitken, 1926) to measure the absolute difference between the log-likelihood value and its asymptotic estimate, that is,

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

where $\ell_{\infty}^{(k+1)}$ defined in McLachlan and Krishnan (2008) is the limiting value of the log-likelihood at the $(k+1)$th iteration. Unless otherwise stated, we terminate the iterations if the pre-specified tolerance $\epsilon = 10^{-5}$ is satisfied or when the maximum number of iterations $K = 10\,000$ is reached.

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 proposed by Hubert and Arabie (1985) as an improvement of Rand index (RI) (Rand, 1997) is commonly used to validate the agreement between two different partitions of the same data. Notice that ARI value generally ranges between 0 and 1, but it can be negative corresponding to very poor agreement. The CCR is calculated for each permutation of the cluster labels and the reported rate is the largest value across all permutations.

## 4 Applications to real data

### 4.1 The chronic kidney disease data

As an illustration, we consider the chronic kidney disease (CKD) data collected from a hospital over a period of nearly two months. This dataset, available from the UCI

learning repository (Lichman, 2013), contains $p = 11$ numeric attributes measured on $n = 400$ patients and one diagnose variable for binary classification, of which 250 instances have the diagnosis of CKD and 150 instances have the diagnose of no CKD. An overview of these 11 attributes, denoted by $X_1, \ldots, X_{11}$, is summarized in Table 1. It is shown that the sample means and standard deviations for the two populations are quite different from each other. In addition, most attributes exhibit mild to strong asymmetry and light to extremely heavy tails, indicating that the two subpopulations deviate substantially from normality. The overall data contain 4 400 measurements in which there are 636 absent cases, leading to a missing rate of 14.45%. There are 185 patients who have at least one missing value, representing a missing rate of 46.25% as considered by the listwise procedure that discards all instances containing missing values.

**Table 1** An overview of chronic kidney disease database

| | | CKD (no CKD) | | | | |
|---|---|---|---|---|---|---|
| Attributes | Description | Number of missing values | Sample mean | Sample sd | Sample skewness | Sample kurtosis |
| $X_1$ | age (years) | 8(1) | 54.5(46.5) | 17.4(15.6) | −1.2(0.1) | 4.2(2.3) |
| $X_2$ | blood pressure (mm/Hg) | 10(2) | 79.6(71.4) | 15.2(8.5) | 1.5(–0.3) | 10.4(1.4) |
| $X_3$ | blood glucose(mgs/dl) | 38(6) | 175.4(107.7) | 92.1(18.6) | 1.3(–0.1) | 4.4(2.0) |
| $X_4$ | blood urea (mgs/dl) | 13(6) | 72.4(32.8) | 58.6(11.5) | 2.0(–0.1) | 8.3(1.7) |
| $X_5$ | serum creatinine (mgs/dl) | 12(5) | 4.4(0.9) | 7.0(0.3) | 6.2(–0.1) | 56.0(1.6) |
| $X_6$ | sodium (mEq/L) | 82(5) | 133.9(141.7) | 12.4(4.8) | −7.0(0.2) | 72.1(2.0) |
| $X_7$ | potassium (mEq/L) | 83(5) | 4.9(4.3) | 4.3(0.6) | 8.6(–0.3) | 78.3(1.5) |
| $X_8$ | haemoglobin (gms) | 46(6) | 10.6(15.2) | 2.2(1.3) | −0.3(0.2) | 3.4(2.1) |
| $X_9$ | packed cell volume | 67(4) | 32.9(46.3) | 7.2(4.1) | −0.2(0.2) | 3.6(1.8) |
| $X_{10}$ | white blood cell count($\times 10^2$) | 99(7) | 90.7(77.1) | 35.8(18.4) | 1.3(0.1) | 6.7(2.0) |
| $X_{11}$ | red blood cell count | 124(7) | 3.9(5.4) | 0.9(0.6) | 0.8(0.2) | 5.9(1.9) |

We implement the MFA, MSNFA and MSTFA approaches with $q$ ranging from 1 to $q_{\max} = 6$ to fitting this dataset. The number of components $g$ is set to 2 corresponding to two manifest group memberships. Table 2 lists the ML results, including the maximized log-likelihood values and the number of parameters together with the BIC and ICL values.

For the purpose of comparing the performance of clustering, the classification agreement measured by ARI and CCR are also reported in the last two columns of Table 2. Observing the table, the MSTFA with $q = 2$ provides the best fitting performance (BIC = 6 351.90 and ICL = 6 386.66) and the best accurate classification (ARI = 0.66; CCR = 0.91) for this dataset. This example demonstrates that the MSTFA provides improvements over the other two classical methods because it provides more flexibility in capturing non-normal features. Figure 1 graphs the factor-2 scores estimated by using equation (3.22) from the three mixture factor models. Comparing the three histograms in Figure 1, we find that the estimated factor scores obtained by the MSTFA model appear to be more suitable as its skewness and kurtosis are larger compared to that of the other two models. A second finding is that the factor score estimates are not robust to the distributional specification of the

**Table 2** Performance of three mixtures of factor models fitted to the CKD data

| Model | $q$ | $\ell_{\max}$ | m | BIC | ICL | ARI | CCR |
|---|---|---|---|---|---|---|---|
| MFA | 1 | −3 415.55 | 67 | 7 232.52 | 7 270.31 | 0.25 | 0.75 |
| | 2 | −3 332.29 | 87 | 7 185.83 | 7 232.78 | 0.13 | 0.69 |
| | 3 | −3 312.31 | 105 | 7 253.72 | 7 291.04 | 0.23 | 0.74 |
| | 4 | −3 301.23 | 121 | 7 327.42 | 7 364.99 | 0.23 | 0.74 |
| | 5 | −3 281.69 | 135 | 7 372.22 | 7 413.13 | 0.18 | 0.72 |
| | 6 | −3 278.19 | 147 | 7 437.13 | 7 477.73 | 0.18 | 0.72 |
| MSNFA | 1 | −3 380.76 | 69 | 7 174.94 | 7 214.04 | 0.22 | 0.74 |
| | 2 | −3 211.31 | 91 | 6 967.84 | 7 016.39 | 0.15 | 0.70 |
| | 3 | −3 193.03 | 111 | 7 051.10 | 7 099.95 | 0.16 | 0.71 |
| | 4 | −3 179.99 | 129 | 7 132.87 | 7 179.25 | 0.16 | 0.71 |
| | 5 | −3 168.48 | 145 | 7 205.73 | 7 250.12 | 0.17 | 0.71 |
| | 6 | −3 155.46 | 159 | 7 263.56 | 7 299.62 | 0.20 | 0.70 |
| MSTFA | 1 | −3 025.50 | 71 | 6 476.40 | 6 526.33 | 0.40 | 0.82 |
| | 2 | −2 897.35 | 93 | **6 351.90** | **6 386.66** | **0.66** | **0.91** |
| | 3 | −2 872.12 | 113 | 6 421.28 | 6 460.80 | 0.58 | 0.88 |
| | 4 | −2 847.50 | 131 | 6 479.89 | 6 513.43 | 0.66 | 0.90 |
| | 5 | −2 864.06 | 147 | 6 608.86 | 6 648.27 | 0.56 | 0.88 |
| | 6 | −2 841.56 | 161 | 6 647.74 | 6 694.60 | 0.50 | 0.85 |

**Note:** The smallest BIC and ICL scores and the largest ARI and CCR values are indicated in bold.

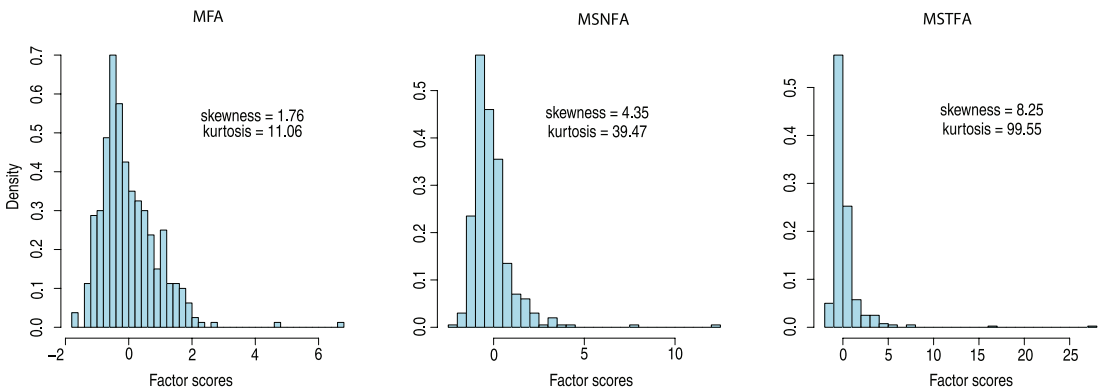


**Figure 1** Histograms of the estimated 2nd factor scores obtained from three mixtures of factor analysis models

latent factors. In this example, the MFA model tends to shrink the large values down although their factor scores do not appear to be normal.

We are also interested in investigating the imputation of missing values from the three mixture factor models. In handling missing data, it is generally believed that an appropriate specification of latent factors can produce more accurate imputations. To examine how consistent the missing values imputed from the MFA model are versus the MSNFA and MSTFA models, Figure 2 depicts the pairwise scatter plots of the 636 missing values predicted by using equation (3.23). We find that the MFA and

MSNFA models offer very similar imputed values, while the MSTFA model provides rather different imputations. Thus, the MSTFA model would seem to be potentially a more realistic model for imputation than the MFA and MSNFA models.

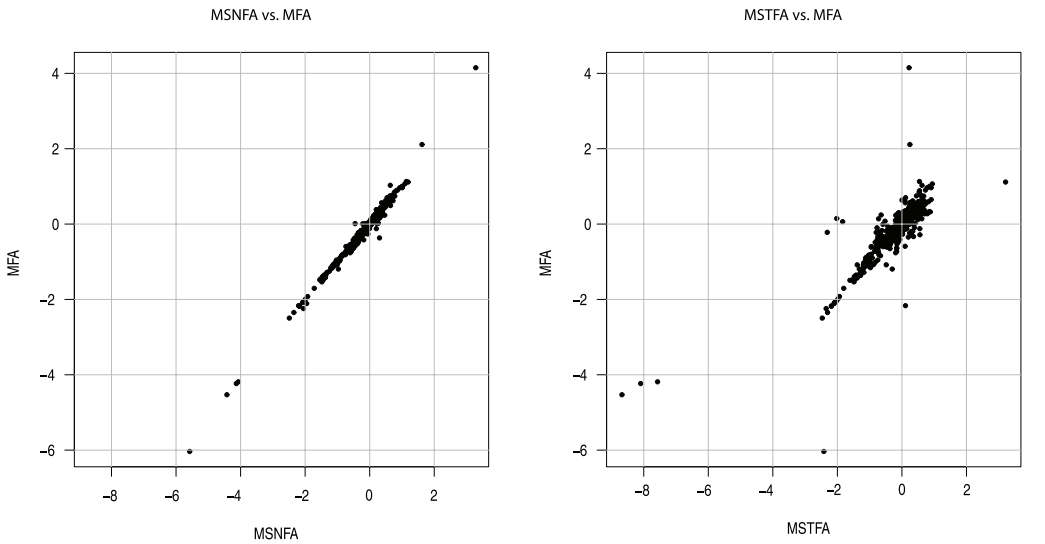


**Figure 2** Scatter plots of imputed missing values via three mixtures of factor analysis models for the CKD data

## 4.2 The Wisconsin prognostic breast cancer data

The second data to which we applied our proposed method is the Wisconsin Prognostic Breast Cancer (WPBC) data (Street et al., 1995), publicly available at the UCI Machine Learning data repository. This dataset consists of 198 patients, each with 32 continuous attributes related to the characteristics of breast cell nuclei presenting in the digitized image of a fine needle aspirate (FNA) test. The first 30 attributes are the means, the standard errors and the means of the 3 worst values of 10 cellular features computed for each FNA digitized image. The last 2 attributes are traditional prognostic variables: the diameter of the excised tumor (tumor size in centimeters) and the number of positive axillary lymph nodes observed at time of surgery (lymph node status). There are 4 missing cases in lymph node status. In this study, patients are partitioned into 2 groups according to a binary diagnostic decision: 47 recurrent patients if the disease is observed within two years versus 151 non-recurrent patients if the disease is never observed beyond a threshold point.

Figure 3 shows the boxplots of normalized measurements for the 32 attributes in the recurrent and non-recurrent patients. It can be observed that the distribution of most attributes is highly skewed or has a rather long tail. This motivates us to advocate the MSTFA model to analyse this dataset knowing that the assumption of normality is violated and outliers are present.

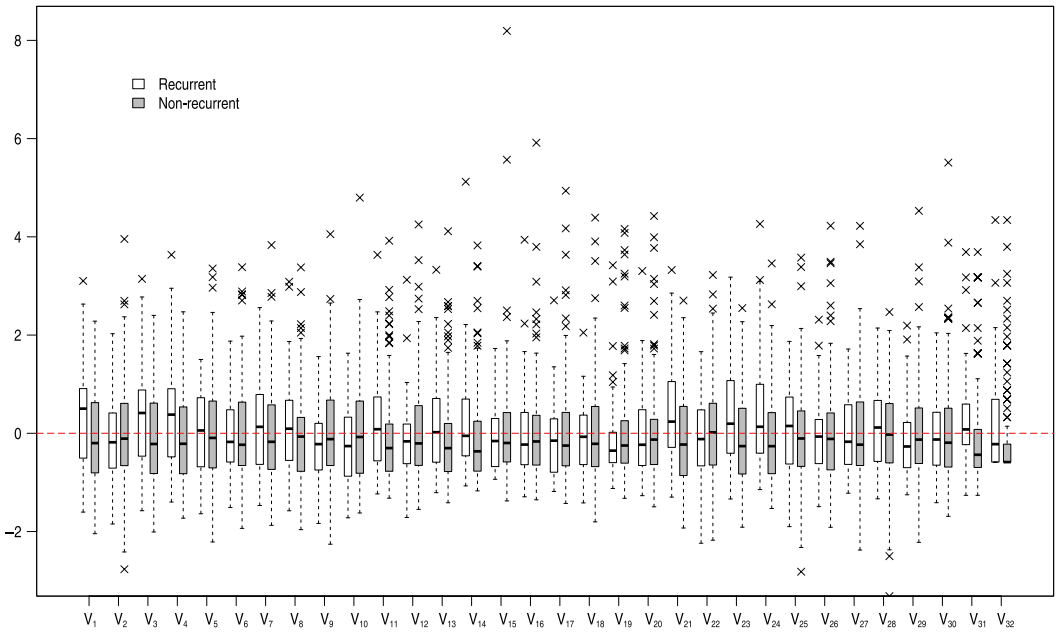


**Figure 3** Boxplots for the 32 attributes of the standardized WPBC data

We consider the fitting of MFA, MSNFA, and MSTFA models to the data where both $g$ and $q$ are taken to be unknown a priori. These models are applied to the WPBC data with $g = 2 - 4$ and $q = 1 - 15$. Comparisons are only made on BIC and CCR since ICL and ARI correspond to the same preference for this example. For clarity, we report in Table 3, the values of BIC and CCR for $q = 6 - 11$. On the basis of the results reported in this table, the best model selected by BIC is MSTFA($g = 2; q = 10$), which also provides the best classification performance (CCR = 0.73).

## 4.3 Simulation based on AIS data with synthetic missing values

Our third example concerns the Australian Institute of Sport (AIS) data (Cook and Weisberg, 1994), containing $p = 11$ physical and hematological attributes measured on 100 female and 102 male athletes, comprising a total of $n = 202$ athletes. A detailed account of these attributes along with their sample skewness and kurtosis (a split by gender) are separately summarized in Table 4. As indicated in the table, there are some attributes which obviously deviate from normality due to strong skewness and high kurtosis.

We conduct a small simulation study to compare the performance of the MFA, MSNFA and MSTFA approaches to accommodating missing values based on the AIS data. To avoid some attributes having a greater impact due to their different scales, we normalize the data such that each variable has zero mean and unit standard deviation. Afterwards, missing values are artificially generated by deleting at random from the

**Table 3** Comparison of three mixtures of factor models fitted to the WPBC data

| Model | Number of factors ($q$) | $g = 2$ | | $g = 3$ | | $g = 4$ | |
|---|---|---|---|---|---|---|---|
| | | BIC | CCR | BIC | CCR | BIC | CCR |
| MFA | 6 | 8 707.31 | 0.62 | 9 322.82 | 0.38 | 9 649.82 | 0.45 |
| | 7 | 8 350.96 | 0.59 | 8 699.37 | 0.39 | 9 614.51 | 0.40 |
| | 8 | 8 130.29 | 0.58 | 8 627.44 | 0.40 | 9 683.10 | 0.39 |
| | 9 | 7 936.38 | 0.60 | 8 649.06 | 0.42 | 9 339.07 | 0.35 |
| | 10 | 7 735.07 | 0.60 | 8 487.85 | 0.44 | 9 547.20 | 0.39 |
| | 11 | 8 395.56 | 0.51 | 8 836.04 | 0.39 | 9 476.50 | 0.39 |
| MSNFA | 6 | 8 591.18 | 0.61 | 9 411.91 | 0.38 | 9 606.95 | 0.45 |
| | 7 | 8 759.43 | 0.57 | 9 071.26 | 0.41 | 9 410.31 | 0.38 |
| | 8 | 8 098.02 | 0.63 | 8 650.47 | 0.38 | 9 820.53 | 0.37 |
| | 9 | 8 009.57 | 0.60 | 9 101.89 | 0.38 | 9 594.44 | 0.42 |
| | 10 | 7 944.39 | 0.61 | 8 961.08 | 0.40 | 9 655.86 | 0.40 |
| | 11 | 7 785.33 | 0.60 | 8 842.95 | 0.41 | 10 113.23 | 0.38 |
| MSTFA | 6 | 7 846.94 | 0.63 | 8 864.27 | 0.40 | 9 410.11 | 0.47 |
| | 7 | 7 603.39 | 0.62 | 8 357.64 | 0.45 | 8 950.99 | 0.44 |
| | 8 | 7 626.57 | 0.73 | 8 211.62 | 0.40 | 9 330.81 | 0.47 |
| | 9 | 7 628.35 | 0.70 | 8 311.09 | 0.44 | 9 597.39 | 0.51 |
| | 10 | **7 465.87** | **0.73** | 8 210.83 | 0.44 | 9 682.83 | 0.52 |
| | 11 | 7 493.80 | 0.70 | 8 221.63 | 0.42 | 9 579.99 | 0.52 |

**Note:** The smallest BIC score and the largest CCR value are indicated in bold.

**Table 4** An overview of 11 attributes of the AIS data

| Variable | Description | Female | | Male | |
|---|---|---|---|---|---|
| | | Skewness | Kurtosis | Skewness | Kurtosis |
| rcc | red cell count | 0.69 | 3.30 | 0.92 | 7.73 |
| wcc | white cell count | 0.75 | 4.00 | 0.86 | 4.58 |
| Hc | Hematocrit | 0.26 | 2.34 | 1.49 | 10.37 |
| Hg | Haemoglobin | 0.09 | 2.18 | 0.97 | 5.31 |
| Fe | plasma ferritin concentration | 1.35 | 5.57 | 0.88 | 3.13 |
| bmi | body mass index | 0.69 | 4.18 | 1.41 | 5.99 |
| ssf | sum of skin folds | 0.78 | 3.64 | 1.39 | 4.79 |
| Bfat | body fat percentage | 0.35 | 2.91 | 1.53 | 5.08 |
| lbm | lean body mass | −0.31 | 3.45 | 0.27 | 3.62 |
| Ht | height (cm) | −0.56 | 4.20 | 0.07 | 3.00 |
| Wt | weight (kg) | −0.17 | 3.13 | 0.39 | 3.41 |

standardized AIS data under low ($r = 10\%$), moderate ($r = 20\%$) and relatively high ($r = 30\%$) rates of missingness. To study the robustness against the presence of underlying noise, we added 10 noisy points drawn from a Student's $t$-distribution with 2 df to the simulated incomplete data. A total of 100 replications are carried out across each combination of $q$=1–3 and $r = 10\%, 20\%$ and $30\%$. Each simulated dataset was fitted under the three considered models. Comparisons were made on the adequacy of overall fitness in terms of BIC and ICL, classification accuracy in terms of ARI and CCR. The precision for reconstructing missing values is assessed by the

**Table 5** Simulation results based on 100 replications

| Missing rate | Criterion | | $q = 1$ | | | $q = 2$ | | | $q = 3$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | MFA | MSNFA | MSTFA | MFA | MSNFA | MSTFA | MFA | MSNFA | MSTFA |
| | | Mean | 5 064.08 | 4 890.41 | 4 198.70 | 4 990.16 | 4 871.31 | 4 224.78 | 4 152.70 | 4 004.29 | 3 786.67 |
| | BIC | Std | 227.76 | 312.25.67 | 110.99 | 321.62 | 320.17 | 108.46 | 447.80 | 352.39 | 59.26 |
| | | Freq | 0 | 0 | 100 | 3 | 2 | 95 | 0 | 12 | 88 |
| | | Mean | 5 791.11 | 5 650.86 | 4 975.86 | 5 717.19 | 5 631.76 | 5 001.94 | 5 030.14 | 4 391.87 | 4 730.97 |
| | ICL | Std | 227.76 | 312.25.67 | 110.99 | 321.62 | 124.68 | 108.46 | 477.80 | 352.39 | 59.26 |
| | | Freq | 0 | 0 | 100 | 3 | 2 | 95 | 20 | 5 | 75 |
| | | Mean | 0.66 | 0.60 | 0.71 | 0.53 | 0.57 | 0.72 | 0.17 | 0.12 | 0.70 |
| $r = 10\%$ | ARI | Std | 0.32 | 0.31 | 0.13 | 0.38 | 0.34 | 0.12 | 0.33 | 0.29 | 0.14 |
| | | Freq | 17 | 33 | 50 | 20 | 32 | 48 | 14 | 3 | 83 |
| | | Mean | 0.87 | 0.85 | 0.92 | 0.81 | 0.83 | 0.92 | 0.60 | 0.58 | 0.92 |
| | CCR | Std | 0.15 | 0.17 | 0.04 | 0.20 | 0.18 | 0.04 | 0.18 | 0.16 | 0.04 |
| | | Freq | 17 | 33 | 50 | 20 | 32 | 48 | 14 | 4 | 82 |
| | | Mean | 0.39 | 0.38 | 0.35 | 0.40 | 0.39 | 0.35 | 0.32 | 0.30 | 0.26 |
| | MSD | Std | 0.08 | 0.08 | 0.07 | 0.09 | 0.09 | 0.06 | 0.14 | 0.07 | 0.05 |
| | | Freq | 13 | 23 | 64 | 13 | 20 | 67 | 7 | 15 | 78 |
| | | Mean | 4 697.55 | 4 624.32 | 4 022.36 | 4 756.27 | 4 632.91 | 4 015.62 | 4 088.38 | 3 925.67 | 37 16.52 |
| | BIC | Std | 269.96 | 241.51 | 98.43 | 240.08 | 286.00 | 106.57 | 436.43 | 349.67 | 63.16 |
| | | Freq | 3 | 1 | 96 | 0 | 1 | 99 | 0 | 22 | 78 |
| | | Mean | 5 425.57 | 5 384.77 | 4 799.53 | 5 483.29 | 5 393.36 | 4 792.78 | 4 965.83 | 4 853.25 | 4 660.82 |
| | ICL | Std | 269.96 | 244.51 | 98.43 | 240.08 | 286.00 | 106.57 | 436.43 | 349.67 | 63.16 |
| | | Freq | 3 | 1 | 96 | 0 | 2 | 98 | 14 | 34 | 52 |
| | | Mean | 0.48 | 0.56 | 0.73 | 0.53 | 0.55 | 0.71 | 0.22 | 0.15 | 0.69 |
| $r = 20\%$ | ARI | Std | 0.37 | 0.32 | 0.12 | 0.35 | 0.33 | 0.14 | 0.36 | 0.31 | 0.04 |
| | | Freq | 26 | 36 | 38 | 15 | 31 | 54 | 12 | 9 | 79 |
| | | Mean | 0.79 | 0.83 | 0.92 | 0.82 | 0.83 | 0.92 | 0.63 | 0.59 | 0.91 |
| | CCR | Std | 0.19 | 0.17 | 0.04 | 0.18 | 0.17 | 0.04 | 0.19 | 10.17 | 0.04 |
| | | Freq | 26 | 36 | 26 | 16 | 33 | 51 | 12 | 9 | 79 |
| | | Mean | 0.43 | 0.42 | 0.37 | 0.43 | 0.42 | 0.37 | 0.35 | 0.34 | 0.28 |
| | MSD | Std | 0.08 | 0.07 | 0.05 | 0.08 | 0.08 | 0.04 | 0.09 | 0.09 | 0.04 |
| | | Freq | 6 | 16 | 78 | 7 | 12 | 81 | 5 | 11 | 84 |
| | | Mean | 4 585.64 | 4 523.86 | 4 087.88 | 4 327.34 | 4 288.73 | 3 789.43 | 3 925.89 | 3 785.00 | 3 607.10 |
| | BIC | Std | 167.40 | 156.55 | 61.49 | 223.32 | 206.04 | 95.21 | 343.84 | 297.91 | 61.06 |
| | | Freq | 0 | 0 | 100 | 1 | 2 | 97 | 2 | 36 | 62 |
| | | Mean | 5 145.53 | 5 100.46 | 4 681.20 | 5 054.36 | 5 049.17 | 4 566.59 | 4 803.34 | 4 712.58 | 4 551.39 |
| | ICL | Std | 167.40 | 156.55 | 61.49 | 223.32 | 206.04 | 95.21 | 343.84 | 297.91 | 61.06 |
| | | Freq | 0 | 0 | 100 | 5 | 1 | 94 | 11 | 38 | 44 |
| | | Mean | 0.73 | 0.73 | 0.82 | 0.49 | 0.55 | 0.74 | 0.23 | 0.16 | 0.71 |
| $r = 30\%$ | ARI | Std | 0.20 | 0.20 | 0.06 | 0.38 | 0.31 | 0.12 | 0.36 | 0.32 | 0.17 |
| | | Freq | 31 | 31 | 38 | 17 | 29 | 54 | 11 | 8 | 81 |
| | | Mean | 0.92 | 0.92 | 0.95 | 0.79 | 0.83 | 0.93 | 0.64 | 0.60 | 0.92 |
| | CCR | Std | 0.07 | 0.07 | 0.02 | 0.20 | 0.17 | 0.04 | 0.05 | 0.18 | 0.20 |
| | | Freq | 34 | 29 | 37 | 16 | 32 | 52 | 14 | 8 | 78 |
| | | Mean | 0.51 | 0.51 | 0.48 | 0.46 | 0.45 | 0.39 | 0.39 | 0.38 | 0.32 |
| | MSD | Std | 0.05 | 0.04 | 0.04 | 0.10 | 0.08 | 0.06 | 0.08 | 0.07 | 0.04 |
| | | Freq | 8 | 15 | 77 | 3 | 7 | 90 | 1 | 9 | 90 |

mean squared deviation (MSD) between the true values $\boldsymbol{y}_j^{\mathrm{m}}$ and the imputed values $\hat{\boldsymbol{y}}_j^{\mathrm{m}}$. The quantity of MSD can be calculated as

$$\mathrm{MSD} = \frac{1}{n^{\mathrm{m}}} \sum_{j=1}^{n} (\boldsymbol{y}_j^{\mathrm{m}} - \hat{\boldsymbol{y}}_j^{\mathrm{m}})^{\top} (\boldsymbol{y}_j^{\mathrm{m}} - \hat{\boldsymbol{y}}_j^{\mathrm{m}}),$$

where $n^{\mathrm{m}} = \sum_{j=1}^{n} (p - p_j^{\mathrm{o}})$ is the number of total missing values. In general, a smaller value of MSD indicates a more accurate imputation result.

Table 5 reports the averages of criteria values (Mean) together with their standard deviations (Std) under every scenario considered. As a guide for determining the most adequate model, the frequencies (Freq) preferred by each criterion are also recorded. In all cases, the MSTFA model provides a better fit, higher classification accuracy and lower MSD than the other two competitors. The MFA as well as MSNFA models are relatively rarely chosen due possibly to a lack of sufficient robustness against serious violations of normality assumption. The experimental study highlights the ability of MSTFA model in providing more precise inference surrounded by the presence of synthetic missing values and background noise.

## 5 Conclusion

This article presents a robust extension of MFA models based on the rMST distribution, called the MSTFA model, as a new tool for flexibly dealing with the data exhibiting patterns of asymmetry, multimodality, heavy-tailed noise and possibly missing values. An effective ECME algorithm is developed for parameter estimation with explicit expressions for all E-step conditional expectations and CM estimators except for the dfs. The prediction of latent factors and imputation of missing values are obtained as a by-product once the ML estimates are obtained. The experimental studies highlight the tractability and superiority of the MSTFA model over the existing MFA and MSNFA approaches. The major strength of MSTFA lies in the fact that the non-negligible effects of asymmetry and outliers involving in each component can be simultaneously taken into full account and thereby achieves improvements.

There are some possible variants and modifications of the current method that deserve further exploration. A natural generalization is to consider a broader family of CFUST distributions as a unified modelling framework with more options. On the other hand, our estimating algorithm can lead to a degenerate covariance matrix when the dimension of variables becomes extremely large. To circumvent this ill-posed problem, one possible extension is to pursue a parsimonious modelling of MSTFA with common factor loadings (Baek et al., 2010; Baek and McLachlan, 2011; Wang, 2013; Murray et al., 2014b; Wang, 2015; Wang and Lin, 2016) and will be the subject of future work. Despite the ease in implementation of the proposed ECME procedure, it still requires an exhaustive search to determine $(g, q)$ from a vast amount of candidates. Another option is to develop an automated learning algorithm for

the joint determination of $(g, q)$ through variational Bayes approximation methods (Ghahramani and Beal, 2000; Wei and Li, 2013; Subedi and McNicholas, 2014).

## Acknowledgements

The authors are grateful to the editor and the anonymous reviewers for their insightful comments and suggestions. This work has been supported by the Ministry of Science and Technology of Taiwan, grant numbers MOST 105-2118-M-005-003-MY2 and MOST 105-2118-M-035-004-MY2.

## References

Aitken AC (1926) On Bernoulli's numerical solution of algebraic equations. *Proceedings of the Royal Society of Edinburgh*, **46**, 289–305.

Arellano-Valle R and Genton M (2005) On fundamental skew distributions. *Journal of Multivariate Analysis*, **96**, 93–116.

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*, **65**, 367–89.

Azzalini A and Dalla Valle A (1996) The multivariate skew-normal distribution. *Biometrika*, **83**, 715–26.

Baek J and McLachlan G (2011) Mixtures of common *t*-factor analyzers for clustering high-dimensional microarray data. *Bioinformatics*, **27**, 1269–76.

Baek J, McLachlan G and Flack L (2010) Mixtures of factor analyzers with common factor loadings: Applications to the clustering and visualization of high-dimensional data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **32**, 1–13.

Bai J and Li K (2012) Statistical analysis of factor models of high dimension. *Annals of Statistics*, **40**, 436–65.

Barndor-Nielsen O and Shephard N (2001) Non-Gaussian Ornstein–Uhlenbeck-based models and some of their uses in financial economics. *Journal of the Royal Statistical Society, Series B*, **63**, 167–241.

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*, **22**, 719–25.

Bolfarine H, Montenegro LC and Lachos VH (2007) Influence diagnostics for skew-normal linear mixed models. *Sankhyā*, **69**, 648–70.

Cook RD and Weisberg S (1994) *An Introduction to Regression Graphics*. New York, NY: Wiley.

Dempster AP, Laird NM and Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, **9**, 1–38.

Ghahramani Z and Beal M (2000) Variational inference for Bayesian mixture of factor analysers. In *S Solla, Leen T and Muller KR, eds. Advances in Neural Information Processing Systems, Vol. 12, pages 449–55. Cambridge, MA: MIT Press.*

Ghahramani Z and Hinton GE (1997) *The EM algorithm for factor analyzers* (Technical report no. CRG-TR-96-1). Toronto, CA: The University of Toronto.

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

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

Keribin C (2000) Consistent estimation of the order of mixture models. *Sankhyā*, **62**, 49–66.

Lee S and McLachlan G (2013) On mixtures of skew normal and skew *t*-distributions. *Advances in Data Analysis and Classification*, **7**, 241–66.

Lee SX and McLachlan GJ (2016) Finite mixtures of canonical fundamental skew *t*-distributions. *Statistics and Computing*, **26**, 573–89.

Lee YW and Poon SH (2011) *Systemic and systematic factors for loan portfolio loss distribution*, pages 1–61. Econometrics and applied economics workshops, School of Social Science, University of Manchester, Manchester, UK.

Lichman M (2013) *UCI Machine Learning Repository* Irvine, CA: University of California, School of Information and Computer Science. URL http://archive.ics.uci.edu/ml

Lin TI, Ho HJ and Lee CR (2014) Flexible mixture modelling using the multivariate skew-t-normal distribution. *Statistics and Computing*, **24**, 531–46.

Lin TI, McLachlan GJ and Lee SX (2016) Extending mixtures of factor models using the restricted multivariate skew-normal distribution. *Journal of Multivariate Analysis*, **143**, 398–413.

Lin TI, Wu PH, McLachlan GJ and Lee SX (2015) A robust factor analysis model using the restricted skew-*t* distribution. *Test*, **24**, 510–31.

Little RJA and Rubin DB (2002) *Statistical Analysis with Missing Data, 2nd edition*. New York, NY: Wiley.

Liu C and Rubin DB (1994) The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika*, **81**, 633–48.

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

McLachlan GJ, Bean RW and Jones BT (2007) Extension of the mixture of factor analyzers model to incorporate the multivariate t-distribution. *Computational Statistics and Data Analysis*, **51**, 5327–38.

McLachlan GJ, Bean RW and Peel D (2002) A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, **18**, 413–22.

McLachlan GJ and Krishnan T (2008) *The EM algorithm and extensions, 2nd edition*. New York, NY: John Wiley and Sons.

McLachlan GJ and Peel D (2000) *Finite Mixture Models*. New York, NY: Wiley

McLachlan GJ, Peel D and Bean RW (2003) Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics and Data Analysis*, **41**, 379–88.

Meng XL and Rubin DB (1993) Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, **80**, 267–78.

Murray PM, Browne RP and McNicholas PD (2014a) Mixtures of skew-*t* factor analyzers. *Computational Statistics and Data Analysis*, **77**, 326–35.

Murray PM, McNicholas PD and Browne RP (2014b) Mixtures of common skew-*t* factor analyzers. *STAT*, **3**, 68–82.

Pyne S, Hu X, Wang K, Rossin E, Lin TI, Maier LM, Baecher-Allan C, McLachlan GJ, Tamayo P, Haer DA, De Jager PL and Mesirov JP (2009) Automated high-dimensional flow cytometric data analysis. *Proceedings of the National Academy of Sciences, USA*, **106**, pages 8519–24.

Rand WM (1997) Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66**, 846–50.

Rubin DB (1976) Inference and missing data. *Biometrika*, **63**, 581–92.

——— (1987) *Multiple Imputation for Nonresponse in Surveys*. New York, NY: Wiley.

Sahu SK, Dey DK and Branco MD (2003) A new class of multivariate skew distributions with application to Bayesian regression models. *Canadian Journal of Statistics*, **31**, 129–50.

Schafer JL (1997) *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Schwarz G (1978) Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–64.

Street WN, Mangasarian OL and Wolberg WH (1995) An inductive learning approach to prognostic prediction. In Prieditis A and Russell S, eds. *Proceedings of the Twelfth International Conference on Machine Learning*, pages 522–30. San Francisco, CA: Morgan Kaufmann.

Subedi S and McNicholas PD (2014) Variational Bayes approximations for clustering via mixtures of normal inverse Gaussian distributions. *Advances in Data Analysis and Classification*, **8**, 167–93.

Tortora C, McNicholas P and Browne R (2016) A mixture of generalized hyperbolic factor analyzers. *Advances in Data Analysis and Classification*, **10**, 423–40.

Ueda N, Nakano R, Ghahramani Z and Hinton GE (2000) Smem algorithm for mixture models. *Neural Computation*, **12**, 2109–28.

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

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

——— (2015) Mixtures of common *t*-factor analyzers for modeling high-dimensional data with missing values. *Computational Statistics and Data Analysis*, **83**, 223–35.

Wang W and Lin T (2016) Flexible clustering via extended mixtures of common *t*-factor analyzers. *AStA Advances in Statistical Analysis*. doi: 10.1007/s10182-016-0281-0.

Wei X and Li C (2013) Bayesian mixtures of common factor analyzers: Model, variational inference, and applications. *Signal Processing*, **93**, 2894–2905.