


# Mixtures of restricted skew- $t$ factor analyzers with common factor loadings

Wan-Lun Wang<sup>1</sup> · Luis M. Castro<sup>2</sup> ·  
Yen-Ting Chang<sup>3</sup> · Tsung-I Lin<sup>3,4</sup> 

Received: 14 February 2017 / Revised: 3 February 2018 / Accepted: 27 February 2018 /  
Published online: 8 March 2018  
© Springer-Verlag GmbH Germany, part of Springer Nature 2018

**Abstract** Mixtures of common  $t$  factor analyzers (MCtFA) have been shown its effectiveness in robustifying mixtures of common factor analyzers (MCFA) when handling model-based clustering of the high-dimensional data with heavy tails. However, the MCtFA model may still suffer from a lack of robustness against observations whose distributions are highly asymmetric. This paper presents a further robust extension of the MCFA and MCtFA models, called the mixture of common restricted skew- $t$  factor analyzers (MCrStFA), by assuming a restricted multivariate skew- $t$  distribution for the common factors. The MCrStFA model can be used to accommodate severely non-normal (skewed and leptokurtic) random phenomena while preserving its parsimony in factor-analytic representation and performing graphical visualization in low-dimensional plots. A computationally feasible expectation conditional maximization either algorithm is developed to carry out maximum likelihood estimation. The numbers of factors and mixture components are simultaneously determined based on common likelihood penalized criteria. The usefulness of our proposed model is illustrated with simulated and real datasets, and experimental results signify its superiority over some existing competitors.

---

✉ Tsung-I Lin  
tilin@nchu.edu.tw

- <sup>1</sup> Department of Statistics, Graduate Institute of Statistics and Actuarial Science, Feng Chia University, Taichung, Taiwan
- <sup>2</sup> Department of Statistics, Pontificia Universidad Católica de Chile, Casilla 306, Correo 22, Santiago, Chile
- <sup>3</sup> Institute of Statistics, National Chung Hsing University, Taichung, Taiwan
- <sup>4</sup> Department of Public Health, China Medical University, Taichung, Taiwan

**Keywords** Clustering · Common factor loadings · Data reduction · ECME algorithm · Factor analyzer · Outliers

**Mathematics Subject Classification** 62H25 · 62H30

## 1 Introduction

Mixtures of factor analyzers (MFA), originally introduced by Ghahramani and Hinton (1997), provide a global non-linear approach to dimension reduction via the adoption of component distributions having a factor-analytic representation for the component-covariance matrices. To substantially reduce the number of parameters in component matrices, especially when the number of components ( $g$ ) or features ( $p$ ) becomes large, Baek et al. (2010) extended the MFA by using common component factor loadings, known as mixtures of common factor analyzers (MCFA), which have now been a popular tool for high-dimensional data analysis. To deal with data with extreme values or outliers commonly observed in microarray experiments, Baek and McLachlan (2011) presented a robust version of MCFA using multivariate Student's- $t$  distributed component errors and factors, called mixtures of common  $t$ -factor analyzers (MCtFA). Recently, Wang (2013, 2015) extended the MCFA and MCtFA approaches to accommodating high-dimensional data with possibly missing values.

The specification of component factors and errors on both MFA and MCFA rests on the assumption of multivariate normality for computational convenience and mathematical tractability, but the two models are highly vulnerable to outliers. Although the use of MCtFA model is less affected by the violation of normality, it may still suffer from the lack of robustness against highly asymmetric observations. In many practical problems, however, the data to be analyzed may contain a group or groups of observations whose distributions are moderately or severely skewed and/or of having heavy tails. As shown in many empirical studies, a slight deviation from normality may seriously affect the estimates of mixture parameters and subsequently lead to spurious groups as well as misleading statistical inference.

Over the past few decades, there has been growing interest in adopting more flexible parametric distributions to accommodate non-normal features such as asymmetry and longer-than-normal tails leading to non-zero skewness and excess kurtosis, see the monograph by Azzalini (2014) for a more comprehensive overview. Lin et al. (2015) proposed a robust extension of factor analysis models based on the restricted multivariate skew- $t$  (rMST) distribution (Pyne et al. 2009). Other related proposals include mixtures of skew-normal/ $t$  factor analyzers (Lin et al. 2016, 2018), mixtures of generalized hyperbolic (GH) factor analyzers (Tortora et al. 2016), mixtures of skew- $t$  factor analyzers (Murray et al. 2014a), and mixtures of common skew- $t$  factor analyzers (Murray et al. 2014b). Besides, Murray et al. (2017a) presented an extended version of MFA with the component factors and errors following the skew- $t$  distribution considered by Sahu et al. (2003), which is referred to as the unrestricted multivariate skew- $t$  (uMST) distribution by Lee and McLachlan (2014).

Note that the rMST and uMST distributions are not nested within each other, and they are equivalent only in the univariate case. Moreover, Sahu et al. (2003) have high-

lighted that the calculation of the uMST density becomes cumbersome as  $p$  increases. The computational difficulty of the uMST formulation was also pointed out by Murray et al. (2017a; Section 5). Azzalini et al. (2016) have provided a detailed comparison between the rMST and uMST distributions in terms of the merits of both distributions for data modeling. When comparing the two distributions in the context of model-based clustering, their illustrative examples indicate that “neither formulation is markedly superior and, if these results were to be taken in favor of either formulation, it would be the classical formulation”, namely the rMST distribution adopted in this paper.

Further, it is interesting to note that the skew- $t$  distribution adopted by Murray et al. (2014a, b), arising from the family of GH distributions (Bardorff-Nielsen and Shephard 2001), is referred to as the generalized hyperbolic skew- $t$  (GHST) distribution henceforth. Its density form is rather different from the rMST distribution and does not include the skew-normal as a limiting case (Lee and Poon 2011). The model proposed by Murray et al. (2014b) is henceforth referred to as mixtures of common generalized hyperbolic skew- $t$  factor analyzers (MCghstFA).

In this paper, we propose an alternative skew extension of the MCtFA model based on the rMST distribution, called the mixture of common restricted skew- $t$  factor analyzers (MCrstFA) model. This new proposal preserves resistance to extremely non-normal effects commonly happen in high-dimensional data. Similar to MCFA and MCtFA models, common factor loadings are utilized for parsimoniously modeling the component-covariance matrices. To portray the observed data into a lower dimensional space and avoid possible singularities, the scale-covariance matrices for component errors ( $\mathbf{D}_i$ ) are generally assumed to be homogeneous ( $\mathbf{D}_i = \mathbf{D}$ ). Under certain circumstances,  $\mathbf{D}_i$  can be relaxed to be unequal or modified to different types such as (isotropic with unequal variances) or (isotropic with equal variance). Lately, Wang and Lin (2017) presented a modification of MCtFA using component-specific  $\mathbf{D}_i$  and empirically demonstrated its advantage in classifying new subjects whose true group labels are unknown in advance.

The rest of the paper is structured as follows. In Sect. 2, we establish the notation and outline some preliminary properties of the rMST distribution. In Sect. 3, we present the specification of MCrstFA model and develop a workable expectation conditional maximization either (ECME) algorithm for carrying maximum likelihood (ML) estimation. In Sect. 4, the initialization along with the stopping rules, the criteria for model selection and clustering performance, and the identifiability issues are discussed. In Sect. 5, we conduct two simulation studies to examine the validity of MCrstFA model. The methodology is illustrated on a real example concerning human liver cancer data in Sect. 6. Concluding remarks and directions for future works are given in Sect. 7. Some detailed proofs and supplementary information are deferred to appendices.

## 2 Notation and prerequisites

We first review the rMST distribution and study its related properties. Let  $\phi_p(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$  be the probability density function (pdf) of a multivariate normal distribution with mean vector  $\boldsymbol{\mu}$  and covariance matrix  $\boldsymbol{\Sigma}$ , denoted by  $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ ;  $\Phi(\cdot)$  the cumulative distribution function (cdf) of the standard normal distribution;  $TN(v, \sigma^2; (a, b))$  the

truncated normal distribution defined as a normal distribution  $N(\mu, \sigma^2)$  lying within an interval  $(a, b)$ ;  $t_p(\cdot; \mu, \Sigma, \nu)$  the pdf of a  $p$ -variate  $t$  distribution with location  $\mu$ , scale-covariance matrix  $\Sigma$  and the degree of freedom (DOF)  $\nu$ ;  $g(x; \alpha, \beta)$  the pdf of gamma distribution given by  $\beta^\alpha x^{\alpha-1} \exp\{-\beta x\}/\Gamma(\alpha)$ ;  $T(\cdot; \nu)$  the cdf of the Student's  $t$  distribution with zero mean, unit scale variance and DOF  $\nu$ ;  $\mathbf{1}_p$  a  $p \times 1$  vector with all elements being 1;  $\mathbf{I}_p$  a  $p \times p$  identity matrix;  $\text{Diag}\{\cdot\}$  a diagonal matrix made by extracting the main diagonal elements of a square matrix or the diagonalization of a vector;  $A^{1/2}$  the square root of a symmetric matrix  $A$ .

Following Pyne et al. (2009), a  $p$ -dimensional random vector  $\mathbf{Y}$  is said to follow the rMST distribution with location vector  $\mu \in \mathbb{R}^p$ , scale-covariance matrix  $\Sigma$ , skewness vector  $\lambda \in \mathbb{R}^p$  and DOF  $\nu \in \mathbb{R}^+$ , denoted as  $\mathbf{Y} \sim rST_p(\mu, \Sigma, \lambda, \nu)$ , if it has the pdf:

$$\psi_p(\mathbf{y}; \mu, \Sigma, \lambda, \nu) = 2t_p(\mathbf{y}; \mu, \Omega, \nu)T\left(M\sqrt{\frac{\nu+p}{\nu+\delta}}; \nu+p\right), \tag{1}$$

where  $\Omega = \Sigma + \lambda\lambda^\top$ ,  $\delta = (\mathbf{y} - \mu)^\top \Omega^{-1}(\mathbf{y} - \mu)$  and  $M = \lambda^\top \Omega^{-1}(\mathbf{y} - \mu)/(1 - \lambda^\top \Omega^{-1}\lambda)^{1/2}$ . Note that the distribution of  $\mathbf{Y}$  is reduced to  $t_p(\mu, \Sigma, \nu)$  by setting  $\lambda = \mathbf{0}$  and to  $rSN_p(\mu, \Sigma, \lambda)$  as  $\nu \rightarrow \infty$ . Furthermore, the family of (1) also includes  $N_p(\mu, \Sigma)$ , obtained by letting  $\lambda = \mathbf{0}$  and  $\nu \rightarrow \infty$ .

Alternatively, the rMST distribution can be hierarchically represented as

$$\begin{aligned} \mathbf{Y} \mid (\gamma, \tau) &\sim N_p(\mu + \lambda\gamma, \Sigma/\tau), \\ \gamma \mid \tau &\sim TN(0, 1/\tau; (0, \infty)), \\ \tau &\sim \text{Gamma}(\nu/2, \nu/2), \end{aligned} \tag{2}$$

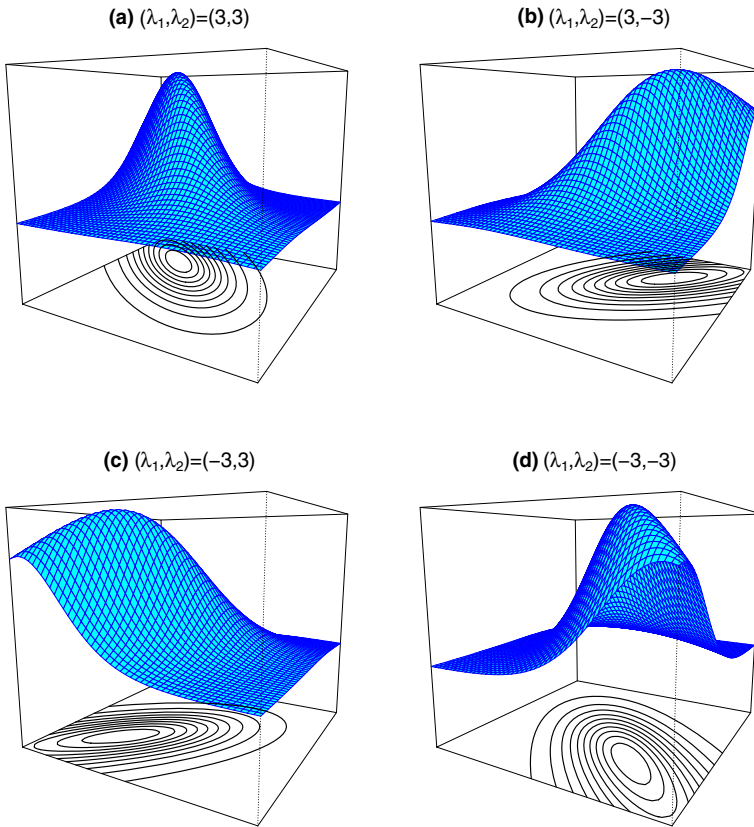
where  $\text{Gamma}(\alpha, \beta)$  stands for the gamma distribution with mean  $\alpha/\beta$ . Figure 1 shows the perspective plots with added contours for rMST densities under  $\mu = (0, 0)^\top$ ,  $\Sigma = \mathbf{I}_2$ ,  $\nu = 4$  and various specifications of  $\lambda = (\lambda_1, \lambda_2)^\top$ . It is clearly seen that these plots are non-elliptical and can be skewed and correlated toward different directions depending on the chosen parameters. Therefore, the rMST distribution provides a flexible mechanism to adapt well to more complicated data.

### 3 Methodology

#### 3.1 Model formulation

Suppose that  $\mathbf{Y} = (\mathbf{Y}_1, \dots, \mathbf{Y}_n)$  forms a random sample of size  $n$  in which each  $\mathbf{Y}_j = (Y_{j1}, \dots, Y_{jp})^\top$  is a  $p$ -dimensional vector of feature variables. Suppose further that these samples come independently from  $g$  distinct subgroups in a heterogeneous population. The MCcrstFA model for each  $\mathbf{Y}_j$  is

$$\mathbf{Y}_j = \mathbf{A}\mathbf{U}_{ij} + \mathbf{e}_{ij} \quad \text{with probability } \pi_i \quad (i = 1, \dots, g), \tag{3}$$



**Fig. 1** The contours of bivariate rMST distribution with  $\mu = (0, 0)^\top$ ,  $\Sigma = I_2$  and  $\nu = 4$  for different values of  $\lambda_1$  and  $\lambda_2$

for  $j = 1, \dots, n$ , where  $A$  is a  $p \times q$  matrix of common factor loadings,  $U_{ij}$  is a  $q$ -dimensional ( $q < p$ ) vector of component factors,  $e_{ij}$  is a  $p$ -dimensional vector of component errors, and  $\pi_i$ s are the mixing proportions subject to  $\sum_{i=1}^g \pi_i = 1$ .

Furthermore, we assume that  $U_{ij}$  and  $e_{ij}$  are jointly distributed as

$$\begin{bmatrix} U_{ij} \\ e_{ij} \end{bmatrix} \sim rST_{p+q} \left( \begin{bmatrix} \xi_i \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \Omega_i & \mathbf{0} \\ \mathbf{0} & D_i \end{bmatrix}, \begin{bmatrix} \lambda_i \\ \mathbf{0} \end{bmatrix}, \nu_i \right), \tag{4}$$

where  $\xi_i$  is a  $q$ -dimensional location vector,  $\Omega_i$  is a  $q \times q$  positive-definite scale covariance matrix,  $\lambda_i \in \mathbb{R}^q$  is a skewness vector,  $D_i$  is a  $p \times p$  positive diagonal matrix, and  $\nu_i$  is the DOF. The specifications of  $D_i$  and  $\nu_i$  in (4) can be either constrained to be equal or allowed to vary among components.

Based on (3) along with assumption (4), the pdf of  $Y_j$  is

$$f(y_j) = \sum_{i=1}^g \pi_i \psi_p(y_j; \mu_i, \Sigma_i, \alpha_i, \nu_i), \tag{5}$$

where

$$\boldsymbol{\mu}_i = \mathbf{A}\boldsymbol{\xi}_i, \quad \boldsymbol{\Sigma}_i = \mathbf{A}\boldsymbol{\Omega}_i\mathbf{A}^\top + \mathbf{D}_i, \quad \boldsymbol{\alpha}_i = \mathbf{A}\boldsymbol{\lambda}_i, \tag{6}$$

and  $\psi_p(\mathbf{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \boldsymbol{\alpha}_i, v_i)$  is the rMST density function defined in (1). Notice that the representations in (6) cannot be uniquely determined because they remain unchanged if the common factor loading matrix  $\mathbf{A}$  is postmultiplied by any nonsingular matrix. Thus, we must impose  $q^2$  constraints to achieve identifiability of  $\mathbf{A}$ . As a result, the number of free parameters in the MCrstFA is

$$d_1 = (g - 1) + pg + q(p + g) + \frac{1}{2}gq(q + 1) - q^2 + gq + g.$$

If  $\mathbf{D}_i$ s are constrained to be homogeneous across components, the number of parameters is

$$d_2 = (g - 1) + p + q(p + g) + \frac{1}{2}gq(q + 1) - q^2 + gq + g;$$

and if component DOFs are further assumed to be identical, the resulting number of parameters is

$$d_3 = (g - 1) + p + q(p + g) + \frac{1}{2}gq(q + 1) - q^2 + gq + 1.$$

We remark that the number of parameters in MCrstFA is increased by  $qg$  involved in  $\boldsymbol{\lambda}_i$  (without adding too much complexity) as compared with MCFA and MCtFA.

To indicate the class membership of observation  $\mathbf{y}_j$ , we introduce allocation variables  $\mathbf{Z}_j = (Z_{1j}, \dots, Z_{gj})^\top$ , defined as

$$Z_{ij} = \begin{cases} 1, & \mathbf{Y}_j \text{ belongs to } i\text{th component;} \\ 0, & \text{otherwise.} \end{cases}$$

Thus, we have  $\mathbf{Z}_j \stackrel{\text{iid}}{\sim} \mathcal{M}(1; \pi_1, \dots, \pi_g)$ , meaning a multinomial distribution with  $g$  possible outcomes which can occur in a single trial, where  $\pi_i = \Pr(Z_{ij} = 1)$  can be regarded as the prior probability of  $\mathbf{y}_j$  belonging to the  $i$ th component.

According to (2) and (3), the MCrstFA model can be formulated by a five-level hierarchical representation:

$$\begin{aligned} \mathbf{Y}_j \mid (\mathbf{U}_{ij}, \gamma_j, \tau_j, Z_{ij} = 1) &\sim N_p(\mathbf{A}\mathbf{U}_{ij}, \tau_j^{-1}\mathbf{D}_i), \\ \mathbf{U}_{ij} \mid (\gamma_j, \tau_j, Z_{ij} = 1) &\sim N_q(\boldsymbol{\xi}_i + \boldsymbol{\lambda}_i\gamma_j, \tau_j^{-1}\boldsymbol{\Omega}_i), \\ \gamma_j \mid (\tau_j, Z_{ij} = 1) &\sim TN(0, \tau_j^{-1}; (0, \infty)), \\ \tau_j \mid (Z_{ij} = 1) &\sim \text{Gamma}\left(\frac{v_i}{2}, \frac{v_i}{2}\right), \\ \mathbf{Z}_j &\sim \mathcal{M}(1; \pi_1, \dots, \pi_g). \end{aligned} \tag{7}$$

By Bayes' rule, it suffices to derive the following conditional distributions, and the proofs of which are sketched in ‘‘Appendix A’’. Specifically,

$$\begin{aligned}
 U_{ij} \mid (\mathbf{y}_j, \gamma_j, \tau_j, Z_{ij} = 1) &\sim N_q \left( \boldsymbol{\xi}_i + \boldsymbol{\lambda}_i \gamma_j + \boldsymbol{\beta}_i^\top (\mathbf{y}_j - \boldsymbol{\mu}_i - \boldsymbol{\alpha}_i \gamma_j), \tau_j^{-1} (\mathbf{I}_q \right. \\
 &\quad \left. - \boldsymbol{\beta}_i^\top \mathbf{A}) \boldsymbol{\Omega}_i \right), \\
 \gamma_j \mid (\mathbf{y}_j, \tau_j, Z_{ij} = 1) &\sim TN(h_{ij}, \tau_j^{-1} \sigma_i^2; (0, \infty)), \\
 f(\tau_j \mid \mathbf{y}_j, Z_{ij} = 1) &= \frac{\Phi(\sqrt{\tau_j} M_{ij})}{T\left(M_{ij} \sqrt{\frac{v_i+p}{v_i+\delta_{ij}}}; v_i+p\right)} g\left(\tau_j; \frac{v_i+p}{2}, \frac{v_i+\delta_{ij}}{2}\right), \\
 Z_{ij} = 1 \mid \mathbf{y}_j &\sim \mathcal{M}(1; \tilde{\pi}_{1j}, \dots, \tilde{\pi}_{gj}),
 \end{aligned} \tag{8}$$

where  $\boldsymbol{\beta}_i = \boldsymbol{\Sigma}_i^{-1} \mathbf{A} \boldsymbol{\Omega}_i$ ,  $\delta_{ij} = (\mathbf{y}_j - \boldsymbol{\mu}_i)^\top \mathbf{V}_i^{-1} (\mathbf{y}_j - \boldsymbol{\mu}_i)$ , and  $M_{ij} = h_{ij} / \sigma_i$  with  $\mathbf{V}_i = \boldsymbol{\Sigma}_i + \boldsymbol{\alpha}_i \boldsymbol{\alpha}_i^\top$ ,  $h_{ij} = \boldsymbol{\alpha}_i^\top \mathbf{V}_i^{-1} (\mathbf{y}_j - \boldsymbol{\mu}_i)$  and  $\sigma_i^2 = 1 - \boldsymbol{\alpha}_i^\top \mathbf{V}_i^{-1} \boldsymbol{\alpha}_i$ . Moreover,

$$\tilde{\pi}_{ij} = P(Z_{ij} = 1 \mid \mathbf{y}_j) = \frac{\pi_i \psi_p(\mathbf{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \boldsymbol{\alpha}_i, v_i)}{\sum_{h=1}^g \pi_h \psi_p(\mathbf{y}_j; \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h, \boldsymbol{\alpha}_h, v_h)}. \tag{9}$$

To simplify the notation, we define  $\mathbf{b}_{ij} = \boldsymbol{\xi}_i + \boldsymbol{\beta}_i^\top (\mathbf{y}_j - \boldsymbol{\mu}_i)$  and  $c_{ij}(r) = \{(v_i + p + r) / (v_i + \delta_{ij})\}^{1/2}$  for  $r = -2, 0, 2$ , and let “ $\mid \dots$ ” represent conditioning on  $\mathbf{Y}_j = \mathbf{y}_j$  and  $Z_{ij} = 1$ . The following proposition summarizes some essential conditional expectations for implementing the ECME algorithm described in the next subsection.

**Proposition 1** Consider the posterior distributions given in (8), we establish the following conditional expectations:

$$\begin{aligned}
 E(\tau_j \mid \dots) &= \{c_{ij}(0)\}^2 \frac{T(M_{ij} c_{ij}(2); v_i + p + 2)}{T(M_{ij} c_{ij}(0); v_i + p)}, \\
 E(\gamma_j \mid \dots) &= h_{ij} + \frac{\sigma_i t(M_{ij} c_{ij}(-2); v_i + p - 2)}{c_{ij}(-2) T(M_{ij} c_{ij}(0); v_i + p)}, \\
 E(\tau_j \gamma_j \mid \dots) &= h_{ij} E(\tau_j \mid \dots) + \sigma_i c_{ij}(0) \frac{t(M_{ij} c_{ij}(0); v_i + p)}{T(M_{ij} c_{ij}(0); v_i + p)}, \\
 E(\tau_j \gamma_j^2 \mid \dots) &= \sigma_i^2 + h_{ij} E(\tau_j \gamma_j \mid \dots), \\
 E(\mathbf{U}_{ij} \mid \dots) &= \mathbf{b}_{ij} + (\boldsymbol{\lambda}_i - \boldsymbol{\beta}_i^\top \boldsymbol{\alpha}_i) E(\gamma_j \mid \dots), \\
 E(\tau_j \mathbf{U}_{ij} \mid \dots) &= \mathbf{b}_{ij} E(\tau_j \mid \dots) + (\boldsymbol{\lambda}_i - \boldsymbol{\beta}_i^\top \boldsymbol{\alpha}_i) E(\tau_j \gamma_j \mid \dots), \\
 E(\tau_j \gamma_j \mathbf{U}_{ij} \mid \dots) &= \mathbf{b}_{ij} E(\tau_j \gamma_j \mid \dots) + (\boldsymbol{\lambda}_i - \boldsymbol{\beta}_i^\top \boldsymbol{\alpha}_i) E(\tau_j \gamma_j^2 \mid \dots), \\
 E(\tau_j \mathbf{U}_{ij} \mathbf{U}_{ij}^\top \mid \dots) &= (\mathbf{I}_q - \boldsymbol{\beta}_i^\top \mathbf{A}) \boldsymbol{\Omega}_i + E(\tau_j \gamma_j \mathbf{U}_{ij} \mid \dots) (\boldsymbol{\lambda}_i - \boldsymbol{\beta}_i^\top \boldsymbol{\alpha}_i)^\top \\
 &\quad + E(\tau_j \mathbf{U}_{ij} \mid \dots) \mathbf{b}_{ij}^\top,
 \end{aligned} \tag{10}$$

and

$$E(\log \tau_j | \dots) = \frac{\int_{-\infty}^{M_{ij}} t \left( x; 0, \frac{v_i + \delta_{ij}}{v_i + p}, v_i + p \right) f_{v_i}(x) dx}{T \left( M_{ij} \sqrt{\frac{v_i + p}{v_i + \delta_{ij}}}; v_i + p \right)} + E(\tau_j | \dots) - \left( \frac{v_i + p}{v_i + \delta_{ij}} \right) + \text{DG} \left( \frac{v_i + p}{2} \right) - \log \left( \frac{v_i + \delta_{ij}}{2} \right), \quad (11)$$

where  $f_{v_i}(x)$  is defined by (B.10).

*Proof* The results follow directly from some fundamental matrix manipulations and the law of iterated expectations. See ‘‘Appendix B’’ for more details.  $\square$

### 3.2 Parameter estimation via the ECME algorithm

The EM algorithm (Dempster et al. 1977) is a popular iterative method for finding ML estimates when the data are incomplete or the model contains latent variables. The main advantage of EM lies in the fact of monotone convergence without sacrificing simplicity. One common limitation of the EM algorithm is that the M-step usually yields no closed forms for estimators of parameters. To overcome this weakness, Meng and Rubin (1993) proposed the expectation conditional maximization (ECM) algorithm to replace the M-step of EM with several computational simpler CM-steps, each of which maximizes the expected complete-data log-likelihood function (known as the  $Q$ -function) sequentially. Importantly, the authors also showed that the ECM algorithm preserves all desiring properties of EM. In certain situations, some of the CM-steps of ECM may be computationally intractable. Liu and Rubin (1994) advanced the ECM algorithm with the CM steps that maximize either the  $Q$ -function, called the CMQ-step, or the corresponding constrained actual log-likelihood function, called the CML-step. The method is referred to as the ECME algorithm.

For notational simplicity, we denote the observed data by  $\mathbf{y} = (y_1, \dots, y_n)$ , allocation indicators by  $\mathbf{Z} = (z_1, \dots, z_n)$ , latent factors by  $\mathbf{U} = (U_1, \dots, U_n)$ , hidden variables  $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_n)$  and scaling weight variables by  $\boldsymbol{\tau} = (\tau_1, \dots, \tau_n)$ . Therefore, the complete data  $\mathbf{y}_c$  comprise the observed data  $\mathbf{y}$  together with missing data  $\mathbf{y}_m = (\mathbf{Z}, \mathbf{U}, \boldsymbol{\gamma}, \boldsymbol{\tau})$ . From (5), it is readily seen that

$$Y_j | (Z_{ij} = 1) \sim rST_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \boldsymbol{\alpha}_i, v_i).$$

Therefore, the joint pdf of  $(\mathbf{Y}, \mathbf{Z})$  is

$$f(\mathbf{y}, \mathbf{z}) = \prod_{j=1}^n \prod_{i=1}^g \{\pi_i \psi_p(y_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \boldsymbol{\alpha}_i, v_i)\}^{z_{ij}}. \quad (12)$$

Let  $\boldsymbol{\theta}_i = (\pi_i, \boldsymbol{\xi}_i, \boldsymbol{\Omega}_i, \mathbf{D}_i, \boldsymbol{\lambda}_i, v_i)$  be the parameter vector belonging to the  $i$ -th component, and  $\boldsymbol{\Theta} = \{\mathbf{A}, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_g\}$  the entire unknown parameters to be estimated. According to (7), the complete-data log-likelihood function is



$$\begin{aligned} \ell_c(\Theta \mid \mathbf{y}_c) = & \sum_{i=1}^g \sum_{j=1}^n z_{ij} \left\{ \log \pi_i - \frac{1}{2} \log |\mathbf{D}_i| - \frac{\tau_j}{2} (\mathbf{y}_j - \mathbf{A}U_{ij})^\top \mathbf{D}_i^{-1} (\mathbf{y}_j - \mathbf{A}U_{ij}) \right. \\ & - \frac{1}{2} \log |\boldsymbol{\Omega}_i| - \frac{\tau_j}{2} (\mathbf{U}_{ij} - \boldsymbol{\xi}_i - \lambda_i \gamma_j)^\top \boldsymbol{\Omega}_i^{-1} (\mathbf{U}_{ij} - \boldsymbol{\xi}_i - \lambda_i \gamma_j) \\ & \left. - \log \Gamma\left(\frac{\nu_i}{2}\right) + \frac{\nu_i}{2} \log\left(\frac{\nu_i}{2}\right) + \frac{\nu_i}{2} \log \tau_j - \frac{\nu_i}{2} \tau_j \right\}. \end{aligned}$$

To evaluate the  $Q$ -function, defined as  $Q(\Theta \mid \hat{\Theta}^{(k)}) = E[\ell_c(\Theta \mid \mathbf{y}_c) \mid \mathbf{y}, \hat{\Theta}^{(k)}]$ , we first define the following conditional expectations:

$$\begin{aligned} \hat{z}_{ij}^{(k)} &= P(Z_{ij} = 1 \mid \mathbf{y}_j, \hat{\Theta}^{(k)}), \quad \hat{\tau}_{ij}^{(k)} = E(\tau_j \mid \mathbf{y}_j, \hat{\Theta}^{(k)}, Z_{ij} = 1), \\ \hat{\kappa}_{ij}^{(k)} &= E(\log \tau_j \mid \mathbf{y}_j, \hat{\Theta}^{(k)}, Z_{ij} = 1), \quad \hat{s}_{1ij}^{(k)} = E(\tau_j \gamma_j \mid \mathbf{y}_j, \hat{\Theta}^{(k)}, Z_{ij} = 1), \\ \hat{s}_{2ij}^{(k)} &= E(\tau_j \gamma_j^2 \mid \mathbf{y}_j, \hat{\Theta}^{(k)}, Z_{ij} = 1), \quad \hat{\eta}_{ij}^{(k)} = E(\tau_j U_{ij} \mid \mathbf{y}_j, \hat{\Theta}^{(k)}, Z_{ij} = 1), \\ \hat{\Psi}_{ij}^{(k)} &= E(\tau_j U_{ij} U_{ij}^\top \mid \mathbf{y}_j, \hat{\Theta}^{(k)}, Z_{ij} = 1), \quad \hat{\zeta}_{ij}^{(k)} = E(\tau_j \gamma_j U_{ij} \mid \mathbf{y}_j, \hat{\Theta}^{(k)}, Z_{ij} = 1) \end{aligned}$$

for  $i = 1, \dots, g$  and  $j = 1, \dots, n$ , which can be evaluated using (9), (10) and (11).

To update the mixture parameters  $\Theta$ , the ECME algorithm proceeds as follows:

E-step: Given  $\Theta = \hat{\Theta}^{(k)}$ , calculate the  $Q$ -function, obtained as

$$\begin{aligned} Q(\Theta \mid \hat{\Theta}^{(k)}) = & \sum_{i=1}^g \sum_{j=1}^n \hat{z}_{ij}^{(k)} \left\{ \log \pi_i - \frac{1}{2} \log |\mathbf{D}_i| - \frac{1}{2} \log |\boldsymbol{\Omega}_i| - \log \Gamma\left(\frac{\nu_i}{2}\right) \right. \\ & \left. + \frac{\nu_i}{2} \log\left(\frac{\nu_i}{2}\right) + \frac{\nu_i}{2} (\hat{\kappa}_{ij}^{(k)} - \hat{\tau}_{ij}^{(k)}) - \frac{1}{2} \text{tr}(\mathbf{D}_i^{-1} \boldsymbol{\Upsilon}_{ij} - \boldsymbol{\Omega}_i^{-1} \mathbf{A}_{ij}) \right\}, \end{aligned} \tag{13}$$

where

$$\boldsymbol{\Upsilon}_{ij} = \boldsymbol{\Upsilon}_{ij}(\mathbf{A}) = \hat{\tau}_{ij}^{(k)} \mathbf{y}_j \mathbf{y}_j^\top - \mathbf{y}_j \hat{\eta}_{ij}^{(k)\top} \mathbf{A}^\top - \mathbf{A} \hat{\eta}_{ij}^{(k)} \mathbf{y}_j^\top + \mathbf{A} \hat{\Psi}_{ij}^{(k)} \mathbf{A}^\top \tag{14}$$

and

$$\begin{aligned} \mathbf{A}_{ij} = \mathbf{A}_{ij}(\boldsymbol{\xi}, \boldsymbol{\lambda}) = & \hat{\Psi}_{ij}^{(k)} - \hat{\eta}_{ij}^{(k)} \boldsymbol{\xi}_i^\top - \hat{\zeta}_{ij}^{(k)} \boldsymbol{\lambda}_i^\top - \boldsymbol{\xi}_i \left( \hat{\eta}_{ij}^{(k)\top} - \hat{\tau}_{ij}^{(k)} \boldsymbol{\xi}_i^\top - \hat{s}_{1ij}^{(k)} \boldsymbol{\lambda}_i^\top \right) \\ & - \boldsymbol{\lambda}_i \left( \hat{\zeta}_{ij}^{(k)\top} - \hat{s}_{1ij}^{(k)} \boldsymbol{\xi}_i^\top - \hat{s}_{2ij}^{(k)} \boldsymbol{\lambda}_i^\top \right). \end{aligned} \tag{15}$$

CM-steps: Maximizing (13) with respect to  $\pi_i, \xi_i, \lambda_i, \mathbf{A}, \mathbf{\Omega}_i$  and  $\mathbf{D}_i$ , we obtain

$$\begin{aligned} \hat{\pi}_i^{(k+1)} &= \frac{1}{n} \sum_{j=1}^n \hat{z}_{ij}^{(k)}, \\ \hat{\xi}_i^{(k+1)} &= \frac{\left(\sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{\eta}_{ij}^{(k)}\right) \left(\sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{s}_{2ij}^{(k)}\right) - \left(\sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{\xi}_{ij}^{(k)}\right) \left(\sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{s}_{1ij}^{(k)}\right)}{\left(\sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{t}_{ij}^{(k)}\right) \left(\sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{s}_{2ij}^{(k)}\right) - \left(\sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{s}_{1ij}^{(k)}\right)^2}, \\ \hat{\lambda}_i^{(k+1)} &= \frac{\left(\sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{t}_{ij}^{(k)}\right) \left(\sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{\xi}_{ij}^{(k)}\right) - \left(\sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{s}_{1ij}^{(k)}\right) \left(\sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{\eta}_{ij}^{(k)}\right)}{\left(\sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{t}_{ij}^{(k)}\right) \left(\sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{s}_{2ij}^{(k)}\right) - \left(\sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{s}_{1ij}^{(k)}\right)^2}, \\ \hat{\mathbf{A}}^{(k+1)} &= \left(\sum_{i=1}^g \sum_{j=1}^n \hat{z}_{ij}^{(k)} \mathbf{y}_j \hat{\eta}_{ij}^{(k)\top}\right) \left(\sum_{i=1}^g \sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{\psi}_{ij}^{(k)}\right)^{-1}, \\ \hat{\mathbf{\Omega}}_i^{(k+1)} &= \frac{\sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{\mathbf{A}}_{ij}^{(k+1)}}{\sum_{j=1}^n \hat{z}_{ij}^{(k)}} \text{ and } \hat{\mathbf{D}}_i^{(k+1)} = \frac{\text{Diag}\{\sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{\Upsilon}_{ij}^{(k+1)}\}}{\sum_{j=1}^n \hat{z}_{ij}^{(k)}}, \end{aligned}$$

where  $\hat{\Upsilon}_{ij}^{(k+1)}$  and  $\hat{\mathbf{A}}_{ij}^{(k+1)}$  are  $\Upsilon_{ij}$  and  $\mathbf{A}_{ij}$  in (14) and (15) with  $\xi_i, \lambda_i$  and  $\mathbf{A}$  replaced by  $\hat{\xi}_i^{(k+1)}, \hat{\lambda}_i^{(k+1)}$  and  $\hat{\mathbf{A}}^{(k+1)}$ , respectively. Moreover, when  $\mathbf{D}_i$ s are assumed to be the same, say  $\mathbf{D}_i = \mathbf{D}$  for all  $i$ , the updated estimator of  $\mathbf{D}$  is given by  $\hat{\mathbf{D}}^{(k+1)} = n^{-1} \text{Diag}\{\sum_{i=1}^g \sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{\Upsilon}_{ij}^{(k+1)}\}$ . The proof of the updated estimators is sketched in ‘‘Appendix C’’.

CML-step: In light of (12), the updated estimator of  $v_i$  can be obtained by solving the following equations:

$$\hat{v}_i^{(k+1)} = \arg \max_{v_i} \left\{ \sum_{j=1}^n \hat{z}_{ij}^{(k+1)} \log \left( \psi_p(\mathbf{y}_j; \hat{\boldsymbol{\mu}}_i^{(k+1)}, \hat{\boldsymbol{\Sigma}}_i^{(k+1)}, \hat{\boldsymbol{\alpha}}_i^{(k+1)}, v_i) \right) \right\}, \tag{16}$$

for  $i = 1, \dots, g$ , where  $\hat{\boldsymbol{\mu}}_i^{(k+1)} = \hat{\mathbf{A}}^{(k+1)} \hat{\xi}_i^{(k+1)}$ ,  $\hat{\boldsymbol{\Sigma}}_i^{(k+1)} = \hat{\mathbf{A}}^{(k+1)} \hat{\mathbf{\Omega}}_i^{(k+1)} \hat{\mathbf{A}}^{(k+1)\top} + \hat{\mathbf{D}}_i^{(k+1)}$  and  $\hat{\boldsymbol{\alpha}}_i^{(k+1)} = \hat{\mathbf{A}}^{(k+1)} \hat{\lambda}_i^{(k+1)}$ .

In the case of assuming common DOFs, say  $v_i = v$  for all  $i$ , the updated estimator of  $v$  is obtained by maximizing the constrained actual log-likelihood function, that is,

$$\hat{v}^{(k+1)} = \arg \max_v \left\{ \sum_{j=1}^n \log \left( \sum_{i=1}^g \hat{\pi}_i^{(k+1)} \psi_p(\mathbf{y}_j; \hat{\boldsymbol{\mu}}_i^{(k+1)}, \hat{\boldsymbol{\Sigma}}_i^{(k+1)}, \hat{\boldsymbol{\alpha}}_i^{(k+1)}, v) \right) \right\}. \tag{17}$$

Herein, we remark that the solutions of (16) and (17) involve carrying out a one-dimensional search using the built-in R function `optim` function over a box constraint

(2, 200). Given an initial guess of parameters  $\hat{\Theta}^{(0)}$ , the above ECME procedure is performed recursively until maximization of the log-likelihood function is achieved. The resulting ML estimates are denoted by  $\hat{\Theta} = (\hat{A}, \hat{\pi}_i, \hat{\xi}_i, \hat{\Omega}_i, \hat{D}_i, \hat{\lambda}_i, \hat{v}_i, i = 1, \dots, g)$ . As a result, the posterior probability of  $y_j$  belonging to the  $i$ -th component of the mixture is calculated by replacing  $\Theta$  in (9) with  $\Theta = \hat{\Theta}$ , denoted by  $\hat{z}_{ij} = P(Z_{ij} = 1 | y_j, \hat{\Theta})$ . Based on the maximum a posteriori (MAP) classification rule,  $y_j$  is assigned to group  $s$  if  $\max\{\hat{z}_{ij}\}_{i=1}^g$  occurs at  $i = s$ .

Consequently, the conditional expectations of the factor scores  $U_{ij}$  given  $y_j$  and the  $i$ -th membership of the mixture meaning that  $Z_{ij} = 1$  can be estimated by  $\hat{u}_{ij} = E(U_{ij} | Y_j = y_j, Z_{ij} = 1, \hat{\Theta})$  which is given in (10) with  $\Theta$  substituted by  $\hat{\Theta}$ . Then, the  $j$ -th estimated factor scores corresponding to  $y_j$  can be calculated as

$$\hat{u}_j = \sum_{i=1}^g \hat{z}_{ij} \hat{u}_{ij}, \quad j = 1 \dots n. \tag{18}$$

An alternative estimator of (18) is given by

$$\hat{u}_j = \sum_{i=1}^g \text{MAP}\{\hat{z}_{ij}\} \hat{u}_{ij}, \tag{19}$$

where  $\text{MAP}\{\hat{z}_{ij}\} = 1$ , if  $\max\{\hat{z}_{hj}\}_{h=1}^g$  occurs at  $h = i$ , and  $\text{MAP}\{\hat{z}_{ij}\} = 0$  otherwise. These estimated factor scores can be used to portray the observed data into a lower dimensional space (Baek et al. 2010; Baek and McLachlan 2011) and be applied to feature extractions (Ueda et al. 2000).

## 4 Practical issues from computational aspects

### 4.1 Initialization and stopping rules

Like other iterative procedures, the ECME algorithm may suffer from convergence difficulties such as singularity of component covariance matrices or undetermined local maximum. To alleviate such problems, one simple strategy is to try many different initial values and select the solution that provides the highest likelihood. To obtain different sets of initial values, this can be done by performing multiple times of  $K$ -means (Hartigan and Wong 1979) clustering or *random starts* (McLachlan and Peel 2000) in the sense that each sample point is randomly assigned to one of clusters. We recommend below a simple way of generating sensible initial values.

1. Given initial memberships obtained by a single run of clustering through  $K$ -means, we set  $\hat{Z}_j^{(0)} = (\hat{z}_{1j}^{(0)}, \dots, \hat{z}_{gj}^{(0)})$ . The initial values of  $\pi_i$ s are

$$\hat{\pi}_i^{(0)} = \frac{1}{n} \sum_{j=1}^n \hat{z}_{ij}^{(0)}, \quad i = 1, \dots, g.$$

- Let  $\mathbf{y}_{(i)}$  be the collection of the  $i$ -th partitioned group. After that, we compute factor scores using the R built-in `factanal` function. The initial estimates of  $\hat{\xi}_i^{(0)}$ ,  $\hat{\Omega}_i^{(0)}$ ,  $\hat{\lambda}_i^{(0)}$  and  $\hat{\nu}_i^{(0)}$ , for  $i = 1, \dots, g$ , are obtained by implementing R `EMMIXskew` package (Wang et al. 2009) for fitting the rMST distribution to the estimated factor scores.
- Perform the *principal components analysis* (PCA) method to obtain the factor loading matrix for  $\mathbf{y}_{(i)}$ , denoted by  $\hat{\mathbf{B}}_i^{(0)}$  for  $i = 1, \dots, g$ . The initial estimate of  $\mathbf{A}$  is specified as

$$\hat{\mathbf{A}}^{(0)} = \sum_{i=1}^g \hat{\pi}_i^{(0)} \hat{\mathbf{B}}_i^{(0)} \hat{\Omega}_i^{(0)-1/2}.$$

- The initial estimate of  $\mathbf{D}_i$  is obtained as a diagonal matrix formed from the diagonal elements of the sample covariance matrix of  $\mathbf{y}_{(i)}$ . For the restricted case of  $\mathbf{D}_i = \mathbf{D}$ , the initial estimate  $\hat{\mathbf{D}}^{(0)}$  is formed as the diagonal elements of the pooled within-cluster sample covariance matrix of  $\mathbf{y}_{(1)}, \dots, \mathbf{y}_{(g)}$ .

Since the ECME algorithm is an iterative method, the stopping rules should be specified. In our experimental studies, we adopt by default the traditional criterion to terminate the algorithm when a predefined the maximum number of iterations  $k_{\max} = 2 \times 10^4$  is reached or when the difference between two successive log-likelihood values is less than  $10^{-6}$ . Alternatively, one can use the Aitken acceleration-based stopping criterion (Aitken 1926; McLachlan and Krishnan 2008), which is at least as strict as lack of progress in likelihood in the neighborhood of a maximum (McNicholas et al. 2010).

## 4.2 Model selection and performance evaluation

The log-likelihood value cannot be adopted as a model selection criterion because it is a nondecreasing function of the number of components ( $g$ ) and the dimension of factors ( $q$ ). We use the Bayesian information criterion (BIC; Schwarz 1978) and the integrated classification likelihood (ICL; Biernacki et al. 2000) to determine the best pair of  $(g, q)$  over a number of candidate models for achieving satisfactory performance (McNicholas and Murphy 2008; Lin et al. 2016). The BIC and ICL are defined as

$$\text{BIC} = d \log n - 2\ell_{\max} \quad \text{and} \quad \text{ICL} = \text{BIC} + 2\text{ENT}(\hat{\mathbf{z}}),$$

where  $d$  is the number of free parameters,  $\ell_{\max}$  is the maximized log-likelihood value, and  $\text{ENT}(\hat{\mathbf{z}}) = -\sum_{i=1}^g \sum_{j=1}^n \hat{z}_{ij} \log \hat{z}_{ij}$  is a penalty term called *entropy* that favors well-separated mixtures. The ICL penalizes complex model seriously and selects more parsimonious models than does BIC.

To evaluate the clustering performance of model-based approach, the adjusted Rand index (ARI; Hubert and Arabie 1985) and the correct classification rate (CCR; Lee et al. 2003) are employed. Typically, the ARI value ranges between 0 and 1 in most cases, but it can be negative corresponding to a poor level of agreement, e.g., fewer instances

are correctly classified than would be expected by chance. The metric of CCR has a value between 0 and 1. The CCR is determined to have the lowest misclassification rate by comparing all permutations of the MAP clustering labels with the true class labels.

### 4.3 Identifiability issues

The mixture model itself suffers from a non-identifiability problem arising from a permutation of the class labels in parameter vectors. The switching issue of class labels is often inherent in Bayesian implementation of mixture models. However, this is not a problem in practice when employing the EM-based algorithm to estimate mixture densities since we can still determine a sequence of ML estimates that are consistent and asymptotically efficient, see McLachlan and Basford (1988).

On the other hand, there is another identifiability problem corresponding to the rotational indeterminacy of common factor loading matrix *A*. As suggested by Baek et al. (2010), a unique solution of *A*, say  $\hat{A}^*$ , can be obtained by postmultiplying a nonsingular matrix for which the solution is orthonormal, i.e.,  $\hat{A}^{*\top} \hat{A}^* = \mathbf{I}_q$ . This can be achieved by adopting the Cholesky decomposition to find the upper triangular matrix *C* of order *q* such that  $\hat{A}^\top \hat{A} = \mathbf{C}^\top \mathbf{C}$ , resulting in  $\hat{A}^* = \hat{A} \hat{\mathbf{C}}^{-1}$ .

Related to the standard errors of the ML estimates, it would be of interest to calculate them using the empirical information matrix for  $\Theta$  in a manner analogous to Wang and Lin (2016). This procedure will be tackled by the authors in a future paper.

## 5 Simulation

We conduct two simulation experiments to demonstrate the proposed techniques. Unless otherwise stated, we shall consider only the case of  $D_i = D$  for all *i* in the later analysis.

### 5.1 Experiment 1

In this experiment, to compare the accuracy of three parsimonious factor-analytic approaches for clustering and representing low-dimensional data, we generate a *p* = 3 dimensional dataset of size *n* = 1000 from a *g* = 2 component mixture of rMST distributions. The presumed mixture parameters as involved in (5) are

$$\begin{aligned} \pi_1 &= 0.5, \quad \pi_2 = 0.5, \quad \boldsymbol{\mu}_1 = (0, 0, 0)^\top, \quad \boldsymbol{\mu}_2 = (1, 1, 3)^\top, \\ \nu_1 &= 4, \quad \nu_2 = 5, \quad \boldsymbol{\alpha}_1 = (-2, -5, -5)^\top, \quad \boldsymbol{\alpha}_2 = (-2, 5, 5)^\top, \\ \boldsymbol{\Sigma}_1 &= \begin{bmatrix} 4 & -1.8 & -1 \\ -1.8 & 2 & 0.9 \\ -1 & 0.9 & 2 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 4 & 1.8 & 0.8 \\ 1.8 & 2 & 0.5 \\ 0.8 & 0.5 & 2 \end{bmatrix}. \end{aligned}$$

**Table 1** Cross-tabulations of true (A, B) and predicted (1, 2) class memberships for three parsimonious factor-analytic approaches for the simulated data

	MCFA		MCtFA		MCRstFA	
	1	2	1	2	1	2
A	61	439	490	10	495	5
B	51	449	38	462	23	477

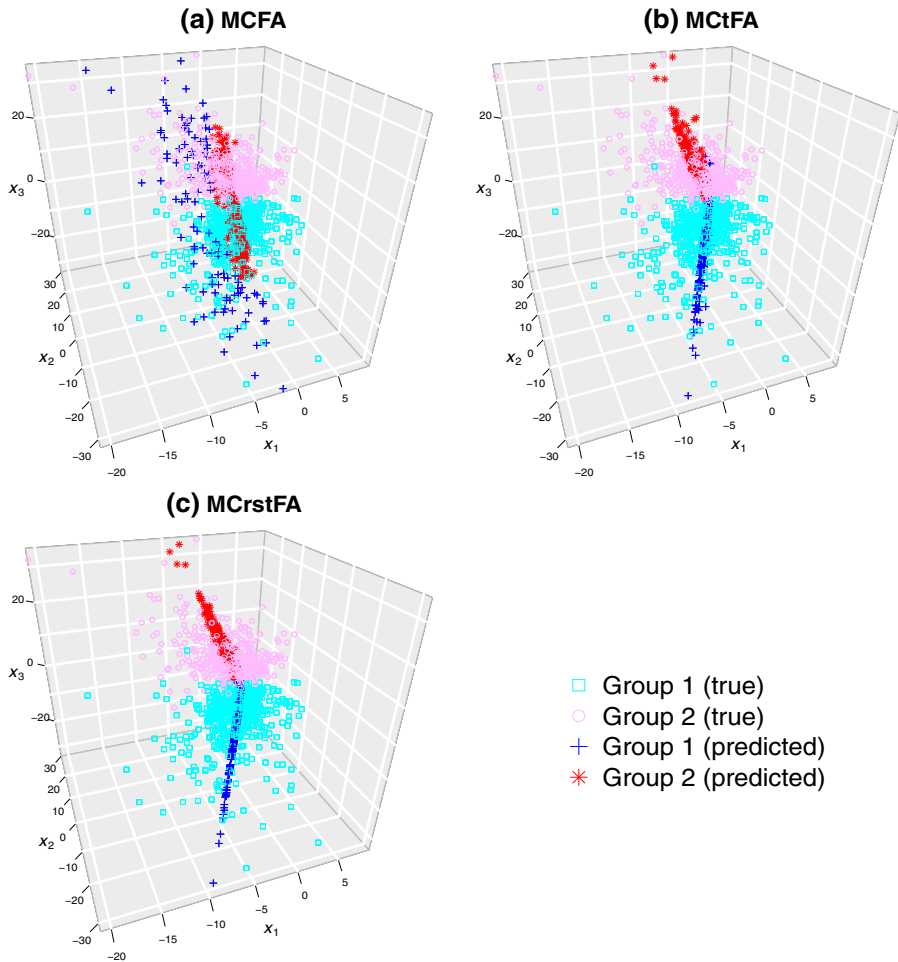
The MCFA, MCtFA and MCRstFA models with  $q = 2$  factors and  $g = 2$  components are fitted via the ECME algorithm to the simulated data. When the parameter estimates and the corresponding factor scores are obtained under each fitted model, we can compare the clustering performance and calculate the predicted values of each observed feature vector  $y_j$ . As anticipated, the MCRstFA approach gives the best clustering result (ARI = 0.891; CCR = 0.972), followed closely by MCtFA (ARI = 0.817; CCR = 0.952). The MCFA has the worst performance (ARI =  $1.78 \times 10^{-6}$ ; CCR = 0.51), indicating a lack of ability to cluster mixtures of skewed data with outliers. A cross-tabulation of the true and predicted class memberships is given in Table 1. As can be seen, the MCRstFA approach provides fewer misclassified observations and outperforms the other two considered approaches, say MCtFA and MCFA.

Figure 2 displays plots of the actual observations  $y_j$  overlaid with predicted observations  $\hat{y}_j$ , calculated as  $\hat{y}_j = \hat{A}\hat{u}_j$ , ( $j = 1, \dots, 1000$ ), where  $\hat{A}$  is the estimated projection matrix, and  $\hat{u}_j$  is the estimated factor scores defined in (18). As shown in Fig. 2a, the MCFA model performs poorly because of a lack of mechanisms to cope with data exhibiting non-normal features. On the other hand, it is clearly observed from Fig. 2b, c that the original scattering structure of two groups can be retrieved quite well using the MCtFA and MCRstFA approaches, but the MCtFA is slightly unfavored due to somewhat poor fit caused by having 20 more misclassified units than the MCRstFA.

### 5.2 Experiment 2

To further demonstrate the validity of the MCRstFA approach for handling the data of higher dimensions, we perform a second simulation experiment in situations where the MCRstFA holds exactly. In this study, data were generated from the 3-component MCRstFA model with  $q = 2$ , and  $p = 10$  and 20. We perform 100 Monte Carlo (MC) repetitions of sample size  $n = 1500$  observations and equal mixing proportions, namely  $\pi_i = 1/3$  for all  $i$ . The elements of  $p \times q$  common factor loadings  $A$  were randomly generated from  $N(0, 1)$ , while the component DOFs are taken as  $(v_1, v_2, v_3) = (4, 6, 9)$ . The location vectors, scale-covariance matrices and skewness parameters of the component factors  $U_{ij}$  are chosen as

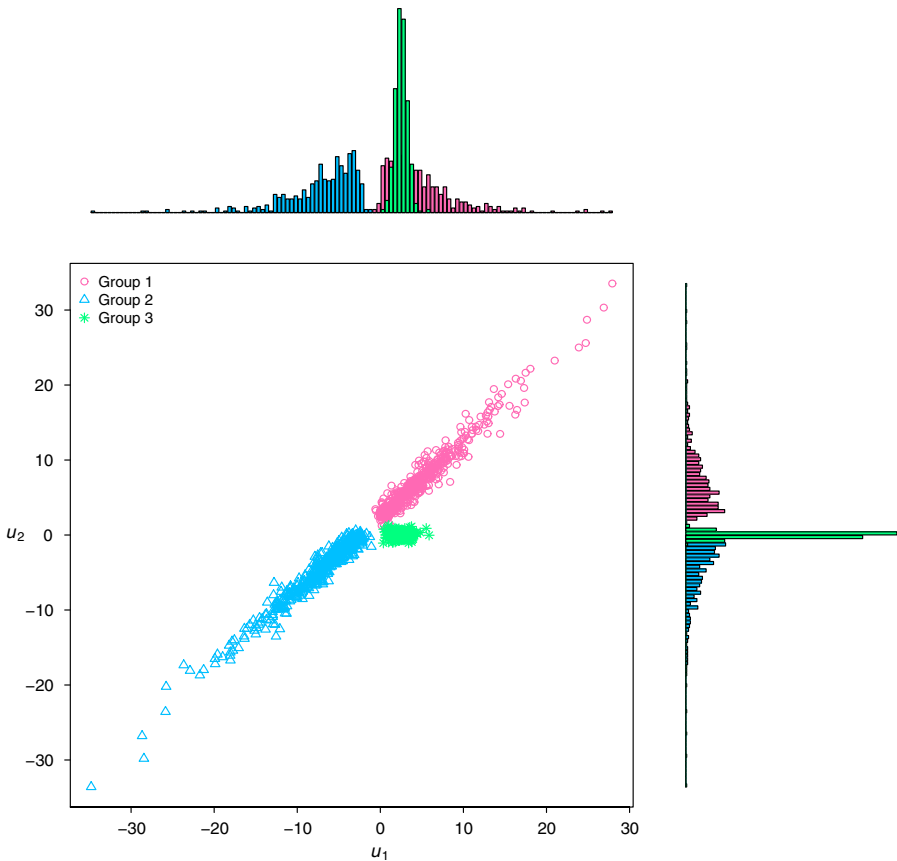
$$\begin{aligned} \xi_1 &= (0, 2.5)^\top, & \xi_2 &= (-2.5, 0)^\top, & \xi_3 &= (2.5, 0)^\top, \\ \lambda_1 &= (5, 5)^\top, & \lambda_2 &= (-5, -5)^\top, & \lambda_3 &= (0, 0)^\top, \\ \Omega_1 &= \begin{bmatrix} 0.1 & 0 \\ 0 & 0.45 \end{bmatrix}, & \Omega_2 &= \begin{bmatrix} 0.45 & 0 \\ 0 & 0.1 \end{bmatrix}, & \Omega_3 &= \begin{bmatrix} 0.45 & 0 \\ 0 & 0.1 \end{bmatrix}. \end{aligned}$$



**Fig. 2** Original observations and the predicted observations by MCFA, MCtFA, and MCcstFA

Figure 3 gives an illustration of the generated bivariate factor scores based on one simulated case for each of the three components. Typically, these component factor scores look somewhat well separated and exhibit non-elliptical scattering patterns and heavy tails. The component error vectors  $e_{ij}$  were drawn independently from  $t_p(\mathbf{0}, \mathbf{D}, \nu_i)$ , where diagonal elements of  $\mathbf{D}$  were randomly generated from a uniform distribution ranging between 0.1 and 0.3.

We process each of 100 MC simulated datasets by fitting the MCFA, MCtFA and MCcstFA models. Comparisons were made on the adequacy of overall fitness in terms of BIC and ICL and the classification agreement on the true and predicted memberships assessed by ARI and CCR. Table 2 lists the average values of criteria together with the corresponding standard deviations (Std) under every scenario considered. As a guide to select the most plausible model, the frequencies (Freq) preferred by these



**Fig. 3** Scatter plot of generated bivariate factors for each of  $g = 3$  components

criteria are also reported. In all cases, the MCrtFA model provides better fits and clustering results than the other two approaches. In particular, the MCFA and MCtFA are seldom or even never chosen by these four indices due to a lack of sufficient robustness against skewness. We have also undertaken the simulation study with a much higher dimension, say  $p = 100$ , and found that the MCrtFA model still works similarly well without degrading its performance.

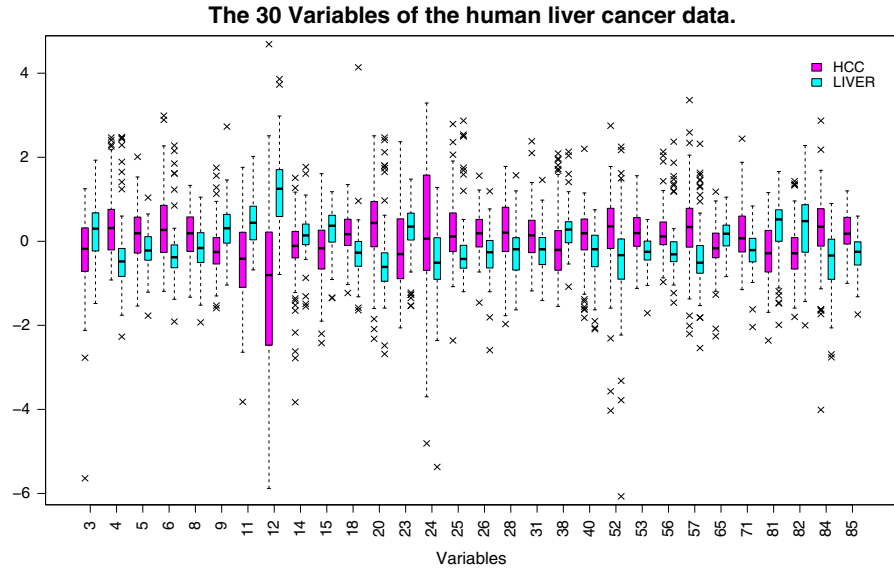
## 6 Application to real data

We applied our method to the human liver cancer data (Chen et al. 2002), which consist of  $p = 85$  gene expressions partitioned into two subpopulations. Hepatocellular carcinoma (HCC) is one of the 10 leading causes of death in the world. Chen et al. (2002) used cDNA microarrays to characterize patterns of gene expression in HCC, from which they found that the expression patterns in HCC and nontumor liver tissues (LIVER) are distinctly different from one another. In the data, there are  $n = 179$  sam-



**Table 2** Comparison of MCFA, MCIFA and MCcstFA models for simulation based on 100 replications

Criteria	$p = 10$			$p = 20$		
	MCFA	MCIFA	MCcstFA	MCFA	MCIFA	MCcstFA
BIC	Mean	-13,344.55	-15,354.82	-15,524.38	-37,598.39	-42,808.07
	Std	70.93	15.62	96.56	52.35	12.62
	Freq	0	0	100	0	0
ICL	Mean	-13,102.19	-14,685.13	-15,384.57	-37,052.25	-42,138.63
	Std	78.76	181.74	394.52	138.97	186.29
	Freq	0	0	100	0	0
ARI	Mean	0.5548	0.7706	0.9213	0.5447	0.7343
	Std	0.0041	0.2114	0.1570	0.0203	0.1701
	Freq	5	8	87	2	4
CCR	Mean	0.6773	0.8603	0.9530	0.6789	0.8425
	Std	0.0230	0.1415	0.1115	0.0352	0.1373
	Freq	6	7	87	4	2



**Fig. 4** Boxplots for the 30 genes in the human liver cancer data. The  $x$ -coordinate indicates the order of original genes

ples in the genomic expression patterns from patients, of which 104 belong to HCC and 75 to LIVER.

Figure 4 depicts the boxplots of top 30 genes which have the most significant difference between two classes obtained by performing the two-sample  $t$ -test. Apparently, the distribution of each selected gene is highly skewed or has a long tail.

We implement the two-component MCFA, MCFa, MCrtFA and MCghstFA approaches with  $q$  ranging from 1 to 10. In the same vein as that of the simulation experiments, we assume  $D_i = D$  for all  $i$ , but place no restrictions on component DOFs. A comparison of some characterizations between the MCrtFA and MCghstFA models is summarized in Table 5. When fitting the MCghstFA model, we implement the ECM algorithm described in “Appendix D”. For clarity, Table 3 presents only the fitting results and classification agreements of each method with  $q$  ranging from 5 to 10. Judging from BIC and ICL, the best fitted model is given by the MCghstFA model with  $q = 8$ . While comparing the classification performance, the MCrtFA model with  $q = 6$  provides the best agreement on predicting the true group memberships (ARI = 0.2427 and CCR = 0.7486) for this dataset. Notice that the best classifier does not necessarily give the best fit to the data. Again, the MCrtFA approach demonstrates its usefulness in clustering high-dimensional data with asymmetry and/or fat tails.

Table 4 compares the best classification results obtained from the fitted MCFA ( $q = 10$ ), MCFa ( $q = 6$ ), MCrtFA ( $q = 6$ ) and MCghstFA ( $q = 10$ ) models. We found that the number of the correctly classified HCC tissues in the fit of MCrtFA is more than those of the other three approaches. However, there is no obvious difference among them in predicting the class memberships of LIVER tissues.

**Table 3** Comparison of fitting results and implied clustering versus the true membership of the human liver cancer data

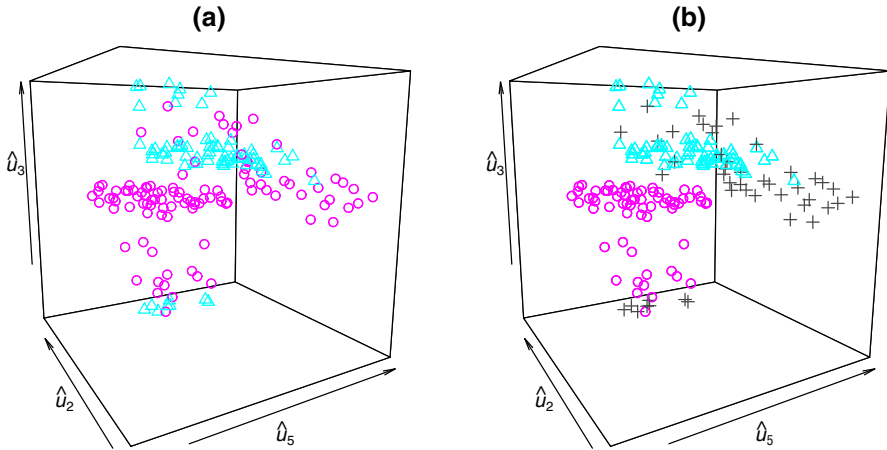
Model	Factors	$\ell_{\max}$	$d$	BIC	ICL	ARI	CCR
MCFA	5	- 15, 110.06	526	<b>32,948.69</b>	<b>32,948.69</b>	- 0.0028	0.5475
	6	- 14, 952.48	614	33,090.02	33,098.36	- 0.0006	0.5531
	7	- 14, 714.43	702	33,070.41	33,077.98	0.0022	0.5587
	8	- 14, 502.58	790	33,103.20	33,108.02	- 0.0121	0.5196
	9	- 14, 355.34	878	33,265.21	33,275.42	0.0490	0.6201
	10	- 14, 178.00	966	33,367.02	33,371.90	<b>0.0700</b>	<b>0.6369</b>
MCtFA	5	- 14, 198.9	528	31,136.73	31,136.77	- 0.0028	0.5475
	6	- 14, 026.77	616	31,248.97	31,278.82	<b>0.1020</b>	<b>0.6648</b>
	7	- 13, 700.85	704	<b>31,053.61</b>	<b>31,055.07</b>	- 0.0143	0.5196
	8	- 13, 491.35	792	31,091.12	31,094.50	- 0.0147	0.5140
	9	- 13, 266.24	880	31,097.39	31,106.04	0.0198	0.5866
	10	- 13, 131.82	968	31,285.04	31,292.45	0.0241	0.5922
MCrStFA	5	- 14, 154.58	538	31,099.97	31,100.02	- 0.0028	0.5475
	6	- 13, 986.08	628	31,229.84	31,253.25	<b>0.2427</b>	<b>0.7486</b>
	7	- 13, 655.81	718	<b>31,036.16</b>	<b>31,049.49</b>	- 0.0074	0.5363
	8	- 13, 443.11	808	31,077.62	31,106.66	0.0801	0.6480
	9	- 13, 204.65	898	31,067.58	31,076.36	0.0286	0.5978
	10	- 13, 069.24	988	31,263.61	31,270.71	0.0241	0.5922
MCghstFA	5	- 14, 131.31	538	31,053.42	31,053.44	- 0.0028	0.5475
	6	- 13, 971.05	628	31,199.77	31,226.22	0.0120	0.5754
	7	- 13, 477.30	718	30,679.15	30,679.19	- 0.0136	0.5251
	8	- 13, 095.40	808	<b>30,382.17</b>	<b>30,382.53</b>	- 0.0147	0.5140
	9	- 13, 079.05	898	30,816.38	30,816.44	- 0.0140	0.5084
	10	- 13, 012.70	988	31,150.53	31,151.09	<b>0.0755</b>	<b>0.6425</b>

The smallest BIC and ICL scores and the largest ARI and CCR values under each family of considered models are indicated in bold

**Table 4** Cross-tabulations of true and predicted (1,2) class memberships for four mixtures of common factor-analytic approaches for the human liver cancer data

	MCFA ( $q = 10$ )		MCtFA ( $q = 6$ )		MCrStFA ( $q = 6$ )		MCghstFA ( $q = 10$ )	
	1	2	1	2	1	2	1	2
HCC	47	57	53	51	67	37	48	56
LIVER	8	67	9	66	8	67	8	67

To visualize the clustering results in a low-dimensional space, Fig. 5 portrays the data in a 3D space using the factor scores estimated by (19). In the plot, we use the second, third and fifth factors in the fit of MCrStFA with  $q = 6$  factors. The estimated factor scores in Fig. 5a, b are plotted according to the true and implied clustering labels, respectively. It can be observed from the two plots that the two clusters are



**Fig. 5** Plot of the (estimated) posterior mean factor scores via the MCcrstFA approach for the human liver cancer data based on **a** the true class labels, and **b** the implied clustering labels. (○) HCC; (△) LIVER; (+) Misclassification

inherently overlapped so that no approach works satisfactorily on classifying these tissues. Most of the misclassified tissues, labelled by ‘plus symbol’ in Fig. 5b, appear in the overlapping area between two clusters.

## 7 Conclusion

We propose an extension of MCFA in which component factors and errors are jointly modeled by the rMST distribution, called the MCcrstFA model, as a new model-based tool for analyzing high-dimensional data with strong degree of abnormality and multimodality. An attractive feature of the MCcrstFA is that the component means, component covariance matrices as well as component skewness parameters are represented by common factor loadings, allowing parsimonious model fitting while preserving its robustness.

We describe an analytically simple ECME procedure developed under a five-level hierarchy for fitting the MCcrstFA. This approach enables us to project high-dimensional clustering results into a low-dimensional space through displaying estimated factor scores. Numerical simulation studies and experimental data demonstrate its usefulness and flexibility on the basis of model fitting and outright clustering.

The techniques presented so far are limited to the likelihood-based approach and focus on complete data analysis. Some possible avenues for future research include building a framework to handle the presence of censoring observations (Castro et al. 2015; Lachos et al. 2017) or the occurrence of missing values (Ouyang et al. 2004; Lin 2014; Wang et al. 2017a, b), both of which are common problems in the analysis of high-dimensional data. Although our estimating procedure is easy to implement, there is a lack of feasible guidelines for a joint determination of  $(g, q)$  within a single run of the training process. Toward this end, variational Bayes (VB) approximations (Waterhouse et al. 1996; Jordan et al. 1999; Beal 2003) have been presented as an

iterative Bayesian alternative to the EM-based algorithm for their fast and deterministic nature. The attractive feature of the VB scheme allows for an automated learning of parameter estimation and model selection. The VB approach has been effectively applied to Gaussian mixtures (Teschendorff et al. 2005), MFA models (Ghahramani and Beal 2000), and mixtures of normal inverse Gaussian distributions (Subedi and McNicholas 2014) for simultaneously estimating model parameters and determining the number of components. Therefore, it is worthwhile to develop a novel VB algorithm for learning the MCrstFA model. Another inspiration for future work is to extend the MCrstFA model based on a broader family of multivariate skew distributions such as the scale mixtures of skew-normal distributions (Cabral et al. 2012; Prates et al. 2013), the multivariate canonical fundamental skew- $t$  distributions (Arellano-Valle and Genton 2005; Lee and McLachlan 2016), and the hidden truncation hyperbolic distributions introduced very recently by Murray et al. (2017b).

**Acknowledgements** The authors gratefully acknowledge the Coordinating Editor, Maurizio Vichi, the Associate Editor and three anonymous referees for their comments and suggestions that greatly improved this paper. W.L. Wang and T.I. Lin would like to acknowledge the support of the Ministry of Science and Technology of Taiwan under Grant Nos. MOST 105-2118-M-035-004-MY2 and MOST 105-2118-M-005-003-MY2, respectively. L.M. Castro acknowledges support from Grant FONDECYT 1170258 from Chilean government.

### Appendix A: Proof of hierarchical representation (8)

It follows from (7) that

$$\begin{aligned}
 & E \left( \mathbf{Y}_j \mathbf{U}_{ij}^\top \mid \gamma_j, \tau_j, Z_{ij} = 1 \right) \\
 &= E \left[ E \left( \mathbf{Y}_j \mathbf{U}_{ij}^\top \mid \mathbf{U}_{ij}, \gamma_j, \tau_j, Z_{ij} = 1 \right) \mid \gamma_j, \tau_j, Z_{ij} = 1 \right] \\
 &= E \left[ E(\mathbf{Y}_j \mid \mathbf{U}_{ij}, \gamma_j, \tau_j, Z_{ij} = 1) \mathbf{U}_{ij}^\top \mid \gamma_j, \tau_j, Z_{ij} = 1 \right] \\
 &= E \left( \mathbf{A} \mathbf{U}_{ij} \mathbf{U}_{ij}^\top \mid \gamma_j, \tau_j, Z_{ij} = 1 \right) \\
 &= \mathbf{A} \left[ \tau_j^{-1} \boldsymbol{\Omega}_i + (\boldsymbol{\xi}_i + \boldsymbol{\lambda}_i \gamma_j)(\boldsymbol{\xi}_i + \boldsymbol{\lambda}_i \gamma_j)^\top \right],
 \end{aligned}$$

and

$$\begin{aligned}
 & \text{cov} \left( \mathbf{Y}_j, \mathbf{U}_{ij}^\top \mid \gamma_j, \tau_j, Z_{ij} = 1 \right) \\
 &= E \left( \mathbf{Y}_j \mathbf{U}_{ij}^\top \mid \gamma_j, \tau_j, Z_{ij} = 1 \right) - E(\mathbf{Y}_j \mid \gamma_j, \tau_j, Z_{ij} = 1) E \left( \mathbf{U}_{ij}^\top \mid \gamma_j, \tau_j, Z_{ij} = 1 \right) \\
 &= \mathbf{A} \left[ \tau_j^{-1} \boldsymbol{\Omega}_i + (\boldsymbol{\xi}_i + \boldsymbol{\lambda}_i \gamma_j)(\boldsymbol{\xi}_i + \boldsymbol{\lambda}_i \gamma_j)^\top \right] - (\boldsymbol{\mu}_i + \boldsymbol{\alpha}_i \gamma_j)(\boldsymbol{\xi}_i + \boldsymbol{\lambda}_i \gamma_j)^\top \\
 &= \mathbf{A} \left[ \tau_j^{-1} \boldsymbol{\Omega}_i + (\boldsymbol{\xi}_i + \boldsymbol{\lambda}_i \gamma_j)(\boldsymbol{\xi}_i + \boldsymbol{\lambda}_i \gamma_j)^\top \right] - \mathbf{A}(\boldsymbol{\xi}_i + \boldsymbol{\lambda}_i \gamma_j)(\boldsymbol{\xi}_i + \boldsymbol{\lambda}_i \gamma_j)^\top \\
 &= \tau_j^{-1} \mathbf{A} \boldsymbol{\Omega}_i.
 \end{aligned}$$

This gives rise to the following joint distribution:

$$\begin{bmatrix} \mathbf{Y}_j \\ \mathbf{U}_{ij} \end{bmatrix} \Big| (\gamma_j, \tau_j, Z_{ij} = 1) \sim N_{p+q} \left( \begin{bmatrix} \boldsymbol{\mu}_i + \boldsymbol{\alpha}_i \gamma_j \\ \boldsymbol{\xi}_i + \boldsymbol{\lambda}_i \gamma_j \end{bmatrix}, \tau_j^{-1} \begin{bmatrix} \boldsymbol{\Sigma}_i & \mathbf{A} \boldsymbol{\Omega}_i \\ \boldsymbol{\Omega}_i^\top \mathbf{A}^\top & \boldsymbol{\Omega}_i \end{bmatrix} \right).$$

We then have the following standard results:

$$\begin{aligned} E(\mathbf{U}_{ij} \mid \mathbf{y}_j, \gamma_j, \tau_j, Z_{ij} = 1) &= (\boldsymbol{\xi}_i + \boldsymbol{\lambda}_i \gamma_j) + \left( \tau_j^{-1} \boldsymbol{\Omega}_i^\top \mathbf{A}^\top \right) (\tau_j^{-1} \boldsymbol{\Sigma}_i)^{-1} (\mathbf{y}_j - \boldsymbol{\mu}_i - \boldsymbol{\alpha}_i \gamma_j) \\ &= \boldsymbol{\xi}_i + \boldsymbol{\lambda}_i \gamma_j + \boldsymbol{\beta}_i^\top (\mathbf{y}_j - \boldsymbol{\mu}_i - \boldsymbol{\alpha}_i \gamma_j), \end{aligned}$$

and

$$\begin{aligned} \text{cov}(\mathbf{U}_{ij} \mid \mathbf{y}_j, \gamma_j, \tau_j, Z_{ij} = 1) &= \tau_j^{-1} \boldsymbol{\Omega}_i - \left( \tau_j^{-1} \boldsymbol{\Omega}_i^\top \mathbf{A}^\top \right) (\tau_j^{-1} \boldsymbol{\Sigma}_i)^{-1} (\tau_j^{-1} \mathbf{A} \boldsymbol{\Omega}_i) \\ &= \tau_j^{-1} \left( \mathbf{I}_q - \boldsymbol{\beta}_i^\top \mathbf{A} \right) \boldsymbol{\Omega}_i, \end{aligned}$$

where  $\boldsymbol{\beta}_i = \boldsymbol{\Sigma}_i^{-1} \mathbf{A} \boldsymbol{\Omega}_i$ . Using the characterization of the multivariate normal distribution, we can obtain

$$\mathbf{U}_{ij} \mid (\mathbf{y}_j, \gamma_j, \tau_j, Z_{ij} = 1) \sim N_q \left( \boldsymbol{\xi}_i + \boldsymbol{\lambda}_i \gamma_j + \boldsymbol{\beta}_i^\top (\mathbf{y}_j - \boldsymbol{\mu}_i - \boldsymbol{\alpha}_i \gamma_j), \tau_j^{-1} \left( \mathbf{I}_q - \boldsymbol{\beta}_i^\top \mathbf{A} \right) \boldsymbol{\Omega}_i \right).$$

With similar arguments, we have

$$\begin{aligned} f(\mathbf{y}_j, \gamma_j, \tau_j \mid z_{ij} = 1) &= f(\mathbf{y}_j \mid \gamma_j, \tau_j, z_{ij} = 1) f(\gamma_j \mid \tau_j, z_{ij} = 1) f(\tau_j \mid z_{ij} = 1) \\ &= \frac{2|\boldsymbol{\Sigma}_i|^{-\frac{1}{2}} \left(\frac{v_i}{2}\right)^{\frac{v_i}{2}}}{(2\pi)^{\frac{p+1}{2}} \Gamma\left(\frac{v_i}{2}\right)} \tau_j^{\frac{p+v_i+1}{2}-1} \exp \left\{ -\frac{\tau_j}{2} \left[ \frac{(\gamma_j - h_i)^2}{\sigma_i^2} + \delta_{ij} + v_i \right] \right\}, \\ f(\mathbf{y}_j, \tau_j \mid z_{ij} = 1) &= \frac{2|\boldsymbol{\Sigma}_i|^{-\frac{1}{2}} \sigma_i \left(\frac{v_i}{2}\right)^{\frac{v_i}{2}}}{(2\pi)^{\frac{p}{2}} \Gamma\left(\frac{v_i}{2}\right)} \tau_j^{\frac{p+v_i}{2}-1} \\ &\quad \exp \left\{ -\frac{\tau_j}{2} [\delta_{ij} + v_i] \right\} \Phi(\sqrt{\tau_j} M_{ij}), \\ f(\gamma_j \mid \mathbf{y}_j, \tau_j, z_{ij} = 1) &= \frac{f(\mathbf{y}_j, \gamma_j, \tau_j \mid z_{ij} = 1)}{f(\mathbf{y}_j, \tau_j \mid z_{ij} = 1)} = \frac{\phi(\gamma_j; h_i, \tau_j^{-1} \sigma_i^2)}{\Phi(\sqrt{\tau_j} M_{ij})}. \end{aligned}$$

Hence, it is trivial to establish that  $\gamma_j \mid (\mathbf{y}_j, \tau_j, Z_{ij} = 1) \sim TN(h_i, \tau_j^{-1} \sigma_i^2; (0, \infty))$ . Furthermore, standard calculation gives

$$\begin{aligned}
 f(\tau_j \mid \mathbf{y}_j, Z_{ij} = 1) &= \frac{f(\mathbf{y}_j, \tau_j \mid Z_{ij} = 1)}{f(\mathbf{y}_j \mid Z_{ij} = 1)} \\
 &= \frac{\tau_j^{\frac{v_i+p}{2}-1}}{\Gamma\left(\frac{v_i+p}{2}\right)} \left(\frac{v_i + \delta_{ij}}{2}\right)^{\frac{v_i+p}{2}} \frac{\Phi(\sqrt{\tau_j} M_{ij})}{T\left(M_{ij} \sqrt{\frac{v_i+p}{v_i+\delta_{ij}}}; v_i + p\right)} \\
 &\quad \times \exp\left\{-\frac{\tau_j}{2} [\delta_{ij} + v_i]\right\} \\
 &= \frac{\Phi(\sqrt{\tau_j} M_{ij})}{T\left(M_{ij} \sqrt{\frac{v_i+p}{v_i+\delta_{ij}}}; v_i + p\right)} g\left(\tau_j; \frac{v_i + p}{2}, \frac{v_i + \delta_{ij}}{2}\right).
 \end{aligned}$$

**Appendix B: Proof of Proposition 1**

(a) Standard calculation of conditional expectation yields

$$\begin{aligned}
 E(\tau_j \mid \mathbf{y}_j, z_{ij} = 1) &= \int_0^\infty \tau_j f(\tau_j \mid \mathbf{y}_j, z_{ij} = 1) d\tau_j \\
 &= \int_0^\infty \tau_j \frac{\Phi(\sqrt{\tau_j} M_{ij})}{T\left(M_{ij} \sqrt{\frac{v_i+p}{v_i+\delta_{ij}}}; v_i + p\right)} g\left(\tau_j; \frac{v_i + p}{2}, \frac{v_i + \delta_{ij}}{2}\right) d\tau_j \\
 &= \frac{\left(\frac{v_i+p}{v_i+\delta_{ij}}\right)}{T\left(M_{ij} \sqrt{\frac{v_i+p}{v_i+\delta_{ij}}}; v_i + p\right)} \int_0^\infty \Phi(\sqrt{\tau_j} M_{ij}) g\left(\tau_j; \frac{v_i + p + 2}{2}, \frac{v_i + \delta_{ij}}{2}\right) d\tau_j \\
 &= \left(\frac{v_i + p}{v_i + \delta_{ij}}\right) \frac{T\left(M_{ij} \sqrt{\frac{v_i+p+2}{v_i+\delta_{ij}}}; v_i + p + 2\right)}{T\left(M_{ij} \sqrt{\frac{v_i+p}{v_i+\delta_{ij}}}; v_i + p\right)}.
 \end{aligned} \tag{B.1}$$

(b) Because  $\gamma_j \mid (\mathbf{y}_j, \tau_j, Z_{ij} = 1) \sim TN(h_i, \tau_j^{-1} \sigma_i^2; (0, \infty))$ , we obtain

$$E(\gamma_j \mid \mathbf{y}_j, \tau_j, z_{ij} = 1) = h_{ij} + \frac{\sigma_i}{\sqrt{\tau_j}} \frac{\phi(\sqrt{\tau_j} M_{ij})}{\Phi(\sqrt{\tau_j} M_{ij})}. \tag{B.2}$$

(c) We first need to show

$$\begin{aligned}
 &E\left(\tau_j^{\frac{k}{2}} \frac{\phi(\sqrt{\tau_j} M_{ij})}{\Phi(\sqrt{\tau_j} M_{ij})} \mid \mathbf{y}_j, z_{ij} = 1\right) \\
 &= \frac{1}{T\left(M_{ij} \sqrt{\frac{v_i+p}{v_i+\delta_{ij}}}; v_i + p\right)} \int_0^\infty \tau_j^{\frac{k}{2}} \phi(\sqrt{\tau_j} M_{ij}) g\left(\tau_j; \frac{v_i + p}{2}, \frac{v_i + \delta_{ij}}{2}\right) d\tau_j \\
 &= \frac{1}{T\left(M_{ij} \sqrt{\frac{v_i+p}{v_i+\delta_{ij}}}; v_i + p\right)} \int_0^\infty \tau_j^{\frac{k-1}{2}} \phi(M_{ij}; 0, \tau_j^{-1}) g\left(\tau_j; \frac{v_i + p}{2}, \frac{v_i + \delta_{ij}}{2}\right) d\tau_j
 \end{aligned}$$

$$\begin{aligned}
 &= \frac{\Gamma\left(\frac{v_i+p+k-1}{2}\right) \int_0^\infty \phi\left(M_{ij}; 0, \tau_j^{-1}\right) g\left(\tau_j; \frac{v_i+p+k-1}{2}, \frac{v_i+\delta_{ij}}{2}\right) d\tau_j}{\Gamma\left(\frac{v_i+p}{2}\right) \left(\frac{v_i+\delta_{ij}}{2}\right)^{\frac{k-1}{2}} T\left(M_{ij} \sqrt{\frac{v_i+p}{v_i+\delta_{ij}}}; v_i+p\right)} \\
 &= \frac{\Gamma\left(\frac{v_i+p+k-1}{2}\right) \sqrt{\frac{v_i+p+k-1}{v_i+\delta_{ij}}} t\left(M_{ij} \sqrt{\frac{v_i+p+k-1}{v_i+\delta_{ij}}}; v_i+p+k-1\right)}{\Gamma\left(\frac{v_i+p}{2}\right) \left(\frac{v_i+\delta_{ij}}{2}\right)^{\frac{k-1}{2}} T\left(M_{ij} \sqrt{\frac{v_i+p}{v_i+\delta_{ij}}}; v_i+p\right)}. \tag{B.3}
 \end{aligned}$$

Applying the result in (B.3) with  $k = -1$  and (B.2) yields

$$\begin{aligned}
 E(\gamma_j | \mathbf{y}_j, z_{ij} = 1) &= E[E(\gamma_j | \mathbf{y}_j, \tau_j, z_{ij} = 1) | \mathbf{y}_j, z_{ij} = 1] \\
 &= E\left[h_{ij} + \frac{\sigma_i}{\sqrt{\tau_j}} \frac{\phi(\sqrt{\tau_j} M_{ij})}{\Phi(\sqrt{\tau_j} M_{ij})} \middle| \mathbf{y}_j, z_{ij} = 1\right] \\
 &= h_{ij} + \sigma_i E\left(\frac{1}{\sqrt{\tau_j}} \frac{\phi(\sqrt{\tau_j} M_{ij})}{\Phi(\sqrt{\tau_j} M_{ij})} \middle| \mathbf{y}_j, z_{ij} = 1\right) \\
 &= h_{ij} + \frac{\sigma_i}{\sqrt{\frac{v_i+p-2}{v_i+\delta_{ij}}}} \frac{t\left(M_{ij} \sqrt{\frac{v_i+p-2}{v_i+\delta_{ij}}}; v_i+p-2\right)}{T\left(M_{ij} \sqrt{\frac{v_i+p}{v_i+\delta_{ij}}}; v_i+p\right)}. \tag{B.4}
 \end{aligned}$$

(d) Using (B.1), (B.2) and (B.3) with  $k = 1$ , we have

$$\begin{aligned}
 E(\tau_j \gamma_j | \mathbf{y}_j, z_{ij} = 1) &= E[E(\tau_j \gamma_j | \mathbf{y}_j, \tau_j, z_{ij} = 1) | \mathbf{y}_j, z_{ij} = 1] \\
 &= E[\tau_j E(\gamma_j | \mathbf{y}_j, \tau_j, z_{ij} = 1) | \mathbf{y}_j, z_{ij} = 1] \\
 &= E\left[\tau_j \left(h_{ij} + \frac{\sigma_i}{\sqrt{\tau_j}} \frac{\phi(\sqrt{\tau_j} M_{ij})}{\Phi(\sqrt{\tau_j} M_{ij})}\right) \middle| \mathbf{y}_j, z_{ij} = 1\right] \\
 &= h_{ij} E(\tau_j | \mathbf{y}_j, z_{ij} = 1) + \sigma_i E\left[\sqrt{\tau_j} \frac{\phi(\sqrt{\tau_j} M_{ij})}{\Phi(\sqrt{\tau_j} M_{ij})} \middle| \mathbf{y}_j, z_{ij} = 1\right] \\
 &= h_{ij} \left[\frac{v_i+p}{v_i+\delta_{ij}} \frac{T\left(M_{ij} \sqrt{\frac{v_i+p+2}{v_i+\delta_{ij}}}; v_i+p+2\right)}{T\left(M_{ij} \sqrt{\frac{v_i+p}{v_i+\delta_{ij}}}; v_i+p\right)}\right] \\
 &\quad + \sigma_i \left[\frac{\sqrt{v_i+p}}{\sqrt{v_i+\delta_{ij}}} \frac{t\left(M_{ij} \sqrt{\frac{v_i+p}{v_i+\delta_{ij}}}; v_i+p\right)}{T\left(M_{ij} \sqrt{\frac{v_i+p}{v_i+\delta_{ij}}}; v_i+p\right)}\right]. \tag{B.5}
 \end{aligned}$$

(e) Using the result of (B.2), the second moment of a truncated normal distribution is given by



$$\begin{aligned}
 E\left(\gamma_j^2 | \mathbf{y}_j, \tau_j, z_{ij} = 1\right) &= h_{ij} E(\gamma_j | \mathbf{y}_j, \tau_j, z_{ij} = 1) + \frac{\sigma_i^2}{\tau_j} \\
 &= h_{ij} \left( h_{ij} + \frac{\sigma_i}{\sqrt{\tau_j}} \frac{\phi(\sqrt{\tau_j} M_{ij})}{\Phi(\sqrt{\tau_j} M_{ij})} \right) + \frac{\sigma_i^2}{\tau_j}. \tag{B.6}
 \end{aligned}$$

(f) Applying the double expectation and using (B.5) and (B.6), we have

$$\begin{aligned}
 E\left(\tau_j \gamma_j^2 | \mathbf{y}_j, z_{ij} = 1\right) &= E\left[E\left(\tau_j \gamma_j^2 | \mathbf{y}_j, \tau_j, z_{ij} = 1\right) | \mathbf{y}_j, z_{ij} = 1\right] \\
 &= E\left[\tau_j E\left(\gamma_j^2 | \mathbf{y}_j, \tau_j, z_{ij} = 1\right) | \mathbf{y}_j, z_{ij} = 1\right] \\
 &= E\left\{\tau_j \left[h_{ij} E(\gamma_j | \mathbf{y}_j, \tau_j, z_{ij} = 1) + \tau_j^{-1} \sigma_i^2\right] | \mathbf{y}_j, z_{ij} = 1\right\} \\
 &= h_{ij} E\left(\tau_j \gamma_j | \mathbf{y}_j, z_{ij} = 1\right) + \sigma_i^2. \tag{B.7}
 \end{aligned}$$

(g) Applying the double expectation and the result of (B.4), we have

$$\begin{aligned}
 E(\mathbf{U}_{ij} | \mathbf{y}_j, z_{ij} = 1) &= E\left[E(\mathbf{U}_{ij} | \mathbf{y}_j, \gamma_j, \tau_j, z_{ij} = 1) | \mathbf{y}_j, z_{ij} = 1\right] \\
 &= E\left[\xi_i + \lambda_i \gamma_j + \beta_i^\top (\mathbf{y}_j - \boldsymbol{\mu}_i - \boldsymbol{\alpha}_i \gamma_j) | \mathbf{y}_j, z_{ij} = 1\right] \\
 &= \xi_i + \beta_i^\top (\mathbf{y}_j - \boldsymbol{\mu}_i) + (\lambda_i - \beta_i^\top \boldsymbol{\alpha}_i) E(\gamma_j | \mathbf{y}_j, z_{ij} = 1). \tag{B.8}
 \end{aligned}$$

(h) Applying the double expectation and using (B.1) and (B.5), we have

$$\begin{aligned}
 E(\tau_j \mathbf{U}_{ij} | \mathbf{y}_j, z_{ij} = 1) &= E\left[E(\tau_j \mathbf{U}_{ij} | \mathbf{y}_j, \gamma_j, \tau_j, z_{ij} = 1) | \mathbf{y}_j, z_{ij} = 1\right] \\
 &= E\left[\tau_j E(\mathbf{U}_{ij} | \mathbf{y}_j, \gamma_j, \tau_j, z_{ij} = 1) | \mathbf{y}_j, z_{ij} = 1\right] \\
 &= E\left\{\tau_j \left[\xi_i + \lambda_i \gamma_j + \beta_i^\top (\mathbf{y}_j - \boldsymbol{\mu}_i - \boldsymbol{\alpha}_i \gamma_j)\right] | \mathbf{y}_j, z_{ij} = 1\right\} \\
 &= \left[\xi_i + \beta_i^\top (\mathbf{y}_j - \boldsymbol{\mu}_i)\right] E(\tau_j | \mathbf{y}_j, z_{ij} = 1) \\
 &\quad + (\lambda_i - \beta_i^\top \boldsymbol{\alpha}_i) E(\tau_j \gamma_j | \mathbf{y}_j, z_{ij} = 1). \tag{B.9}
 \end{aligned}$$

(i) Applying the double expectation and using (B.5) and (B.7), we have

$$\begin{aligned}
 E\left(\tau_j \gamma_j \mathbf{U}_{ij} | \mathbf{y}_j, z_{ij} = 1\right) &= E\left[E\left(\tau_j \gamma_j \mathbf{U}_{ij} | \mathbf{y}_j, \gamma_j, \tau_j, z_{ij} = 1\right) | \mathbf{y}_j, z_{ij} = 1\right] \\
 &= E\left[\tau_j \gamma_j E(\mathbf{U}_{ij} | \mathbf{y}_j, \gamma_j, \tau_j, z_{ij} = 1) | \mathbf{y}_j, z_{ij} = 1\right] \\
 &= E\left\{\tau_j \gamma_j \left[\xi_i + \lambda_i \gamma_j + \beta_i^\top (\mathbf{y}_j - \boldsymbol{\mu}_i - \boldsymbol{\alpha}_i \gamma_j)\right] | \mathbf{y}_j, z_{ij} = 1\right\} \\
 &= \left[\xi_i + \beta_i^\top (\mathbf{y}_j - \boldsymbol{\mu}_i)\right] E(\tau_j \gamma_j | \mathbf{y}_j, z_{ij} = 1) \\
 &\quad + (\lambda_i - \beta_i^\top \boldsymbol{\alpha}_i) E(\tau_j \gamma_j^2 | \mathbf{y}_j, z_{ij} = 1).
 \end{aligned}$$

(j) Applying the double expectation and using (B.8) and (B.9), we have

$$\begin{aligned}
 E\left(\tau_j \mathbf{U}_{ij} \mathbf{U}_{ij}^\top | \mathbf{y}_j, z_{ij} = 1\right) &= E\left[E\left(\tau_j \mathbf{U}_{ij} \mathbf{U}_{ij}^\top | \mathbf{y}_j, \gamma_j, \tau_j, z_{ij} = 1\right) | \mathbf{y}_j, z_{ij} = 1\right] \\
 &= E\left[\tau_j E\left(\mathbf{U}_{ij} \mathbf{U}_{ij}^\top | \mathbf{y}_j, \gamma_j, \tau_j, z_{ij} = 1\right) | \mathbf{y}_j, z_{ij} = 1\right] \\
 &= E\left\{\tau_j [E\left(\mathbf{U}_{ij} | \mathbf{y}_j, \gamma_j, \tau_j, z_{ij} = 1\right) E\left(\mathbf{U}_{ij}^\top | \mathbf{y}_j, \gamma_j, \tau_j, z_{ij} = 1\right) \right. \\
 &\quad \left. + \text{cov}\left(\mathbf{U}_{ij} | \mathbf{y}_j, \gamma_j, \tau_j, z_{ij} = 1\right)] | \mathbf{y}_j, z_{ij} = 1\right\} \\
 &= E\left\{\tau_j [E\left(\mathbf{U}_{ij} | \mathbf{y}_j, \gamma_j, \tau_j, z_{ij} = 1\right) (\boldsymbol{\xi}_i + \boldsymbol{\lambda}_i \gamma_j + \boldsymbol{\beta}_i^\top (\mathbf{y}_j - \boldsymbol{\mu}_i - \boldsymbol{\alpha}_i \gamma_j))^\top \right. \\
 &\quad \left. + \tau_j^{-1} (\mathbf{I}_q - \boldsymbol{\beta}_i^\top \mathbf{A}) \boldsymbol{\Omega}_i] | \mathbf{y}_j, z_{ij} = 1\right\} \\
 &= E\left(\gamma_j \tau_j \mathbf{U}_{ij} | \mathbf{y}_j, z_{ij} = 1\right) (\boldsymbol{\lambda}_i - \boldsymbol{\beta}_i^\top \boldsymbol{\alpha}_i)^\top \\
 &\quad + E\left(\tau_j \mathbf{U}_{ij} | \mathbf{y}_j, z_{ij} = 1\right) [\boldsymbol{\xi}_i + \boldsymbol{\beta}_i^\top (\mathbf{y}_j - \boldsymbol{\mu}_i)]^\top + (\mathbf{I}_q - \boldsymbol{\beta}_i^\top \mathbf{A}) \boldsymbol{\Omega}_i.
 \end{aligned}$$

(k) It is known that  $\int_0^\infty f(\tau_j | \mathbf{y}_j, Z_{ij} = 1) d\tau_j = 1$ , that is,

$$\int_0^\infty \frac{\Phi(\sqrt{\tau_j} M_{ij})}{T\left(M_{ij} \sqrt{\frac{v_i+p}{v_i+\delta_{ij}}}; v_i+p\right)} \frac{\left(\frac{v_i+\delta_{ij}}{2}\right)^{\left(\frac{v_i+p}{2}\right)}}{\Gamma\left(\frac{v_i+p}{2}\right)} \exp\left\{-\frac{v_i+\delta_{ij}}{2} \tau_j\right\} d\tau_j = 1.$$

Then

$$\frac{d}{dv_i} \int_0^\infty b_j \Phi(\sqrt{\tau_j} M_{ij}) \exp\left\{-\frac{v_i+\delta_{ij}}{2} \tau_j\right\} d\tau_j = 0,$$

where

$$b_j = \frac{\left(\frac{v_i+\delta_{ij}}{2}\right)^{(v_i+p)/2}}{\Gamma\left(\frac{v_i+p}{2}\right) T\left(M_{ij} \sqrt{\frac{v_i+p}{v_i+\delta_{ij}}}; v_i+p\right)}.$$

By Leibnitz’s rule, we can obtain

$$\begin{aligned}
 E(\log \tau_j | \mathbf{y}_j, z_{ij} = 1) - E(\tau_j | \mathbf{y}_j, z_{ij} = 1) + \log\left(\frac{v_i+\delta_{ij}}{2}\right) + \left(\frac{v_i+p}{v_i+\delta_{ij}}\right) \\
 - \text{DG}\left(\frac{v_i+p}{2}\right) - \frac{\int_{-\infty}^{M_{ij}} t\left(x; 0, \frac{v_i+\delta_{ij}}{v_i+p}, v_i+p\right) f_{v_i}(x) dx}{T\left(M_{ij} \sqrt{\frac{v_i+p}{v_i+\delta_{ij}}}; v_i+p\right)} = 0,
 \end{aligned}$$

where

$$\begin{aligned}
 f_{v_i}(x) &= \text{DG}\left(\frac{v_i+p+1}{2}\right) - \text{DG}\left(\frac{v_i+p}{2}\right) - \frac{1}{\pi(v_i+\delta_{ij})} \\
 &\quad - \log\left(1 + \frac{x^2}{v_i+\delta_{ij}}\right) + \frac{(v_i+p+1)x^2}{(v_i+\delta_{ij})(x^2+v_i+\delta_{ij})}. \tag{B.10}
 \end{aligned}$$

It follows that

$$E(\log \tau_j | \mathbf{y}_j, z_{ij} = 1) = E(\tau_j | \mathbf{y}_j, z_{ij} = 1) - \log \left( \frac{v_i + \delta_{ij}}{2} \right) - \left( \frac{v_i + p}{v_i + \delta_{ij}} \right) \\ + \text{DG} \left( \frac{v_i + p}{2} \right) + \frac{\int_{-\infty}^{M_{ij}} t \left( x; 0, \frac{v_i + \delta_{ij}}{v_i + p}, v_i + p \right) f_{v_i}(x) dx}{T \left( M_{ij} \sqrt{\frac{v_i + p}{v_i + \delta_{ij}}}; v_i + p \right)}.$$

### Appendix C: Proof of CM-steps

(a) By the Lagrange multiplier method, we define

$$L(\pi_i, \lambda) = Q(\Theta | \hat{\Theta}^{(k)}) - \lambda \left( \sum_{i=1}^g \pi_i - 1 \right),$$

and then take partial derivatives, yielding

$$\frac{\partial L(\pi_i, \lambda)}{\partial \pi_i} = \sum_{j=1}^n \hat{z}_{ij}^{(k)} \frac{1}{\pi_i} - \lambda = 0, \quad \text{and} \quad \frac{\partial L(\pi_i, \lambda)}{\partial \lambda} = - \left( \sum_{i=1}^g \pi_i - 1 \right) = 0.$$

Since  $\sum_{i=1}^g \sum_{j=1}^n \hat{z}_{ij}^{(k)} = n$ , we obtain  $\hat{\pi}_i^{(k+1)} = \sum_{j=1}^n \hat{z}_{ij}^{(k)} / n$ .

(b) Differentiating  $Q(\Theta | \hat{\Theta}^{(k)})$  with respect to  $\xi_i$  leads to

$$\frac{\partial Q}{\partial \xi_i} = -\frac{1}{2} \frac{\partial}{\partial \xi_i} \sum_{j=1}^n \hat{z}_{ij}^{(k)} \Omega_i^{-1} \text{tr} \left[ -\hat{\eta}_{ij}^{(k)} \xi_i^\top - \xi_i \hat{\eta}_{ij}^{(k)\top} \right. \\ \left. + \xi_i \hat{\tau}_{ij}^{(k)} \xi_i^\top + \xi_i \hat{s}_{1ij}^{(k)} \lambda_i^\top + \lambda_i \hat{s}_{1ij}^{(k)} \xi_i^\top \right] \\ = \text{tr} \left\{ \Omega_i^{-1} \sum_{j=1}^n \hat{z}_{ij}^{(k)} \left[ \hat{\eta}_{ij}^{(k)} - \hat{\tau}_{ij}^{(k)} \xi_i - \hat{s}_{1ij}^{(k)} \lambda_i \right] \right\}.$$

Moreover, the partial derivative of  $Q(\Theta | \hat{\Theta}^{(k)})$  with respect to  $\lambda_i$  is

$$\frac{\partial Q}{\partial \lambda_i} = -\frac{1}{2} \frac{\partial}{\partial \lambda_i} \sum_{j=1}^n \hat{z}_{ij}^{(k)} \Omega_i^{-1} \text{tr} \left[ -\hat{\xi}_{ij}^{(k)} \lambda_i^\top + \xi_i \hat{s}_{1ij}^{(k)} \lambda_i^\top \right. \\ \left. - \lambda_i \hat{\xi}_{ij}^{(k)\top} + \lambda_i \hat{s}_{1ij}^{(k)} \xi_i^\top + \lambda_i \hat{s}_{2ij}^{(k)} \lambda_i^\top \right] \\ = \text{tr} \left\{ \Omega_i^{-1} \sum_{j=1}^n \hat{z}_{ij}^{(k)} \left[ \hat{\xi}_{ij}^{(k)} - \hat{s}_{1ij}^{(k)} \xi_i - \hat{s}_{2ij}^{(k)} \lambda_i \right] \right\}.$$

Solving the above two equations, we get

$$\frac{\partial Q}{\partial \xi_i} = \sum_{j=1}^n \hat{z}_{ij}^{(k)} \Omega_i^{-1} (\hat{\eta}_{ij}^{(k)} - \hat{s}_{1ij}^{(k)} \lambda_i) - \sum_{j=1}^n \hat{z}_{ij}^{(k)} \Omega_i^{-1} \hat{\tau}_{ij}^{(k)} \xi_i = \mathbf{0}, \tag{C.1}$$

$$\frac{\partial Q}{\partial \lambda_i} = \sum_{j=1}^n \hat{z}_{ij}^{(k)} \Omega_i^{-1} (\hat{\xi}_{ij}^{(k)} - \hat{s}_{1ij}^{(k)} \xi_i) - \sum_{j=1}^n \hat{z}_{ij}^{(k)} \Omega_i^{-1} \hat{s}_{2ij}^{(k)} \lambda_i = \mathbf{0}. \tag{C.2}$$

After rearrangement, (C.1) and (C.2) can be rewritten as

$$\begin{aligned} \sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{\tau}_{ij}^{(k)} \xi_i + \sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{s}_{1ij}^{(k)} \lambda_i &= \sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{\eta}_{ij}^{(k)}, \\ \sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{s}_{1ij}^{(k)} \xi_i + \sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{s}_{2ij}^{(k)} \lambda_i &= \sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{\xi}_{ij}^{(k)}. \end{aligned}$$

Using Cramer’s law, the solutions of the two linear equations are

$$\hat{\xi}_i^{(k+1)} = \frac{\left(\sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{\eta}_{ij}^{(k)}\right) \left(\sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{s}_{2ij}^{(k)}\right) - \left(\sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{\xi}_{ij}^{(k)}\right) \left(\sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{s}_{1ij}^{(k)}\right)}{\left(\sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{\tau}_{ij}^{(k)}\right) \left(\sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{s}_{2ij}^{(k)}\right) - \left(\sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{s}_{1ij}^{(k)}\right)^2},$$

and

$$\hat{\lambda}_i^{(k+1)} = \frac{\left(\sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{\tau}_{ij}^{(k)}\right) \left(\sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{\xi}_{ij}^{(k)}\right) - \left(\sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{s}_{1ij}^{(k)}\right) \left(\sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{\eta}_{ij}^{(k)}\right)}{\left(\sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{\tau}_{ij}^{(k)}\right) \left(\sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{s}_{2ij}^{(k)}\right) - \left(\sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{s}_{1ij}^{(k)}\right)^2}.$$

(c) The partial derivative of  $Q(\Theta \mid \hat{\Theta}^{(k)})$  with respect to  $A$  is

$$\begin{aligned} \frac{\partial Q}{\partial A} &= -\frac{1}{2} \frac{\partial}{\partial A} \sum_{i=1}^g \sum_{j=1}^n \hat{z}_{ij}^{(k)} \text{tr} \left( -D_i^{-1} y_j \hat{\eta}_{ij}^{(k)\top} A^\top \right. \\ &\quad \left. - D_i^{-1} A \hat{\eta}_{ij}^{(k)} y_j^\top + D_i^{-1} A \hat{\psi}_{ij}^{(k)} A^\top \right) \\ &= \text{tr} \left( \sum_{i=1}^g \sum_{j=1}^n \hat{z}_{ij}^{(k)} D_i^{-1} y_j \hat{\eta}_{ij}^{(k)\top} - \sum_{i=1}^g \sum_{j=1}^n \hat{z}_{ij}^{(k)} D_i^{-1} \hat{\psi}_{ij}^{(k)} A \right). \tag{C.3} \end{aligned}$$

Equating (C.3) to the zero matrix, we have

$$\hat{A}^{(k+1)} = \left( \sum_{i=1}^g \sum_{j=1}^n \hat{z}_{ij}^{(k)} y_j \hat{\eta}_{ij}^{(k)\top} \right) \left( \sum_{i=1}^g \sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{\psi}_{ij}^{(k)} \right)^{-1}.$$

(d) The partial derivative of  $Q(\Theta \mid \hat{\Theta}^{(k)})$  with respect to  $\Omega_i$  is

$$\begin{aligned} \frac{\partial Q}{\partial \Omega_i^{-1}} &= \frac{1}{2} \frac{\partial}{\partial \Omega_i^{-1}} \sum_{j=1}^n \hat{z}_{ij}^{(k)} \left\{ \log |\Omega_i^{-1}| - \text{tr} \left( \Omega_i^{-1} \mathbf{A}_{ij} \right) \right\} \\ &= \frac{1}{2} \sum_{j=1}^n \hat{z}_{ij}^{(k)} \left[ 2\Omega_i - \text{Diag}\{\Omega_i\} - (2\mathbf{A}_{ij} - \text{Diag}\{\mathbf{A}_{ij}\}) \right]. \end{aligned} \tag{C.4}$$

Equating (C.4) to the zero vector gives

$$\hat{\Omega}_i^{(k+1)} = \frac{\sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{\mathbf{A}}_{ij}^{(k+1)}}{\sum_{j=1}^n \hat{z}_{ij}^{(k)}}.$$

(e) Taking the partial derivative of  $Q(\Theta \mid \hat{\Theta}^{(k)})$  with respect to  $\mathbf{D}_i$  yields

$$\begin{aligned} \frac{\partial Q}{\partial \mathbf{D}_i^{-1}} &= \frac{1}{2} \frac{\partial}{\partial \mathbf{D}_i^{-1}} \sum_{j=1}^n \hat{z}_{ij}^{(k)} \left[ \log |\mathbf{D}_i^{-1}| - \text{tr} \left( \mathbf{D}_i^{-1} \mathbf{r}_{ij} \right) \right] \\ &= \frac{1}{2} \sum_{j=1}^n \hat{z}_{ij}^{(k)} (\mathbf{D}_i - \mathbf{r}_{ij}). \end{aligned} \tag{C.5}$$

We have the following estimator

$$\hat{\mathbf{D}}_i^{(k+1)} = \frac{\text{Diag} \left\{ \sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{\mathbf{r}}_{ij}^{(k+1)} \right\}}{\sum_{j=1}^n \hat{z}_{ij}^{(k)}}$$

obtained by equating (C.5) to the zero matrix.

### Appendix D: Parameter estimation for the MCghstFA model using the ECM algorithm

According to Table 5, the MCghstFA model admits a three-level hierarchy:

$$\begin{aligned} \left[ \begin{matrix} \mathbf{Y}_j \\ \mathbf{U}_{ij} \end{matrix} \right] \mid (W_j, Z_{ij} = 1) &\sim N_{p+q} \left( \left[ \begin{matrix} \mathbf{A}\xi_i + W_j \mathbf{A}\lambda_i \\ \xi_i + W_j \lambda_i \end{matrix} \right], W_j \left[ \begin{matrix} \mathbf{A}\Omega_i \mathbf{A} + \mathbf{D} & \mathbf{A}\Omega_i \\ \Omega_i \mathbf{A}^\top & \Omega_i \end{matrix} \right] \right), \\ W_j \mid Z_{ij} = 1 &\sim \Gamma^{-1} \left( \frac{v_i}{2}, \frac{v_i}{2} \right), \\ \mathbf{Z}_j &\sim \mathcal{M}(1; \pi_1, \dots, \pi_g). \end{aligned} \tag{D.1}$$

From (D.1), it can be verified that

$$\mathbf{Y}_j \mid (W_j, Z_{ij} = 1) \sim N_p(\mathbf{A}\xi_i + W_j \mathbf{A}\lambda_i, W_j(\mathbf{A}\Omega_i \mathbf{A} + \mathbf{D})),$$

**Table 5** Comparison of some characterizations between the MCstFA and MCghstFA models

Model	MCstFA	MCghstFA
Formulation	$Y_j = AU_{ij} + e_{ij} \text{ with probability } \pi_i$ $\begin{bmatrix} U_{ij} \\ e_{ij} \end{bmatrix} \sim rST_{p+q} \left( \begin{bmatrix} \xi_j \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \lambda_j \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \lambda_j \\ \mathbf{0} \end{bmatrix}, v_i \right).$	$Y_j = AU_{ij} + e_{ij} \text{ with probability } \pi_i$ $U_{ij} \mid W_j \sim N_q(\xi_i + W_j \lambda_i, W_j \Omega_i), e_{ij} \mid W_j \sim N_p(\mathbf{0}, W_j D),$ $W_j^{-1} \sim \text{Gamma}(v_i/2, v_i/2), U_{ij} \mid W_j \perp e_{ij} \mid W_j.$
Hierarchical representation	$Y_j \mid (U_{ij}, \gamma_j, \tau_j, Z_{ij} = 1) \sim N_p(AU_{ij}, \tau_j^{-1} D),$ $U_{ij} \mid (\gamma_j, \tau_j, Z_{ij} = 1) \sim N_q(\xi_i + \gamma_j \lambda_i, \tau_j^{-1} \Omega_i)$ $\gamma_j \mid (\tau_j, Z_{ij} = 1) \sim TN(0, \tau_j^{-1}; (0, \infty))$ $\tau_j \mid (Z_{ij} = 1) \sim \text{Gamma}(\frac{p}{2}, \frac{p}{2})$ $Z_j \sim \mathcal{M}(1; \pi_1, \dots, \pi_g)$	$Y_j \mid (U_{ij}, W_j, Z_{ij} = 1) \sim N_p(AU_{ij}, W_j D),$ $U_{ij} \mid (W_j, Z_{ij} = 1) \sim N_q(\xi_i + W_j \lambda_i, W_j \Omega_i)$ $W_j^{-1} \mid (Z_{ij} = 1) \sim \text{Gamma}(\frac{p}{2}, \frac{p}{2})$ $Z_j \sim \mathcal{M}(1; \pi_1, \dots, \pi_g)$
Marginal density	$f(y_j) = \sum_{i=1}^g \pi_i \psi_p(y_j; A\xi_i, A\Omega_i A^T + D, A\lambda_i, v_i)$ $\psi_p: \text{restricted MST density function}$	$f(y_j) = \sum_{i=1}^g \pi_i \xi_p(y_j; A\xi_i, A\Omega_i A^T + D, A\lambda_i, v_i)$ $\xi_p: \text{generalized hyperbolic MST density function}$
Free parameters	$(g-1) + p + q(p+g) + \frac{1}{2} gq(q+1) - q^2 + gq + g$	The same

and

$$U_{ij} \mid (y_j, W_j, Z_{ij} = 1) \sim N_q(\boldsymbol{\mu}_{2,1}, \boldsymbol{\Sigma}_{22,1}), \tag{D.2}$$

where  $\boldsymbol{\mu}_{2,1} = \boldsymbol{\xi}_i + W_j \boldsymbol{\lambda}_i + \boldsymbol{\Omega}_i \mathbf{A}^\top (\mathbf{A} \boldsymbol{\Omega}_i \mathbf{A}^\top + \mathbf{D})^{-1} (y_j - \mathbf{A} \boldsymbol{\xi}_i - W_j \mathbf{A} \boldsymbol{\lambda}_i)$  and  $\boldsymbol{\Sigma}_{22,1} = W_j (\boldsymbol{\Omega}_i - \boldsymbol{\Omega}_i \mathbf{A}^\top (\mathbf{A} \boldsymbol{\Omega}_i \mathbf{A}^\top + \mathbf{D})^{-1} \mathbf{A} \boldsymbol{\Omega}_i) = W_j (\boldsymbol{\Omega}_i^{-1} - \mathbf{A}^\top \mathbf{D}^{-1} \mathbf{A})^{-1}$ .

A positive random variable  $X$  is said to follow the Generalized Inverse Gaussian (GIG) distribution (Good 1953), denoted by  $W \sim \text{GIG}(\psi, \chi, r)$ , if it has the pdf

$$f_{\text{GIG}}(w; \psi, \chi, r) = \frac{\chi^{-r} (\sqrt{\chi \psi})^r}{2 K_r(\sqrt{\chi \psi})} w^{q-1} \exp \left\{ -\frac{1}{2} (\chi w^{-1} + \psi w) \right\}, \tag{D.3}$$

where  $\psi, \chi \in \mathbb{R}^+$ ,  $r \in \mathbb{R}$ , and  $K_q$  is the modified Bessel function of the third kind with index  $r$ . Some particular moments of the GIG distribution have tractable forms, for instance,

$$E(W^a) = (\chi/\psi)^{a/2} \frac{K_{r+a}(\sqrt{\psi \chi})}{K_r(\sqrt{\psi \chi})}, \quad a \in \mathbb{R}, \tag{D.4}$$

and

$$E(\log W) = \log(\chi/\psi)^{1/2} + \frac{K'_r(\sqrt{\psi \chi})}{K_r(\sqrt{\psi \chi})}, \tag{D.5}$$

where

$$K'_r(x) = \frac{dK_r(x)}{dx} = \frac{1}{2} \int_0^\infty \log(y) y^{r-1} \exp \left\{ -\frac{x}{2} \left( y + \frac{1}{y} \right) \right\} dy, \quad x > 0. \tag{D.6}$$

By Bayes' Theorem, the conditional pdf of  $W_j$  given  $y_j$  can be written as

$$f(w_j \mid y_j, Z_{ij} = 1) \propto w_j^{-(v_i+p)/2-1} \exp \left\{ -\frac{1}{2} \left[ (v_i + \Delta_{ij}) w_j^{-1} + (\boldsymbol{\lambda}_i^\top \mathbf{A}^\top (\mathbf{A} \boldsymbol{\Omega}_i \mathbf{A}^\top + \mathbf{D})^{-1} \mathbf{A} \boldsymbol{\lambda}_i) w_j \right] \right\},$$

where  $\Delta_{ij} = (y_j - \mathbf{A} \boldsymbol{\xi}_i)^\top (\mathbf{A} \boldsymbol{\Omega}_i \mathbf{A}^\top + \mathbf{D})^{-1} (y_j - \mathbf{A} \boldsymbol{\xi}_i)$ . It follows from (D.3) that

$$W_j \mid (y_j, Z_{ij} = 1) \sim \text{GIG} \left( \boldsymbol{\lambda}_i^\top \mathbf{A}^\top (\mathbf{A} \boldsymbol{\Omega}_i \mathbf{A}^\top + \mathbf{D})^{-1} \mathbf{A} \boldsymbol{\lambda}_i, v_i + \Delta_{ij}, -\frac{v_i + p}{2} \right).$$

Alternatively, the MCghstFA model can be represented by a four-level hierarchy:

$$\begin{aligned} Y_j \mid (U_{ij}, W_j, Z_{ij} = 1) &\sim N_p(\mathbf{A} U_{ij}, W_j \mathbf{D}), \\ U_{ij} \mid (W_j, Z_{ij} = 1) &\sim N_q(\boldsymbol{\xi}_i + W_j \boldsymbol{\lambda}_i, W_j \boldsymbol{\Omega}_i), \end{aligned}$$

$$\begin{aligned}
 W_j \mid Z_{ij} = 1 &\sim \Gamma^{-1}\left(\frac{\nu_i}{2}, \frac{\nu_i}{2}\right), \\
 \mathbf{Z}_j &\sim \mathcal{M}(1; \pi_1, \dots, \pi_g).
 \end{aligned}
 \tag{D.7}$$

From (D.7), the complete-data log-likelihood function for  $\Theta$  on the basis of  $Y_c = \{y_j, \mathbf{U}_{ij}, W_j, \mathbf{Z}_j\}_{j=1}^n$ , for  $i = 1, \dots, g$ , is given by

$$\begin{aligned}
 \ell_c(\Theta \mid Y_c) &= \sum_{i=1}^g \sum_{j=1}^n Z_{ij} \log \left\{ \pi_i \log \phi_p(y_j \mid \mathbf{A}\mathbf{U}_{ij}, W_j \mathbf{D}) \phi_p(\mathbf{U}_{ij} \mid \xi_i \right. \\
 &\quad \left. + W_j \lambda_i, W_j \boldsymbol{\Omega}_i) \times f\left(w_j \mid \frac{\nu_i}{2}, \frac{\nu_i}{2}\right) \right\} \\
 &= \sum_{i=1}^g \sum_{j=1}^n Z_{ij} \left\{ \log \pi_i - \frac{1}{2} \log |W_j \mathbf{D}| - \frac{1}{2} \log |W_j \boldsymbol{\Omega}_i| \right. \\
 &\quad \left. - \frac{W_j^{-1}}{2} \left[ (y_j - \mathbf{A}\mathbf{U}_{ij})^\top \mathbf{D}^{-1} (y_j - \mathbf{A}\mathbf{U}_{ij}) \right. \right. \\
 &\quad \left. \left. + (\mathbf{U}_{ij} - \xi_i - W_j \lambda_i)^\top \boldsymbol{\Omega}_i^{-1} (\mathbf{U}_{ij} - \xi_i - W_j \lambda_i) \right] \right. \\
 &\quad \left. + \frac{\nu_i}{2} \log \left(\frac{\nu_i}{2}\right) - \log \Gamma\left(\frac{\nu_i}{2}\right) - \left(\frac{\nu_i}{2} + 1\right) \log W_j - \frac{\nu_i}{2W_j} \right\}.
 \end{aligned}
 \tag{D.8}$$

To evaluate the expected value of (D.8), called the  $Q$  function, we first calculate

$$\hat{z}_{ij} = E\left(Z_{ij} \mid y_j, \hat{\Theta}\right) = \frac{\hat{\pi}_i \zeta\left(y_j; \hat{\mathbf{A}} \hat{\xi}_i, \hat{\mathbf{A}} \hat{\boldsymbol{\Omega}}_i \hat{\mathbf{A}}^\top + \hat{\mathbf{D}}, \hat{\mathbf{A}} \hat{\lambda}_i, \hat{\nu}_i\right)}{f\left(y_j; \hat{\Theta}\right)},
 \tag{D.9}$$

which is the posterior probability of  $y_j$  belonging to the  $i$ th component of the mixture. In addition, we utilize the results (D.4) and (D.5) to calculate of the following conditional expectations:

$$\begin{aligned}
 \hat{s}_{1ij} &= E(W_j \mid y_j, Z_{ij} = 1, \hat{\Theta}) \\
 &= \left( \frac{\hat{\nu}_i + \hat{\Delta}_{ij}}{\hat{\lambda}_i^\top \hat{\mathbf{A}}^\top (\hat{\mathbf{A}} \hat{\boldsymbol{\Omega}}_i \hat{\mathbf{A}}^\top + \hat{\mathbf{D}})^{-1} \hat{\mathbf{A}} \hat{\lambda}_i} \right)^{1/2} \frac{K_{-\frac{(\hat{\nu}+p)}{2}+1}(\hat{\omega}_{ij})}{K_{-\frac{(\hat{\nu}+p)}{2}}(\hat{\omega}_{ij})}, \\
 \hat{s}_{2ij} &= E(W_j^{-1} \mid y_j, Z_{ij} = 1, \hat{\Theta}) \\
 &= \left( \frac{\hat{\nu}_i + \hat{\Delta}_{ij}}{\hat{\lambda}_i^\top \hat{\mathbf{A}}^\top (\hat{\mathbf{A}} \hat{\boldsymbol{\Omega}}_i \hat{\mathbf{A}}^\top + \hat{\mathbf{D}})^{-1} \hat{\mathbf{A}} \hat{\lambda}_i} \right)^{-1/2} \frac{K_{-\frac{(\hat{\nu}+p)}{2}-1}(\hat{\omega}_{ij})}{K_{-\frac{(\hat{\nu}+p)}{2}}(\hat{\omega}_{ij})}, \\
 \hat{s}_{3ij} &= E(\log W_j \mid y_j, Z_{ij} = 1, \hat{\Theta}) \\
 &= \log \left( \frac{\hat{\nu}_i + \hat{\Delta}_{ij}}{\hat{\lambda}_i^\top \hat{\mathbf{A}}^\top (\hat{\mathbf{A}} \hat{\boldsymbol{\Omega}}_i \hat{\mathbf{A}}^\top + \hat{\mathbf{D}})^{-1} \hat{\mathbf{A}} \hat{\lambda}_i} \right)^{1/2} + \frac{K'_{-\frac{(\hat{\nu}+p)}{2}}(\hat{\omega}_{ij})}{K_{-\frac{(\hat{\nu}+p)}{2}}(\hat{\omega}_{ij})},
 \end{aligned}
 \tag{D.10}$$



where  $\hat{\omega}_{ij} = \sqrt{(\hat{\nu}_i + \hat{\Delta}_{ij})\hat{\lambda}_i^\top \hat{A}^\top (\hat{A}\hat{\Omega}_i\hat{A}^\top + \hat{D})^{-1}\hat{A}\hat{\lambda}_i}$  and  $K'_{-\frac{(\hat{\nu}_i+p)}{2}}(\hat{\omega}_{ij})$  is evaluated via (D.6). By (D.2), we obtain

$$\begin{aligned} \hat{u}_{ij} &= E\left(\mathbf{U}_{ij} \mid \mathbf{y}_j, Z_{ij} = 1, \hat{\Theta}\right) = E\left[E(\mathbf{U}_{ij} \mid \mathbf{y}_j, W_j, Z_{ij} = 1) \mid \mathbf{y}_j, Z_{ij} = 1, \hat{\Theta}\right] \\ &= \hat{\xi}_i + \hat{s}_{1ij}\hat{\lambda}_i + \hat{\gamma}_i^\top(\mathbf{y}_j - \hat{A}\hat{\xi}_i - \hat{s}_{1ij}\hat{A}\hat{\lambda}_i), \end{aligned} \tag{D.11}$$

$$\begin{aligned} \hat{\eta}_{ij} &= E\left(W_j^{-1}\mathbf{U}_{ij} \mid \mathbf{y}_j, Z_{ij} = 1, \hat{\Theta}\right) \\ &= E\left[W_j^{-1}E(\mathbf{U}_{ij} \mid \mathbf{y}_j, W_j, Z_{ij} = 1) \mid \mathbf{y}_j, Z_{ij} = 1, \hat{\Theta}\right] \\ &= \hat{\lambda}_i - \hat{\gamma}_i^\top \hat{A}\hat{\lambda}_i + \hat{s}_{2ij}\left(\hat{\xi}_i + \hat{\gamma}_i^\top(\mathbf{y}_j - \hat{A}\hat{\xi}_i)\right), \end{aligned} \tag{D.12}$$

and

$$\begin{aligned} \hat{\Psi}_{ij} &= E\left(W_j^{-1}\mathbf{U}_{ij}\mathbf{U}_{ij}^\top \mid \mathbf{y}_j, Z_{ij} = 1, \hat{\Theta}\right) \\ &= E\left[W_j^{-1}E(\mathbf{U}_{ij}\mathbf{U}_{ij}^\top \mid \mathbf{y}_j, W_j, Z_{ij} = 1) \mid \mathbf{y}_j, Z_{ij} = 1, \hat{\Theta}\right] \\ &= E\left[W_j^{-1}\left[E(\mathbf{U}_{ij} \mid \mathbf{y}_j, W_j, Z_{ij} = 1)E(\mathbf{U}_{ij}^\top \mid \mathbf{y}_j, W_j, Z_{ij} = 1)\right.\right. \\ &\quad \left.\left. + \text{cov}(\mathbf{U}_{ij} \mid \mathbf{y}_j, W_j, Z_{ij} = 1)\right] \mid \mathbf{y}_j, Z_{ij} = 1, \hat{\Theta}\right] \\ &= \hat{\eta}_{ij}\left(\hat{\xi}_i + \hat{\gamma}_i^\top(\mathbf{y}_j - \hat{A}\hat{\xi}_i)\right)^\top + \hat{u}_{ij}\left(\hat{\lambda}_i - \hat{\gamma}_i^\top \hat{A}\hat{\lambda}_i\right)^\top \\ &\quad + \left(\hat{\Omega}_i^{-1} + \hat{A}^\top \hat{D}^{-1} \hat{A}\right)^{-1}, \end{aligned} \tag{D.13}$$

where  $\hat{\gamma}_i = (\hat{A}\hat{\Omega}_i\hat{A}^\top + \hat{D})^{-1}\hat{A}\hat{\Omega}_i$ .

After some algebraic manipulations, the resulting  $Q$  function that gets rid of the constants is given by

$$\begin{aligned} Q(\Theta \mid \hat{\Theta}) &= \sum_{i=1}^g \sum_{j=1}^n \hat{z}_{ij} \left\{ \log \pi_i - \frac{1}{2} \log |\mathbf{D}| - \frac{1}{2} \log |\Omega_i| - \frac{1}{2} \text{tr}\left(\mathbf{D}^{-1}\left[\hat{s}_{2ij}\mathbf{y}_j\mathbf{y}_j^\top\right.\right.\right. \\ &\quad \left.\left.\left. - \mathbf{y}_j\hat{\eta}_{ij}^\top\mathbf{A}^\top - \mathbf{A}\hat{\eta}_{ij}\mathbf{y}_j^\top + \mathbf{A}\hat{\Psi}_{ij}\mathbf{A}^\top\right]\right) \\ &\quad - \frac{1}{2} \text{tr}\left(\Omega_i^{-1}\left[\hat{\Psi}_{ij} - \hat{\eta}_{ij}\hat{\xi}_i^\top - \hat{\xi}_i\hat{\eta}_{ij}^\top\right.\right. \\ &\quad \left.\left. + \hat{s}_{2ij}\hat{\xi}_i\hat{\xi}_i^\top + \hat{s}_{1ij}\hat{\lambda}_i\hat{\lambda}_i^\top - (\hat{u}_{ij} - \hat{\xi}_i)\hat{\lambda}_i^\top - \hat{\lambda}_i(\hat{u}_{ij} - \hat{\xi}_i)^\top\right]\right) \\ &\quad \left. + \left(\frac{\nu_i}{2}\right) \log\left(\frac{\nu_i}{2}\right) - \frac{\nu_i}{2}\hat{s}_{2ij} - \frac{\nu_i}{2}\hat{s}_{3ij} - \log \Gamma\left(\frac{\nu_i}{2}\right)\right\}. \end{aligned} \tag{D.14}$$

Taking partial derivatives of (D.14) with respect to  $\xi_i$  and  $\lambda_i$  and equating them to zero vectors yield

$$\sum_{j=1}^n \hat{z}_{ij} (\hat{\eta}_{ij} - \hat{s}_{2ij}\xi_i - \lambda_i) = \mathbf{0}, \tag{D.15}$$

$$\sum_{j=1}^n \hat{z}_{ij} (\hat{u}_{ij} - \xi_i - \hat{s}_{1ij}\lambda_i) = \mathbf{0}. \tag{D.16}$$

In summary, the ECM algorithm for estimating the parameters of MCghstFA proceeds as follows:

- E-step: Given the current value  $\Theta = \hat{\Theta}$ , compute  $\hat{z}_{ij}, \hat{s}_{1ij}, \hat{s}_{2ij}, \hat{s}_{3ij}, \hat{u}_{ij}, \hat{\eta}_{ij}$  and  $\hat{\Psi}_{ij}$  as defined in (D.9)–(D.13) for  $i = 1, \dots, g$  and  $j = 1, \dots, n$ .
- CM step 1: Maximizing (D.14) with respect to  $\pi_i$  and using the Lagrange multiplier method, this gives  $\hat{\pi}_i = \hat{n}_i/n$ , where  $\hat{n}_i = \sum_{j=1}^n \hat{z}_{ij}$ .
- CM step 2: Update parameters  $\xi_i$  and  $\lambda_i$  by solving simultaneous Eqs. (D.15) and (D.16). Simple matrix algebra yields

$$\hat{\xi}_i = \frac{(\sum_{j=1}^n \hat{z}_{ij}\hat{s}_{1ij})(\sum_{j=1}^n \hat{z}_{ij}\hat{\eta}_{ij}) - \hat{n}_i(\sum_{j=1}^n \hat{z}_{ij}\hat{u}_{ij})}{(\sum_{j=1}^n \hat{z}_{ij}\hat{s}_{1ij})(\sum_{j=1}^n \hat{z}_{ij}\hat{s}_{2ij}) - \hat{n}_i^2}$$

and

$$\hat{\lambda}_i = \frac{(\sum_{j=1}^n \hat{z}_{ij}\hat{s}_{2ij})(\sum_{j=1}^n \hat{z}_{ij}\hat{u}_{ij}) - \hat{n}_i(\sum_{j=1}^n \hat{z}_{ij}\hat{\eta}_{ij})}{(\sum_{j=1}^n \hat{z}_{ij}\hat{s}_{1ij})(\sum_{j=1}^n \hat{z}_{ij}\hat{s}_{2ij}) - \hat{n}_i^2}.$$

CM-step3: The updates for  $\mathbf{A}, \mathbf{\Omega}_i$  and  $\mathbf{D}$  are given by

$$\begin{aligned} \hat{\mathbf{A}} &= \left( \sum_{i=1}^g \sum_{j=1}^n \hat{z}_{ij} \hat{\mathbf{y}}_j \hat{\eta}_{ij}^\top \right) \left( \sum_{i=1}^g \sum_{j=1}^n \hat{z}_{ij} \hat{\Psi}_{ij} \right)^{-1}, \\ \hat{\mathbf{\Omega}}_i &= \frac{1}{\hat{n}_i} \sum_{j=1}^n \hat{z}_{ij} \left[ \hat{\Psi}_{ij} - \hat{\eta}_{ij} \hat{\xi}_i^\top - \hat{\xi}_i \hat{\eta}_{ij}^\top + \hat{s}_{2ij} \hat{\xi}_i \hat{\xi}_i^\top + \hat{s}_{1ij} \hat{\lambda}_i \hat{\lambda}_i^\top \right. \\ &\quad \left. - (\hat{u}_{ij} - \hat{\xi}_i) \hat{\lambda}_i^\top - \hat{\lambda}_i (\hat{u}_{ij} - \hat{\xi}_i)^\top \right], \\ \hat{\mathbf{D}} &= \frac{1}{n} \text{Diag} \left\{ \sum_{i=1}^g \sum_{j=1}^n \hat{z}_{ij} (\hat{s}_{2ij} \mathbf{y}_j \mathbf{y}_j^\top - \mathbf{y}_j \hat{\eta}_{ij}^\top \hat{\mathbf{A}}^\top) \right\}. \end{aligned}$$

CM step 4: Calculate  $\hat{v}_i$  by solving the root of the following equation:

$$\log\left(\frac{v_i}{2}\right) - \text{DG}\left(\frac{v_i}{2}\right) + 1 - \frac{1}{\hat{n}_i} \sum_{j=1}^n \hat{z}_{ij} (\hat{s}_{2ij} + \hat{s}_{3ij}) = 0.$$

## References

- Aitken AC (1926) On Bernoulli's numerical solution of algebraic equations. *Proc R Soc Edinb* 46:289–305
- Arellano-Valle RB, Genton MG (2005) On fundamental skew distributions. *J Multivar Anal* 96:93–116
- Azzalini A (2014) The skew-normal and related families. IMS monographs series. Cambridge University Press, Cambridge
- Azzalini A, Browne RP, Genton MG, McNicholas PD (2016) On nomenclature for, and the relative merits of, two formulations of skew distributions. *Stat Probab Lett* 110:201–206
- Baek J, McLachlan GJ (2011) Mixtures of common  $t$ -factor analyzers for clustering high-dimensional microarray data. *Bioinformatics* 27:1269–1276
- Baek J, McLachlan GJ, Flack LK (2010) Mixtures of factor analyzers with common factor loadings: applications to the clustering and visualization of high-dimensional data. *IEEE Trans Pattern Anal Mach Intell* 32:1–13
- Barndorff-Nielsen O, Shephard N (2001) Non-Gaussian Ornstein–Uhlenbeck-based models and some of their uses in financial economics. *J Roy Stat Soc Ser B* 63:167–241
- Beal MJ (2003) Variational algorithms for approximate Bayesian inference. Ph.D. thesis, The University of London, London, UK
- Biernacki C, Celeux G, Govaert G (2000) Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans Pattern Anal Mach Intell* 22:719–725
- Cabral CR, Lachos VH, Prates MO (2012) Multivariate mixture modeling using skew-normal independent distributions. *Comput Stat Data Anal* 56:126–142
- Castro LM, Costa DR, Prates MO, Lachos VH (2015) Likelihood-based inference for Tobit confirmatory factor analysis using the multivariate Student- $t$  distribution. *Stat Comput* 25:1163–1183
- Chen X, Cheung ST, So S, Fan ST, Barry C, Higgins J, Lai KM, Ji J, Dudoit S, Ng IO, Van De Rijn M, Botstein D, Brown PO (2002) Gene expression patterns in human liver cancers. *Mol Biol Cell* 13:1929–1939
- Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J R Stat Soc B* 9:1–38
- Ghahramani Z, Beal M (2000) Variational inference for Bayesian mixture of factor analysers. In: Solla S, Leen T, Muller K-R (eds) *Advances in neural information processing systems*. MIT Press, Cambridge
- Ghahramani Z, Hinton GE (1997) The EM algorithm for factor analyzers. Technical Report No. CRG-TR-96-1, The University of Toronto, Toronto
- Hartigan JA, Wong MA (1979) Algorithm AS 136: a K-means clustering algorithm. *J R Stat Soc C* 28:100–108
- Hubert LJ, Arabie P (1985) Comparing partitions. *J Classif* 2:193–218
- Jordan MI, Ghahramani Z, Jaakkola TS, Saul LK (1999) An introduction to variational methods for graphical models. *Mach Learn* 37:183–233
- Lachos VH, Morenoa EJJ, Chen K, Cabral CRB (2017) Finite mixture modeling of censored data using the multivariate Student- $t$  distribution. *J Multivar Anal* 159:151–167
- Lee SX, McLachlan GJ (2014) Finite mixtures of multivariate skew  $t$ -distributions: some recent and new results. *Stat Comp* 24:181–202
- Lee SX, McLachlan GJ (2016) Finite mixtures of canonical fundamental skew  $t$ -distributions: the unication of the restricted and unrestricted skew  $t$ -mixture models. *Stat Comp* 26:573–589
- Lee YW, Poon SH (2011) Systemic and systematic factors for loan portfolio loss distribution. *Econometrics and applied economics workshops*, pp 1–61. School of Social Science, University of Manchester
- Lee WL, Chen YC, Hsieh KS (2003) Ultrasonic liver tissues classification by fractal feature vector based on M-band wavelet transform. *IEEE Trans Med Imaging* 22:382–392
- Lin TI (2014) Learning from incomplete data via parameterized  $t$  mixture models through eigenvalue decomposition. *Comput Stat Data Anal* 71:183–195
- Lin TI, Wu PH, McLachlan GJ, Lee SX (2015) A robust factor analysis model using the restricted skew- $t$  distribution. *TEST* 24:510–531
- Lin TI, McLachlan GJ, Lee SX (2016) Extending mixtures of factor models using the restricted multivariate skew-normal distribution. *J Multivar Anal* 143:398–413
- Lin TI, Wang WL, McLachlan GJ, Lee SX (2018) Robust mixtures of factor analysis models using the restricted multivariate skew- $t$  distribution. *Stat Model* 28:50–72
- Liu C, Rubin DB (1994) The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. *Biometrika* 81:33–648

- McLachlan GJ, Basford KE (1988) Mixture models: inference and application to clustering. Marcel Dekker, New York
- McLachlan GJ, Krishnan T (2008) The EM algorithm and extensions, 2nd edn. Wiley, New York
- McLachlan GJ, Peel D (2000) Finite mixture models. Wiley, New York
- McNicholas PD, Murphy TB (2008) Parsimonious Gaussian mixture models. *Stat Comp* 18:285–296
- McNicholas PD, Murphy TB, McDaid AF, Frost D (2010) Serial and parallel implementations of model-based clustering via parsimonious Gaussian mixture models. *Comput Stat Data Anal* 54:711–723
- Meng XL, Rubin DB (1993) Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* 80:267–278
- Murray PM, Browne RP, McNicholas PD (2014a) Mixtures of skew- $t$  factor analyzers. *Comput Stat Data Anal* 77:326–335
- Murray PM, McNicholas PD, Browne RP (2014b) Mixtures of common skew- $t$  factor analyzers. *Stat* 3:68–82
- Murray PM, Browne RP, McNicholas PD (2017a) A mixture of SDB skew- $t$  factor analyzers. *Econom Stat* 3:160–168
- Murray PM, Browne RP, McNicholas PD (2017b) Hidden truncation hyperbolic distributions, finite mixtures thereof, and their application for clustering. *J Multivar Anal* 161:141–156
- Ouyang M, Welsh W, Georgopoulos P (2004) Gaussian mixture clustering and imputation of microarray data. *Bioinformatics* 20:917–923
- Prates MO, Cabral CR, Lachos VH (2013) mixsmsn: fitting finite mixture of scale mixture of skew-normal distributions. *J Stat Soft* 54:1–20
- Pyne S, Hu X, Wang K, Rossin E, Lin TI, Maier LM, Baecher-Allan C, McLachlan GJ, Tamayo P, Hafler DA, De Jager PL, Mesirov JP (2009) Automated high-dimensional flow cytometric data analysis. *Proc Natl Acad Sci USA* 106:8519–8524
- Sahu SK, Dey DK, Branco MD (2003) A new class of multivariate skew distributions with application to Bayesian regression models. *Can J Stat* 31:129–150
- Schwarz G (1978) Estimating the dimension of a model. *Ann Stat* 6:461–464
- Subedi S, McNicholas PD (2014) Variational Bayes approximations for clustering via mixtures of normal inverse Gaussian distributions. *Adv Data Anal Classif* 8:167–193
- Teschendorff A, Wang Y, Barbosa-Morais N, Brenton J, Caldas C (2005) A variational Bayesian mixture modelling framework for cluster analysis of gene-expression data. *Bioinformatics* 21:3025–3033
- Tortora C, McNicholas P, Browne R (2016) A mixture of generalized hyperbolic factor analyzers. *Adv Data Anal Classif* 10:423–440
- Ueda N, Nakano R, Ghahramani Z, Hinton GE (2000) SMEM algorithm for mixture models. *Neural Comput* 12:2109–2128
- Wang WL (2013) Mixtures of common factor analyzers for high-dimensional data with missing information. *J Multivar Anal* 117:120–133
- Wang WL (2015) Mixtures of common  $t$ -factor analyzers for modeling high-dimensional data with missing values. *Comput Stat Data Anal* 83:223–235
- Wang WL, Lin TI (2016) Maximum likelihood inference for the multivariate  $t$  mixture model. *J Multivar Anal* 149:54–64
- Wang WL, Lin TI (2017) Flexible clustering via extended mixtures of common  $t$ -factor analyzers. *ASTA Adv Stat Anal* 101:227–252
- Wang K, McLachlan GJ, Ng SK, Peel D (2009) EMMIX-skew: EM algorithm for mixture of multivariate skew normal/ $t$  distributions. R package version 1.0-12
- Wang WL, Castro LM, Lin TI (2017a) Automated learning of  $t$  factor analysis models with complete and incomplete data. *J Multivar Anal* 161:157–171
- Wang WL, Liu M, Lin TI (2017b) Robust skew- $t$  factor analysis models for handling missing data. *Stat Methods Appl* 26:649–672
- Waterhouse S, MacKay D, Robinson T (1996) Bayesian methods for mixture of experts. In: Touretzky DS, Mozer MC, Hasselmo ME (eds) *Advances in neural information processing systems*, vol 8. MIT Press, Cambridge