AStA Adv Stat Anal (2017) 101:227–252
DOI 10.1007/s10182-016-0281-0

ORIGINAL PAPER

Flexible clustering via extended mixtures of common *t*-factor analyzers

Wan-Lun Wang[1] · **Tsung-I Lin**[2,3]

Received: 2 March 2016 / Accepted: 21 October 2016 / Published online: 2 November 2016

Abstract Mixtures of *t*-factor analyzers have been broadly used for model-based density estimation and clustering of high-dimensional data from a heterogeneous population with longer-than-normal tails or atypical observations. To reduce the number of parameters in the component covariance matrices, the mixtures of common *t*-factor analyzers (MCtFA) have been recently proposed by assuming a common factor loading across different components. In this paper, we present an extended version of MCtFA using distinct covariance matrices for component errors. The modified mixture model offers a more appropriate way to represent the data in a graphical fashion. Two flexible EM-type algorithms are developed for iteratively computing maximum likelihood estimates of parameters. Practical considerations for the specification of starting values, model-based clustering, classification of new subject and identification of potential outliers are also provided. We demonstrate the superiority of the proposed methodology by analyzing the Italian wine data and a simulation study.

Keywords Clustering · Classification · Factor loadings · Mixture models · Outlier detection

Mathematics Subject Classification 62H25 · 62H30

✉ Tsung-I Lin
tilin@nchu.edu.tw

Wan-Lun Wang
wlunwang@fcu.edu.tw

1 Department of Statistics, Graduate Institute of Statistics and Actuarial Science, Feng Chia University, Taichung 40724, Taiwan

2 Institute of Statistics, National Chung Hsing University, Taichung, Taiwan

3 Department of Public Health, China Medical University, Taichung, Taiwan

1 Introduction

Mixtures of factor analyzers (MFA) originally introduced by Ghahramani and Hinton (1997) have become the most popular tool for clustering and local dimensionality reduction of high-dimensional data, especially when the number of observations is not relatively large than their dimension. The MFA along with their applications have been extensively studied by Hinton et al. (1997), McLachlan and Peel (2000) and McLachlan et al. (2002, 2003), among others. To reduce the number of parameters, especially when the number of components or features is quite large, Baek et al. (2010) extended the MFA using common component-factor loadings, called mixtures of common factor analyzers (MCFA), and described an alternating expectation conditional maximization (AECM) algorithm (Meng and Dyk 1997) for conducting maximum likelihood (ML) estimation. Wang (2013) further studied an extension of the MCFA approach, which allows practitioners to handle model-based density estimation, clustering, visualization and discriminant analysis of high-dimensional data in the presence of missing values.

A number of different Bayesian strategies have been developed for inferring finite mixture models and its extensions through factor-analytic representations. Diebolt and Robert (1994) presented a Gibbs-sampling scheme to perform posterior inference on Gaussian mixture (GMIX) models. Zhang et al. (2004) advocated the use of the reversible jump Markov chain Monte Carlo (MCMC) algorithm (Green 1995; Richardson and Green 1997) for fitting GMIX models with unknown number of components. Lopes and West (2004) explored feasible MCMC methods for Bayesian model assessments in factor analysis models. Bayesian treatments on MFA have been investigated through a variational Bayes (VB) approximation (Ghahramani and Beal 2000) and a stochastic simulation procedure (Fokouè and Titterington 2003), where there is uncertainty about the dimensionality of the latent spaces, i.e., the unknown number of mixture components and common factors. Recently, Wei and Li (2013) proposed a VB algorithm for learning MCFA from a Bayesian perspective.

In the MFA and MCFA frameworks, component factors and errors are routinely assumed to be normally distributed for mathematical convenience and computational tractability. However, the normality assumption is not always realistic because of its known sensitivity to outliers. Furthermore, a poor fit for the data with longer than normal tails may subsequently yield a wrong clustering identification. To cope with such an obstacle, McLachlan et al. (2007) proposed the mixtures of t-factor analyzers (MtFA), whereby the multivariate t family (Kotz and Nadarajah 2004) with dimension p, mean vector $\boldsymbol{\mu}(\nu > 1)$, covariance matrix $\nu(\nu-2)^{-1}\boldsymbol{\Sigma}(\nu > 2)$, and degrees of freedom (df) ν, denoted by $t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$, is used to be the underlying distribution for both component factors and errors. The multivariate t density is

$$t_p(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{\Gamma\left(\frac{\nu+p}{2}\right)|\boldsymbol{\Sigma}|^{-1/2}}{(\pi\nu)^{p/2}\Gamma\left(\frac{\nu}{2}\right)}\left[1+(\boldsymbol{y}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})/\nu\right]^{-(\nu+p)/2}, \quad \boldsymbol{y}\in\mathscr{R}^p,$$

where the df ν may be viewed as a robustness turning parameter that is used to control the fatness of the tails of the probability distribution.

Specifically, let $\boldsymbol{y}_j = (y_{j1}, \ldots, y_{jp})^{\mathrm{T}}$, $j = 1, \ldots, n$, be np-dimensional vectors of feature variables. The MtFA approach formulates $\boldsymbol{y}_j$ as:

$$\boldsymbol{y}_j = \boldsymbol{\mu}_i + \boldsymbol{B}_i \boldsymbol{u}_{ij} + \boldsymbol{e}_{ij} \quad \text{with probability } \pi_i \quad (i = 1, \ldots, g), \tag{1}$$

where $\boldsymbol{\mu}_i$ is a $p \times 1$ vector of component location, $\boldsymbol{B}_i$ is a $p \times q$ matrix of component factor loadings, $\boldsymbol{u}_{ij}$ is a q-dimensional vector of component factors, and $\boldsymbol{e}_{ij}$ is a p-dimensional vector of component errors. Here, $(\boldsymbol{u}_{ij}^{\mathrm{T}}, \boldsymbol{e}_{ij}^{\mathrm{T}})^{\mathrm{T}}$ is assumed to jointly follow a multivariate t distribution with zero mean, a block-diagonal scale-covariance matrix $\mathrm{diag}\{\boldsymbol{I}_q, \boldsymbol{D}_i\}$, and the df ν_i, where $\boldsymbol{I}_q$ is an identity matrix of size q and $\boldsymbol{D}_i$ is a diagonal matrix. Consequently, the density of $\boldsymbol{y}_j$ for the MtFA is

$$f(\boldsymbol{y}_j) = \sum_{i=1}^{g} \pi_i t_p(\boldsymbol{y}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \nu_i),$$

where $\boldsymbol{\Sigma}_i = \boldsymbol{B}_i \boldsymbol{B}_i^{\mathrm{T}} + \boldsymbol{D}_i$. Note that MtFA will be reduced to MFA as all component dfs ν_i's tend to infinity. McNicholas and Murphy (2008) introduced a new class of Gaussian mixture models with several parsimonious covariance structures, called parsimonious Gaussian mixture models (PGMM). Andrews and McNicholas (2011) investigated a restricted MtFA model, which is obtained by imposing constraints on the df, the factor loadings, and the error covariance matrices. Recently, Wang and Lin (2013) proposed an ad-hoc expectation conditional maximization (ECM; Meng and Rubin 1993) algorithm on the basis of a much smaller hidden data space for fast ML estimation of MtFA. They have also done a simulation study to show that their new procedure can substantially outperform the commonly used expectation maximization (EM; Dempster et al. 1977) algorithm and the AECM algorithm used in McLachlan et al. (2007) in most situations, regardless of how the convergence speed is assessed by the computing time and/or number of iterations.

For model-based clustering of high-dimensional data, in practice, the dimension p is sometime quite large and/or the number of components (clusters) g is sometimes not small. Therefore, the number of parameters in model (1) might be unmanageable and, thus, encounters near-singular estimates or inestimable component covariance matrices. As a robust extension of MCFA, Baek and McLachlan (2011) proposed a parsimonious version of the MtFA, named as mixtures of common t-factor analyzers (MCtFA), which utilizes common factor loadings to reduce further the number of parameters in the specification of the component-covariance matrices. For the consideration of different covariance matrices for latent factors, this paper presents an extended version of MCtFA, called the EMCtFA, studies its essential properties and describes two variants of the EM algorithm, including the ECM and the expectation conditional maximization *either* (ECME; Liu and Rubin 1994) algorithms for ML estimation of model parameters.

As an alternative to exact ML methods, the simulated ML estimation can be implemented for the model using the Monte Carlo (MC) or importance sampling (IS) methods, known as the MCEM and ISEM algorithms. One drawback of simulated ML methods is that the model fitting procedure relies on MC estimates which can be

difficult to implement due to the heavy computational burden. Another issue is that an increase in log-likelihood at each iteration is not guaranteed because of MC errors (McLachlan and Krishnan 2008). Our proposed EM-type algorithms have exactly closed-form expressions in the E-step and analytically reduced expressions in CM-steps, yielding more accurate estimates than the simulated ML methods.

In this work, we provide a guideline for choosing a set of suitable initial values. Furthermore, the probabilistic classification of new subjects and estimation of latent factors are also investigated. Under the assumption of non-normality, importantly, there is also a problem of outlier detection in mixture modeling. Outliers usually lead to overestimating the number of components to offer a good presentation of the data (Fraley and Raftery 2002). We also offer a rule for identifying which observations are suspected outliers under the EMCtFA framework.

The remainder of this paper is structured as follows. In Sect. 2, we establish the notation and formulate the EMCtFA model. Section 3 presents two EM-type algorithms for fitting EMCtFA and outlines a simple way of setting the initialization. Section 4 describes some practical tools, including model-based clustering, classification, outlier detection and model selection. In Sect. 5, the application of the proposed methodology is illustrated through analyzing the Italian wine data. In Sect. 6, we conduct a simulation study to compare the performance of our recommended initialization procedure with the existing method. We conclude the paper with a short summary in Sect. 7. The detailed derivations are sketched in "Appendix".

2 Extended mixtures of common t-factor analyzers (EMCtFA)

Consider n independent p-dimensional feature vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, which come independently from a nonhomogeneous population with g subgroups. In the sense of dimensionality reduction, q must be smaller than p. The EMCtFA model for continuous features $\boldsymbol{y}_j$ can be described as:

$$\boldsymbol{y}_j = \boldsymbol{A}\boldsymbol{u}_{ij} + \boldsymbol{e}_{ij} \quad \text{with probability } \pi_i \quad (i = 1, \ldots, g), \tag{2}$$

where $\boldsymbol{A}$ is a $p \times q$ matrix of common factor loadings, $\boldsymbol{e}_{ij}$ is a p-dimensional vector of component errors, and $\pi_i \in (0, 1)$ is the mixing proportion subject to $\sum_{i=1}^{g} \pi_i = 1$. The joint distribution of $\boldsymbol{u}_{ij}$ and $\boldsymbol{e}_{ij}$ for the ith component is assumed to be

$$\begin{bmatrix} \boldsymbol{u}_{ij} \\ \boldsymbol{e}_{ij} \end{bmatrix} \sim t_{q+p}\left(\begin{bmatrix} \boldsymbol{\beta}_i \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Omega}_i & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D}_i \end{bmatrix}, \nu_i \right), \tag{3}$$

where $\boldsymbol{\beta}_i$ is a q-dimensional location vector, $\boldsymbol{\Omega}_i$ is a $q \times q$ positive-definite scale-covariance matrix, $\boldsymbol{D}_i$ is a $p \times p$ diagonal covariance matrix, and ν_i is the df. We further assume that the joint distributions of $(\boldsymbol{u}_{ij}^{\mathrm{T}}, \boldsymbol{e}_{ij}^{\mathrm{T}})^{\mathrm{T}}$ for distinct subjects are independent. When $\boldsymbol{D}_i = \boldsymbol{D}$ for all i, the EMCtFA reduces to the original MCtFA model (Baek and McLachlan 2011). Generally, the component dfs are allowed to vary for flexibly controlling possibly different degrees of the tail thickness of component distributions. The special case of equal df, say $\nu_i = \nu$ for all i, is usually considered for the sake

of parsimony and fast convergence of the algorithm. The MCtFA includes the MCFA as a limiting/special case when all component dfs approach infinity simultaneously. It can also be shown that MCtFA is a special case of MtFA by virtue of Eqs. (17)–(21) in Baek et al. (2010).

For the MtFA in (1), $q(q-1)/2$ uniqueness constraints are imposed for component factor loadings $\boldsymbol{B}_i$ and, thus, the total number of parameters in (1) is

$$d_1 = (2g-1) + 2gp + g[pq - q(q-1)/2].$$

For the EMCtFA in (2) along with assumption (3), the common factor loadings $\boldsymbol{A}$ must be unique only up to postmultiply by a nonsingular matrix such that its number of free parameters is $pq - q^2$. As a result, the total number of parameters in (2) is

$$d_2 = (2g-1) + pg + q(p-q+g) + gq(q+1)/2,$$

while that in the MCtFA (Baek and McLachlan 2011) is

$$d_3 = (2g-1) + p + q(p-q+g) + gq(q+1)/2.$$

It follows straightforwardly that the difference in numbers of parameters between MtFA and EMCtFA is $d_1 - d_2 = (g-1)q(p-q) + g(p-q)$, which is nonnegative when $p \geq q$ and $g \geq 1$. Meanwhile, the difference in numbers of parameters between the EMCtFA and MCtFA is $d_2 - d_3 = (g-1)p$, which is also nonnegative when $g \geq 1$. Therefore, we have $d_1 \geq d_2 \geq d_3$ if and only if $p \geq q$ and $g \geq 1$. Clearly, the EMCtFA reaches a compromise between the MtFA and MCtFA approaches through the specification of distinct covariance matrices for component errors. The EMCtFA as well as MCtFA are preferable to the MtFA model if the dimension p or the number of component g is relatively large to suffer from the convergence problems. Furthermore, unlike MtFA, the estimated posterior means of factor scores of EMCtFA can be used to portray the data in low-dimensional subspaces.

Let $\boldsymbol{\Theta} = \{\boldsymbol{A}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g\}$ denote the entire unknown model parameters where $\boldsymbol{\theta}_i = (\pi_i, \boldsymbol{\beta}_i, \boldsymbol{\Omega}_i, \boldsymbol{D}_i, \nu_i)$, $i = 1, \ldots, g$, represents the parameter vector for the ith component. According to (2) and (3), the probability density function (pdf) of $\boldsymbol{y}_j$ is

$$f(\boldsymbol{y}_j \mid \boldsymbol{\Theta}) = \sum_{i=1}^{g} \pi_i t_p(\boldsymbol{y}_j \mid \boldsymbol{A}\boldsymbol{\beta}_i, \boldsymbol{\Sigma}_i, \nu_i),$$

where $\boldsymbol{\Sigma}_i = \boldsymbol{A}\boldsymbol{\Omega}_i\boldsymbol{A}^{\mathrm{T}} + \boldsymbol{D}_i$. Therefore, the ML estimates $\hat{\boldsymbol{\Theta}}$ based on a set of independent observations $\boldsymbol{y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ is $\hat{\boldsymbol{\Theta}} = \mathrm{argmax}_{\boldsymbol{\Theta}}\, \ell(\boldsymbol{\Theta} \mid \boldsymbol{y})$, where $\ell(\boldsymbol{\Theta}|\boldsymbol{y}) = \sum_{j=1}^{n} \log f(\boldsymbol{y}_j|\boldsymbol{\Theta})$ is the observed log-likelihood function. Unfortunately, there are no explicit analytical solutions for ML estimator of $\boldsymbol{\Theta}$. In this case, we resort to the EM-type algorithm (Dempster et al. 1977), which is popular iterative device for ML estimation in models incorporating new hidden variables.

In the EM framework on supporting the interpretation of missing data problem, it is convenient to introduce a set of allocation variables $\boldsymbol{Z}_j = (z_{1j}, \ldots, z_{gj})$, $j =$

$1, \ldots, n$, where the component membership $z_{ij} = 1$ if $\boldsymbol{y}_j$ belongs to the ith component and $z_{ij} = 0$ otherwise. This indicates that $\boldsymbol{Z}_j$ independently follows a multinomial distribution with one trial and mixing properties $(\pi_1, \ldots, \pi_g)$ subject to $\sum_{i=1}^{g} \pi_i = 1$, denoted as $\boldsymbol{Z}_j \sim \mathscr{M}(1; \pi_1, \ldots, \pi_g)$. Based on the essential property of multivariate t distribution, we also utilize the scaling variables τ_js following the gamma distribution with shape $\nu_i/2$ and rate $\nu_i/2$ conditioning on $z_{ij} = 1$. Through introducing the latent variables $\boldsymbol{Z}_j$ and τ_j, for $j = 1, \ldots, n$, three hierarchical representations of the EMCtFA are sketched in "Appendix".

As a consequence, we establish Proposition 1, which is useful for evaluating the conditional expectations involved in the ECME algorithm described in the next section.

Proposition 1 *Given the hierarchical representations* (18)–(20)*, we have*

$$\begin{bmatrix} \boldsymbol{y}_j \\ \boldsymbol{u}_{ij} \end{bmatrix} \Bigg| (\tau_j, z_{ij} = 1) \sim \mathscr{N}_{p+q}\left(\begin{bmatrix} \boldsymbol{A}\boldsymbol{\beta}_i \\ \boldsymbol{\beta}_i \end{bmatrix}, \tau_j^{-1} \begin{bmatrix} \boldsymbol{\Sigma}_i & \boldsymbol{A}\boldsymbol{\Omega}_i \\ \boldsymbol{\Omega}_i \boldsymbol{A}^{\mathrm{T}} & \boldsymbol{\Omega}_i \end{bmatrix} \right).$$

It follows that

$$\begin{bmatrix} \boldsymbol{y}_j \\ \boldsymbol{u}_{ij} \end{bmatrix} \Bigg| (z_{ij} = 1) \sim t_{p+q}\left(\begin{bmatrix} \boldsymbol{A}\boldsymbol{\beta}_i \\ \boldsymbol{\beta}_i \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_i & \boldsymbol{A}\boldsymbol{\Omega}_i \\ \boldsymbol{\Omega}_i \boldsymbol{A}^{\mathrm{T}} & \boldsymbol{\Omega}_i \end{bmatrix}, \nu_i \right).$$

A simple algebra shows that

$$\boldsymbol{u}_{ij} \mid (\boldsymbol{y}_j, \tau_j, z_{ij} = 1) \sim \mathscr{N}_q(\boldsymbol{\beta}_i + \boldsymbol{\gamma}_i^{\mathrm{T}}(\boldsymbol{y}_j - \boldsymbol{A}\boldsymbol{\beta}_i), \tau_j^{-1}(\boldsymbol{I}_q - \boldsymbol{\gamma}_i^{\mathrm{T}}\boldsymbol{A})\boldsymbol{\Omega}_i), \tag{4}$$

$$\tau_j \mid (\boldsymbol{y}_j, z_{ij} = 1) \sim \text{Gamma}\left(\frac{\nu_i + p}{2}, \frac{\nu_i + \delta_{ij}}{2}\right), \tag{5}$$

where $\boldsymbol{\gamma}_i = \boldsymbol{\Sigma}_i^{-1}\boldsymbol{A}\boldsymbol{\Omega}_i$ *and* $\delta_{ij} = (\boldsymbol{y}_j - \boldsymbol{A}\boldsymbol{\beta}_i)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_j - \boldsymbol{A}\boldsymbol{\beta}_i)$ *denotes the Mahalanobis distance between the observation* $\boldsymbol{y}_j$ *and the component mean* $\boldsymbol{A}\boldsymbol{\beta}_i$*. Subsequently, multiplying* (4) *by* (5) *and then integrating out* τ_j *implies*

$$\boldsymbol{u}_{ij} \mid (\boldsymbol{y}_j, z_{ij} = 1) \sim t_q\left(\boldsymbol{\beta}_i + \boldsymbol{\gamma}_i^{\mathrm{T}}(\boldsymbol{y}_j - \boldsymbol{A}\boldsymbol{\beta}_i), \left(\frac{\nu_i + \delta_{ij}}{\nu_i + p}\right)(\boldsymbol{I}_q - \boldsymbol{\gamma}_i^{\mathrm{T}}\boldsymbol{A})\boldsymbol{\Omega}_i, \nu_i + p\right).$$

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

3 Parameter estimation

3.1 ML estimation via the ECM and ECME algorithms

The EM algorithm has several appealing features including simplicity of implementation and monotone convergence with each iteration increasing the likelihood. However, the EM algorithm is not straightforward for ML estimation of the model (2) because its M-step is computationally difficult. To go further, we exploit a variant of the EM

algorithm, called the ECME (Liu and Rubin 1994) algorithm. The ECME algorithm proceeds to estimate parameters by replacing the M-steps of EM with *either* CM-steps that maximize a sequence of constrained Q functions, as in ECM, *or* CML-steps that maximize the correspondingly constrained actual likelihood function. Furthermore, it shares the appealing features of EM (Dempster et al. 1977) and ECM (Meng and Rubin 1993), and possesses typically a faster convergence rate than either EM or ECM in terms of CPU time and/or number of iterations.

For notational convenience, we denote the allocation variables by $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$, the scaling variables by $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_n\}$ and unobservable factors by $\boldsymbol{U} = \{\boldsymbol{u}_{ij}; i = 1, \ldots, g, j = 1, \ldots, n\}$. Treating $(\boldsymbol{Z}, \boldsymbol{\tau}, \boldsymbol{U})$ as the "missing" data and combining them with the observed data $\boldsymbol{y}$ as the "complete" data, the complete-data log-likelihood function of $\boldsymbol{\Theta}$ based on hierarchy (20) is

$$\begin{aligned}\ell_c(\boldsymbol{\Theta}|\boldsymbol{y}, \boldsymbol{Z}, \boldsymbol{\tau}, \boldsymbol{U}) = \sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij}\{&\log \pi_i + \log \phi_p(\boldsymbol{y}_j \mid \boldsymbol{A}\boldsymbol{u}_{ij}, \tau_j^{-1}\boldsymbol{D}_i) \\ &+ \log \phi_q(\boldsymbol{u}_{ij} \mid \boldsymbol{\beta}_i, \tau_j^{-1}\boldsymbol{\Omega}_i) + \log \mathscr{G}(\tau_j \mid \nu_i/2, \nu_i/2)\}, \quad (6)\end{aligned}$$

where $\phi_p(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ stands for the pdf of the p-variate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and $\mathscr{G}(\cdot|a, b)$ denotes the pdf of the gamma distribution with mean a/b and variance a/b^2.

Let $\hat{\boldsymbol{\Theta}}^{(k)} = (\hat{\boldsymbol{A}}^{(k)}, \hat{\pi}_i^{(k)}, \hat{\boldsymbol{\beta}}_i^{(k)}, \hat{\boldsymbol{\Omega}}_i^{(k)}, \hat{\boldsymbol{D}}_i^{(k)}, \hat{\nu}_i^{(k)}, i = 1, \ldots, g)$ be the estimates of $\boldsymbol{\Theta}$ at the kth iteration. In the E-step of ECME, one needs to evaluate the conditional expectation of (6) at $\boldsymbol{\Theta} = \hat{\boldsymbol{\Theta}}^{(k)}$, which is the so-called Q-function:

$$Q(\boldsymbol{\Theta}|\hat{\boldsymbol{\Theta}}^{(k)}) = E(\ell_c(\boldsymbol{\Theta}|\boldsymbol{y}, \boldsymbol{Z}, \boldsymbol{\tau}, \boldsymbol{U})|\boldsymbol{y}, \hat{\boldsymbol{\Theta}}^{(k)}). \quad (7)$$

All necessary conditional expectations in (7) can result from Eq. (21). The CM-steps, each of which maximizes the constrained Q-function or the constrained log-likelihood function over $\boldsymbol{\Theta}$ step-by-step but conditioned on some vector functions of $\boldsymbol{\Theta}$ being estimated at its previous step, proceed as follows:

CM-step 1 for ECM and ECME Fix $\nu_i = \hat{\nu}_i^{(k)} (i = 1, \ldots, g)$, and update $\hat{\pi}_i^{(k)}$, $\hat{\boldsymbol{A}}^{(k)}$, $\hat{\boldsymbol{\beta}}_i^{(k)}$, $\hat{\boldsymbol{\Omega}}_i^{(k)}$, and $\hat{\boldsymbol{D}}_i^{(k)}$ by maximizing (7), which gives

$$\begin{aligned}\hat{\pi}_i^{(k+1)} &= \sum_{j=1}^{n} \hat{z}_{ij}^{(k)}/n, \\ \hat{\boldsymbol{A}}^{(k+1)} &= \left\{\sum_{j=1}^{n}\sum_{i=1}^{g} \hat{z}_{ij}^{(k)}\hat{\tau}_{ij}^{(k)}\boldsymbol{y}_j\left[\hat{\boldsymbol{\beta}}_i^{(k)\mathrm{T}} + \hat{\boldsymbol{y}}_{ij}^{(k)\mathrm{T}}\hat{\boldsymbol{\gamma}}_i^{(k)}\right]\right\} \\ &\quad \times \left\{\sum_{j=1}^{n}\sum_{i=1}^{g} \hat{z}_{ij}^{(k)}\left[(\boldsymbol{I}_q - \hat{\boldsymbol{\gamma}}_i^{(k)\mathrm{T}}\hat{\boldsymbol{A}}^{(k)})\hat{\boldsymbol{\Omega}}_i^{(k)} + \hat{\tau}_{ij}^{(k)}\hat{\boldsymbol{u}}_{ij}^{(k)}\hat{\boldsymbol{u}}_{ij}^{(k)\mathrm{T}}\right]\right\}^{-1}, \quad (8)\end{aligned}$$

$$\hat{\boldsymbol{\beta}}_i^{(k+1)} = \hat{\boldsymbol{\beta}}_i^{(k)} + \frac{\sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{\tau}_{ij}^{(k)} \hat{\boldsymbol{\gamma}}_i^{(k)\mathrm{T}} \hat{\mathbf{y}}_{ij}^{(k)}}{\sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{\tau}_{ij}^{(k)}}, \tag{9}$$

$$\hat{\boldsymbol{\Omega}}_i^{(k+1)} = \frac{\sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{\tau}_{ij}^{(k)} \hat{\boldsymbol{\gamma}}_i^{(k)\mathrm{T}} \hat{\mathbf{y}}_{ij}^{(k)} \hat{\mathbf{y}}_{ij}^{(k)\mathrm{T}} \hat{\boldsymbol{\gamma}}_i^{(k)}}{\sum_{j=1}^n \hat{z}_{ij}^{(k)}} + (\boldsymbol{I}_q - \hat{\boldsymbol{\gamma}}_i^{(k)\mathrm{T}} \hat{\boldsymbol{A}}^{(k)}) \hat{\boldsymbol{\Omega}}_i^{(k)}, \tag{10}$$

and

$$\hat{\boldsymbol{D}}_i^{(k+1)} = \mathrm{diag}\left\{ \hat{\boldsymbol{D}}_i^{(k)} \left(\boldsymbol{I}_p - \hat{\boldsymbol{\Sigma}}_i^{(k)^{-1}} \hat{\boldsymbol{D}}_i^{(k)}\right) + \frac{\hat{\boldsymbol{D}}_i^{(k)} \hat{\boldsymbol{\Sigma}}_i^{(k)-1} \left(\sum_{j=1}^n \hat{z}_{ij}^{(k)} \hat{\tau}_{ij}^{(k)} \hat{\mathbf{y}}_{ij}^{(k)} \hat{\mathbf{y}}_{ij}^{(k)\mathrm{T}}\right) \hat{\boldsymbol{\Sigma}}_i^{(k)-1} \hat{\boldsymbol{D}}_i^{(k)}}{\sum_{j=1}^n \hat{z}_{ij}^{(k)}} \right\}, \tag{11}$$

for $i = 1, \ldots, g$. When $\boldsymbol{D}_i$s are assumed to be the same across components, that is $\boldsymbol{D}_1 = \cdots = \boldsymbol{D}_g = \boldsymbol{D}$, the updated formula for $\boldsymbol{D}$ is given by

$$\hat{\boldsymbol{D}}^{(k+1)} = \mathrm{diag}\left\{ \frac{\sum_{j=1}^n \sum_{i=1}^g \hat{z}_{ij}^{(k)} \hat{\boldsymbol{D}}^{(k)} - \sum_{j=1}^n \sum_{i=1}^g \hat{z}_{ij}^{(k)} \hat{\boldsymbol{D}}^{(k)} \hat{\boldsymbol{\Sigma}}_i^{(k)-1} \hat{\boldsymbol{D}}^{(k)}}{\sum_{j=1}^n \sum_{i=1}^g \hat{z}_{ij}^{(k)}} + \frac{\sum_{j=1}^n \sum_{i=1}^g \hat{z}_{ij}^{(k)} \hat{\tau}_{ij}^{(k)} \hat{\boldsymbol{D}}^{(k)} \hat{\boldsymbol{\Sigma}}_i^{(k)-1} \hat{\mathbf{y}}_{ij}^{(k)} \hat{\mathbf{y}}_{ij}^{(k)\mathrm{T}} \hat{\boldsymbol{\Sigma}}_i^{(k)-1} \hat{\boldsymbol{D}}^{(k)}}{\sum_{j=1}^n \sum_{i=1}^g \hat{z}_{ij}^{(k)}} \right\}.$$

CM-step 2 for ECM Solve the roots of the following equation, which maximizes the constrained Q-functions:

$$\log\left(\frac{\nu_i}{2}\right) + 1 - \mathscr{D}_{\mathscr{G}}\left(\frac{\nu_i}{2}\right) + \frac{\sum_{j=1}^n \hat{z}_{ij}^{(k)} (\hat{\kappa}_{ij}^{(k)} - \hat{\tau}_{ij}^{(k)})}{\sum_{j=1}^n \hat{z}_{ij}^{(k)}} = 0, \quad i = 1, \ldots, g. \tag{12}$$

As $\nu_1 = \cdots = \nu_g = \nu$, we obtain $\hat{\nu}^{(k+1)}$ as the solution of the following equation:

$$\log\left(\frac{\nu}{2}\right) + 1 - \mathscr{D}_{\mathscr{G}}\left(\frac{\nu}{2}\right) + \frac{\sum_{i=1}^g \sum_{j=1}^n \hat{z}_{ij}^{(k)} (\hat{\kappa}_{ij}^{(k)} - \hat{\tau}_{ij}^{(k)})}{n} = 0. \tag{13}$$

CM-step 2 for ECME Alternatively, to improve the convergence, we may exploit the advantage of the ECME step. Given current estimates, calculate $\hat{\nu}_i^{(k+1)}$ by maximizing the constrained log-likelihood function, i.e.,

$$\hat{\nu}_i^{(k+1)} = \arg\max_{\nu_i} \left\{ \sum_{j=1}^{n} \log \left(\hat{\pi}_i^{(k+1)} t_p(\boldsymbol{y}_j | \hat{\boldsymbol{A}}^{(k+1)} \hat{\boldsymbol{\beta}}_i^{(k+1)}, \hat{\boldsymbol{\Sigma}}_i^{(k+1)}, \nu_i) \right) \right\}. \quad (14)$$

Similarly, as the case of common dfs ($\nu_1 = \cdots = \nu_g = \nu$), we calculate

$$\hat{\nu}^{(k+1)} = \arg\max_{\nu} \left\{ \sum_{j=1}^{n} \log \left(\sum_{i=1}^{g} \hat{\pi}_i^{(k+1)} t_p(\boldsymbol{y}_j | \hat{\boldsymbol{A}}^{(k+1)} \hat{\boldsymbol{\beta}}_i^{(k+1)}, \hat{\boldsymbol{\Sigma}}_i^{(k+1)}, \nu) \right) \right\}. \quad (15)$$

Note that the solutions of Eqs. (12) and (13) involve a one-dimensional search for ν_i one at a time or for the common df ν, which can be directly done by employing the uniroot routine built in the R package (R Development Core Team 2009) constrained within an appropriate $[2, 200]$ interval. The procedures (14) and (15) can be implemented straightforwardly using the optim routine with starting value $\hat{\nu}_i^{(k)}$ at each iteration. Given a set of suitable initial values $\hat{\boldsymbol{\Theta}}^{(0)}$ recommended in the next subsection, the ECM or ECME algorithms are performed to obtain the ML estimates $\hat{\boldsymbol{\Theta}} = (\hat{\boldsymbol{A}}, \hat{\pi}_i, \hat{\boldsymbol{\beta}}_i, \hat{\boldsymbol{\Omega}}_i, \hat{\boldsymbol{D}}_i, \hat{\nu}_i, i = 1, \ldots, g)$ iteratively until the user's specified stopping rule is achieved. While carrying out quantitative analysis of experimental data, the stopping rule $\ell(\hat{\boldsymbol{\Theta}}^{(k+1)}|\boldsymbol{y}) - \ell(\hat{\boldsymbol{\Theta}}^{(k)}|\boldsymbol{y}) < 10^{-6}$ is employed.

3.2 Initialization

The EM-type algorithm, like other iteration-based methods, may suffer from computational difficulties such as slow or even non-convergence. In particular, when the data are too sparse or the dimension of latent factors is over-specified, a poor choice of initial values $\hat{\boldsymbol{\Theta}}^{(0)}$ may lead to the convergence in the boundary of the parameter space. To alleviate such potential problems, a simple way of automatically generating a set of suitable initial values is recommended below:

1. Perform a K-means clustering (Hartigan and Wong 1979) initialized with respect to a random start. Specify the zero-one component indicator $\hat{\boldsymbol{Z}}_j^{(0)} = (\hat{z}_{1j}^{(0)}, \ldots, \hat{z}_{gj}^{(0)})$ according to the K-means results. The initial values of the mixing proportions π_is are taken as

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

2. Let $\boldsymbol{y}_{(i)}$ be the data in the ith subpopulation (group). Perform the ordinary factor analysis (Spearman 1904) for $\boldsymbol{y}_{(i)}$. The initial estimate of $\hat{\boldsymbol{\Omega}}_i^{(0)}$ is chosen as the sample variance–covariance matrix of the estimated factor scores.
3. Obtain the factor loading matrix for $\boldsymbol{y}_{(i)}$ via the *principle components analysis* (PCA; Flury 1984) method, denoted by $\hat{\boldsymbol{B}}_i^{(0)}$ for $i = 1, \ldots, g$. Set the initial estimate of $\boldsymbol{A}$ as

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

4. As for initial estimate of $\boldsymbol{\beta}_i$, set $\hat{\boldsymbol{\beta}}_i^{(0)} = \hat{\boldsymbol{A}}^{(0)} \bar{\boldsymbol{y}}_i$, where $\bar{\boldsymbol{y}}_i$ is the sample mean vector of $\boldsymbol{y}_{(i)}$, $i = 1, \ldots, g$.
5. The initial estimate of $\boldsymbol{D}_i$ is obtained as a diagonal matrix formed from the diagonal elements of the sample covariance matrix of $\boldsymbol{y}_{(i)}$. As $\boldsymbol{D}_1 = \cdots = \boldsymbol{D}_g = \boldsymbol{D}$, we set $\hat{\boldsymbol{D}}^{(0)}$ as a diagonal matrix formed from the diagonal elements of the pooled within-cluster sample covariance matrix of g partitioned groups of the data.
6. With regard to the initial estimate of ν_i, we recommend setting a relatively large initial value, say $\hat{\nu}_i^{(0)} = 50$, $\forall i$, which corresponds to an initial assumption of near-normality for the component factors and errors.

When implementing ECM and ECME for the EMCtFA, it is advantageous to use the Sherman–Morrison–Woodbury formula (Golub and Loan 1989) to avoid inverting any large $p \times p$ matrix. That is, the inversion of the $p \times p$ matrix $(\boldsymbol{A}\boldsymbol{\Omega}_i\boldsymbol{A}^{\mathrm{T}} + \boldsymbol{D}_i)$ can be undertaken using the following result:

$$\left(\boldsymbol{A}\boldsymbol{\Omega}_i\boldsymbol{A}^{\mathrm{T}} + \boldsymbol{D}_i\right)^{-1} = \boldsymbol{D}_i^{-1} - \boldsymbol{D}_i^{-1}\boldsymbol{A}\left(\boldsymbol{\Omega}_i^{-1} + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{D}_i^{-1}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{D}_i^{-1},$$

which involves only the inverse of $q \times q$ matrix on the right hand side. It follows that $\boldsymbol{\gamma}_i$ can be rewritten as $\boldsymbol{D}_i^{-1}\boldsymbol{A}(\boldsymbol{\Omega}_i^{-1} + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{D}_i^{-1}\boldsymbol{A})^{-1}$. Moreover, to obtain the unique solution of $\boldsymbol{A}$, as suggested by Baek et al. (2010), we perform the Cholesky decomposition on $\hat{\boldsymbol{A}}$ such that $\hat{\boldsymbol{A}}^{\mathrm{T}}\hat{\boldsymbol{A}} = \boldsymbol{C}^{\mathrm{T}}\boldsymbol{C}$, where $\boldsymbol{C}$ is the upper triangular matrix of order q. If we replace $\hat{\boldsymbol{A}}$ by $\hat{\boldsymbol{A}}\boldsymbol{C}^{-1}$, then the orthonormal estimate of $\boldsymbol{A}$, which satisfies the condition of $\hat{\boldsymbol{A}}^{\mathrm{T}}\hat{\boldsymbol{A}} = \boldsymbol{I}_q$, can be obtained. Consequently, the limiting estimates $\hat{\boldsymbol{\beta}}_i$ and $\hat{\boldsymbol{\Omega}}_i$ are given as $\boldsymbol{C}\hat{\boldsymbol{\beta}}_i$ and $\boldsymbol{C}\hat{\boldsymbol{\Omega}}_i\boldsymbol{C}^{\mathrm{T}}$, respectively.

Notably, the EM-based procedures can get trapped in one of the many local maxima of the likelihood function, and such a phenomenon may still occur in the estimation of the EMCtFA, especially when the number of latent factors is over-specified. To circumvent such a limitation, we recommend initializing the algorithm with a variety of slightly different initial values by performing the K-means allocation of subjects with various random starts. The global optimum is obtained by choosing the one with the largest log-likelihood value.

4 Computational aspects

4.1 Clustering

The estimation of the component labels $\boldsymbol{Z}_j$ and factor scores $\boldsymbol{u}_j$ is meaningful for clustering each observation $\boldsymbol{y}_j$ to a suitable cluster and displaying the high-dimensional data in lower-dimensional plots. Once the EMCtFA model has been fitted, a probabilistic clustering of the data into g clusters can be determined based on the maximum a posteriori (MAP) of component membership. That is, $\hat{z}_{ij}^{(k)}$ evaluated at $\boldsymbol{\Theta} = \hat{\boldsymbol{\Theta}}$, denoted by $\hat{z}_{ij}$, indicates the estimated posterior probability that $\boldsymbol{y}_j$ belongs to the ith component. A natural assignment is achieved by assigning each observation to the component which has the highest estimated posterior probability.

From Eq. (21), we calculate the estimated conditional expectation of component factors $\boldsymbol{u}_{ij}$ corresponding to $\boldsymbol{y}_j$ evaluated at $\boldsymbol{\Theta} = \hat{\boldsymbol{\Theta}}$, denoted by $\hat{\boldsymbol{u}}_{ij}$. Then, it is straightforward to estimate the jth factor scores corresponding to $\boldsymbol{y}_j$ as

$$\hat{\boldsymbol{u}}_j = \sum_{i=1}^{g} \hat{z}_{ij} \hat{\boldsymbol{u}}_{ij}. \tag{16}$$

Let $\tilde{z}_{ij} = 1$ if $\hat{z}_{ij} \geq \hat{z}_{hj}$ for $h \neq i$, $i, h = 1, \ldots, g$, and $\tilde{z}_{ij} = 0$ otherwise. Alternatively, substituting $\tilde{z}_{ij}$ for $\hat{z}_{ij}$ in (16) leads to the other posterior estimates of factor scores. Therefore, we can display the p-dimensional observations $\boldsymbol{y}_j$ in a q-dimensional subspace by plotting the corresponding values $\hat{\boldsymbol{u}}_j$. In addition, the fitted values of $\boldsymbol{y}_j$ can be calculated as $\hat{\boldsymbol{y}}_j = \hat{\boldsymbol{A}}\hat{\boldsymbol{u}}_j$.

4.2 Classification for new subjects

It is also of interest to classify a new subject using the EMCtFA approach. For this purpose, let $\boldsymbol{y}_{\text{new}} = (y_{\text{new}1}, \ldots, y_{\text{new}p})^{\text{T}}$ be the observations for a new subject. Suppose that the model for $\boldsymbol{y}_{\text{new}}$ can be written as:

$$\boldsymbol{y}_{\text{new}} = \boldsymbol{A}\boldsymbol{u}_{i,\text{new}} + \boldsymbol{e}_{i,\text{new}} \quad \text{with probability} \quad \pi_i \ (i = 1, \ldots, g),$$

where the joint distribution of $\boldsymbol{u}_{i,\text{new}}$ and $\boldsymbol{e}_{i,\text{new}}$ satisfies the assumption (3). We now turn our attention to diagnose the allocated group of the new subject and characterize its predictive density. The work of classification of the new subject is based on a fitted (i) conditional distribution of the observed vector $\boldsymbol{y}_{\text{new}}$ given an appropriate predictor of factor scores (*conditional prediction*) and (ii) marginal distribution of the observed vector $\boldsymbol{y}_{\text{new}}$ (*marginal prediction*).

Given the model parameters, the strength of allocating $\boldsymbol{y}_{\text{new}}$ to the ith group is characterized by a predictive density $p(\boldsymbol{y}_{\text{new}}|\boldsymbol{A}, \boldsymbol{\theta}_i)$ whose estimated expression is discussed below. The predictive density of $\boldsymbol{y}_{\text{new}}$ is

$$\hat{p}(\boldsymbol{y}_{\text{new}}|\boldsymbol{\Theta}) = \sum_{i=1}^{g} \pi_i \hat{p}(\boldsymbol{y}_{\text{new}}|\boldsymbol{A}, \boldsymbol{\theta}_i)$$

in which the predictive density belonging to component i, say $\hat{p}(\boldsymbol{y}_{\text{new}}|\boldsymbol{A}, \boldsymbol{\theta}_i)$, can be estimated by the conditional and marginal predictions described below.

For *conditional prediction*, the predictive density $p(\boldsymbol{y}_{\text{new}}|\boldsymbol{A}, \boldsymbol{\theta}_i)$ is the conditional density of $\boldsymbol{y}_{\text{new}}$ given the estimated factor scores $\hat{\boldsymbol{u}}_{i,\text{new}}$. Specifically,

$$\hat{p}(\boldsymbol{y}_{\text{new}}|\boldsymbol{A}, \boldsymbol{\theta}_i) = t_p(\boldsymbol{y}_{\text{new}}|\boldsymbol{A}\hat{\boldsymbol{u}}_{i,\text{new}}, \boldsymbol{D}_i, \nu_i).$$

As in (16), a suitable estimate of component factors is the conditional mean of $\boldsymbol{u}_{i,\text{new}}$ which is calculated using an expression analogous to $\hat{\boldsymbol{u}}_{ij}^{(k)}$ with $\boldsymbol{y}_j, \boldsymbol{u}_{ij}$ and $\hat{\boldsymbol{\Theta}}^{(k)}$ replaced by $\boldsymbol{y}_{\text{new}}, \boldsymbol{u}_{i,\text{new}}$ and $\hat{\boldsymbol{\Theta}}$, respectively.

For *marginal prediction*, the predictive density $p(\boldsymbol{y}_{\text{new}}|\boldsymbol{A}, \boldsymbol{\theta}_i)$ is the marginal density of $\boldsymbol{y}_{\text{new}}$, where the term 'marginal' reflects the fact that the component factors $\boldsymbol{u}_{i,\text{new}}$ are integrated out from the joint density of $(\boldsymbol{y}_{\text{new}}^{\text{T}}, \boldsymbol{u}_{i,\text{new}}^{\text{T}})^{\text{T}}$. We, thus, have

$$\hat{p}(\boldsymbol{y}_{\text{new}}|\boldsymbol{A}, \boldsymbol{\theta}_i) = t_p(\boldsymbol{y}_{\text{new}}|\boldsymbol{A}\boldsymbol{\beta}_i, \boldsymbol{\Sigma}_i, \nu_i).$$

Subsequently, the estimated allocation of the new subject to group i is according to a combination of the prior probabilities $\pi_1, \ldots, \pi_g$ and the estimated values of predictive densities $\hat{p}(\boldsymbol{y}_{\text{new}}|\boldsymbol{A}, \boldsymbol{\theta}_1), \ldots, \hat{p}(\boldsymbol{y}_{\text{new}}|\boldsymbol{A}, \boldsymbol{\theta}_g)$, given by

$$\hat{\mathscr{P}}_{i,\text{new}} = \pi_i \hat{p}(\boldsymbol{y}_{\text{new}}|\boldsymbol{A}, \boldsymbol{\theta}_i) \big/ \hat{p}(\boldsymbol{y}_{\text{new}}|\boldsymbol{\Theta}), \quad i = 1, \ldots, g.$$

Within the likelihood-based approach, all model parameters are estimated by the ML estimates $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{\theta}}_i$. Consequently, based on the MAP classification rule, the feature vector $\boldsymbol{y}_{\text{new}}$ is classified to group i if $\hat{\mathscr{P}}_{i,\text{new}} > \hat{\mathscr{P}}_{h,\text{new}}$, for $h \neq i, i, h = 1, \ldots, g$.

4.3 Outlier identification

Identification of outliers is an important issue because few outliers may produce poor clustering results. Just like the use of allocation indicator z_{ij}, introducing the scaling variable τ_j not only facilitates the implementation of the EM-type algorithm but also enables the interpretation of the estimated model. As can be seen from (8) to (11), $\hat{\tau}_{ij}^{(k)}$ can be treated as the weight in the estimation of $\boldsymbol{A}, \boldsymbol{\beta}_i, \boldsymbol{\Omega}_i$ and $\boldsymbol{D}_i$. Because the estimated value of τ_j is negatively correlated with the estimated Mahalanobis distance δ_{ij} between $\boldsymbol{y}_j$ and $\boldsymbol{A}\boldsymbol{\beta}_i$, a small value of $\hat{\tau}_{ij}$ (i.e., $\hat{\tau}_{ij}^{(k)}$ at convergence) would downweight the influence of the corresponding subject, which can be thought of as a suspected outlier.

To explicitly identify which subject should be an outlier, we follow the idea of Lo and Gottardo (2012) to establish a convenient rule of judging a subject with the associated $\hat{\tau}_j = \sum_{i=1}^{g} \tilde{z}_{ij}\hat{\tau}_{ij}$ value smaller than a critical value, where $\tilde{z}_{ij} = 1$ if $\hat{z}_{ij} \geq \hat{z}_{hj}$ for $h \neq i$, and $\tilde{z}_{ij} = 0$ otherwise. From a viewpoint of hypothesis testing,

if we treat $\hat{\tau}_j$ as a test statistic, then the critical value should be theoretically selected based on some standard distributions. Given $z_{ij} = 1$, $\boldsymbol{y}_j$ follows a p-dimensional t-distribution with location $\boldsymbol{A}\boldsymbol{\beta}_i$, scale-covariance $\boldsymbol{\Sigma}_i$ and df ν_i, and the Mahalanobis distance δ_{ij} follows $p\mathscr{F}(p, \nu_i)$, where $\mathscr{F}(a, b)$ denotes a F distribution with dfs a and b. Thus, $\hat{\tau}_j$ has a scale Beta distribution, say $(1 + p/\nu_i)\mathscr{B}eta(\nu_i/2, p/2)$. Under a significance level of α, the critical value is determined as:

$$c = (1 + p/\nu_i)\mathscr{B}_\alpha(\nu_i/2, p/2), \tag{17}$$

where $\mathscr{B}_\alpha(\cdot, \cdot)$ denotes the α quantile of the Beta distribution such that $P(B \geq \mathscr{B}_\alpha) = 1 - \alpha$. Consequently, given $\boldsymbol{y}_j$ belonging to the ith group, if $\hat{\tau}_j < c$ then the corresponding subject will be treated as a suspect outlier.

4.4 Model selection

To choose the preferred models and determine the numbers of latent factors q and components g, we adopt two widely used model selection criteria. Let $\ell_{\max}$ be the maximized log-likelihood, and m the number of free parameters in the model. The Bayesian information criterion (BIC; Schwarz 1978), defined as

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

is the most commonly employed approach to identifying which model gives the best approximation to the underlying density. Accordingly, models with smaller BIC scores are preferred. Under certain regularity conditions, Keribin (2000) presented a theoretical justification for the efficacy of the BIC in determining the number of components of a mixture model. Fraley and Raftery (2002) gave some empirical evidence that the BIC performs well in model-based clustering tasks.

As argued by Biernacki et al. (2000), BIC may not be an ideal way of identifying the number of clusters. Indeed, BIC favors models with more mixture components to provide a good density estimation of the data. Instead they proposed an alternative promising measure for estimating the proper number of clusters based on the integrated completed likelihood (ICL), calculated as:

$$\text{ICL} = \text{BIC} + 2EN(\boldsymbol{z}),$$

where $EN(\boldsymbol{z}) = -\sum_{j=1}^{n}\sum_{i=1}^{g}\hat{z}_{ij}\log\hat{z}_{ij}$ is the entropy of the classification matrix with the (i, j)th entry being $\hat{z}_{ij}$. In the same vein, the smaller the ICL value, the better the model. Typically, the ICL is preferable to BIC for EMCtFA as it leads to fewer factors since it places a higher penalty on more complex models. Nevertheless, there is no unanimity about which criterion is always the best, and a combined use of BIC and ICL could be of help in screening reasonable candidate models.

From a classification viewpoint, the accuracy of classification can be taken as an alternative measure of fitness of data in some sense. To measure the agreement between a clustering of the data and their true group labels, we employ the leave-one-out (LOO)

cross validation of the MAP classifications against the true group labels to evaluate the correct classification rate (CCR; Lee et al. 2003) and the adjusted Rand index (ARI; Hubert and Arabie 1985). The LOO technique is to take one out of subjects and use the remaining subjects as the training data to update the parameters. The CCR, which ranges from zero to one, is computed as the proportion of correct clusters with respect to the true group labels. As a measure of class agreement between two data clustering, the ARI has an expected value of zero under random classification and takes the maximized value one for perfect classification.

5 Application: the Italian wine data

Forina et al. (1986) reported 28 chemical and physical properties of three types of Italian wine, including 59 Barolo, 71 Grignolino and 48 Barbera. A subset of $p = 13$ of these variables (listed in the first column of Table 2) for $n = 178$ wines is available as part of the gclus package (Hurley 2004) of R software. The proposed techniques are demonstrated on the analysis of these Italian wines.

For the sake of comparison, in addition to the EMCtFA model, the MFA, MtFA, MCFA, MCtFA and EMCFA (extended MCFA, which is the original MCFA with distinct variance–covariance matrices for latent factors) approaches are also fitted to the data. Prior to analyses, each variable is standardized to have zero mean and unit standard deviation. For the MtFA, EMCtFA, and MCtFA, the assumption of equal and unequal dfs is imposed on the component factors and errors. Henceforth, their counterparts in the case of equal dfs, say $\nu_i = \nu$ for all i, are named as the 'MtFAe', 'EMCtFAe', and 'MCtFAe', respectively. For supervised learning of the wine data with three class labels, the nine candidate models are fitted with $g = 3$ components and q varying from 1 to 8, where the choice of maximum $q = 8$ satisfies the restriction of $(p - q)^2 \geq (p + q)$, as recommended by McLachlan and Peel (2000, Chapter 8). All models are trained by the proposed ECME algorithm over five trials of different *K-means* initializations. The optimal solution is the one providing the largest log-likelihood value.

Table 1 reports the number of model parameters m and the values of BIC and ICL for the considered 72 scenarios in terms of the specification of models and the number of factors q. In light of BIC and ICL, the t-based models outperform their normal counterparts except for the case of $q = 1$. Furthermore, it is evident that both criteria give a consistent preference in the study, that is, the best fit to the data is EMCtFAe ($q = 4$), followed by EMCtFA ($q = 4$), MCtFA ($q = 4$) and MCtFAe ($q = 4$).

The resulting ML estimates of common factor loadings $\hat{\boldsymbol{A}}$ and component means $\hat{\boldsymbol{\mu}}_i = \hat{\boldsymbol{A}}\hat{\boldsymbol{\beta}}_i$ ($i = 1, 2, 3$) together with the empirical sample means $\bar{\boldsymbol{y}}_i$ for the best model are presented in Table 2. Herein, the names for the cluster components are matched with the shortest Euclidean norm of the distance between the sample class means $\bar{\boldsymbol{y}}_i$ and the estimated component means $\hat{\boldsymbol{\mu}}_i$, for $i = 1, 2, 3$. The estimates of mixing proportions are $\hat{\pi}_1 = 0.331$, $\hat{\pi}_2 = 0.382$ and $\hat{\pi}_3 = 0.287$, respectively, and they are very close to the proportions of the corresponding groups of wine data. Besides, the estimate of common df ($\hat{\nu} = 12.658$) is somewhat small, signifying that the heavy-tailed behavior exhibits within the multi-dimensional Italian wine data.

Table 1 Model comparison based on BIC and ICL for the wine data

Model	Criteria	Number of factors q							
		1	2	3	4	5	6	7	8
MFA	m[a]	119	155	188	218	245	269	290	308
	BIC	5344.055	5375.926	5376.443	5432.848	5517.902	5601.253	5677.090	5748.157
	ICL	5350.027	5379.506	5377.866	5434.698	5519.517	5602.502	5678.179	5749.465
MtFAe	m	120	156	189	219	246	270	291	309
	BIC	5376.098	5305.470	5321.324	5395.514	5478.975	5550.865	5658.809	5724.323
	ICL	5388.640	5312.149	5324.695	5397.838	5480.836	5552.562	5660.413	5725.792
MtFA	m	122	158	191	221	248	272	293	311
	BIC	5382.408	5303.722	5318.125	5381.371	5477.953	5551.949	5665.408	5725.309
	ICL	5395.045	5307.217	5320.120	5383.157	5479.509	5553.620	5667.162	5727.160
EMCFA	m	59	78	98	119	141	164	188	213
	BIC	5823.975	5394.782	5300.277	5264.029	5289.268	5335.393	5339.032	5401.752
	ICL	5888.161	5433.193	5319.124	5274.928	5291.791	5338.842	5341.644	5419.296
EMCtFAe	m	60	79	99	120	142	165	189	214
	BIC	5796.207	5332.388	5243.569	**5182.898**	5253.118	5276.933	5324.928	5407.735
	ICL	5870.694	5366.638	5266.744	**5196.593**	5262.271	5285.593	5328.208	5412.315
EMCtFA	m	62	81	101	122	144	167	191	216
	BIC	5806.261	5342.419	5255.913	5192.263	5252.605	5274.697	5322.150	5400.859
	ICL	5879.232	5375.589	5277.238	5205.502	5258.867	5276.509	5323.926	5405.616
MCFA	m	33	52	72	93	115	138	162	187
	BIC	5888.355	5481.908	5334.233	5287.007	5303.143	5321.781	5323.837	5375.313
	ICL	5976.910	5522.299	5353.530	5300.869	5310.620	5327.262	5329.954	5381.429

Table 1 continued

Model	Criteria	Number of factors q							
		1	2	3	4	5	6	7	8
MCtFAe	m	34	53	73	94	116	139	163	188
	BIC	5796.230	5348.065	5230.698	5208.511	5230.544	5257.994	5272.962	5325.552
	ICL	5865.856	5393.808	5265.121	5231.834	5249.186	5270.822	5284.037	5336.421
MCtFA	m	36	55	75	96	118	141	165	190
	BIC	5803.504	5355.837	5235.408	5206.592	5230.828	5250.944	5265.068	5342.882
	ICL	5866.577	5392.161	5254.845	5215.239	5235.709	5255.048	5268.937	5346.341

The smallest BIC and ICL scores are indicated in bold

[a] m is the number of model parameters

Table 2 Estimation results for the fitted EMCtFAe with $g = 3$ and $q = 4$ on the wine data

Variable	Common factor loadings: $\hat{\boldsymbol{A}}$				Barolo		Grignolino		Barbera	
	$\boldsymbol{a}_1$	$\boldsymbol{a}_2$	$\boldsymbol{a}_3$	$\boldsymbol{a}_4$	$\bar{\boldsymbol{y}}_1$	$\hat{\boldsymbol{\mu}}_1$	$\bar{\boldsymbol{y}}_2$	$\hat{\boldsymbol{\mu}}_2$	$\bar{\boldsymbol{y}}_3$	$\hat{\boldsymbol{\mu}}_3$
Alcohol	−0.357	−0.203	−0.308	0.124	0.917	0.856	−0.889	−0.756	0.189	0.135
Malic	−0.220	−0.052	0.325	0.021	−0.292	−0.505	−0.361	−0.222	0.893	0.858
Ash	−0.555	0.559	−0.017	−0.255	0.325	0.257	−0.444	−0.390	0.257	0.205
Alcalinity	−0.175	0.356	0.397	−0.196	−0.736	−0.735	0.223	0.160	0.575	0.539
Magnesium	−0.280	0.070	−0.221	0.016	0.462	0.550	−0.364	−0.441	−0.030	0.011
Phenols	−0.083	−0.219	−0.025	−0.491	0.871	0.838	−0.058	0.029	−0.985	−1.025
Flavanoids	−0.014	−0.158	−0.108	−0.497	0.954	0.923	0.052	0.125	−1.249	−1.252
Nonflavanoid	−0.099	0.260	0.118	0.183	−0.577	−0.586	0.015	−0.071	0.688	0.759
Proanthocyanins	−0.043	−0.264	0.091	−0.422	0.539	0.531	0.069	0.071	−0.764	−0.738
Intensity	−0.475	−0.457	0.161	0.102	0.203	0.198	−0.850	−0.854	1.009	0.985
Hue	0.229	0.296	−0.398	−0.107	0.457	0.563	0.433	0.360	−1.202	−1.130
OD280	0.157	0.069	−0.195	−0.378	0.769	0.682	0.245	0.371	−1.307	−1.315
Proline	−0.293	−0.024	−0.579	0.128	1.171	1.226	−0.722	−0.686	−0.372	−0.366

In the model-based classification framework, the main objective is to estimate the group memberships for new subjects with unknown group memberships. To empirically demonstrate the performance of the proposed classification procedure described in Sect. 4.2, we calculate the CCR and ARI of the classification results based on both marginal and conditional predictions. Table 3 tabulates the classification performance of the best fitting models under each class of EMCtFA and MCtFA models, respectively, namely EMCtFAe ($q = 4$) and MCtFA ($q = 4$). The EMCtFAe model shows a slight improvement on the classification accuracy by virtue of having higher CCR and ARI values compared to the MCtFA.

Figure 1 shows the 3D scatter-ellipsoids plots of two triple estimated factor scores calculated using (16) for the best model, where the colors of the dots correspond to the true class labels. It is interesting to see that the three groups of wines can be visually separated by mapping the estimated factor scores to a low-dimensional space. Furthermore, it is of interest to detect outlying observations based on the identification rule described in Sect. 4.2. With a significance level of $\alpha = 0.05$, any subject with the estimated $\hat{\tau}_j$ which is less than the critical value $c = 0.552$, calculated by (17), will be deemed as an outlier. Using such an identification rule indicates that Barolo wine 14, Grignolino wines 62, 69, 70, 74, 96, 97, 111 and 122, and Barbera wines 159 and 160 can be thought of as potential outliers. The finding is consistent with the estimate of df, reflecting that the wine data have longer-than-normal noises.

6 Simulation

In this section, we conduct a small-scale simulation study to compare the performance of the initialization method presented in Sect. 3.2 (Method 1) with the strategy described in the Appendix of Baek et al. (2010) (Method 2). The computation was carried out by R package 2.13.1 in win 64 environment of desktop PC machine with 3.40 GHz/Intel Core(TM) i7-2600 CPU Processor and 8.0 GB RAM. We generate 100 artificial data points in $\mathscr{R}^{10+p_2}$ of size $n = 100$ and 250 from a five-component EMCtFA model with $q = 2$. The dimension for noise variables p_2 is set to 0 and 20, so the numbers of total variables p are equal to 10 and 30, respectively. Specifically, the artificial data were generated from

$$\boldsymbol{y}_j = (\boldsymbol{A}_1^{\mathrm{T}}, \boldsymbol{A}_2^{\mathrm{T}})^{\mathrm{T}} \boldsymbol{u}_{ij} + \boldsymbol{e}_{ij} \quad \text{with probability } \pi_i, (i = 1, \ldots, 5),$$

in which the distributional assumption for $(\boldsymbol{u}_{ij}^{\mathrm{T}}, \boldsymbol{e}_{ij}^{\mathrm{T}})^{\mathrm{T}}$ satisfies Eq. (3). The presumed model parameters are the same with those specified in Section 6 of Baek et al. (2010), except for $\nu_i = 5$ for $i = 1, \ldots, 5$. Each simulated dataset was fitted with the EMCtFA ($g = 5$, $q = 2$) by implementing the ECME algorithm with parameters initialized once from each of the two methods. A total of 100 independent replications were run for each simulated case.

Table 4 lists the averages of required numbers of iterations, consumed CPU time (in seconds) until convergence, initial and maximized (converged) log-likelihood values, and the CCR and ARI values for clustering results along with the number of non-convergence cases (in parentheses) over 100 trials. Those non-convergence cases occur

Table 3 Summary of outcome groups against classification results under the EMCtFAe ($q = 4$) and MCtFA ($q = 4$) models with CCR and ARI values (based on marginal prediction and conditional prediction) together with the corresponding mixing proportions

Classify to			EMCtFAe			MCtFA		
			1	2	3	1	2	3
Marginal	True group	1: Barolo	58	1	0	59	0	0
		2: Grignolino	1	67	3	5	66	0
		3: Barbera	0	0	48	0	1	47
	$\hat{\pi}_i$		0.331	0.382	0.287	0.360	0.376	0.264
	CCR		0.972			0.966		
	ARI		0.898			0.895		
Conditional	True group	1: Barolo	58	1	0	59	0	0
		2: Grignolino	4	66	1	8	62	1
		3: Barbera	0	0	48	0	1	47
	$\hat{\pi}_i$		0.348	0.377	0.275	0.376	0.354	0.270
	CCR		0.966			0.944		
	ARI		0.895			0.830		

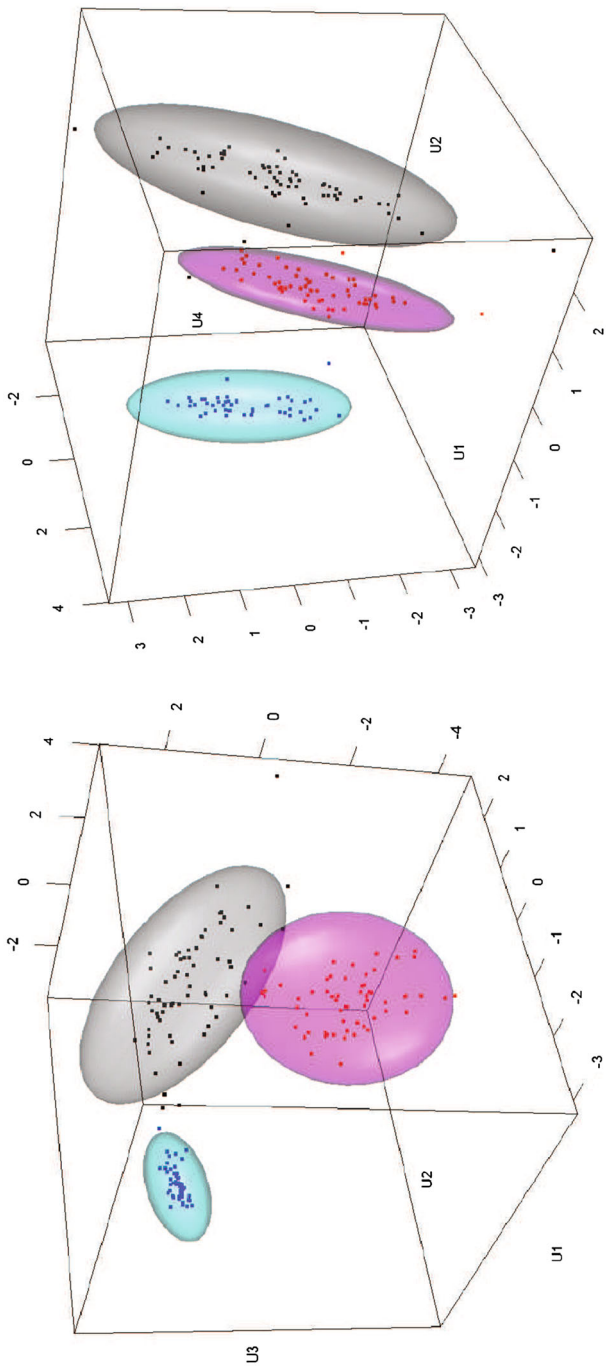

Fig. 1 3D Scatterplots along with 95% confidence ellipsoids of the triple estimated factor scores colored according to the classified clusters under the fitted EMCtFAe with $g = 3$ and $q = 4$ for the wine data

Table 4 Averages of numbers of iterations, CPU time (in seconds) for convergence, the log-likelihood values at the initial and converged iterations, and the CCR and ARI values for clustering results along with the number of non-convergence cases (in parentheses) over 100 replicates using ECME algorithm with two different initialization methods

Sample size n	Dimension p	Initialization method	(NA)	Convergence		Log-likelihood		Clustering	
				Iteration	CPU time	$\ell(\hat{\boldsymbol{\Theta}}^{(0)}\|\boldsymbol{y})$	$\ell_{\max}$	CCR	ARI
100	10	1	(1)	852.740	46.045	−1579.639	−990.340	0.811	0.625
		2	(5)	1744.660	94.258	−2081.302	−1001.861	0.657	0.454
	30	1	(2)	662.571	112.630	−4360.225	−3434.794	0.765	0.570
		2	(17)	1687.093	290.829	−4923.601	−3462.028	0.590	0.370
250	10	1	(2)	350.816	48.087	−3797.980	−2525.842	0.840	0.651
		2	(10)	1344.390	182.907	−5084.119	−2543.675	0.724	0.513
	30	1	(3)	279.639	100.752	−10910.475	−8668.990	0.804	0.605
		2	(14)	1450.347	526.999	−12223.312	−8700.376	0.667	0.453

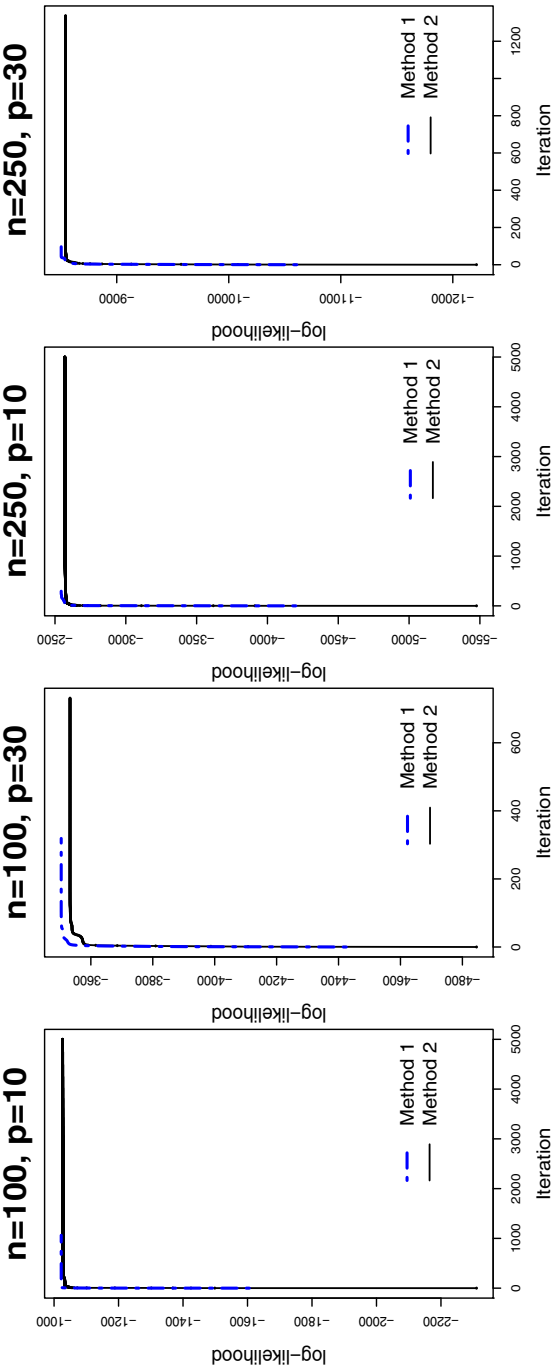

Fig. 2 Typical evolvements of log-likelihood values fitted to one of simulated datasets by the EMCtFA with $g = 5$ and $q = 2$ for each case through implementing the ECME with three different initialization methods

due mostly to the singularity of scale-covariance matrices $\boldsymbol{\Sigma}_i$ during the iterations before convergence. Figure 2 displays the typical evolvements of log-likelihood values for one of the 100 replicates for each considered case. The numerical results indicate that the initialization of Method 1 leads to much faster convergence speed as it requires smaller numbers of iterations and less CPU time than Method 2. Meanwhile, Method 1 can obtain higher starting log-likelihood values (closer to the maximized log-likelihood values upon convergence of ECME) and maximized log-likelihood values as well as better classification performance in terms of CCR and ARI. This study illustrates the effectiveness of our recommended initialization procedure. The poor performance of Method 2 is largely attributable to the fact that it generates inappropriate initial values simply from the standard normal distribution for each entry of $\hat{\boldsymbol{A}}^{(0)}$.

7 Conclusion

The MtFA approach indeed provides a more flexible formulation of the component scale-covariances and the component means without restrictions. Hence, it is useful for analyzing the high-dimensional data with heavy tails or atypical observations. In this paper, we have studied a comparable approach, named as EMCtFA, using a factor-analytic representation of the multivariate t-component scale-covariance matrices with common factor loadings and distinct covariance matrices for latent factors and errors. The EMCtFA approach, which contains the MCtFA as a special case, achieves a compromised reduction in the number of parameters, particularly when the dimension p and the number of clusters g are not small. This approach is very well suited for clustering a wide variety of high-dimensional data into several clusters and provides robustness (less sensitive to outliers) in the sense of resulting number of clusters.

In this work, we have developed two computationally flexible EM-type algorithms and offered a simple way of generating suitable initial values for carrying out ML estimation of the EMCtFA model within a convenient complete data framework. The utility of the proposed approach has been demonstrated through experimental studies based on the real and simulated datasets. Numerical results have also shown that the proposed techniques perform reasonably well for the Italian wine data and outperform some common existing approaches.

To alleviate some limitations associated with the deterministic likelihood-based approach, one may resort to the VB approximation method working with maximization of a lower bound on the marginal log-likelihood (Jordan 1999; Corduneanu and Bishop 2001; Tzikas et al. 2008; Zhao and Yu 2009). The VB strategy has been shown effective to simultaneous estimate model parameters and determine the number of components for the MFA (Ghahramani and Beal 2000) and MCFA (Wei and Li 2013) models. Therefore, it is worthwhile to establish a novel VB scheme for learning the EMCtFA model under an approximated Bayesian paradigm. Besides, it is of interest to extend the EMCtFA based on a broader mixture family of component densities such as the multivariate skew t (Lin 2010; Lee and McLachlan 2014) and the canonical fundamental multivariate skew t (Lee and McLachlan 2016) distributions.

Acknowledgements The authors are grateful to the Chief Editor, the Associate Editor, and two anonymous reviewers for their insightful comments and suggestions that greatly improved this article. This work was partially supported by 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.

Appendix A: Hierarchies and some properties for the EMCtFA

For the convenience of developing ML estimation, we have the first hierarchy of model (2) with assumption (3):

$$\begin{aligned} \boldsymbol{y}_j \mid (z_{ij} = 1) &\sim t_p(\boldsymbol{A}\boldsymbol{\beta}_i, \boldsymbol{\Sigma}_i, \nu_i), \\ \boldsymbol{Z}_j &\sim \mathscr{M}(1; \pi_1, \ldots, \pi_g). \end{aligned} \tag{18}$$

Based on the characterization of multivariate t distributions, the second hierarchy can be expressed as:

$$\begin{aligned} \boldsymbol{y}_j \mid (\tau_j, z_{ij} = 1) &\sim \mathscr{N}_p(\boldsymbol{A}\boldsymbol{\beta}_i, \tau_j^{-1}\boldsymbol{\Sigma}_i), \\ \tau_j \mid (z_{ij} = 1) &\sim \text{Gamma}\,(\nu_i/2, \nu_i/2), \\ \boldsymbol{Z}_j &\sim \mathscr{M}(1; \pi_1, \ldots, \pi_g). \end{aligned} \tag{19}$$

From (3), the zero covariance of $(\boldsymbol{u}_{ij}^{\mathrm{T}}, \boldsymbol{e}_{ij}^{\mathrm{T}})^{\mathrm{T}}$ implicitly implies that $\boldsymbol{u}_{ij} \mid \tau_j$ and $\boldsymbol{e}_{ij} \mid \tau_j$ are assumed to be independent. The third hierarchy can be written as:

$$\begin{aligned} \boldsymbol{y}_j \mid (\boldsymbol{u}_{ij}, \tau_j, z_{ij} = 1) &\sim \mathscr{N}_p(\boldsymbol{A}\boldsymbol{u}_{ij}, \tau_j^{-1}\boldsymbol{D}_i), \\ \boldsymbol{u}_{ij} \mid (\tau_j, z_{ij} = 1) &\sim \mathscr{N}_q(\boldsymbol{\beta}_i, \tau_j^{-1}\boldsymbol{\Omega}_i), \\ \tau_j \mid (z_{ij} = 1) &\sim \text{Gamma}\,(\nu_i/2, \nu_i/2), \\ \boldsymbol{Z}_j &\sim \mathscr{M}(1; \pi_1, \ldots, \pi_g). \end{aligned} \tag{20}$$

Furthermore, we need the following conditional moments of latent variables for E-step of the ECM and ECME procedures. According to hierarchies (18)–(20), it follows from Proposition 1 that

$$\begin{aligned} \hat{z}_{ij}^{(k)} &= E(z_{ij}|\boldsymbol{y}_j, \hat{\boldsymbol{\Theta}}^{(k)}) = \hat{\pi}_i^{(k)} t_p(\boldsymbol{y}_j|\hat{\boldsymbol{A}}^{(k)}\hat{\boldsymbol{\beta}}_i^{(k)}, \hat{\boldsymbol{\Sigma}}_i^{(k)}, \hat{\nu}_i^{(k)}) \big/ f(\boldsymbol{y}_j|\hat{\boldsymbol{\Theta}}^{(k)}), \\ \hat{\tau}_{ij}^{(k)} &= E(\tau_j|\boldsymbol{y}_j, z_{ij} = 1, \hat{\boldsymbol{\Theta}}^{(k)}) = (\hat{\nu}_i^{(k)} + p)/(\hat{\nu}_i^{(k)} + \hat{\delta}_{ij}^{(k)}), \\ \hat{\kappa}_{ij}^{(k)} &= E(\log \tau_j|\boldsymbol{y}_j, z_{ij} = 1, \hat{\boldsymbol{\Theta}}^{(k)}) = \mathscr{D}\mathscr{G}\left(\frac{\hat{\nu}_i^{(k)} + p}{2}\right) - \log\left(\frac{\hat{\nu}_i^{(k)} + \hat{\delta}_{ij}^{(k)}}{2}\right), \\ \hat{\boldsymbol{u}}_{ij}^{(k)} &= E(\boldsymbol{u}_{ij}|\boldsymbol{y}_j, z_{ij} = 1, \hat{\boldsymbol{\Theta}}^{(k)}) = \hat{\boldsymbol{\beta}}_i^{(k)} + \hat{\boldsymbol{\gamma}}_i^{(k)\mathrm{T}}\hat{\boldsymbol{y}}_{ij}^{(k)}, \\ \hat{\boldsymbol{\Psi}}_{ij}^{(k)} &= E(\boldsymbol{u}_{ij}\boldsymbol{u}_{ij}^{\mathrm{T}}|\boldsymbol{y}_j, z_{ij} = 1, \hat{\boldsymbol{\Theta}}^{(k)}) = \hat{\boldsymbol{u}}_{ij}^{(k)}\hat{\boldsymbol{u}}_{ij}^{(k)\mathrm{T}} + \hat{\tau}_{ij}^{(k)-1}(\boldsymbol{I}_q - \hat{\boldsymbol{\gamma}}_i^{(k)\mathrm{T}}\hat{\boldsymbol{A}}^{(k)})\hat{\boldsymbol{\Omega}}_i^{(k)}, \end{aligned} \tag{21}$$

where $\hat{\boldsymbol{\Sigma}}_i^{(k)} = \hat{\boldsymbol{A}}^{(k)}\hat{\boldsymbol{\Omega}}_i^{(k)}\hat{\boldsymbol{A}}^{(k)^{\mathrm{T}}} + \hat{\boldsymbol{D}}_i^{(k)}$, $\hat{\boldsymbol{y}}_i^{(k)} = \hat{\boldsymbol{\Sigma}}_i^{(k)^{-1}}\hat{\boldsymbol{A}}^{(k)}\hat{\boldsymbol{\Omega}}_i^{(k)}$, $\hat{\delta}_{ij}^{(k)} = \hat{\boldsymbol{y}}_{ij}^{(k)\mathrm{T}}\hat{\boldsymbol{\Sigma}}_i^{(k)^{-1}}\hat{\boldsymbol{y}}_{ij}^{(k)}$, $\hat{\boldsymbol{y}}_{ij}^{(k)} = \boldsymbol{y}_j - \hat{\boldsymbol{A}}^{(k)}\hat{\boldsymbol{\beta}}_i^{(k)}$, and $\mathscr{D}_{\mathscr{G}}(\cdot)$ is the digamma function.

References

Andrews, J.L., McNicholas, P.D.: Extending mixtures of multivariate t-factor analyzers. Stat. Comput. **21**(3), 361–373 (2011)

Baek, J., McLachlan, G.J.: Mixtures of common t-factor analyzers for clustering high-dimensional microarray data. Bioinformatics **27**(9), 1269–1276 (2011)

Baek, J., McLachlan, G.J., Flack, L.K.: 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**(7), 1–13 (2010)

Biernacki, C., Celeux, G., Govaert, G.: Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Trans. Pattern Anal. Mach. Intell. **22**, 719–725 (2000)

Corduneanu, A., Bishop, C.: Variational Bayesian model selection for mixture distributions. In: Proceedings of the AI and Statistics Conference, pp. 27–34 (2001)

Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. R. Stat. Soc. B **39**(1), 1–38 (1977)

Diebolt, J., Robert, C.: Estimation of finite mixtures through Bayesian sampling. J. R. Stat. Soc. B **56**, 363–375 (1994)

Flury, B.N.: Common principle components in k groups. J. Am. Stat. Assoc. **79**(388), 892–898 (1984)

Fokouè, E., Titterington, D.M.: Mixtures of factor analyzers. Bayesian estimation and inference by stochastic simulation. Mach. Learn. **50**, 73–94 (2003)

Forina, M., Armanino, C., Castino, M., Ubigli, M.: Multivariate data analysis as a discriminating method of the origin of wines. Vitis **25**(3), 189–201 (1986)

Fraley, C., Raftery, A.E.: Model-based clustering, discriminant analysis and density estimation. J. Am. Stat. Assoc. **97**(458), 611–631 (2002)

Ghahramani, Z., Beal, M.: Variational inference for Bayesian mixture of factor analysers. In: Solla, S., Leen, T., Muller, K.-R (eds.) Advances in Neural Information Processing Systems, vol. 12. MIT Press, Cambridge, pp. 449–455 (2000)

Ghahramani, Z., Hinton, G.E.: The EM algorithm for factor analyzers. In: Technical Report No. CRG-TR-96-1. The University of Toronto, Toronto (1997)

Golub, G.H., Van Loan, C.F.: Matrix Computations, 2nd edn. Johns Hopkins University Press, Baltimore (1989)

Green, P.J.: Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika **82**(4), 711–732 (1995)

Hartigan, J.A., Wong, M.A.: Algorithm AS 136: a K-means clustering algorithm. Appl. Stat. **28**(1), 100–108 (1979)

Hinton, G., Dayan, P., Revow, M.: Modeling the manifolds of images of handwritten digits. IEEE Trans. Neural Netw. **8**(1), 65–73 (1997)

Hubert, L., Arabie, P.: Comparing partitions. J. Classif. **2**(1), 193–218 (1985)

Hurley, C.: Clustering visualizations of multivariate data. J. Comput. Graph. Stat. **13**(4), 788–806 (2004)

Jordan, M.I.: An introduction to variational methods for graphical models. Mach. Learn. **37**, 183–233 (1999)

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

Kotz, S., Nadarajah, S.: Multivariate t Distributions and Their Applications. Cambridge University Press, Cambridge (2004)

Lee, S., McLachlan, G.J.: Finite mixtures of multivariate skew t-distributions: some recent and new results. Stat. Comput. **24**, 181–202 (2014)

Lee, S.X., McLachlan, G.J.: Finite mixtures of canonical fundamental skew t-distributions. Stat. Comput. **26**, 573–589 (2016)

Lee, W.L., Chen, Y.C., Hsieh, K.S.: Ultrasonic liver tissues classification by fractal feature vector based on M-band wavelet transform. IEEE Trans. Med. Imaging **22**, 382–392 (2003)

Lin, T.I.: Robust mixture modeling using multivariate skew t distributions. Stat. Comput. **20**, 343–356 (2010)

Liu, C., Rubin, D.B.: The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. Biometrika **81**(4), 633–648 (1994)
Lo, K., Gottardo, R.: Flexible mixture modeling via the multivariate t distribution with the Box–Cox transformation: an alternative to the skew-t distribution. Stat. Comput. **22**, 33–52 (2012)
Lopes, H.F., West, M.: Bayesian model assessment in factor analysis. Stat. Sin. **14**, 41–67 (2004)
McLachlan, G.J., Bean, R.W., Jones, L.B.T.: Extension of the mixture of factor analyzers model to incorporate the multivariate t-distribution. Comput. Stat. Data Anal. **51**(11), 5327–5338 (2007)
McLachlan, G.J., Bean, R.W., Peel, D.: A mixture model-based approach to the clustering of microarray expression data. Bioinformatics **18**(3), 413–422 (2002)
McLachlan, G.J., Krishnan, T.: The EM Algorithm and Extensions, 2nd edn. Wiley, New York (2008)
McLachlan, G.J., Peel, D., Bean, R.W.: Modelling high-dimensional data by mixtures of factor analyzers. Comput. Stat. Data Anal. **41**, 379–388 (2003)
McNicholas, P.D., Murphy, T.B.: Parsimonious Gaussian mixture models. Stat. Comput. **18**(3), 285–296 (2008)
McLachlan, G.J., Peel, D.: Finite Mixture Models. Wiley, New York (2000)
Meng, X.L., Rubin, D.B.: Maximum likelihood estimation via the ECM algorithm: a general framework. Biometrika **80**(2), 267–278 (1993)
Meng, X.L., van Dyk, D.: The EM algorithm—an old folk-song sung to a fast new tune. J. R. Stat. Soc. B **59**, 511–567 (1997)
R Development Core Team: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna (ISBN 3-900051-07-0). http://www.R-project.org (2009)
Richardson, S., Green, P.J.: On Bayesian analysis of mixtures with an unknown number of components. J. R. Stat. Soc. B **59**, 731–792 (1997)
Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**, 461–464 (1978)
Spearman, C.: 'General intelligence', objectively determined and measured. Am. J. Psychol. **15**, 201–292 (1904)
Tzikas, D.G., Likas, A.C., Galatsanos, N.P.: The variational approximation for Bayesian inference. IEEE Signal Process. **25**, 131–146 (2008)
Wang, W.L.: Mixtures of common factor analyzers for high-dimensional data with missing information. J. Multivar. Anal. **117**, 120–133 (2013)
Wang, W.L., Lin, T.I.: An efficient ECM algorithm for maximum likelihood estimation in mixtures of t-factor analyzers. Comput. Stat. **28**, 751–769 (2013)
Wei, X., Li, C.: Bayesian mixtures of common factor analyzers: model, variational inference, and applications. Signal Process. **93**, 2894–2905 (2013)
Zhang, Z., Chan, K.L., Wu, Y., Chen, C.: Learning a multivariate Gaussian mixture model with the reversible jump MCMC algorithm. Stat. Comput. **14**, 343–355 (2004)
Zhao, J., Yu, P.L.H.: A note on variational Bayesian factor analysis. Neural Netw. **22**, 988–997 (2009)