Journal of Multivariate Analysis 143 (2016) 398–413

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

# Extending mixtures of factor models using the restricted multivariate skew-normal distribution

Tsung-I Lin [a,b,*], Geoffrey J. McLachlan [c], Sharon X. Lee [c]

[a] Institute of Statistics, National Chung Hsing University, Taichung 402, Taiwan
[b] Department of Public Health, China Medical University, Taichung 404, Taiwan
[c] Department of Mathematics, University of Queensland, St Lucia, 4072, Australia

ARTICLE INFO

*Article history:*
Received 27 January 2014
Available online 19 October 2015

*AMS subject classifications:*
62H25
62H30
65C60

*Keywords:*
Clustering
Data reduction
ECM algorithm
Factor analyzer
rMSN distribution
Skewness

ABSTRACT

The mixture of factor analyzers (MFA) model provides a powerful tool for analyzing high-dimensional data as it can reduce the number of free parameters through its factor-analytic representation of the component covariance matrices. This paper extends the MFA model to incorporate a restricted version of the multivariate skew-normal distribution for the latent component factors, called mixtures of skew-normal factor analyzers (MSNFA). The proposed MSNFA model allows us to relax the need of the normality assumption for the latent factors in order to accommodate skewness in the observed data. The MSNFA model thus provides an approach to model-based density estimation and clustering of high-dimensional data exhibiting asymmetric characteristics. A computationally feasible Expectation Conditional Maximization (ECM) algorithm is developed for computing the maximum likelihood estimates of model parameters. The potential of the proposed methodology is exemplified using both real and simulated data.



## 1. Introduction

Factor analysis (FA) is a popular technique for explaining the covariance relationships among many variables through a fewer number of unobservable random quantities known as *latent factors*. Finite mixture models (FMMs) have been widely used as flexible means to model heterogeneous data, in particular, for density estimation and clustering. There are a number of monographs on mixture models; see, for example, [14,19,26,38,46,50,57,68] and the references contained therein. Mixtures of factor analyzers (MFAs), introduced by Ghahramani and Hinton [28], 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; see also [51]. McLachlan et al. [48,52] exploited the MFA model for clustering microarray gene-expression profiles. For data with clusters having longer than the normal tails, McLachlan et al. [47] adopted the family of multivariate $t$-distributions for the component factors and errors to establish a robust extension of MFA. More recently, Baek et al. [9] proposed mixtures of common factor analyzers (MCFA) in which the factors are taken to have a common distribution before their transformation to be white noise. A robust version of MCFA using $t$-component distributions, called mixtures of common factor $t$ analyzers (MCtFA), was subsequently provided by Baek et al. [8]. Wang [72,73] extended the MCFA and MCtFA approaches to accommodate high-dimensional data with possibly missing values. Bayesian treatments of the MFA model have been investigated by Ghahramani and Beal [27] via a variational approximation and Utsugi and Kumagai [70] using the Gibbs sampler and a deterministic algorithm.

* Corresponding author at: Institute of Statistics, National Chung Hsing University, Taichung 402, Taiwan.
*E-mail address:* tilin@nchu.edu.tw (T.-I. Lin).

http://dx.doi.org/10.1016/j.jmva.2015.09.025
0047-259X/

For computational convenience and mathematical tractability, component errors and latent factors in the traditional MFA model are routinely assumed to follow multivariate normal distributions. However, in many applied problems, the data to be analyzed may contain a group or groups of observations whose distributions are moderately or severely skewed. Just like other normal-based mixture models, a slight deviation from normality may seriously affect the estimates of mixture parameters and/or lead to spurious groups, subsequently misleading inference from the data. Wall et al. [71] conducted several simulation studies to explore the influence of non-normal latent factors in the estimation of parameters.

In recent years, there has been growing interest in studying mixtures of skew-normal distributions [37,40], both in the univariate and multivariate cases, as a more general tool for handling heterogeneous data involving asymmetric behavior across sub-populations. Pyne et al. [65] proposed mixtures of multivariate skew-normal and $t$-distributions based on a restricted variant of the skew-elliptical family of distributions of Sahu et al. [66], which we shall refer to as the restricted multivariate skew-normal (rMSN) distribution. The use of "restricted" was adopted by Lee and McLachlan [33] since it is obtained by imposing the restriction that the $p$ latent skewing variables are all equal in the form of the class of skew elliptical distributions proposed by Sahu et al. [66]. The latter class without this restriction was referred to as "unrestricted". The rMSN distribution is equivalent to the skew normal distribution proposed by Azzalini and Dalla Valle [7]. Lee and McLachlan [34] gave a systematic overview of various existing multivariate skew distributions and clarified their conditioning-type and convolution-type representations. Also, Lee and McLachlan [35] have provided the `EMMIXuskew` package, which implements a closed-form expectation–maximization (EM) algorithm for computing the maximum likelihood (ML) estimates of the parameters for mixtures of unrestricted skew-normal and skew-$t$ distributions.

There have been a few different proposals of mixtures of skew factor models in the literature, see, for instance, mixtures of shifted asymmetric Laplace factor analyzers of Franczak et al. [24], mixtures of generalized hyperbolic factor analyzers of Tortora et al. [69], and mixtures of skew-$t$ factor analyzers (MSTFA) of Murray et al. [61]. An unrestricted version of MSTFA was considered by Murray et al. [62]. Notice that the form of the skew-$t$ distribution used in Murray et al. [61] arises as a special case of the generalized hyperbolic distribution [10], called the generalized hyperbolic skew-$t$ (GHST) distribution. More recently, Murray et al. [63] have put forward a skew version of the MCFA model in which the common factors follow the GHST distribution. The model is henceforth referred to as mixtures of common skew-$t$ factor analyzers (MCSTFA). We should emphasize that the GHST distribution differs from the restricted skew-$t$ distribution in a number of ways, such as different behavior in its tails, for example in the univariate case, with one polynomial and the other exponential [1]. Also, it does not become a skew normal distribution as a limiting case [36].

In this paper, we propose mixtures of skew-normal factor analyzers (MSNFA) where the latent component factors are assumed to follow the family of rMSN distributions in an attempt to model the data adequately in the presence of skewed sub-populations. The proposed model, which is a generalization of the MFA model, can be viewed as a novel approach to achieving dimensionality reduction and representing appropriately non-normal data. ML estimates of the parameters in the model can be computed via the closed-form EM implementations [16,58], and the estimated factor scores can be obtained as by-products within the estimation procedure. The asymptotic covariance matrix of the estimated mixture parameters is obtained by inverting an approximation to the observed information matrix [30].

The rest of the paper is organized as follows. In Section 2, we establish notation and provide a preliminary account of the rMSN distribution. In Section 3, we briefly present the formulation of the skew-normal factor analysis (SNFA) model and study its related properties. Section 4 extends the work to the MSNFA model and presents an EM-type algorithm for obtaining the ML estimates of model parameters. Section 5 describes some practical issues, including the specification of starting values, the stopping rule, model selection and two indices for performance evaluation. The proposed methodology is illustrated through both real and simulated data in Section 6. Some concluding remarks are given in Section 7.

## 2. The restricted multivariate skew-normal distribution

We begin with a brief review of the rMSN distribution and a study of some essential properties. A unification of families of MSN distributions and several variants and extensions can be found in [2,4]. To establish notation, let $\phi_p(\cdot\,;\boldsymbol{\mu},\boldsymbol{\Sigma})$ be the probability density function (pdf) corresponding to $N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$, a $p$-dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and variance–covariance matrix $\boldsymbol{\Sigma}$, and $\Phi(\cdot)$ the cumulative distribution function (cdf) of the standard normal distribution. Further, let $TN(\mu,\sigma^2;(a,b))$ denote the truncated normal distribution for $N(\mu,\sigma^2)$ lying within a truncated interval $(a,b)$.

Following Lee and McLachlan [33], a $p\times 1$ random vector $\boldsymbol{X}$ is said to follow a rMSN distribution with location vector $\boldsymbol{\mu}$, dispersion matrix $\boldsymbol{\Sigma}$ and skewness vector $\boldsymbol{\lambda}$, denoted by $\boldsymbol{X}\sim rSN_p(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda})$, if it can be represented as

$$\boldsymbol{X}=\boldsymbol{\lambda}|U_1|+\boldsymbol{U}_2,\qquad U_1\perp\boldsymbol{U}_2, \tag{1}$$

where $U_1\sim N(0,1)$, $\boldsymbol{U}_2\sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$ and the symbol '$\perp$' indicates independence. Letting $W=|U_1|$, a two-level hierarchical representation of (1) is

$$\begin{aligned}\boldsymbol{X}\mid(W=w)&\sim N_p(\boldsymbol{\mu}+\boldsymbol{\lambda}w,\boldsymbol{\Sigma}),\\ W&\sim TN(0,1;(0,\infty))\,.\end{aligned} \tag{2}$$

For computing the moments of $W$, we use the following proposition.

**Proposition 1.** *Let $W \sim TN(\mu, \sigma^2; (0, \infty))$. The density of $W$ is*

$$f(w) = \frac{\phi(w; \mu, \sigma^2)}{\Phi(\mu/\sigma)} I(w > 0),$$

*where $I(\cdot)$ is an indicator function. For positive integer $k$, the moments of $W$ are given by*

$$E(W) = \mu + \sigma \frac{\phi(\mu/\sigma)}{\Phi(\mu/\sigma)} \quad \text{for } k = 1,$$
$$E(W^k) = (k-1)\sigma^2 E(W^{k-2}) + \mu E(W^{k-1}) \quad \text{for } k \geq 2.$$

The pdf of $\boldsymbol{X}$, expressed as a product of a multivariate normal density and a univariate normal cdf, is given by

$$f(\boldsymbol{x}) = 2\phi_p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Omega})\Phi(\xi/\sigma), \tag{3}$$

where $\boldsymbol{\Omega} = \boldsymbol{\Sigma} + \boldsymbol{\lambda}\boldsymbol{\lambda}^\top$, $\xi = \boldsymbol{\lambda}^\top \boldsymbol{\Omega}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$, and $\sigma^2 = (1 + \boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda})^{-1} = 1 - \boldsymbol{\lambda}^\top \boldsymbol{\Omega}^{-1}\boldsymbol{\lambda}$. The rMSN distribution falls into the class of fundamental skew-normal (FUSN) distribution [3]. In addition, it can be treated as a simplified version of Sahu et al. [66] or a modification of the traditional version of Azzalini and Dalla Valle [5,7] via a reparameterization. The version allows us to develop computationally feasible EM-type algorithms for parameter estimation in SNFA and MSNFA models.

Using Proposition 1 and the law of iterative expectations, it follows from (1) that the mean and covariance matrix of $\boldsymbol{X}$ are

$$E(\boldsymbol{X}) = \boldsymbol{\mu} + c\boldsymbol{\lambda} \quad \text{and} \quad \text{cov}(\boldsymbol{X}) = \boldsymbol{\Sigma} + (1 - c^2)\boldsymbol{\lambda}\boldsymbol{\lambda}^\top, \tag{4}$$

where $c = \sqrt{2/\pi}$. The higher order moments of $\boldsymbol{X}$ can be derived from the moment generating function (mgf) given in the following proposition.

**Proposition 2.** *If $\boldsymbol{X} \sim rSN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$, then the mgf of $\boldsymbol{X}$ is*

$$M_{\boldsymbol{X}}(\boldsymbol{t}) = 2\exp\left(\boldsymbol{t}^\top \boldsymbol{\mu} + \frac{1}{2}\boldsymbol{t}^\top \boldsymbol{\Omega}\boldsymbol{t}\right)\Phi(\boldsymbol{\lambda}^\top \boldsymbol{t}), \quad \boldsymbol{t} \in \mathbb{R}^p.$$

The following result shows an appealing closure property of the rMSN distribution under affine transformations, which is useful for later methodological developments.

**Proposition 3.** *Let $\boldsymbol{X} \sim rSN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$. For any full rank matrix $\boldsymbol{L} \in \mathbb{R}^{q\times p}$ $(1 \leqslant q \leqslant p)$, the distribution of the linear transformation $\boldsymbol{LX}$ is*

$$\boldsymbol{LX} \sim rSN_q(\boldsymbol{L}\boldsymbol{\mu}, \boldsymbol{L}\boldsymbol{\Sigma}\boldsymbol{L}^\top, \boldsymbol{L}\boldsymbol{\lambda}).$$

The proof follows directly by applying Proposition 2 to the transformation $\boldsymbol{LX}$.

## 3. The skew-normal factor analysis model

### 3.1. The model

We consider a generalization of the traditional FA model, namely the SNFA model, in which the hidden factors are assumed to follow an rMSN distribution within the family defined by (1). Suppose that $\boldsymbol{Y} = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n\}$ is a random sample of $n$ $p$-dimensional observations. The SNFA model can be written as

$$\begin{cases} \boldsymbol{Y}_j = \boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{U}_j + \boldsymbol{\varepsilon}_j, & \boldsymbol{U}_j \perp \boldsymbol{\varepsilon}_j, \\ \boldsymbol{U}_j \overset{\text{iid}}{\sim} rSN_q(-c\boldsymbol{\Delta}^{-1/2}\boldsymbol{\lambda}, \boldsymbol{\Delta}^{-1}, \boldsymbol{\Delta}^{-1/2}\boldsymbol{\lambda}), & \boldsymbol{\varepsilon}_j \overset{\text{iid}}{\sim} N_p(\boldsymbol{0}, \boldsymbol{D}), \end{cases} \tag{5}$$

for $j = 1, \ldots, n$, where $\boldsymbol{\mu}$ is a $p$-dimensional location vector, $\boldsymbol{B}$ is a $p \times q$ matrix of factor loadings, $\boldsymbol{U}_j$ is a $q$-dimensional vector $(q < p)$ of latent variables called *factors*, $\boldsymbol{\varepsilon}_j$ is a $p$-dimensional vector of errors, and $\boldsymbol{\Delta} = \boldsymbol{I}_q + (1 - c^2)\boldsymbol{\lambda}\boldsymbol{\lambda}^{\mathrm{T}}$ is a scale matrix. The elements of the factor loadings $\boldsymbol{B}$ indicate the strength of dependence of each variable on each factor. Moreover, $\boldsymbol{D}$ is a positive diagonal matrix and $\boldsymbol{I}_q$ stands for an identity matrix of order $q$.

Under model (5), an appealing property is that

$$\mathrm{E}(\boldsymbol{U}_j) = \boldsymbol{0} \quad \text{and} \quad \text{cov}(\boldsymbol{U}_j) = \boldsymbol{I}_q. \tag{6}$$

Hence, the chosen distributional assumption for $\boldsymbol{U}_j$ makes the factor score estimates of FA and SNFA models comparable. By Proposition 3, we can deduce that

$$\boldsymbol{Y}_j \sim rSN_p(\boldsymbol{\mu} - c\boldsymbol{\alpha}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}),$$

where $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{\Delta}^{-1}\boldsymbol{B}^{\top} + \boldsymbol{D}$ and $\boldsymbol{\alpha} = \boldsymbol{B}\boldsymbol{\Delta}^{-1/2}\boldsymbol{\lambda}$. Clearly, the marginal distribution of $\boldsymbol{Y}_j$ belongs to the family of rMSN distributions in which the skewness parameter $\boldsymbol{\alpha}$ depends both on $\boldsymbol{B}$ and $\boldsymbol{\lambda}$. It follows immediately from (4) that

$$E(\boldsymbol{Y}_j) = \boldsymbol{\mu} \quad \text{and} \quad \text{cov}(\boldsymbol{Y}_j) = \boldsymbol{B}\boldsymbol{B}^{\top} + \boldsymbol{D}. \tag{7}$$

Another interesting feature of this model is that the parameter estimates of $\boldsymbol{\mu}$, $\boldsymbol{B}$ and $\boldsymbol{D}$ can be used to recover the sample mean and sample covariance for both FA and SNFA models. The important characteristics (6) and (7) were not considered neither in Montanari and Viroli [60] nor in other developments in the literature.

### 3.2. Identifiability issues

For a hidden dimensionality $q > 1$, there is an identifiability issue associated with the rotational invariance of the factor loading matrix $\boldsymbol{B}$. For any orthogonal matrix $\boldsymbol{P}$ of order $q$, model (5) still holds when $\boldsymbol{B}$ is replaced by $\boldsymbol{B}\boldsymbol{P}$ and the latent $\boldsymbol{U}_j$ is changed to $\boldsymbol{P}^{\top}\boldsymbol{U}_j$. Moreover, such an orthogonal transformation will leave the covariance matrix in (7) invariant since $\boldsymbol{B}\boldsymbol{P}(\boldsymbol{B}\boldsymbol{P})^{\top} = \boldsymbol{B}\boldsymbol{B}^{\top}$.

To circumvent this identifiability problem (rotational indeterminacy), one of the most commonly used techniques is to constrain the loading matrix $\boldsymbol{B}$ so that the upper-right triangle is zero and the diagonal entries are strictly positive, see, for example, Fokoué and Titterington [21] and Lopes and West [42]. This means that $q(q-1)/2$ elements of $\boldsymbol{B}$ are constrained. The number of free parameters to be estimated is $m = p(q+2) + q - q(q-1)/2$.

The mixture model itself poses another identifiability problem raised by relabeling of components. More precisely, the likelihood is invariant under a permutation of the class labels in parameter vectors, and thus a label switching problem can occur when some labels of the mixture classes permute [50]. However, the switching of class labels is not a concern with the use of the EM algorithm and its variants to compute the ML estimates.

## 4. Mixture of restricted skew-normal factors

### 4.1. Model formulation

Let $\boldsymbol{Y}_j = (Y_{j1}, \ldots, Y_{jp})^{\top}$ be a $p$-dimensional vector of $p$ feature variables ($j = 1, \ldots, n$), where $\boldsymbol{Y}_j$ comes from a heterogeneous population with a finite number, say $g$, of groups. To denote which component $\boldsymbol{Y}_j$ belongs in this finite mixture framework, we introduce the latent membership-indicator vectors, $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$. Here $Z_{ij} = (\boldsymbol{Z}_j)_i$ is one or zero, according to whether $\boldsymbol{Y}_j$ belongs or does not belong to the $i$th component ($i = 1, \ldots, g$; $j = 1, \ldots, n$). Accordingly, we have

$$\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n \overset{\text{iid}}{\sim} \mathcal{M}(1; \pi_1, \ldots, \pi_g),$$

where the pdf of the multinomial variate $\boldsymbol{Z}_j$ is given by

$$f(\boldsymbol{z}_j; \boldsymbol{\pi}) = \pi_1^{z_{1j}} \pi_2^{z_{2j}} \cdots \pi_g^{z_{gj}}, \quad \text{for } j = 1, \ldots, n,$$

and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)^{\top}$ subject to $\sum_{i=1}^{g} \pi_i = 1$.

The MSNFA model is a generalization of the MFA model by postulating a mixture of $g$ SNFA sub-models for the distribution of $\boldsymbol{Y}_j$. We consider the use of the MSNFA model in an attempt to accommodate skewness arising frequently in high-dimensional data without performing transformation.

Given $Z_{ij} = 1$, each $\boldsymbol{Y}_j$ can be modeled as

$$\boldsymbol{Y}_j = \boldsymbol{\mu}_i + \boldsymbol{B}_i\boldsymbol{U}_{ij} + \boldsymbol{\epsilon}_{ij}, \quad \text{with probability } \pi_i \ (i = 1, \ldots, g), \tag{8}$$

for $j = 1, \ldots, n$, where the factors $\boldsymbol{U}_{i1}, \ldots, \boldsymbol{U}_{in} \overset{\text{iid}}{\sim} rSN_q(-c\boldsymbol{\Delta}_i^{-1/2}\boldsymbol{\lambda}_i, \boldsymbol{\Delta}_i^{-1}, \boldsymbol{\Delta}_i^{-1/2}\boldsymbol{\lambda}_i)$, independently of the errors $\boldsymbol{\epsilon}_{ij}$, which are distributed independently as $N_p(\boldsymbol{0}, \boldsymbol{D}_i)$, where $\boldsymbol{\Delta}_i = \boldsymbol{I}_q + (1-c^2)\boldsymbol{\lambda}_i\boldsymbol{\lambda}_i^{\top}$ and $\boldsymbol{D}_i$ is a positive diagonal matrix.

From model (8), the marginal pdf of $\boldsymbol{Y}_j$ is

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

where $\psi(\boldsymbol{y}_j; \boldsymbol{\theta}_i)$ is the pdf of rMSN distribution defined in (3), $\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \boldsymbol{B}_i, \boldsymbol{D}_i, \boldsymbol{\lambda}_i)$ is composed of the unknown parameters of the $i$th mixture component, and $\boldsymbol{\Theta} = \{\pi_1, \ldots, \pi_{g-1}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g\}$ represents all the unknown parameters of the mixture model. Given a set of $n$ observations $\boldsymbol{y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$, ML estimation can be undertaken by maximizing the log-likelihood function of $\boldsymbol{\Theta}$, given by

$$\ell(\boldsymbol{\Theta}; \boldsymbol{y}) = \sum_{j=1}^{n} \log \left( \sum_{i=1}^{g} \pi_i \psi(\boldsymbol{y}_j; \boldsymbol{\theta}_i) \right). \tag{9}$$

Unfortunately, it is not straightforward to derive explicit analytical solutions for ML estimator of $\boldsymbol{\Theta}$. To cope with this obstacle, one usually resorts to the EM-type algorithm [16], which is a popular iterative device for ML estimation in models involving latent variables or missing data.

Under model (8), it can be shown that

$$\boldsymbol{Y}_j \mid (Z_{ij} = 1) \sim rSN_p(\boldsymbol{\mu}_i - c\boldsymbol{\alpha}_i, \boldsymbol{\Sigma}_i, \boldsymbol{\alpha}_i), \tag{10}$$

where $\boldsymbol{\Sigma}_i = \boldsymbol{B}_i\boldsymbol{\Delta}_i^{-1}\boldsymbol{B}_i^\top + \boldsymbol{D}_i$ and $\boldsymbol{\alpha}_i = \boldsymbol{B}_i\boldsymbol{\Delta}_i^{-1/2}\boldsymbol{\lambda}_i$. To facilitate the derivation of our inference procedure, we adopt the following scaling transformation:

$$\tilde{\boldsymbol{B}}_i \overset{\triangle}{=} \boldsymbol{B}_i\boldsymbol{\Delta}_i^{-1/2} \quad \text{and} \quad \tilde{\boldsymbol{U}}_{ij} \overset{\triangle}{=} \boldsymbol{\Delta}_i^{1/2}\boldsymbol{U}_j.$$

Based on (2) and (10), a four-level hierarchical representation of model (8) is

$$\begin{aligned}
\boldsymbol{Y}_j \mid (\tilde{\boldsymbol{u}}_{ij}, w_j, Z_{ij} = 1) &\sim N_p(\boldsymbol{\mu}_i + \tilde{\boldsymbol{B}}_i\tilde{\boldsymbol{u}}_{ij}, \boldsymbol{D}_i),\\
\tilde{\boldsymbol{U}}_{ij} \mid (w_j, Z_{ij} = 1) &\sim N_q\big((w_j - c)\boldsymbol{\lambda}_i, \boldsymbol{I}_q\big),\\
W_j \mid (Z_{ij} = 1) &\sim TN\big(0, 1; (0, \infty)\big),\\
\boldsymbol{Z}_j &\sim \mathcal{M}(1; \pi_1, \ldots, \pi_g).
\end{aligned} \tag{11}$$

In the EM framework, the augmented quadruples $\{\boldsymbol{Y}_j, \boldsymbol{Z}_j, \tilde{\boldsymbol{U}}_{ij}, w_j\}_{j=1}^n$ are referred to as the complete data. Using Bayes' Theorem, it suffices to show that

$$\begin{aligned}
&\tilde{\boldsymbol{U}}_{ij} \mid (Z_{ij} = 1, w_j, \boldsymbol{y}_j) \sim N_q\big(\boldsymbol{q}_{ij}, \boldsymbol{C}_i\big),\\
&W_j \mid (Z_{ij} = 1, \boldsymbol{y}_j) \sim TN\big(a_{ij}, 1 - \boldsymbol{\alpha}_i^\top\boldsymbol{\Omega}_i^{-1}\boldsymbol{\alpha}_i; (0, \infty)\big),
\end{aligned} \tag{12}$$

where $\boldsymbol{q}_{ij} = \boldsymbol{C}_i[\boldsymbol{v}_{ij} + \boldsymbol{\lambda}_i(w_j - c)]$, $\boldsymbol{v}_{ij} = \tilde{\boldsymbol{B}}_i^\top\boldsymbol{D}_i^{-1}(\boldsymbol{y}_j - \boldsymbol{\mu}_i)$, $\boldsymbol{C}_i = (\boldsymbol{I}_q + \tilde{\boldsymbol{B}}_i^\top\boldsymbol{D}_i^{-1}\tilde{\boldsymbol{B}}_i)^{-1}$, $a_{ij} = \boldsymbol{\alpha}_i^\top\boldsymbol{\Omega}_i^{-1}(\boldsymbol{y}_j - \boldsymbol{\mu}_i + c\boldsymbol{\alpha}_i)$ and $\boldsymbol{\Omega}_i = \boldsymbol{\Sigma}_i + \boldsymbol{\alpha}_i\boldsymbol{\alpha}_i^\top$. As an immediate consequence, we establish the following proposition, which is crucial for the calculation of some conditional expectations involved in the proposed ECM algorithm.

**Proposition 4.** *Given the hierarchical representation (12), we have the following (the symbol "$\mid \cdots$" denotes conditioning on $Z_{ij} = 1$ and $\boldsymbol{Y}_j = \boldsymbol{y}_j$):*

(a) *The conditional expectation of $Z_{ij}$ given $\boldsymbol{Y}_j = \boldsymbol{y}_j$ is*

$$E(Z_{ij} \mid \boldsymbol{y}_j) = \frac{\pi_i\psi(\boldsymbol{y}_j; \boldsymbol{\theta}_i)}{f(\boldsymbol{y}_j; \boldsymbol{\Theta})}. \tag{13}$$

(b) *Some specific conditional expectations related to $W_j$ and $\boldsymbol{U}_j$ are*

$$E(W_j \mid \cdots) = (1 - \boldsymbol{\alpha}_i^\top\boldsymbol{\Omega}_i^{-1}\boldsymbol{\alpha}_i)^{1/2}\left(A_{ij} + \frac{\phi(A_{ij})}{\Phi(A_{ij})}\right), \tag{14}$$

$$E(W_j^2 \mid \cdots) = (1 - \boldsymbol{\alpha}_i^\top\boldsymbol{\Omega}_i^{-1}\boldsymbol{\alpha}_i)\left[1 + A_{ij}\left(A_{ij} + \frac{\phi(A_{ij})}{\Phi(A_{ij})}\right)\right], \tag{15}$$

$$E(\tilde{\boldsymbol{U}}_{ij} \mid \cdots) = \boldsymbol{C}_i\big(\boldsymbol{v}_{ij} + \boldsymbol{\lambda}_i(E(W_j \mid \cdots) - c)\big), \tag{16}$$

$$E(W_j\tilde{\boldsymbol{U}}_{ij} \mid \cdots) = \boldsymbol{C}_i\Big\{\boldsymbol{v}_{ij}E(W_j \mid \cdots) + \boldsymbol{\lambda}_i\big[E(W_j^2 \mid \cdots) - cE(W_j \mid \cdots)\big]\Big\}, \tag{17}$$

*and*

$$E(\tilde{\boldsymbol{U}}_{ij}\tilde{\boldsymbol{U}}_{ij}^\top \mid \cdots) = \big\{\boldsymbol{I}_q + E(\tilde{\boldsymbol{U}}_{ij} \mid \cdots)\boldsymbol{v}_{ij}^\top + \big[E(W_j\tilde{\boldsymbol{U}}_{ij} \mid \cdots) - cE(\tilde{\boldsymbol{U}}_{ij} \mid \cdots)\big]\boldsymbol{\lambda}_i^\top\big\}\boldsymbol{C}_i, \tag{18}$$

*where $A_{ij} = (1 - \boldsymbol{\alpha}_i^\top\boldsymbol{\Omega}_i^{-1}\boldsymbol{\alpha}_i)^{-1/2}a_{ij}$.*

### 4.2. ML estimation via the ECM algorithm

The EM algorithm has several attractive features such as simplicity of implementation and monotonic convergence properties. However, it cannot be directly applied for ML estimation of the MSNFA model because the M-step is difficult to compute. To proceed further, we exploit a variant of the EM algorithm, called the ECM algorithm [58], which is easy to implement and more broadly applicable than EM. The key feature of ECM is to replace the M-step of the EM algorithm with a sequence of simpler constrained or conditional maximization (CM) steps. Moreover, it shares all appealing features of EM and can show faster convergence in terms of number of iterations or total CPU time.

For notational convenience, let $\boldsymbol{u} = (\boldsymbol{u}_1^\top, \ldots, \boldsymbol{u}_n^\top)^\top$, $\boldsymbol{w} = (w_1, \ldots, w_n)^\top$ and $\boldsymbol{z} = (\boldsymbol{z}_1^\top, \ldots, \boldsymbol{z}_n^\top)^\top$, which are treated as missing data in the EM framework. According to (11), the log-likelihood function of $\boldsymbol{\Theta}$that can be formed from the complete-data vector $\boldsymbol{y}_c = (\boldsymbol{y}^\top, \boldsymbol{u}^\top, \boldsymbol{w}^\top, \boldsymbol{z}^\top)^\top$, aside from additive terms not involving the parameters, is

$$\ell_c(\boldsymbol{\Theta}; \boldsymbol{y}_c) = \sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij}\left\{\log \pi_i - \frac{1}{2}\left[\log|\boldsymbol{D}_i| + \operatorname{tr}\left(\boldsymbol{D}_i^{-1}\boldsymbol{\Upsilon}_{ij}\right) + (w_j - c)^2\boldsymbol{\lambda}_i^\top\boldsymbol{\lambda}_i - 2(w_j - c)\boldsymbol{\lambda}_i^\top\tilde{\boldsymbol{u}}_{ij}\right]\right\}, \tag{19}$$

where $\boldsymbol{\Upsilon}_{ij} = (\boldsymbol{y}_j - \boldsymbol{\mu}_i - \tilde{\boldsymbol{B}}_i\tilde{\boldsymbol{u}}_{ij})(\boldsymbol{y}_j - \boldsymbol{\mu}_i - \tilde{\boldsymbol{B}}_i\tilde{\boldsymbol{u}}_{ij})^\top$.

In the E-step of the algorithm, we need to calculate the $Q$-function, denoted by $Q(\boldsymbol{\Theta}; \hat{\boldsymbol{\Theta}}^{(k)})$, which is the conditional expectation of (19) given the observed data $\boldsymbol{y}$ and the current estimate $\hat{\boldsymbol{\Theta}}^{(k)}$. To evaluate the $Q$-function, the necessary conditional expectations include $\hat{z}_{ij}^{(k)} = E(Z_{ij} \mid \boldsymbol{y}_j, \hat{\boldsymbol{\Theta}}^{(k)})$, $\hat{w}_{1ij}^{(k)} = E(W_j \mid Z_{ij} = 1, \boldsymbol{y}_j, \hat{\boldsymbol{\Theta}}^{(k)})$, $\hat{w}_{2ij}^{(k)} = E(W_j^2 \mid Z_{ij} = 1, \boldsymbol{y}_j, \hat{\boldsymbol{\Theta}}^{(k)})$, $\hat{\boldsymbol{\kappa}}_{ij}^{(k)} = E(W_j\tilde{\boldsymbol{U}}_{ij} \mid \boldsymbol{y}_j, \hat{\boldsymbol{\Theta}}^{(k)})$, $\hat{\boldsymbol{\eta}}_{ij}^{(k)} = E(\tilde{\boldsymbol{U}}_{ij} \mid \boldsymbol{y}_j, \hat{\boldsymbol{\Theta}}^{(k)})$ and $\hat{\boldsymbol{\Psi}}_{ij}^{(k)} = E(\tilde{\boldsymbol{U}}_{ij}\tilde{\boldsymbol{U}}_{ij}^\top \mid \boldsymbol{y}_j, \hat{\boldsymbol{\Theta}}^{(k)})$. Therefore, we have

$$Q(\boldsymbol{\Theta}; \hat{\boldsymbol{\Theta}}^{(k)}) = \sum_{i=1}^{g}\sum_{j=1}^{n} \hat{z}_{ij}^{(k)}\left\{\log \pi_i - \frac{1}{2}\left[\log|\boldsymbol{D}_i| + \operatorname{tr}\left(\boldsymbol{D}_i^{-1}\boldsymbol{\Upsilon}_{ij}^{(k)}\right) + \hat{h}_{ij}^{(k)}\boldsymbol{\lambda}_i^\top\boldsymbol{\lambda}_i - 2\boldsymbol{\lambda}_i^\top\hat{\boldsymbol{\zeta}}_{ij}^{(k)}\right]\right\}, \tag{20}$$

where $\hat{h}_{ij}^{(k)} = \hat{w}_{2ij}^{(k)} - 2c\hat{w}_{1ij}^{(k)} + c^2$, $\boldsymbol{\zeta}_{ij}^{(k)} = \hat{\boldsymbol{\kappa}}_{ij}^{(k)} - c\hat{\boldsymbol{\eta}}_{ij}^{(k)}$ and

$$\boldsymbol{\Upsilon}_{ij}^{(k)} = (\boldsymbol{y}_j - \boldsymbol{\mu}_i - \tilde{\boldsymbol{B}}_i\hat{\boldsymbol{\eta}}_{ij}^{(k)})(\boldsymbol{y}_j - \boldsymbol{\mu}_i - \tilde{\boldsymbol{B}}_i\hat{\boldsymbol{\eta}}_{ij}^{(k)})^\top + \tilde{\boldsymbol{B}}_i(\hat{\boldsymbol{\Psi}}_{ij}^{(k)} - \hat{\boldsymbol{\eta}}_{ij}^{(k)}\hat{\boldsymbol{\eta}}_{ij}^{(k)^\top})\tilde{\boldsymbol{B}}_i^\top, \tag{21}$$

which involves free parameters $\boldsymbol{\mu}_i$ and $\tilde{\boldsymbol{B}}_i$ for $i = 1, \ldots, g$.

In summary, the implementation of the ECM algorithm proceeds as follows:

E-step: Given $\boldsymbol{\Theta} = \hat{\boldsymbol{\Theta}}^{(k)}$, compute $\hat{z}_{ij}^{(k)}$, $\hat{w}_{1ij}^{(k)}$, $\hat{w}_{2ij}^{(k)}$, $\hat{\boldsymbol{\kappa}}_{ij}^{(k)}$, $\hat{\boldsymbol{\eta}}_{ij}^{(k)}$ and $\hat{\boldsymbol{\Psi}}_{ij}^{(k)}$ by using (13)–(18), for $i = 1, \ldots, g$ and $j = 1, \ldots, n$.

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

CM-step 2: Update $\hat{\boldsymbol{\mu}}_i^{(k)}$ by maximizing (20) over $\boldsymbol{\mu}_i$, which gives

$$\hat{\boldsymbol{\mu}}_i^{(k+1)} = \frac{1}{\hat{n}_i^{(k)}}\sum_{j=1}^{n}\hat{z}_{ij}^{(k)}\left(\boldsymbol{y}_j - \hat{\tilde{\boldsymbol{B}}}_i^{(k)}\hat{\boldsymbol{\eta}}_{ij}^{(k)}\right).$$

CM-step 3: Fix $\boldsymbol{\mu}_i = \hat{\boldsymbol{\mu}}_i^{(k+1)}$, update $\tilde{\boldsymbol{B}}_i^{(k)}$ by maximizing (20) over $\tilde{\boldsymbol{B}}_i$, which gives

$$\hat{\tilde{\boldsymbol{B}}}_i^{(k+1)} = \sum_{j=1}^{n}\hat{z}_{ij}^{(k)}\left[(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_i^{(k+1)})\hat{\boldsymbol{\eta}}_{ij}^{(k)^\top}\right]\left(\sum_{j=1}^{n}\hat{z}_{ij}^{(k)}\hat{\boldsymbol{\Psi}}_{ij}^{(k)}\right)^{-1}.$$

CM-step 4: Fix $\boldsymbol{\mu} = \hat{\boldsymbol{\mu}}_i^{(k+1)}$ and $\tilde{\boldsymbol{B}}_i = \hat{\tilde{\boldsymbol{B}}}_i^{(k+1)}$, update $\hat{\boldsymbol{D}}_i^{(k)}$ by maximizing (20) over $\boldsymbol{D}_i$, which leads to

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

where $\hat{\boldsymbol{\Upsilon}}_{ij}^{(k)}$ is $\boldsymbol{\Upsilon}_{ij}^{(k)}$ in (21) with $(\boldsymbol{\mu}_i, \tilde{\boldsymbol{B}}_i)$ replaced by $(\hat{\boldsymbol{\mu}}_i^{(k+1)}, \hat{\tilde{\boldsymbol{B}}}_i^{(k+1)})$, respectively.

CM-step 5: Update $\hat{\boldsymbol{\lambda}}_i^{(k)}$ by maximizing (20) over $\boldsymbol{\lambda}_i$, which gives

$$\hat{\boldsymbol{\lambda}}_i^{(k+1)} = \frac{\sum_{j=1}^{n}\hat{z}_{ij}^{(k)}\hat{\boldsymbol{\zeta}}_{ij}^{(k)}}{\sum_{j=1}^{n}\hat{z}_{ij}^{(k)}\hat{h}_{ij}^{(k)}}.$$

The E- and CM-steps are alternated repeatedly until a suitable convergence rule is satisfied, e.g., the difference in successive values of the log-likelihood is less than a tolerance value. Upon convergence, the ML estimate of $\boldsymbol{\Theta}$ is denoted by $\hat{\boldsymbol{\Theta}} = \{\hat{\pi}_i, \hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{B}}_i, \hat{\boldsymbol{D}}_i, \hat{\boldsymbol{\lambda}}_i\}_{i=1}^{g}$, where $\hat{\boldsymbol{B}}_i = \hat{\tilde{\boldsymbol{B}}}_i\hat{\boldsymbol{\Delta}}_i^{1/2}$ and $\hat{\boldsymbol{\Delta}}_i = \boldsymbol{I}_q + (1 - c^2)\hat{\boldsymbol{\lambda}}_i\hat{\boldsymbol{\lambda}}_i^\top$. Consequently, the *conditional prediction* of factor scores is estimated by

$$\hat{\boldsymbol{U}}_j = \sum_{i=1}^{g}\hat{\pi}_i\hat{\boldsymbol{\Delta}}_i^{-1/2}\hat{\boldsymbol{\eta}}_{ij}, \tag{22}$$

where $\hat{\boldsymbol{\eta}}_{ij}$ can be calculated through (16) with $\boldsymbol{\Theta}$ evaluated at $\hat{\boldsymbol{\Theta}}$.

### 4.3. Computing standard errors via numerical differentiation

The asymptotic covariance matrix of the ML estimator can be approximated by the inverse of the observed information matrix; see Efron and Hinkley [18]. Specifically, the observed information matrix

$$\boldsymbol{I}(\hat{\boldsymbol{\Theta}}; \boldsymbol{y}) = -\frac{\partial^2 \ell(\boldsymbol{\Theta}; \boldsymbol{y})}{\partial \boldsymbol{\Theta} \partial \boldsymbol{\Theta}^\top}\Big|_{\boldsymbol{\Theta}=\hat{\boldsymbol{\Theta}}}$$

is a $m \times m$ matrix of the negative of second-order partial derivatives of the log-likelihood function with respect to each parameter, where $m$ is the number of distinct parameters in $\boldsymbol{\Theta}$. The asymptotic standard errors of $\hat{\boldsymbol{\Theta}}$ can be calculated by taking the square roots of the diagonal elements of $[\boldsymbol{I}(\hat{\boldsymbol{\Theta}}; \boldsymbol{y})]^{-1}$.

In the literature, there have been a few strategies recommended for efficiently computing $\boldsymbol{I}(\hat{\boldsymbol{\Theta}}; \boldsymbol{y})$ when implementing the ECM algorithm; see, for example, Louis [43] and Meng and Rubin [59]. A problem raising from these methods is that they require the second-order derivatives of the $Q$-function, which is rather cumbersome to calculate in FA models.

To approximate $\boldsymbol{I}(\hat{\boldsymbol{\Theta}}; \boldsymbol{y})$ numerically, Jamshidian [30] suggested using the central difference. Let $\boldsymbol{s}(\boldsymbol{\Theta}; \boldsymbol{y}) = \partial \ell(\boldsymbol{\Theta}; \boldsymbol{y})/\partial \boldsymbol{\Theta}$ be the score vector of $\ell(\boldsymbol{\Theta}; \boldsymbol{y})$ and $\boldsymbol{s}_c(\boldsymbol{\Theta}; \boldsymbol{y}) = \partial \ell_c(\boldsymbol{\Theta}; \boldsymbol{y}_c)/\partial \boldsymbol{\Theta}$ be the complete-data score of $\ell_c(\boldsymbol{\Theta}; \boldsymbol{y}_c)$. Moreover, it can be verified that $\boldsymbol{s}(\boldsymbol{\Theta}; \boldsymbol{y}) = E[\boldsymbol{s}_c(\boldsymbol{\Theta}; \boldsymbol{y}_c) \mid \boldsymbol{y}]$, see McLachlan and Peel [50]. Explicit expressions for the elements of $\boldsymbol{s}(\boldsymbol{\Theta}; \boldsymbol{y})$ are available upon request.

Let $\boldsymbol{G} = [\boldsymbol{g}_1 \mid \cdots \mid \boldsymbol{g}_m]$ be a $m \times m$ matrix with the $r$th column being

$$\boldsymbol{g}_r = \frac{\boldsymbol{s}(\hat{\boldsymbol{\Theta}} + h_r^* \boldsymbol{e}_r; \boldsymbol{y}) - \boldsymbol{s}(\hat{\boldsymbol{\Theta}} - h_r^* \boldsymbol{e}_r; \boldsymbol{y})}{2h_r^*}, \quad r = 1, \ldots, m,$$

where $\boldsymbol{e}_r$ is a unit vector corresponding to the $r$th element. The values of $h_r^*$ are small numbers chosen based on the scale of problem. In later data analysis, we use $h_r^* = \max(\eta, \eta|\hat{\boldsymbol{\Theta}}_r|)$ with $\hat{\boldsymbol{\Theta}}_r$ denoting the $r$th of element of $\hat{\boldsymbol{\Theta}}$, where values such as $\eta = 10^{-4}$ should be sufficiently small to approximate and large enough to avoid the roundoff error. Since $\boldsymbol{G}$ may not be symmetric, it is suggested to use $\tilde{\boldsymbol{I}}(\hat{\boldsymbol{\Theta}}; \boldsymbol{y}) = -(\boldsymbol{G} + \boldsymbol{G}^\top)/2$ to approximate $\boldsymbol{I}(\hat{\boldsymbol{\Theta}}; \boldsymbol{y})$.

## 5. Strategies for implementation

### 5.1. Initialization

As described in Section 4, the MSNFA parameters are estimated through the ECM algorithm. However, the EM-type algorithm has an intrinsic limitation that there is no guarantee of convergence to the global optimum [77]. For modeling multi-model distributions, the iterations may converge to a local maximum or to a saddle point. Sometimes, the quality of the final solution depends heavily on starting values. To cope with such potential problems, we recommend a simple way of obtaining suitable initial values for the ECM algorithm below.

1. Perform the $k$-means algorithm initialized with a random seed. Then, initialize the zero–one membership indicator $\hat{\boldsymbol{z}}_j^{(0)} = \{\hat{z}_{ij}^{(0)}\}_{i=1}^g$ according to the $k$-means clustering result. The initial values for the mixing proportions and component locations are then given by $\hat{\pi}_i^{(0)} = n^{-1} \sum_{j=1}^n \hat{z}_{ij}^{(0)}$ and $\hat{\boldsymbol{\mu}}_i^{(0)} = \sum_{j=1}^n \hat{z}_{ij}^{(0)} \boldsymbol{y}_j / \sum_{j=1}^n \hat{z}_{ij}^{(0)}$.
2. Subtract each observation from its initial cluster means. Then, do a FA fit to these $k$ "centralized samples" via the ML estimation (default) or the PCA method. The resulting estimates of factor loadings and error covariance matrices are taken as the initial values, namely $\hat{\boldsymbol{B}}_i^{(0)}$ and $\hat{\boldsymbol{D}}_i^{(0)}$ for $i = 1, \ldots, g$. Next, compute the corresponding factor scores of each cluster via the *conditional prediction* method such as (22). The initial values for the skewness parameters $\hat{\boldsymbol{\lambda}}_i^{(0)}$ are obtained by fitting the rMSN distribution to the $k$ samples of factor scores via the R package `EMMIXskew` [74].

The $k$-means is the most widely used method for getting an initial partition of groups, but it can sometimes be very inadequate for non-spherical data, especially when the dimension of the data is high [45,55]. The use of *deterministic* initialization such as the agglomerative nesting [76], the model-based hierarchical clustering of [22] implemented using the R package `mclust` [23] and the multistage procedure of [44] provide further choices of starting values. There are other popular approaches based on *stochastic* initialization schemes which may alleviate the potential drawbacks of $k$-means. For instance, the *emEM* [12] employs some *short runs* of EM algorithm from a number of random initializations. Each short run is stopped according to a loose convergence criterion. The solution with the highest log-likelihood value is used as a starter of the single *long* EM with a strict convergence criterion. Maitra [44] proposed a simple modification of *emEM*, called the *ranEM*, which skips running the short-EM by just evaluating the likelihood of each valid initial random start and choosing the parameters with the highest log-likelihood value as the initializer for the long-EM. Other extraordinary strategies of searching optimal starting values to promote algorithmic efficiency can be referred to [31,56,64].

The above procedure provides a quick and convenient strategy to initialize the parameters. Once the ECM algorithm has converged, we can determine the cluster membership according to the maximum *a posteriori* (MAP) classification rule. That is, each observation $\boldsymbol{y}_j$ is assigned to the component with the highest posterior probability.

The ECM procedure can get stuck in one of the many local maxima of the likelihood function [58]. To overcome such a flaw, it is recommended to initialize the algorithm with various choices of starting values for searching for all local maxima [49]. This can be done by specifying a variety of other starting points such as *random starts* [50]. The ML estimate $\hat{\boldsymbol{\Theta}}$ can be taken to be the maximizer corresponding to the highest log-likelihood value.

### 5.2. Model selection

A number of information criteria have been proposed to facilitate identifying an appropriate model. The most frequently employed index is the Bayesian Information Criterion (BIC) [67]

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

where $m$ is the number of free parameters, and $\ell_{\max}$ is the maximized log-likelihood value. Empirical evidence [8,9,53] has shown that BIC is useful in choosing the true number of classes of a given mixture model and an ideal number of latent factors.

As outlined by Biernacki et al. [13], an alternative promising measure for estimating the proper number of clusters is based on the integrated completed likelihood (ICL), defined as

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

where $\text{ENT}(\hat{\boldsymbol{z}}) = -\sum_{i=1}^{g}\sum_{j=1}^{n} \hat{z}_{ij} \log \hat{z}_{ij}$ is the entropy used to measure the overlap of clusters, where $\hat{z}_{ij}$ is the posterior probability of $\boldsymbol{y}_j$ classified to class $i$. Notably, ICL penalizes complex models more severely than BIC and thus favors the models with fewer latent classes.

In general, a smaller BIC or ICL value indicates a better fitted model. We note by passing that there is no clear consensus regarding which criterion is better to use. This depends on the problem at hand and usually a combined use would be of help to screen reasonable candidate models.

### 5.3. Convergence assessment

To monitor the convergence by using the likelihood increasing property of the ECM algorithm, the default stopping rule is $\ell(\hat{\boldsymbol{\Theta}}^{(k)}|\boldsymbol{y})/\ell(\hat{\boldsymbol{\Theta}}^{(k-1)}|\boldsymbol{y}) - 1 < \epsilon$, where $\epsilon$ is a user-specified tolerance. Another recommendation is to adopt the Aitken's acceleration criterion [49] which estimates the asymptotic maximum of the likelihood and allows to detect an early convergence. In our analysis, the algorithm is terminated if the maximum number of iterations $k_{\max} = 5000$ is reached or when the relative difference between two successive log-likelihood values is less than $\epsilon = 10^{-8}$.

### 5.4. Performance evaluation

To assess the model-based classification accuracy, we use the correct classification rate (CCR) and the adjusted Rand index (ARI) as proposed by Hubert and Arabie [29]. The CCR is calculated by considering all permutations of the class labels and the one with the lowest misclassification error can be treated as the final class membership assignment. As a measure of class agreement, the ARI accounts for the fact that a random classification may correctly classify some instances. The ARI has expected values of 0 under random classification and 1 for perfect classification. For both CCR and ARI, larger values indicate better classification results.

## 6. Application

### 6.1. Fishers' Iris data

As a motivating example, we use the Versicolor subset of Fisher's Iris data [20]. There are a total of 50 samples with each containing four-dimensional measurements in centimeters on the attributes of petal length, petal width, sepal length and sepal width. First, we employ a one-factor FA using the ML method for the data. Fig. 1 depicts the histogram of estimated factor scores in which the patterns are markedly skewed to the left with sample skewness equal to $-0.52$. Table 1 reports the ML results obtained by fitting the FA and SNFA models with $q = 1$ to the Versicolor data. The proportion of the total sample variances explained by the factor is larger under SNFA (69.7%) than under FA (66.6%). The ML estimate of the skewness parameter $\lambda$ is $-5.68$ and its standard error is 0.29, supporting strongly non-normality of the underlying factor.

Since the maximized likelihood values of the two fitted models are obtained, we perform the likelihood ratio test (LRT) for testing the hypothesis $H_0 : \lambda = 0$ (FA) against $H_1 : \lambda \neq 0$ (SNFA). The resulting LRT statistic is 4.52 with $p$-value 0.034, which is significant compared with a $\chi^2_1$ distribution, giving the other indication that the SNFA model is superior to the conventional FA. The $\chi^2_1$ distribution would be the limiting null distribution if regularity conditions hold [17]. Moreover, the sample skewness of the factor scores estimated by SNFA is –0.65, which exhibits a stronger left skew than does FA. In this regard, the "missed skewness" by the FA is then corrected to some extent by the SNFA.

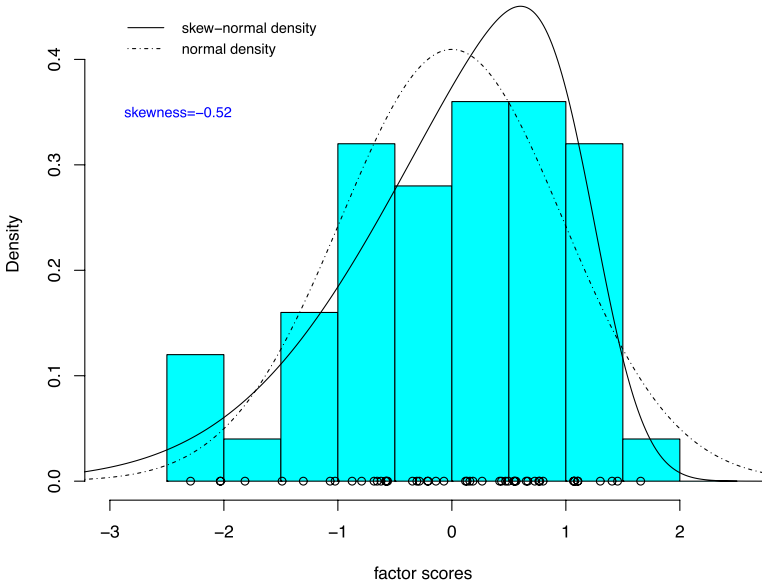


**Fig. 1.** Histogram of the factor scores obtained by fitting FA ($q = 1$) together with the fitted skew-normal (solid line) and normal (dot-dashed line) densities.

We consider also the fitting of the MSNFA model to the full set of Fisher's Iris data, which contains four geometric measurements of 50 samples from each of the three species of Iris (Setosa, Versicolor, andVirginica). For this illustration, the true number of clusters is taken to be unknown. Hence, the MSNFA model was applied to the data with $g$ ranging from 1 to 4. The number of latent factors $q$ is fixed at 1 to satisfy the restriction $(p-q)^2 \geq (p+q)$ as given by Eq. (8.5) of McLachlan and Peel [50]. For comparison, we also implement the MSTFA and MCSTFA models via the alternating expectation conditional maximization (AECM) algorithms described by [58,57], respectively. When implementing the estimating procedure, the component dfs are restricted to be equal for stabilizing the convergence. A summary of the results is listed in Table 2.

To compare the clustering performance of these models, we report in Table 2 the associated ARI and CCR values for each model considered. It can be observed that BIC selects the correct number of clusters ($g = 3$) for the MSNFA model and it attains its highest CCR and ARI for $g = 3$ (CCR $= 0.980$ and ARI $= 0.941$). The MSTFA model also attains its highest values for $g = 3$, which are not as high as for the MSNFA model (CCR $= 0.973$ and ARI $= 0.922$). Also, BIC suggests $g = 2$ rather than $g = 3$ clusters for MSTFA. The use of the ICL criterion selects $g = 2$ clusters for both the MSNFA and MSTFA models. The performance of the MCSTFA model can be seen to be much poorer than that for the MSNFA and MCSTFA models. Cross-tabulation of the true and predicted class memberships (Table 3) shows that both models can perfectly separate Setosa and Virginica samples from the other two species. The MCSTFA approach does not perform relatively well for this dataset as not a large number of parameters are needed to characterize the structure of clusters.

### 6.2. The WDBC dataset

Breast cancer is a major cause of death for women. Early detection of breast cancer through classification can avoid unnecessary surgery. As another illustration, we applied our method to the Wisconsin Diagnostic Breast Cancer (WDBC) data,which are available from the UCI Machine Learning data repository [25]. These data consist of $n = 569$ instances with a total of 32 different attributes. The first two attributes correspond to the ID number and the diagnosis status, of which 357 have the diagnosis benign and 212 have the diagnose malignant. The rest $p = 30$ attributes are ten real-valued measurements (Radius, Texture, Perimeter, Area, Smoothness, Compactness, Concavity, Concave points, Symmetry and Fractal dimension) computed from a digitized mammography image of a fine needle aspirates (FNA) of breast tissue, together with their associated mean, standard error and the mean of the three largest ('worst') values, respectively. Fig. 2 displays the scatterplots of the first 10 quantitative features. One can observe that many of these plots are apparently bimodal and appear to have rather non-elliptical patterns for both benign and malignant samples.

Fig. 3 shows the histograms of sample skewness of the 30 variables in benign and malignant samples. It is readily seen that most of the variables exhibit highly positive skewness. There are indeed over half of the variables with skewness greater than one. This motivates us to advocate the use of MSNFA model to analyze this dataset.

Since there are two known classes, we implemented two-component MFA and MSNFA models with $q$ ranging from 1 to 10. To fit the models via the ML method, the ECM algorithm developed in Section 4.2 was employed under twenty different initializations for the parameters. The resulting ML solutions, including the maximized log-likelihood values, the number of parameters together with the BIC and ICL values are listed in Table 4. To compare the classification accuracy, we also

**Table 1**
ML results for the Versicolor subset of the Iris data. Values within parentheses are the corresponding standard errors of ML estimates.

| Variable | MFA ($g = q = 1$) | | | MSNFA ($g = q = 1$) | | |
|---|---|---|---|---|---|---|
| | $\boldsymbol{\mu}$ | $\boldsymbol{B}$ | $\boldsymbol{d}$ | $\boldsymbol{\mu}$ | $\boldsymbol{B}$ | $\boldsymbol{d}$ |
| Sepal Length | 5.936 (0.072) | 0.395 (0.063) | 0.105 (0.025) | 5.942 (0.090) | 0.403 (0.069) | 0.105 (0.029) |
| Sepal Width | 2.770 (0.044) | 0.198 (0.041) | 0.057 (0.012) | 2.773 (0.054) | 0.201 (0.053) | 0.058 (0.013) |
| Petal Length | 4.260 (0.066) | 0.442 (0.052) | 0.021 (0.015) | 4.267 (0.090) | 0.454 (0.097) | 0.019 (0.005) |
| Petal Width | 1.326 (0.028) | 0.162 (0.023) | 0.012 (0.003) | 1.329 (0.038) | 0.163 (0.027) | 0.013 (0.003) |
| Proportion of variance explained | | 0.666 | | | 0.697 | |
| $m$ | | 12 | | | 13 | |
| $\ell_{\max}$ | | −16.488 | | | −14.229 | |
| LRT ($p$-value) | | | 4.517 | (0.034) | | |

**Table 2**
Comparison of the fitted MSNFA, MSTFA and MCSTFA models on the Iris data.

| Model | $g$ | $\ell_{\max}$ | $m$ | BIC | ICL | ARI | CCR |
|---|---|---|---|---|---|---|---|
| MSNFA | 1 | −419.3 | 13 | 903.8 | 903.8 | 0.000 | 0.333 |
| | 2 | −231.6 | 27 | 598.4 | **598.4** | 0.568 | 0.667 |
| | 3 | −192.8 | 41 | **591.2** | 600.6 | **0.941** | **0.980** |
| | 4 | −184.6 | 55 | 644.6 | 672.8 | 0.757 | 0.820 |
| MSTFA[a] | 1 | −387.4 | 17 | 860.0 | 860.0 | 0.000 | 0.333 |
| | 2 | −214.4 | 34 | **599.2** | **599.2** | 0.568 | 0.667 |
| | 3 | −176.8 | 51 | 609.2 | 617.6 | **0.922** | **0.973** |
| | 4 | −170.7 | 68 | 680.8 | 692.6 | 0.727 | 0.807 |
| MCSTFA[b] | 1 | −700.3 | 14 | **1470.8** | **1470.8** | 0.000 | 0.333 |
| | 2 | −686.4 | 21 | 1478.0 | 1583.8 | 0.185 | **0.553** |
| | 3 | −680.5 | 28 | 1501.2 | 1624.8 | 0.140 | 0.460 |
| | 4 | −676.1 | 35 | 1527.6 | 1698.2 | **0.238** | 0.440 |

MSTFA[a] and MCSTFA[b] indicate the mixture of skew-$t$ factor analyzers [61] and the mixture of common skew-$t$ factor analyzers [63], respectively, based on the generalized hyperbolic skew-$t$ distribution.

**Table 3**
Cross-tabulations of true and predicted class memberships for the selected MSNFA and MSTFA models on the Iris data.

| | MSNFA | | | MSTFA | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 1 | 2 | 3 |
| Setosa | 50 | 0 | 0 | 50 | 0 | 0 |
| Versicolor | 0 | 47 | 3 | 0 | 46 | 4 |
| Virginica | 0 | 0 | 50 | 0 | 0 | 50 |

computed the ARI and CCR for each $q$. As can be seen, the best fitted model is MSNFA with $q = 9$, no matter which model selection criterion was used. In addition, the resulting ARI (0.712) and CCR (0.923) under the fitted MSNFA ($q = 9$) are higher than all those under MFA models, although the MSNFA reaches its best ARI (0.762) and CCR (0.937) when $q = 7$. The result confirms that the MSNFA is more appropriate for this dataset, providing more accurate classification for this dataset, which exhibits a departure from normality. Finally, we did attempt to compare our MSNFA method with the MSTFA and MCSTFA models [61,63], but we encountered certain convergence problem when implementing the latter two models for this dataset.

### 6.3. Seeds data

Our third example concerns the seeds dataset analyzed by Charytanowicz and Niewczas [15]. Seven geometric features (area, perimeter, compactness, length of kernel, width of kernel, asymmetry coefficient, and length of kernel groove) were measured from the $X$-ray images of 210 wheat kernels. These grains belong to three different wheat varieties, namely Kama, Rosa, and Canadian. We consider the fitting of the MFA, MSNFA, MSTFA, and MCSTFA models to this dataset, with $q$ varying from 1 to 3. Focusing first on the case where $g$ is *a priori* known to be 3, it can be observed from Table 5 that the model with $q = 3$ is preferred by both the BIC and ICL for the MFA and MSTFA models. For the GHST factor models (see Table 6), the

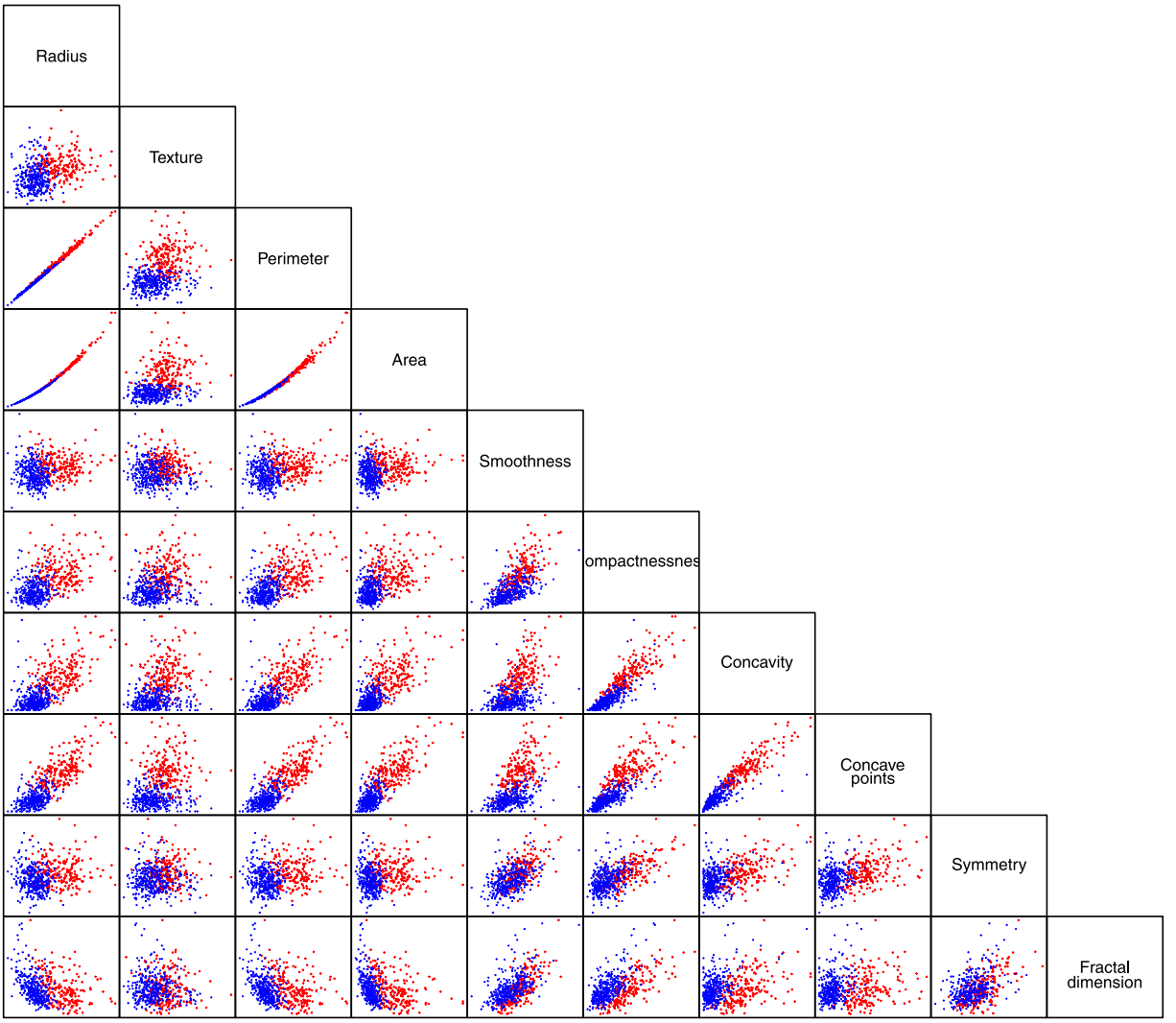


**Fig. 2.** Pairwise scatterplots of the first 10 quantitative features. The blue dot represents benign samples, and the red dot represents malignant samples.

MSTFA model obtained its lowest BIC and ICL values when $q = 3$, whereas $q = 2$ is preferred by BIC and ICL for the MCSTFA model. However, their performances in clustering were relatively poor in terms of ARI and CCR.

We consider also the fitting of these factor models to the seeds data when $g$ is taken to be unknown. The MFA, MSNFA, MSTFA, and MCSTFA models were applied to the data with $g$ ranging from 1 to 4. As above, the number of latent factors $q$ varies from 1 to 3. On comparing their results reported in Tables 5 and 6, it can be observed that the model corresponding to $g = 3$ and $q = 3$ is preferred by both BIC and ICL for the MFA and MSNFA model, with the latter obtaining lower BIC and ICL values. For the MSTFA and MCSFTA models, a model with $g = 2$ would be chosen based on BIC and ICL. In this example, the highest ARI and CCR is given by the MSNFA model with $q = 3$ (ARI $= 0.7505$ and CCR $= 0.9095$), which coincides with the model selected by BIC and ICL. We note in Table 6 that the likelihood does not always increase with $g$ and/or $q$ for the MSTFA and MCSTFA models, indicating the convergence problems we encountered in the fitting of these two models.

### *6.4. A simulation study*

We undertook a simulation study to examine the goodness of fit and clustering ability in simulated data by applying the proposed MSNFA model. To conduct experimental studies, we generated data sets in $\mathbb{R}^{10}$ of size $n$ each from a 3-component MSNFA model with $q = 2$ factors. The presumed parameters are given as

$$w_1 = w_2 = w_3 = 1/3, \qquad \boldsymbol{\mu}_1 = 1\mathbf{1}_{10}, \qquad \boldsymbol{\mu}_2 = 2\mathbf{1}_{10}, \qquad \boldsymbol{\mu}_3 = 3\mathbf{1}_{10}$$
$$\boldsymbol{B}_i = \text{Unif}(10, 2), \qquad \boldsymbol{D}_i = \text{diag}\{\text{Unif}(10, 1)\}, \qquad \boldsymbol{\lambda}_i = \lambda\mathbf{1}_2, \quad (i = 1, 2, 3),$$

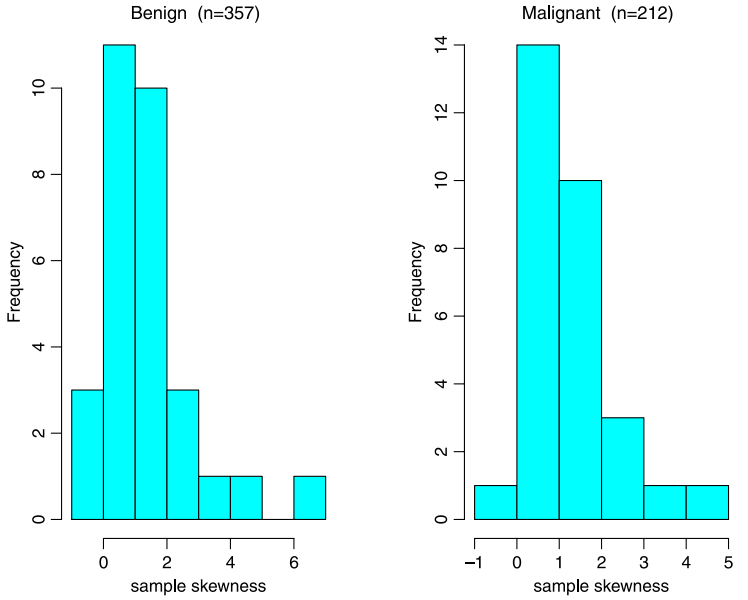


**Fig. 3.** Histograms of sample skewness of the 30 variables in benign and malignant samples.

**Table 4**
Comparison of MFA and MSNFA fitting results and implied clustering versus the true membership of WDBC data.

| Model | $q$ | $\ell_{\max}$ | $m$ | BIC | ICL | ARI | CCR |
|---|---|---|---|---|---|---|---|
| MFA | 1 | 9,624.8 | 181 | −18,101.4 | −18,083.8 | 0.520 | 0.861 |
| | 2 | 12,362.7 | 239 | −23,209.2 | −23,193.2 | 0.396 | 0.817 |
| | 3 | 13,962.5 | 295 | −26,053.6 | −26,043.4 | 0.359 | 0.803 |
| | 4 | 15,616.8 | 349 | −29,019.6 | −29,013.6 | 0.658 | 0.907 |
| | 5 | 15,726.5 | 401 | −28,909.2 | −28,897.4 | 0.595 | 0.888 |
| | 6 | 16,691.4 | 451 | −30,521.6 | −30,513.4 | 0.630 | 0.898 |
| | 7 | 17,017.2 | 499 | −30,868.8 | −30,862.0 | 0.670 | 0.910 |
| | 8 | 17,248.6 | 545 | −31,039.8 | −31,030.6 | 0.595 | 0.888 |
| | 9 | 18,467.3 | 589 | −33,198.0 | −33,190.8 | 0.700 | 0.919 |
| | 10 | 17,692.3 | 631 | −31,381.6 | −31,370.0 | 0.624 | 0.896 |
| MSNFA | 1 | 9632.8 | 183 | −18,104.8 | −18,086.2 | 0.515 | 0.859 |
| | 2 | 12,441.3 | 243 | −23,341.0 | −23,325.2 | 0.373 | 0.808 |
| | 3 | 14,117.8 | 301 | −26,326.2 | −26,317.2 | 0.397 | 0.817 |
| | 4 | 15,700.5 | 357 | −29,136.2 | −29,127.6 | 0.658 | 0.907 |
| | 5 | 15,830.1 | 411 | −29,053.0 | −29,042.6 | 0.618 | 0.895 |
| | 6 | 16,933.3 | 463 | −30,929.4 | −30,918.6 | 0.718 | 0.924 |
| | 7 | 17,486.8 | 513 | −31,719.2 | −31,712.0 | **0.762** | **0.937** |
| | 8 | 17,572.5 | 561 | −31,586.0 | −31,579.0 | 0.681 | 0.914 |
| | 9 | **18,598.8** | 607 | **−33,347.0** | **−33,340.4** | 0.712 | 0.923 |
| | 10 | 18,000.9 | 651 | −31,872.0 | −31,862.8 | 0.700 | 0.919 |

where Unif$(r, s)$ denotes a $r \times s$ matrix of random numbers drawn from a uniform distribution on the unit (0, 1) interval and $\mathbf{1}_p$ is a $p \times 1$ vector of ones.

As pointed out by Wall et al. [71], the skewness of generated latent factors may become much smaller than the actual population values when the sample size $n$ is not large enough. In this study, we therefore consider somewhat large sample sizes ($n = 600$, 1200, and 2400) to enhance the skewness effects. The specific values of $\lambda$ are chosen as 0, 5, and 10. The higher the value of $\lambda$, the stronger the departure from the normality, while the zero-skewness $\lambda = 0$ corresponds to normal factors.

For comparison purposes, each simulated data set was fitted under the MFA and MSNFA scenarios along with the MSTFA model of Murray et al. [61] with $g = 3$ and $q = 2$. The total numbers of free parameters in the three models are 119, 125 and 152, respectively. Compared with the formulation of MSNFA, the MSTFA model involves a much larger number of unknown parameters because its factor analytic representation applies to the error terms rather than the latent factors.

A total of 100 replications were run across each combination of $n$ and $\lambda$. The comparison between the three models is made using BIC and ARI, which are commonly adopted to evaluate model fitting and classification performances, respectively. Table 7 lists the average BIC and ARI values and the corresponding standard deviations. To evaluate the objective use of the criteria, the frequencies preferred by BIC and ARI are also listed in the table. When $\lambda = 0$, it is not surprising that

**Table 5**
Comparison of the fitted MFA and MSNFA models on the seeds data.

| Model | $g$ | $q$ | $\ell_{\max}$ | $m$ | BIC | ICL | ARI | CCR |
|---|---|---|---|---|---|---|---|---|
| | 1 | 1 | 75.74 | 21 | −39.20 | −156.80 | 0.0000 | 0.3333 |
| | 1 | 2 | 105.05 | 27 | −65.73 | −262.93 | 0.0000 | 0.3333 |
| | 1 | 3 | 680.75 | 32 | −1190.39 | −4761.58 | 0.0000 | 0.3333 |
| | 2 | 1 | 265.34 | 43 | −300.75 | −1176.00 | 0.4191 | 0.6571 |
| | 2 | 2 | 274.18 | 55 | −254.28 | −998.27 | 0.4388 | 0.6571 |
| MFA | 2 | 3 | 1026.80 | 65 | −1706.03 | −6814.15 | 0.4685 | 0.6667 |
| | 3 | 1 | 419.77 | 65 | −491.98 | −1945.11 | 0.4261 | 0.7667 |
| | 3 | 2 | 385.98 | 83 | −328.14 | −1286.14 | 0.4189 | 0.7619 |
| | 3 | 3 | 1201.36 | 98 | **−1878.71** | **−7503.86** | **0.6875** | **0.8810** |
| | 4 | 1 | 396.86 | 87 | −328.52 | −1294.94 | 0.4891 | 0.7429 |
| | 4 | 2 | 540.57 | 111 | −487.61 | −1925.27 | 0.4430 | 0.6476 |
| | 4 | 3 | 1255.09 | 131 | −1809.72 | −7218.28 | 0.6043 | 0.7381 |
| | 1 | 1 | 80.68 | 22 | −43.73 | −174.92 | 0.0000 | 0.3333 |
| | 1 | 2 | 112.00 | 29 | −68.94 | −275.77 | 0.0000 | 0.3333 |
| | 1 | 3 | 685.99 | 35 | −1184.82 | −4739.29 | 0.0000 | 0.3333 |
| | 2 | 1 | 303.42 | 45 | −366.21 | −1416.58 | 0.2868 | 0.6000 |
| | 2 | 2 | 278.44 | 59 | −241.39 | −937.19 | 0.4257 | 0.6524 |
| MSNFA | 2 | 3 | 1034.95 | 71 | −1690.26 | −6751.42 | 0.4720 | 0.6667 |
| | 3 | 1 | 505.52 | 68 | −647.44 | −2570.41 | 0.1253 | 0.5238 |
| | 3 | 2 | 476.50 | 89 | −477.10 | −1878.65 | 0.2363 | 0.6333 |
| | 3 | 3 | 1207.38 | 107 | **−1842.61** | **−7357.09** | **0.7505** | **0.9095** |
| | 4 | 1 | 368.88 | 91 | −251.18 | −961.32 | 0.4517 | 0.7238 |
| | 4 | 2 | 816.30 | 119 | −996.29 | −3951.51 | 0.2632 | 0.5143 |
| | 4 | 3 | 1280.77 | 143 | −1796.90 | −7167.82 | 0.5971 | 0.7333 |

**Table 6**
Results of the fitted MSTFA and MCSTFA models on the seeds data.

| Model | $g$ | $q$ | $\ell_{\max}$ | $m$ | BIC | ICL | ARI | CCR |
|---|---|---|---|---|---|---|---|---|
| | 1 | 1 | 535.65 | 29 | −916.24 | −3664.95 | 0.0000 | 0.3333 |
| | 1 | 2 | 565.63 | 35 | −944.12 | −3776.47 | 0.0000 | 0.3333 |
| | 1 | 3 | 673.56 | 40 | −1133.25 | −4532.98 | 0.0000 | 0.3333 |
| | 2 | 1 | 809.71 | 59 | −1303.94 | −5193.91 | 0.4268 | 0.6524 |
| | 2 | 2 | 733.04 | 71 | −1086.44 | −4331.91 | 0.4372 | 0.6286 |
| MSTFA | 2 | 3 | 1047.10 | 81 | **−1661.07** | **−6624.61** | 0.4685 | 0.6667 |
| | 3 | 1 | 469.99 | 89 | −464.09 | −1808.96 | 0.5216 | 0.8238 |
| | 3 | 2 | 829.82 | 107 | −1087.50 | −4314.84 | 0.4664 | 0.7476 |
| | 3 | 3 | 1071.08 | 122 | −1489.81 | −5940.30 | 0.3975 | 0.6333 |
| | 4 | 1 | 542.17 | 119 | −448.04 | −1656.63 | 0.3651 | 0.6333 |
| | 4 | 2 | 754.36 | 143 | −744.08 | −2939.44 | **0.5776** | **0.7762** |
| | 1 | 1 | −609.55 | 23 | 1342.08 | 5368.33 | 0.0000 | 0.3333 |
| | 1 | 2 | 113.49 | 30 | −66.58 | −266.30 | 0.0000 | 0.3333 |
| | 1 | 3 | −301.87 | 36 | 796.23 | 3184.93 | 0.0000 | 0.3333 |
| | 2 | 1 | −607.92 | 34 | 1397.64 | 5594.50 | 0.0000 | 0.3381 |
| MCSTFA | 2 | 2 | −751.58 | 44 | 1738.43 | 6953.74 | 2.0096 | 0.3333 |
| | 2 | 3 | 540.83 | 54 | **−792.92** | **−3167.05** | 0.0002 | 0.3429 |
| | 3 | 1 | −978.40 | 45 | 2197.43 | 9006.83 | 0.2199 | 0.5571 |
| | 3 | 2 | 175.25 | 58 | −40.38 | −27.31 | **0.5103** | **0.8048** |
| | 3 | 3 | −893.33 | 72 | 2171.64 | 8821.90 | 0.1243 | 0.5000 |

MFA is more likely to be selected. When focusing on the cases of non-zero skewness, the BIC scores provide a 53%–100% agreement with the specification of MSNFA, and the percentage of correctly choosing the true model increases with the sample size and the value of skewness parameter. In this study, the MSTFA model does not work well as it is strongly penalized due to over-fitting.

With regard to the MAP classification, the results indicate that when the latent factors approach normality ($\lambda = 0$), all three models produce comparable ARI values. When the latent factors are moderately and highly skewed ($\lambda = 5, 10$), the MSNFA model yields slightly higher classification accuracies and is preferred more often than the other two models. Such a phenomenon becomes apparent as the sample size increases. In summary, the MNSFA model can provide greater flexibility in model fitting and superiority for clustering in the presence of skew factors, at least for the setting of parameters used in this study.

## 7. Conclusion

We have proposed the MSNFA model by replacing the normal latent factors in the classical MFA model with the rMSN distributed factors for each component. This family of mixture factor analyzers has emerged as an attractive tool since

**Table 7**
Simulation results based on 100 replications.

| | | $\lambda = 0$ | | | $\lambda = 5$ | | | $\lambda = 10$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | MFA | MSNFA | MSTFA[a] | MFA | MSNFA | MSTFA[a] | MFA | MSNFA | MSTFA[a] |
| | $n = 600$ | | | | | | | | | |
| BIC | Mean | −7746.56 | −7762.45 | −7821.87 | −7746.58 | −7743.88 | −7816.90 | −7871.05 | −7865.28 | −7939.25 |
| | Std | 341.70 | 342.46 | 342.47 | 335.16 | 362.39 | 337.12 | 358.66 | 362.49 | 359.65 |
| | Freq | 98 | 2 | 0 | 47 | 53 | 0 | 40 | 60 | 0 |
| ARI | Mean | 0.717 | 0.716 | 0.706 | 0.719 | 0.732 | 0.714 | 0.680 | 0.697 | 0.672 |
| | Std | 0.092 | 0.092 | 0.099 | 0.088 | 0.085 | 0.092 | 0.097 | 0.095 | 0.104 |
| | Freq | 37 | 34 | 29 | 15 | 65 | 20 | 15 | 69 | 16 |
| | $n = 1200$ | | | | | | | | | |
| BIC | Mean | −15,319.82 | −15,338.41 | −15,409.79 | −15,246.60 | −15,228.07 | −15,318.34 | −15,330.90 | −15,307.24 | −15,399.61 |
| | Std | 785.44 | 785.75 | 785.97 | 804.30 | 807.57 | 805.04 | 784.52 | 792.12 | 787.92 |
| | Freq | 98 | 2 | 0 | 24 | 76 | 0 | 15 | 85 | 0 |
| ARI | Mean | 0.715 | 0.715 | 0.708 | 0.728 | 0.742 | 0.729 | 0.713 | 0.730 | 0.713 |
| | Std | 0.080 | 0.080 | 0.082 | 0.089 | 0.084 | 0.089 | 0.094 | 0.088 | 0.094 |
| | Freq | 25 | 44 | 31 | 11 | 68 | 21 | 10 | 75 | 15 |
| | $n = 2400$ | | | | | | | | | |
| BIC | Mean | −30,409.19 | −30,428.81 | −30,513.46 | −30,178.59 | −30,129.58 | −30,245.06 | −30,423.53 | −30,350.51 | −30,488.06 |
| | Std | 1407.30 | 1405.85 | 1405.87 | 1390.61 | 1400.14 | 1393.94 | 1422.68 | 1433.58 | 1426.50 |
| | Freq | 98 | 2 | 0 | 1 | 99 | 0 | 0 | 100 | 0 |
| ARI | Mean | 0.727 | 0.727 | 0.723 | 0.727 | 0.739 | 0.730 | 0.721 | 0.739 | 0.726 |
| | Std | 0.075 | 0.075 | 0.077 | 0.071 | 0.067 | 0.069 | 0.087 | 0.082 | 0.085 |
| | Freq | 35 | 37 | 28 | 4 | 76 | 20 | 1 | 84 | 15 |

[a] Murray et al.'s [61] approach based on the generalized hyperbolic skew-$t$ distribution.

it can account for groups in the data exhibiting patterns of asymmetry and multimodality which are commonly seen in high-dimensional data. For estimating parameters, an analytically simple ECM algorithm is developed under a four-level hierarchical framework. Some computational strategies related to the specification of starting values, convergence assessment and provision of standard errors are provided. Two main identification problems regarding invariant likelihood caused by factor indeterminacy and label switching are also discussed. We should mention that both of which do not affect the clustering results. Numerical results on model choice based on information-based criteria and apparent error rate for summarizing classification accuracy indicate the effectiveness and superiority of the proposed method when compared with the traditional MFA.

There are a number of possible extensions of the current work. While the proposed MSNFA has shown its flexibility in modeling asymmetric features among heterogeneous data, its robustness against outliers could still be unduly influenced by heavy-tailed observations. Mixtures of factor analyzers based on a more general family of distributions such as the skew $t$-distribution and its variants [6,32,39,65,66] would be of interest for future research. For identifying the optimal number of clusters, an effective method is to design a mixture component merging procedure using entropy as the criterion suggested by Baudry et al. [11]. Melnykov [54] further derived the asymptotic distribution of entropy and applied it to find good cluster partitions. Another worthwhile task is to develop workable Markov chain Monte Carlo algorithms for drawing inferences under a Bayesian paradigm. Although the proposed ECM procedure is quite easy to implement, its convergence can be slow in certain situations. Therefore, pursuing some modified algorithms such as [41,75,78] toward fast convergence deserves further investigation.

## Acknowledgments

We gratefully acknowledge the Chief Editor, the Associate Editor and two anonymous referees for their comments and suggestions, which have led to a much improved version of this article. This research was supported by MOST 103-2118-M-005-001-MY2 awarded by the Ministry of Science and Technology of Taiwan.

## References

[1] K. Aas, I.H. Haff, The generalized hyperbolic skew Student's $t$-distribution, J. Financ. Econom. 4 (2005) 275–309.
[2] R. Arellano-Valle, A. Azzalini, On the unification of families of skew-normal distributions, Scand. J. Stat. 33 (2006) 561–574.
[3] R. Arellano-Valle, M. Genton, On fundamental skew distributions, J. Multivariate Anal. 96 (2005) 93–116.
[4] A. Azzalini, The skew-normal distribution and related multivariate families, Scand. J. Stat. 32 (2005) 159–188.
[5] A. Azzalini, A. Capitanio, Statistical applications of the multivariate skew normal distribution, J. R. Stat. Soc. Ser. B 61 (1999) 579–602.
[6] A. Azzalini, A. Capitanio, Distributions generated by perturbation of symmetry with emphasis on a multivariate skew $t$-distribution, J. R. Stat. Soc. Ser. B 65 (2003) 367–389.
[7] A. Azzalini, A. Dalla Valle, The multivariate skew-normal distribution, Biometrika 83 (1996) 715–726.
[8] J. Baek, G.J. McLachlan, Mixtures of common $t$-factor analyzers for clustering high-dimensional microarray data, Bioinformatics 27 (2011) 1269–1276.

[9] J. Baek, G.J. McLachlan, L.K. Flack, 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 (2010) 1–13.
[10] O. Barndorff-Nielsen, N. Shephard, Non-Gaussian Ornstein–Uhlenbeck-based models and some of their uses in financial economics, J. R. Stat. Soc. Ser. B 63 (2001) 167–241.
[11] J.P. Baudry, A.E. Raftery, G. Celeux, K. Lo, R. Gottardo, Combining mixture components for clustering, J. Comput. Graph. Stat. 9 (2010) 332–353.
[12] C. Biernacki, G. Celeux, G. Govaert, Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models, Comput. Statist. Data Anal. 41 (2003) 561–575.
[13] C. Biernacki, G. Celeux, G. Govaert, Assessing a mixture model for clustering with the integrated completed likelihood, IEEE Trans. Pattern Anal. Mach. Intell. 22 (2000) 719–725.
[14] D. Böhning, Computer-Assisted Analysis of Mixtures and Applications: Meta-Analysis, Disease Mapping and Others, Chapman and Hall, New York, 1999.
[15] M. Charytanowicz, J. Niewczas, P. Kulczycki, P. Kowalski, S. Lukasik, S. Zak, A complete gradient clustering algorithm for features analysis of X-ray images, in: E. Pietka, J. Kawa (Eds.), Information Technologies in Biomedicine, Springer, Berlin, 2010, pp. 15–24.
[16] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm (with discussion), J. Roy. Statist. Soc. B 39 (1977) 1–38.
[17] M. Drton, Likelihood ratio tests and singularities, Ann. Statist. 37 (2009) 979–1012.
[18] B. Efron, D.V. Hinkley, Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information, Biometrika 65 (1978) 457–482.
[19] B.S. Everitt, D.J Hand, Finite Mixture Distributions, Chapman & Hall, London, 1981.
[20] R. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugenics 7 (1936) 179–188.
[21] E. Fokoué, D.M. Titterington, Mixtures of factor analyzers. Bayesian estimation and inference by stochastic simulation, Mach. Learn. 50 (2003) 73–94.
[22] C. Fraley, Algorithms for model-based Gaussian hierarchical clustering, SIAM J. Sci. Comput. 20 (1998) 270–281.
[23] C. Fraley, A.E. Raftery, T.B. Murphy, L. Scrucca, MCLUST Version 4 for R: Normal Mixture Modeling and Model-Based Clustering, Technical Report 597, University of Washington, Department of Statistics, Seattle, WA, 2012.
[24] B.C. Franczak, P.D. McNicholas, R.P. Browne, P.M. Murray, Parsimonious shifted asymmetric Laplace mixtures, 2013, Preprint arXiv:1311.0317v1.
[25] A. Frank, A. Asuncion, UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science, 2010. http://archive.ics.uci.edu/ml.
[26] S. Frühwirth-Schnatter, Finite Mixture and Markov Switching Models, Springer, New York, 2006.
[27] Z. Ghahramani, M. Beal, Variational inference for Bayesian mixture of factor analysers, in: S. Solla, T. Leen, K.-R Muller (Eds.), in: Adv. Neur. Info. Proc. Sys., vol. 12, MIT Press, Cambridge, 2000, pp. 449–455.
[28] Z. Ghahramani, G.E. Hinton, The EM Algorithm for Mixtures of Factor Analyzers, Technical Report No. CRG-TR-96-1, University of Toronto, 1997.
[29] L. Hubert, P. Arabie, Comparing partitions, J. Classification 2 (1985) 193–218.
[30] M. Jamshidian, An EM algorithm for ML factor analysis with missing data, in: M. Berkane (Ed.), Latent Variable Modeling and Applications to Causality, Springer Verlag, New York, 1997, pp. 247–258.
[31] D. Karlis, E. Xekalaki, Choosing initial values for the EM algorithm for finite mixtures, Comput. Statist. Data Anal. 41 (2003) 577–590.
[32] V.H. Lachos, P. Ghosh, R.B. Arellano-Valle, Likelihood based inference for skew-normal/independent linear mixed model, Stat. Sin. 20 (2010) 303–322.
[33] S.X. Lee, G.J. McLachlan, On mixtures of skew normal and skew $t$-distributions, Adv. Data Anal. Classif. 7 (2013) 241–266.
[34] S. Lee, G.J. McLachlan, Finite mixtures of multivariate skew $t$-distributions: some recent and new results, Statist. Comput. 24 (2014) 181–202.
[35] S.X. Lee, G.J. McLachlan, EMMIX-uskew: an R package for fitting mixtures of multivariate skew $t$-distributions via the EM algorithm, J. Statist. Software 55 (12) (2013).
[36] Y.W. Lee, S.H. Poon, Systemic and systematic factors for loan portfolio loss distribution, in: Econometrics and Applied Economics Workshops, School of Social Science, University of Manchester, 2011, pp. 1–61.
[37] T.I. Lin, Maximum likelihood estimation for multivariate skew normal mixture models, J. Multivariate Anal. 100 (2009) 257–265.
[38] B.G. Lindsay, Mixture Models: Theory, Geometry, and Applications, in: NSF-CBMS Regional Conference Series in probability and Statistics, vol. 5, Institute of Mathematical Statistics, Hayward, CA, 1995.
[39] T.I. Lin, H.J. Ho, C.R. Lee, Flexible mixture modelling using the multivariate skew-$t$-normal distribution, Statist. Comput. 24 (2014) 531–546.
[40] T.I. Lin, J.C. Lee, S.Y. Yen, Finite mixture modelling using the skew normal distribution, Stat. Sin. 17 (2007) 909–927.
[41] C.H. Liu, D.B. Rubin, Y.N. Wu, Parameter expansion to accelerate EM: the PX-EM algorithm, Biometrika 85 (1998) 755–770.
[42] H.F. Lopes, M. West, Bayesian model assessment in factor analysis, Stat. Sin. 4 (2004) 41–67.
[43] T.A. Louis, Finding the observed information when using the EM algorithm, J. R. Stat. Soc. Ser. B 44 (1982) 226–232.
[44] R. Maitra, Initializing partition-optimization algorithms, IEEE ACM Trans. Comput. Biol. Bioinformatics 6 (2009) 144–157.
[45] R. Maitra, V. Melnykov, Simulating data to study performance of finite mixture modeling and clustering algorithms, J. Comput. Graph. Stat. 19 (2010) 354–376.
[46] G.J. McLachlan, K.E. Basford, Mixture Models: Inference and Application to Clustering, Marcel Dekker, New York, 1988.
[47] G.J. McLachlan, R.W. Bean, L.B.T. Jones, Extension of the mixture of factor analyzers model to incorporate the multivariate $t$-distribution, Comput. Statist. Data Anal. 51 (2007) 5327–5338.
[48] G.J. McLachlan, R.W. Bean, D. Peel, A mixture model-based approach to the clustering of microarray expression data, Bioinformatics 18 (2002) 413–422.
[49] G.J. McLachlan, T. Krishnan, The EM Algorithm and Extensions, second ed., Wiley, New York, 2008.
[50] G.J. McLachlan, D. Peel, Finite Mixture Models, Wiley, New York, 2000.
[51] G.J. McLachlan, D. Peel, Mixtures of factor analyzers, in: Proceedings of the Seventeenth International Conference on Machine Learning, Morgan Kaufmann, San Francisco, 2000, pp. 599–606.
[52] G.J. McLachlan, D. Peel, R.W. Bean, Modelling high-dimensional data by mixtures of factor analyzers, Comput. Statist. Data Anal. 41 (2003) 379–388.
[53] P.D. McNicholas, T.B. Murphy, Parsimonious Gaussian mixture models, Statist. Comput. 18 (2008) 285–296.
[54] V. Melnykov, On the distribution of posterior probabilities in finite mixture models with application in clustering, J. Multivariate Anal. 122 (2013) 175–189.
[55] V. Melnykov, W.C. Chen, R. Maitra, MixSim: an R package for simulating data to study performance of clustering algorithms, J. Stat. Softw. 51 (2012) 1–25.
[56] V. Melnykov, I. Melnykov, Initializing the EM algorithm in Gaussian mixture models with an unknown number of components, Comput. Statist. Data Anal. 56 (2012) 1381–1395.
[57] K. Mengersen, C. Robert, D.M. Titterington, Mixtures: Estimation and Applications, John Wiley & Sons, Chichester, UK, 2011.
[58] X.L. Meng, D.B. Rubin, Maximum likelihood estimation via the ECM algorithm: a general framework, Biometrika 80 (1993) 267–278.
[59] X.L. Meng, D.B. Rubin, Using EM to obtain asymptotic variance–covariance matrices: The SEM algorithm, J. Amer. Statist. Assoc. 86 (1991) 899–909.
[60] A. Montanari, C. Viroli, A skew-normal factor model for the analysis of student satisfaction towards university courses, J. Appl. Statist. 37 (2010) 473–487.
[61] P.M. Murray, R. Browne, P.D. McNicholas, Mixtures of skew-$t$ factor analyzers, Comput. Statist. Data Anal. 77 (2014) 326–335.
[62] P. Murray, R. Browne, P.D. McNicholas, Mixtures of 'unrestricted' skew-$t$ factor analyzers, 2013, Preprint arXiv:1310.6224v1.
[63] P.M. Murray, P.D. McNicholas, R. Browne, Mixtures of common skew-$t$ factor analyzers, Stat 3 (2014) 68–82.
[64] A. O'Hagan, T. Murphy, I. Gormley, Computational aspects of fitting mixture models via the expectation–maximization algorithm, Comput. Statist. Data Anal. 56 (2012) 3843–3864.
[65] S. Pyne, X. Hu, K. Wang, E. Rossin, T.I. Lin, L.M. Maier, C. Baecher-Allan, G.J. McLachlan, P. Tamayo, D.A. Hafler, P.L. De Jager, J.P. Mesirov, Automated high- dimensional flow cytometric data analysis, Proc. Natl. Acad. Sci USA 106 (2009) 8519–8524.

[66] S.K. Sahu, D.K. Dey, M.D. Branco, A new class of multivariate skew distributions with application to Bayesian regression models, Canad. J. Statist. 31 (2003) 129–150.
[67] G. Schwarz, Estimating the dimension of a model, Ann. Statist. 6 (1978) 461–464.
[68] D.M. Titterington, A.F.M. Smith, U.E. Markov, Statistical Analysis of Finite Mixture Distributions, Wiley, New York, 1985.
[69] C. Tortora, P.D. McNicholas, R. Browne, A mixture of generalized hyperbolic factor analyzers, 2013, arXiv:1311.6530v1.
[70] A. Utsugi, T. Kumagai, Bayesian analysis of mixtures of factor analyzers, Neural Comp. 13 (2001) 993–1002.
[71] M.M. Wall, J. Guo, Y. Amemiya, Mixture factor analysis for approximating a non-normally distributed continuous latent factor with continuous and dichotomous observed variables, Multivar. Behav. Res. 47 (2012) 276–313.
[72] W.L. Wang, Mixtures of common factor analyzers for high-dimensional data with missing information, J. Multivariate Anal. 117 (2013) 120–133.
[73] W.L. Wang, Mixtures of common $t$-factor analyzers for modeling high-dimensional data with missing values, Comput. Statist. Data Anal. 83 (2015) 223–235.
[74] K. Wang, EMMIX-skew (R package version 1.0-12): EM algorithm for mixture of multivariate skew Normal/$t$ distributions, 2009.
[75] W.L. Wang, T.I. Lin, An efficient ECM algorithm for maximum likelihood estimation in mixtures of $t$-factor analyzers, Comput. Statist. 8 (2013) 751–769.
[76] J.H. Ward, Hierarchical grouping to optimize an objective function, J. Amer. Statist. Assoc. 58 (1963) 236–244.
[77] C.F.J. Wu, On the convergence properties of the EM algorithm, Ann. Stat. 11 (1983) 95–103.
[78] J.H. Zhao, P.L.H. Yu, Fast ML estimation for the mixture of factor analyzers via an ECM algorithm, IEEE. Trans. Neural Netw. 19 (2008) 1956–1961.