Comput Stat (2010) 25:183–201
DOI 10.1007/s00180-009-0169-5

ORIGINAL PAPER

Supervised learning of multivariate skew normal mixture models with missing information

Tzy-Chy Lin · Tsung-I Lin

Received: 9 November 2008 / Accepted: 26 August 2009 / Published online: 19 September 2009

Abstract We establish computationally flexible tools for the analysis of multivariate skew normal mixtures when missing values occur in data. To facilitate the computation and simplify the theoretical derivation, two auxiliary permutation matrices are incorporated into the model for the determination of observed and missing components of each observation and are manifestly effective in reducing the computational complexity. We present an analytically feasible EM algorithm for the supervised learning of parameters as well as missing observations. The proposed mixture analyzer, including the most commonly used Gaussian mixtures as a special case, allows practitioners to handle incomplete multivariate data sets in a wide range of considerations. The methodology is illustrated through a real data set with varying proportions of synthetic missing values generated by MCAR and MAR mechanisms and shown to perform well on classification tasks.

Keywords Classifier · EM algorithm · Ignorable · Incomplete data · MSN model · Multivariate truncated normal

1 Introduction

Finite mixture models have become a flexible and powerful probabilistic learning tool for heterogeneous multivariate data and been used extensively in classification and

T.-C. Lin
Institute of Statistics, National Chiao Tung University, Hsinchu 301, Taiwan

T.-I. Lin (✉)
Department of Applied Mathematics and Institute of Statistics, National Chung Hsing University, Taichung 402, Taiwan
e-mail: tilin@amath.nchu.edu.tw

clustering. During the last two decades, the usefulness of Gaussian mixture (GMIX) (Pearson 1894) and Student's t mixture (TMIX) models, see Peel and McLachlan (2000), Shoham (2002), Shoham et al. (2003) and Lin et al. (2004), are being more frequently applied in various fields such as pattern recognition, data mining, computer vision, signal and image processing, machine learning and bioinformatics, etc. For a comprehensive introduction to mixture models and their applications, see monographs by Titterington et al. (1985), McLachlan and Basford (1988), McLachlan and Peel (2000), Frühwirth-Schnatter (2006), and the references therein. Recently, mixtures of univariate skew normal and skew t distributions as natural extensions of univariate Gaussian mixtures have been considered by Lin et al. (2007a,b).

In practice data may exhibit highly asymmetric observations and thus statistical inferences drawn from the ordinary Gaussian assumptions may yield unreliable results. To reduce unduly skewness encountered in general practice, one commonly adopted approach is through the best known data-based power transformation proposed by Box and Cox (1964). Although such a treatment is very convenient to use, the achievement of joint normality is rarely satisfied and the transformed variables become more difficult to interpret. Instead of applying transformations, there has been a growing interest in proposing a wider class of distributions, namely, the multivariate skew normal (MSN) distribution, which contains an extra vector of parameters in regulating skewness and includes the Gaussian family as a special case.

The MSN distribution was originally studied by Azzalini and Dalla Valle (1996) and some further attractive features and applications are given in Azzalini and Capitanio (1999). Based on this class of distributions, a number of extensions or alternative proposals have appeared during the last decade. Arellano-Vallea and Genton (2005) studied the family of fundamental skew normal (FUSN) distributions, giving a unified scheme to obtain MSN distributions starting from symmetric ones. Subsequently, Arellano-Valle and Azzalini (2006) provided a survey on some of its extensions and variants. Sahu et al. (2003) defined a new class of MSN distributions and remarked that this sort of formulation is more flexible in terms of adjusting the correlation structure than the MSN of Azzalini and Dalla Valle (1996). Recently, Lin (2009a) introduced a new mixture modeling framework with component densities using the MSN distribution of Sahu et al. (2003) and showed its great flexibility in modeling asymmetrical data.

Learning mixture models from incomplete data has become a powerful tool to handle real-world multivariate data sets with complex missing patterns. The work was pioneered by Ghahramani and Jordan (1994), who applied the expectation maximization (EM) algorithm (Dempster et al. 1977) to compute maximum likelihood (ML) estimates of the GMIX model with arbitrary patterns of missingness. Lin et al. (2006) extended their approach by introducing some efficient learning strategies from both ML and Bayesian perspectives. Wang et al. (2004) presented an ordinary EM algorithm for ML estimation of TMIX models with missing information. Related work on using the parameter expanded Expectation Maximization (PX-EM) algorithm (Liu et al. 1998) for the supervised learning of TMIX models with incomplete data was done by Lin et al. (2009).

In this paper, we consider the learning of multivariate skew normal mixture (MSN-MIX) models when missing values may occur in the data. The probability distri-

bution of a p-dimensional random vector $\boldsymbol{X}$ is MSN with location vector $\boldsymbol{\xi} \in \mathbb{R}^p$, scale covariance matrix $\boldsymbol{\Sigma}$ (a $p \times p$ positive definite matrix) and skewness matrix $\boldsymbol{\Lambda}$ (a p-dimensional diagonal matrix), denoted as $\boldsymbol{X} \sim SN_p(\boldsymbol{\xi}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda})$. The probability density function (pdf) of this distribution is

$$\psi_p(\boldsymbol{X} \mid \boldsymbol{\xi}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}) = 2^p \, \phi_p(\boldsymbol{X} \mid \boldsymbol{\xi}, \boldsymbol{\Omega}) \Phi_p(\boldsymbol{\Lambda}\boldsymbol{\Omega}^{-1}(\boldsymbol{X} - \boldsymbol{\xi}) | \boldsymbol{\Delta}), \tag{1}$$

where $\boldsymbol{\Omega} = \boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2$, $\boldsymbol{\Delta} = (\boldsymbol{I}_p + \boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})^{-1} = \boldsymbol{I}_p - \boldsymbol{\Lambda}\boldsymbol{\Omega}^{-1}\boldsymbol{\Lambda}$ ($\boldsymbol{I}_p$ is a p-dimensional identity matrix) and the notations $\phi_p(\cdot \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\Phi_p(\cdot \mid \boldsymbol{\Sigma})$, respectively, stand for the pdf of $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and cumulative density function (cdf) of $N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$.

According to Proposition 1 of Arellano-Valle et al. (2007), the MSN distribution has the stochastic representation

$$\boldsymbol{X} = \boldsymbol{\xi} + \boldsymbol{\Lambda} | \boldsymbol{\zeta}_1 | + \boldsymbol{\Sigma}^{1/2} \boldsymbol{\zeta}_2,$$

where $\boldsymbol{\zeta}_1$ and $\boldsymbol{\zeta}_2$ are two independent $N_p(\boldsymbol{0}, \boldsymbol{I}_p)$ random vectors. Let $\boldsymbol{\gamma} = |\boldsymbol{\zeta}_1| = (|\zeta_{11}|, \ldots, |\zeta_{1p}|)^\top$. Then $\boldsymbol{\gamma}$ represents a vector consisting of p independent standard half-normal random variables. A further hierarchical representation of MSN can be written as

$$\begin{aligned} \boldsymbol{X} \mid \boldsymbol{\gamma} &\sim N_p(\boldsymbol{\xi} + \boldsymbol{\Lambda}\boldsymbol{\gamma}, \boldsymbol{\Sigma}), \\ \boldsymbol{\gamma} &\sim TN_p(\boldsymbol{0}, \boldsymbol{I}_p; \mathbb{R}^p_+), \end{aligned} \tag{2}$$

where $\mathbb{R}^p_+$ denotes the Euclidean vector space of all p-tuples of positive real numbers and $TN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbb{A})$ denotes a p-variate truncated normal distribution for $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ lying within the hyperplane $\mathbb{A}$.

In what follows we assume that the missing data are missing at random (MAR) with an ignorable mechanism (cf. Rubin 1976; Schafer 1997; Little and Rubin 2002). In this case, the missingness is unrelated to the missing values and likelihood inference can ignore the missing data mechanism. For computational aspects, we offer an analytically tractable EM algorithm coupled with some useful model-based tools to handle data with general missing patterns in the class of MSNMIX model. Note that the proposed strategy is also valid if mechanism is missing completely at random (MCAR), which is a special case of MAR. To reduce complications during the EM procedure, we introduce two permutation matrices for indexing the observed and missing components of each datum. Under this model, we also offer a conditional predictor to retrieve the missing components and a classifier for allocating partially observed vectors.

The outline of the paper is as follows. In Sect. 2, we describe the model, establish the notation, and study some important statistical properties of the model. In Sect. 3, a computationally feasible EM algorithm is used to compute the ML estimates from incomplete data. Statistical principles regarding classification and prediction of incomplete features are also investigated. In Sect. 4, the proposed methodologies are applied to a real data set with varying proportions of synthetic missing values. Some concluding remarks are given in Sect. 5, and the technical derivation is sketched in Appendix.

2 Skew normal mixtures with missing information

In the MSNMIX model, we let $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$ be a set of p-dimensional random samples arising from a population with g subclasses $\mathcal{C}_1, \ldots, \mathcal{C}_g$. That is, each $\boldsymbol{Y}_j$ has the density

$$f(\boldsymbol{Y}_j|\boldsymbol{\Theta}) = \sum_{i=1}^{g} w_i \psi_p(\boldsymbol{Y}_j \mid \boldsymbol{\xi}_i, \boldsymbol{\Sigma}_i, \boldsymbol{\Lambda}_i), \quad w_i \geq 0, \quad \sum_{i=1}^{g} \omega_i = 1, \tag{3}$$

where $\boldsymbol{\Lambda}_i = \text{Diag}(\boldsymbol{\lambda}_i)$ with $\boldsymbol{\lambda}_i = (\lambda_{i1}, \ldots, \lambda_{ip})^\top$ and the unknown parameter vector $\boldsymbol{\Theta}$ contains the mixing probabilities w_i $(i = 1, \ldots, g-1)$, the elements of component locations $\boldsymbol{\xi}_i$'s, the distinct elements of component scale covariance matrices $\boldsymbol{\Sigma}_i$'s and the skewness vectors $\boldsymbol{\lambda}_i$'s. Note that the notation $\psi_p(\cdot \mid \boldsymbol{\xi}_i, \boldsymbol{\Sigma}_i, \boldsymbol{\Lambda}_i)$ is the MSN density defined in (1) and Diag$(\cdot)$ denotes a diagonal matrix created by extracting the main diagonal elements from a square matrix or the diagonalization of a vector. The mean and covariance of $\boldsymbol{Y}_j$ are given by

$$\begin{aligned} E(\boldsymbol{Y}_j) &= \sum_{i=1}^{g} w_i \boldsymbol{\mu}_i, \\ \text{Cov}(\boldsymbol{Y}_j) &= \sum_{i=1}^{g} \left\{ w_i \left(1 - w_i\right) \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\top + w_i \boldsymbol{\Upsilon}_i \right\} - \sum_{i \neq j}^{g} w_i w_j \boldsymbol{\mu}_i \boldsymbol{\mu}_j^\top, \end{aligned}$$

where $\boldsymbol{\mu}_i = \boldsymbol{\xi}_i + \sqrt{2/\pi}\boldsymbol{\lambda}_i$ and $\boldsymbol{\Upsilon}_i = \boldsymbol{\Sigma}_i + (1 - 2/\pi)\boldsymbol{\Lambda}_i^2$ are the mean vector and covariance matrix of $SN_p(\boldsymbol{\xi}_i, \boldsymbol{\Sigma}_i, \boldsymbol{\Lambda}_i)$, respectively.

To pose model (3) into an EM framework, we introduce allocation variables $\boldsymbol{Z}_j = (Z_{1j}, \ldots, Z_{gj})^\top$, one for each individual $\boldsymbol{Y}_j$, whose role is to encode which component has generated $\boldsymbol{Y}_j$. Specifically, the indicators $\boldsymbol{Z}_j$ $(j = 1, \ldots, n)$ are a $g \times 1$ vector of binary variables, whose elements are

$$Z_{sj} = \begin{cases} 1 & \text{if } \boldsymbol{Y}_j \text{ belongs to group } s, \\ 0 & \text{otherwise,} \end{cases}$$

subject to $\sum_{i=1}^{g} Z_{ij} = 1$. This implies $\boldsymbol{Z}_j$ follows a multinomial random vector with 1 trial and cell probabilities $w_1, \ldots, w_g$, denoted by $\boldsymbol{Z}_j \sim \mathcal{M}(1; w_1, \ldots, w_g)$.

A three-level hierarchical representation of (3) can be expressed by

$$\begin{aligned} \boldsymbol{Y}_j \mid (\boldsymbol{\gamma}_j, Z_{ij} = 1) &\sim N_p(\boldsymbol{\xi}_i + \boldsymbol{\Lambda}_i \boldsymbol{\gamma}_j, \boldsymbol{\Sigma}_i), \\ \boldsymbol{\gamma}_j \mid (Z_{ij} = 1) &\sim TN_p(\boldsymbol{0}, \boldsymbol{I}_p; \mathbb{R}_+^p), \\ \boldsymbol{Z}_j &\sim \mathcal{M}(1; w_1, \ldots, w_g), \end{aligned} \tag{4}$$

for $i = 1, \ldots, g$ and $j = 1, \ldots, n$. From (4), we declare the complete data vector to be $(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\gamma})$, where $\boldsymbol{Y} = (\boldsymbol{Y}_1^\top, \ldots, \boldsymbol{Y}_n^\top)^\top$, $\boldsymbol{Z} = (\boldsymbol{Z}_1^\top, \ldots, \boldsymbol{Z}_n^\top)^\top$ and

$\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^\top, \ldots, \boldsymbol{\gamma}_n^\top)^\top$. From (4), the likelihood function for $\boldsymbol{\Theta}$ based on complete data is

$$L_c(\boldsymbol{\Theta} \mid \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\gamma}) \propto \prod_{i=1}^{g} \prod_{j=1}^{n} \left(w_i \phi_p(\boldsymbol{Y}_j \mid \boldsymbol{\xi}_i + \boldsymbol{\Lambda}_i \boldsymbol{\gamma}_j, \boldsymbol{\Sigma}_i)\right)^{Z_{ij}}. \tag{5}$$

We are interested in ML estimation of model (3) when $\boldsymbol{Y}$ may be partially observed. The underlying missingness mechanism is assumed to be MAR.

Following Lin et al. (2006), we partition $\boldsymbol{Y}_j$ into two components $(\boldsymbol{Y}_j^{\mathrm{o}\top}, \boldsymbol{Y}_j^{\mathrm{m}\top})^\top$, where $\boldsymbol{Y}_j^{\mathrm{o}}$ ($p_j^{\mathrm{o}} \times 1$) and $\boldsymbol{Y}_j^{\mathrm{m}}$ ($(p - p_j^{\mathrm{o}}) \times 1$) denote the observed and missing components of $\boldsymbol{Y}_j$, respectively. To facilitate the computation, we introduce two types of the binary indicator matrices, denoted by $\boldsymbol{O}_j$ ($p_j^{\mathrm{o}} \times p$) and $\boldsymbol{M}_j$ ($(p - p_j^{\mathrm{o}}) \times p$) such that $\boldsymbol{Y}_j^{\mathrm{o}} = \boldsymbol{O}_j \boldsymbol{Y}_j$ and $\boldsymbol{Y}_j^{\mathrm{m}} = \boldsymbol{M}_j \boldsymbol{Y}_j$, which can be extracted from a p-dimensional identity matrix $\boldsymbol{I}_p$ corresponding to row positions of $\boldsymbol{Y}_j^{\mathrm{o}}$ and $\boldsymbol{Y}_j^{\mathrm{m}}$ in $\boldsymbol{Y}_j$, respectively. It is straightforward to verify that (a) $\boldsymbol{Y}_j = \boldsymbol{O}_j^\top \boldsymbol{Y}_j^{\mathrm{o}} + \boldsymbol{M}_j^\top \boldsymbol{Y}_j^{\mathrm{m}}$; (b) $\boldsymbol{O}_j^\top \boldsymbol{O}_j + \boldsymbol{M}_j^\top \boldsymbol{M}_j = \boldsymbol{I}_p$. Furthermore, we can establish the following results.

Theorem 1 *Let $\boldsymbol{Y}_j \sim \sum_{i=1}^{g} w_i \psi_p(\boldsymbol{Y}_j \mid \boldsymbol{\xi}_i, \boldsymbol{\Sigma}_i, \boldsymbol{\Lambda}_i)$, and let $\boldsymbol{Y}_j^{\mathrm{o}}$ and $\boldsymbol{Y}_j^{\mathrm{m}}$ be the observed and the missing components corresponding to $\boldsymbol{Y}_j$, respectively. We have*

(a) *The marginal density of $\boldsymbol{Y}_j^{\mathrm{o}}$ is $\sum_{i=1}^{g} w_i \psi_{p_j^{\mathrm{o}}}(\boldsymbol{Y}_j^{\mathrm{o}} | \boldsymbol{\xi}_{ij}^{\mathrm{o}}, \boldsymbol{\Sigma}_{ij}^{\mathrm{oo}}, \boldsymbol{\Lambda}_{ij}^{\mathrm{oo}})$, where $\boldsymbol{\xi}_{ij}^{\mathrm{o}} = \boldsymbol{O}_j \boldsymbol{\xi}_i$, $\boldsymbol{\Sigma}_{ij}^{\mathrm{oo}} = \boldsymbol{O}_j \boldsymbol{\Sigma}_i \boldsymbol{O}_j^\top$ and $\boldsymbol{\Lambda}_{ij}^{\mathrm{oo}} = \boldsymbol{O}_j \boldsymbol{\Lambda}_i \boldsymbol{O}_j^\top$.*

(b) *The conditional density of $\boldsymbol{Y}_j^{\mathrm{m}}$ given $\boldsymbol{Y}_j^{\mathrm{o}}$ is*

$$f(\boldsymbol{Y}_j^{\mathrm{m}} \mid \boldsymbol{Y}_j^{\mathrm{o}}) = 2^p \sum_{i=1}^{g} \tilde{w}_{ij} \phi_{p-p_j^{\mathrm{o}}}\left(\boldsymbol{Y}_j^{\mathrm{m}} \mid \boldsymbol{\xi}_{ij}^{\mathrm{m}\cdot\mathrm{o}}, \boldsymbol{\Omega}_{ij}^{\mathrm{mm}\cdot\mathrm{o}}\right) \Phi_p\left(\boldsymbol{\Lambda}_i \boldsymbol{\Omega}_i^{-1}(\boldsymbol{Y}_j - \boldsymbol{\xi}_i) \mid \boldsymbol{\Delta}_i\right),$$

where

$$\begin{aligned}
\tilde{w}_{ij} &= w_i \phi_{p_j^{\mathrm{o}}}\left(\boldsymbol{Y}_j^{\mathrm{o}} \mid \boldsymbol{\xi}_{ij}^{\mathrm{o}}, \boldsymbol{\Omega}_{ij}^{\mathrm{oo}}\right) / \sum_{h=1}^{g} w_h \psi_{p_j^{\mathrm{o}}}\left(\boldsymbol{Y}_j^{\mathrm{o}} \mid \boldsymbol{\xi}_{hj}^{\mathrm{o}}, \boldsymbol{\Sigma}_{hj}^{\mathrm{oo}}, \boldsymbol{\Lambda}_{hj}^{\mathrm{oo}}\right), \\
\boldsymbol{\xi}_{ij}^{\mathrm{m}\cdot\mathrm{o}} &= \boldsymbol{M}_j\left(\boldsymbol{\xi}_i + \boldsymbol{\Omega}_i \boldsymbol{C}_{ij}^{\mathrm{oo}}(\boldsymbol{Y}_j - \boldsymbol{\xi}_i)\right), \\
\boldsymbol{\Omega}_{ij}^{\mathrm{mm}\cdot\mathrm{o}} &= \boldsymbol{M}_j(\boldsymbol{I}_p - \boldsymbol{\Omega}_i \boldsymbol{C}_{ij}^{\mathrm{oo}})\boldsymbol{\Omega}_i \boldsymbol{M}_j^\top,
\end{aligned}$$

with $\boldsymbol{\Omega}_{ij}^{\mathrm{oo}} = \boldsymbol{O}_j \boldsymbol{\Omega}_i \boldsymbol{O}_j^\top$ and $\boldsymbol{C}_{ij}^{\mathrm{oo}} = \boldsymbol{O}_j^\top \boldsymbol{\Omega}_{ij}^{\mathrm{oo}^{-1}} \boldsymbol{O}_j$.

Theorem 2 *Given (4), we have the following conditional distributions:*

(a) *The conditional distribution of $\boldsymbol{Y}_j^{\mathrm{o}}$ given $\boldsymbol{\gamma}_j$ and $Z_{ij} = 1$ is*

$$\boldsymbol{Y}_j^{\mathrm{o}} \mid (\boldsymbol{\gamma}_j, Z_{ij} = 1) \sim N_{p_j^{\mathrm{o}}}\left(\boldsymbol{\mu}_{ij}^{\mathrm{o}}, \boldsymbol{\Sigma}_{ij}^{\mathrm{oo}}\right),$$

where $\boldsymbol{\mu}_{ij}^{\mathrm{o}} = \boldsymbol{O}_j(\boldsymbol{\xi}_i + \boldsymbol{\Lambda}_i \boldsymbol{\gamma}_j)$ and $\boldsymbol{\Sigma}_{ij}^{\mathrm{oo}} = \boldsymbol{O}_j \boldsymbol{\Sigma}_i \boldsymbol{O}_j^\top$.

(b) *The conditional distribution of* $\boldsymbol{Y}_j^{\mathrm{m}}$ *given* $\boldsymbol{Y}_j^{\mathrm{o}}$, $\boldsymbol{\gamma}_j$, *and* $Z_{ij} = 1$ *is*

$$\boldsymbol{Y}_j^{\mathrm{m}} \mid (\boldsymbol{Y}_j^{\mathrm{o}}, \boldsymbol{\gamma}_j, Z_{ij} = 1) \sim N_{p-p_j^{\mathrm{o}}}(\boldsymbol{\mu}_{ij}^{\mathrm{m\cdot o}}, \boldsymbol{\Sigma}_{ij}^{\mathrm{mm\cdot o}}),$$

where

$$\begin{aligned}\boldsymbol{\mu}_{ij}^{\mathrm{m\cdot o}} &= \boldsymbol{M}_j(\boldsymbol{\xi}_i + \boldsymbol{\Lambda}_i\boldsymbol{\gamma}_j + \boldsymbol{\Sigma}_i\boldsymbol{S}_{ij}^{\mathrm{oo}}(\boldsymbol{Y}_j - \boldsymbol{\xi}_i - \boldsymbol{\Lambda}_i\boldsymbol{\gamma}_j)),\\ \boldsymbol{\Sigma}_{ij}^{\mathrm{mm\cdot o}} &= \boldsymbol{M}_j(\boldsymbol{I}_p - \boldsymbol{\Sigma}_i\boldsymbol{S}_{ij}^{\mathrm{oo}})\boldsymbol{\Sigma}_i\boldsymbol{M}_j^{\top},\end{aligned}$$

and $\boldsymbol{S}_{ij}^{\mathrm{oo}} = \boldsymbol{O}_j^{\top}(\boldsymbol{O}_j\boldsymbol{\Sigma}_i\boldsymbol{O}_j^{\top})^{-1}\boldsymbol{O}_j$.

(c) *The conditional distribution of* $\boldsymbol{\gamma}_j$ *given* $\boldsymbol{Y}_j^{\mathrm{o}}$ *and* $Z_{ij} = 1$ *is*

$$\boldsymbol{\gamma}_j \mid (\boldsymbol{Y}_j^{\mathrm{o}}, Z_{ij} = 1) \sim TN_p\left(\boldsymbol{\Lambda}_i\boldsymbol{C}_{ij}^{\mathrm{oo}}(\boldsymbol{Y}_j - \boldsymbol{\xi}_i), \boldsymbol{I}_p - \boldsymbol{\Lambda}_i\boldsymbol{C}_{ij}^{\mathrm{oo}}\boldsymbol{\Lambda}_i; \mathbb{R}_+^p\right).$$

Let $E\left(\boldsymbol{\gamma}_j|\boldsymbol{Y}_j^{\mathrm{o}}, Z_{ij} = 1\right) = \boldsymbol{\eta}_{ij}$ and $E\left(\boldsymbol{\gamma}_j\boldsymbol{\gamma}_j^{\top}|\boldsymbol{Y}_j^{\mathrm{o}}, Z_{ij} = 1\right) = \boldsymbol{\Psi}_{ij}$. Both of which are implicit functions of parameters $\boldsymbol{\xi}_i$, $\boldsymbol{\Sigma}_i$ and $\boldsymbol{\Lambda}_i$ and can easily be evaluated by using formulas (10) and (11) in Lin (2009a). The following corollary is a direct implication of Theorem 2.

Corollary 1 *Recall that* $\boldsymbol{Y}_j = \boldsymbol{O}_j^{\top}\boldsymbol{Y}_j^{\mathrm{o}} + \boldsymbol{M}_j^{\top}\boldsymbol{Y}_j^{\mathrm{m}}$ *and* $\boldsymbol{O}_j^{\top}\boldsymbol{O}_j + \boldsymbol{M}_j^{\top}\boldsymbol{M}_j = \boldsymbol{I}_p$. *These give rise to* $\boldsymbol{O}_j^{\top}\boldsymbol{O}_j(\boldsymbol{I}_p - \boldsymbol{\Sigma}_i\boldsymbol{S}_{ij}^{\mathrm{oo}}) = \boldsymbol{0}$. *Then we can obtain*

(a) $E(\boldsymbol{Y}_j \mid \boldsymbol{Y}_j^{\mathrm{o}}, Z_{ij} = 1) = \boldsymbol{\xi}_i + \boldsymbol{\Lambda}_i\boldsymbol{\eta}_{ij} + \boldsymbol{\Sigma}_i\boldsymbol{S}_{ij}^{\mathrm{oo}}(\boldsymbol{Y}_j - \boldsymbol{\xi}_i - \boldsymbol{\Lambda}_i\boldsymbol{\eta}_{ij})$.

(b) $\mathrm{Cov}(\boldsymbol{Y}_j|\boldsymbol{Y}_j^{\mathrm{o}}, Z_{ij}=1) = (\boldsymbol{I}_p - \boldsymbol{\Sigma}_i\boldsymbol{S}_{ij}^{\mathrm{oo}})\left(\boldsymbol{\Sigma}_i + \boldsymbol{\Lambda}_i(\boldsymbol{\Psi}_{ij} - \boldsymbol{\eta}_{ij}\boldsymbol{\eta}_{ij}^{\top})\boldsymbol{\Lambda}_i(\boldsymbol{I}_p - \boldsymbol{S}_{ij}^{\mathrm{oo}}\boldsymbol{\Sigma}_i)\right)$.

(c) $E(\boldsymbol{Y}_j\boldsymbol{\gamma}_j^{\top} \mid \boldsymbol{Y}_j^{\mathrm{o}}, Z_{ij} = 1) = (\boldsymbol{I}_p - \boldsymbol{\Sigma}_i\boldsymbol{S}_{ij}^{\mathrm{oo}})(\boldsymbol{\xi}_i\boldsymbol{\eta}_{ij}^{\top} + \boldsymbol{\Lambda}_i\boldsymbol{\Psi}_{ij}) + \boldsymbol{\Sigma}_i\boldsymbol{S}_{ij}^{\mathrm{oo}}\boldsymbol{Y}_j\boldsymbol{\eta}_{ij}^{\top}$.

For a p-dimensional observation with the probability of missingness greater than or equal to zero for each attribute, there are $2^p - 1$ unique patterns of missingness. In general, completely missing pattern does not occur. This indicates that the missing rate should be smaller than $(p-1)/p$. To lessen the computational load, Lin et al. (2006) have described a simple and feasible procedure by rearranging $\boldsymbol{Y}$ according to unique missing patterns of data.

3 ML estimation via the EM algorithm

The EM algorithm of Dempster et al. (1977) has been widely used in literature to carry out ML estimation in a variety of incomplete data problems. We offer an efficient EM algorithm for learning model (3) from incomplete data. From (5), the log-likelihood function of $\boldsymbol{\Theta}$ based on complete data, aside from additive constant terms, can be written by

$$\begin{aligned}&\ell_c(\boldsymbol{\Theta} \mid \boldsymbol{Y}^{\mathrm{o}}, \boldsymbol{Y}^{\mathrm{m}}, \boldsymbol{Z}, \boldsymbol{\gamma})\\ &= \sum_{i=1}^{g}\sum_{j=1}^{n} Z_{ij}\log w_i + \frac{1}{2}\sum_{i=1}^{g}\left(\log|\boldsymbol{\Sigma}_i^{-1}|\left(\sum_{j=1}^{n} Z_{ij}\right) - \mathrm{tr}\left(\boldsymbol{\Sigma}_i^{-1}\sum_{j=1}^{n}\boldsymbol{H}_{ij}\right)\right), \quad (6)\end{aligned}$$

where

$$\boldsymbol{H}_{ij} = Z_{ij}(\boldsymbol{Y}_j - \boldsymbol{\xi}_i - \boldsymbol{\Lambda}_i\boldsymbol{\gamma}_j)(\boldsymbol{Y}_j - \boldsymbol{\xi}_i - \boldsymbol{\Lambda}_i\boldsymbol{\gamma}_j)^\top. \tag{7}$$

At the kth iteration of the E-step, we need to calculate the Q-function, defined as

$$Q(\boldsymbol{\Theta} \mid \hat{\boldsymbol{\Theta}}^{(k)}) = E\left(\ell_c(\boldsymbol{\Theta} \mid \boldsymbol{Y}^{\mathrm{o}}, \boldsymbol{Y}^{\mathrm{m}}, \boldsymbol{Z}, \boldsymbol{\gamma}) \mid \boldsymbol{Y}^{\mathrm{o}}, \hat{\boldsymbol{\Theta}}^{(k)}\right)$$

which is the conditional expectation of (6) with respect to the distribution of $(\boldsymbol{Y}^{\mathrm{m}}, \boldsymbol{Z}, \boldsymbol{\gamma})$ given the observed data $\boldsymbol{Y}^{\mathrm{o}}$ and the current estimate $\hat{\boldsymbol{\Theta}}^{(k)}$.

Let $\hat{Z}_{ij}^{(k)}$ denote the posterior probability that $\boldsymbol{Y}_j$ arises from the ith component. By Bayes' rule, at the kth iteration, we have

$$\hat{Z}_{ij}^{(k)} = \Pr(Z_{ij} = 1 \mid \boldsymbol{Y}_j^{\mathrm{o}}, \hat{\boldsymbol{\Theta}}^{(k)}) = \frac{\hat{w}_i^{(k)}\psi_{p_j^{\mathrm{o}}}\left(\boldsymbol{Y}_j^{\mathrm{o}} \mid \hat{\boldsymbol{\xi}}_{ij}^{\mathrm{o}(k)}, \hat{\boldsymbol{\Sigma}}_{ij}^{\mathrm{oo}(k)}, \hat{\boldsymbol{\Lambda}}_{ij}^{\mathrm{oo}(k)}\right)}{\sum_{h=1}^{g}\hat{w}_h^{(k)}\psi_{p_j^{\mathrm{o}}}\left(\boldsymbol{Y}_j^{\mathrm{o}} \mid \hat{\boldsymbol{\xi}}_{hj}^{\mathrm{o}(k)}, \hat{\boldsymbol{\Sigma}}_{hj}^{\mathrm{oo}(k)}, \hat{\boldsymbol{\Lambda}}_{hj}^{\mathrm{oo}(k)}\right)}, \tag{8}$$

where $\hat{\boldsymbol{\xi}}_{ij}^{\mathrm{o}(k)} = \boldsymbol{O}_j\hat{\boldsymbol{\xi}}_i^{(k)}$, $\hat{\boldsymbol{\Sigma}}_{ij}^{\mathrm{oo}(k)} = \boldsymbol{O}_j\hat{\boldsymbol{\Sigma}}_i^{(k)}\boldsymbol{O}_j^\top$ and $\hat{\boldsymbol{\Lambda}}_{ij}^{\mathrm{oo}(k)} = \boldsymbol{O}_j\hat{\boldsymbol{\Lambda}}_i^{(k)}\boldsymbol{O}_j^\top$ for $i = 1, \ldots, g$ and $j = 1, \ldots, n$. Furthermore, it can be shown that the expected value of (7) conditional on $\boldsymbol{Y}_j^{\mathrm{o}}$ and current estimates $\hat{\boldsymbol{\Theta}}^{(k)}$ is

$$\begin{aligned}
&\hat{\boldsymbol{H}}_{ij}^{(k)}(\boldsymbol{\xi}_i, \boldsymbol{\Lambda}_i) \\
&= \hat{Z}_{ij}^{(k)}\left(\left(\boldsymbol{I}_p - \hat{\boldsymbol{\Sigma}}_i^{(k)}\hat{\boldsymbol{S}}_{ij}^{\mathrm{oo}(k)}\right)\hat{\boldsymbol{\Sigma}}_i^{(k)} + \left(\hat{\boldsymbol{Y}}_{ij}^{(k)} - \boldsymbol{\xi}_i - \boldsymbol{\Lambda}_i\hat{\boldsymbol{\eta}}_{ij}^{(k)}\right)\left(\hat{\boldsymbol{Y}}_{ij}^{(k)} - \boldsymbol{\xi}_i - \boldsymbol{\Lambda}_i\hat{\boldsymbol{\eta}}_{ij}^{(k)}\right)^\top\right. \\
&\quad \left. + \left(\hat{\boldsymbol{\Lambda}}_{ij}^{(k)} - \boldsymbol{\Lambda}_i\right)\left(\hat{\boldsymbol{\Psi}}_{ij}^{(k)} - \hat{\boldsymbol{\eta}}_{ij}^{(k)}\hat{\boldsymbol{\eta}}_{ij}^{(k)\top}\right)\left(\hat{\boldsymbol{\Lambda}}_{ij}^{(k)} - \boldsymbol{\Lambda}_i\right)^\top\right),
\end{aligned} \tag{9}$$

where $\hat{\boldsymbol{S}}_{ij}^{\mathrm{oo}(k)} = \boldsymbol{O}_j^\top\left(\boldsymbol{O}_j\hat{\boldsymbol{\Sigma}}_i^{(k)}\boldsymbol{O}_j^\top\right)^{-1}\boldsymbol{O}_j$, $\hat{\boldsymbol{\Lambda}}_{ij}^{(k)} = \left(\boldsymbol{I}_p - \hat{\boldsymbol{\Sigma}}_i^{(k)}\hat{\boldsymbol{S}}_{ij}^{\mathrm{oo}(k)}\right)\hat{\boldsymbol{\Lambda}}_i^{(k)}$ and

$$\begin{aligned}
\hat{\boldsymbol{Y}}_{ij}^{(k)} &= E\left(\boldsymbol{Y}_j \mid Z_{ij} = 1, \boldsymbol{Y}^{\mathrm{o}}, \hat{\boldsymbol{\Theta}}^{(k)}\right) \\
&= \hat{\boldsymbol{\Sigma}}_i^{(k)}\hat{\boldsymbol{S}}_{ij}^{\mathrm{oo}(k)}\boldsymbol{Y}_j + \left(\boldsymbol{I}_p - \hat{\boldsymbol{\Sigma}}_i^{(k)}\hat{\boldsymbol{S}}_{ij}^{\mathrm{oo}(k)}\right)\left(\hat{\boldsymbol{\xi}}_i^{(k)} + \hat{\boldsymbol{\Lambda}}_i^{(k)}\hat{\boldsymbol{\eta}}_{ij}^{(k)}\right).
\end{aligned} \tag{10}$$

Hence, the Q-function can be written by

$$\begin{aligned}
Q(\boldsymbol{\Theta} \mid \hat{\boldsymbol{\Theta}}^{(k)}) = &\sum_{i=1}^{g}\sum_{j=1}^{n}\hat{Z}_{ij}^{(k)}\log w_i \\
&+\frac{1}{2}\sum_{i=1}^{g}\left(\log|\boldsymbol{\Sigma}_i^{-1}|\left(\sum_{j=1}^{n}\hat{Z}_{ij}^{(k)}\right) - \mathrm{tr}\left(\boldsymbol{\Sigma}_i^{-1}\sum_{j=1}^{n}\hat{\boldsymbol{H}}_{ij}^{(k)}(\boldsymbol{\xi}_i, \boldsymbol{\Lambda}_i)\right)\right).
\end{aligned} \tag{11}$$

In summary, the EM algorithm can be implemented as follows:
E-step: Given $\boldsymbol{\Theta} = \hat{\boldsymbol{\Theta}}^{(k)}$, compute $\hat{Z}_{ij}^{(k)}$, $\hat{\boldsymbol{H}}_{ij}^{(k)}$ and $\hat{\boldsymbol{Y}}_{ij}^{(k)}$ for $i = 1, \ldots, g$ and $j = 1, \ldots, n$, by using Eqs. 8–10, respectively.
M-step:

1. Update $\hat{w}_i^{(k)}$ by maximizing (11) over w_i subject to their sum being unity, which gives

$$\hat{w}_i^{(k+1)} = n^{-1} \sum_{j=1}^{n} \hat{Z}_{ij}^{(k)}.$$

2. Update $\hat{\boldsymbol{\xi}}_i^{(k)}$ by

$$\hat{\boldsymbol{\xi}}_i^{(k+1)} = \left(\sum_{j=1}^{n} \hat{Z}_{ij}^{(k)} \right)^{-1} \left(\sum_{j=1}^{n} \hat{Z}_{ij}^{(k)} \hat{\boldsymbol{Y}}_{ij}^{(k)} - \hat{\boldsymbol{\Lambda}}_i^{(k)} \sum_{j=1}^{n} \hat{Z}_{ij}^{(k)} \hat{\boldsymbol{\eta}}_{ij}^{(k)} \right).$$

3. Update $\hat{\boldsymbol{\Sigma}}_i^{(k)}$ by

$$\hat{\boldsymbol{\Sigma}}_i^{(k+1)} = \frac{\sum_{j=1}^{n} \hat{\boldsymbol{H}}_{ij}^{(k)} (\hat{\boldsymbol{\xi}}^{(k+1)}, \hat{\boldsymbol{\Lambda}}^{(k)})}{\sum_{j=1}^{n} \hat{Z}_{ij}^{(k)}}.$$

4. Update $\hat{\boldsymbol{\Lambda}}_i^{(k)}$ by

$$\hat{\boldsymbol{\Lambda}}_i^{(k+1)} = \text{Diag} \left\{ \left(\hat{\boldsymbol{\Sigma}}_i^{(k+1)^{-1}} \odot \sum_{j=1}^{n} \hat{Z}_{ij}^{(k)} \hat{\boldsymbol{\Psi}}_{ij}^{(k)} \right)^{-1} \left(\hat{\boldsymbol{\Sigma}}_i^{(k+1)^{-1}} \odot \sum_{j=1}^{n} \hat{Z}_{ij}^{(k)} \hat{\mathbf{C}}_{ij}^{(k)} \right) \mathbf{1}_p \right\},$$

where $\hat{\mathbf{C}}_{ij}^{(k)} = \left(\hat{\boldsymbol{\Psi}}_{ij}^{(k)} - \hat{\boldsymbol{\eta}}_{ij}^{(k)} \hat{\boldsymbol{\eta}}_{ij}^{(k)\top} \right) \hat{\mathbf{A}}_{ij}^{\top(k)} + \hat{\boldsymbol{\eta}}_{ij}^{(k)} \left(\hat{\boldsymbol{Y}}_{ij}^{(k)} - \hat{\boldsymbol{\xi}}_i^{(k+1)} \right)^{\top}$, $\mathbf{1}_p$ denotes a p-dimensional vector of ones and the operator '$\odot$' represents the elementwise product of two matrices with the same dimension.

Since the stability and monotone convergence of EM are maintained, the iterations are repeated until a suitable convergence rule is satisfied, e.g., $\|\hat{\boldsymbol{\Theta}}^{(k+1)} - \hat{\boldsymbol{\Theta}}^{(k)}\|$ is sufficiently small. When the convergence is achieved, the resulting estimates are denoted by $\hat{\boldsymbol{\Theta}} = (\hat{w}_1, \ldots, \hat{w}_{g-1}, \hat{\boldsymbol{\xi}}_1, \ldots, \hat{\boldsymbol{\xi}}_g, \hat{\boldsymbol{\Sigma}}_1, \ldots, \hat{\boldsymbol{\Sigma}}_g, \hat{\boldsymbol{\Lambda}}_1, \ldots, \hat{\boldsymbol{\Lambda}}_g)$. Therefore, the posterior probability of the $\boldsymbol{Y}_j$ belonging to group i can be estimated by

$$\hat{w}_{ij}^{*} = P(Z_{ij} = 1 | \boldsymbol{Y}^{\text{o}}, \hat{\boldsymbol{\Theta}}) = \frac{\hat{w}_i \psi_{p_j^{\text{o}}} \left(\boldsymbol{Y}_j^{\text{o}} \mid \hat{\boldsymbol{\xi}}_{ij}^{\text{o}}, \hat{\boldsymbol{\Sigma}}_{ij}^{\text{oo}}, \hat{\boldsymbol{\Lambda}}_{ij}^{\text{oo}} \right)}{\sum_{i=1}^{g} \hat{w}_i \psi_{p_j^{\text{o}}} \left(\boldsymbol{Y}_j^{\text{o}} \mid \hat{\boldsymbol{\xi}}_{ij}^{\text{o}}, \hat{\boldsymbol{\Sigma}}_{ij}^{\text{oo}}, \hat{\boldsymbol{\Lambda}}_{ij}^{\text{oo}} \right)}. \tag{12}$$

According to the ML classification theory of Basford and McLachlan (1985), $\boldsymbol{Y}_j$ is assigned to group s if $\hat{w}^*_{sj} > \hat{w}^*_{ij}$ for $i = 1, \ldots, g$ and $i \neq s$. Consequently, the ML predictor for the missing component $\boldsymbol{Y}^{\mathrm{m}}_j$ is given by

$$\begin{aligned}\hat{\boldsymbol{Y}}^{\mathrm{m}}_j &= E\left(\boldsymbol{Y}^{\mathrm{m}}_j | \boldsymbol{Y}^{\mathrm{o}}_j, \hat{\boldsymbol{\Theta}}\right) \\ &= \boldsymbol{M}_j \sum_{i=1}^{g} \hat{w}^*_{ij} \left(\hat{\boldsymbol{\xi}}_i + \hat{\boldsymbol{\Lambda}}_i \hat{\boldsymbol{\eta}}_{ij} + \hat{\boldsymbol{\Sigma}}_i \hat{\boldsymbol{S}}^{\mathrm{oo}}_{ij} (\boldsymbol{Y}_j - \hat{\boldsymbol{\xi}}_i - \hat{\boldsymbol{\Lambda}}_i \hat{\boldsymbol{\eta}}_{ij})\right). \end{aligned} \tag{13}$$

4 Experimental results

For illustration purposes, we apply the techniques presented so far to a subset of the Australian Institute of Sport (AIS) data, including 13 physical attributes measured on 102 male and 100 female athletes, which are treated as two intrinsic classes. The data were originally reported by Cook and Weisberg (1994) and have already been analyzed by Azzalini and Dalla Valle (1996), Azzalini and Capitanio (1999) and Azzalini (2005), among others. They pointed out the AIS data are better suited to the MSN distribution than Gaussian, but neglected the situation where patterns of multimodality occur. In this example, we select three attributes: X_1: body mass index (BMI), X_2 : the percentage of body fat (Bfat) and X_3 : lean body mass (LBM). Detailed explanations of these variables can be found at http://en.wikipedia.org/wiki/Search.

Figure 1 depicts pairwise bivariate scatter plots of the data with superimposed contours of the fitted 2-component MSNMIX distribution. It can be observed from the figure that the scatter plots and fitted densities reveal a pattern of asymmetric bimodality for each pair of attributes except that the bimodality of BMI versus LBM is not apparent because they are highly correlated. The Pearson's correlation coefficient between these two attributes is 0.71.

To conduct experimental studies, synthetic missing values are generated according to both MCAR and MAR mechanisms, as shown in Zio et al. (2007). In the MCAR experiment, missing items are obtained by deleting at random $r\%$ of the experimental data where each datum retains at least one observed attribute. The missing rates of the synthetic data range from 0.1 up to 0.4 by increments of 0.1. In the MAR case, missing items are only drawn form the attributes (X_1, X_2) depending on the observed values of X_3 under the assumption that the higher the value of X_3 the lower is the nonresponse propensity. Let q_i be the ith quartile of the empirical distribution of X_3 and x_i be the observed value of X_i. The nonresponse probabilities for (X_1, X_2) are 0.25 if $x_3 < q_1$, 0.2 if $x_3 \in [q_1, q_2)$, 0.15 if $x_3 \in [q_2, q_3)$ and 0.1 if $x_3 \geq q_3$. Data were simulated with a total of 500 and 100 replications under the MCAR and MAR settings, respectively. A relative difference of 10^{-5} in successive values of the log-likelihood is used as a stopping guideline for the EM algorithm.

We fit a MSNMIX model with density (1) to 500 synthetic MCAR data sets for $g = 1$ and $g = 2$. Here the number of components $g = 1$ corresponds to the MSN model (a special case of MSNMIX model with a single component) of Sahu et al. (2003), which cannot capture the bimodality and $g = 2$ corresponds to a 2-component

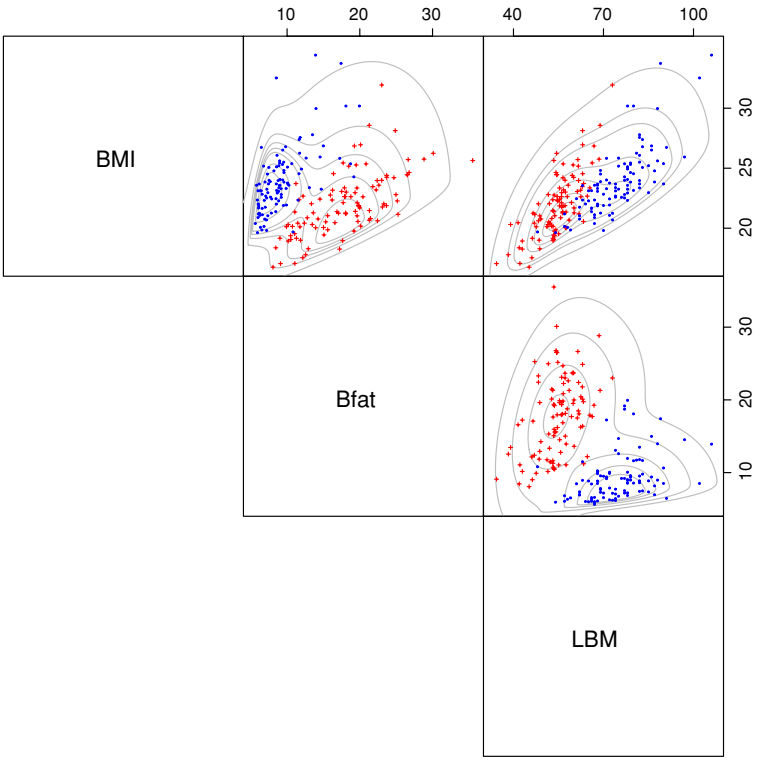

Fig. 1 AIS data: bivariate scatter plots and fitted 2-component MSNMIX contours. (*Plus sign* female, *filled circle* male, *BMI* body mass index, *Bfat* the percentage of body fat, *LBM* lean body mass)

MSNMIX model as written below

$$f(\boldsymbol{Y}_j|\boldsymbol{\Theta}) = wf(\boldsymbol{Y}_j|\boldsymbol{\xi}_1, \boldsymbol{\Sigma}_1, \boldsymbol{\Lambda}_1) + (1-w)f(\boldsymbol{Y}_j|\boldsymbol{\xi}_2, \boldsymbol{\Sigma}_2, \boldsymbol{\Lambda}_2) \quad (j = 1, \ldots, 202),$$

where

$$\boldsymbol{\xi}_i = \begin{bmatrix} \xi_{i1} \\ \xi_{i2} \\ \xi_{i3} \end{bmatrix}, \quad \boldsymbol{\Sigma}_i = \begin{bmatrix} \sigma_{i,11} & \sigma_{i,12} & \sigma_{i,13} \\ \sigma_{i,12} & \sigma_{i,22} & \sigma_{i,23} \\ \sigma_{i,13} & \sigma_{i,23} & \sigma_{i,33} \end{bmatrix} \quad \boldsymbol{\Lambda}_i = \begin{bmatrix} \lambda_{i,11} & 0 & 0 \\ 0 & \lambda_{i,22} & 0 \\ 0 & 0 & \lambda_{i,33} \end{bmatrix} \quad (i = 1, 2).$$

For comparison, we test the null hypothesis $H_0 : g = 1$ (MSN) *versus* the alternative hypothesis $H_1 : g = 2$ (MSNMIX). The numbers of free parameters under H_0 and H_1 are 12 and 25, respectively. The likelihood ratio test (LRT) statistic, given by the difference in values of -2 times the log-likelihood between two test models, is used to judge which of the two models is more suitable for this data set. Figure 2 displays the histograms of converged log-likelihood values of the null and the alternative models along with a summarized box plot for their LRT statistics. It is readily seen that the LRT statistics are highly significant compared with the χ^2_{13} distribution for all cases.

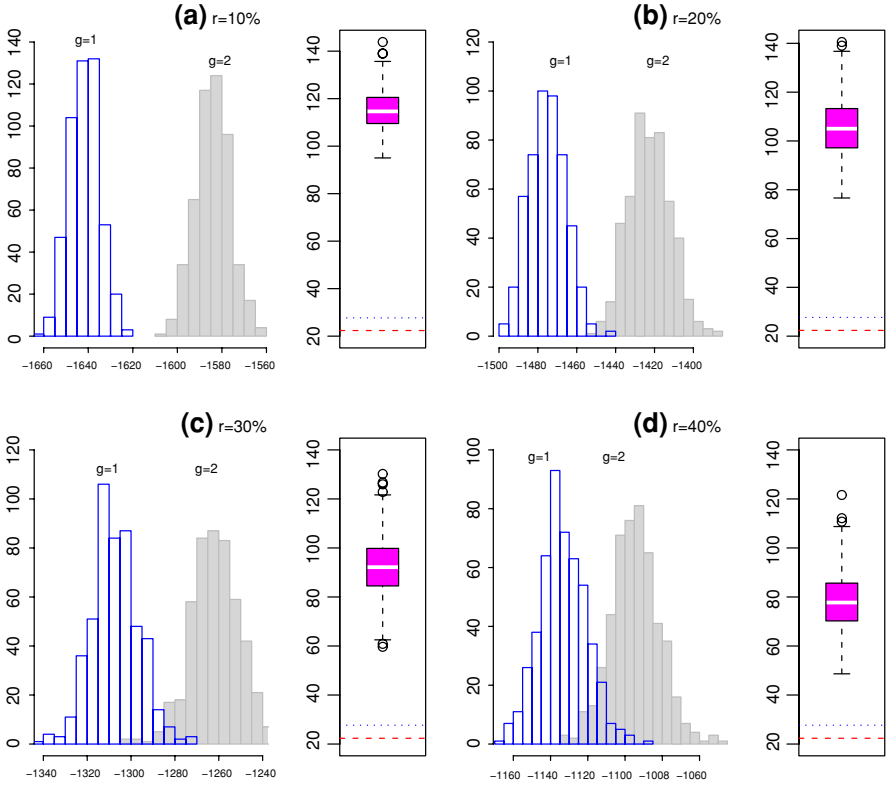

Fig. 2 A comparison of converged log-likelihood values of the null ($g = 1$) and the alternative ($g = 2$) models and their LRT statistics, where *dotted line* ($= \chi^2_{13}(0.99) = 27.69$ and *dashed line* ($= \chi^2_{13}(0.95) = 22.36$, for various proportions of missing values. (Replications=500)

To exemplify the predictive accuracies on the imputation of missing values, we compare the MSN and MSNMIX predictors; see Eq. 13, together with the traditional randomization-based mean imputation (MI) predictor, known as a common heuristic by filling in a single value for each missing value with the observed sample mean of the associated attribute. As a measure of precision, the mean absolute error (MAE) and the mean absolute relative error (MARE) are used to evaluate the prediction discrepancy. They are defined as

$$\text{MAE} = \frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{p}|y_{ij} - \hat{y}_{ij}| \quad \text{and} \quad \text{MARE} = \frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{p}\left|\frac{y_{ij} - \hat{y}_{ij}}{y_{ij}}\right|,$$

where m is the number of missing entries, y_{ij} is the actual value and $\hat{y}_{ij}$ is the respective predictive value.

Comparison results for both MCAR and MAR scenarios are listed in Table 1. The relative improvement percentage (RIP) in Table 1 is defined as the percentage decrease in the relative prediction error when comparing MSN and MSNMIX predictors. In this

Table 1 A comparison of averaged prediction accuracies and the associated standard deviations in parentheses for three imputation methods with varying proportions of missing values

Mechanism	Missing rate (%)	MAE				MARE			
		MI	MSN	MSNMIX	RIP (%)	MI	MSN	MSNMIX	RIP (%)
MCAR	10	6.114 (0.733)	3.476 (0.495)	3.326 (0.499)	4.32	0.243 (0.032)	0.154 (0.024)	0.141 (0.023)	8.44
	20	6.124 (0.493)	3.777 (0.321)	3.637 (0.337)	3.71	0.246 (0.024)	0.166 (0.032)	0.155 (0.018)	6.63
	30	6.115 (0.380)	4.096 (0.300)	3.967 (0.319)	3.15	0.244 (0.018)	0.177 (0.015)	0.167 (0.015)	5.65
	40	6.115 (0.302)	4.382 (0.284)	4.275 (0.307)	2.44	0.245 (0.017)	0.190 (0.014)	0.182 (0.015)	4.21
MAR		3.845 (0.320)	2.572 (0.252)	2.436 (0.260)	5.29	0.270 (0.029)	0.191 (0.029)	0.178 (0.031)	6.87

The relative improvement percentage (RIP) is measured by [(MSN-MSNMIX)/MSN]×100% MI mean imputation, MAE the mean absolute error, $MARE$ the mean absolute relative error

Table 2 A comparison of average misclassification rates (%) between GMIX and MSNMIX models with standard deviations in parentheses

Mechanism	Missing rate (%)	Trivariate case			Bivariate case		
		GMIX	MSNMIX	RIP (%)	GMIX	MSNMIX	RIP (%)
MCAR	10	5.35 (0.015)	4.77 (0.010)	10.8	27.36 (0.129)	10.93 (0.022)	60.1
	20	6.99 (0.024)	6.24 (0.014)	12.0	31.71 (0.136)	14.60 (0.028)	54.0
	30	9.84 (0.034)	8.61 (0.018)	12.5	34.75 (0.123)	18.68 (0.036)	46.2
	40	13.11 (0.037)	11.63 (0.023)	11.3	36.20 (0.105)	20.52 (0.020)	43.3
MAR		5.33 (0.009)	4.85 (0.009)	9.0	20.55 (0.120)	7.91 (0.013)	61.5

The relative improvement percentage (RIP) is measured by [(GMIX-MSNMIX)/GMIX]×100%
Note: The misclassification rates based on the MSNMIX classifier are all significantly lower than those of the GMIX classifier as measured by the Wilcoxon rank-sum test

study, we found that both model-based predictors substantially outperform MI for all cases. Furthermore, the MSNMIX predictor exhibits considerable promising accuracy in the prediction of missing values when compared with those of MSN imputations over a wide range of missing rates.

As another illustration, we compare the supervised learning of classification accuracies between the GMIX and MSNMIX classifiers; see Eq. 12. Comparisons are made on the trivariate data and a bivariate sample on attributes BMI and LBM. Table 2 shows the average misclassification rates from these models. As seen in the table, the misclassification rates of the MSNMIX classifier are all smaller than those of the GMIX classifier, especially for the bivariate sample with RIPs ranging between 43.3 and 61.5%. Alternatively, the Wilcoxon rank sum procedure (Hollander and Wolfe 1999) can be performed to test whether misclassification rate of MSNMIX classifier, P_S, is significantly lower than that of the GMIX classifier, P_G. In other words, the null hypothesis is $H_0 : P_S \leq P_G$. If this p-value is less than 0.05, we can reject the above null hypothesis at the 5% significance level. Once again, the MSNMIX classifier gave highly significant reduction in misclassification rate (all p-values are far less than 0.05) as measured by the Wilcoxon rank sum test. These observations signify the MSNMIX model provides a sound basis for the classification of partially observed features.

5 Concluding remarks

We establish some properties related to the MSNMIX model in a missing information framework. The proposed model is very flexible in dealing with heterogeneous data that involve strong skewness and is robust to the presence of missing observations. We discuss in detail how the EM algorithm coupled with auxiliary matrices can be applied on learning models from incomplete data in an efficient manner. Experimental results indicate that the MSNMIX model performs well for imputations as well as classification when asymmetric multimodality and missing outcomes simultaneously occur in the input data.

There are a number of possible extensions of the current work. We highlight that, with the growing advances of modern stochastic computing technology and inexpensive high-speed computer power, it is worthwhile to pursue a fully Bayesian treatment (e.g., Hastings 1970; Tanner and Wong 1987; Diebolt and Robert 1994; Escobar and West 1995) in this context for enriching up-to-date account of the theory and applicability. Furthermore, it is also of interest to generalize all existing approaches to learning multivariate skew t mixture models (Lin 2009b; Pyne et al. 2009) from incomplete data.

Acknowledgments The authors would like to express his deepest gratitude to the Chief Editor, the Associate Editor and two anonymous referees for their valuable comments and suggestions that greatly improved this paper. This research was supported by the National Science Council of Taiwan (Grant No. NSC97-2118-M-005-001-MY2).

Appendix A: Proof of Theorem 1

(a) Let $\boldsymbol{Y}_j$, $\boldsymbol{\xi}_i$, $\boldsymbol{\Sigma}_i$ and $\boldsymbol{\Lambda}_i$ be partitioned as

$$\boldsymbol{Y}_j = \begin{bmatrix} \boldsymbol{Y}_j^{\mathrm{o}} \\ \boldsymbol{Y}_j^{\mathrm{m}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{O}_j \boldsymbol{Y}_j \\ \boldsymbol{M}_j \boldsymbol{Y}_j \end{bmatrix}, \quad \boldsymbol{\xi}_i = \begin{bmatrix} \boldsymbol{\xi}_{ij}^{\mathrm{o}} \\ \boldsymbol{\xi}_{ij}^{\mathrm{m}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{O}_j \boldsymbol{\xi}_i \\ \boldsymbol{M}_j \boldsymbol{\xi}_i \end{bmatrix},$$
$$\boldsymbol{\Sigma}_i = \begin{bmatrix} \boldsymbol{\Sigma}_{ij}^{\mathrm{oo}} & \boldsymbol{\Sigma}_{ij}^{\mathrm{om}} \\ \boldsymbol{\Sigma}_{ij}^{\mathrm{mo}} & \boldsymbol{\Sigma}_{ij}^{\mathrm{mm}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{O}_j \boldsymbol{\Sigma}_i \boldsymbol{O}_j^\top & \boldsymbol{O}_j \boldsymbol{\Sigma}_i \boldsymbol{M}_j^\top \\ \boldsymbol{M}_j \boldsymbol{\Sigma}_i \boldsymbol{O}_j^\top & \boldsymbol{M}_j \boldsymbol{\Sigma}_i \boldsymbol{M}_j^\top \end{bmatrix}, \quad \text{and}$$
$$\boldsymbol{\Lambda}_i = \begin{bmatrix} \boldsymbol{\Lambda}_{ij}^{\mathrm{oo}} & \boldsymbol{0}_{p_j^{\mathrm{o}} \times p_j^{\mathrm{m}}} \\ \boldsymbol{0}_{p_j^{\mathrm{m}} \times p_j^{\mathrm{o}}} & \boldsymbol{\Lambda}_{ij}^{\mathrm{mm}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{O}_j \boldsymbol{\Lambda}_i \boldsymbol{O}_j^\top & \boldsymbol{0}_{p_j^{\mathrm{o}} \times p_j^{\mathrm{m}}} \\ \boldsymbol{0}_{p_j^{\mathrm{m}} \times p_j^{\mathrm{o}}} & \boldsymbol{M}_j \boldsymbol{\Lambda}_i \boldsymbol{M}_j^\top \end{bmatrix}.$$

Thus, we obtain

$$\boldsymbol{\Omega}_i = \boldsymbol{\Sigma}_i + \boldsymbol{\Lambda}_i^2 = \begin{bmatrix} \boldsymbol{O}_j \left(\boldsymbol{\Sigma}_i + \boldsymbol{\Lambda}_i^2\right) \boldsymbol{O}_j^\top & \boldsymbol{O}_j \boldsymbol{\Sigma}_i \boldsymbol{M}_j^\top \\ \boldsymbol{M}_j \boldsymbol{\Sigma}_i \boldsymbol{O}_j^\top & \boldsymbol{M}_j \left(\boldsymbol{\Sigma}_i + \boldsymbol{\Lambda}_i^2\right) \boldsymbol{M}_j^\top \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Omega}_{ij}^{\mathrm{oo}} & \boldsymbol{\Omega}_{ij}^{\mathrm{om}} \\ \boldsymbol{\Omega}_{ij}^{\mathrm{mo}} & \boldsymbol{\Omega}_{ij}^{\mathrm{mm}} \end{bmatrix}.$$

Note that $\boldsymbol{\Sigma}_i$ and $\boldsymbol{\Lambda}_i$ are symmetric matrices. It follows that $\boldsymbol{\Omega}_i$, $\boldsymbol{\Omega}_{ij}^{\mathrm{oo}}$ and $\boldsymbol{\Omega}_{ij}^{\mathrm{mm}}$ are symmetric matrices and $\boldsymbol{\Omega}_{ij}^{\mathrm{om}^\top} = \boldsymbol{\Omega}_{ij}^{\mathrm{mo}}$. Furthermore,

$$\begin{aligned} \boldsymbol{\Lambda}_i \boldsymbol{\Omega}_i^{-1} &= \begin{bmatrix} \boldsymbol{\Lambda}_{ij}^{\mathrm{oo}} \left(\boldsymbol{\Omega}_{ij}^{\mathrm{oo}^{-1}} \boldsymbol{\Omega}_{ij}^{\mathrm{om}} \boldsymbol{\Omega}_{ij}^{\mathrm{mm}\cdot\mathrm{o}^{-1}} \boldsymbol{\Omega}_{ij}^{\mathrm{mo}} \boldsymbol{\Omega}_{ij}^{\mathrm{oo}^{-1}} + \boldsymbol{\Omega}_{ij}^{\mathrm{oo}^{-1}} \right) & -\boldsymbol{\Lambda}_{ij}^{\mathrm{oo}} \boldsymbol{\Omega}_{ij}^{\mathrm{oo}^{-1}} \boldsymbol{\Omega}_{ij}^{\mathrm{om}} \boldsymbol{\Omega}_{ij}^{\mathrm{mm}\cdot\mathrm{o}^{-1}} \\ -\boldsymbol{\Lambda}_{ij}^{\mathrm{mm}} \boldsymbol{\Omega}_{ij}^{\mathrm{mm}\cdot\mathrm{o}^{-1}} \boldsymbol{\Omega}_{ij}^{\mathrm{mo}} \boldsymbol{\Omega}_{ij}^{\mathrm{oo}^{-1}} & \boldsymbol{\Lambda}_{ij}^{\mathrm{mm}} \boldsymbol{\Omega}_{ij}^{\mathrm{mm}\cdot\mathrm{o}^{-1}} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{B}_{i1} & \boldsymbol{B}_{i2} \end{bmatrix}, \end{aligned}$$

where $\boldsymbol{\Omega}_{ij}^{\mathrm{mm}\cdot\mathrm{o}} = \boldsymbol{\Omega}_{ij}^{\mathrm{mm}} - \boldsymbol{\Omega}_{ij}^{\mathrm{mo}} \boldsymbol{\Omega}_{ij}^{\mathrm{oo}^{-1}} \boldsymbol{\Omega}_{ij}^{\mathrm{om}}$. That is,

$$\boldsymbol{B}_{i1} = \begin{bmatrix} \boldsymbol{\Lambda}_{ij}^{\mathrm{oo}} \left(\boldsymbol{\Omega}_{ij}^{\mathrm{oo}^{-1}} \boldsymbol{\Omega}_{ij}^{\mathrm{om}} \boldsymbol{\Omega}_{ij}^{\mathrm{mm}\cdot\mathrm{o}^{-1}} \boldsymbol{\Omega}_{ij}^{\mathrm{mo}} + \boldsymbol{I}_{p_j^{\mathrm{o}}} \right) \\ -\boldsymbol{\Lambda}_{ij}^{\mathrm{mm}} \boldsymbol{\Omega}_{ij}^{\mathrm{mm}\cdot\mathrm{o}^{-1}} \boldsymbol{\Omega}_{ij}^{\mathrm{mo}} \end{bmatrix} \boldsymbol{\Omega}_{ij}^{\mathrm{oo}^{-1}}$$

and

$$\boldsymbol{B}_{i2} = \begin{bmatrix} -\boldsymbol{\Lambda}_{ij}^{\mathrm{oo}} \boldsymbol{\Omega}_{ij}^{\mathrm{oo}^{-1}} \boldsymbol{\Omega}_{ij}^{\mathrm{om}} \\ \boldsymbol{\Lambda}_{ij}^{\mathrm{mm}} \end{bmatrix} \boldsymbol{\Omega}_{ij}^{\mathrm{mm\cdot o}^{-1}}.$$

It suffices to show that

$$\boldsymbol{B}_{i1} + \boldsymbol{B}_{i2} \boldsymbol{\Omega}_{ij}^{\mathrm{mo}} \boldsymbol{\Omega}_{ij}^{\mathrm{oo}^{-1}} = \begin{bmatrix} \boldsymbol{\Lambda}_{ij}^{\mathrm{oo}} \\ \mathbf{0}_{p_j^{\mathrm{m}} \times p_j^{\mathrm{o}}} \end{bmatrix} \boldsymbol{\Omega}_{ij}^{\mathrm{oo}^{-1}}.$$

Moreover, we have the following results:

$$\begin{aligned} \boldsymbol{\Lambda}_i \boldsymbol{\Omega}_i^{-1} \boldsymbol{\Lambda}_i &= \begin{bmatrix} \boldsymbol{B}_{i1} & \boldsymbol{B}_{i2} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Lambda}_{ij}^{\mathrm{oo}} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda}_{ij}^{\mathrm{mm}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{B}_{i1} \boldsymbol{\Lambda}_{ij}^{\mathrm{oo}} & \boldsymbol{B}_{i2} \boldsymbol{\Lambda}_{ij}^{\mathrm{mm}} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{\Lambda}_{ij}^{\mathrm{oo}} \left(\boldsymbol{\Omega}_{ij}^{\mathrm{oo}^{-1}} \boldsymbol{\Omega}_{ij}^{\mathrm{om}} \boldsymbol{\Omega}_{ij}^{\mathrm{mm\cdot o}^{-1}} \boldsymbol{\Omega}_{ij}^{\mathrm{mo}} + \boldsymbol{I}_{p_j^{\mathrm{o}}} \right) \boldsymbol{\Omega}_{ij}^{\mathrm{oo}^{-1}} \boldsymbol{\Lambda}_{ij}^{\mathrm{oo}} & -\boldsymbol{\Lambda}_{ij}^{\mathrm{oo}} \boldsymbol{\Omega}_{ij}^{\mathrm{oo}^{-1}} \boldsymbol{\Omega}_{ij}^{\mathrm{om}} \boldsymbol{\Omega}_{ij}^{\mathrm{mm\cdot o}^{-1}} \boldsymbol{\Lambda}_{ij}^{\mathrm{mm}} \\ -\boldsymbol{\Lambda}_{ij}^{\mathrm{mm}} \boldsymbol{\Omega}_{ij}^{\mathrm{mm\cdot o}^{-1}} \boldsymbol{\Omega}_{ij}^{\mathrm{mo}} \boldsymbol{\Omega}_{ij}^{\mathrm{oo}^{-1}} \boldsymbol{\Lambda}_{ij}^{\mathrm{oo}} & \boldsymbol{\Lambda}_{ij}^{\mathrm{mm}} \boldsymbol{\Omega}_{ij}^{\mathrm{mm\cdot o}^{-1}} \boldsymbol{\Lambda}_{ij}^{\mathrm{mm}} \end{bmatrix} \end{aligned} \tag{A.1}$$

and

$$\begin{aligned} &\boldsymbol{B}_{i2} \boldsymbol{\Omega}_{ij}^{\mathrm{mm\cdot o}} \boldsymbol{B}_{i2}^{\top} \\ &= \begin{bmatrix} \boldsymbol{\Lambda}_{ij}^{\mathrm{oo}} \boldsymbol{\Omega}_{ij}^{\mathrm{oo}^{-1}} \boldsymbol{\Omega}_{ij}^{\mathrm{om}} \boldsymbol{\Omega}_{ij}^{\mathrm{mm\cdot o}^{-1}} \boldsymbol{\Omega}_{ij}^{\mathrm{mo}} \boldsymbol{\Omega}_{ij}^{\mathrm{oo}^{-1}} \boldsymbol{\Lambda}_{ij}^{\mathrm{oo}} & -\boldsymbol{\Lambda}_{ij}^{\mathrm{oo}} \boldsymbol{\Omega}_{ij}^{\mathrm{oo}^{-1}} \boldsymbol{\Omega}_{ij}^{\mathrm{om}} \boldsymbol{\Omega}_{ij}^{\mathrm{mm\cdot o}^{-1}} \boldsymbol{\Lambda}_{ij}^{\mathrm{mm}} \\ -\boldsymbol{\Lambda}_{ij}^{\mathrm{mm}} \boldsymbol{\Omega}_{ij}^{\mathrm{mm\cdot o}^{-1}} \boldsymbol{\Omega}_{ij}^{\mathrm{mo}} \boldsymbol{\Omega}_{ij}^{\mathrm{oo}^{-1}} \boldsymbol{\Lambda}_{ij}^{\mathrm{oo}} & \boldsymbol{\Lambda}_{ij}^{\mathrm{mm}} \boldsymbol{\Omega}_{ij}^{\mathrm{mm\cdot o}^{-1}} \boldsymbol{\Lambda}_{ij}^{\mathrm{mm}} \end{bmatrix}. \end{aligned} \tag{A.2}$$

Since

$$\boldsymbol{\Delta}_i + \boldsymbol{B}_{i2} \boldsymbol{\Omega}_{ij}^{\mathrm{mm\cdot o}} \boldsymbol{B}_{i2}^{\top} = \boldsymbol{I}_p - \boldsymbol{\Lambda}_i \boldsymbol{\Omega}_i^{-1} \boldsymbol{\Lambda}_i + \boldsymbol{B}_{i2} \boldsymbol{\Omega}_{ij}^{\mathrm{mm\cdot o}} \boldsymbol{B}_{i2}^{\top}, \tag{A.3}$$

substituting (A.1) and (A.2) into (A.3) leads to

$$\boldsymbol{\Delta}_i + \boldsymbol{B}_{i2} \boldsymbol{\Omega}_{ij}^{\mathrm{mm\cdot o}} \boldsymbol{B}_{i2}^{\top} = \begin{bmatrix} \boldsymbol{I}_{p_j^{\mathrm{o}}} - \boldsymbol{\Lambda}_{ij}^{\mathrm{oo}} \boldsymbol{\Omega}_{ij}^{\mathrm{oo}^{-1}} \boldsymbol{\Lambda}_{ij}^{\mathrm{oo}} & \mathbf{0}_{p_j^{\mathrm{o}} \times p_j^{\mathrm{m}}} \\ \mathbf{0}_{p_j^{\mathrm{m}} \times p_j^{\mathrm{o}}} & \boldsymbol{I}_{p_j^{\mathrm{m}}} \end{bmatrix}.$$

Thus, we have

$$\begin{aligned} f(\boldsymbol{Y}_j^{\mathrm{o}}, \boldsymbol{Y}_j^{\mathrm{m}} | Z_{ij} = 1, \boldsymbol{\Theta}) = 2^p \phi_{p_j^{\mathrm{o}}} \left(\boldsymbol{Y}_j^{\mathrm{o}} | \boldsymbol{\xi}_{ij}^{\mathrm{o}}, \boldsymbol{\Omega}_{ij}^{\mathrm{oo}} \right) \phi_{p-p_j^{\mathrm{o}}} \left(\boldsymbol{Y}_j^{\mathrm{m}} | \boldsymbol{\xi}_{ij}^{\mathrm{m\cdot o}}, \boldsymbol{\Omega}_{ij}^{\mathrm{mm\cdot o}} \right) \\ \times \Phi_p \left(\boldsymbol{B}_{i1} \left(\boldsymbol{Y}_j^{\mathrm{o}} - \boldsymbol{\xi}_{ij}^{\mathrm{o}} \right) + \boldsymbol{B}_{i2} \left(\boldsymbol{Y}_j^{\mathrm{m}} - \boldsymbol{\xi}_{ij}^{\mathrm{m}} \right) \mid \boldsymbol{\Delta}_i \right), \end{aligned}$$

where $\boldsymbol{\xi}_{ij}^{\mathrm{m\cdot o}} = \boldsymbol{\xi}_{ij}^{\mathrm{m}} + \boldsymbol{\Omega}_{ij}^{\mathrm{mo}} \boldsymbol{\Omega}_{ij}^{\mathrm{oo}^{-1}} \left(\boldsymbol{Y}_j^{\mathrm{o}} - \boldsymbol{\xi}_{ij}^{\mathrm{o}} \right)$ and $\boldsymbol{\Omega}_{ij}^{\mathrm{mm\cdot o}} = \boldsymbol{\Omega}_{ij}^{\mathrm{mm}} - \boldsymbol{\Omega}_{ij}^{\mathrm{mo}} \boldsymbol{\Omega}_{ij}^{\mathrm{oo}^{-1}} \boldsymbol{\Omega}_{ij}^{\mathrm{om}}$.

The marginal density of $\boldsymbol{Y}_j^{\mathrm{o}}$ is given by

$$f(\boldsymbol{Y}_j^{\mathrm{o}}|Z_{ij}=1,\boldsymbol{\Theta}) = 2^p \phi_{p_j^{\mathrm{o}}}\left(\boldsymbol{Y}_j^{\mathrm{o}}|\boldsymbol{\xi}_{ij}^{\mathrm{o}},\boldsymbol{\Omega}_{ij}^{\mathrm{oo}}\right)\int \phi_{p_j^{\mathrm{m}}}\left(\boldsymbol{Y}_j^{\mathrm{m}}|\boldsymbol{\xi}_{ij}^{\mathrm{m\cdot o}},\boldsymbol{\Omega}_{ij}^{\mathrm{mm\cdot o}}\right) \times \Phi_p\left(\boldsymbol{B}_{i1}\left(\boldsymbol{Y}_j^{\mathrm{o}}-\boldsymbol{\xi}_{ij}^{\mathrm{o}}\right)+\boldsymbol{B}_{i2}\left(\boldsymbol{Y}_j^{\mathrm{m}}-\boldsymbol{\xi}_{ij}^{\mathrm{m}}\right)\mid \boldsymbol{\Delta}_i\right) d\boldsymbol{Y}_j^{\mathrm{m}}.$$

Let $\boldsymbol{z}=\boldsymbol{Y}_j^{\mathrm{m}}-\boldsymbol{\xi}_{ij}^{\mathrm{m\cdot o}}$, then $\boldsymbol{Y}_j^{\mathrm{m}}=\boldsymbol{z}+\boldsymbol{\xi}_{ij}^{\mathrm{m\cdot o}}=\boldsymbol{z}+\boldsymbol{\xi}_{ij}^{\mathrm{m}}+\boldsymbol{\Omega}_{ij}^{\mathrm{mo}}\boldsymbol{\Omega}_{ij}^{\mathrm{oo}^{-1}}\left(\boldsymbol{Y}_j^{\mathrm{o}}-\boldsymbol{\xi}_{ij}^{\mathrm{o}}\right)$. It can be shown that $\phi_{p_j^{\mathrm{m}}}\left(\boldsymbol{Y}_j^{\mathrm{m}}|\boldsymbol{\xi}_{ij}^{\mathrm{m\cdot o}},\boldsymbol{\Omega}_{ij}^{\mathrm{mm\cdot o}}\right)=\phi_{p_j^{\mathrm{m}}}\left(\boldsymbol{z}|\boldsymbol{0},\boldsymbol{\Omega}_{ij}^{\mathrm{mm\cdot o}}\right)$ and $\Phi_p\left(\boldsymbol{B}_{i1}\left(\boldsymbol{Y}_j^{\mathrm{o}}-\boldsymbol{\xi}_{ij}^{\mathrm{o}}\right)+\boldsymbol{B}_{i2}\left(\boldsymbol{Y}_j^{\mathrm{m}}-\boldsymbol{\xi}_{ij}^{\mathrm{m}}\right)\mid\boldsymbol{\Delta}_i\right)=\Phi_p\left(\left(\boldsymbol{B}_{i1}+\boldsymbol{B}_{i2}\boldsymbol{\Omega}_{ij}^{\mathrm{mo}}\boldsymbol{\Omega}_{ij}^{\mathrm{oo}^{-1}}\right)\left(\boldsymbol{Y}_j^{\mathrm{o}}-\boldsymbol{\xi}_{ij}^{\mathrm{o}}\right)+\boldsymbol{B}_{i2}\boldsymbol{z}\mid\boldsymbol{\Delta}_i\right)$.

By Lemma 2.1 of Arellano-Vallea and Genton (2005), we have

$$\begin{aligned} &f(\boldsymbol{Y}_j^{\mathrm{o}}\mid Z_{ij}=1,\boldsymbol{\Theta})\\ &=2^p\phi_{p_j^{\mathrm{o}}}\left(\boldsymbol{Y}_j^{\mathrm{o}}|\boldsymbol{\xi}_{ij}^{\mathrm{o}},\boldsymbol{\Omega}_{ij}^{\mathrm{oo}}\right)\\ &\quad\times\int\phi_{p_j^{\mathrm{m}}}\left(\boldsymbol{z}|\boldsymbol{0},\boldsymbol{\Omega}_{ij}^{\mathrm{mm\cdot o}}\right)\Phi_p\left(\left(\boldsymbol{B}_{i1}+\boldsymbol{B}_{i2}\boldsymbol{\Omega}_{ij}^{\mathrm{mo}}\boldsymbol{\Omega}_{ij}^{\mathrm{oo}^{-1}}\right)\left(\boldsymbol{Y}_j^{\mathrm{o}}-\boldsymbol{\xi}_{ij}^{\mathrm{o}}\right)+\boldsymbol{B}_{i2}\boldsymbol{z}\mid\boldsymbol{\Delta}_i\right)\mathrm{d}\boldsymbol{z}\\ &=2^{p_j^{\mathrm{o}}}\phi_{p_j^{\mathrm{o}}}\left(\boldsymbol{Y}_j^{\mathrm{o}}|\boldsymbol{\xi}_{ij}^{\mathrm{o}},\boldsymbol{\Omega}_{ij}^{\mathrm{oo}}\right)\ \Phi_{p_j^{\mathrm{o}}}\left(\boldsymbol{\Lambda}_{ij}^{\mathrm{oo}}\boldsymbol{\Omega}_{ij}^{\mathrm{oo}^{-1}}\left(\boldsymbol{Y}_j^{\mathrm{o}}-\boldsymbol{\xi}_{ij}^{\mathrm{o}}\right)\mid\boldsymbol{I}_{p_j^{\mathrm{o}}}-\boldsymbol{\Lambda}_{ij}^{\mathrm{oo}}\boldsymbol{\Omega}_{ij}^{\mathrm{oo}^{-1}}\boldsymbol{\Lambda}_{ij}^{\mathrm{oo}}\right). \end{aligned}$$

Thus, $\boldsymbol{Y}_j^{\mathrm{o}}\mid(Z_{ij}=1,\boldsymbol{\Theta})\sim SN_{p_j^{\mathrm{o}}}\left(\boldsymbol{\xi}_{ij}^{\mathrm{o}},\boldsymbol{\Sigma}_{ij}^{\mathrm{oo}},\boldsymbol{\Lambda}_{ij}^{\mathrm{oo}}\right)$. It implies that

$$f(\boldsymbol{Y}_j^{\mathrm{o}}\mid\boldsymbol{\Theta})=\sum_{i=1}^{g}f(\boldsymbol{Y}_j^{\mathrm{o}}\mid Z_{ij}=1,\boldsymbol{\Theta})p(Z_{ij}=1)=\sum_{i=1}^{g}w_i\psi_{p_j^{\mathrm{o}}}\left(\boldsymbol{Y}_j^{\mathrm{o}}\mid\boldsymbol{\xi}_{ij}^{\mathrm{o}},\boldsymbol{\Sigma}_{ij}^{\mathrm{oo}},\boldsymbol{\Lambda}_{ij}^{\mathrm{oo}}\right).$$

(b) By virtue of $\phi_p(\boldsymbol{Y}_j|\boldsymbol{\xi}_i,\boldsymbol{\Omega}_i)=\phi_{p_j^{\mathrm{o}}}(\boldsymbol{Y}_j^{\mathrm{o}}|\boldsymbol{\xi}_{ij}^{\mathrm{o}},\boldsymbol{\Omega}_{ij}^{\mathrm{oo}})\phi_{p-p_j^{\mathrm{o}}}(\boldsymbol{Y}_j^{\mathrm{m}}|\boldsymbol{\xi}_{ij}^{\mathrm{m\cdot o}},\boldsymbol{\Omega}_{ij}^{\mathrm{mm\cdot o}})$, see Theorem 2.5.1 of Anderson (2003), we can deduce that

$$f(\boldsymbol{Y}_j^{\mathrm{m}}\mid\boldsymbol{Y}_j^{\mathrm{o}})=2^p\sum_{i=1}^{g}\tilde{w}_{ij}\phi_{p-p_j^{\mathrm{o}}}(\boldsymbol{Y}_j^{\mathrm{m}}\mid\boldsymbol{\xi}_{ij}^{\mathrm{m\cdot o}},\boldsymbol{\Omega}_{ij}^{\mathrm{mm\cdot o}})\Phi_p(\boldsymbol{\Lambda}_i\boldsymbol{\Omega}_i^{-1}(\boldsymbol{Y}_j-\boldsymbol{\xi}_i)\mid\boldsymbol{\Delta}_i).$$

Appendix B

The proofs of part (a) and part (b) are straightforward and hence are omitted. We only show the proof of part (c). From (4), we have the following densities

$$f(\boldsymbol{Y}_j^{\mathrm{o}}|\boldsymbol{\gamma}_j,Z_{ij}=1,\boldsymbol{\Theta})\propto|\boldsymbol{\Sigma}_{ij}^{\mathrm{oo}}|^{-1/2}\exp\left\{-\frac{1}{2}\left(\boldsymbol{Y}_j^{\mathrm{o}}-\boldsymbol{\mu}_{ij}^{\mathrm{o}}\right)^{\top}\boldsymbol{\Sigma}_{ij}^{\mathrm{oo}^{-1}}(\boldsymbol{Y}_j^{\mathrm{o}}-\boldsymbol{\mu}_{ij}^{\mathrm{o}})\right\}$$

and

$$f(\boldsymbol{\gamma}_j|Z_{ij}=1) \propto \exp\left\{-\frac{1}{2}\boldsymbol{\gamma}_j^\top\boldsymbol{\gamma}_j\right\} I_{\mathbb{R}_+^p}(\boldsymbol{\gamma}_j).$$

Multiplying $f\left(\boldsymbol{Y}_j^{\mathrm{o}}|\boldsymbol{\gamma}_j, Z_{ij}=1, \boldsymbol{\Theta}\right)$ by $f\left(\boldsymbol{\gamma}_j|Z_{ij}=1\right)$ gives

$$\begin{aligned}
&f(\boldsymbol{Y}_j^{\mathrm{o}}, \boldsymbol{\gamma}_j|Z_{ij}=1, \boldsymbol{\Theta})\\
&\quad\propto |\boldsymbol{\Sigma}_{ij}^{\mathrm{oo}}|^{-1/2}\exp\left\{-\frac{1}{2}\left(\left(\boldsymbol{Y}_j^{\mathrm{o}}-\boldsymbol{\mu}_{ij}^{\mathrm{o}}\right)^\top\boldsymbol{\Sigma}_{ij}^{\mathrm{oo}^{-1}}\left(\boldsymbol{Y}_j^{\mathrm{o}}-\boldsymbol{\mu}_{ij}^{\mathrm{o}}\right)+\boldsymbol{\gamma}_j^\top\boldsymbol{\gamma}_j\right)\right\}\mathbb{I}_{\mathbb{R}_+^p}(\boldsymbol{\gamma}_j)\\
&\quad\propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\gamma}_j-\boldsymbol{\Lambda}_i^{-1}(\boldsymbol{Y}_j-\boldsymbol{\xi}_i)\right)^\top\boldsymbol{\Lambda}_i\boldsymbol{S}_{ij}^{\mathrm{oo}}\boldsymbol{\Lambda}_i\left(\boldsymbol{\gamma}_j-\boldsymbol{\Lambda}_i^{-1}(\boldsymbol{Y}_j-\boldsymbol{\xi}_i)\right)\right\}\\
&\quad\times|\boldsymbol{\Sigma}_{ij}^{\mathrm{oo}}|^{-1/2}\exp\left\{-\frac{1}{2}\boldsymbol{\gamma}_j^\top\boldsymbol{\gamma}_j\right\}\mathbb{I}_{\mathbb{R}_+^p}(\boldsymbol{\gamma}_j)\\
&\quad\propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\gamma}_j-\boldsymbol{\Lambda}_i\boldsymbol{C}_{ij}^{\mathrm{oo}}(\boldsymbol{Y}_j-\boldsymbol{\xi}_i)\right)^\top(\boldsymbol{I}_p-\boldsymbol{\Lambda}_i\boldsymbol{C}_{ij}^{\mathrm{oo}}\boldsymbol{\Lambda}_i)^{-1}\right.\\
&\qquad\left.\left(\boldsymbol{\gamma}_j-\boldsymbol{\Lambda}_i\boldsymbol{C}_{ij}^{\mathrm{oo}}(\boldsymbol{Y}_j-\boldsymbol{\xi}_i)\right)\right\}\\
&\quad\times|\boldsymbol{\Sigma}_{ij}^{\mathrm{oo}}|^{-1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{Y}_j-\boldsymbol{\xi}_i)^\top\boldsymbol{C}_{ij}^{\mathrm{oo}}(\boldsymbol{Y}_j-\boldsymbol{\xi}_i)\right\}\mathbb{I}_{\mathbb{R}_+^p}(\boldsymbol{\gamma}_j),
\end{aligned}$$

where the last two equalities follow from

$$\begin{aligned}
&(\boldsymbol{Y}_j^{\mathrm{o}}-\boldsymbol{\mu}_{ij}^{\mathrm{o}})^\top\boldsymbol{\Sigma}_{ij}^{\mathrm{oo}^{-1}}\left(\boldsymbol{Y}_j^{\mathrm{o}}-\boldsymbol{\mu}_{ij}^{\mathrm{o}}\right)+\boldsymbol{\gamma}_j^\top\boldsymbol{\gamma}_j\\
&\quad=(\boldsymbol{O}_j\boldsymbol{Y}_j-\boldsymbol{O}_j(\boldsymbol{\xi}_i+\boldsymbol{\Lambda}_i\boldsymbol{\gamma}_j))^\top\boldsymbol{\Sigma}_{ij}^{\mathrm{oo}^{-1}}(\boldsymbol{O}_j\boldsymbol{Y}_j-\boldsymbol{O}_j(\boldsymbol{\xi}_i+\boldsymbol{\Lambda}_i\boldsymbol{\gamma}_j))+\boldsymbol{\gamma}_j^\top\boldsymbol{\gamma}_j\\
&\quad=(\boldsymbol{Y}_j-\boldsymbol{\xi}_i-\boldsymbol{\Lambda}_i\boldsymbol{\gamma}_j)^\top\boldsymbol{O}_j^\top\boldsymbol{\Sigma}_{ij}^{\mathrm{oo}^{-1}}\boldsymbol{O}_j(\boldsymbol{Y}_j-\boldsymbol{\xi}_i-\boldsymbol{\Lambda}_i\boldsymbol{\gamma}_j)+\boldsymbol{\gamma}_j^\top\boldsymbol{\gamma}_j\\
&\quad=(\boldsymbol{Y}_j-\boldsymbol{\xi}_i-\boldsymbol{\Lambda}_i\boldsymbol{\gamma}_j)^\top\boldsymbol{S}_{ij}^{\mathrm{oo}}(\boldsymbol{Y}_j-\boldsymbol{\xi}_i-\boldsymbol{\Lambda}_i\boldsymbol{\gamma}_j)+\boldsymbol{\gamma}_j^\top\boldsymbol{\gamma}_j\\
&\quad=\left(\boldsymbol{\gamma}_j-\boldsymbol{\Lambda}_i^{-1}(\boldsymbol{Y}_j-\boldsymbol{\xi}_i))^\top\boldsymbol{\Lambda}_i\boldsymbol{S}_{ij}^{\mathrm{oo}}\boldsymbol{\Lambda}_i(\boldsymbol{\gamma}_j-\boldsymbol{\Lambda}_i^{-1}(\boldsymbol{Y}_j-\boldsymbol{\xi}_i)\right)+\boldsymbol{\gamma}_j^\top\boldsymbol{\gamma}_j\\
&\quad=\left(\boldsymbol{\gamma}_j-\boldsymbol{\Lambda}_i\boldsymbol{C}_{ij}^{\mathrm{oo}}(\boldsymbol{Y}_j-\boldsymbol{\xi}_i)\right)^\top(\boldsymbol{I}_p-\boldsymbol{\Lambda}_i\boldsymbol{C}_{ij}^{\mathrm{oo}}\boldsymbol{\Lambda}_i)^{-1}\left(\boldsymbol{\gamma}_j-\boldsymbol{\Lambda}_i^{-1}(\boldsymbol{Y}_j-\boldsymbol{\xi}_i)\right)\\
&\qquad+(\boldsymbol{Y}_j-\boldsymbol{\xi}_i)^\top\boldsymbol{C}_{ij}^{\mathrm{oo}}(\boldsymbol{Y}_j-\boldsymbol{\xi}_i).
\end{aligned}$$

Moreover, it is not difficult to show the identity $|\boldsymbol{\Sigma}_{ij}^{\mathrm{oo}}|=|\boldsymbol{\Omega}_{ij}^{\mathrm{oo}}||\boldsymbol{I}_p-\boldsymbol{\Lambda}_i\boldsymbol{C}_{ij}^{\mathrm{oo}}\boldsymbol{\Lambda}_i|$. By Bayes' rule, the conditional density of $\boldsymbol{\gamma}_j$ given $\boldsymbol{Y}_j^{\mathrm{o}}$ and $Z_{ij}=1$ is

$$\begin{aligned}
&f(\boldsymbol{\gamma}_j|\boldsymbol{Y}_j^{\mathrm{o}}, Z_{ij}=1, \boldsymbol{\Theta})\propto|\boldsymbol{I}_p-\boldsymbol{\Lambda}_i\boldsymbol{C}_{ij}^{\mathrm{oo}}\boldsymbol{\Lambda}_i|^{-1/2}\\
&\qquad\times\exp\left\{-\frac{1}{2}\left(\boldsymbol{\gamma}_j-\boldsymbol{q}_{ij}^{\mathrm{o}})^\top(\boldsymbol{I}_p-\boldsymbol{\Lambda}_i\boldsymbol{C}_{ij}^{\mathrm{oo}}\boldsymbol{\Lambda}_i)^{-1}(\boldsymbol{\gamma}_j-\boldsymbol{q}_{ij}^{\mathrm{o}}\right)\right\}\mathbb{I}_{\mathbb{R}_+^p}(\boldsymbol{\gamma}_j),
\end{aligned}$$

where $\boldsymbol{q}_{ij}^{\mathrm{o}} = \boldsymbol{\Lambda}_i \boldsymbol{C}_{ij}^{\mathrm{oo}} (\boldsymbol{Y}_j - \boldsymbol{\xi}_i)$. This implies that

$$\boldsymbol{\gamma}_j | (\boldsymbol{Y}_j^{\mathrm{o}}, Z_{ij} = 1, \boldsymbol{\Theta}) \sim TN_p \left(\boldsymbol{\Lambda}_i \boldsymbol{C}_{ij}^{\mathrm{oo}} (\boldsymbol{Y}_j - \boldsymbol{\xi}_i), \boldsymbol{I}_p - \boldsymbol{\Lambda}_i \boldsymbol{C}_{ij}^{\mathrm{oo}} \boldsymbol{\Lambda}_i, \mathbb{R}_+^p \right).$$

References

Anderson TW (2003) An introduction to multivariate statistical analysis, 3rd edn. Wiley and Sons, New York

Arellano-Valle RB, Azzalini A (2006) On the unification of families of skew-normal distributions. Scand J Statist 33:561–574

Arellano-Valle RB, Bolfarine H, Lachos VH (2007) Bayesian inference for skew-normal linear mixed models. J Appl Stat 34:663–682

Arellano-Vallea RB, Genton MG (2005) On fundamental skew distributions. J Multivariate Anal 96:93–116

Azzalini A (2005) The skew-normal distribution and related multivariate families (with discussion). Scand J Statist 32:159–200

Azzalini A, Capitanio A (1999) Statistical applications of the multivariate skew-normal distribution. J R Stat Soc Ser B 61:579–602

Azzalini A, Dalla Valle A (1996) The multivariate skew-normal distribution. Biometrika 83:715–726

Basford KE, McLachlan GJ (1985) Estimation of allocation rates in a cluster analysis text. J Am Stat Assoc 80:286–293

Box GEP, Cox DR (1964) An analysis of transformation. J R Stat Soc Ser A 26:211–252

Cook RD, Weisberg S (1994) An introduction to regression graphics. Wiley, New York

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). J R Stat Soc Ser B 39:1–38

Diebolt J, Robert CP (1994) Estimation of finite mixture distributions through Bayesian sampling. J R Stat Soc Ser B 56:363–375

Escobar MD, West M (1995) Bayesian density estimation and inference using mixtures. J Am Stat Assoc 90:577–588

Frühwirth-Schnatter S (2006) Finite mixture and Markov switching models. Springer, New York

Ghahramani Z, Jordan MI (1994) Supervised learning from incomplete data via an EM approach. In: Cowan JD, Tesarro G, Alspector J (eds) Advances in neural information processing systerms, vol 6. Morgan Kaufmann Publishers, San Francisco pp 120–127

Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109

Hollander M, Wolfe DA (1999) Nonparametric statistical methods, 2nd edn. Wiley, New York

Lin TI (2009a) Maximum likelihood estimation for multivariate skew normal mixture models. J Multivariate Anal 100:257–265

Lin TI (2009b) Robust mixture modeling using multivariate skew t distributions. Stat Comput. doi:10.1007/s11222-009-9128-9 (in press)

Lin TI, Lee JC, Ni HF (2004) Bayesian analysis of mixture modelling using the multivariate t distribution. Stat Comput 14:119–130

Lin TI, Lee JC, Ho HJ (2006) On fast supervised learning for normal mixture models with missing information. Pattern Recogn 39:1177–1187

Lin TI, Lee JC, Hsieh WJ (2007a) Robust mixture modeling using the skew t distribution. Stat Comput 17:81–92

Lin TI, Lee JC, Yen SY (2007b) Finite mixture modelling using the skew normal distribution. Statist Sinica 17:909–927

Lin TI, Ho HJ, Shen PS (2009) Computationally efficient learning of multivariate t mixture models with missing information. Comp Stat 24:375–392

Little RJA, Rubin DB (2002) Statistical analysis with missing data, 2nd edn. Wiley, New York

Liu CH, Rubin DB, Wu Y (1998) Parameter expansion to accelerate EM: the PX-EM algorithm. Biometrika 85:755–770

McLachlan GJ, Basford KE (1988) Mixture models: inference and application to clustering. Marcel Dekker, New York

McLachlan GJ, Peel D (2000) Finite mixture models. Wiely, New York

Pearson K (1894) Contributions to the theory of mathematical evolution, Phi. Trans Roy Soc London A 185:71–110

Peel D, McLachlan GJ (2000) Robust mixture modeling using the *t* distribution. Stat Comput 10:339–348

Pyne S, Hu X, Wang K, Rossin E, Lin TI, Maier L, Baecher-Allan C, McLachlan GJ, Tamayo P, Hafler DA, De Jager PL, Mesirov JP (2009) Automated high-dimensional flow cytometric data analysis. Proc Natl Acad Sci USA 106:8519–8524

Rubin DB (1976) Inference and missing data. Biometrika 63:581–592

Sahu SK, Dey DK, Branco MD (2003) A new class of multivariate skew distributions with applications to bayesian regression models. Canad J Statist 31:129–150

Schafer JL (1997) Analysis of incomplete multivariate data. Chapman and Hall, London

Shoham S (2002) Robust clustering by deterministic agglomeration EM of mixtures of multivariate *t*-distributions. Pattern Recogn 35:1127–1142

Shoham S, Fellows MR, Normann RA (2003) Robust, automatic spike sorting using mixtures of multivariate *t*-distributions. J Neurosci Methods 127:111–122

Tanner MA, Wong WH (1987) The calculation of posterior distributions by data augmentation (with discussion). J Am Stat Assoc 82:528–550

Titterington DM, Smith AFM, Markov UE (1985) Statistical analysis of finite mixture distributions. Wiely, New York

Wang HX, Zhang QB, Luo B, Wei S (2004) Robust mixture modelling using multivariate *t* distribution with missing information. Pattern Recogn Lett 25:701–710

Zio MD, Guarnera U, Luzi O (2007) Imputation through finite Gaussian mixture models. Comp Stat Data Anal 51:5305–5316