Available online at www.sciencedirect.com

SCIENCE DIRECT®

Pattern Recognition 39 (2006) 1177–1187

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

On fast supervised learning for normal mixture models with missing information

Tsung I. Lin[a,*], Jack C. Lee[b], Hsiu J. Ho[c]

[a]*Department of Applied Mathematics, National Chung Hsing University, Taichung 402, Taiwan*
[b]*Institute of Statistics and Graduate Institute of Finance, National Chiao Tung University, Hsinchu 300, Taiwan*
[c]*Institute of Statistical Science, Academia Sinica, Taipei, Taiwan*

Received 1 July 2005; received in revised form 7 November 2005; accepted 21 December 2005

Abstract

It is an important research issue to deal with mixture models when missing values occur in the data. In this paper, computational strategies using auxiliary indicator matrices are introduced for efficiently handling mixtures of multivariate normal distributions when the data are missing at random and have an arbitrary missing data pattern, meaning that missing data can occur anywhere. We develop a novel EM algorithm that can dramatically save computation time and be exploited in many applications, such as density estimation, supervised clustering and prediction of missing values. In the aspect of multiple imputations for missing data, we also offer a data augmentation scheme using the Gibbs sampler. Our proposed methodologies are illustrated through some real data sets with varying proportions of missing values.

Keywords: Bayesian classifier; Data augmentation; EM algorithm; Incomplete features; Rao-Blackwellization

1. Introduction

Finite mixture models are known as powerful and flexible tools, which have been fully developed and applied in various theoretic and real problems as they are capable of modelling a wide range of densities, see for example Refs. [1–3]. However, missing values frequently appear in many real-world multivariate data sets that complicate data analyses and statistical inferences for practitioners. Missing data imputation techniques under the assumption of multivariate normal model have been well studied by Refs. [4,5]. Recently, learning mixture models from incomplete data becomes an important research issue in multivariate analysis. The work on the use of Gaussian component was pioneered by Ghahramani and Jordan [6], denoted by GJ hereafter. They present how to implement the expectation–maximization (EM) algorithm [7] to compute maximum likelihood (ML) estimates from multivariate data with arbitrary pattern of missingness. They also compare the performance of EM imputation with a common mean imputation (MI) heuristic for the supervised classification of incomplete features.

Due to rapid advance of computational developments, Bayesian sampling-based approaches are usually considered as an alternative way in dealing with mixture models. There are plenty of papers in the literature to address the problem of fitting normal mixture models under Bayesian treatments. For example, Diebolt and Robert [8] employ the data augmentation (DA) technique of Tanner and Wong [9] as an approximation method for evaluating the posterior distribution and show a duality principle. Escobar and West [10] present a nonparametric Bayesian density estimation for Dirichlet process mixture models. Richardson and Green [11] and Zhang et al. [12] propose a full Bayesian inference for a normal mixture model with unknown number of components using the reversible jump MCMC algorithm proposed by Green [13]. Stephens [14] and Fruhwirth-Schnatter [15] demonstrate Bayesian strategies for the elimination of label switching problems.

[*] Corresponding author. Tel.: +886 4 22850420; fax: +886 4 22873028.
E-mail address: tilin@amath.nchu.edu.tw (T.I. Lin).

0031-3203/$30.00 © 2006 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2005.12.014

In this paper, we offer an efficient EM algorithm for the fitting of a likelihood-based normal mixture model using partially observed data. To reduce computational burden during the EM iterations, we incorporate two types of auxiliary binary indicator matrices corresponding to the observed and unobserved components of each datum. With strategies similar to EM, we also offer a DA computational technique for efficiently imputting missing values and learning parameters using the Gibbs sampler [16], which constructs a Markov chain that converges to a tractable posterior distribution. The feature of the chosen prior distributions is weakly informative to avoid mathematical and computational pitfalls of using improper priors in mixture model, see Celeux et al. [17].

The rest of the paper proceeds as follows. In the next section, we describe the model and its notations, and present some important statistical properties based on the missing information framework. In Sections 3 and 4, two efficient EM and DA algorithms are developed to cope with ML and Bayesian estimation, respectively. We also investigate two issues regarding classification and prediction of incomplete features from both ML and Bayesian perspectives. In Section 5, some real data sets are utilized to illustrate our proposed methodologies with varying proportions of artificially missing values. Also, empirical comparisons between ML and Bayesian approaches in terms of classification and prediction accuracies for incomplete features are demonstrated. Finally, some concluding remarks are given in Section 6.

2. A normal mixture model with missing information

In the normal mixture model, we assume that $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$ form a p-dimensional random sample from a population with g subclasses $\mathscr{C}_1, \ldots, \mathscr{C}_g$, and each $\boldsymbol{Y}_j$ has the density

$$f(\boldsymbol{Y}_j \mid \boldsymbol{\Theta}) = \sum_{i=1}^{g} w_i \phi_p\left(\boldsymbol{Y}_j \,\middle|\, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right), \quad w_i \geqslant 0, \quad \sum_{i=1}^{g} w_i = 1, \tag{1}$$

where w_i's are mixing probabilities, $\phi_p(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a p-dimensional multivariate normal component density with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and $\boldsymbol{\Theta} = (w_1, \ldots, w_g, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_g)$ is the vector of mixture model parameters subject to $\sum_{i=1}^{g} w_i = 1$ and $\boldsymbol{\Sigma}_i$'s are positive definite matrices. Thus, there are $g(p+1)(p+2)/2 - 1$ distinct parameters in model (1).

Typically, in the EM framework, mixture models can be characterized as having an incomplete data structure. It is convenient to formalize the missing part as a set of membership labels $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$ with each label $\boldsymbol{Z}_j = (Z_{1j}, \ldots, Z_{gj})$ being a binary vector such that $Z_{ij} = 1$ if $\boldsymbol{Y}_j$ belongs to component i and $Z_{ij} = 0$ otherwise. Given the mixing probabilities $\boldsymbol{\omega}, \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ independently follow a multinomial distribution. We shall write $\boldsymbol{Z}_j \sim \mathscr{M}(1; w_1, \ldots, w_g)$.

For notational simplicity, let

$$\Delta_{ij} = (\boldsymbol{Y}_j - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{Y}_j - \boldsymbol{\mu}_i), \tag{2}$$

denote the Mahalanobis distance for $\boldsymbol{Y}_j$ with respect to mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$. The complete likelihood function for $\boldsymbol{\Theta}$ is

$$L_c(\boldsymbol{\Theta}|\boldsymbol{Y}, \boldsymbol{Z}) \propto \prod_{j=1}^{n} \prod_{i=1}^{g} \left(w_i |\boldsymbol{\Sigma}_i|^{-1/2} \exp\left(-\frac{1}{2} \Delta_{ij} \right) \right)^{Z_{ij}}. \tag{3}$$

We consider the maximum likelihood estimation problem of model (1) when $\boldsymbol{Y}$ are not completely observed. We further assume that the patterns of missingness are arbitrary and missing at random (MAR), see Refs. [18,19] for more details. Generally speaking, MAR refers to the missingness depending only on observed values but not on missing values.

Let $\boldsymbol{Y}_j$ be partitioned into two components $(\boldsymbol{Y}_j^{\mathrm{o}}, \boldsymbol{Y}_j^{\mathrm{m}})$, 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 EM algorithm, it is advantageous to introduce two types of binary indicator matrices, denoted by $\boldsymbol{O}_j$ and $\boldsymbol{M}_j$ hereafter, corresponding to $\boldsymbol{Y}_j$ such that $\boldsymbol{Y}_j^{\mathrm{o}} = \boldsymbol{O}_j \boldsymbol{Y}_j$ and $\boldsymbol{Y}_j^{\mathrm{m}} = \boldsymbol{M}_j \boldsymbol{Y}_j$, respectively. Notice that $\boldsymbol{O}_j$ and $\boldsymbol{M}_j$ are $p_j^{\mathrm{o}} \times p$ and $(p - p_j^{\mathrm{o}}) \times p$ matrices 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. We then have the following two propositions.

Proposition 1. *Suppose $\boldsymbol{Y}_j$ is partitioned into two components $(\boldsymbol{Y}_j^{\mathrm{o}}, \boldsymbol{Y}_j^{\mathrm{m}})$, where $\boldsymbol{Y}_j^{\mathrm{o}} = \boldsymbol{O}_j \boldsymbol{Y}_j$ and $\boldsymbol{Y}_j^{\mathrm{m}} = \boldsymbol{M}_j \boldsymbol{Y}_j$. We thus have*

$$\boldsymbol{Y}_j = \begin{cases} \boldsymbol{Y}_j^{\mathrm{o}}, & \text{if } p_j^{\mathrm{o}} = p; \\ \boldsymbol{O}_j^\top \boldsymbol{Y}_j^{\mathrm{o}} + \boldsymbol{M}_j^\top \boldsymbol{Y}_j^{\mathrm{m}}, & \text{if } 1 \leqslant p_j^{\mathrm{o}} < p, \end{cases}$$

and $\boldsymbol{O}_j^\top \boldsymbol{O}_j + \boldsymbol{M}_j^\top \boldsymbol{M}_j = \boldsymbol{I}_p$.

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

Proposition 2. *Let $\boldsymbol{Y}_j \sim \sum_{i=1}^{g} w_i \phi_p(\boldsymbol{Y}_j|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, and let $\boldsymbol{Y}_j^{\mathrm{o}}$ and $\boldsymbol{Y}_j^{\mathrm{m}}$ be the observed and missing components corresponding $\boldsymbol{Y}_j$, respectively. The marginal distribution of $\boldsymbol{Y}_j^{\mathrm{o}}$ is denoted by $\boldsymbol{Y}_j^{\mathrm{o}} \sim \sum_{i=1}^{g} w_i \phi_{p_j^{\mathrm{o}}}(\boldsymbol{Y}_j^{\mathrm{o}}|\boldsymbol{\mu}_{ij}^{\mathrm{o}}, \boldsymbol{\Sigma}_{ij}^{\mathrm{oo}})$, where*

$$\phi_{p_j^{\mathrm{o}}}(\boldsymbol{Y}_j^{\mathrm{o}}|\boldsymbol{\mu}_{ij}^{\mathrm{o}}, \boldsymbol{\Sigma}_{ij}^{\mathrm{oo}}) = (2\pi)^{-p_j^{\mathrm{o}}/2} |\boldsymbol{\Sigma}_{ij}^{\mathrm{oo}}|^{-1/2} \exp(-\tfrac{1}{2}\Delta_{ij}^{\mathrm{o}}),$$

and

$$\begin{aligned} &\boldsymbol{\mu}_{ij}^{\mathrm{o}} = \boldsymbol{O}_j \boldsymbol{\mu}_i, \quad \boldsymbol{\Sigma}_{ij}^{\mathrm{oo}} = \boldsymbol{O}_j \boldsymbol{\Sigma}_i \boldsymbol{O}_j^\top, \\ &\Delta_{ij}^{\mathrm{o}} = (\boldsymbol{Y}_j - \boldsymbol{\mu}_i)^\top \boldsymbol{S}_{ij}^{\mathrm{oo}} (\boldsymbol{Y}_j - \boldsymbol{\mu}_i), \\ &\boldsymbol{S}_{ij}^{\mathrm{oo}} = \boldsymbol{O}_j^\top (\boldsymbol{O}_j \boldsymbol{\Sigma}_i \boldsymbol{O}_j^\top)^{-1} \boldsymbol{O}_j. \end{aligned} \tag{4}$$

Consequently, $\boldsymbol{Y}_j^{\mathrm{m}}|\boldsymbol{Y}_j^{\mathrm{o}} \sim \sum_{i=1}^{g} w_{ij}^{*}\phi_{p-p_j^{\mathrm{o}}}(\boldsymbol{Y}_j^{\mathrm{m}}|\boldsymbol{\mu}_{ij}^{\mathrm{m\cdot o}}, \boldsymbol{\Sigma}_{ij}^{\mathrm{mm\cdot o}})$, *where*

$$\phi_{p-p_j^{\mathrm{o}}}(\boldsymbol{Y}_j^{\mathrm{m}}|\boldsymbol{\mu}_{ij}^{\mathrm{m\cdot o}}, \boldsymbol{\Sigma}_{ij}^{\mathrm{mm\cdot o}}) = (2\pi)^{-(p-p_j^{\mathrm{o}})/2}|\boldsymbol{\Sigma}_{ij}^{\mathrm{mm\cdot o}}|^{-1/2} \times \exp(-\tfrac{1}{2}\Delta_{ij}^{\mathrm{m\cdot o}}),$$

and

$$w_{ij}^{*} = w_i\phi_{p_j^{\mathrm{o}}}(\boldsymbol{Y}_j^{\mathrm{o}}|\boldsymbol{\mu}_{ij}^{\mathrm{o}}, \boldsymbol{\Sigma}_{ij}^{\mathrm{oo}}) / \sum_{h=1}^{g} w_h\phi_{p_j^{\mathrm{o}}}(\boldsymbol{Y}_j^{\mathrm{o}}|\boldsymbol{\mu}_{hj}^{\mathrm{o}}, \boldsymbol{\Sigma}_{hj}^{\mathrm{oo}}),$$

$$\begin{aligned}
\boldsymbol{\mu}_{ij}^{\mathrm{m\cdot o}} &= \boldsymbol{M}_j(\boldsymbol{\mu}_i + \boldsymbol{\Sigma}_i\boldsymbol{S}_{ij}^{\mathrm{oo}}(\boldsymbol{Y}_j - \boldsymbol{\mu}_i)),\\
\boldsymbol{\Sigma}_{ij}^{\mathrm{mm\cdot o}} &= \boldsymbol{E}_{ij}\boldsymbol{\Sigma}_i\boldsymbol{M}_j^{\top},\\
\boldsymbol{E}_{ij} &= \boldsymbol{M}_j(\boldsymbol{I}_p - \boldsymbol{\Sigma}_i\boldsymbol{S}_{ij}^{\mathrm{oo}}),\\
\Delta_{ij}^{\mathrm{m\cdot o}} &= (\boldsymbol{Y}_j - \boldsymbol{\mu}_i)^{\top}\boldsymbol{S}_{ij}^{\mathrm{mm\cdot o}}(\boldsymbol{Y}_j - \boldsymbol{\mu}_i),\\
\boldsymbol{S}_{ij}^{\mathrm{mm\cdot o}} &= \boldsymbol{E}_{ij}^{\top}(\boldsymbol{E}_{ij}\boldsymbol{\Sigma}_i\boldsymbol{M}_j^{\top})^{-1}\boldsymbol{E}_{ij}.
\end{aligned} \tag{5}$$

Proof. The sketch of the proof is given in Appendix A. □

To enhance the computational efficiency for estimation, we suggest to rearrange $\boldsymbol{Y}$ according to unique missing patterns of the data. The procedure can be implemented as follows:

(a) Build a binary $n \times p$ indicator matrix, $\boldsymbol{R} = [r_{ij}]$, with each entry $r_{ij} = 1$ if Y_{ij} is missing and $r_{ij} = 0$ otherwise.
(b) Build a $p \times 1$ vector $\boldsymbol{z} = \boldsymbol{R}\boldsymbol{b}$, where $\boldsymbol{b} = (2^1, 2^2, \ldots, 2^p)^{\top}$. Notice that the number of unique missing patterns is equal to the number of unique elements in $\boldsymbol{z}$.
(c) Rank $\boldsymbol{z}$ in an ascending or descending order, denoted by $\boldsymbol{z}^*$. Rearrange $\boldsymbol{Y}$ according to the row positions of $\boldsymbol{z}^*$ in $\boldsymbol{z}$. This will yield clustering of identical patterns of missingness in $\boldsymbol{Y}$ which are adjacent to each other.

3. An efficient EM procedure for ML estimation

Let $\boldsymbol{Y}^{\mathrm{o}} = (\boldsymbol{Y}_1^{\mathrm{o}}, \ldots, \boldsymbol{Y}_n^{\mathrm{o}})$ and $\boldsymbol{Y}^{\mathrm{m}} = (\boldsymbol{Y}_1^{\mathrm{m}}, \ldots, \boldsymbol{Y}_n^{\mathrm{m}})$ denote the observed portion and missing portion of the data, respectively. The complete-data log-likelihood function can be reexpressed by

$$\begin{aligned}
&\ell_c(\boldsymbol{\Theta}\,|\,\boldsymbol{Y}^{\mathrm{o}}, \boldsymbol{Y}^{\mathrm{m}}, \boldsymbol{Z})\\
&\quad= \ell_{c_1}(\boldsymbol{w}\,|\,\boldsymbol{Y}^{\mathrm{o}}, \boldsymbol{Y}^{\mathrm{m}}, \boldsymbol{Z}) + \ell_{c_2}(\boldsymbol{\Psi}\,|\,\boldsymbol{Y}^{\mathrm{o}}, \boldsymbol{Y}^{\mathrm{m}}, \boldsymbol{Z})\\
&\quad= \sum_{i=1}^{g}\sum_{j=1}^{n} Z_{ij}\log w_i\\
&\qquad+ \frac{1}{2}\sum_{i=1}^{g}\left(\log|\boldsymbol{\Sigma}_i^{-1}|\sum_{j=1}^{n} Z_{ij} - \sum_{j=1}^{n} Z_{ij}(\Delta_{ij}^{\mathrm{o}} + \Delta_{ij}^{\mathrm{m\cdot o}})\right),
\end{aligned} \tag{6}$$

where $\boldsymbol{\omega} = (w_1, \ldots, w_g)$ and $\boldsymbol{\Psi} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_g)$. From Eq. (5), it is easy to verify that $\boldsymbol{\Sigma}_i^{-1} = \boldsymbol{S}_{ij}^{\mathrm{oo}} + \boldsymbol{S}_{ij}^{\mathrm{mm\cdot o}}$ and $\boldsymbol{O}_j^{\top}\boldsymbol{O}_j(\boldsymbol{I}_p - \boldsymbol{\Sigma}_i\boldsymbol{S}_{ij}^{\mathrm{oo}}) = \boldsymbol{0}$. Hence, we have the following result.

Proposition 3. *The conditional expectation of Eq.* (6) *is given by*

$$\begin{aligned}
Q(\boldsymbol{\Theta}|\hat{\boldsymbol{\Theta}}^{(k)}) &= E(\ell_c(\boldsymbol{\Theta}\mid\boldsymbol{Y}^{\mathrm{o}}, \boldsymbol{Y}^{\mathrm{m}}, \boldsymbol{Z})\mid\boldsymbol{Y}^{\mathrm{o}}, \hat{\boldsymbol{\Theta}}^{(k)})\\
&= Q_1(\boldsymbol{w}|\hat{\boldsymbol{\Theta}}^{(k)}) + Q_2(\boldsymbol{\Psi}|\hat{\boldsymbol{\Theta}}^{(k)}).
\end{aligned}$$

It follows that

$$Q_1(\boldsymbol{w}|\hat{\boldsymbol{\Theta}}^{(k)}) = \sum_{i=1}^{g}\sum_{j=1}^{n}\hat{Z}_{ij}^{(k)}\log w_i, \tag{7}$$

$$Q_2(\boldsymbol{\Psi}|\hat{\boldsymbol{\Theta}}^{(k)}) = \frac{1}{2}\sum_{i=1}^{g}\left(\log|\boldsymbol{\Sigma}_i^{-1}|\sum_{j=1}^{n}\hat{Z}_{ij}^{(k)} - \mathrm{tr}\left(\boldsymbol{\Sigma}_i^{-1}\sum_{j=1}^{n}\boldsymbol{\Omega}_{ij}^{(k)}\right)\right), \tag{8}$$

where

$$\boldsymbol{\Omega}_{ij}^{(k)} = \hat{Z}_{ij}^{(k)}((\hat{\boldsymbol{Y}}_{ij}^{(k)} - \boldsymbol{\mu}_i)(\hat{\boldsymbol{Y}}_{ij}^{(k)} - \boldsymbol{\mu}_i)^{\top} + (\boldsymbol{I}_p - \hat{\boldsymbol{\Sigma}}_i^{(k)}\hat{\boldsymbol{S}}_{ij}^{\mathrm{oo}(k)})\hat{\boldsymbol{\Sigma}}_i^{(k)}), \tag{9}$$

$$\hat{Z}_{ij}^{(k)} = \frac{\hat{w}_i^{(k)}\phi_{p_j^{\mathrm{o}}}(\boldsymbol{Y}_j^{\mathrm{o}}|\hat{\boldsymbol{\mu}}_{ij}^{\mathrm{o}(k)}, \hat{\boldsymbol{\Sigma}}_{ij}^{\mathrm{oo}(k)})}{\sum_{h=1}^{g}\hat{w}_h^{(k)}\phi_{p_j^{\mathrm{o}}}(\boldsymbol{Y}_j^{\mathrm{o}}|\hat{\boldsymbol{\mu}}_{hj}^{\mathrm{o}(k)}, \hat{\boldsymbol{\Sigma}}_{hj}^{\mathrm{oo}(k)})}, \tag{10}$$

$$\hat{\boldsymbol{Y}}_{ij}^{(k)} = \hat{\boldsymbol{\mu}}_i^{(k)} + \hat{\boldsymbol{\Sigma}}_i^{(k)}\hat{\boldsymbol{S}}_{ij}^{\mathrm{oo}(k)}(\boldsymbol{Y}_j - \hat{\boldsymbol{\mu}}_i^{(k)}), \tag{11}$$

and $\hat{\boldsymbol{S}}_{ij}^{\mathrm{oo}(k)}$ *is* $\boldsymbol{S}_{ij}^{\mathrm{oo}}$ *given in Eq.* (4) *with* $\boldsymbol{\Sigma}_i$ *replaced by* $\hat{\boldsymbol{\Sigma}}_i^{(k)}$.

Proof. The detailed proof is shown in Appendix B. □

By these propositions, a modification of GJ's EM algorithm can be implemented as follows:

E-step: Given $\boldsymbol{\Theta} = \hat{\boldsymbol{\Theta}}^{(k)}$, impute $\hat{Z}_{ij}^{(k)}$ and $\hat{\boldsymbol{Y}}_{ij}^{(k)}$ for $i = 1, \ldots, g$ and $j = 1, \ldots, n$, using Eqs. (10) and (11).

M-Step:

1. Update $\hat{w}_i^{(k)}$ by maximizing Eq. (7) over w_i subject to their sum is unity, which gives

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

2. Fix $\boldsymbol{\Sigma}_i$ at $\hat{\boldsymbol{\Sigma}}_i^{(k)}$, update $\hat{\boldsymbol{\mu}}_i^{(k)}$ by maximizing Eq. (8) over $\boldsymbol{\mu}_i$, which leads to

$$\hat{\boldsymbol{\mu}}_i^{(k+1)} = \frac{\sum_{j=1}^{n}\hat{Z}_{ij}^{(k)}\hat{\boldsymbol{Y}}_{ij}^{(k)}}{\sum_{j=1}^{n}\hat{Z}_{ij}^{(k)}}.$$

3. Fix $\boldsymbol{\mu}_i$ at $\hat{\boldsymbol{\mu}}_i^{(k+1)}$, update $\hat{\boldsymbol{\Sigma}}_i^{(k)}$ by maximizing constrained Eq. (8) over $\boldsymbol{\Sigma}_i$, which leads to

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

where $\hat{\boldsymbol{\Omega}}_{ij}^{(k)}$ is $\boldsymbol{\Omega}_{ij}^{(k)}$ in Eq. (9) with $\boldsymbol{\mu}_i$ replaced by $\hat{\boldsymbol{\mu}}_i^{(k+1)}$. We remark two major advantages of the above EM algorithm:

(a) With auxiliary matrices $\boldsymbol{O}_j$'s obtained at the initiation, there is no need to take care of the associated row positions of missing values at each iteration.
(b) The implementation of the M-step has low computational cost as it is similar to the case of no missing values. Therefore, the modified EM algorithm is more straightforward than the version of GJ.

Applying Bayes' theorem, the posterior probability of the $\boldsymbol{Y}_j$ belonging to $\mathscr{C}_i$ can be estimated by

$$\hat{w}_{ij}^* = \Pr(Z_{ij}=1|\boldsymbol{Y}^{\mathrm{o}}, \hat{\boldsymbol{\Theta}}) = \frac{\hat{w}_i\, \phi_{p_j^{\mathrm{o}}}(\boldsymbol{Y}_j^{\mathrm{o}}|\hat{\boldsymbol{\mu}}_{ij}^{\mathrm{o}}, \hat{\boldsymbol{\Sigma}}_{ij}^{\mathrm{oo}})}{\sum_{h=1}^g \hat{w}_h \phi_{p_j^{\mathrm{o}}}(\boldsymbol{Y}_j^{\mathrm{o}}|\hat{\boldsymbol{\mu}}_{hj}^{\mathrm{o}}, \hat{\boldsymbol{\Sigma}}_{hj}^{\mathrm{oo}})}. \quad (12)$$

By the ML classification theory [20], $\boldsymbol{Y}_j$ is assigned to $\mathscr{C}_s$ if $\hat{w}_{sj}^* > \hat{w}_{ij}^*$ $(i=1,\ldots,g;\ i\neq s)$.

Consequently, an ML predictor for the missing component $\boldsymbol{Y}_j^{\mathrm{m}}$ is given by

$$\hat{\boldsymbol{Y}}_j^{\mathrm{m}} = E(\boldsymbol{Y}_j^{\mathrm{m}}|\boldsymbol{Y}^{\mathrm{o}}, \hat{\boldsymbol{\Theta}}) = \boldsymbol{M}_j \sum_{i=1}^g \hat{w}_{ij}^* (\hat{\boldsymbol{\mu}}_i + \hat{\boldsymbol{\Sigma}}_i \hat{\boldsymbol{S}}_{ij}^{\mathrm{oo}} (\boldsymbol{Y}_j - \hat{\boldsymbol{\mu}}_i)). \quad (13)$$

4. A data augmentation scheme for Bayesian sampling

The DA algorithm [9] is a general and effective algorithm for producing multiple imputation of missing data. The DA has been broadly applied in a variety of missing data problems, see Refs. [4,19] and references therein. In this section, we construct an efficient DA algorithm that combines the latent variables $\boldsymbol{Z}$ and unobserved data $\boldsymbol{Y}^{\mathrm{m}}$ for simulating the posterior density of $\boldsymbol{\Theta}$.

The DA algorithm consists of the imputation step (I-step) and the posterior step (P-step). At the kth iteration of the DA algorithm, the I-step is defined by drawing imputations of $\boldsymbol{Z}_j^{(k)}$ and $\boldsymbol{Y}_j^{\mathrm{m}^{(k)}}$ from the predictive distributions $p(\boldsymbol{Z}_j \mid \boldsymbol{Y}^{\mathrm{o}}, \boldsymbol{\Theta}^{(k)})$ and $p(\boldsymbol{Y}_j^{\mathrm{m}} \mid \boldsymbol{Y}^{\mathrm{o}}, \boldsymbol{Z}_j, \boldsymbol{\Theta}^{(k)})$, respectively for all j, and the P-step refers to generating $\boldsymbol{\Theta}^{(k+1)}$ from $p(\boldsymbol{\Theta} \mid \boldsymbol{Y}^{\mathrm{o}}, \boldsymbol{Y}^{\mathrm{m}^{(k+1)}}, \boldsymbol{Z}^{(k+1)})$. If iterations are performed by a sufficiently long burn-in period, then the simulations $\boldsymbol{Z}_j^{(k)}, \boldsymbol{Y}_j^{\mathrm{m}^{(k)}}$ and $\boldsymbol{\Theta}^{(k)}$ are distributed according to $p(\boldsymbol{Z}_j \mid \boldsymbol{Y}^{\mathrm{o}})$, $p(\boldsymbol{Y}_j^{\mathrm{m}} \mid \boldsymbol{Y}^{\mathrm{o}})$ and $p(\boldsymbol{\Theta} \mid \boldsymbol{Y}^{\mathrm{o}})$, respectively for all k. To perform the Bayesian inference for mixture models, it is necessary to choose a proper prior distribution for each parameter to avoid yielding improper posterior distributions [17].

In various mixture contexts, a vague Dirichlet distribution, denoted by $\mathscr{D}(\delta,\ldots,\delta)$, is the most natural prior for mixing probabilities $\boldsymbol{w}$. Its density is proportional to $w_1^{\delta-1}\cdots w_{g-1}^{\delta-1}(1-w_1-\cdots-w_{g-1})^{\delta-1}$. For component mean vectors $\boldsymbol{\mu}_i$, it is standard to adopt conjugate Gaussian priors. As for the inverse covariance matrix $\boldsymbol{\Sigma}_i^{-1}$, the Wishart distribution is often chosen as a conjugate prior. A p-dimensional Wishart distribution with parameters v and $\boldsymbol{A}$ $(p\times p)$ is denoted by $\mathscr{W}_p(v,\boldsymbol{A})$, and for $v\geqslant p$ the density is

$$f(\boldsymbol{U}|\boldsymbol{A}) \propto |\boldsymbol{A}|^{-v/2}|\boldsymbol{U}|^{(v-p-1)/2}\exp\left(-\frac{1}{2}\operatorname{tr}(\boldsymbol{U}\boldsymbol{A}^{-1})\right),$$

where $\boldsymbol{A}$ is called a *hyperparameter* matrix if $\boldsymbol{U}$ is considered random in Bayesian treatments.

Following the suggestion of [11,14] who base their recommendation on the use of conjugate priors, our chosen priors are

$$\begin{aligned}
&\boldsymbol{w} \sim \mathscr{D}(\delta,\ldots,\delta),\\
&\boldsymbol{\mu}_i \sim \mathscr{N}_p\left(\boldsymbol{\xi}, \boldsymbol{\kappa}^{-1}\right) \quad (i=1,\ldots,g),\\
&\boldsymbol{\Sigma}_i^{-1}\,|\,\boldsymbol{B} \sim \mathscr{W}_p\left(2\,\alpha, (2\boldsymbol{B})^{-1}\right) \quad (i=1,\ldots,g),\\
&\boldsymbol{B} \sim \mathscr{W}_p\left(2\gamma, (2\mathbf{H})^{-1}\right),
\end{aligned}$$

where $\boldsymbol{B}$ is a hyperparameter matrix with a conjugate Wishart distribution, and $(\boldsymbol{\kappa}, \mathbf{H}, \boldsymbol{\xi}, \alpha, \delta, \gamma)$ are fixed as appropriate quantities to reflect the flatness of priors. The joint prior distribution function of $\boldsymbol{\Theta}$ and $\boldsymbol{B}$ is

$$\begin{aligned}
\pi(\boldsymbol{\Theta}, \boldsymbol{B}) \propto\ & |\boldsymbol{B}|^{g\alpha+(2\gamma-p-1)/2}\exp(-\operatorname{tr}(\mathbf{H}\boldsymbol{B}))\prod_{i=1}^g w_i^{\delta-1}\\
&\times\prod_{i=1}^g |\boldsymbol{\Sigma}_i^{-1}|^{(2\alpha-p-1)/2}\exp\\
&\times\left(-\frac{1}{2}(\boldsymbol{\mu}_i-\boldsymbol{\xi})^\top\boldsymbol{\kappa}(\boldsymbol{\mu}_i-\boldsymbol{\xi}) - \operatorname{tr}\left(\boldsymbol{B}\boldsymbol{\Sigma}_i^{-1}\right)\right). \quad (14)
\end{aligned}$$

Upon multiplying Eqs. (3) and (14), we have the following joint posterior density:

$$\begin{aligned}
&p(\boldsymbol{\Theta}, \boldsymbol{B}, \boldsymbol{Y}^{\mathrm{m}}, \boldsymbol{Z}|\boldsymbol{Y}^{\mathrm{o}})\\
&\quad\propto w_1^{\delta-1}\cdots w_g^{\delta-1}\,|\,\boldsymbol{B}|^{g\alpha+(2\gamma-p-1)/2}\exp(-\operatorname{tr}(\mathbf{H}\boldsymbol{B}))\\
&\qquad\times\prod_{i=1}^g \exp\left(-\frac{1}{2}(\boldsymbol{\mu}_i-\boldsymbol{\xi})^\top\boldsymbol{\kappa}(\boldsymbol{\mu}_i-\boldsymbol{\xi})\right)\\
&\qquad\times\left|\boldsymbol{\Sigma}_i^{-1}\right|^{(2\alpha-p-1)/2}\exp(-\operatorname{tr}(\boldsymbol{B}\boldsymbol{\Sigma}_i^{-1}))\\
&\qquad\times\prod_{j=1}^n\prod_{i=1}^g\left(w_i\,|\,\boldsymbol{\Sigma}_i^{-1}|^{1/2}\exp\left(-\frac{1}{2}(\Delta_{ij}^{\mathrm{o}}+\Delta_{ij}^{\mathrm{m\cdot o}})\right)\right)^{Z_{ij}},
\end{aligned} \quad (15)$$

where Δ_{ij}^{o} and $\Delta_{ij}^{\mathrm{m\cdot o}}$ are given in Eqs. (4) and (5), respectively.

Proposition 4. *The full conditional posteriors of* $\boldsymbol{\Theta}$, $\boldsymbol{B}$, $\boldsymbol{Z}$ *and* $\boldsymbol{Y}^{\mathrm{m}}$ *are as follows (the symbol* "$|\cdots$" *denotes conditioning on all other variables):*

$$p(\boldsymbol{Z}_j|\boldsymbol{Y}^{\mathrm{o}},\boldsymbol{\Theta}) \propto \prod_{i=1}^{g}(w_i\phi_{p_j^{\mathrm{o}}}(\boldsymbol{Y}_j^{\mathrm{o}}|\boldsymbol{\mu}_{ij}^{\mathrm{o}},\boldsymbol{\Sigma}_{ij}^{\mathrm{oo}})^{Z_{ij}},$$

$$p(\boldsymbol{Y}_j^{\mathrm{m}}|Z_{ij}=1,\cdots) \propto \exp\left(-\frac{1}{2}(\boldsymbol{Y}_j^{\mathrm{m}}-\boldsymbol{\mu}_{ij}^{\mathrm{m\cdot o}})^{\top}\boldsymbol{\Sigma}_{ij}^{\mathrm{mm\cdot o}^{-1}}(\boldsymbol{Y}_j^{\mathrm{m}}-\boldsymbol{\mu}_{ij}^{\mathrm{m\cdot o}})\right),$$

$$p(\boldsymbol{w}|\cdots) \propto \prod_{i=1}^{g} w_i^{\sum_{j=1}^{n}Z_{ij}+\delta-1},$$

$$p(\boldsymbol{\mu}_i|\cdots) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_i^*)^{\top}\boldsymbol{\Sigma}_i^{*^{-1}}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_i^*)\right),$$

$$p(\boldsymbol{B}|\cdots) \propto |\boldsymbol{B}|^{(2(g\alpha+\gamma)-p-1)/2} \times \exp\left(-\mathrm{tr}\left(\boldsymbol{B}\left(\mathbf{H}+\sum_{i=1}^{g}\boldsymbol{\Sigma}_i^{-1}\right)\right)\right),$$

$$p(\boldsymbol{\Sigma}_i^{-1}|\cdots) \propto \left|\boldsymbol{\Sigma}_i^{-1}\right|^{(\alpha^*-p-1)/2}\exp\left(-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}_i^{-1}\boldsymbol{A}_i)\right),$$

where $\boldsymbol{\mu}_{ij}^{\mathrm{m\cdot o}}$ *and* $\boldsymbol{\Sigma}_{ij}^{\mathrm{mm\cdot o}}$ *are given by Eq.* (5), *and*

$$\boldsymbol{\Sigma}_i^* = \left(\boldsymbol{\Sigma}_i^{-1}\sum_{j=1}^{n}Z_{ij}+\boldsymbol{\kappa}\right)^{-1}, \tag{16}$$

$$\boldsymbol{\mu}_i^* = \boldsymbol{\Sigma}_i^*\left(\boldsymbol{\Sigma}_i^{-1}\sum_{j=1}^{n}Z_{ij}\boldsymbol{Y}_j+\boldsymbol{\kappa}\boldsymbol{\xi}\right), \tag{17}$$

$$\alpha_i^* = \sum_{j=1}^{n}Z_{ij}+2\alpha, \tag{18}$$

$$\boldsymbol{A}_i = 2\boldsymbol{B}+\sum_{j=1}^{n}Z_{ij}\left(\boldsymbol{Y}_j-\boldsymbol{\mu}_i\right)\left(\boldsymbol{Y}_j-\boldsymbol{\mu}_i\right)^{\top}, \tag{19}$$

for $i=1,\ldots,g$ *and* $j=1,\ldots,n$.

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

In the simulation process, samples for $\boldsymbol{Z}$, $\boldsymbol{Y}^{\mathrm{m}}$, $\boldsymbol{B}$ and $\boldsymbol{\Theta}$ are alternately generated, the DA algorithm using the Gibbs sampler can be implemented as follows:

I-Step:

1. Given $\boldsymbol{\Theta}$, $\boldsymbol{Y}^{\mathrm{m}}$ and $\boldsymbol{Y}^{\mathrm{o}}$, generate $\boldsymbol{Z}_j$ from $\mathscr{M}(1; r_{1j},\ldots,r_{gj})$, where
$$r_{ij} = \frac{w_i\phi_{p_j^{\mathrm{o}}}\left(\boldsymbol{Y}_j^{\mathrm{o}}|\boldsymbol{\mu}_{ij}^{\mathrm{o}},\boldsymbol{\Sigma}_{ij}^{\mathrm{oo}}\right)}{\sum_{h=1}^{g}w_h\phi_{p_j^{\mathrm{o}}}\left(\boldsymbol{Y}_j^{\mathrm{o}}|\boldsymbol{\mu}_{hj}^{\mathrm{o}},\boldsymbol{\Sigma}_{hj}^{\mathrm{oo}}\right)}.$$
2. Generate $\boldsymbol{Y}_j^{\mathrm{m}}$ given $Z_{ij}=1$, $\boldsymbol{\Theta}$ and $\boldsymbol{Y}^{\mathrm{o}}$, from $N_{p-p_j^{\mathrm{o}}}\left(\boldsymbol{\mu}_{ij}^{\mathrm{m\cdot o}},\boldsymbol{\Sigma}_{ij}^{\mathrm{mm\cdot o}}\right)$, where $\boldsymbol{\mu}_{ij}^{\mathrm{m\cdot o}}$ and $\boldsymbol{\Sigma}_{ij}^{\mathrm{mm\cdot o}}$ are as in Eq. (5).

P-Step:

1. Generate $\boldsymbol{w}$ given $\boldsymbol{Z}$ from $\mathscr{D}(n_1+\delta,\ldots,n_g+\delta)$, where $n_i=\sum_{j=1}^{n}Z_{ij}$.
2. Generate $\boldsymbol{\mu}_i$ given $\boldsymbol{Z}$, $\boldsymbol{\Sigma}_i$, $\boldsymbol{Y}^{\mathrm{o}}$ and $\boldsymbol{Y}^{\mathrm{m}}$ from $\mathscr{N}_p\left(\boldsymbol{\mu}_i^*,\boldsymbol{\Sigma}_i^*\right)$ with $\boldsymbol{\mu}_i^*$ and $\boldsymbol{\Sigma}_i^*$ given in Eqs. (17) and (16), respectively.
3. Generate $\boldsymbol{B}$ given $\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_g$ from $\mathscr{W}_p(2\gamma^*,(2\mathbf{H}^*)^{-1})$, where $\gamma^*=g\alpha+\gamma$ and $\mathbf{H}^*=\mathbf{H}+\sum_{i=1}^{g}\boldsymbol{\Sigma}_i^{-1}$.
4. Generate $\boldsymbol{\Sigma}_i^{-1}$ given $\boldsymbol{Z}$, $\boldsymbol{\mu}_i$, $\boldsymbol{Y}^{\mathrm{o}}$ and $\boldsymbol{Y}^{\mathrm{m}}$ from $\mathscr{W}_p(\alpha_i^*,\boldsymbol{A}_i^{-1})$, where α_i^* and $\boldsymbol{A}_i$ are given in Eqs. (18) and (19), respectively.

To satisfy the "Principle of Stable Estimation" of Edwards et al. [21] in the Bayesian treatment, we need to specify $(\boldsymbol{\xi},\boldsymbol{\kappa},\alpha,\gamma,\mathbf{H})$ so as to be insensitive to changes of the prior. Specifically, it is often to choose $\delta=1$. For $\boldsymbol{\xi}$ and $\boldsymbol{\kappa}$, we let $\boldsymbol{\xi}$ be the empirical mean vector and $\boldsymbol{\kappa}^{-1}=(1-\eta)^{-1}\mathrm{diag}\{R_1^2,\ldots,R_p^2\}$, where η is the percentage of missing values of the data which is used to adjust the flatness and R_i is the range of the observed values for the ith attribute. This specification makes a weak prior information for $\boldsymbol{\mu}_i$. As a generalization of [11], we take $\alpha=p+1$, $\gamma=(p+1)/10$ and $\mathbf{H}=10\boldsymbol{\kappa}$.

We are interested in the classification and prediction problems for incomplete features. Under certain conditions, quantities based on Rao-Blackwellization [22] often greatly improve the precision of Monte Carlo estimates. Given a set of converged Monte Carlo DA samples $\boldsymbol{\Theta}^{(\ell)}$ $(\ell=1,\ldots,L)$, a Bayesian predictor for $\boldsymbol{Y}_j^{\mathrm{m}}$ is given by

$$\begin{aligned}\widetilde{\boldsymbol{Y}}_j^{\mathrm{m}} &= \frac{1}{L}\sum_{\ell=1}^{L}E(\boldsymbol{Y}_j^{\mathrm{m}}|\boldsymbol{Y}_j^{\mathrm{o}},\boldsymbol{\Theta}^{(\ell)})\\ &=\boldsymbol{M}_j\,\frac{1}{L}\sum_{\ell=1}^{L}\left(\sum_{i=1}^{g}r_{ij}^{(\ell)}(\boldsymbol{\mu}_i^{(\ell)}+\boldsymbol{\Sigma}_i^{(\ell)}\boldsymbol{S}_{ij}^{\mathrm{oo}^{(\ell)}}(\boldsymbol{Y}_j-\boldsymbol{\mu}_i^{(\ell)}))\right),\end{aligned} \tag{20}$$

where

$$r_{ij}^{(\ell)} = \frac{w_i^{(\ell)}\phi_{p_j^{\mathrm{o}}}(\boldsymbol{Y}_j^{\mathrm{o}}|\boldsymbol{\mu}_{ij}^{\mathrm{o}^{(\ell)}},\boldsymbol{\Sigma}_{ij}^{\mathrm{oo}^{(\ell)}})}{\sum_{h=1}^{g}w_h^{(\ell)}\phi_{p_j^{\mathrm{o}}}(\boldsymbol{Y}_j^{\mathrm{o}}|\boldsymbol{\mu}_{hj}^{\mathrm{o}^{(\ell)}},\boldsymbol{\Sigma}_{hj}^{\mathrm{oo}^{(\ell)}})}.$$

Consequently, a Bayesian classifier for $\boldsymbol{Y}_j$ can be estimated by averaging over the draws of $\boldsymbol{\Theta}^{(\ell)}$

$$\hat{r}_{ij}^* = \Pr(Z_{ij}=1|\boldsymbol{Y}_j^{\mathrm{o}}) \approx \frac{1}{L}\sum_{\ell=1}^{L}r_{ij}^{(\ell)}. \tag{21}$$

By the Bayesian classification rule, $\boldsymbol{Y}_j$ is assigned to $\mathscr{C}_s$ if $\hat{r}_{sj}^*>\hat{r}_{ij}^*$ $(i=1,\ldots,g;\ i\neq s)$.

Table 1
A comparison of CPU time (in seconds) and relative reduced time (RRT) between GJ-EM algorithm (old) and the proposed procedure (new) under various missing rates (Replications = 500)

Data	$\eta = 10\%$			$\eta = 20\%$			$\eta = 30\%$		
	Old	New	RRT (%)	Old	New	RRT (%)	Old	New	RRT (%)
Iris	12.47	1.22	90.2	21.51	1.61	92.5	56.21	3.61	93.6
Crabs	34.72	3.27	90.6	78.77	6.78	91.4	265.01	20.68	92.2

RRT = (old − new)/old × 100%

Table 2
A comparison of prediction accuracies for MI, EM and DA imputations with the standard deviations in parentheses for the *iris* data set (Replications=500)

η (%)	MAE			MARE			RMSE		
	MI	EM	DA	MI	EM	DA	MI	EM	DA
10	0.812 (0.081)	0.213 (0.026)	0.210 (0.026)	0.697 (0.186)	0.100 (0.027)	0.099 (0.027)	1.062 (0.096)	0.285 (0.050)	0.280 (0.050)
20	0.816 (0.053)	0.237 (0.025)	0.233 (0.025)	0.675 (0.129)	0.114 (0.031)	0.113 (0.031)	1.071 (0.065)	0.331 (0.060)	0.326 (0.060)
30	0.820 (0.046)	0.268 (0.023)	0.259 (0.022)	0.684 (0.097)	0.138 (0.033)	0.132 (0.032)	1.078 (0.058)	0.395 (0.061)	0.380 (0.060)
40	0.819 (0.035)	0.301 (0.030)	0.278 (0.026)	0.683 (0.082)	0.161 (0.038)	0.154 (0.036)	1.077 (0.041)	0.448 (0.065)	0.428 (0.063)
50	0.817 (0.029)	0.346 (0.031)	0.325 (0.028)	0.675 (0.084)	0.198 (0.043)	0.188 (0.041)	1.074 (0.036)	0.522 (0.063)	0.495 (0.060)

Table 3
A comparison of prediction accuracies for MI, EM and DA imputations with the standard deviations in parentheses for the *crabs* data set (Replications=500)

η (%)	MAE			MARE			RMSE		
	MI	EM	DA	MI	EM	DA	MI	EM	DA
10	4.063 (0.337)	0.421 (0.055)	0.415 (0.050)	0.202 (0.018)	0.024 (0.003)	0.023 (0.003)	5.391 (0.427)	0.611 (0.114)	0.598 (0.105)
20	4.008 (0.227)	0.484 (0.041)	0.474 (0.037)	0.200 (0.012)	0.027 (0.002)	0.026 (0.002)	5.343 (0.305)	0.714 (0.090)	0.693 (0.083)
30	4.037 (0.169)	0.568 (0.044)	0.550 (0.041)	0.202 (0.009)	0.030 (0.002)	0.029 (0.002)	5.384 (0.225)	0.846 (0.096)	0.812 (0.091)
40	4.036 (0.138)	0.662 (0.044)	0.632 (0.042)	0.203 (0.007)	0.035 (0.002)	0.033 (0.002)	5.381 (0.188)	0.977 (0.092)	0.932 (0.094)
50	4.039 (0.108)	0.768 (0.052)	0.728 (0.050)	0.202 (0.006)	0.039 (0.002)	0.037 (0.002)	5.386 (0.142)	1.120 (0.102)	1.058 (0.100)

5. Experimental results

For illustration purposes, we start to apply results developed in Sections 2–4 to two famous multivariate data sets. One is the *iris* data taken from Anderson [23] or Fisher [24]. It consists of four-dimensional measurements in centimeters on the attributes of petal length, petal width, sepal length and sepal width for 50 flower specimens of each of three species: setsosa, versicolor, and virginica. The other is the *crabs* data of Campbell and Mahon [25] on the *gensus Leptograpsus*. It consists of

Table 4
A comparison of average misclassification rates (%) between ML and Bayesian classifiers (replicates = 500)

η (%)	Iris		Crabs	
	ML	Bayesian	ML	Bayesian
0	3.33	3.00	7.50	7.30
10	3.85	3.75	9.75	9.50
20	5.20	5.00	13.66	13.55
30	6.90	6.10	19.22	18.80
40	10.15	9.20	26.75	25.20
50	13.42	12.30	35.21	33.00

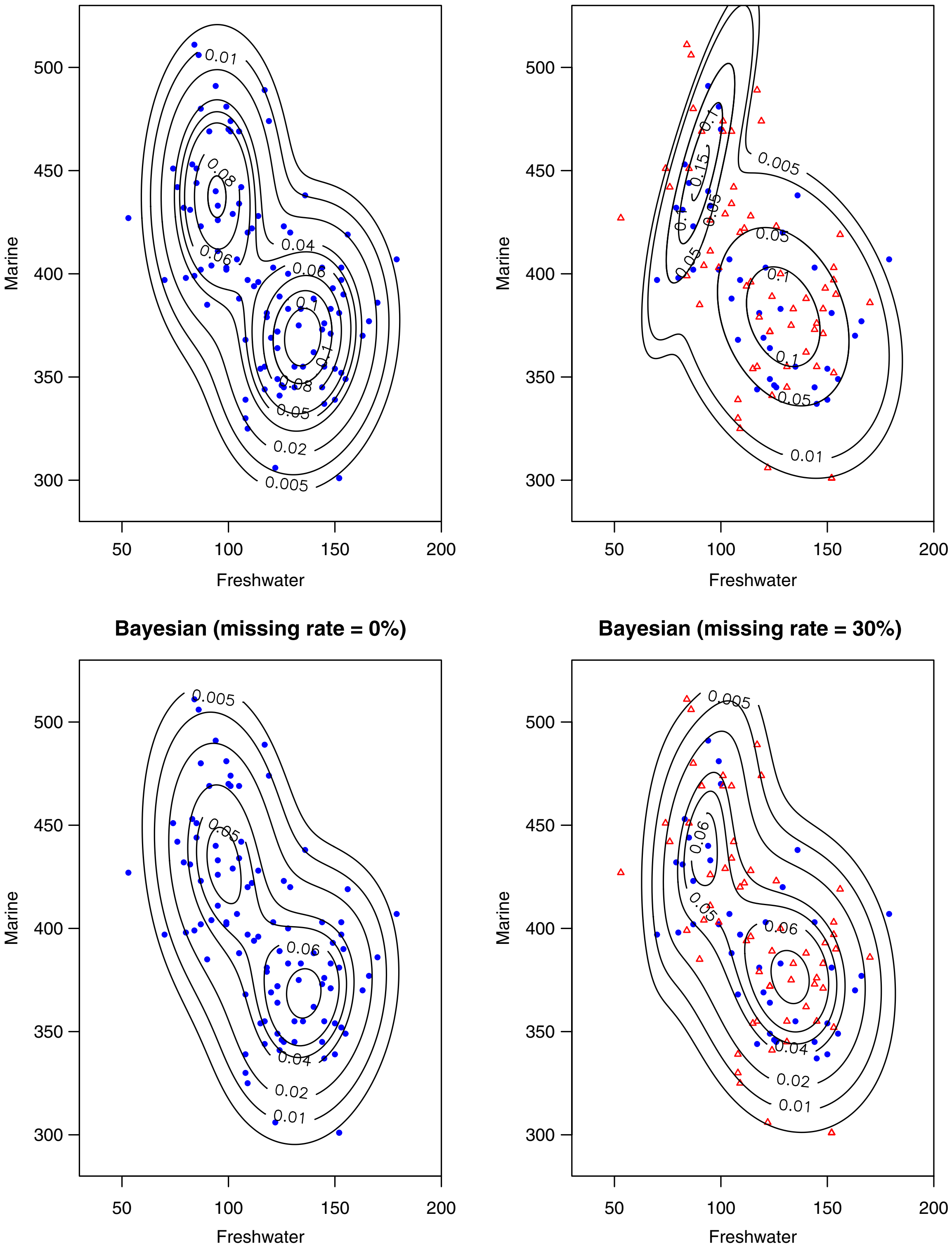

Fig. 1. ML and Bayesian density estimation for the two-component salmon data (●, both attributes are completely observed; △, one of the two attributes is missing).

five-dimensional morphological measurements on the attributes of width of frontal lip, rear width, length along the mid-line of the carapace, maximum width of the carapace and body depth for 50 crabs of each of four groups: blue male, blue female, orange male and orange female. Both data sets are included as a part of the R package, which is freely available at the web site http://cran.r-project.org.

To conduct experimental studies, we first generate 500 artificially missing data sets by deleting at random from the three data sets under various specified missing rate η (proportion of missing values) while we maintain each datum to have at least one observed attribute. Table 1 presents the computation times of our developed EM algorithm and those of using GJ-EM. All computations are solely carried out by R package in the environment of a desktop PC (CPU: 3Gb-MHz/Intel Pentium 4 Processor; RAM: 1Gb/DDR-400). Since the programming implementations have many characteristics (e.g., vector or matrix subroutines instead of loops), the CPU times in Table 1 might not be directly comparable, but provide a sense of their actual performances in a practical setting. As seen in the table, all computation times are dramatically reduced over 90% by using the new EM procedure.

To exemplify the predictive performance for the EM and DA imputation methods, see Eqs. (13) and (20), together with the traditional mean imputation (MI) method, known as “filling-in” with the sample mean of the associated attribute, we utilize the pseudo-cross-validation (PSV) of Stone [26] to evaluate these three approaches. A relative tolerance of 10^{-8} for the log-likelihood function and parameter estimates are used as the convergence criterion for the EM algorithm. As for the DA algorithm, we take the ML estimates as the initialization and carry out 2000 iterations with the first 1000 iterations as burn-in and the remaining 1000 iterations as inference samples. It is noted that our chosen burn-in number is much larger than needed based on checking the *multivariate potential scale reduction factor* (MPSRF) of Brooks and Gelman [27]. As for discrepancy measures, we use the mean absolute error (MAE), the mean absolute relative error (MARE) and root mean square error (RMSE). Comparison results are listed in Tables 2 and 3. As seen in the tables, we found that both EM and DA substantially outperform MI for all cases. Furthermore, DA imputation exhibits considerable promising accuracy in the prediction of missing values when compared to the EM imputation, especially as the size of observed values becomes small (i.e., missing rate increases).

As another illustration, we attempt to explore classification accuracies between the ML classifier Eq. (12) and the Bayesian classifier Eq. (21) via PSV. Experimental results in Table 4 indicate that both classifiers are comparable at low-level missing, but Bayesian classifier yields lower misclassification rates as the missing rate increases, though improvements are not substantial.

Finally, we are interested in comparing behaviors of density estimation from both ML-fitted and Bayesian posterior predictive aspects. To illustrate this, we use the *salmon* data taken from Johnson and Wichern [28]. This data set has two attributes, the diameter of rings for the first-year freshwater growth and the diameter of rings for the first-year marine growth (both measured in hundredths of an inch), for each of 50 Alaskan-born and Canadian-born salmon fishes. The ML-fitted density estimation is obtained by plugging the ML estimates into Eq. (1). As for Bayesian predictive density, it can be approximated by the use of Rao-Blackwellization

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{Y}^{\mathrm{o}}) &= \int p(\boldsymbol{y}|\boldsymbol{Y}^{\mathrm{o}}, \boldsymbol{\Theta})\, p(\boldsymbol{\Theta}|\boldsymbol{Y}^{\mathrm{o}})\, \mathrm{d}\boldsymbol{\Theta} \\ &\approx \frac{1}{L}\sum_{\ell=1}^{L} p(\boldsymbol{y}|\boldsymbol{\Theta}^{(\ell)}) \\ &= \frac{1}{L}\sum_{\ell=1}^{L}\left(\sum_{i=1}^{g} w_i^{(\ell)}\left((2\pi)^{-p/2}|\boldsymbol{\Sigma}_i^{(\ell)}|^{-1/2}\right.\right. \\ &\quad \left.\left.\times\ \exp\left(-\frac{1}{2}\Delta_{ij}^{(\ell)}\right)\right)\right), \qquad (22) \end{aligned}$$

where $\Delta_{ij}^{(\ell)} = (\boldsymbol{y} - \boldsymbol{\mu}_i^{(\ell)})^{\top}\boldsymbol{\Sigma}_i^{(\ell)^{-1}}(\boldsymbol{y} - \boldsymbol{\mu}_i^{(\ell)})$ and $\boldsymbol{\Theta}^{(\ell)}$ $(\ell = 1, \ldots, L)$ is a set of converged Monte Carlo samples generated from the DA algorithm.

The contour plots obtained by the ML-fitting and Bayesian predictive densities Eq. (22) for both completely observed data ($\eta = 0\%$) and partially observed data ($\eta = 30\%$) are depicted in Fig. 1, respectively. Both look similar when data are not missing but using Eq. (22) seems to have a relatively smoother appearance. In addition, we found that the ML-fitted contour shapes tend to be distorted at high-level missing and even for moderate-level missing ($\eta = 30\%$). However, the distortion rarely happened while using Eq. (22). This indicates that Bayesian learning is more resistant to missing values.

6. Conclusions

In this paper, two novel EM and DA computational algorithms for learning normal mixture models under a missing information framework are presented. It should be emphasized that our proposed procedures offer neat ways to program with low-cost computation. Experimental results indicate that Bayesian treatment is a worthwhile tool for mixture modelling under a considerable extent of missing information.

Recently, Bayesian and non-Bayesian robust mixture model modelling using the t distribution has received notable attentions, see Refs. [29–32]. Future work will make some kind of comparisons theoretically or empirically among various competitive models.

Acknowledgments

The authors are grateful to the Editors and one anonymous referee for their helpful comments and suggestions that substantially improve the presentation of the paper. This work was supported by the National Science Council of Taiwan.

Appendix A. Proof of Proposition 2

Suppose $\boldsymbol{Y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then for any $q \times p$ matrix $\boldsymbol{A}$ with rank $q(q \leqslant p)$, we can obtain $\boldsymbol{AY} \sim N_p(\boldsymbol{A\mu}, \boldsymbol{A\Sigma A}^\top)$. With similar arguments, the marginal distributions of $\boldsymbol{Y}_j^{\mathrm{o}}$ and $\boldsymbol{Y}_j^{\mathrm{m}}$ are:

$$\boldsymbol{Y}_j^{\mathrm{o}} = \boldsymbol{O}_j\boldsymbol{Y}_j \sim \sum_{i=1}^{g} w_i\, \phi_{p_j^{\mathrm{o}}}(\boldsymbol{\mu}_{ij}^{\mathrm{o}}, \boldsymbol{\Sigma}_{ij}^{\mathrm{oo}}), \quad \boldsymbol{\mu}_{ij}^{\mathrm{o}} = \boldsymbol{O}_j\boldsymbol{\mu}_i,$$
$$\boldsymbol{\Sigma}_{ij}^{\mathrm{oo}} = \boldsymbol{O}_j\boldsymbol{\Sigma}_i\boldsymbol{O}_j^\top,$$
$$\boldsymbol{Y}_j^{\mathrm{m}} = \boldsymbol{M}_j\boldsymbol{Y}_j \sim \sum_{i=1}^{g} w_i\, \phi_{p-p_j^{\mathrm{o}}}(\boldsymbol{\mu}_{ij}^{\mathrm{m}}, \boldsymbol{\Sigma}_{ij}^{\mathrm{mm}}), \quad \boldsymbol{\mu}_{ij}^{\mathrm{m}} = \boldsymbol{M}_j\boldsymbol{\mu}_i,$$
$$\boldsymbol{\Sigma}_{ij}^{\mathrm{mm}} = \boldsymbol{M}_j\boldsymbol{\Sigma}_i\boldsymbol{M}_j^\top.$$

Note that the Δ_{ij} in Eq. (2) can be reexpressed as

$$\Delta_{ij} = \begin{bmatrix} \boldsymbol{Y}_j^{\mathrm{o}} - \boldsymbol{\mu}_{ij}^{\mathrm{o}} \\ \boldsymbol{Y}_j^{\mathrm{m}} - \boldsymbol{\mu}_{ij}^{\mathrm{m}} \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{\Sigma}_{ij}^{\mathrm{oo}} & \boldsymbol{\Sigma}_{ij}^{\mathrm{om}} \\ \boldsymbol{\Sigma}_{ij}^{\mathrm{mo}} & \boldsymbol{\Sigma}_{ij}^{\mathrm{mm}} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{Y}_j^{\mathrm{o}} - \boldsymbol{\mu}_{ij}^{\mathrm{o}} \\ \boldsymbol{Y}_j^{\mathrm{m}} - \boldsymbol{\mu}_{ij}^{\mathrm{m}} \end{bmatrix}, \qquad (23)$$

where $\boldsymbol{\Sigma}_{ij}^{\mathrm{om}} = \boldsymbol{O}_j\boldsymbol{\Sigma}_i\boldsymbol{M}_j^\top$ and $\boldsymbol{\Sigma}_{ij}^{\mathrm{mo}} = \boldsymbol{M}_j\boldsymbol{\Sigma}_i\boldsymbol{O}_j^\top$. Also, the second and third factors on the right hand side of Eq. (23) can be represented by

$$\begin{bmatrix} \boldsymbol{\Sigma}_{ij}^{\mathrm{oo}} & \boldsymbol{\Sigma}_{ij}^{\mathrm{om}} \\ \boldsymbol{\Sigma}_{ij}^{\mathrm{mo}} & \boldsymbol{\Sigma}_{ij}^{\mathrm{mm}} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{\Sigma}_{ij}^{\mathrm{oo}^{-1}}\boldsymbol{\Sigma}_{ij}^{\mathrm{om}} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_{ij}^{\mathrm{oo}^{-1}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_{ij}^{\mathrm{mm\cdot o}^{-1}} \end{bmatrix} \times \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ -\boldsymbol{\Sigma}_{ij}^{\mathrm{mo}}\boldsymbol{\Sigma}_{ij}^{\mathrm{oo}^{-1}} & \boldsymbol{I} \end{bmatrix},$$

and

$$\begin{bmatrix} \boldsymbol{Y}_j^{\mathrm{o}} - \boldsymbol{\mu}_{ij}^{\mathrm{o}} \\ \boldsymbol{Y}_j^{\mathrm{m}} - \boldsymbol{\mu}_{ij}^{\mathrm{m}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{O}_j(\boldsymbol{Y}_j - \boldsymbol{\mu}_i) \\ \boldsymbol{M}_j(\boldsymbol{Y}_j - \boldsymbol{\mu}_i) \end{bmatrix} = \begin{bmatrix} \boldsymbol{O}_j \\ \boldsymbol{M}_j \end{bmatrix} (\boldsymbol{Y}_j - \boldsymbol{\mu}_i).$$

We then have the following standard results:

$$\begin{aligned} \boldsymbol{\Sigma}_{ij}^{\mathrm{mm\cdot o}} &= \boldsymbol{\Sigma}_{ij}^{\mathrm{mm}} - \boldsymbol{\Sigma}_{ij}^{\mathrm{mo}}\boldsymbol{\Sigma}_{ij}^{\mathrm{oo}^{-1}}\boldsymbol{\Sigma}_{ij}^{\mathrm{om}} \\ &= \boldsymbol{M}_j\boldsymbol{\Sigma}_i\boldsymbol{M}_j^\top - \boldsymbol{M}_j\boldsymbol{\Sigma}_i\boldsymbol{O}_j^\top(\boldsymbol{O}_j\boldsymbol{\Sigma}_i\boldsymbol{O}_j^\top)^{-1}\boldsymbol{O}_j\boldsymbol{\Sigma}_i\boldsymbol{M}_j^\top \\ &= \boldsymbol{M}_j(\boldsymbol{I}_p - \boldsymbol{\Sigma}_i\boldsymbol{S}_{ij}^{\mathrm{oo}})\boldsymbol{\Sigma}_i\boldsymbol{M}_j^\top = \boldsymbol{E}_{ij}\boldsymbol{\Sigma}_i\boldsymbol{M}_j^\top, \end{aligned}$$

where $\boldsymbol{E}_{ij} = \boldsymbol{M}_j(\boldsymbol{I}_p - \boldsymbol{\Sigma}_i\boldsymbol{S}_{ij}^{\mathrm{oo}})$, $\boldsymbol{S}_{ij}^{\mathrm{oo}} = \boldsymbol{O}_j^\top(\boldsymbol{O}_j\boldsymbol{\Sigma}_i\boldsymbol{O}_j^\top)^{-1}\boldsymbol{O}_j$.

Since

$$\begin{aligned} -\boldsymbol{\Sigma}_{ij}^{\mathrm{mo}}\boldsymbol{\Sigma}_{ij}^{\mathrm{oo}^{-1}}\boldsymbol{O}_j + \boldsymbol{M}_j &= \boldsymbol{M}_j - \boldsymbol{M}_j\boldsymbol{\Sigma}_i\boldsymbol{O}_j^\top(\boldsymbol{O}_j\boldsymbol{\Sigma}_i\boldsymbol{O}_j^\top)^{-1}\boldsymbol{O}_j \\ &= \boldsymbol{M}_j(\boldsymbol{I}_p - \boldsymbol{\Sigma}_i\boldsymbol{S}_{ij}^{\mathrm{oo}}) \\ &= \boldsymbol{E}_{ij} \end{aligned}$$

and

$$\begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ -\boldsymbol{\Sigma}_{ij}^{\mathrm{mo}}\boldsymbol{\Sigma}_{ij}^{\mathrm{oo}^{-1}} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{O}_j \\ \boldsymbol{M}_j \end{bmatrix} = \begin{bmatrix} \boldsymbol{O}_j \\ \boldsymbol{E}_{ij} \end{bmatrix},$$

it suffices to show that

$$\begin{aligned} \boldsymbol{\mu}_{ij}^{\mathrm{m\cdot o}} &= \boldsymbol{\mu}_{ij}^{\mathrm{m}} + \boldsymbol{\Sigma}_{ij}^{\mathrm{mo}}\boldsymbol{\Sigma}_{ij}^{\mathrm{oo}^{-1}}(\boldsymbol{Y}_j^{\mathrm{o}} - \boldsymbol{\mu}_{ij}^{\mathrm{o}}) \\ &= \boldsymbol{M}_j\boldsymbol{\mu}_i + \boldsymbol{M}_j\boldsymbol{\Sigma}_i\boldsymbol{O}_j^\top(\boldsymbol{O}_j\boldsymbol{\Sigma}_i\boldsymbol{O}_j^\top)^{-1}\boldsymbol{O}_j(\boldsymbol{Y}_j - \boldsymbol{\mu}_i) \\ &= \boldsymbol{M}_j(\boldsymbol{\mu}_i + \boldsymbol{\Sigma}_i\boldsymbol{S}_{ij}^{\mathrm{oo}}(\boldsymbol{Y}_j - \boldsymbol{\mu}_i)). \end{aligned}$$

Hence,

$$\begin{aligned} \Delta_{ij} &= (\boldsymbol{Y}_j^{\mathrm{o}} - \boldsymbol{\mu}_{ij}^{\mathrm{o}})^\top\boldsymbol{\Sigma}_{ij}^{\mathrm{oo}^{-1}}(\boldsymbol{Y}_j^{\mathrm{o}} - \boldsymbol{\mu}_{ij}^{\mathrm{o}}) \\ &\quad + (\boldsymbol{Y}_j^{\mathrm{m}} - \boldsymbol{\mu}_{ij}^{\mathrm{m\cdot o}})^\top\boldsymbol{\Sigma}_{ij}^{\mathrm{mm\cdot o}^{-1}}(\boldsymbol{Y}_j^{\mathrm{m}} - \boldsymbol{\mu}_{ij}^{\mathrm{m\cdot o}}) \\ &= (\boldsymbol{Y}_j - \boldsymbol{\mu}_i)^\top(\boldsymbol{S}_{ij}^{\mathrm{oo}} + \boldsymbol{S}_{ij}^{\mathrm{mm\cdot o}})(\boldsymbol{Y}_j - \boldsymbol{\mu}_i) \\ &= \Delta_{ij}^{\mathrm{o}} + \Delta_{ij}^{\mathrm{m\cdot o}}, \end{aligned}$$

where

$$\begin{aligned} \Delta_{ij}^{\mathrm{o}} &= (\boldsymbol{Y}_j^{\mathrm{o}} - \boldsymbol{\mu}_{ij}^{\mathrm{o}})^\top\boldsymbol{\Sigma}_{ij}^{\mathrm{oo}^{-1}}(\boldsymbol{Y}_j^{\mathrm{o}} - \boldsymbol{\mu}_{ij}^{\mathrm{o}}) \\ &= (\boldsymbol{Y}_j - \boldsymbol{\mu}_i)^\top\boldsymbol{S}_{ij}^{\mathrm{oo}}(\boldsymbol{Y}_j - \boldsymbol{\mu}_i), \end{aligned}$$

$$\begin{aligned} \Delta_{ij}^{\mathrm{m\cdot o}} &= (\boldsymbol{Y}_j^{\mathrm{m}} - \boldsymbol{\mu}_{ij}^{\mathrm{m\cdot o}})^\top\boldsymbol{\Sigma}_{ij}^{\mathrm{mm\cdot o}^{-1}}(\boldsymbol{Y}_j^{\mathrm{m}} - \boldsymbol{\mu}_{ij}^{\mathrm{m\cdot o}}) \\ &= (\boldsymbol{Y}_j - \boldsymbol{\mu}_i)^\top\boldsymbol{S}_{ij}^{\mathrm{mm\cdot o}}(\boldsymbol{Y}_j - \boldsymbol{\mu}_i), \end{aligned}$$

$$\boldsymbol{S}_{ij}^{\mathrm{mm\cdot o}} = \boldsymbol{E}_{ij}^\top(\boldsymbol{E}_{ij}\boldsymbol{\Sigma}_i\boldsymbol{M}_j^\top)^{-1}\boldsymbol{E}_{ij}.$$

Using the fact that $|\boldsymbol{\Sigma}_i| = |\boldsymbol{\Sigma}_{ij}^{\mathrm{oo}}||\boldsymbol{\Sigma}_{ij}^{\mathrm{mm\cdot o}}|$ and above results, we have

$$\begin{aligned} f(\boldsymbol{Y}_j^{\mathrm{m}}|\boldsymbol{Y}_j^{\mathrm{o}}) &= \frac{f(\boldsymbol{Y}_j)}{f(\boldsymbol{Y}_j^{\mathrm{o}})} \\ &= \frac{\sum_{i=1}^{g} w_i\,\phi_{p_j^{\mathrm{o}}}(\boldsymbol{Y}_j^{\mathrm{o}}|\boldsymbol{\mu}_{ij}^{\mathrm{o}}, \boldsymbol{\Sigma}_{ij}^{\mathrm{oo}})\phi_{p-p_j^{\mathrm{o}}}(\boldsymbol{Y}_j^{\mathrm{m}}|\boldsymbol{Y}_j^{\mathrm{o}}, \boldsymbol{\mu}_{ij}^{\mathrm{m\cdot o}}, \boldsymbol{\Sigma}_{ij}^{\mathrm{mm\cdot o}})}{\sum_{i=1}^{g} w_i\,\phi_{p_j^{\mathrm{o}}}(\boldsymbol{Y}_j^{\mathrm{o}}|\boldsymbol{\mu}_{ij}^{\mathrm{o}}, \boldsymbol{\Sigma}_{ij}^{\mathrm{oo}})} \\ &= \sum_{i=1}^{g} w_{ij}^{*}\,\phi_{p-p_j^{\mathrm{o}}}(\boldsymbol{Y}_j^{\mathrm{m}}|\boldsymbol{Y}_j^{\mathrm{o}}, \boldsymbol{\mu}_{ij}^{\mathrm{m\cdot o}}, \boldsymbol{\Sigma}_{ij}^{\mathrm{mm\cdot o}}), \end{aligned}$$

where $w_{ij}^{*} = w_i\phi_{p_j^{\mathrm{o}}}(\boldsymbol{Y}_j^{\mathrm{o}}|\boldsymbol{\mu}_{ij}^{\mathrm{o}}, \boldsymbol{\Sigma}_{ij}^{\mathrm{oo}}) / \sum_{h=1}^{g} w_h\,\phi_{p_j^{\mathrm{o}}}(\boldsymbol{Y}_j^{\mathrm{o}}|\boldsymbol{\mu}_{hj}^{\mathrm{o}}, \boldsymbol{\Sigma}_{hj}^{\mathrm{oo}})$.

Appendix B. Proof of Proposition 3

Letting $\hat{Z}_{ij}^{(k)} = E(Z_{ij}|\boldsymbol{Y}^{\mathrm{o}}, \hat{\boldsymbol{\Theta}}^{(k)})$, $\hat{\boldsymbol{\xi}}_{ij}^{(k)} = E(Z_{ij}\boldsymbol{Y}_j|\boldsymbol{Y}^{\mathrm{o}}, \hat{\boldsymbol{\Theta}}^{(k)})$ and $\hat{\boldsymbol{\Phi}}_{ij}^{(k)} = E(Z_{ij}\boldsymbol{Y}_j\boldsymbol{Y}_j^\top|\boldsymbol{Y}^{\mathrm{o}}, \hat{\boldsymbol{\Theta}}^{(k)})$, we can show that

$$\begin{aligned} \hat{Z}_{ij}^{(k)} &= \Pr(Z_{ij} = 1|\boldsymbol{Y}_j^{\mathrm{o}}, \hat{\boldsymbol{\Theta}}^{(k)}) \\ &= \frac{\hat{w}_i^{(k)}\,\phi_{p_j^{\mathrm{o}}}(\boldsymbol{Y}_j^{\mathrm{o}}|\hat{\boldsymbol{\mu}}_{ij}^{\mathrm{o}(k)}, \hat{\boldsymbol{\Sigma}}_{ij}^{\mathrm{oo}(k)})}{\sum_{h=1}^{g} \hat{w}_h^{(k)}\,\phi_{p_j^{\mathrm{o}}}(\boldsymbol{Y}_j^{\mathrm{o}}|\hat{\boldsymbol{\mu}}_{hj}^{\mathrm{o}(k)}, \hat{\boldsymbol{\Sigma}}_{hj}^{\mathrm{oo}(k)})}, \end{aligned}$$

$$\begin{aligned} \hat{\boldsymbol{\xi}}_{ij}^{(k)} &= \Pr(Z_{ij} = 1|\boldsymbol{Y}^{\mathrm{o}}, \hat{\boldsymbol{\Theta}}^{(k)})E[\boldsymbol{Y}_j|Z_{ij} = 1, \boldsymbol{Y}^{\mathrm{o}}, \hat{\boldsymbol{\Theta}}^{(k)}] \\ &= E(Z_{ij}|\boldsymbol{Y}^{\mathrm{o}}, \hat{\boldsymbol{\Theta}}^{(k)})E(\boldsymbol{Y}_j|Z_{ij} = 1, \boldsymbol{Y}^{\mathrm{o}}, \hat{\boldsymbol{\Theta}}^{(k)}) \\ &= \hat{Z}_{ij}^{(k)}\hat{\boldsymbol{Y}}_{ij}^{(k)}, \end{aligned}$$

and

$$\begin{aligned}\hat{\boldsymbol{\Phi}}_{ij}^{(k)} &= E(Z_{ij}\boldsymbol{Y}_j\boldsymbol{Y}_j^{\top}|\boldsymbol{Y}^{\mathrm{o}}, \hat{\boldsymbol{\Theta}}^{(k)})\\ &= E(Z_{ij}|\boldsymbol{Y}^{\mathrm{o}}, \hat{\boldsymbol{\Theta}}^{(k)})E(\boldsymbol{Y}_j\boldsymbol{Y}_j^{\top}|Z_{ij}=1, \boldsymbol{Y}^{\mathrm{o}}, \hat{\boldsymbol{\Theta}}^{(k)})\\ &= \hat{Z}_{ij}^{(k)}(\hat{\boldsymbol{Y}}_{ij}^{(k)}\hat{\boldsymbol{Y}}_{ij}^{(k)\top} + \mathrm{Cov}(\boldsymbol{Y}_j|Z_{ij}=1, \boldsymbol{Y}^{\mathrm{o}}, \hat{\boldsymbol{\Theta}}^{(k)}))\\ &= \hat{Z}_{ij}^{(k)}(\hat{\boldsymbol{Y}}_{ij}^{(k)}\hat{\boldsymbol{Y}}_{ij}^{(k)\top} + (\boldsymbol{I}_p - \hat{\boldsymbol{\Sigma}}_i^{(k)}\hat{\boldsymbol{S}}_{ij}^{\mathrm{oo}(k)})\hat{\boldsymbol{\Sigma}}_i^{(k)}).\end{aligned}$$

Since $\boldsymbol{Y}_j = \boldsymbol{O}_j^{\top}\boldsymbol{Y}_j^{\mathrm{o}} + \boldsymbol{M}_j^{\top}\boldsymbol{Y}_j^{\mathrm{m}}$ and $\boldsymbol{O}_j^{\top}\boldsymbol{O}_j(\boldsymbol{I}_p - \boldsymbol{\Sigma}_i\boldsymbol{S}_{ij}^{\mathrm{oo}}) = \boldsymbol{0}$, we have

$$\begin{aligned}\hat{\boldsymbol{Y}}_{ij}^{(k)} &= E(\boldsymbol{Y}_j|Z_{ij}=1, \boldsymbol{Y}^{\mathrm{o}}, \hat{\boldsymbol{\Theta}}^{(k)})\\ &= E(\boldsymbol{O}_j^{\top}\boldsymbol{Y}_j^{\mathrm{o}} + \boldsymbol{M}_j^{\top}\boldsymbol{Y}_j^{\mathrm{m}}|Z_{ij}=1, \boldsymbol{Y}^{\mathrm{o}}, \hat{\boldsymbol{\Theta}}^{(k)})\\ &= \boldsymbol{O}_j^{\top}\boldsymbol{Y}_j^{\mathrm{o}} + \boldsymbol{M}_j^{\top}E(\boldsymbol{Y}_j^{\mathrm{m}}|Z_{ij}=1, \boldsymbol{Y}^{\mathrm{o}}, \hat{\boldsymbol{\Theta}}^{(k)})\\ &= \boldsymbol{O}_j^{\top}\boldsymbol{Y}_j^{\mathrm{o}} + \boldsymbol{M}_j^{\top}\hat{\boldsymbol{\mu}}_{ij}^{\mathrm{m\cdot o}(k)}\\ &= \boldsymbol{O}_j^{\top}\boldsymbol{O}_j\boldsymbol{Y}_j + \boldsymbol{M}_j^{\top}\boldsymbol{M}_j(\hat{\boldsymbol{\mu}}_i^{(k)} + \hat{\boldsymbol{\Sigma}}_i^{(k)}\hat{\boldsymbol{S}}_{ij}^{\mathrm{oo}(k)}(\boldsymbol{Y}_j - \hat{\boldsymbol{\mu}}_i^{(k)}))\\ &= \boldsymbol{O}_j^{\top}\boldsymbol{O}_j\boldsymbol{Y}_j + (\boldsymbol{I}_p - \boldsymbol{O}_j^{\top}\boldsymbol{O}_j)(\hat{\boldsymbol{\mu}}_i^{(k)}\\ &\quad + \hat{\boldsymbol{\Sigma}}_i^{(k)}\hat{\boldsymbol{S}}_{ij}^{\mathrm{oo}(k)}(\boldsymbol{Y}_j - \hat{\boldsymbol{\mu}}_i^{(k)}))\\ &= \hat{\boldsymbol{\mu}}_i^{(k)} + \hat{\boldsymbol{\Sigma}}_i^{(k)}\hat{\boldsymbol{S}}_{ij}^{\mathrm{oo}(k)}(\boldsymbol{Y}_j - \hat{\boldsymbol{\mu}}_i^{(k)})\\ &\quad + \boldsymbol{O}_j^{\top}\boldsymbol{O}_j(\boldsymbol{I}_p - \hat{\boldsymbol{\Sigma}}_i^{(k)}\hat{\boldsymbol{S}}_{ij}^{\mathrm{oo}(k)})(\boldsymbol{Y}_j - \hat{\boldsymbol{\mu}}_i^{(k)})\\ &= \hat{\boldsymbol{\mu}}_i^{(k)} + \hat{\boldsymbol{\Sigma}}_i^{(k)}\hat{\boldsymbol{S}}_{ij}^{\mathrm{oo}(k)}(\boldsymbol{Y}_j - \hat{\boldsymbol{\mu}}_i^{(k)}),\end{aligned}$$

and

$$\begin{aligned}&\mathrm{Cov}(\boldsymbol{Y}_j|Z_{ij}=1, \boldsymbol{Y}^{\mathrm{o}}, \hat{\boldsymbol{\Theta}}^{(k)})\\ &\quad = \mathrm{Cov}(\boldsymbol{O}_j^{\top}\boldsymbol{Y}_j^{\mathrm{o}} + \boldsymbol{M}_j^{\top}\boldsymbol{Y}_j^{\mathrm{m}}|Z_{ij}=1, \boldsymbol{Y}^{\mathrm{o}}, \hat{\boldsymbol{\Theta}}^{(k)})\\ &\quad = \boldsymbol{M}_j^{\top}\mathrm{Cov}(\boldsymbol{Y}_j^{\mathrm{m}}|Z_{ij}=1, \boldsymbol{Y}^{\mathrm{o}}, \hat{\boldsymbol{\Theta}}^{(k)})\boldsymbol{M}_j\\ &\quad = \boldsymbol{M}_j^{\top}\hat{\boldsymbol{\Sigma}}_{ij}^{\mathrm{mm\cdot o}(k)}\boldsymbol{M}_j\\ &\quad = \boldsymbol{M}_j^{\top}\hat{\boldsymbol{E}}_{ij}^{(k)}\hat{\boldsymbol{\Sigma}}_i^{(k)}\boldsymbol{M}_j^{\top}\boldsymbol{M}_j\\ &\quad = \boldsymbol{M}_j^{\top}\boldsymbol{M}_j(\boldsymbol{I}_p - \hat{\boldsymbol{\Sigma}}_i^{(k)}\hat{\boldsymbol{S}}_{ij}^{\mathrm{oo}(k)})\hat{\boldsymbol{\Sigma}}_i^{(k)}\boldsymbol{M}_j^{\top}\boldsymbol{M}_j\\ &\quad = (\boldsymbol{I}_p - \boldsymbol{O}_j^{\top}\boldsymbol{O}_j)(\boldsymbol{I}_p - \hat{\boldsymbol{\Sigma}}_i^{(k)}\hat{\boldsymbol{S}}_{ij}^{\mathrm{oo}(k)})\hat{\boldsymbol{\Sigma}}_i^{(k)}(\boldsymbol{I}_p - \boldsymbol{O}_j^{\top}\boldsymbol{O}_j)\\ &\quad = (\boldsymbol{I}_p - \hat{\boldsymbol{\Sigma}}_i^{(k)}\hat{\boldsymbol{S}}_{ij}^{\mathrm{oo}(k)})\hat{\boldsymbol{\Sigma}}_i^{(k)}(\boldsymbol{I}_p - \boldsymbol{O}_j^{\top}\boldsymbol{O}_j)\\ &\quad = (\boldsymbol{I}_p - \hat{\boldsymbol{\Sigma}}_i^{(k)}\hat{\boldsymbol{S}}_{ij}^{\mathrm{oo}(k)})\hat{\boldsymbol{\Sigma}}_i^{(k)}.\end{aligned}$$

Therefore,

$$\begin{aligned}\boldsymbol{\Omega}_{ij}^{(k)} &= \hat{\boldsymbol{\Phi}}_{ij}^{(k)} - \hat{\boldsymbol{\xi}}_{ij}^{(k)}\boldsymbol{\mu}_i^{\top} - \boldsymbol{\mu}_i\hat{\boldsymbol{\xi}}_{ij}^{(k)\top} + \hat{Z}_{ij}^{(k)}\boldsymbol{\mu}_i\boldsymbol{\mu}_i^{\top}\\ &= \hat{Z}_{ij}^{(k)}(\hat{\boldsymbol{Y}}_{ij}^{(k)}\hat{\boldsymbol{Y}}_{ij}^{(k)\top} + (\boldsymbol{I}_p - \hat{\boldsymbol{\Sigma}}_i^{(k)}\hat{\boldsymbol{S}}_{ij}^{\mathrm{oo}(k)})\hat{\boldsymbol{\Sigma}}_i^{(k)})\\ &\quad - 2\hat{Z}_{ij}^{(k)}\hat{\boldsymbol{Y}}_{ij}^{(k)}\boldsymbol{\mu}_i^{\top} + \hat{Z}_{ij}^{(k)}\boldsymbol{\mu}_i\boldsymbol{\mu}_i^{\top}\\ &= \hat{Z}_{ij}^{(k)}((\hat{\boldsymbol{Y}}_{ij}^{(k)} - \boldsymbol{\mu}_i)(\hat{\boldsymbol{Y}}_{ij}^{(k)} - \boldsymbol{\mu}_i)^{\top}\\ &\quad + (\boldsymbol{I}_p - \hat{\boldsymbol{\Sigma}}_i^{(k)}\hat{\boldsymbol{S}}_{ij}^{\mathrm{oo}(k)})\hat{\boldsymbol{\Sigma}}_i^{(k)}).\end{aligned}$$

References

[1] D.M. Titterington, A.F.M. Smith, U.E. Markov, Statistical Analysis of Finite Mixture Distributions, Wiley, New York, 1985.

[2] G.J. McLachlan, K.E. Basford, Mixture Models: Inference and Application to Clustering, Marcel Dekker, New York, 1988.

[3] G.J. McLachlan, D. Peel, Finite Mixture Model, Wiley, New York, 2000.

[4] J.L. Schafer, Analysis of Incomplete Multivariate Data, Chapman & Hall, London, 1997.

[5] C.H. Liu, Efficient ML estimation of multivariate normal distribution from incomplete data, J. Multivariate. Anal. 69 (1999) 206–217.

[6] Z. Ghahramani, M.I. Jordan, Supervised learning from incomplete data via an EM approach, in: J.D. Cowan, G. Tesarro, J. Alspector (Eds.), Advances in Neural Information Processing Systems, vol. 6, Morgan Kaufmann, San Francisco, 1994, pp. 120–127.

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

[8] J. Diebolt, C.P. Robert, Estimation of finite mixture distributions through Bayesian sampling, J. R. Statist. Soc. B. 56 (1994) 363–375.

[9] M.A. Tanner, W.H. Wong, The calculation of posterior distributions by data augmentation (with discussion), J. Am. Statist. Assoc. 82 (1987) 528–550.

[10] M.D. Escobar, M. West, Bayesian density estimation and inference using mixtures, J. Amer. Statist. Assoc. 90 (1995) 577–588.

[11] S. Richardson, P.J. Green, On Bayesian analysis of mixtures with an unknown number of components, J. R. Statist. Soc. B. 59 (1997) 731–792.

[12] Z.H. Zhang, K.L. Chan, Y.M. Wu, C.B. Chen, Learning a multivariate gaussian mixture model with the reversible jump MCMC algorithm, Statist. Comput. 14 (2004) 343–355.

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

[14] M. Stephens, Bayesian analysis of mixture models with an unknown number of components—an alternative to reversible jump methods, Ann. Statist. 28 (2000) 40–74.

[15] S. Fruhwirth-Schnatter, Markov Chain Monte Carlo estimation of classical and dynamic switching and mixture models, J. Amer. Statist. Assoc. 96 (2001) 194–209.

[16] S. Geman, D. Geman, Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, IEEE Trans. Pattern Anal. Mach. Intell. 6 (1984) 721–741.

[17] G. Celeux, M. Hurn, C.P. Robert, Computational and inferential difficulties with mixture posterior distributions, J. Amer. Statist. Assoc. 95 (2000) 957–970.

[18] D.B. Rubin, Inference and missing data, Biometrika 63 (1976) 581–592.

[19] R.J.A. Little, D.B. Rubin, Statistical Analysis with Missing Data, second ed., Wiley, New York, 2002.

[20] K.E. Basford, G.J. McLachlan, Estimation of allocation rates in a cluster analysis text, J. Amer. Statist. Assoc. 80 (1985) 286–293.

[21] W.H. Edwards, H. Lindman, L.J. Savage, Bayesian statistical inference for psychological research, Psycol. Rev. 70 (1963) 193–242.

[22] A.E. Gelfand, A.F.M. Smith, Sampling based approaches to calculate marginal densities, J. Amer. Statist. Assoc. 85 (1990) 398–409.

[23] E. Anderson, The irises of the Gaspé Peninsula, Bull. Am. Iris Soc. 59 (1935) 2–5.

[24] R.A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugenics 7 (1936) 179–188.

[25] N.A. Campbell, R.J. Mahon, A multivariate study of variation in two species of rock crab of genus Leptograpsus, Aust. J. Zool. 22 (1974) 417–425.

[26] M. Stone, Cross-validatory choice and assessment of statistical prediction (with discussion), J. R. Statist. Soc. B. 36 (1974) 111–147.

[27] S.P. Brooks, A. Gelman, General method s for monitoring convergence of iterative simulations, J. Comput. Graph. Statist. 7 (1998) 434–455.

[28] R.A. Johnson, D.W. Wichern, Applied Multivariate Statistical Analysis, fifth ed., Prentice-Hall, Englewood Cliffs, NJ, 2002.

[29] D. Peel, G.J. McLachlan, Robust mixture modeling using the t distribution, Statist. Comput. 10 (2000) 339–348.

[30] S. Shoham, Robust clustering by deterministic agglomeration EM of mixtures of multivariate t-distributions, Pattern Recognition 35 (2002) 1127–1142.

[31] T.I. Lin, J.C. Lee, H.F. Ni, Bayesian Analysis of mixture modelling using the multivariate t distribution, Statist. Comput. 14 (2004) 119–130.

[32] H.X. Wang, Q.B. Zhang, B. Luo, S. Wei, Robust mixture modelling using multivariate t distribution with missing information, Pattern Recognition Lett. 25 (2004) 701–710.

About the Author—TSUNG I. LIN received his B.A. in applied mathematics from National Chung Hsing University, Taiwan in 1993, the M.S. in statistics from National Tsing Hua University, Taiwan in 1997 and Ph.D. in statistics from National Chiao Tung University, Taiwan in 2003. He is at present an assistant professor of National Chung Hsing University. Dr. Lin published papers in statistics and finance. His recent research includes computational statistics, robust mixture modelling and Bayesian analysis.

About the Author—JACK C. LEE received his B.A. in management from National Taiwan University, Taiwan in 1964, the M.S. in economics from the University of Rochester in 1969 and Ph.D. in statistics from the State University of New York at Buffalo in 1972. Dr. Lee is a fellow of the American Statistical Association and an elected member of the International Statistical Institute. He is at present a University Chair Professor of National Chiao Tung University. Dr. Lee published over seventy papers in statistics, engineering and finance. His recent research includes multivariate analysis, speech recognition and finance.

About the Author—HSIU J. Ho received his B.A. in 2003 and M.S. in 2005, both in statistics from Tunghai University, Taiwan. He is currently a research assistant under the supervision of Dr. Yi-Hau Chen in the Institute of Statistical Science at Academia Sinica, Taiwan. His research interests include pattern recognition and feature selection.