Stat Comput (2010) 20: 343–356
DOI 10.1007/s11222-009-9128-9

Robust mixture modeling using multivariate skew t distributions

Tsung-I Lin

Received: 6 May 2008 / Accepted: 20 April 2009 / Published online: 5 May 2009

Abstract This paper presents a robust mixture modeling framework using the multivariate skew t distributions, an extension of the multivariate Student's t family with additional shape parameters to regulate skewness. The proposed model results in a very complicated likelihood. Two variants of Monte Carlo EM algorithms are developed to carry out maximum likelihood estimation of mixture parameters. In addition, we offer a general information-based method for obtaining the asymptotic covariance matrix of maximum likelihood estimates. Some practical issues including the selection of starting values as well as the stopping criterion are also discussed. The proposed methodology is applied to a subset of the Australian Institute of Sport data for illustration.

Keywords MCEM-type algorithms · MSN · MST · Multivariate truncated normal · Multivariate truncated t · Outliers

1 Introduction

Finite mixture models have become more frequently used to provide a natural framework for unobserved heterogeneity in a population. Most importantly, mixture modeling has already been a promising statistical tool for density estimation, supervised classification, unsupervised clustering and a wide variety of other problems due to its analytical tractability. For a comprehensive introduction to fundamental theory of mixture models encountered in practice, see the earlier review paper by Redner and Walker (1984) and monographs by Titterington et al. (1985), McLachlan and Basford (1988), McLachlan and Peel (2000), Frühwirth-Schnatter (2006).

In the past decades, the mixture of multivariate t distributions considered by Peel and McLachlan (2000) provides a natural robust extension of normal mixtures to model the data involving longer than normal tails or discrepant distributions. In many practical problems, however, the robustness of t mixtures may not be ideal in the presence of highly asymmetric observations. Lin et al. (2007a) proposed a novel (univariate) skew t mixture (STMIX) model, which allows for accommodation of both skewness and thick tails for making robust inferences. In particular, it includes the usual normal and t mixtures as well as the skew normal mixture (SNMIX) model proposed by Lin et al. (2007b) as special cases. Despite having sound experimental results using STMIX, its application is still limited to the data with univariate outcomes.

In this paper, we propose a multivariate version of the STMIX (MSTMIX) model, composed of a weighed sum of g-component multivariate skew t (MST) distributions, where g denotes the number of components and is held fixed. The MST distribution we adopt belongs to a new class of multivariate skew elliptical distributions introduced by Sahu et al. (2003). Heuristically, it has a convenient stochastic representation that is useful for conducting maximum likelihood (ML) estimation under a workable complete data framework.

Recently, some attempts have been made to an automated flow cytometry analysis via robust model-based clustering methods. Lo et al. (2008) proposed a flexible approach to identifying cell populations in flow cytometry data based on t-mixture models coupled with a Box-Cox transformation.

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

Pyne et al. (2009) presented a novel FLow analysis with Automated Multivariate Estimation (FLAME) approach to deal with high-dimensional cytometric data using mixtures of a variant of Sahu et al.'s (2003) skew t distribution. Both approaches generalize the use of popular normal mixtures to account for outliers and skewness in cell populations and are capable of separating nonelliptical clusters.

The EM algorithm (Dempster et al. 1977) and its extensions such as the expectation conditional maximization (ECM) algorithm (Meng and Rubin 1993) and the expectation conditional maximization either (ECME) algorithm (Liu and Rubin 1994) have been exploited as useful tools for conducting ML estimation in a variety of mixture model specifications. While carrying out ML estimation in MSTMIX models, we suffer from some difficulties in evaluating the E-step due to the intractability of the target distribution. We develop two extensions the Monte Carlo EM algorithm (Wei and Tanner 1990), say MCEM-type algorithms, with its M-step simplified by using the concepts of ECM and ECME, namely the MCECM and the MCECME algorithms. In particular, the proposed algorithms involve a simple way of generating random samples from the exact conditional distribution and thus preserve the flexibility in gauging Monte Carlo errors.

The rest of the paper is organized as follows. In Sect. 2, we establish notation and outline some main results. Section 3 presents the implementation of MCEM-type algorithms for ML estimation of the MST distribution. Section 4 discusses the specification of MSTMIX model and presents the implementation of MCEM-type algorithms for obtaining the ML estimates of the parameters. In Sect. 5, the application of the proposed methodology is illustrated through a real-world data set. Some concluding remarks are given in Sect. 6.

2 Preliminaries

We begin with a review of the multivariate skew normal (MSN) and MST distributions and a study of some related properties. Next, we briefly introduce the multivariate truncated normal and truncated t distributions. A simple way of generating random samples from the multivariate truncated t distribution is also discussed. Throughout this paper, we shall use the following notation: $\mathbf{M}^{1/2}$ denotes the square root of a symmetric matrix $\mathbf{M}$; $\mathbf{1}_p$ denotes a $p \times 1$ vector of ones; $\boldsymbol{I}_p$ denotes the $p \times p$ identity matrix; Diag$\{\cdot\}$ denotes a diagonal matrix created by extracting the main diagonal elements of a square matrix or the diagonalization of a vector; and diag$\{\cdot\}$ denotes a vector containing the diagonal elements of a square matrix.

2.1 The MSN and MST distributions

As considered in Sahu et al. (2003), a random vector $\boldsymbol{Z}$ is said to follow a p-variate skew normal distribution with location vector $\boldsymbol{\xi}$, scale covariance matrix $\boldsymbol{\Sigma}$ and skewness matrix $\boldsymbol{\Lambda} = \text{Diag}\{\boldsymbol{\lambda}\}$, where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)^{\mathrm{T}}$, if it has density

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

where $\boldsymbol{\Omega} = \boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2$, $\boldsymbol{\Delta} = (\boldsymbol{I}_p + \boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})^{-1} = \boldsymbol{I}_p - \boldsymbol{\Lambda}\boldsymbol{\Omega}^{-1}\boldsymbol{\Lambda}$, and $\phi_p(\cdot \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\Phi_p(\cdot \mid \boldsymbol{\Sigma})$ denote the pdf of $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and cdf of $N_p(\mathbf{0}, \boldsymbol{\Sigma})$, respectively. In usual notation, we shall write $\boldsymbol{Z} \sim SN_p(\boldsymbol{\xi}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda})$ for a random vector with density (1). The subscript will be omitted henceforth if $p = 1$. Note that if $\boldsymbol{\Lambda} = \mathbf{0}$, then the density of $\boldsymbol{Z}$ reduces to the $N_p(\boldsymbol{\xi}, \boldsymbol{\Sigma})$ density.

Following Arellano-Valle et al. (2007), the MSN distribution has a convenient stochastic representation

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

where $\boldsymbol{\zeta}_1$ and $\boldsymbol{\zeta}_2$ are two independent $N_p(\mathbf{0}, \boldsymbol{I}_p)$ random vectors. Essentially, $|\boldsymbol{\zeta}_1|$ follows a p-dimensional standard half-normal (HN) distribution, denoted by $|\boldsymbol{\zeta}_1| \sim HN_p(\mathbf{0}, \boldsymbol{I}_p)$. Sahu et al. (2003) have shown that the mean and covariance of $\boldsymbol{Z}$ are given, respectively, by

$$E(\boldsymbol{Z}) = \boldsymbol{\xi} + \sqrt{\frac{2}{\pi}}\boldsymbol{\lambda} \quad \text{and} \quad \text{cov}(\boldsymbol{Z}) = \boldsymbol{\Sigma} + \left(1 - \frac{2}{\pi}\right)\boldsymbol{\Lambda}^2. \tag{3}$$

Now, we start by defining the MST distribution and its hierarchical formulation and then introduce some further properties. For notational simplicity, we denote $t_p(\cdot \mid \boldsymbol{\xi}, \boldsymbol{\Sigma}, \nu)$ as the p-dimensional multivariate t distribution with location vector $\boldsymbol{\xi}$, scale covariance matrix $\boldsymbol{\Sigma}$ and degrees of freedom (df) ν, and $T_p(\cdot \mid \boldsymbol{\Sigma}; \nu)$ as the cdf of $t_p(\mathbf{0}, \boldsymbol{\Sigma}, \nu)$.

A random vector $\boldsymbol{Y}$ is said to follow a p-variate skew t distribution, denoted by $\boldsymbol{Y} \sim St_p(\boldsymbol{\xi}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}, \nu)$, if it can be represented by

$$\begin{aligned} &\boldsymbol{Y} = \boldsymbol{\xi} + \frac{\boldsymbol{Z}}{\sqrt{\tau}}, \quad \boldsymbol{Z} \sim SN_p(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}), \\ &\tau \sim \Gamma(\nu/2, \nu/2), \ \boldsymbol{Z} \perp \tau, \end{aligned} \tag{4}$$

where $\Gamma(\alpha, \beta)$ stands for a gamma distribution with mean α/β and the symbol '$\perp$' indicates independence. Note that the distribution of $\boldsymbol{Y}$ reduces to $t_p(\boldsymbol{\xi}, \boldsymbol{\Sigma}, \nu)$ as $\boldsymbol{\Lambda} = \mathbf{0}$ and to $SN_p(\boldsymbol{\xi}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda})$ as $\nu \to \infty$.

The following result is an extension property of Azzalini and Capitaino (1999, Lemma 1), which is crucial for evaluating some integrations in this paper.

Proposition 1 *If $\tau \sim \Gamma(\alpha, \beta)$, then for any $\boldsymbol{a} \in \mathbb{R}^p$*

$$E\big(\Phi_p(\boldsymbol{a}\sqrt{\tau}|\boldsymbol{\Delta})\big) = T_p\left(\boldsymbol{a}\sqrt{\frac{\alpha}{\beta}} \,\middle|\, \boldsymbol{\Delta}; 2\alpha\right).$$

Proof See Appendix A. □

From (4), it can be observed that $\boldsymbol{Y}|\tau \sim SN_p(\boldsymbol{\xi}, \boldsymbol{\Sigma}/\tau, \boldsymbol{\Lambda}/\sqrt{\tau})$. By Proposition 1, the density of $\boldsymbol{Y}$ is

$$f(\boldsymbol{y}|\boldsymbol{\xi}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}, \nu) = 2^p t_p(\boldsymbol{y}|\boldsymbol{\xi}, \boldsymbol{\Omega}, \nu) \times T_p\left(\boldsymbol{q}\sqrt{\frac{\nu+p}{U+\nu}}\Big|\boldsymbol{\Delta}; \nu+p\right), \qquad (5)$$

where $\boldsymbol{q} = \boldsymbol{\Lambda}\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\xi})$ and $U = (\boldsymbol{y}-\boldsymbol{\xi})^{\mathrm{T}}\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\xi})$. From (3) and using the law of iterative expectations, we obtain

$$\boldsymbol{\xi} + \sqrt{\frac{\nu}{\pi}}\frac{\Gamma((\nu-1)/2)}{\Gamma(\nu/2)}\boldsymbol{\lambda}$$

and

$$\mathrm{cov}(\boldsymbol{Y}) = \frac{\nu}{\nu-2}\left[\boldsymbol{\Sigma} + \left(1-\frac{2}{\pi}\right)\boldsymbol{\Lambda}^2\right] + \frac{2}{\pi}\left[\frac{\nu}{\nu-2} - \left(\frac{\Gamma((\nu-1)/2)}{\Gamma(\nu/2)}\right)^2\frac{\nu}{2}\right]\boldsymbol{\lambda}\boldsymbol{\lambda}^{\mathrm{T}}. \qquad (6)$$

The mean and covariance matrix do not exist when $\nu \le 1$ and $\nu \le 2$, respectively. Notice that there is an error in the expression of $\mathrm{cov}(\boldsymbol{Y})$ reported in Sahu et al. (2003, p. 137) that leads to $\boldsymbol{\Lambda}^2$ instead of $\boldsymbol{\lambda}\boldsymbol{\lambda}^{\mathrm{T}}$ in (6).

2.2 The multivariate truncated normal and truncated t distributions

Let $TN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbb{A})$ denote a p-variate truncated normal distribution for $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ lying within a left-truncated hyperplane region $\mathbb{A} = \{\boldsymbol{x} = (x_1, \ldots, x_p)^{\mathrm{T}}|x_1 \ge a_1, \ldots, x_p \ge a_p\}$ and let the notation $\prod_{i=1}^p \int_{a_i}^\infty = \int_{a_1}^\infty \cdots \int_{a_p}^\infty$ stand for the abbreviation of multiple integrals. When a p-dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_p)^{\mathrm{T}}$ has the density

$$f(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbb{A}) = \frac{\phi_p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\prod_{r=1}^p \int_{a_r}^\infty \phi_p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})d\boldsymbol{x}} I_{\mathbb{A}}(\boldsymbol{x}), \qquad (7)$$

where $I_{\mathbb{A}}(\boldsymbol{x})$ is the indicator function whose value equals one if $\boldsymbol{x} \in \mathbb{A}$ and zero elsewhere. We shall use the notation $\boldsymbol{X} \sim TN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbb{A})$ if $\boldsymbol{X}$ has density (7).

Let $Tt_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu; \mathbb{A})$ represent a p-variate truncated t distribution for $t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ lying within a left-truncated hyperplane $\mathbb{A}$. Specifically, we say $\boldsymbol{X} \sim Tt_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu; \mathbb{A})$, if its density is given by

$$f(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu; \mathbb{A}) = \frac{t_p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)}{\prod_{r=1}^p \int_{a_r}^\infty t_p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)d\boldsymbol{x}} I_{\mathbb{A}}(\boldsymbol{x}). \qquad (8)$$

We now present a flexible Gibbs procedure to draw random samples from (8). Let $\boldsymbol{L}$ be the Cholesky factor such that $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}$. If $\boldsymbol{Z} = \boldsymbol{L}^{-1}\boldsymbol{X}$, then

$$\boldsymbol{Z} \sim Tt_p(\boldsymbol{\mu}^*, \boldsymbol{I}_p, \nu; \mathbb{B}), \quad \mathbb{B} = \{\boldsymbol{L}\boldsymbol{z} \ge \boldsymbol{a}\},$$

where $\boldsymbol{\mu}^* = \boldsymbol{L}^{-1}\boldsymbol{\mu} = (\mu_1^*, \ldots, \mu_p^*)^{\mathrm{T}}$. Based on the property of the multivariate t distribution concerning its conditional distribution, the full conditionals required for the Gibbs sampler are

$$Z_r|(\boldsymbol{Z}_{-r} = \boldsymbol{z}_{-r}) \sim Tt\left(\mu_r^*, \frac{\nu+\delta_{-r}}{\nu+p-1}, \nu+p-1; \mathbb{B}_r\right), \quad r = 1, \ldots, p, \qquad (9)$$

and each of which follows a univariate truncated t distribution with truncation $\mathbb{B}_r$. Here $\delta_{-r} = (\boldsymbol{z}_{-r} - \boldsymbol{\mu}_{-r}^*)^{\mathrm{T}}(\boldsymbol{z}_{-r} - \boldsymbol{\mu}_{-r}^*)$, $\mathbb{B}_r = \{Z_r \in \mathbb{R} \mid \boldsymbol{L}\boldsymbol{z} \ge \boldsymbol{a}\}$ and $\boldsymbol{z}_{-r}$ is a subvector of $\boldsymbol{z}$ by disposing of the rth entry. Random variates of (9) can be easily generated by using the program of Nadarajah and Kotz (2007), implemented in the R package (R Development Core Team 2008). The Gibbs sampler proceeds as follows:

1. Choose an acceptable initial value $\boldsymbol{X}^{(0)}$. Set $\boldsymbol{Z}^{(0)} = \boldsymbol{L}^{-1}\boldsymbol{X}^{(0)}$ and $k = 1$.
2. Generate $Z_r^{(k)}$, for $r = 1, \ldots, p$, from $f(z_r|z_1^{(k)}, \ldots, z_{r-1}^{(k)}, z_{r+1}^{(k-1)}, \ldots, z_p^{(k-1)})$, the full conditional density of the truncated t distribution defined in (9).
3. Return $\boldsymbol{X}^{(k)} = \boldsymbol{L}\boldsymbol{Z}^{(k)}$ and set $k = k+1$. Go to Step 2.

3 ML estimation of the MST distribution

In this section, we demonstrate how to employ two MCEM-type algorithms for ML estimation of the MST distribution, which can be viewed as a single component MSTMIX model.

3.1 The model and likelihood

From (2) and (4), n independent observations $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ from $St_p(\boldsymbol{\xi}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}, \nu)$ can be hierarchically represented by

$$\begin{aligned} &\boldsymbol{Y}_j \mid (\boldsymbol{\gamma}_j, \tau_j) \sim N_p(\boldsymbol{\xi} + \boldsymbol{\Lambda}\boldsymbol{\gamma}_j, \tau_j^{-1}\boldsymbol{\Sigma}), \\ &\boldsymbol{\gamma}_j \mid \tau_j \sim HN_p(\boldsymbol{0}, \tau_j^{-1}\boldsymbol{I}_p), \quad \tau_j \sim \Gamma(\nu/2, \nu/2). \end{aligned} \qquad (10)$$

From (10), the joint pdf of $\boldsymbol{Y}_j$, $\boldsymbol{\gamma}_j$ and τ_j is given by

$$f(\boldsymbol{y}_j, \boldsymbol{\gamma}_j, \tau_j) = \frac{(\nu/2)^{\nu/2}|\boldsymbol{\Sigma}|^{-1/2}}{\pi^p\Gamma(\nu/2)}\tau_j^{(\nu/2+p)-1} \times \exp\left\{-\frac{\tau_j}{2}\left[(\boldsymbol{\gamma}_j - \boldsymbol{q}_j)^{\mathrm{T}}\boldsymbol{\Delta}^{-1}(\boldsymbol{\gamma}_j - \boldsymbol{q}_j) + U_j + \nu\right]\right\}, \qquad (11)$$

where $\boldsymbol{q}_j = \boldsymbol{\Lambda}\boldsymbol{\Omega}^{-1}(\boldsymbol{y}_j - \boldsymbol{\xi})$ and $U_j = (\boldsymbol{y}_j - \boldsymbol{\xi})^{\mathrm{T}}\boldsymbol{\Omega}^{-1} \times (\boldsymbol{y}_j - \boldsymbol{\xi})$. Integrating out $\boldsymbol{\gamma}_j$ and τ_j, respectively, in (11) yields

$$\begin{aligned} f(\boldsymbol{y}_j, \tau_j) &= \left(\frac{2}{\pi}\right)^{p/2} \frac{(\nu/2)^{\nu/2}|\boldsymbol{\Sigma}|^{-1/2}}{\Gamma(\nu/2)} |\boldsymbol{\Delta}|^{1/2} \tau_j^{(\nu+p)/2-1} \\ &\quad \times \exp\left\{-\frac{\tau_j}{2}(U_j+\nu)\right\} \Phi_p(\boldsymbol{q}\sqrt{\tau_j} \mid \boldsymbol{\Delta}), \end{aligned} \tag{12}$$

and

$$\begin{aligned} f(\boldsymbol{y}_j, \boldsymbol{\gamma}_j) &= \frac{(\nu/2)^{\nu/2}|\boldsymbol{\Sigma}|^{-1/2}\Gamma((\nu+2p)/2)}{\pi^p \Gamma(\nu/2)} \\ &\quad \times \left[\frac{(\boldsymbol{\gamma}_j - \boldsymbol{q}_j)^{\mathrm{T}}\boldsymbol{\Delta}^{-1}(\boldsymbol{\gamma}_j - \boldsymbol{q}_j) + U_j + \nu}{2}\right]^{-(\nu+2p)/2}. \end{aligned} \tag{13}$$

Then, dividing (11) by (12) gives

$$\begin{aligned} f(\boldsymbol{\gamma}_j | \tau_j, \boldsymbol{y}_j) &= \frac{(2\pi)^{-p/2}|\boldsymbol{\Delta}|^{-1/2}\tau_j^{p/2}}{\Phi_p(\boldsymbol{q}_j\sqrt{\tau_j} \mid \boldsymbol{\Delta})} \\ &\quad \times \exp\left\{-\frac{\tau_j}{2}(\boldsymbol{\gamma}_j - \boldsymbol{q}_j)^{\mathrm{T}}\boldsymbol{\Delta}^{-1}(\boldsymbol{\gamma}_j - \boldsymbol{q}_j)\right\}. \end{aligned} \tag{14}$$

It follows from (14) that the conditional distribution of $\boldsymbol{\gamma}_j$ given τ_j and $\boldsymbol{y}_j$ is

$$\boldsymbol{\gamma}_j \mid \tau_j, \boldsymbol{y}_j \sim TN_p(\boldsymbol{q}_j, \tau_j^{-1}\boldsymbol{\Delta}; \mathbb{R}_+^p),$$

where $\mathbb{R}_+^p$ denotes the Euclidean vector space of all p-tuples of positive real numbers.

Additionally, dividing (11) by (13) gives

$$\begin{aligned} f(\tau_j | \boldsymbol{\gamma}_j, \boldsymbol{y}_j) &= \frac{\tau_j^{(\nu+2p)/2-1}}{\Gamma((\nu+2p)/2)} \\ &\quad \times \left[\frac{(\boldsymbol{\gamma}_j - \boldsymbol{q}_j)^{\mathrm{T}}\boldsymbol{\Delta}^{-1}(\boldsymbol{\gamma}_j - \boldsymbol{q}_j) + U_j + \nu}{2}\right]^{(\nu+2p)/2} \\ &\quad \times \exp\left\{-\frac{\tau_j}{2}\left[(\boldsymbol{\gamma}_j - \boldsymbol{q}_j)^{\mathrm{T}}\boldsymbol{\Delta}^{-1}(\boldsymbol{\gamma}_j - \boldsymbol{q}_j) + U_j + \nu\right]\right\} \end{aligned}$$

implying that

$$\begin{aligned} &\tau_j \mid (\boldsymbol{\gamma}_j, \boldsymbol{y}_j) \\ &\quad \sim \Gamma\left(\frac{\nu+2p}{2}, \frac{(\boldsymbol{\gamma}_j - \boldsymbol{q}_j)^{\mathrm{T}}\boldsymbol{\Delta}^{-1}(\boldsymbol{\gamma}_j - \boldsymbol{q}_j) + U_j + \nu}{2}\right). \end{aligned} \tag{15}$$

From (12), applying Proposition 1 leads to

$$\begin{aligned} f(\tau_j | \boldsymbol{y}_j) &= C\tau_j^{(\nu+p)/2-1} \exp\left\{-\frac{\tau_j}{2}(U_j+\nu)\right\} \\ &\quad \times \Phi_p(\boldsymbol{q}_j\sqrt{\tau_j} \mid \boldsymbol{\Delta}), \end{aligned} \tag{16}$$

where

$$C = \frac{(\frac{U_j+\nu}{2})^{(\nu+p)/2}}{\Gamma(\frac{\nu+p}{2})T_p(\boldsymbol{q}_j\sqrt{\frac{\nu+p}{U_j+\nu}} \mid \boldsymbol{\Delta}; \nu+p)} \tag{17}$$

is the normalizing constant. Further, it can be observed from (13) that

$$f(\boldsymbol{\gamma}_j | \boldsymbol{y}_j) \propto \left(1 + \frac{(\boldsymbol{\gamma}_j - \boldsymbol{q}_j)^{\mathrm{T}}\boldsymbol{\Delta}^{-1}(\boldsymbol{\gamma}_j - \boldsymbol{q}_j)}{U_j + \nu}\right)^{-(\nu+2p)/2}, \quad \boldsymbol{\gamma}_j \in \mathbb{R}_+^p,$$

implying that

$$\boldsymbol{\gamma}_j \mid \boldsymbol{y}_j \sim Tt_p\left(\boldsymbol{q}_j, \frac{U_j+\nu}{p+\nu}\boldsymbol{\Delta}, \nu+p;\ \mathbb{R}_+^p\right). \tag{18}$$

In what follows, we denote $c_j(r) = (\nu+p+r)/(U_j+\nu)$, where $r = 0, 2$. As a immediate consequence, we establish the following proposition.

Proposition 2 *From the conditional density* (16), *we have the following*:

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

$$E(\tau_j \mid \boldsymbol{y}_j) = c_j(0)\frac{T_p(\boldsymbol{q}_j\sqrt{c_j(2)} \mid \boldsymbol{\Delta}; \nu+p+2)}{T_p(\boldsymbol{q}_j\sqrt{c_j(0)} \mid \boldsymbol{\Delta}; \nu+p)}.$$

(b) *The conditional expectation of* $\log\tau_j$ *given* $\boldsymbol{Y}_j = \boldsymbol{y}_j$ *is*

$$\begin{aligned} &E(\log\tau_j \mid \boldsymbol{y}_j) \\ &\quad = \mathrm{DG}\left(\frac{\nu+p}{2}\right) - \log\left(\frac{U_j+\nu}{2}\right) \\ &\qquad + c_j(0)\left[\frac{T_p(\boldsymbol{q}_j\sqrt{c_j(2)} \mid \boldsymbol{\Delta}; \nu+p+2)}{T_p(\boldsymbol{q}_j\sqrt{c_j(0)} \mid \boldsymbol{\Delta}; \nu+p)} - 1\right] \\ &\qquad + T_p^{-1}\left(\boldsymbol{q}_j\sqrt{c_j(0)} \mid \boldsymbol{\Delta}; \nu+p\right) \\ &\qquad \times \prod_{r=1}^{p}\int_{-\infty}^{q_{rj}} g_\nu(\boldsymbol{x})t_p\left(\boldsymbol{x} \,\middle|\, \frac{\boldsymbol{\Delta}}{c_j(0)}; \nu+p\right)d\boldsymbol{x}, \end{aligned}$$

where

$$g_\nu(\boldsymbol{x}) = \mathrm{DG}\left(\frac{\nu}{2}+p\right) - \mathrm{DG}\left(\frac{\nu+p}{2}\right) - \frac{p}{p+\nu} + \frac{p(U_j-p)}{(\nu+p)(U_j+\nu)} - \log\left(1+\frac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Delta}^{-1}\boldsymbol{x}}{U_j+\nu}\right) + \frac{(\nu+2p)(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Delta}^{-1}\boldsymbol{x})}{(U_j+\nu)(U_j+\nu+\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Delta}^{-1}\boldsymbol{x})},$$

and q_{rj} *is the* r*th element of* $\boldsymbol{q}_j$ *and* $\mathrm{DG}(x) = \Gamma'(x)/\Gamma(x)$ *is the digamma function.*

Proof See Appendix B. □

3.2 Parameter estimation

We offer two MCEM-type algorithms, which are employed to calculate fully parametric ML estimates based on (10). Let $\boldsymbol{y} = (\boldsymbol{y}_1, \dots, \boldsymbol{y}_n)$, $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_n)$ and $\boldsymbol{\tau} = (\tau_1, \dots, \tau_n)$ for notational convenience. The complete data log-likelihood function of $\boldsymbol{\theta} = (\boldsymbol{\xi}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}, \nu)$ given $(\boldsymbol{y}, \boldsymbol{\gamma}, \boldsymbol{\tau})$, aside from additive constants, is given by

$$\begin{aligned}\ell_c(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{\gamma},\boldsymbol{\tau}) &= \frac{n\nu}{2}\log\left(\frac{\nu}{2}\right) - n\log\Gamma\left(\frac{\nu}{2}\right) - \frac{n}{2}\log|\boldsymbol{\Sigma}| \\ &\quad + \left(\frac{\nu}{2}+p-1\right)\sum_{j=1}^n \log\tau_j \\ &\quad - \sum_{j=1}^n \frac{\tau_j}{2}\Big\{(\boldsymbol{y}_j-\boldsymbol{\xi}_i-\boldsymbol{\Lambda}_i\boldsymbol{\gamma}_j)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_j-\boldsymbol{\xi}_i-\boldsymbol{\Lambda}_i\boldsymbol{\gamma}_j) \\ &\quad + \nu + \boldsymbol{\gamma}_j^{\mathrm{T}}\boldsymbol{\gamma}_j\Big\}. \end{aligned} \tag{19}$$

Hence, the expected value of complete data log-likelihood (19) evaluated with $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^{(k)}$, which we shall denote the Q-function, is

$$Q(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(k)}) = E\big(\ell_c(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{\gamma}, \boldsymbol{\tau}) \mid \boldsymbol{y}, \hat{\boldsymbol{\theta}}^{(k)}\big). \tag{20}$$

At the kth iteration, we let $\hat{\tau}_j^{(k)} = E(\tau_j|\boldsymbol{y}_j, \hat{\boldsymbol{\theta}}^{(k)})$, $\hat{\kappa}_j^{(k)} = E(\log\tau_j|\boldsymbol{y}_j, \hat{\boldsymbol{\theta}}^{(k)})$, $\hat{\boldsymbol{\eta}}_j^{(k)} = E(\tau_j\boldsymbol{\gamma}_j|\boldsymbol{y}_j, \hat{\boldsymbol{\theta}}^{(k)})$ and $\hat{\boldsymbol{\Psi}}_j^{(k)} = E(\tau_j\boldsymbol{\gamma}_j\boldsymbol{\gamma}_j^{\mathrm{T}}|\boldsymbol{y}_j, \hat{\boldsymbol{\theta}}^{(k)})$, which themselves are the necessary conditional expectations involved in (20). Note that (20) cannot be written in a closed form since $\hat{\boldsymbol{\eta}}_j^{(k)}$ and $\hat{\boldsymbol{\Psi}}_j^{(k)}$ are analytically intractable, while $\hat{\tau}_j^{(k)}$ and $\hat{\kappa}_j^{(k)}$ can be exactly computed through Proposition 2.

Using the Monte Carlo method, (20) can be approximated by

$$\hat{Q}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(k)}) = \frac{1}{M}\sum_{m=1}^M \ell_c(\boldsymbol{\theta} \mid \boldsymbol{y}, \hat{\boldsymbol{\gamma}}_{[m]}^{(k)}, \hat{\boldsymbol{\tau}}_{[m]}^{(k)}), \tag{21}$$

where $\hat{\boldsymbol{\gamma}}_{[m]}^{(k)} = (\hat{\boldsymbol{\gamma}}_{1,m}^{(k)}, \dots, \hat{\boldsymbol{\gamma}}_{n,m}^{(k)})$ and $\hat{\boldsymbol{\tau}}_{[m]}^{(k)} = (\hat{\tau}_{1,m}^{(k)}, \dots, \hat{\tau}_{n,m}^{(k)})$, $m = 1, \dots, M$, are a set of independent random samples generated from $f(\boldsymbol{\gamma}_j, \tau_j|\boldsymbol{y}_j)$, $j = 1, \dots, n$, given $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^{(k)}$. A perfect scheme for exact sampling is based on the marginal posterior distribution (15) and the conditional posterior distribution (18). More specifically, we can generate independent samples $(\hat{\boldsymbol{\gamma}}_{j,m}^{(k+1)}, \hat{\tau}_{j,m}^{(k+1)})$, $m = 1, \dots, M$, from

$$\hat{\boldsymbol{\gamma}}_{j,m}^{(k+1)}|\boldsymbol{y}_j \sim Tt_p\left(\hat{\boldsymbol{q}}_j^{(k)}, \frac{\hat{U}_j^{(k)}+\hat{\nu}^{(k)}}{p+\hat{\nu}^{(k)}}\hat{\boldsymbol{\Delta}}^{(k)}, \hat{\nu}^{(k)}+p;\ \mathbb{R}_+^p\right),$$

using the Gibbs sampler described in Sect. 2 and then drawing $\hat{\tau}_{j,m}^{(k+1)}$ given $\hat{\boldsymbol{\gamma}}_{j,m}^{(k+1)}$ and $\boldsymbol{y}_j$ from

$$\Gamma\left(\frac{\hat{\nu}^{(k)}+2p}{2}, \frac{(\hat{\boldsymbol{\gamma}}_{j,m}^{(k+1)}-\hat{\boldsymbol{q}}_j^{(k)})^{\mathrm{T}}\hat{\boldsymbol{\Delta}}^{(k)-1}(\hat{\boldsymbol{\gamma}}_{j,m}^{(k+1)}-\hat{\boldsymbol{q}}_j^{(k)}) + \hat{U}_j^{(k)} + \hat{\nu}^{(k)}}{2}\right),$$

where $\hat{\boldsymbol{q}}_j^{(k)} = \hat{\boldsymbol{\Lambda}}^{(k)}\hat{\boldsymbol{\Omega}}^{(k)-1}(\boldsymbol{y}_j - \hat{\boldsymbol{\xi}}^{(k)})$, $\hat{U}_j^{(k)} = (\boldsymbol{y}_j - \hat{\boldsymbol{\xi}}^{(k)})^{\mathrm{T}}\hat{\boldsymbol{\Omega}}^{(k)-1}(\boldsymbol{y}_j - \hat{\boldsymbol{\xi}}^{(k)})$, $\hat{\boldsymbol{\Omega}}^{(k)} = \hat{\boldsymbol{\Sigma}}^{(k)} + \hat{\boldsymbol{\Lambda}}^{2(k)}$ and $\hat{\boldsymbol{\Delta}}^{(k)} = \boldsymbol{I}_p - \hat{\boldsymbol{\Lambda}}^{(k)}\hat{\boldsymbol{\Omega}}^{(k)-1}\hat{\boldsymbol{\Lambda}}^{(k)}$.

Therefore, we can utilize them to form the Monte Carlo estimates of all foregoing conditional expectations. That is,

$$\begin{aligned} &\hat{\tau}_j^{(k)} \simeq \frac{1}{M}\sum_{m=1}^M \hat{\tau}_{j,m}^{(k)}, \qquad \hat{\kappa}_j^{(k)} \simeq \frac{1}{M}\sum_{m=1}^M \log\hat{\tau}_{j,m}^{(k)}, \\ &\hat{\boldsymbol{\eta}}_j^{(k)} \simeq \frac{1}{M}\sum_{m=1}^M \hat{\tau}_{j,m}^{(k)}\hat{\boldsymbol{\gamma}}_{j,m}^{(k)}, \\ &\hat{\boldsymbol{\Psi}}_j^{(k)} \simeq \frac{1}{M}\sum_{m=1}^M \hat{\tau}_{j,m}^{(k)}\hat{\boldsymbol{\gamma}}_{j,m}^{(k)}\hat{\boldsymbol{\gamma}}_{j,m}^{(k)\mathrm{T}}. \end{aligned} \tag{22}$$

In summary, the implementation of the MCECM algorithm is as follows:

MCE-step: Given $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^{(k)}$, compute Monte Carlo expectations $\hat{\tau}_j^{(k)}$, $\hat{\kappa}_j^{(k)}$, $\hat{\boldsymbol{\eta}}_j^{(k)}$ and $\hat{\boldsymbol{\Psi}}_j^{(k)}$ for $j = 1, \dots, n$ by using (22).

CM-steps:

CM-step 1: Update $\hat{\boldsymbol{\xi}}^{(k)}$ by maximizing (21) over $\boldsymbol{\xi}$, which leads to

$$\hat{\boldsymbol{\xi}}^{(k+1)} = \frac{\sum_{j=1}^{n} \hat{\tau}_j^{(k)} \boldsymbol{y}_j - \hat{\boldsymbol{\Lambda}}^{(k)} \sum_{j=1}^{n} \hat{\boldsymbol{\eta}}_j^{(k)}}{\sum_{j=1}^{n} \hat{\tau}_j^{(k)}}.$$

CM-step 2: Update $\hat{\boldsymbol{\lambda}}^{(k)}$ by maximizing (21) over $\boldsymbol{\lambda}$, which gives

$$\hat{\boldsymbol{\Lambda}}^{(k+1)} = \mathrm{diag}\big\{\big(\hat{\boldsymbol{\Sigma}}^{(k)^{-1}} \odot \hat{\mathbf{B}}_1^{(k)}\big)^{-1}\big(\hat{\boldsymbol{\Sigma}}^{(k)^{-1}} \odot \hat{\mathbf{B}}_2^{(k)}\big)\mathbf{1}_p\big\},$$

where $\hat{\mathbf{B}}_1^{(k)} = \sum_{j=1}^{n} \hat{\boldsymbol{\Psi}}_j^{(k)}, \hat{\mathbf{B}}_2^{(k)} = \sum_{j=1}^{n} \hat{\boldsymbol{\eta}}_j^{(k)}(\boldsymbol{y}_j - \hat{\boldsymbol{\xi}}^{(k+1)})^{\mathrm{T}}$, and the operator '$\odot$' stands for the elementwise product of two matrices with the same dimensions.

CM-step 3: Update $\hat{\boldsymbol{\Sigma}}^{(k)}$ by maximizing (21) over $\boldsymbol{\Sigma}$, which gives

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}^{(k+1)} = \frac{1}{n}\Bigg[&\sum_{j=1}^{n} \hat{\tau}_j^{(k)}(\boldsymbol{y}_j - \hat{\boldsymbol{\xi}}^{(k+1)})(\boldsymbol{y}_j - \hat{\boldsymbol{\xi}}^{(k+1)})^{\mathrm{T}} \\ &+ \hat{\boldsymbol{\Lambda}}^{(k+1)}\hat{\mathbf{B}}_1^{(k)}\hat{\boldsymbol{\Lambda}}^{(k+1)} \\ &- \hat{\boldsymbol{\Lambda}}^{(k+1)}\hat{\mathbf{B}}_2^{(k)} - \hat{\mathbf{B}}_2^{(k)\mathrm{T}}\hat{\boldsymbol{\Lambda}}^{(k+1)}\Bigg].\end{aligned}$$

CM-step 4: Obtain $\hat{\nu}^{(k+1)}$ as the solution of the equation

$$\log\left(\frac{\nu}{2}\right) + 1 - \mathrm{DG}\left(\frac{\nu}{2}\right) + \frac{1}{n}\sum_{j=1}^{n}(\hat{\kappa}_j^{(k)} - \hat{\tau}_j^{(k)}) = 0.$$

Note that the CM-Step 4 consists of a one-dimensional search for the root of ν, which can be easily solved by using the built-in 'uniroot' R routine. To accelerate the convergence of the above algorithm, one can take the advantage of the ECME algorithm, which refers to some CM-steps of the ECM algorithm replaced by steps that maximize a restricted actual log-likelihood function, called the 'CML-step', without sacrificing simplicity. We call such kind of generalization the MCECME algorithm. Accordingly, the above CM-step 4 can be modified into the following CML-step.

CML-step: Obtain $\hat{\nu}^{(k+1)}$ by maximizing the constrained log-likelihood function

$$\begin{aligned}\hat{\nu}^{(k+1)} = \underset{\nu}{\mathrm{argmax}} \sum_{j=1}^{n} \log\Bigg\{ & t_p(\boldsymbol{y}_j | \hat{\boldsymbol{\xi}}^{(k+1)}, \hat{\boldsymbol{\Omega}}^{(k+1)}, \nu) \\ & \times T_p\left(\hat{\boldsymbol{q}}_j^{(k+1)}\sqrt{\frac{\nu + p}{\hat{U}_j^{(k+1)} + \nu}} \,\Bigg|\, \hat{\boldsymbol{\Delta}}^{(k+1)}; \nu + p\right)\Bigg\}.\end{aligned}$$

3.3 Notes on implementation

We address some important issues when employing the above procedures. It is crucial to get good starting values of $\boldsymbol{\theta}$ to achieve convergence swiftly. A simple way of obtaining reasonable starting values is described below.

1. For a given sample $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$, compute sample mean vector $\bar{\boldsymbol{y}}$ and sample covariance matrix $\boldsymbol{S} = [s_{ij}]$.
2. Specify a constant a and set $\hat{\boldsymbol{\Sigma}}^{(0)} = \boldsymbol{S} + (a-1)\mathrm{Diag}\{\boldsymbol{S}\}$, for $0 < a < 1$.
3. Set $\hat{\lambda}_i^{(0)} = \pm\sqrt{(1-a)s_{ii}/(1-2/\pi)}$ for $i = 1, \ldots, p$, whose sign is determined by the sample skewness of the associated variable.
4. Set $\hat{\boldsymbol{\xi}}^{(0)} = \bar{\boldsymbol{y}} - \sqrt{2/\pi}\hat{\boldsymbol{\lambda}}^{(0)}$ and a relatively large initial value for ν, e.g., $\hat{\nu}^{(0)} = 40$.

Generally, there may exist multiple modes in the log-likelihood function. A convenient way to circumvent such limitations is to try several starting values, which can be obtained by specifying various constants a's ranged uniformly in (0, 1). If there exist several modes, one can search for the global optimum by comparing their relative log-likelihood values.

Because Monte Carlo errors involve at the E-step, an increase in likelihood is not guaranteed at each iteration. We follow the suggestions in McCulloch (1994) and Booth and Hobert (1999) to increase the Monte Carlo sample size M properly to gauge Monte Carlo errors. In practice, we choose M as few as possible at the beginning of the algorithm and systematically increase M with the number of iterations. For example, as in McCulloch (1994), one may take $M = 50$ for iterations 1–19, $M = 100$ for iterations 20–39 and $M = 5000$ for iteration 40 and over.

To assess the convergence of the algorithm, an anonymous referee recommended the use of Aitken acceleration-based stopping criterion (see, McLachlan and Krishnan 2008; Chap. 4.9). The Aitken's acceleration at iteration k is defined by

$$a^{(k)} = \frac{l^{(k+1)} - l^{(k)}}{l^{(k)} - l^{(k-1)}},$$

where here for brevity of notation $l^{(k)}$ means the observed log-likelihood evaluated at $\hat{\boldsymbol{\theta}}^{(k)}$. The asymptotic estimate of the log-likelihood at iteration $k+1$ is given by

$$l_\infty^{(k+1)} = l^{(k)} + \frac{1}{1 - a^{(k)}}(l^{(k+1)} - l^{(k)}).$$

As a rule of thumb, Lindsay (1995) suggested the algorithm can be considered to have converged when $|l_\infty^{(k+1)} - l^{(k+1)}| < \epsilon$, where ϵ is the desired tolerance. For later analysis, $\epsilon = 10^{-3}$ is employed.

4 The multivariate skew t mixture model

4.1 Model formulation

Consider n independent random vectors $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$, which are taken from a mixture of MST distributions. The pdf of a g-component MSTMIX model is

$$f(\boldsymbol{y}_j|\boldsymbol{\Theta}) = \sum_{i=1}^{g} w_i \psi(\boldsymbol{y}_j \mid \boldsymbol{\xi}_i, \boldsymbol{\Sigma}_i, \boldsymbol{\Lambda}_i, \nu_i), \tag{23}$$

where $\psi(\boldsymbol{y}_j \mid \boldsymbol{\xi}_i, \boldsymbol{\Sigma}_i, \boldsymbol{\Lambda}_i, \nu_i)$ is the MST density defined in (5) and $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g)$ represents all unknown parameters. Note that the component parameters $\boldsymbol{\theta}_i$ consist of $(w_i, \boldsymbol{\xi}_i, \boldsymbol{\Sigma}_i, \boldsymbol{\Lambda}_i, \nu_i)$, where w_i's are the mixing proportions subject to $\sum_{i=1}^{g} w_i = 1$.

To pose this mixture model to an incomplete data problem, it is conceivable to introduce allocation variables $\boldsymbol{Z}_j = (Z_{1j}, \ldots, Z_{gj})^{\mathrm{T}}$, $j = 1, \ldots, n$, whose values are a set of binary variables with

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

and satisfying $\sum_{i=1}^{g} Z_{ij} = 1$. This implies $\boldsymbol{Z}_j$ follows a multinomial random vector with 1 trial and cell probabilities $w_1, \ldots, w_g$, denoted by $\boldsymbol{Z}_j \sim \mathcal{M}(1; w_1, \ldots, w_g)$. With the inclusion of allocation variables $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$ and latent variables $\boldsymbol{\gamma}_j$'s and τ_j's, a hierarchical representation of (23) is given by

$$\begin{aligned}
&\boldsymbol{Y}_j \mid (\boldsymbol{\gamma}_j, \tau_j, Z_{ij} = 1) \sim N_p(\boldsymbol{\xi}_i + \boldsymbol{\Lambda}_i \boldsymbol{\gamma}_j, \tau_j^{-1} \boldsymbol{\Sigma}_i), \\
&\boldsymbol{\gamma}_j \mid (\tau_j, Z_{ij} = 1) \sim HN_p(\boldsymbol{0}, \tau_j^{-1} \boldsymbol{I}_p), \\
&\tau_j \mid (Z_{ij} = 1) \sim \Gamma(\nu_i/2, \nu_i/2), \\
&\boldsymbol{Z}_j \sim \mathcal{M}(1; w_1, \ldots, w_g).
\end{aligned} \tag{24}$$

4.2 Computational aspects

It follows from the hierarchical structure (24) on the basis of the observed data $\boldsymbol{y}$ and latent data $\boldsymbol{\gamma}$, $\boldsymbol{\tau}$ and $\boldsymbol{Z}$ that the complete data log-likelihood function of $\boldsymbol{\Theta}$, ignoring constants, is

$$\begin{aligned}
&\ell_c(\boldsymbol{\Theta}|\boldsymbol{y}, \boldsymbol{\gamma}, \boldsymbol{\tau}, \boldsymbol{Z}) \\
&\quad = \sum_{i=1}^{g} \sum_{j=1}^{n} Z_{ij} \bigg\{ \log w_i + \frac{\nu_i}{2} \log\left(\frac{\nu_i}{2}\right) - \log \Gamma\left(\frac{\nu_i}{2}\right) \\
&\qquad - \frac{1}{2} \log |\boldsymbol{\Sigma}_i| + \left(\frac{\nu_i}{2} + p - 1\right) \log \tau_j \\
&\qquad - \frac{\tau_j}{2} \big[(\boldsymbol{y}_j - \boldsymbol{\xi}_i - \boldsymbol{\Lambda}_i \boldsymbol{\gamma}_j)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{y}_j - \boldsymbol{\xi}_i - \boldsymbol{\Lambda}_i \boldsymbol{\gamma}_j) \\
&\qquad + \nu_i + \boldsymbol{\gamma}_j^{\mathrm{T}} \boldsymbol{\gamma}_j \big] \bigg\}.
\end{aligned} \tag{25}$$

Let $\hat{z}_{ij}^{(k)}$ denote the posterior probability that $\boldsymbol{y}_j$ belongs to the ith component of the mixture using the current estimates $\hat{\boldsymbol{\Theta}}^{(k)}$ for $\boldsymbol{\Theta}$ ($i = 1, \ldots, g$; $j = 1, \ldots, n$). Specifically, it can be calculated as

$$\begin{aligned}
\hat{z}_{ij}^{(k)} &= \Pr(Z_{ij} = 1 \mid \boldsymbol{y}, \hat{\boldsymbol{\Theta}}^{(k)}) \\
&= \frac{\hat{w}_i^{(k)} \psi(\boldsymbol{y}_j | \hat{\boldsymbol{\xi}}_i^{(k)}, \hat{\boldsymbol{\Sigma}}_i^{(k)}, \hat{\boldsymbol{\Lambda}}_i^{(k)}, \hat{\nu}_i^{(k)})}{f(\boldsymbol{y}_j \mid \hat{\boldsymbol{\Theta}}^{(k)})}.
\end{aligned} \tag{26}$$

To obtain the Q-function with respect to (25), we need to compute the other four conditional expectations, namely $\hat{\tau}_{ij}^{(k)} = E(Z_{ij}\tau_j|\boldsymbol{y}_j, \hat{\boldsymbol{\Theta}}^{(k)})$, $\hat{\kappa}_{ij}^{(k)} = E(Z_{ij} \log \tau_j|\boldsymbol{y}_j, \hat{\boldsymbol{\Theta}}^{(k)})$, $\hat{\boldsymbol{\eta}}_{ij}^{(k)} = E(Z_{ij}\tau_j\boldsymbol{\gamma}_j|\boldsymbol{y}_j, \hat{\boldsymbol{\Theta}}^{(k)})$ and $\hat{\boldsymbol{\Psi}}_{ij}^{(k)} = E(Z_{ij}\tau_j\boldsymbol{\gamma}_j\boldsymbol{\gamma}_j^{\mathrm{T}}| \boldsymbol{y}_j, \hat{\boldsymbol{\Theta}}^{(k)})$, respectively.

A Monte Carlo estimate of the Q-function can be evaluated as

$$\hat{Q}(\boldsymbol{\Theta}|\hat{\boldsymbol{\Theta}}^{(k)}) = \frac{1}{M} \sum_{m=1}^{M} \ell_c(\boldsymbol{\Theta} \mid \boldsymbol{y}, \hat{\boldsymbol{\gamma}}_{[m]}^{*(k)}, \hat{\boldsymbol{\tau}}_{[m]}^{*(k)}, \boldsymbol{Z}),$$

where $\hat{\boldsymbol{\gamma}}_{[m]}^{*(k)} = \{\hat{\boldsymbol{\gamma}}_{ij,m}^{*(k)}, i = 1, \ldots, g; j = 1, \ldots, n\}$ and $\hat{\boldsymbol{\tau}}_{[m]}^{*(k)} = \{\hat{\tau}_{ij,m}^{*(k)}, i = 1, \ldots, g; j = 1, \ldots, n\}$ for $m = 1, \ldots, M$, are a set of independent random samples generated from $f(\boldsymbol{\gamma}_j, \tau_j|\boldsymbol{y}_j, Z_{ij} = 1)$. The sampling procedure can be conveniently implemented through

$$\begin{aligned}
&\hat{\boldsymbol{\gamma}}_{ij,m}^{(k+1)}|(\boldsymbol{y}_j, Z_{ij} = 1) \\
&\quad \sim Tt_p\left(\hat{\boldsymbol{q}}_{ij}^{(k)}, \frac{\hat{U}_{ij}^{(k)} + \hat{\nu}_i^{(k)}}{p + \hat{\nu}_i^{(k)}} \hat{\boldsymbol{\Delta}}_i^{(k)}, \hat{\nu}_i^{(k)} + p;\ \mathbb{R}_+^p\right), \\
&\hat{\tau}_{ij,m}^{(k+1)}|(\hat{\boldsymbol{\gamma}}_{ij,m}^{(k+1)}, \boldsymbol{y}_j, Z_{ij} = 1) \\
&\quad \sim \Gamma\Bigg(\frac{\hat{\nu}_i^{(k)} + 2p}{2}, \\
&\qquad \frac{(\hat{\boldsymbol{\gamma}}_{ij,m}^{(k+1)} - \hat{\boldsymbol{q}}_{ij}^{(k)})^{\mathrm{T}} \hat{\boldsymbol{\Delta}}_i^{(k)-1} (\hat{\boldsymbol{\gamma}}_{ij,m}^{(k+1)} - \hat{\boldsymbol{q}}_{ij}^{(k)}) + \hat{U}_{ij}^{(k)} + \hat{\nu}_i^{(k)}}{2}\Bigg),
\end{aligned}$$

where $\hat{\boldsymbol{q}}_{ij}^{(k)} = \hat{\boldsymbol{\Lambda}}_i^{(k)} \hat{\boldsymbol{\Omega}}_i^{(k)^{-1}} (\boldsymbol{y}_j - \hat{\boldsymbol{\xi}}_i^{(k)})$, $\hat{U}_{ij}^{(k)} = (\boldsymbol{y}_j - \hat{\boldsymbol{\xi}}_i^{(k)})^{\mathrm{T}} \hat{\boldsymbol{\Omega}}_i^{(k)^{-1}} (\boldsymbol{y}_j - \hat{\boldsymbol{\xi}}_i^{(k)})$, $\hat{\boldsymbol{\Omega}}_i^{(k)} = \hat{\boldsymbol{\Sigma}}_i^{(k)} + \hat{\boldsymbol{\Lambda}}_i^{2(k)}$ and $\hat{\boldsymbol{\Delta}}_i^{(k)} = \boldsymbol{I}_p - \hat{\boldsymbol{\Lambda}}_i^{(k)} \hat{\boldsymbol{\Omega}}_i^{(k)-1} \hat{\boldsymbol{\Lambda}}_i^{(k)}$. Therefore, the conditional expectations defined previously can be readily approximated as

$$\begin{aligned}
\hat{\tau}_{ij}^{(k)} &\simeq \frac{\hat{z}_{ij}^{(k)}}{M} \sum_{m=1}^{M} \hat{\tau}_{ij,m}^{(k)}, \\
\hat{\kappa}_{ij}^{(k)} &\simeq \frac{\hat{z}_{ij}^{(k)}}{M} \sum_{m=1}^{M} \log \hat{\tau}_{ij,m}^{(k)}, \\
\hat{\boldsymbol{\eta}}_{ij}^{(k)} &\simeq \frac{\hat{z}_{ij}^{(k)}}{M} \sum_{m=1}^{M} \hat{\tau}_{ij,m}^{(k)} \hat{\boldsymbol{\gamma}}_{ij,m}^{(k)}, \\
\hat{\boldsymbol{\Psi}}_{ij}^{(k)} &\simeq \frac{\hat{z}_{ij}^{(k)}}{M} \sum_{m=1}^{M} \hat{\tau}_{ij,m}^{(k)} \hat{\boldsymbol{\gamma}}_{ij,m}^{(k)} \hat{\boldsymbol{\gamma}}_{ij,m}^{(k)\mathrm{T}}.
\end{aligned} \tag{27}$$

The resulting MCECM algorithm that includes one MCE-step and five CM-steps is described as follows:

MCE-step: Given $\boldsymbol{\Theta} = \hat{\boldsymbol{\Theta}}^{(k)}$, compute $\hat{z}_{ij}^{(k)}$ as in (26), and $\hat{\tau}_{ij}^{(k)}, \hat{\kappa}_{ij}^{(k)}, \hat{\boldsymbol{\eta}}_{ij}^{(k)}$ and $\hat{\boldsymbol{\Psi}}_{ij}^{(k)}$ listed in (27) for each $i = 1, \ldots, g$ and $j = 1, \ldots, n$.

CM-steps:

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

CM-step 2: Calculate $\hat{\boldsymbol{\xi}}_i^{(k+1)} = (\sum_{j=1}^{n} \hat{\tau}_{ij}^{(k)} \boldsymbol{y}_j - \hat{\boldsymbol{\Lambda}}_i^{(k)} \sum_{j=1}^{n} \hat{\boldsymbol{\eta}}_{ij}^{(k)}) / \sum_{j=1}^{n} \hat{\tau}_{ij}^{(k)}$.

CM-step 3: Calculate $\hat{\boldsymbol{\Lambda}}_i^{(k+1)} = \mathrm{diag}\{(\hat{\boldsymbol{\Sigma}}_i^{(k)^{-1}} \odot \hat{\mathbf{B}}_{1i}^{(k)})^{-1} (\hat{\boldsymbol{\Sigma}}_i^{(k)^{-1}} \odot \hat{\mathbf{B}}_{2i}^{(k)}) \mathbf{1}_p\}$, where $\hat{\mathbf{B}}_{1i}^{(k)} = \sum_{j=1}^{n} \hat{\boldsymbol{\Psi}}_{ij}^{(k)}$ and $\hat{\mathbf{B}}_{2i}^{(k)} = \sum_{j=1}^{n} \hat{\boldsymbol{\eta}}_{ij}^{(k)} (\boldsymbol{y}_j - \hat{\boldsymbol{\xi}}_i^{(k+1)})^{\mathrm{T}}$.

CM-step 4: Calculate

$$\hat{\boldsymbol{\Sigma}}_i^{(k+1)} = \frac{1}{\sum_{j=1}^{n} \hat{z}_{ij}^{(k)}} \Bigg[\sum_{j=1}^{n} \hat{\tau}_{ij}^{(k)} (\boldsymbol{y}_j - \hat{\boldsymbol{\xi}}_i^{(k+1)})(\boldsymbol{y}_j - \hat{\boldsymbol{\xi}}_i^{(k+1)})^{\mathrm{T}} + \hat{\boldsymbol{\Lambda}}_i^{(k+1)} \hat{\mathbf{B}}_{1i}^{(k)} \hat{\boldsymbol{\Lambda}}_i^{(k+1)} - \hat{\boldsymbol{\Lambda}}_i^{(k+1)} \hat{\mathbf{B}}_{2i}^{(k)} - \hat{\mathbf{B}}_{2i}^{(k)^{\mathrm{T}}} \hat{\boldsymbol{\Lambda}}_i^{(k+1)} \Bigg].$$

CM-step 5: Obtain $\hat{\nu}_i^{(k+1)}$ as the solution of the equation

$$\log\left(\frac{\nu_i}{2}\right) + 1 - \mathrm{DG}\left(\frac{\nu_i}{2}\right) + \frac{1}{\sum_{j=1}^{n} \hat{z}_{ij}^{(k)}} \sum_{j=1}^{n} (\hat{\kappa}_{ij}^{(k)} - \hat{\tau}_{ij}^{(k)}) = 0.$$

If the dfs are assumed to be identical, i.e., $\nu_1 = \cdots = \nu_g = \nu$, the above CM-step 5 is suggested to switch to the following CML-step:

CML-step: Update $\hat{\nu}^{(k)}$ by

$$\hat{\nu}^{(k+1)} = \underset{\nu}{\mathrm{argmax}} \sum_{j=1}^{n} \log\left(\sum_{i=1}^{g} \hat{w}_i^{(k+1)} \times \psi(\boldsymbol{y}_j \mid \hat{\boldsymbol{\xi}}_i^{(k+1)}, \hat{\boldsymbol{\Sigma}}_i^{(k+1)}, \hat{\boldsymbol{\Lambda}}_i^{(k+1)}, \nu) \right).$$

The MCEM-type algorithms do not provide directly the asymptotic covariance matrix of the estimates. An information-based method is considered for evaluating the standard error estimates. The details are sketched in Appendix C. We also offer a simple way of automatically generating suitable initial values. The method is enumerated in order as follows:

(a) Partition the input data into g components based on a K-means clustering.
(b) Compute the zero-one component membership indicator $\hat{\boldsymbol{z}}_j^{(0)} = \{\hat{z}_{ij}^{(0)}\}_{i=1}^{g}$ according to the a K-means clustering result.
(c) Initialize the mixing proportions as $\hat{w}_i^{(0)} = n^{-1} \sum_{j=1}^{n} \hat{z}_{ij}^{(0)}$.
(d) For each partitioned component, compute the initial values $\hat{\boldsymbol{\xi}}_i^{(0)}$, $\hat{\boldsymbol{\Sigma}}_i^{(0)}$, $\hat{\boldsymbol{\Lambda}}_i^{(0)}$ and $\hat{\nu}_i^{(0)}$ according to the method described in Sect. 3.3.

4.3 Model comparison

To identify the best selected model, the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are two most widely used criteria:

$$\mathrm{AIC} = 2m - 2\ell(\hat{\boldsymbol{\Theta}} \mid \boldsymbol{y}) \quad \text{and} \quad \mathrm{BIC} = m \log n - 2\ell(\hat{\boldsymbol{\Theta}} \mid \boldsymbol{y}),$$

where $\ell(\hat{\boldsymbol{\Theta}} \mid \boldsymbol{y})$ is the maximized log-likelihood and m is the number of free parameters in the model. Both of which can be applied to non-nested and nested models, but do not always lead to the same choice. So there is no clear consensus regarding which criterion is better to use. Keribin (2000) presented a theoretical justification for the consistency of BIC on determining the number of components of a mixture model. Fraley and Raftery (1998, 2002) and more recently McNicholas and Murphy (2008) provided empirical evidence that BIC may be useful in practice. Conceptually, a combined use of AIC and BIC would be of help to screening reasonable candidate models.

5 An illustration

We apply the proposed techniques to a subset of the Australian Institute of Sport (AIS) data, including 13 physical variables on 102 male and 100 female athletes. The AIS data were originally reported by Cook and Weisberg (1994) and subsequently analyzed by Azzalini and Dalla Valle (1996), Azzalini and Capitaino (1999, 2003) and Azzalini (2005), among others. In this example, our attention focuses on a bivariate sample of two variables, (BMI, Bfat), which represent the body mass index and the percentage of body fat, respectively. In particular, this bivariate data set exhibits an apparent bimodal asymmetric mixture pattern with some outlying observations. This motivates us to advocate the use of MSTMIX model as a promising tool to analyze this data set.

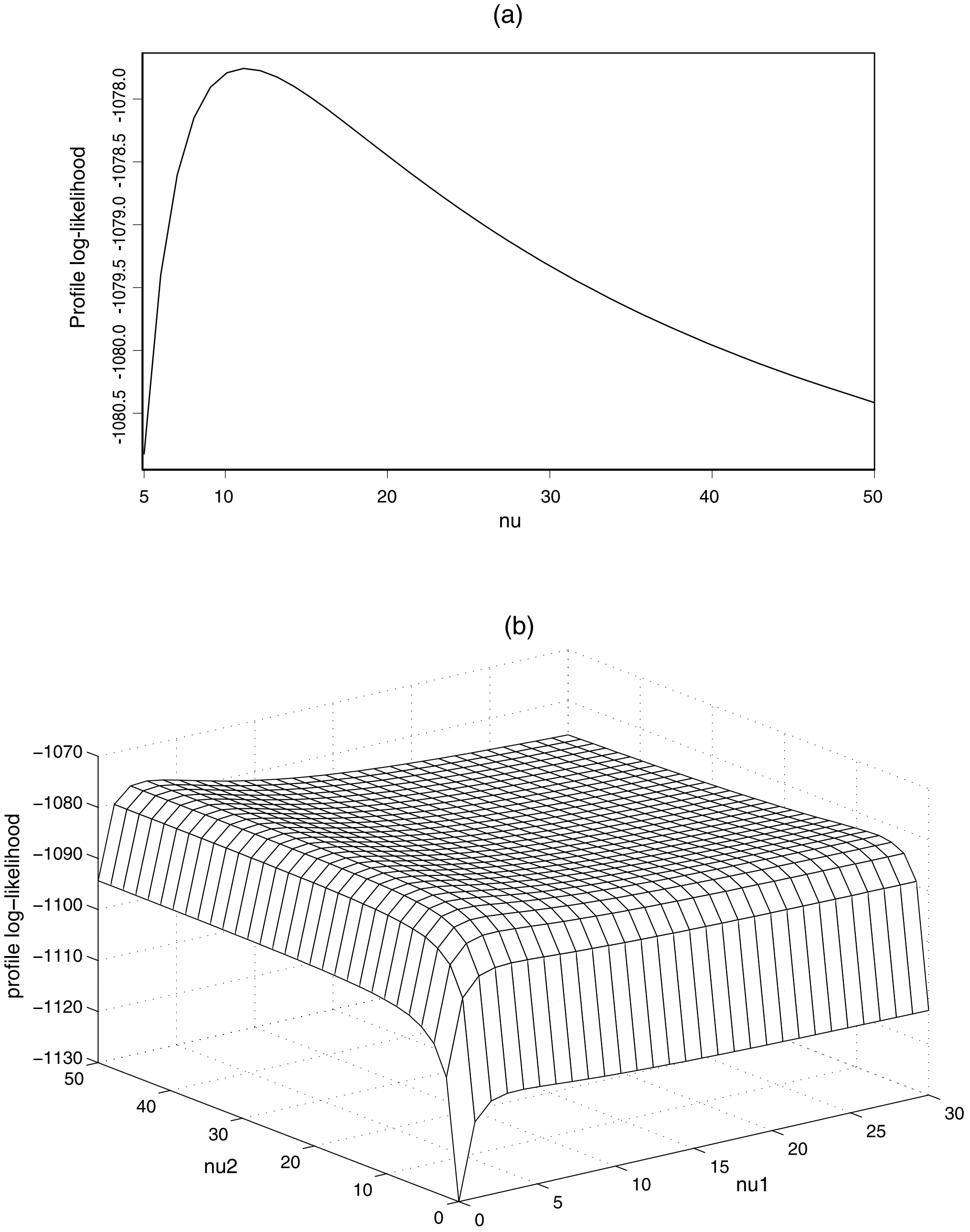

Fig. 1 Plot of the profile log-likelihood for ν_1 and ν_2 for the AIS data set with a two component MSTMIX model with (**a**) equal degrees of freedom ($\nu_1 = \nu_2 = \nu$), (**b**) unequal degrees of freedom ν_1 and ν_2

Specifically, a 2-component MSTMIX model can be written as

$$f(\boldsymbol{y}_j|\boldsymbol{\Theta}) = w\psi(\boldsymbol{y}_j|\boldsymbol{\xi}_1, \boldsymbol{\Sigma}_1, \boldsymbol{\Lambda}_1, \nu_1) + (1-w)\psi(\boldsymbol{y}_j|\boldsymbol{\xi}_2, \boldsymbol{\Sigma}_2, \boldsymbol{\Lambda}_2, \nu_2), \tag{28}$$

where

$$\boldsymbol{\xi}_i = (\xi_{i1}, \xi_{i2})^{\mathrm{T}}, \qquad \boldsymbol{\Sigma}_i = \begin{bmatrix} \sigma_{i,11} & \sigma_{i,12} \\ \sigma_{i,12} & \sigma_{i,22} \end{bmatrix} \quad \text{and}$$

$$\boldsymbol{\Lambda}_i = \begin{bmatrix} \lambda_{i,11} & 0 \\ 0 & \lambda_{i,22} \end{bmatrix},$$

for $i = 1, 2$ and $j = 1, \ldots, 202$.

A preliminary examination of the fitting of model (28) suggests that it is reasonable to assume that the dfs are equal. To explain this, we compute the profile log-likelihood function in the cases of equal and unequal dfs. It can be seen from Fig. 1 that the curve has a significant drop at the peak of the profile log-likelihood in the case of equal df, while it is fairly flat on one side for the case of unequal dfs. Notably,

Table 1 Summary results from fitting various mixture models on the AIS data

Parameter	MVNMIX		MVTMIX		MSNMIX		MSTMIX	
	mle	se	mle	se	mle	se	mle	se
w	0.35	0.04	0.45	0.05	0.45	0.06	0.48	0.07
ξ_{11}	23.11	0.23	23.37	0.26	22.00	2.52	21.68	0.29
ξ_{12}	7.96	0.20	8.32	0.23	5.90	0.14	5.92	0.06
ξ_{21}	22.87	0.39	22.05	0.77	19.31	0.38	19.28	0.35
ξ_{22}	16.49	0.70	17.32	1.20	13.95	1.73	17.40	1.15
$\sigma_{1,11}$	2.88	0.70	3.79	0.92	3.19	2.99	2.82	0.41
$\sigma_{1,12}$	1.56	0.55	2.28	0.63	0.51	0.31	0.55	0.43
$\sigma_{1,22}$	2.12	0.66	3.16	0.77	0.11	0.11	0.12	1.03
$\sigma_{2,11}$	10.98	1.47	5.61	1.12	2.77	1.05	2.40	0.52
$\sigma_{2,12}$	4.96	2.08	6.59	1.86	7.14	2.14	7.01	1.10
$\sigma_{2,22}$	32.07	4.97	24.31	4.20	20.41	8.96	23.62	0.77
$\lambda_{1,11}$	–	–	–	–	1.15	3.22	1.59	0.30
$\lambda_{1,22}$	–	–	–	–	3.41	0.56	3.12	0.11
$\lambda_{2,11}$	–	–	–	–	4.81	0.45	4.17	1.79
$\lambda_{2,22}$	–	–	–	–	4.60	1.90	0.70	6.30
ν	–	–	5.82	1.69	–	–	11.68	5.23
m	11		12		15		16	
$\ell(\hat{\mathbf{\Theta}})$	−1097.79		−1093.59		−1080.65		−1077.58	
AIC	2217.58		2211.17		2191.29		2187.17	
BIC	2253.97		2250.87		2240.92		2240.10	

the ML estimates of ν_1 and ν_2 in the case of unequal dfs are 4.2 and 44.1, respectively, meaning that component 2 corresponds to an assumption of near-normality. Therefore, we restrict ourselves to the situation of equal df, say $\nu_1 = \nu_2 = \nu$. For comparison purposes, mixtures of multivariate (normal; t; skew normal) distributions are also applied to model this data set. Note that model (28) will include multivariate normal mixtures (MVNMIX; $\mathbf{\Lambda}_1 = \mathbf{\Lambda}_2 = \mathbf{0}$, $\nu \to \infty$), multivariate t mixtures (MVTMIX; $\mathbf{\Lambda}_1 = \mathbf{\Lambda}_2 = \mathbf{0}$), and multivariate skew normal mixtures (MSNMIX; $\nu \to \infty$) proposed by Lin (2009) as special cases.

A summary of ML fitting results is given in Table 1, where the values of AIC and BIC are also presented. When comparing fitted objects, the model with the smallest AIC or BIC value is taken to be the best one. In light of these two criteria, the results show that the best-fit to the data is MSTMIX followed by MSNMIX, MVTMIX and MVNMIX. Because the BIC values of the fitted MSNMIX and MSTMIX models are very close, it is suggested to use the likelihood ratio test statistic to judge which of the two models is more appropriate for this data set. The difference in values of −2 times of the log-likelihood between these two fitted models is 6.14, which is highly significant compared to a χ_1^2 distribution, indicating the existence of heavy tails in addition to the effect of skewness. Note that the smaller the df, the better the preference of the t model. It follows from Table 1 that the estimates of df in MVTMIX and MSTMIX models are 5.82 and 11.68, signifying the appropriateness of the use of heavy-tailed t distributions.

To demonstrate the graphical representation for these fitted models, we display the scatter plot superimposed on their corresponding estimated contour densities in Fig. 2. Visually, the MSTMIX model adapts the shape of the scattering pattern more adequately than the other candidate models. This observation manifests the superiority of MSTMIX in the capability of dealing with heterogeneous data involving both asymmetric and heavily tailed behaviors.

6 Concluding remarks

We have proposed a new robust approach to finite mixture models based on MST distributions, called the MSTMIX model, which offers a great deal of flexibility that accommodates asymmetry and heavy tails simultaneously. This model-based tool allows practitioners to analyze heterogeneous multivariate data in a broad variety of considerations. We have described a four-level hierarchical representation for MSTMIX models and developed computationally flexible MCEM-type algorithms for conducting ML estimation. Numerical results show that the MSTMIX model performs reasonably well for the experimental data.

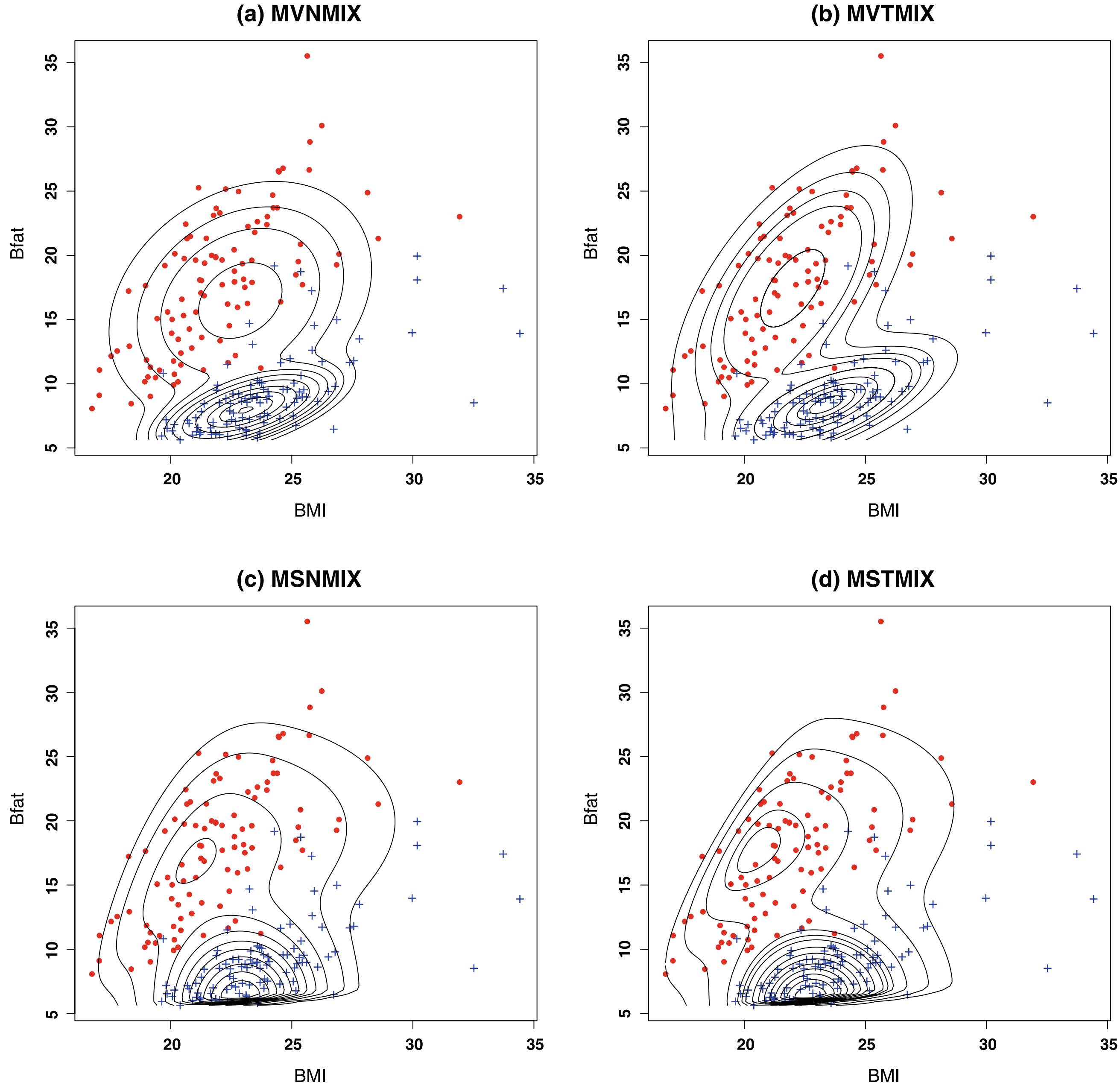

Fig. 2 Scatter plot of BMI and Bfat with superimposed contours of two-component various models in (**a**)–(**d**). The sex is indexed by the symbol • for female and + for male

When data are of high dimension, the presented algorithms can be computationally intensive or encounter some convergence problems. To avoid such problems, one may contemplate generalizing the reported methods in a more parsimonious setting (McNicholas and Murphy 2008). During the last two decades, many researchers (e.g., Diebolt and Robert 1994; Escobar and West 1995; Richardson and Green 1997; Zhang et al. 2004; Dellaportas and Papageorgiou 2006) have attracted the attention to the problem of Bayesian mixture modeling due to the popularity of Markov chain Monte Carlo techniques. Fraley and Raftery (2002) summarized three reasons why statisticians might be interested in adopting a Bayesian approach. Therefore, it is worthwhile to investigate the applicability of a fully Bayesian treatment in this mixture context.

Acknowledgements The author 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. I am also grateful to Ms. Shu-Chun Yang for her skillful assistance in the initial simulation study. This research was supported by the National Science Council of Taiwan (Grant NO. NSC97-2118-M-005-001-MY2).

Appendix A: Proof of Proposition 1

Let $\boldsymbol{V} \sim N_p(\mathbf{0}, \boldsymbol{\Delta})$, where $\boldsymbol{\Delta}$ is a $p \times p$ positive matrix. Assuming that $\boldsymbol{V}$ and τ are independent, then

$$\begin{aligned}
&E_\tau(\Phi_p(\boldsymbol{a}\sqrt{\tau}|\boldsymbol{\Delta}))\\
&\quad = E_\tau\big(P(\boldsymbol{V} \le \boldsymbol{a}\sqrt{\tau}|\tau)\big)\\
&\quad = E_\tau\left(P\left(\frac{\boldsymbol{V}}{\sqrt{\tau\beta/\alpha}} \le \boldsymbol{a}\sqrt{\frac{\alpha}{\beta}}\right)\middle|\tau\right)\\
&\quad = P\left(\boldsymbol{T}^* \le \boldsymbol{a}\sqrt{\frac{\alpha}{\beta}}\right),
\end{aligned}$$

where $\boldsymbol{T}^* \sim t_p(\mathbf{0}, \boldsymbol{\Delta}; 2\alpha)$ is the quoted p-variate t distribution with scale covariance matrix $\boldsymbol{\Delta}$ and degrees of freedom 2α.

Appendix B: Proof of Proposition 2

(a) By Proposition 1, standard calculation of conditional expectation yields

$$\begin{aligned}
&E(\tau_j|\boldsymbol{y}_j)\\
&\quad = \int_0^\infty \tau_j f(\tau_j|\boldsymbol{y}_j)d\tau_j\\
&\quad = C\,\Gamma\left(\frac{\nu+p+2}{2}\right)\left(\frac{U_j+\nu}{2}\right)^{-(\nu+p+2)/2}\\
&\qquad \times \int_0^\infty g\left(\tau_j\middle|\frac{\nu+p+2}{2}, \frac{U_j+\nu}{2}\right)\\
&\qquad \times \Phi_p(\boldsymbol{q}\sqrt{\tau_j}|\boldsymbol{\Delta})d\tau_j\\
&\quad = c_j(0)\frac{T_p(\boldsymbol{q}_j\sqrt{c_j(2)} \mid \boldsymbol{\Delta}; \nu+p+2)}{T_p(\boldsymbol{q}_j\sqrt{c_j(0)} \mid \boldsymbol{\Delta}; \nu+p)},
\end{aligned}$$

where $g(\cdot|\alpha,\beta)$ denotes the density of $\Gamma(\alpha,\beta)$ and C is given in (17).

(b) From (16), it is true that

$$\begin{aligned}
&\frac{d}{d\nu}\int_0^\infty C\,\tau_j^{(\nu+p)/2-1}\\
&\qquad \times \exp\left\{-\frac{\tau_j}{2}(U_j+\nu)\right\}\Phi_p(\boldsymbol{q}_j\sqrt{\tau_j}|\boldsymbol{\Delta})d\tau_j\\
&\quad = 0.
\end{aligned}$$

Note that

$$\begin{aligned}
&\frac{\partial}{\partial\nu}T_p^{-1}(\boldsymbol{q}_j\sqrt{c_j(0)} \mid \boldsymbol{\Delta}; \nu+p)\\
&\quad = -2T_p^{-2}(\boldsymbol{q}_j\sqrt{c_j(0)} \mid \boldsymbol{\Delta}; \nu+p)\\
&\qquad \times \prod_{r=1}^{p}\int_{-\infty}^{q_{rj}}\left(\mathrm{DG}\left(\frac{\nu}{2}+p\right) - \mathrm{DG}\left(\frac{\nu+p}{2}\right)\right.\\
&\qquad - \frac{p}{p+\nu} + \frac{p(U_j-p)}{(\nu+p)(U_j+\nu)}\\
&\qquad - \log\left[1+\frac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Delta}^{-1}\boldsymbol{x}}{U_j+\nu}\right]\\
&\qquad \left. + \frac{(\nu+2p)(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Delta}^{-1}\boldsymbol{x})}{(\nu+U_j)(U_j+\nu+\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Delta}^{-1}\boldsymbol{x})}\right)\\
&\qquad \times t_p\left(\boldsymbol{x}\middle|\frac{\boldsymbol{\Delta}}{c_j(0)}; \nu+p\right)d\boldsymbol{x}.
\end{aligned}$$

By Leibnitz's rule, we can obtain

$$\begin{aligned}
&\log\left(\frac{U_j+\nu}{2}\right) + \left(\frac{\nu+p}{U_j+\nu}\right)\\
&\quad - \mathrm{DG}\left(\frac{\nu+p}{2}\right) + E(\log\tau_j|\boldsymbol{y}_j) - E(\tau_j|\boldsymbol{y}_j)\\
&\quad - T_p^{-1}(\boldsymbol{q}_j\sqrt{c_j(0)} \mid \boldsymbol{\Delta}; \nu+p)\\
&\quad \times \prod_{r=1}^{p}\int_{-\infty}^{q_{rj}} g_\nu(\boldsymbol{x})t_p\left(\boldsymbol{x}\middle|\frac{\boldsymbol{\Delta}}{c_j(0)}; \nu+p\right)d\boldsymbol{x} = 0.
\end{aligned}$$

Together with the result (a) immediately completes the proof.

Appendix C: Estimation of standard errors

We follow the information-based method exploited by Basford et al. (1997) to compute the asymptotic covariance of the ML estimates. The empirical information matrix, according to Meilijson's (1989) formula, is defined as

$$\begin{aligned}
\boldsymbol{I}_e(\boldsymbol{\Theta} \mid \boldsymbol{y}) = {} & \sum_{j=1}^{n}\boldsymbol{s}(\boldsymbol{y}_j \mid \boldsymbol{\Theta})\boldsymbol{s}^{\mathrm{T}}(\boldsymbol{y}_j \mid \boldsymbol{\Theta})\\
& - n^{-1}\boldsymbol{S}(\boldsymbol{y} \mid \boldsymbol{\Theta})\boldsymbol{S}^{\mathrm{T}}(\boldsymbol{y} \mid \boldsymbol{\Theta}), \qquad \text{(C.1)}
\end{aligned}$$

where $\boldsymbol{S}(\boldsymbol{y} \mid \boldsymbol{\Theta}) = \sum_{j=1}^{n}\boldsymbol{s}(\boldsymbol{y}_j \mid \boldsymbol{\Theta})$. It is noted from the result of Louis (1982) that the individual score can be determined as

$$\begin{aligned}
\boldsymbol{s}(\boldsymbol{y}_j \mid \boldsymbol{\Theta}) &= \frac{\partial\log f(\boldsymbol{y}_j|\boldsymbol{\Theta})}{\partial\boldsymbol{\Theta}}\\
&= E\left(\frac{\partial\ell_{cj}(\boldsymbol{\Theta} \mid \boldsymbol{y}_j, \boldsymbol{\gamma}_j, \tau_j)}{\partial\boldsymbol{\Theta}}\middle|\boldsymbol{y}_j, \boldsymbol{\Theta}\right),
\end{aligned}$$

where $\ell_{cj}(\boldsymbol{\Theta}|\boldsymbol{y}_j, \boldsymbol{Z}_j, \boldsymbol{\gamma}_j, \tau_j)$ is the complete data log-likelihood formed from the single observation $\boldsymbol{y}_j$.

Let vech$(\cdot)$ be the matrix operator which stacks only the distinct elements of a symmetric matrix into a single vector. Substituting the ML estimates $\hat{\boldsymbol{\Theta}}$ into $\boldsymbol{\Theta}$, (C.1) is reduced to

$$\boldsymbol{I}_e(\hat{\boldsymbol{\Theta}} \mid \boldsymbol{y}) = \sum_{j=1}^{n} \hat{\boldsymbol{s}}_j \hat{\boldsymbol{s}}_j^{\mathrm{T}}, \tag{C.2}$$

where $\hat{\boldsymbol{s}}_j$ is an individual score vector containing elements of

$$(\hat{s}_{j,w_1}, \ldots, \hat{s}_{j,w_{g-1}}, \hat{\boldsymbol{s}}_{j,\boldsymbol{\xi}_1}^{\mathrm{T}}, \ldots, \hat{\boldsymbol{s}}_{j,\boldsymbol{\xi}_g}^{\mathrm{T}},$$
$$\hat{\boldsymbol{s}}_{j,\boldsymbol{\sigma}_1}^{\mathrm{T}}, \ldots, \hat{\boldsymbol{s}}_{j,\boldsymbol{\sigma}_g}^{\mathrm{T}}, \hat{\boldsymbol{s}}_{j,\boldsymbol{\lambda}_1}^{\mathrm{T}}, \ldots, \hat{\boldsymbol{s}}_{j,\boldsymbol{\lambda}_g}^{\mathrm{T}}, \hat{s}_{j,\nu_1}, \ldots, \hat{s}_{j,\nu_g})^{\mathrm{T}}$$

with $\boldsymbol{\sigma}_i = \mathrm{vech}(\boldsymbol{\Sigma}_i)$. Explicit expressions for the elements of $\hat{\boldsymbol{s}}_j$ are

$$\hat{s}_{j,w_r} = \frac{\hat{z}_{rj}}{\hat{w}_r} - \frac{\hat{z}_{gj}}{\hat{w}_g} \quad (r = 1, \ldots, g-1),$$

$$\hat{\boldsymbol{s}}_{j,\boldsymbol{\xi}_i} = \hat{\boldsymbol{\Sigma}}_i^{-1}\big[\hat{\tau}_{ij}(\boldsymbol{y}_j - \hat{\boldsymbol{\xi}}_i) - \hat{\boldsymbol{\Lambda}}_i \hat{\boldsymbol{\eta}}_{ij}\big],$$

$$\hat{\boldsymbol{s}}_{j,\boldsymbol{\sigma}_i} = \mathrm{vech}\bigg(\frac{1}{2}\Big[2\hat{\mathbf{A}}_{ij} - \mathrm{Diag}\{\hat{\mathbf{A}}_{ij}\}\Big]\bigg),$$

$$\hat{\boldsymbol{s}}_{j,\boldsymbol{\lambda}_i} = \big[\hat{\boldsymbol{\Sigma}}_i^{-1} \odot \hat{\boldsymbol{\eta}}_{ij}(\boldsymbol{y}_j - \hat{\boldsymbol{\xi}}_i)^{\mathrm{T}}\big]\mathbf{1}_p$$
$$\qquad - (\hat{\boldsymbol{\Sigma}}_i^{-1} \odot \hat{\boldsymbol{\Psi}}_{ij})\hat{\boldsymbol{\lambda}}_i,$$

$$\hat{s}_{j,\nu_i} = \frac{\hat{z}_{ij}}{2}\bigg[\log\Big(\frac{\hat{\nu}_i}{2}\Big) + 1 - \mathrm{DG}\Big(\frac{\hat{\nu}_i}{2}\Big)\bigg]$$
$$\qquad + \frac{1}{2}(\hat{\kappa}_{ij} - \hat{\tau}_{ij}),$$

where $\hat{\mathbf{A}}_{ij} = \hat{\boldsymbol{\Sigma}}_i^{-1}[\hat{\tau}_{ij}(\boldsymbol{y}_j - \hat{\boldsymbol{\xi}}_i)(\boldsymbol{y}_j - \hat{\boldsymbol{\xi}}_i)^{\mathrm{T}} - (\boldsymbol{y}_j - \hat{\boldsymbol{\xi}}_i)\hat{\boldsymbol{\eta}}_{ij}^{\mathrm{T}}\hat{\boldsymbol{\Lambda}}_i - \hat{\boldsymbol{\Lambda}}_i\hat{\boldsymbol{\eta}}_{ij}(\boldsymbol{y}_j - \hat{\boldsymbol{\xi}}_i)^{\mathrm{T}} + \hat{\boldsymbol{\Lambda}}_i\hat{\boldsymbol{\Psi}}_{ij}\hat{\boldsymbol{\Lambda}}_i]\hat{\boldsymbol{\Sigma}}_i^{-1} - \hat{\boldsymbol{\Sigma}}_i^{-1}$, $\hat{z}_{ij}$ is $\hat{z}_{ij}^{(k)}$ in (26) with $\hat{\boldsymbol{\Theta}}^{(k)}$ replaced by $\hat{\boldsymbol{\Theta}}$, and $\hat{\tau}_{ij}$, $\hat{\kappa}_{ij}$, $\hat{\boldsymbol{\eta}}_{ij}$ and $\hat{\boldsymbol{\Psi}}_{ij}$ are those Monte Carlo estimates in (27) evaluated at $\hat{\boldsymbol{\Theta}}$. If $\nu_1 = \cdots = \nu_g = \nu$, it follows that $\hat{s}_{j,\nu} = \sum_{i=1}^{g} \hat{s}_{j,\nu_i}$. Standard error estimates of $\hat{\boldsymbol{\Theta}}$ can be obtained by inverting (C.2).

References

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

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

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

Azzalini, A., Capitaino, A.: Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. J. R. Stat. Soc. Ser. B **65**, 367–389 (2003)

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

Basford, K.E., Greenway, D.R., McLachlan, G.J., Peel, D.: Standard errors of fitted means under normal mixture. Comput. Stat. **12**, 1–17 (1997)

Booth, G.J., Hobert, P.J.: Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. J. R. Stat. Soc. Ser. B **61**, 265–285 (1999)

Cook, R.D., Weisberg, S.: An Introduction to Regression Graphics. Wiley, New York (1994)

Dellaportas, P., Papageorgiou, I.: Multivariate mixtures of normals with unknown number of components. Stat. Comput. **16**, 57–68 (2006)

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

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

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

Fraley, C., Raftery, A.E.: How many clusters? Which clustering method? Answers via model-based cluster analysis. Comput. J. **41**, 578–588 (1998)

Fraley, C., Raftery, A.E.: Model-based clustering, discriminant analysis, and density estimation. J. Am. Stat. Assoc. **97**, 611–612 (2002)

Frühwirth-Schnatter, S.: Finite Mixture and Markov Switching Models. Springer, New York (2006)

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

Lin, T.I.: Maximum likelihood estimation for multivariate skew normal mixture models. J. Multivar. Anal. **100**, 257–265 (2009)

Lin, T.I., Lee, J.C., Hsieh, W.J.: Robust mixture modeling using the skew t distribution. Stat. Comput. **17**, 81–92 (2007a)

Lin, T.I., Lee, J.C., Yen, S.Y.: Finite mixture modelling using the skew normal distribution. Stat. Sin. **17**, 909–927 (2007b)

Lindsay, B.: Mixture Models: Theory, Geometry and Applications. Institute of Mathematical Statistics, Hayward (1995)

Liu, C.H., Rubin, D.B.: The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. Biometrika **81**, 633–648 (1994)

Lo, K., Brinkman, R.R., Gottardo, R.: Automated gating of flow cytometry data via robust model-based clustering. Cytometry Part A **73**, 321–332 (2008)

Louis, T.A.: Finding the observed information when using the EM algorithm. J. R. Stat. Soc. Ser. B **44**, 226–232 (1982)

McCulloch, C.E.: Maximum likelihood variance components estimation for binary data. J. Am. Stat. Assoc. **89**, 330–335 (1994)

McLachlan, G.J., Basford, K.E.: Mixture Models: Inference and Application to Clustering. Dekker, New York (1988)

McLachlan, G.J., Krishnan, T.: The EM Algorithm and Extensions, 2nd edn. Wiley, New York (2008)

McLachlan, G.J., Peel, D.: Finite Mixture Models. Wiley, New York (2000)

McNicholas, P.D., Murphy, T.B.: Parsimonious Gaussian mixture models. Stat. Comput. **18**, 285–296 (2008)

Meilijson, I.: A fast improvement to the EM algorithm to its own terms. J. R. Stat. Soc. Ser. B **51**, 127–138 (1989)

Meng, X.L., Rubin, D.B.: Maximum likelihood estimation via the ECM algorithm: a general framework. Biometrika **80**, 267–278 (1993)

Nadarajah, S., Kotz, S.: Programs in R for computing truncated t distributions. Qual. Reliab. Eng. Int. **23**, 273–278 (2007)

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

Pyne, S., Hu, X., Wang, K., Rossin, E., Lin, T.I., Maier, L., Baecher-Allan, C., McLachlan, G.J., Tamayo, P., Hafler, D.A., De Jager, P.L., Mesirov, J.P.: Automated high-dimensional flow cytometric data analysis. Proc. Natl. Acad. Sci. USA (2009). doi:10.1073/pnas.0903028106

Redner, R.A., Walker, H.F.: Mixture densities, maximum likelihood and the EM algorithm. SIAM Rev. **26**, 195–239 (1984)

R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna (2008)

Richardson, S., Green, P.J.: On Bayesian analysis of mixtures with an unknown number of components (with discussion). J. R. Stat. Soc. Ser. B **59**, 731–792 (1997)

Sahu, S.K., Dey, D.K., Branco, M.D.: A new class of multivariate skew distributions with application to Bayesian regression models. Can. J. Stat. **31**, 129–150 (2003)

Titterington, D.M., Smith, A.F.M., Markov, U.E.: Statistical Analysis of Finite Mixture Distributions. Wiley, New York (1985)

Wei, G.C.G., Tanner, M.A.: A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. J. Am. Stat. Assoc. **85**, 699–704 (1990)

Zhang, Z., Chan, K.L., Wu, Y., Cen, C.B.: Learning a multivariate Gaussian mixture model with the reversible Jump MCMC algorithm. Stat. Comput. **14**, 343–355 (2004)