Stat Comput (2007) 17:81–92
DOI 10.1007/s11222-006-9005-8

Robust mixture modeling using the skew t distribution

Tsung I. Lin · Jack C. Lee · Wan J. Hsieh

Received: May 2006 / Accepted: October 2006 / Published online: 27 January 2007

Abstract A finite mixture model using the Student's t distribution has been recognized as a robust extension of normal mixtures. Recently, a mixture of skew normal distributions has been found to be effective in the treatment of heterogeneous data involving asymmetric behaviors across subclasses. In this article, we propose a robust mixture framework based on the skew t distribution to efficiently deal with heavy-tailedness, extra skewness and multimodality in a wide range of settings. Statistical mixture modeling based on normal, Student's t and skew normal distributions can be viewed as special cases of the skew t mixture model. We present analytically simple EM-type algorithms for iteratively computing maximum likelihood estimates. The proposed methodology is illustrated by analyzing a real data example.

Keywords EM-type algorithms · Maximum likelihood · Outlying observations · PX-EM algorithm · Skew t mixtures · Truncated normal

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

J. C. Lee
Graduate Institute of Finance, National Chiao Tung University, Taiwan

W. J. Hsieh
Institute of Statistics, National Chiao Tung University, Taiwan

1 Introduction

The normal mixture (NORMIX) model has been found to be one of the most popular model-based approaches to dealing with data in the presence of *population heterogeneity* in the sense that data intrinsically consist of unlabelled observations, each of which is thought to belong to one of g classes (or components). For a comprehensive list of applications and an abundant literature survey on this area, see Titterington et al. (1985), McLachlan and Basford (1988), and McLachlan and Peel (2000). It is well known that the Student's t distribution involves an additional tuning parameter (the degrees of freedom) that is useful for outlier accommodation. Over the past few years, there has been considerable attention to a robust mixture context based on the Student's t distribution, which we call the t mixture (TMIX) model. Recent developments about TMIX models include Peel and McLachlan (2000), Shoham (2002), Shoham et al. (2003), Lin et al. (2004), and Wang et al. (2004).

While NORMIX and TMIX models have been well recognized as useful in many practical applications, data with varying degrees of extreme skewness among subclasses may not be well modeled. In attempting to appropriately model a set of data arising from a class or several classes with asymmetric observations, Lin et al. (2007) recently introduced a new mixture model with each unseen component following a skew normal distribution (Azzalini, 1985, 1986). A skew normal mixture (SNMIX) model for a continuous random variable Y is of the form

$$Y \sim \sum_{i=1}^{g} w_i f\big(y|\xi_i, \sigma_i^2, \lambda_i\big), \quad \omega_i \geq 0, \quad \sum_{i=1}^{g} \omega_i = 1, \qquad (1)$$

where g is the number of components, w_i's are mixing probabilities and

$$f\left(y|\xi_i,\sigma_i^2,\lambda_i\right)=\frac{2}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{(y-\xi_i)^2}{2\sigma_i^2}\right)\times\int_{-\infty}^{\lambda_i\frac{(y-\xi_i)}{\sigma_i}}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)dx$$

is the skew normal density function with location parameter $\xi_i\in\mathbb{R}$, scale parameter $\sigma_i^2>0$ and skewness parameter $\lambda_i\in\mathbb{R}$. As described in Lin et al. (2007), the SNMIX model (1) can be represented by a normal-truncated normal-multinomial hierarchial structure. Such representation leads to a convenient implementation for maximum likelihood (ML) estimation under a complete-data framework.

Although model (1) offers great flexibility in modeling data with varying asymmetric behaviors, it may suffer from a lack of robustness in the presence of extreme outlying observations. In general, the skewness parameters could be unduly affected by observations that are atypical within components in model (1) being fitted. This motivates us to develop a wider class of mixture distributions to accommodate asymmetry and long tails simultaneously. In this paper, we are devoted to the fitting of mixture of skew t distributions, introduced by Azzalini and Capitaino (2003), allowing for heavy tails in addition to skewness as a natural extension of Lin et al. (2007). With this skew t mixture (STMIX) model approach, the NORMIX, TMIX and SNMIX models can be treated as special cases in this family.

The rest of the paper is organized as follows. Section 2 briefly outlines some preliminary properties of the skew t distribution. Section 3 presents the implementation of ML estimation for fitting the skew t distribution via three simple extensions/modifications of the EM algorithm (Dempster et al., 1977), including the ECM algorithm (Meng and Rubin, 1993), the ECME algorithm (Liu and Rubin, 1994), and the PX-EM algorithm (Liu et al., 1998). Section 4 discusses the STMIX model and presents the implementation of EM-type algorithms for obtaining ML estimates of the parameters. Moreover, we offer a simple way to calculate the information-based standard errors instead of using computationally intensive resampling techniques. In Section 5, the application of the proposed methodology is illustrated through real data of body mass indices measuring from the U.S. male adults. Some concluding remarks are given in Section 6.

2 Preliminaries

For computational ease and notational simplicity, throughout this paper we denote by $\phi(\cdot)$ and $\Phi(\cdot)$ respectively the probability density function (pdf) and the cumulative distribution function (cdf) of the standard normal distribution and denote by $t_\nu(\cdot)$ and $T_\nu(\cdot)$ respectively the pdf and the cdf of the Student's t distribution with degrees of freedom ν. We start by defining the skew t distribution and its hierarchical formulation and then introduce some further properties.

A random variable Y is said to follow the skew t distribution $\mathcal{ST}(\xi,\sigma^2,\lambda,\nu)$ with location parameter $\xi\in\mathbb{R}$, scale parameter $\sigma^2\in(0,\infty)$, skewness parameter $\lambda\in\mathbb{R}$ and degrees of freedom $\nu\in(0,\infty)$ if it has the following representation:

$$Y=\xi+\sigma\frac{Z}{\sqrt{\tau}},\quad Z\sim\mathcal{SN}(\lambda),\quad \tau\sim\Gamma(\nu/2,\nu/2),\quad Z\perp\tau,\tag{2}$$

where $\mathcal{SN}(\lambda)$ stands for the standard skew normal distribution with pdf given by $f(z)=2\phi(z)\Phi(\lambda z)$, $z\in\mathbb{R}$, $\Gamma(\alpha,\beta)$ is the gamma distribution with mean α/β, and the symbol '$\perp$' indicates independence.

The following result, as provided by Azzalini and Capitanio (2003), is useful for evaluating some integrals that we use in the rest of the paper:

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

$$E\Big(\Phi(a\sqrt{\tau})\Big)=T_{2\alpha}\left(a\sqrt{\frac{\alpha}{\beta}}\right).$$

By Proposition 1, integrating τ from the joint density of (Y,τ) will lead to the following marginal density of Y

$$f(y)=\frac{2}{\sigma}t_\nu(\eta)T_{\nu+1}\left(\lambda\eta\sqrt{\frac{\nu+1}{\eta^2+\nu}}\right),\quad \eta=\frac{y-\xi}{\sigma}.\tag{3}$$

Note that as $\nu\to\infty$, $\tau\to1$ with probability 1 and $Y=\xi+\sigma Z$.

As shown by Azzalini (1986, p. 201) and Henze (1986, Theorem 1), a stochastic representation of $Z\sim\mathcal{SN}(\lambda)$ is $Z=\delta_\lambda|U_1|+\sqrt{1-\delta_\lambda^2}U_2$, where $\delta_\lambda=\lambda/\sqrt{1+\lambda^2}$, and U_1 and U_2 are independent $N(0,1)$ random variables. This yields a further hierarchical representation of (2) in the following:

$$Y\mid\gamma,\tau\sim\mathcal{N}\left(\xi+\delta_\lambda\gamma,\frac{1-\delta_\lambda^2}{\tau}\sigma^2\right),$$

$$\gamma\mid\tau\sim\mathcal{TN}\left(0,\frac{\sigma^2}{\tau};(0,\infty)\right),\quad\tau\sim\Gamma(\nu/2,\nu/2),\tag{4}$$

where $\mathcal{TN}(\mu, \sigma^2; (a, b))$ represents the truncated normal distribution with $\mathcal{N}(\mu, \sigma^2)$ lying within the truncated interval (a, b).

From (4), the joint pdf of Y, γ, τ is given by

$$\begin{aligned} f(\gamma, \tau, y) &= \frac{1}{\pi\sqrt{1-\delta_\lambda^2\sigma^2}} \frac{(\nu/2)^{\frac{\nu}{2}}}{\Gamma(\nu/2)} \tau^{\frac{\nu}{2}} \\ &\quad \times \exp\left(-\frac{\tau}{2(1-\delta_\lambda^2)}\eta^2\right) \exp\left(-\frac{\tau}{2}\nu\right) \\ &\quad \times \exp\left(-\frac{\gamma^2\tau}{2(1-\delta_\lambda^2)\sigma^2} + \frac{\gamma\tau}{(1-\delta_\lambda^2)\sigma^2}\delta_\lambda(y-\xi)\right). \end{aligned} \tag{5}$$

Integrating out γ in (5), we get

$$f(\tau, y) = \sqrt{\frac{2}{\pi}} \frac{1}{\sigma} \tau^{\frac{\nu-1}{2}} \frac{(\nu/2)^{\frac{\nu}{2}}}{\Gamma(\nu/2)} \exp\left(-\frac{\tau}{2}(\eta^2+\nu)\right) \Phi\left(\lambda\eta\sqrt{\tau}\right). \tag{6}$$

Dividing (5) by (6) gives

$$\begin{aligned} f(\gamma \mid \tau, y) &= \frac{1}{\sqrt{2\pi}} \frac{\sqrt{\tau}}{\sigma\sqrt{1-\delta_\lambda^2}} \\ &\quad \times \exp\left(-\frac{\tau\left(\gamma-(y-\xi)\delta_\lambda\right)^2}{2(1-\delta^2)\sigma^2}\right) \Phi^{-1}\left(\lambda\eta\sqrt{\tau}\right). \end{aligned} \tag{7}$$

It follows from (7) that the conditional distribution of γ given τ and Y is

$$\gamma \mid \tau, Y \sim \mathcal{TN}\left(\delta_\lambda(y-\xi), \frac{(1-\delta_\lambda^2)\sigma^2}{\tau}; (0, \infty)\right). \tag{8}$$

From (6), applying Proposition 1 yields the conditional density of τ given Y

$$f(\tau \mid y) = b\tau^{(\nu-1)/2} \exp\left(-\frac{\tau}{2}(\eta^2+\nu)\right) \Phi\left(\lambda\eta\sqrt{\tau}\right), \tag{9}$$

where

$$b = \left(\frac{\eta^2+\nu}{2}\right)^{(\nu+1)/2} \left\{\Gamma\left(\frac{\nu+1}{2}\right) T_{\nu+1}\left(\lambda\eta\sqrt{\frac{\nu+1}{\eta^2+\nu}}\right)\right\}^{-1} \tag{10}$$

is the normalizing constant.

Proposition 2. *Given the hierarchical representation* (4), *we have the following:*

(a) *The conditional expectation of* τ *given* $Y = y$ *is*

$$E(\tau|y) = \left(\frac{\nu+1}{\eta^2+\nu}\right) \frac{T_{\nu+3}\left(M\sqrt{\frac{\nu+3}{\nu+1}}\right)}{T_{\nu+1}(M)},$$

where $M = \lambda\eta\sqrt{\frac{\nu+1}{\eta^2+\nu}}$.

(b) *The conditional expectation of* $\gamma\tau$ *given* $Y = y$ *is*

$$\begin{aligned} &E(\gamma\tau|y) \\ &= \delta_\lambda(y-\xi)E(\tau|y) + \frac{\sqrt{1-\delta_\lambda^2}}{\pi f_Y(y)} \left(\frac{\eta^2}{\nu(1-\delta_\lambda^2)} + 1\right)^{-(\nu/2+1)}. \end{aligned}$$

(c) *The conditional expectation of* $\gamma^2\tau$ *given* $Y = y$ *is*

$$\begin{aligned} &E(\gamma^2\tau|y) \\ &\quad = \delta_\lambda^2(y-\xi)^2E(\tau|y) + \left(1-\delta_\lambda^2\right)\sigma^2 \\ &\qquad + \frac{\delta_\lambda(y-\xi)\sqrt{1-\delta_\lambda^2}}{\pi f_Y(y)} \left(\frac{\eta^2}{\nu(1-\delta_\lambda^2)} + 1\right)^{-(\nu/2+1)}. \end{aligned}$$

(d) *The conditional expectation of* $\log(\tau)$ *given* $Y = y$ *is*

$$\begin{aligned} &E\left(\log(\tau)|y\right) \\ &\quad = \mathrm{DG}\left(\frac{\nu+1}{2}\right) - \log\left(\frac{\eta^2+\nu}{2}\right) \\ &\qquad + \frac{\nu+1}{\eta^2+\nu}\left(\frac{T_{\nu+3}\left(M\sqrt{\frac{\nu+3}{\nu+1}}\right)}{T_{\nu+1}(M)} - 1\right) \\ &\qquad + \frac{\lambda\eta(\eta^2-1)}{\sqrt{(\nu+1)(\nu+\eta^2)^3}} \frac{t_{\nu+1}(M)}{T_{\nu+1}(M)} \\ &\qquad + \frac{1}{T_{\nu+1}(M)} \int_{-\infty}^{M} g_\nu(x) t_{\nu+1}(x) dx, \end{aligned}$$

where

$$\begin{aligned} g_\nu(x) &= \mathrm{DG}\left(\frac{\nu+2}{2}\right) - \mathrm{DG}\left(\frac{\nu+1}{2}\right) - \log\left(1 + \frac{x^2}{\nu+1}\right) \\ &\quad + \frac{(\nu+1)x^2 - \nu - 1}{(\nu+1)(\nu+1+x^2)}, \end{aligned} \tag{11}$$

and $\mathrm{DG}(x) = \Gamma'(x)/\Gamma(x)$ *is the digamma function.*

Proof: See Appendix. □

3 ML estimation of the skew t distribution

In this section, we demonstrate how to employ the EM-type algorithms for ML estimation of the skew t distribution, which can be viewed as a single component skew t mixture model that we shall discuss in the next section. From the representation (4), n independent observations from $\mathcal{ST}(\xi, \sigma^2, \lambda, \tau)$ can be expressed by

$$
\begin{aligned}
Y_j \mid \gamma_j, \tau_j &\overset{\text{ind}}{\sim} \mathcal{N}\left(\xi + \delta_\lambda \gamma_j, \frac{1-\delta_\lambda^2}{\tau_j}\sigma^2\right),\\
\gamma_j \mid \tau_j &\overset{\text{ind}}{\sim} \mathcal{TN}\left(0, \frac{\sigma^2}{\tau_j}; (0,\infty)\right),\\
\tau_j &\overset{\text{ind}}{\sim} \Gamma(\nu/2, \nu/2) \qquad (j = 1, \ldots, n).
\end{aligned}
$$

Letting $\boldsymbol{y} = (y_1, \ldots, y_n)$, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)$ and $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n)$, the complete data log-likelihood function of $\boldsymbol{\theta} = (\xi, \sigma^2, \lambda, \nu)$ given $(\boldsymbol{y}, \boldsymbol{\gamma}, \boldsymbol{\tau})$, ignoring additive constant terms, is given by

$$
\begin{aligned}
&\ell_c(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{\gamma}, \boldsymbol{\tau})\\
&\quad = -\frac{\nu}{2}\sum_{i=1}^{n}\tau_j - \sum_{j=1}^{n}\left(\frac{\eta_j^2\tau_j}{2(1-\delta_\lambda^2)}\right) + \sum_{j=1}^{n}\left(\frac{\delta_\lambda\eta_j\gamma_j\tau_j}{(1-\delta_\lambda^2)\sigma}\right)\\
&\qquad - \sum_{j=1}^{n}\left(\frac{{\gamma_j}^2\tau_j}{2(1-\delta_\lambda^2)\sigma^2}\right) - n\log\sigma^2 - \frac{n}{2}\log(1-\delta_\lambda^2)\\
&\qquad + \frac{n\nu}{2}\log\left(\frac{\nu}{2}\right) - n\log\Gamma\left(\frac{\nu}{2}\right) + \frac{\nu}{2}\sum_{j=1}^{n}\log\tau_j,
\end{aligned}
$$

where $\eta_j = (y_j - \xi)/\sigma$.

By Proposition 2, given the current estimate $\hat{\boldsymbol{\theta}}^{(k)} = (\hat{\xi}^{(k)}, \hat{\sigma}^{2^{(k)}}, \hat{\lambda}^{(k)}, \hat{\nu}^{(k)})$ at the kth iteration, the expected complete data log-likelihood function or the Q-function as asserted in Dempster et al. (1977) is

$$
\begin{aligned}
&Q(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(k)})\\
&\quad = -\frac{\nu}{2}\sum_{j=1}^{n}\hat{s}_{1j}^{(k)} - \sum_{j=1}^{n}\left(\frac{\eta_j^2\hat{s}_{1j}^{(k)}}{2(1-\delta_\lambda^2)}\right) + \sum_{j=1}^{n}\left(\frac{\delta_\lambda\eta_j\hat{s}_{2j}^{(k)}}{(1-\delta_\lambda^2)\sigma}\right)\\
&\qquad - \sum_{j=1}^{n}\left(\frac{\hat{s}_{3j}^{(k)}}{2(1-\delta_\lambda^2)\sigma^2}\right) - n\log\sigma^2 - \frac{n}{2}\log(1-\delta_\lambda^2)\\
&\qquad + \frac{n\nu}{2}\log\left(\frac{\nu}{2}\right) - n\log\Gamma\left(\frac{\nu}{2}\right) + \frac{\nu}{2}\sum_{j=1}^{n}\hat{s}_{4j}^{(k)}, \qquad (12)
\end{aligned}
$$

where

$$
\begin{aligned}
\hat{s}_{1j}^{(k)} &= E\big(\tau_j|y_j, \hat{\boldsymbol{\theta}}^{(k)}\big)\\
&= \left(\frac{\hat{\nu}^{(k)}+1}{\hat{\eta}_j^{2^{(k)}} + \hat{\nu}^{(k)}}\right)\frac{T_{\hat{\nu}^{(k)}+3}\left(\hat{M}_j^{(k)}\sqrt{\frac{\hat{\nu}^{(k)}+3}{\hat{\nu}^{(k)}+1}}\right)}{T_{\hat{\nu}^{(k)}+1}\big(\hat{M}_j^{(k)}\big)}, \qquad (13)
\end{aligned}
$$

$$
\begin{aligned}
\hat{s}_{2j}^{(k)} &= E\big(\gamma_j\tau_j|y_j, \hat{\boldsymbol{\theta}}^{(k)}\big) = \hat{\delta}_\lambda^{(k)}(y_j - \hat{\xi}^{(k)})\hat{s}_{1j}^{(k)}\\
&\quad + \frac{\sqrt{1-\hat{\delta}_\lambda^{2^{(k)}}}}{\pi\hat{f}_{Y_j}^{(k)}(y_j)}\left(\frac{\hat{\eta}_j^{2^{(k)}}}{\hat{\nu}^{(k)}\big(1-\delta_\lambda^{2^{(k)}}\big)} + 1\right)^{-(\hat{\nu}^{(k)}/2+1)}, \qquad (14)
\end{aligned}
$$

$$
\begin{aligned}
\hat{s}_{3j}^{(k)} &= E\big(\gamma_j^2\tau_j|y_j, \hat{\boldsymbol{\theta}}^{(k)}\big)\\
&= \hat{\delta}_\lambda^{2^{(k)}}\big(y_j - \hat{\xi}^{(k)}\big)^2\hat{s}_{1j}^{(k)} + \big(1-\hat{\delta}_\lambda^{2^{(k)}}\big)\hat{\sigma}^{2^{(k)}}\\
&\quad + \frac{\hat{\delta}_\lambda^{(k)}\big(y_j - \hat{\xi}^{(k)}\big)\sqrt{1-\hat{\delta}_\lambda^{2^{(k)}}}}{\pi\hat{f}_{Y_j}^{(k)}(y_j)}\\
&\quad \times\left(\frac{\hat{\eta}_j^{2^{(k)}}}{\hat{\nu}^{(k)}\big(1-\hat{\delta}_\lambda^{2^{(k)}}\big)} + 1\right)^{-(\hat{\nu}^{(k)}/2+1)}, \qquad (15)
\end{aligned}
$$

and

$$
\begin{aligned}
\hat{s}_{4j}^{(k)} &= E(\log\tau_j|y_j, \hat{\boldsymbol{\theta}}^{(k)})\\
&= \mathrm{DG}\left(\frac{\hat{\nu}^{(k)}+1}{2}\right)\\
&\quad + \frac{\hat{\nu}^{(k)}+1}{\hat{\eta}_j^{2^{(k)}} + \hat{\nu}^{(k)}}\left(\frac{T_{\hat{\nu}^{(k)}+3}\left(\hat{M}_j^{(k)}\sqrt{\frac{\hat{\nu}^{(k)}+3}{\hat{\nu}^{(k)}+1}}\right)}{T_{\hat{\nu}^{(k)}+1}\big(\hat{M}_j^{(k)}\big)} - 1\right)\\
&\quad - \log\left(\frac{\hat{\eta}_j^{2^{(k)}} + \hat{\nu}^{(k)}}{2}\right) + \frac{\hat{\lambda}^{(k)}\hat{\eta}_j^{(k)}\big(\hat{\eta}_j^{2^{(k)}} - 1\big)}{\sqrt{(\hat{\nu}^{(k)}+1)\big(\hat{\nu}^{(k)} + \hat{\eta}_j^{2^{(k)}}\big)^3}}\\
&\quad \times\left(\frac{t_{\hat{\nu}^{(k)}+1}(\hat{M}_j^{(k)})}{T_{\hat{\nu}^{(k)}+1}(\hat{M}_j^{(k)})}\right) + \frac{1}{T_{\hat{\nu}^{(k)}+1}\big(\hat{M}_j^{(k)}\big)}\\
&\quad \times\int_{-\infty}^{\hat{M}_j^{(k)}} g_{\hat{\nu}^{(k)}}(x)t_{\hat{\nu}^{(k)}+1}(x)dx, \qquad (16)
\end{aligned}
$$

with

$$
\hat{\eta}_j^{(k)} = \frac{y_j - \hat{\xi}^{(k)}}{\hat{\sigma}^{(k)}}, \quad \hat{\delta}_\lambda^{(k)} = \frac{\hat{\lambda}^{(k)}}{\sqrt{1+\hat{\lambda}^{2^{(k)}}}},
$$

$$
\hat{M}_j^{(k)} = \hat{\lambda}^{(k)}\hat{\eta}_j^{(k)}\sqrt{\frac{\hat{\nu}^{(k)}+1}{\hat{\eta}_j^{2^{(k)}} + \hat{\nu}^{(k)}}},
$$

$$
\hat{f}_{Y_j}^{(k)}(y_j) = \frac{2}{\hat{\sigma}^{(k)}}t_{\hat{\nu}^{(k)}}\big(\hat{\eta}_j^{(k)}\big)T_{\hat{\nu}^{(k)}+1}\big(\hat{M}_j^{(k)}\big).
$$

Our proposed ECM algorithm for the skew t distribution consists of one E-step and four CM-steps as described below:

E-step: Given $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^{(k)}$, compute $\hat{s}_{1j}^{(k)}$, $\hat{s}_{2j}^{(k)}$, $\hat{s}_{3j}^{(k)}$ and $\hat{s}_{4j}^{(k)}$ in Eqs. (13)–(16) for $j = 1, \ldots, n$.

CM-step 1: Update $\hat{\xi}^{(k)}$ by maximizing (12) over ξ, which leads to

$$\hat{\xi}^{(k+1)} = \frac{\sum_{j=1}^{n} \hat{s}_{1j}^{(k)} y_j - \hat{\delta}_\lambda^{(k)} \sum_{j=1}^{n} \hat{s}_{2j}^{(k)}}{\sum_{j=1}^{n} \hat{s}_{1j}^{(k)}}.$$

CM-step 2: Fix $\xi = \hat{\xi}^{(k+1)}$, update $\hat{\sigma}^{2^{(k)}}$ by maximizing (12) over σ^2, which gives

$$\hat{\sigma}^{2^{(k+1)}} = \frac{\sum_{j=1}^{n} \left(\hat{s}_{1j}^{(k)} \left(y_j - \hat{\xi}^{(k+1)}\right)^2 - 2\hat{\delta}_\lambda^{(k)} \hat{s}_{2j}^{(k)} \left(y_j - \hat{\xi}^{(k+1)}\right) + \hat{s}_{3j}^{(k)}\right)}{2n\left(1 - \hat{\delta}_\lambda^{2^{(k)}}\right)}.$$

CM-step 3: Fix $\xi = \hat{\xi}^{(k+1)}$ and $\sigma^2 = \hat{\sigma}^{2^{(k+1)}}$, obtain $\hat{\lambda}^{(k+1)}$ as the solution of

$$n\delta_\lambda\left(1 - \delta_\lambda^2\right) - \delta_\lambda \left(\sum_{j=1}^{n} \frac{\hat{s}_{1j}^{(k)}\left(y_j - \hat{\xi}^{(k+1)}\right)^2}{\hat{\sigma}^{2(k+1)}} + \sum_{j=1}^{n} \frac{\hat{s}_{3j}^{(k)}}{\hat{\sigma}^{2(k+1)}}\right) + \left(1 + \delta_\lambda^2\right) \sum_{j=1}^{n} \frac{\hat{s}_{2j}^{(k)}\left(y_j - \hat{\xi}^{(k+1)}\right)}{\hat{\sigma}^{2(k+1)}} = 0.$$

CM-step 4: Fix $\xi = \hat{\xi}^{(k+1)}$, $\sigma^2 = \hat{\sigma}^{2^{(k+1)}}$ and $\lambda = \hat{\lambda}^{(k+1)}$, obtain $\hat{\nu}^{(k+1)}$ as the solution of

$$\log\left(\frac{\nu}{2}\right) + 1 - \mathrm{DG}\left(\frac{\nu}{2}\right) + \frac{1}{n} \sum_{j=1}^{n} \left(\hat{s}_{4j}^{(k)} - \hat{s}_{1j}^{(k)}\right) = 0.$$

Note that the CM-Steps 3 and 4 require a one-dimensional search for the root of λ and ν, respectively, which can be easily achieved by using the 'uniroot' function built in R. As pointed out by Liu and Rubin (1994), the one-dimensional search involved in CM-steps 3 and 4 can be very slow in some situations. To circumvent this obstacle, one may use a more efficient ECME algorithm, which refers to some conditional maximization (CM) steps of the ECM algorithm replaced by steps that maximize a restricted actual log-likelihood function, called the 'CML-step'. With the simple modifications, the ECME algorithm for fitting the skew t distribution can be performed by changing CM-steps 3 and 4 of the above ECM algorithm to a single CML-step as follows:

CML-step: Update $\lambda^{(k)}$ and $\nu^{(k)}$ by optimizing the following constrained actual log-likelihood function:

$$\left(\lambda^{(k+1)}, \nu^{(k+1)}\right) = \underset{\lambda,\nu}{\mathrm{argmax}} \sum_{j=1}^{n} \log\left\{t_\nu\left(\eta_j^{(k+1)}\right) T_{\nu+1}\left(\lambda \eta_j^{(k+1)} \sqrt{\frac{\nu+1}{\eta_j^{2^{(k+1)}} + \nu}}\right)\right\}.$$

Another strategy for speeding up convergence rate is to use the PX-EM algorithm of Liu et al. (1998), which can be simply done by replacing the CM-steps 2 and 4 in the previous ECM algorithm with the following PX.CM steps:

PX.CM-step 2:

$$\hat{\sigma}^{2^{(k+1)}} = \frac{\sum_{j=1}^{n} \left(\hat{s}_{1j}^{(k)} \left(y_j - \hat{\xi}^{(k+1)}\right)^2 - 2\hat{\delta}_\lambda^{(k)} \hat{s}_{2j}^{(k)} \left(y_j - \hat{\xi}^{(k+1)}\right) + \hat{s}_{3j}^{(k)}\right)}{2\left(1 - \hat{\delta}_\lambda^{2^{(k)}}\right) \sum_{j=1}^{n} \hat{s}_{1j}^{(k)}}.$$

PX.CM-step 4:

$$\log\left(\frac{n\nu}{2\sum_{j=1}^{n} \hat{s}_{1j}^{(k)}}\right) - \mathrm{DG}\left(\frac{\nu}{2}\right) + \frac{1}{n} \sum_{j=1}^{n} \hat{s}_{4j}^{(k)} = 0.$$

Assuming that the regularity conditions in Zacks (1971, Chap. 5) hold, these guarantee that asymptotic covariance of the ML estimates can be estimated by the inverse of the *observed information matrix*, $I_o(\hat{\boldsymbol{\theta}}; \boldsymbol{y}) = \sum_{j=1}^{n} \hat{\boldsymbol{u}}_j \hat{\boldsymbol{u}}_j^{\mathrm{T}}$, where

$$\hat{\boldsymbol{u}}_j = \left.\frac{\partial \log f(y_j)}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}}$$

is the score vector corresponding to the single observation y_j evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$.

Expressions for the elements of the score vector with respect to ξ, σ^2, λ and ν are given by

$$\frac{\partial \log f(y_j)}{\partial \xi} = \frac{\eta_j}{\sigma}\left(\frac{\nu+1}{\eta_j^2 + \nu}\right) - \frac{\lambda\nu}{\sigma} \sqrt{\frac{\nu+1}{(\eta_j^2 + \nu)^3}} \frac{t_{\nu+1}\left(M_j\right)}{T_{\nu+1}\left(M_j\right)},$$

$$\frac{\partial \log f(y_j)}{\partial \sigma} = \frac{\nu}{\sigma}\left(\frac{\eta_j^2 - 1}{\eta_j^2 + \nu}\right) - \frac{\lambda\nu\eta_j}{\sigma} \sqrt{\frac{\nu+1}{(\eta_j^2 + \nu)^3}} \frac{t_{\nu+1}\left(M_j\right)}{T_{\nu+1}\left(M_j\right)},$$

$$\frac{\partial \log f(y_j)}{\partial \lambda} = \eta_j \sqrt{\frac{\nu+1}{\eta_j^2 + \nu}} \frac{t_{\nu+1}\left(M_j\right)}{T_{\nu+1}\left(M_j\right)},$$

$$\begin{aligned}\frac{\partial \log f(y_j)}{\partial \nu} = \frac{1}{2}\Bigg\{&\mathrm{DG}\Big(\frac{\nu+1}{2}\Big) - \mathrm{DG}\Big(\frac{\nu}{2}\Big)\\ &- \log\Big(1+\frac{\eta_j^2}{\nu}\Big) + \frac{\eta_j^2-1}{\eta_j^2+\nu}\\ &+ \frac{\lambda\eta_j(\eta_j^2-1)}{\sqrt{(\nu+1)(\eta_j^2+\nu)^3}}\frac{t_{\nu+1}(M_j)}{T_{\nu+1}(M_j)}\\ &+ \frac{1}{T_{\nu+1}(M_j)}\int_{-\infty}^{M_j} g_\nu(x)t_{\nu+1}(x)dx\Bigg\},\end{aligned}$$

where $\eta_j = \sigma^{-1}(y_j - \xi)$ and $M_j = \lambda\eta_j\sqrt{\frac{\nu+1}{\eta_j^2+\nu}}$.

4 The skew *t* mixture model

We consider a g-component mixture model ($g > 1$) in which a set of random sample $Y_1, \dots, Y_n$ arises from a mixture of skew t distributions, given by

$$\begin{aligned}\psi(y_j \mid \boldsymbol{\Theta}) &= \sum_{i=1}^{g} w_i f(y_j \mid \xi_i, \sigma_i^2, \lambda_i, \nu_i),\\ w_i &\geq 0, \quad \sum_{i=1}^{g} w_i = 1, \qquad (17)\end{aligned}$$

where $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_g)$ with $\boldsymbol{\theta}_i = (w_i, \xi_i, \sigma_i^2, \lambda_i, \nu_i)$ denoting the unknown parameters of component i, and w_i's being the mixing probabilities. In the mixture context, it naturally provides a flexible framework for modeling *unobserved population heterogeneity* in the collected sample. With this phenomenon, for each Y_j, it is convenient to introduce a set of zero-one indicator variables $\boldsymbol{Z}_j = (Z_{1j}, \dots, Z_{gj})^{\mathrm{T}}$ ($j = 1, \dots, n$) to describe the unknown population membership. Each $\boldsymbol{Z}_j$ is a multinomial random vector with 1 trial and cell probabilities $w_1, \dots, w_g$, denoted as $\boldsymbol{Z}_j \sim \mathcal{M}(1;\ w_1, \dots, w_g)$. Note that the rth element $z_{rj} = 1$ if Y_j arises from the component r. With the inclusion of indicator variables $Z_j's$, a hierarchical representation of (17) is given by

$$\begin{aligned}Y_j \mid \gamma_j, \tau_j, z_{ij} = 1 &\sim \mathcal{N}\Big(\xi_i + \delta_{\lambda_i}\gamma_j, \frac{1-\delta_{\lambda_i}{}^2}{\tau_j}\sigma_i^2\Big),\\ \gamma_j \mid \tau_j, z_{ij} = 1 &\sim \mathcal{TN}\Big(0, \frac{\sigma_i^2}{\tau_j}; (0,\infty)\Big),\\ \tau_j \mid z_{ij} = 1 &\sim \Gamma(\nu_i/2, \nu_i/2),\\ \boldsymbol{Z}_j &\sim \mathcal{M}(1; w_1, w_2, \dots, w_g). \qquad (18)\end{aligned}$$

It follows from the hierarchical structure (18) on the basis of the observed data $\boldsymbol{y}$ and latent variables $\boldsymbol{\gamma}$, $\boldsymbol{\tau}$ and $\boldsymbol{Z}_j$'s that the complete data log-likelihood function of $\boldsymbol{\Theta}$, ignoring constants, is

$$\begin{aligned}\ell_c(\boldsymbol{\Theta}) = \sum_{j=1}^{n}\sum_{i=1}^{g} Z_{ij}\Bigg\{&\log w_i - \frac{\nu_i\tau_j}{2} - \frac{\tau_j\eta_{ij}^2}{2(1-\delta_{\lambda_i}^2)}\\ &+ \frac{\delta_{\lambda_i}\eta_{ij}\gamma_j\tau_j}{(1-\delta_{\lambda_i}^2)\sigma_i} - \frac{\gamma_j{}^2\tau_j}{2(1-\delta_{\lambda_i}^2)\sigma_i^2}\\ &- \frac{1}{2}\log(1-\delta_{\lambda_i}^2) - \log\sigma_i^2 + \frac{\nu_i}{2}\log\frac{\nu_i}{2}\\ &- \log\Gamma\Big(\frac{\nu_i}{2}\Big) + \frac{\nu_i}{2}\log\tau_j\Bigg\}, \qquad (19)\end{aligned}$$

where $\eta_{ij} = (y_j - \xi_i)/\sigma_i$ and $\delta_{\lambda_i} = \lambda_i/\sqrt{1+\lambda_i^2}$.

Let $\hat{z}_{ij}^{(k)} = E(Z_{ij}|y_j, \hat{\boldsymbol{\Theta}}^{(k)})$, $\hat{s}_{1ij}^{(k)} = E(Z_{ij}\tau_j|y_j, \hat{\boldsymbol{\Theta}}^{(k)})$, $\hat{s}_{2ij}^{(k)} = E(Z_{ij}\gamma_j\tau_j|y_j, \hat{\boldsymbol{\Theta}}^{(k)})$ $\hat{s}_{3ij}^{(k)} = E(Z_{ij}\gamma_j^2\tau_j|y_j, \hat{\boldsymbol{\Theta}}^{(k)})$ and $\hat{s}_{4ij}^{(k)} = E(Z_{ij}\log\tau_j|y_j, \hat{\boldsymbol{\Theta}}^{(k)})$ be the necessary conditional expectations of (19) for obtaining the Q-function at the kth iteration. These expressions, for $i = 1, \dots, g$ and $j = 1, \dots, n$, are given by

$$\hat{z}_{ij}^{(k)} = \frac{w_i^{(k)} f(y_j \mid \xi_i^{(k)}, \sigma_i^{2^{(k)}}, \lambda_i^{(k)}, \nu_i^{(k)})}{\psi(y_j|\hat{\boldsymbol{\Theta}}^{(k)})}, \qquad (20)$$

$$\hat{s}_{1ij}^{(k)} = \hat{z}_{ij}^{(k)}\left(\frac{\hat{\nu}_i^{(k)}+1}{\hat{\eta}_{ij}^{2^{(k)}}+\hat{\nu}_i^{(k)}}\right)\frac{T_{\hat{\nu}_i^{(k)}+3}\left(\hat{M}_{ij}^{(k)}\sqrt{\frac{\hat{\nu}_i^{(k)}+3}{\hat{\nu}_i^{(k)}+1}}\right)}{T_{\hat{\nu}_i^{(k)}+1}(\hat{M}_{ij}^{(k)})}, \qquad (21)$$

$$\begin{aligned}\hat{s}_{2ij}^{(k)} = {}&\hat{\delta}_{\lambda_i}^{(k)}(y_j - \hat{\xi}_i^{(k)})\hat{s}_{1ij}^{(k)}\\ &+\hat{z}_{ij}^{(k)}\left\{\frac{\sqrt{1-\hat{\delta}_{\lambda_i}^{2^{(k)}}}}{\pi\psi(y_j|\hat{\boldsymbol{\Theta}}^{(k)})}\left(\frac{\hat{\eta}_{ij}^{2^{(k)}}}{\hat{\nu}_i^{(k)}(1-\hat{\delta}_{\lambda_i}^{2^{(k)}})}+1\right)^{-(\hat{\nu}_i^{(k)}/2+1)}\right\}, \qquad (22)\end{aligned}$$

$$\begin{aligned}\hat{s}_{3ij}^{(k)} = {}&\hat{\delta}_{\lambda_i}^{2^{(k)}}(y_j - \hat{\xi}_i^{(k)})^2\hat{s}_{1ij}^{(k)} + \hat{z}_{ij}^{(k)}\Bigg\{(1-\hat{\delta}_{\lambda_i}^{2^{(k)}})\hat{\sigma}_i^{2^{(k)}}\\ &+\frac{\hat{\delta}_{\lambda_i}^{(k)}(y_j - \hat{\xi}_i^{(k)})\sqrt{1-\hat{\delta}_{\lambda_i}^{2^{(k)}}}}{\pi\psi(y_j|\hat{\boldsymbol{\Theta}}^{(k)})}\\ &\times\left(\frac{\hat{\eta}_{ij}^{2^{(k)}}}{\hat{\nu}_i^{(k)}(1-\hat{\delta}_{\lambda_i}^{2^{(k)}})}+1\right)^{-(\hat{\nu}_i^{(k)}/2+1)}\Bigg\}, \qquad (23)\end{aligned}$$

and

$$\begin{aligned}\hat{s}_{4ij}^{(k)} = \hat{z}_{ij}^{(k)}\Bigg\{&\mathrm{DG}\left(\frac{\hat{\nu}_i^{(k)}+1}{2}\right)+\frac{\hat{\nu}_i^{(k)}+1}{\hat{\eta}_{ij}^{2^{(k)}}+\hat{\nu}_i^{(k)}}\\ &\times\left(\frac{T_{\hat{\nu}_i^{(k)}+3}\left(\hat{M}_{ij}^{(k)}\sqrt{\frac{\hat{\nu}_i^{(k)}+3}{\hat{\nu}_i^{(k)}+1}}\right)}{T_{\hat{\nu}_i^{(k)}+1}\left(\hat{M}_{ij}^{(k)}\right)}-1\right)\\ &-\log\left(\frac{\hat{\eta}_{ij}^{2^{(k)}}+\hat{\nu}_i^{(k)}}{2}\right)+\frac{\hat{\lambda}_i^{(k)}\hat{\eta}_{ij}^{(k)}\left(\hat{\eta}_{ij}^{2^{(k)}}-1\right)}{\sqrt{\left(\hat{\nu}_i^{(k)}+1\right)\left(\hat{\nu}_i^{(k)}+\hat{\eta}_{ij}^{2^{(k)}}\right)^3}}\\ &\times\frac{t_{\hat{\nu}_i^{(k)}+1}\left(\hat{M}_{ij}^{(k)}\right)}{T_{\hat{\nu}_i^{(k)}+1}\left(\hat{M}_{ij}^{(k)}\right)}+\frac{1}{T_{\hat{\nu}_i^{(k)}+1}\left(\hat{M}_{ij}^{(k)}\right)}\\ &\times\int_{-\infty}^{\hat{M}_{ij}^{(k)}} g_{\hat{\nu}_i^{(k)}}(x)t_{\hat{\nu}_i^{(k)}+1}(x)dx\Bigg\}\end{aligned} \tag{24}$$

with

$$\hat{\eta}_{ij}^{(k)} = \frac{y_j-\hat{\xi}_i^{(k)}}{\hat{\sigma}_i^{(k)}}, \quad \hat{\delta}_{\lambda_i}^{(k)} = \frac{\hat{\lambda}_i^{(k)}}{\sqrt{1+\hat{\lambda}_i^{2^{(k)}}}},$$

$$\hat{M}_{ij}^{(k)} = \hat{\lambda}_i^{(k)}\hat{\eta}_{ij}^{(k)}\sqrt{\frac{\hat{\nu}_i^{(k)}+1}{\hat{\eta}_{ij}^{2^{(k)}}+\hat{\nu}_i^{(k)}}},$$

$\psi(y_j|\hat{\mathbf{\Theta}}^{(k)})$ is $\psi(y_j|\mathbf{\Theta})$ in (17) with $\mathbf{\Theta}$ replaced by $\hat{\mathbf{\Theta}}^{(k)}$ and $g_{\hat{\nu}_i^{(k)}}(x)$ is $g_\nu(x)$ in (11) with ν replaced by $\hat{\nu}_i^{(k)}$. The ECM algorithm for the skew t mixture model is as follows:

E-step: Given $\mathbf{\Theta} = \hat{\mathbf{\Theta}}^{(k)}$, compute $\hat{z}_{ij}^{(k)}, \hat{s}_{1ij}^{(k)}, \hat{s}_{2ij}^{(k)}, \hat{s}_{3ij}^{(k)}$ and $\hat{s}_{4ij}^{(k)}$ in Eqs. (20)–(24) for $i = 1, \ldots, g$ and $j = 1, \ldots, n$.

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

CM-step 2: Calculate

$$\hat{\xi}_i^{(k+1)} = \frac{\sum_{j=1}^n \hat{s}_{1ij}^{(k)} y_i - \hat{\delta}_{\lambda_i}^{(k)}\sum_{j=1}^n \hat{s}_{2ij}^{(k)}}{\sum_{j=1}^n \hat{s}_{1ij}^{(k)}}.$$

CM-step 3: Calculate

$$\hat{\sigma}_i^{2^{(k+1)}} = \frac{\sum_{j=1}^n\left(\hat{s}_{1ij}^{(k)}\left(y_j-\hat{\xi}_i^{(k+1)}\right)^2-2\hat{\delta}_{\lambda_i}^{(k)}\hat{s}_{2ij}^{(k)}\left(y_j-\hat{\xi}_i^{(k+1)}\right)+\hat{s}_{3ij}^{(k)}\right)}{2\left(1-\hat{\delta}_{\lambda_i}^{2^{(k)}}\right)\sum_{j=1}^n \hat{z}_{ij}^{(k)}}.$$

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

$$\begin{aligned}&\delta_{\lambda_i}\left(1-\delta_{\lambda_i}^2\right)\sum_{j=1}^n \hat{z}_{ij}^{(k)}\\ &-\delta_{\lambda_i}\left(\sum_{j=1}^n \frac{\hat{s}_{1ij}^{(k)}\left(y_i-\hat{\xi}_i^{(k+1)}\right)^2}{\hat{\sigma}_i^{2(k+1)}}+\sum_{j=1}^n \frac{\hat{s}_{3ij}^{(k)}}{\hat{\sigma}_i^{2^{(k+1)}}}\right)\\ &+\left(1+\delta_{\lambda_i}^2\right)\sum_{j=1}^n \frac{\hat{s}_{2ij}^{(k)}\left(y_j-\hat{\xi}_i^{(k+1)}\right)}{\hat{\sigma}_i^{2^{(k+1)}}} = 0.\end{aligned}$$

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

$$\log\left(\frac{\nu_i}{2}\right)+1-\mathrm{DG}\left(\frac{\nu_i}{2}\right)+\frac{\sum_{j=1}^n\left(\hat{s}_{4ij}^{(k)}-\hat{s}_{1ij}^{(k)}\right)}{\sum_{j=1}^n \hat{z}_{ij}^{(k)}} = 0.$$

If the degrees of freedom are assumed to be identical, i.e. $\nu_1 = \cdots = \nu_g = \nu$, we suggest that the CM-step 5 of the above ECM algorithm be switched to a simple CML step as follows:

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

$$\begin{aligned}\hat{\nu}^{(k+1)} = \operatorname*{argmax}_{\nu}\sum_{j=1}^n \log\\ \left(\sum_{i=1}^g \hat{w}_i^{(k+1)} f\left(y_j \mid \hat{\xi}_i^{(k+1)},\ \hat{\sigma}_i^{2(k+1)}, \lambda_i^{(k+1)}, \nu\right)\right).\end{aligned}$$

Following similar ideas as Liu et al. (1998), the PX-EM algorithm for the STMIX model can be obtained by replacing the CM-steps 3 and 5 in the previous ECM algorithm with the following PX.CM steps:

PX.CM-step3:

$$\hat{\sigma}_i^{2(k+1)} = \frac{\sum_{j=1}^n \hat{s}_{1ij}^{(k)}\left(y_j-\hat{\xi}_i^{(k+1)}\right)^2-2\hat{\delta}_i^{(k)}\sum_{j=1}^n \hat{s}_{2ij}^{(k)}\left(y_j-\hat{\xi}_i^{(k+1)}\right)+\sum_{j=1}^n \hat{s}_{3ij}^{(k)}}{2\left(1-\hat{\delta}_i^{2(k)}\right)\sum_{j=1}^n \hat{s}_{1ij}^{(k)}}.$$

PX.CM-step5:

$$\log\left(\frac{\nu_i\sum_{j=1}^n \hat{z}_{ij}^{(k)}}{2\sum_{j=1}^n \hat{s}_{1ij}^{(k)}}\right)-\mathrm{DG}\left(\frac{\nu_i}{2}\right)+\frac{\sum_{j=1}^n \hat{s}_{4ij}^{(k)}}{\sum_{j=1}^n \hat{z}_{ij}^{(k)}} = 0.$$

Besides being simple in implementation while maintaining the simplicity and stability properties of the EM algorithm, the PX-EM algorithm is desirable since its convergence is always faster and often much faster than the original algorithm.

The iterations of the above algorithm are repeated until a suitable convergence rule is satisfied, e.g., $\|\hat{\mathbf{\Theta}}^{(k+1)} - \hat{\mathbf{\Theta}}^{(k)}\|$ is sufficiently small. An oft-voiced criticism is that the EM-type procedure tends to get stuck in local modes. A convenient way to circumvent such limitations is to try several EM iterations with a variety of starting values that are representative of the parameter space. If there exist several modes, one can find the global mode by comparing their relative masses and log-likelihood values.

Under some general regularity conditions, we follow Basford et al. (1997) to provide an information-based method to obtain the asymptotic covariance of ML estimates of mixture model parameters. By a similar argument as noted earlier, we define by $I_o(\hat{\boldsymbol{\Theta}}; \boldsymbol{y}) = \sum_{j=1}^n \hat{\boldsymbol{u}}_j \hat{\boldsymbol{u}}_j^{\mathrm{T}}$ the observed information matrix, where $\boldsymbol{u}_j = \partial\psi(y_j|\boldsymbol{\Theta})/\partial\boldsymbol{\Theta}$ is the complete-data score statistic corresponding to the single observation y_j $(j = 1, \ldots, n)$.

Corresponding to the vector of all $5g - 1$ unknown parameters in $\boldsymbol{\Theta}$, let $\hat{\boldsymbol{u}}_j$ be a vector containing

$$\begin{aligned}&(\hat{u}_{j,w_1}, \ldots, \hat{u}_{j,w_{g-1}}, \hat{u}_{j,\xi_1}, \ldots, \hat{u}_{j,\xi_g}, \hat{u}_{j,\sigma_1}, \ldots, \hat{u}_{j,\sigma_g},\\ &\quad \hat{u}_{j,\lambda_1}, \ldots, \hat{u}_{j,\lambda_g}, \hat{u}_{j,\nu_1}, \ldots, \hat{u}_{j,\nu_g})^{\mathrm{T}}.\end{aligned}$$

The elements of $\hat{\boldsymbol{u}}_j$ are given by

$$\begin{aligned}
\hat{u}_{j,w_r} &= \frac{\hat{z}_{rj}}{\hat{w}_r} - \frac{\hat{z}_{gj}}{\hat{w}_g},\\
\hat{u}_{j,\xi_r} &= \frac{\hat{z}_{rj}}{\hat{\sigma}_r}\left(\frac{\hat{\nu}_r + 1}{\hat{\eta}_{rj}^2 + \hat{\nu}_r}\right)\\
&\quad \times \left[\hat{\eta}_{rj} - \frac{\hat{\lambda}_r \hat{\nu}_r}{\sqrt{(\hat{\nu}_r + 1)\big(\hat{\eta}_{rj}^2 + \hat{\nu}_r\big)}} \frac{t_{\hat{\nu}_r+1}(\hat{M}_{rj})}{T_{\hat{\nu}_r+1}(\hat{M}_{rj})}\right],\\
\hat{u}_{j,\sigma_r} &= \frac{\hat{z}_{rj}}{\hat{\sigma}_r}\left[\frac{\hat{\nu}_r(\hat{\eta}_{rj}^2 - 1)}{\hat{\eta}_{rj}^2 + \hat{\nu}_r}\right.\\
&\quad \left. - \hat{\eta}_{rj}\frac{\hat{\lambda}_r\hat{\nu}_r}{\hat{\sigma}_r}\sqrt{\frac{\hat{\nu}_r + 1}{\big(\hat{\eta}_{rj}^2 + \hat{\nu}_r\big)^3}}\frac{t_{\hat{\nu}_r+1}(\hat{M}_{rj})}{T_{\hat{\nu}_r+1}(\hat{M}_{rj})}\right],\\
\hat{u}_{j,\lambda_r} &= \hat{z}_{rj}\hat{\eta}_{rj}\sqrt{\frac{\hat{\nu}_r + 1}{\hat{\eta}_{rj}^2 + \hat{\nu}_r}}\frac{t_{\hat{\nu}_r+1}(\hat{M}_{rj})}{T_{\hat{\nu}_r+1}(\hat{M}_{rj})},\\
\hat{u}_{j,\nu_r} &= \frac{\hat{z}_{rj}}{2}\left[\mathrm{DG}\left(\frac{\hat{\nu}_r + 1}{2}\right) - \mathrm{DG}\left(\frac{\hat{\nu}_r}{2}\right) - \log\left(\frac{\hat{\nu}_r + \hat{\eta}_{rj}^2}{\hat{\nu}_r}\right)\right.\\
&\quad + \frac{\hat{\eta}_{rj}^2 - 1}{\hat{\eta}_{rj}^2 + \hat{\nu}_r} + \frac{\hat{\lambda}_r\hat{\eta}_{rj}\big(\hat{\eta}_{rj}^2 - 1\big)}{\sqrt{\big(\hat{\nu}_r + 1\big)(\hat{\eta}_{rj}^2 + \hat{\nu}_r)^3}}\frac{t_{\hat{\nu}_r+1}(\hat{M}_{rj})}{T_{\hat{\nu}_r+1}(\hat{M}_{rj})}\\
&\quad \left. + \frac{1}{T_{\hat{\nu}_r+1}(\hat{M}_{rj})}\int_{-\infty}^{\hat{M}_{rj}} g_{\hat{\nu}_r}(x_j)t_{\hat{\nu}_r+1}(x_j)dx_j\right],
\end{aligned}$$

where $\hat{z}_{rj} = \hat{w}_r f(y_j|\hat{\xi}_r, \hat{\sigma}_r^2, \hat{\lambda}_r, \hat{\nu}_r)/\psi(y_j|\hat{\boldsymbol{\Theta}})$ for $r = 1, \ldots, g$. If the degrees of freedom are assumed to be equal, say $\nu_1 = \cdots = \nu_g = \nu$, we have $\hat{u}_{j,\nu} = \sum_{r=1}^g \hat{u}_{j,\nu_r}$.

5 An illustrative example

Obesity is one of the key factors for many chronic diseases and the trend in the prevalence of obesity in the U.S. continues to increase (Flegal et al., 2002). Body mass index (BMI; kg/m^2), calculated by the ratio of body weight in kilograms and body height in meters squared, has become the medical standard used to measure overweight and obesity. For adults, overweight is defined as a BMI value between 25 to 29.9, and obesity is defined as a BMI value greater than or equal to 30.

In America, the National Center for Health Statistics (NCHS) of the Center for Disease Control (CDC) has conducted a national health and nutrition examination survey (NHANES) annually since 1999. The survey data are released in a two-year cycle.

For illustration, we consider the BMI for men aged 18 to 80 years in the two recent releases NHANES 1999–2000 and NHANES 2001–2002. There are 4,579 participants (adult men) with BMI records. Of these participants, the correlation coefficient between BMI and body weight is 0.914, indicating they are highly correlated. To explore a mixture pattern of BMI arising from two intrinsic groups of body weights, participants with weights ranging between 70.1(kg) to 95.0 (kg) were dropped in our analyses. The remaining data, namely *bmimen*, consist of 1,069 and 1,054 participants with body weights lying within [39.50 kg, 70.00 kg] and [95.01 kg, 196.80 kg], respectively.

For comparison purposes, we fit the data with a two-component mixture model using normal, Student' t, skew normal, and skew t as component densities, while the degrees of freedom are assumed to be equal. To be more specific, a two-component STMIX model with equal degrees of freedom can be written as

$$\psi(y|\boldsymbol{\Theta}) = wf\big(y|\xi_1, \sigma_1^2, \lambda_1, \nu\big) + (1 - \omega)f\big(y|\xi_2, \sigma_2^2, \lambda_2, \nu\big). \tag{25}$$

Of course, model (25) will include NORMIX ($\lambda_1 = \lambda_2 = 0$; $\nu = \infty$), TMIX ($\lambda_1 = \lambda_2 = 0$), and SNMIX ($\nu = \infty$) as special cases.

For comparing the fitting results, the ML estimates and the associated information-based standard errors together with the log-likelihood, and AIC and BIC values for NORMIX, TMIX, SNMIX and STMIX models are summarized in Table 1. When comparing these fitted models, we notice that the smaller the AIC and BIC values, the better the fit. It is evidently seen that the STMIX model has the best fitting result. Comparing STMIX with SNMIX, it is observed that using a heavy-tailed t distribution will reduce the skewness effects. In Fig. 1, we plot the profile log-likelihood of the degrees of freedom ν for the STMIX model to illustrate that the SNMIX

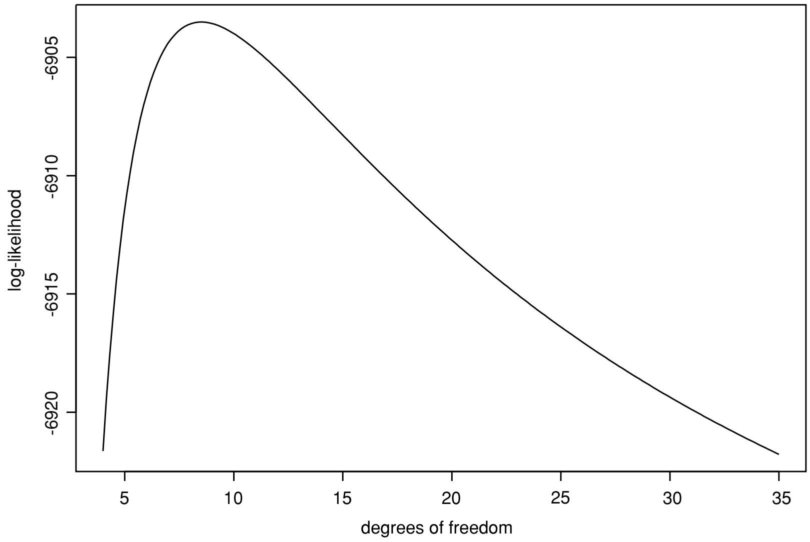

Fig. 1 Plot of the profile log-likelihood of the degrees of freedom ν for fitting the *bmimen* data with a two component STMIX model with equal degrees of freedom ($\nu_1 = \nu_2 = \nu$)

Table 1 ML estimation results for fitting various mixture models on the BMI adult men example

	NORMIX		TMIX		SNMIX		STMIX	
Parameter	Mle	Se	Mle	Se	Mle	Se	Mle	Se
w	0.397	0.0188	0.438	0.017	0.531	0.013	0.539	0.017
ξ_1	21.443	0.0465	21.591	0.089	19.567	0.036	19.672	0.330
ξ_2	32.565	0.1845	33.030	0.264	28.760	0.009	29.173	0.182
σ_1	2.021	0.0866	1.956	0.083	3.731	0.288	3.482	0.350
σ_2	6.422	0.1584	5.006	0.242	7.960	0.159	6.679	0.232
λ_1	—	—	—	—	1.834	0.344	1.782	0.257
λ_2	—	—	—	—	10.184	2.615	5.912	1.400
ν	—	—	7.075	1.314	—	—	8.502	1.441
$\ell(\hat{\Theta})$	−6958.37		−6934.69		−6916.26		−6903.51	
AIC	13926.74		13881.38		13846.52		13823.02	
BIC	13955.04		13915.34		13886.14		13868.30	

model is not favorable for this data set since the profile log-likelihood has a significant drop at the peak value of 8.5.

To compare these four mixture models in density estimation, we display the fitting results superimposed on a single set of coordinate axes in Fig. 2. Based on the graphical visualization, it appears that the STMIX fitted density performs more adequately than the other three fitted densities. It is of interest to emphasize that the SNMIX fitting leads to an increase in skewness for coping with excessive heavy tailedness. Furthermore, the fitted SNMIX density is very close to the fitted STMIX density in the tail region.

To study the validity of the four hypothesized mixture models, we perform the Kolmogorov-Smirnov's (K-S) goodness of fit test. The procedure for calculating the K-S test statistic D_n, which is defined as the maximum value of the absolute difference between the empirical and estimated cumulative distributions, and the corresponding p-values are described below.

Step 1: Ordering n data values yields $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}$.

Step 2: Compute the K-S test statistic

$$D_n = \max_{j=1,\ldots,n} \left\{ \frac{j}{n} - \hat{F}(y_{(j)}), \hat{F}(y_{(j)}) - \frac{j-1}{n} \right\},$$

where $\hat{F}(\cdot)$ is the fitted cdf of a hypothesized mixture distribution.

Step 3: Generate n random random numbers from $U(0, 1)$ and order them, we have $u^{(i)}_{(1)} \leq u^{(i)}_{(2)} \leq \cdots \leq u^{(i)}_{(n)}$.

Step 4: Compute

$$d^{(i)} = \max_{j=1,\ldots,n} \left\{ \frac{j}{n} - u^{(i)}_{(j)}, u^{(i)}_{(j)} - \frac{j-1}{n} \right\}.$$

Step 5: Let $I_i = 1$ if $d^{(i)} \geq D_n$ and 0 otherwise. Repeat *Steps* 3 and 4 N times, we get $I_1, \ldots, I_N$. The p-value is estimated by $\sum_{i=1}^{N} I_i/N$.

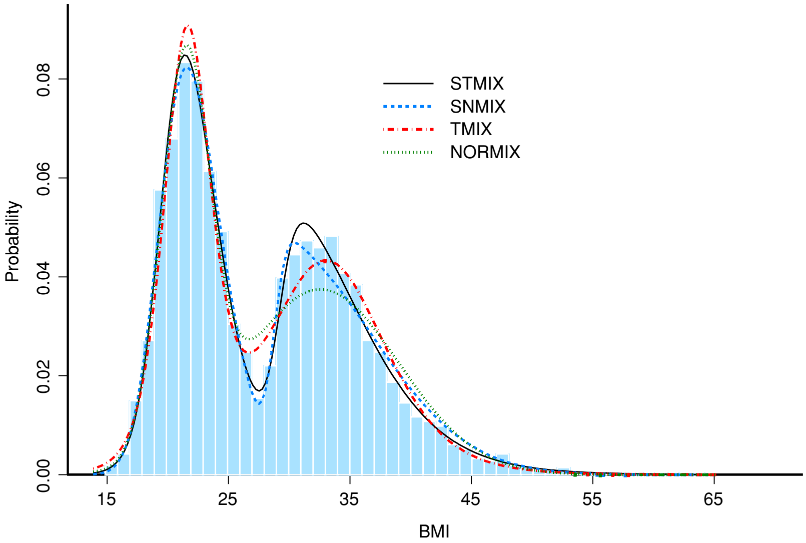

Fig. 2 Histogram of the *bmimen* data with overlaid four ML-fitted two component mixture densities (normal, Student's t, skew normal and skew t)

The resulting K-S tests are listed in Table 2 . The reported p-values can be used as a similarity assessment of the experimental data against the fitted distribution. Of the four mixture models, the best fit is STMIX with a p-value of 0.971. That is, it strongly suggests that the *bmimen* data follow a mixture of skew t distributions.

6 Concluding remarks

We have proposed a robust approach to a finite mixture model based on the skew t distribution, called the STMIX model, which accommodates both asymmetry and heavy tails jointly that allows practitioners for analyzing data in a wide variety of considerations. We have described a normal-truncated normal-gamma-multinomial hierarchy for the STMIX model and presented some modern EM-type algorithms for ML estimation in a flexible complete-data framework. We demonstrate our approach with a real data set and show that the STMIX model has better performance than the other competitors.

Due to recent advances in computational technology, it is worthwhile to carry out Bayesian treatments via Markov chain Monte Carlo (MCMC) sampling methods in the context of STMIX model. The basic idea is to explore the joint posterior distributions of the model parameters together with latent variables $\boldsymbol{\gamma}$ and $\boldsymbol{\tau}$, and allocation variables $\boldsymbol{Z}$ when informative priors are employed. Other extensions of the current work include, for example, a generalization of STMIX to multivariate settings (e.g., Azzalini and Capitanio, 2003; Jones and Faddy, 2003) and determination of the number of components in skew t mixtures via reversible jump MCMC (e.g., Richardson and Green, 1997; Zhang et al., 2004; Dellaportas and Papageorgiou, 2006).

Table 2 The K-S test results for the four fitted mixture models

Model	NORMIX	TMIX	SNMIX	STMIX
D_n	0.0322	0.0190	0.0240	0.0106
p-value	0.022	0.425	0.176	0.971

Appendix: Proof of Proposition 2

(a) Standard calculation of conditional expectation yields

$$\begin{aligned}E(\tau \mid y) &= \int_0^\infty \tau f(\tau \mid y) d\tau \\ &= \int_0^\infty b\tau^{\frac{\nu+1}{2}} \exp\Big(-\frac{\tau}{2}(\eta^2+\nu)\Big)\Phi\big(\lambda\eta\sqrt{\tau}\big)d\tau \\ &= b\frac{\Gamma\left(\frac{\nu+3}{2}\right)}{\left(\frac{\eta^2+\nu}{2}\right)^{(\nu+3)/2}} \\ &\quad \times \int_0^\infty \gamma\left(\tau \,\middle|\, \frac{\nu+3}{2}, \frac{\eta^2+\nu}{2}\right)\Phi\big(\lambda\eta\sqrt{\tau}\big)d\tau,\end{aligned}$$

where $\gamma(\cdot|\alpha,\beta)$ denotes the density of $\Gamma(\alpha,\beta)$ and b is given in (10).

By Proposition 1, it suffices to show

$$E(\tau \mid y) = \left(\frac{\nu+1}{\eta^2+\nu}\right)\frac{T_{\nu+3}\left(\lambda\eta\sqrt{\frac{\nu+3}{\eta^2+\nu}}\right)}{T_{\nu+1}\left(\lambda\eta\sqrt{\frac{\nu+1}{\eta^2+\nu}}\right)}.$$

(b) We first need to show the following:

$$
\begin{aligned}
&E\left(\sqrt{\tau}\frac{\phi(\lambda\eta\sqrt{\tau})}{\Phi(\lambda\eta\sqrt{\tau})}\bigg| y\right)\\
&\quad=\int_0^\infty \sqrt{\tau}\frac{\phi(\lambda\eta\sqrt{\tau})}{\Phi(\lambda\eta\sqrt{\tau})}\frac{f(\tau,y)}{f(y)}d\tau\\
&\quad=\frac{(\nu/2)^{\nu/2}}{\pi\sigma\Gamma(\nu/2)f(y)}\\
&\qquad\times\int_0^\infty \tau^{(\nu/2+1)-1}\exp\left(-\frac{\tau}{2}\left(\frac{\eta^2}{1-\delta_\lambda^2}+\nu\right)\right)d\tau\\
&\quad=\frac{1}{\pi\sigma f(y)}\left(\frac{\eta^2}{\nu(1-\delta_\lambda^2)}+1\right)^{-(\nu/2+1)}. \qquad (A.1)
\end{aligned}
$$

From (8), the expectation of a truncated normal distribution is given by

$$
E(\gamma \mid y,\tau)=\delta_\lambda(y-\xi)+\frac{\phi(\lambda\eta\sqrt{\tau})}{\Phi(\lambda\eta\sqrt{\tau})}\sqrt{\frac{1-\delta_\lambda^2}{\tau}}\sigma. \qquad (A.2)
$$

Applying the law of iterated expectation and using (A.1) and (A.2), we get

$$
\begin{aligned}
E(\gamma\tau\mid y)&=E\left(\tau E(\gamma\mid y,\tau)\mid y\right)\\
&=\delta_\lambda(y-\xi)E(\tau\mid y)\\
&\quad+\sqrt{1-\delta_\lambda^2}\sigma E\left(\sqrt{\tau}\frac{\phi(\lambda\eta\sqrt{\tau})}{\Phi(\lambda\eta\sqrt{\tau})}\bigg| y\right)\\
&=\delta_\lambda(y-\xi)E(\tau|y)\\
&\quad+\frac{\sqrt{1-\delta_\lambda^2}}{\pi f(y)}\left(\frac{\eta^2}{\nu(1-\delta_\lambda^2)}+1\right)^{-(\nu/2+1)}.
\end{aligned}
$$

(c) Similarly, it is easy to verify that

$$
\begin{aligned}
E(\gamma^2\mid y,\tau)&=\delta_\lambda^2(y-\xi)^2+\frac{(1-\delta_\lambda^2)\sigma^2}{\tau}\\
&\quad+\sigma\delta_\lambda(y-\xi)\sqrt{\frac{1-\delta_\lambda^2}{\tau}}\frac{\phi(\lambda\eta\sqrt{\tau})}{\Phi(\lambda\eta\sqrt{\tau})}. \qquad (A.3)
\end{aligned}
$$

From (A.1) and (A.3), applying the law of iterated expectation gives

$$
\begin{aligned}
E(\gamma^2\tau\mid y)&=\delta_\lambda^2(y-\xi)^2E(\tau|y)+(1-\delta_\lambda^2)\sigma^2\\
&\quad+\frac{\delta_\lambda(y-\xi)\sqrt{1-\delta_\lambda^2}}{\pi f(y)}\left(\frac{\eta^2}{\nu(1-\delta_\lambda^2)}+1\right)^{-(\nu/2+1)}.
\end{aligned}
$$

(d) From (9), it is true that

$$
\begin{aligned}
&\frac{d}{d\nu}\int_0^\infty f(\tau\mid y)d\tau\\
&\quad=\frac{d}{d\nu}\int_0^\infty b\tau^{(\nu-1)/2}\exp\left(-\frac{\tau}{2}(\eta^2+\nu)\right)\Phi(\lambda\eta\sqrt{\tau})d\tau=0.
\end{aligned}
$$

By Leibnitz's rule, we can get

$$
\begin{aligned}
&\log\left(\frac{\eta^2+\nu}{2}\right)+\left(\frac{\nu+1}{\eta^2+\nu}\right)-\mathrm{DG}\left(\frac{\nu+1}{2}\right)\\
&\quad-\frac{1}{T_{\nu+1}(M)}\int_{-\infty}^{M}g_\nu(x)t_{\nu+1}(x)dx\\
&\quad+\lambda(\nu+1)^{-\frac{1}{2}}\eta(\eta^2-1)(\eta^2+\nu)^{-\frac{3}{2}}\frac{t_{\nu+1}(M)}{T_{\nu+1}(M)}\\
&\quad+E(\log\tau\mid y)-E(\tau\mid y)=0.
\end{aligned}
$$

Hence

$$
\begin{aligned}
&E\left(\log(\tau)|y\right)\\
&\quad=\mathrm{DG}\left(\frac{\nu+1}{2}\right)-\log\left(\frac{\eta^2+\nu}{2}\right)+\frac{\nu+1}{\eta^2+\nu}\\
&\qquad\times\left(\frac{T_{\nu+3}\left(M\sqrt{\frac{\nu+3}{\nu+1}}\right)}{T_{\nu+1}(M)}-1\right)\\
&\qquad+\frac{\lambda\eta(\eta^2-1)}{\sqrt{(\nu+1)(\nu+\eta^2)^3}}\frac{t_{\nu+1}(M)}{T_{\nu+1}(M)}\\
&\qquad+\frac{1}{T_{\nu+1}(M)}\int_{-\infty}^{M}g_\nu(x)t_{\nu+1}(x)dx.
\end{aligned}
$$

Acknowledgments The authors are grateful to the editors and one anonymous referee for their valuable comments and suggestions that led to significant improvement of the paper. This research work was supported in part by grants NSC94-2118-M-005-003 and NSC94-2118-M-009-001 from the National Science Council of Taiwan.

References

Azzalini A. 1985. A class of distributions which includes the normal ones. Scandinavian Journal of Statistics 12: 171–178.

Azzalini A. 1986. Further results on a class of distributions which includes the normal ones. Statistica 46: 199–208.

Azzalini A. and Capitaino A. 2003. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. Journal of the Royal Statistical Society, Series B 65: 367–389.

Basford K.E., Greenway D.R., McLachlan G.J., and Peel, D. 1997. Standard errors of fitted means under normal mixture. Computational Statistics 12: 1–17.

Dellaportas P. and Papageorgiou I. 2006. Multivariate mixtures of normals with unknown number of components. Statistics and Computing 16: 57–68.

Dempster A.P., Laird N.M., and Rubin D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society, Series B 39: 1–38.

Flegal K.M., Carroll M.D., Ogden C.L., and Johnson C.L. 2002. Prevalence and trends in obesity among US adults, 1999–2000. Journal of the American Medical Association 288: 1723–1727.

Henze N. 1986. A probabilistic representation of the skew-normal distribution. Scandinavian Journal of Statistics 13: 271–275.

Jones M.C. and Faddy M.J. 2003. A skew extension of the t-distribution, with applications. Journal of the Royal Statistical Society, Series B 65: 159–174.

Lin T.I., Lee J.C., and Ni H.F. 2004. Bayesian analysis of mixture modelling using the multivariate t distribution. Statistics and Computing 14: 119–130.

Lin T.I., Lee J.C., and Yen S.Y. 2007. Finite mixture modelling using the skew normal distribution. Statistica Sinica (In press).

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

Liu C.H., Rubin D.B., and Wu, Y. 1998. Parameter expansion to accelerate EM: the PX-EM algorithm. Biometrika 85: 755–770.

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

McLachlan G.J. and Peel D. 2000. Finite Mixture Models, Wiely, New York.

Meng X.L. and Rubin D.B. 1993. Maximum likelihood estimation via the ECM algorithm: a general framework. Biometrika 80: 267–78.

Peel D. and McLachlan G. J. 2000. Robust mixture modeling using the t distribution. Statistics and Computing 10: 339–348.

Richardson S. and Green P.J. 1997. On Bayesian analysis of mixtures with an unknown number of components (with discussion). Journal of the Royal Statistical Society, Series B 59: 731–792.

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

Shoham S., Fellows M.R., and Normann R.A. 2003. Robust, automatic spike sorting using mixtures of multivariate t-distributions. Journal of Neuroscience Methods 127: 111–122.

Titterington D.M., Smith A.F.M., and Markov U.E. 1985. Statistical Analysis of Finite Mixture Distributions, Wiely, New York.

Wang H.X., Zhang Q.B., Luo B., and Wei S. 2004. Robust mixture modelling using multivariate t distribution with missing information. Pattern Recognition Letter 25: 701–710.

Zacks S. 1971. The Theory of Statistical Inference, New York, Wiley.

Zhang Z., Chan K.L., Wu Y., and Cen C.B. 2004. Learning a multivariate Gaussian mixture model with the reversible Jump MCMC algorithm. Statistics and Computing 14: 343–355.