Statistica Sinica **17**(2007), 909-927

FINITE MIXTURE MODELLING USING THE SKEW NORMAL DISTRIBUTION

Tsung I. Lin[1], Jack C. Lee[2] and Shu Y. Yen[2]

[1]*National Chung Hsing University and* [2]*National Chiao Tung University*

Abstract: Normal mixture models provide the most popular framework for modelling heterogeneity in a population with continuous outcomes arising in a variety of subclasses. In the last two decades, the skew normal distribution has been shown beneficial in dealing with asymmetric data in various theoretic and applied problems. In this article, we address the problem of analyzing a mixture of skew normal distributions from the likelihood-based and Bayesian perspectives, respectively. Computational techniques using EM-type algorithms are employed for iteratively computing maximum likelihood estimates. Also, a fully Bayesian approach using the Markov chain Monte Carlo method is developed to carry out posterior analyses. Numerical results are illustrated through two examples.

Key words and phrases: ECM algorithm, ECME algorithm, Fisher information, Markov chain Monte Carlo, maximum likelihood estimation, skew normal mixtures.

1. Introduction

Finite mixture models have been broadly developed and widely applied to classification, clustering, density estimation and pattern recognition problems, as shown by Titterington, Smith and Markov (1985), McLachlan and Basord (1988), McLachlan and Peel (2000), and the references therein. With the growing advances of computational methods, especially for the development of Markov chain Monte Carlo (MCMC) techniques, many works are also devoted to Bayesian mixture modelling issues, including Diebolt and Robert (1994), Escobar and West (1995), Richardson and Green (1997) and Stephens (2000), among others.

In many applied problems, the shapes of fitted mixture normal components may be distorted, and inferences can be misleading when the data involves highly asymmetric observations. In particular, the normal mixture (NORMIX) model tends to *overfit* when additional components are included to capture the skewness. Sometimes, increasing the number of pseudo-components may lead to difficulties and inefficiencies in computations. Instead, we consider using the skew normal distributions proposed by Azzalini (1985) as component densities to overcome the potential weakness of normal mixtures. The skew normal distribution is a new class of density functions dependent on an additional shape parameter, and includes the normal density as a special case. It provides a more flexible approach to the fitting of asymmetric observations and uses fewer components in the fitting of mixture models. A comprehensive coverage of the fundamental theory and new developments for skew-elliptical distributions is given by Genton (2004).

It is not easy to deal with computational aspects of parameter estimation for the fitting of skew normal mixture (SNMIX) models. For simplicity, we treat the number of components as known and describe how to employ EM-type algorithms for finding the maximum likelihood (ML) estimates. In addition, Bayesian sampling methods for SNMIX are considered as an alternative modelling strategy. Priors and hyperparameters are chosen as weakly informative to avoid nonidentifiability problems in the mixture context.

The rest of the paper unfolds as follows. Section 2 briefly outlines some preliminaries of the skew normal distribution. Azzalini and Capitaino (1999) pointed out that the ML estimates might be improved by a few EM iterations, but detailed expressions of the EM algorithm are not available in the literature. We thus show how to compute the ML estimates for the skew normal distribution using two EM-type algorithms. In Section 3 we show a hierarchical representation for the SNMIX model by incorporating two latent variables. Based on the model, we also derive the corresponding EM-type algorithms for ML estimation. Meanwhile, the information-based standard errors are also presented. In Section 4, we develop the MCMC sampling algorithm used in simulating posterior distributions to carry out Bayesian inferences. In Section 5, two examples are given, and in Section 6 we provide some concluding remarks.

2. The Skew Normal Distribution

2.1. Preliminaries

As developed by Azzalini (1985, 1986), a random variable Y follows a univariate skew normal distribution with location parameter ξ, scale parameter σ^2 and skewness parameter $\lambda \in \mathbb{R}$ if it has the density

$$\psi(y \mid \xi, \sigma^2, \lambda) = \frac{2}{\sigma}\phi\Big(\frac{y-\xi}{\sigma}\Big)\Phi\Big(\lambda\frac{y-\xi}{\sigma}\Big), \tag{1}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal density function and cumulative distribution function, respectively; then, for brevity, we say that $Y \sim SN(\xi, \sigma^2, \lambda)$. Note that if $\lambda = 0$, the density of Y reduces to the $N(\xi, \sigma^2)$ density.

Lemma 1. *If $Y \sim SN(\xi, \sigma^2, \lambda)$ and $X \sim N(\xi, \sigma^2/(1+\lambda^2))$, we have*
(i) $E(X^{n+1}) = \xi E(X^n) + [\sigma^2/(1+\lambda^2)][dE(X^n)/d\xi]$.

(ii) $E(Y^{n+1}) = \xi E(Y^n) + \sigma^2[dE(Y^n)/d\xi] + \sqrt{2/\pi}\delta(\lambda)\sigma E(X^n)$.

(iii) $E\big\{Y - \mathrm{E}\,(Y)\big\}^{n+1} = \sigma^2[dE\{Y - E(Y)\}^n/d\xi] + n\sigma^2 E\big\{Y - E(Y)\big\}^{n-1}$ $- \big\{E(Y) - \xi\big\}E\big\{Y - E(Y)\big\}^n + \sqrt{2/\pi}\delta(\lambda)\sigma E\big\{X - E(Y)\big\}^n$.

Lemma 1 provides a simple way of obtaining higher moments without using the moment generating function. With some basic algebraic manipulations, we can easily obtain

$$E(Y) = \xi + \sqrt{\frac{2}{\pi}}\delta(\lambda)\sigma, \quad \mathrm{var}(Y) = \Big\{1 - \frac{2}{\pi}\delta^2(\lambda)\Big\}\sigma^2,$$
$$\gamma_Y = \frac{\sqrt{2}(4-\pi)\lambda^3}{\Big\{\pi + (\pi-2)\lambda^2\Big\}^{3/2}}, \quad \kappa_Y = 3 + \frac{8(\pi-3)\lambda^4}{\Big\{\pi + (\pi-2)\lambda^2\Big\}^2}, \tag{2}$$

where $\delta(\lambda) = \lambda/\sqrt{1+\lambda^2}$, and γ_Y and κ_Y are the measures of skewness and kurtosis, respectively. It is easily shown that γ_Y is in $(-0.9953,\ 0.9953)$ and κ_Y is in $(3,\ 3.8692)$. Henze (1986) showed that the odd moments of the standard skew normal variable $Z = (Y - \xi)/\sigma$ have the expression

$$E(Z^{2k+1}) = \sqrt{\frac{2}{\pi}}\lambda(1+\lambda^2)^{-(k+0.5)}2^{-k}(2k+1)!\sum_{j=0}^{k}\frac{j!(2\lambda)^{2j}}{(2j+1)!(k-j)!},$$

while the even moments coincide with those of standard normal, as $Z^2 \sim \chi_1^2$ (Roberts and Geisser (1966)).

From (2), Arnold, Beaver, Groeneveld and Meeker (1993) showed the following method of moments estimators:

$$\tilde{\xi} = m_1 - a_1\left(\frac{m_3}{b_1}\right)^{\frac{1}{3}},$$
$$\tilde{\sigma}^2 = m_2 + a_1^2\left(\frac{m_3}{b_1}\right)^{\frac{2}{3}},$$
$$\tilde{\delta}(\lambda) = \left\{a_1^2 + m_2\left(\frac{b_1}{m_3}\right)^{\frac{2}{3}}\right\}^{-\frac{1}{2}}, \tag{3}$$

where $a_1 = \sqrt{2/\pi}$, $b_1 = (4/\pi - 1)a_1$, $m_1 = n^{-1}\sum_{i=1}^n Y_i$, $m_2 = (n-1)^{-1}\sum_{i=1}^n(Y_i - \bar{Y}_i)^2$, and $m_3 = (n-1)^{-1}\sum_{i=1}^n(Y_i - \bar{Y}_i)^3$.

2.2. Parameter estimation using EM-type algorithms

In this subsection, we show how to exploit two extensions of the EM algorithm (Dempster, Laird and Rubin (1977)), the ECM algorithm (Meng and Rubin (1993)) and the ECME algorithm (Liu and Rubin (1994)), for ML estimation of the skew normal distribution. A key feature of these two EM-type algorithms is that they preserve the stability of the EM algorithm with their monotone convergence. In order to represent the skew normal model in an incomplete data framework, we extend the result of Azzalini (1986, p.201) and (Henze (1986, Thm. 1)) to show that if $Y_j \sim SN(\xi, \sigma^2, \lambda)$, then

$$Y_j = \xi + \delta(\lambda)\tau_j + \sqrt{1-\delta^2(\lambda)}U_j, \tag{4}$$

with $\tau_j \sim TN(0, \sigma^2)I\{\tau_j > 0\}$, $U_j \sim N(0, \sigma^2)$, where τ_j and U_j are independent, $TN(\cdot, \cdot)$ denotes the truncated normal distribution, and $I\{\cdot\}$ represents an indicator function. Letting $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ and $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n)$, the complete-data log-likelihood of $\boldsymbol{\theta} = (\xi, \sigma^2, \lambda)$ given $(\boldsymbol{Y}, \boldsymbol{\tau})$, after omitting additive constants, is

$$\begin{aligned}\ell_c(\boldsymbol{\theta}) = &-n\log(\sigma^2) - \frac{n}{2}\log\Big(1-\delta^2(\lambda)\Big)\\ &-\frac{\sum_{j=1}^n \tau_j^2 - 2\delta(\lambda)\sum_{j=1}^n \tau_j(y_j-\xi) + \sum_{j=1}^n (y_j-\xi)^2}{2\sigma^2\Big(1-\delta^2(\lambda)\Big)}.\end{aligned} \tag{5}$$

Obviously, the posterior distribution of τ_j is

$$\tau_j | Y_j = y_j \sim TN(\mu_{\tau_j}, \sigma_\tau^2)I\{\tau_j > 0\}, \tag{6}$$

where $\mu_{\tau_j} = \delta(\lambda)(y_j - \xi)$ and $\sigma_\tau = \sigma\sqrt{1-\delta^2(\lambda)}$.

Lemma 2. *Let $X \sim TN(\mu, \sigma^2)I\{a_1 < x < a_2\}$ be a truncated normal distribution with the density*

$$f(x|\mu, \sigma^2) = \Big\{\Phi(\alpha_2) - \Phi(\alpha_1)\Big\}^{-1}\frac{1}{\sqrt{2\pi}\sigma}\exp\Big\{-\frac{1}{2\sigma^2}(x-\mu)^2\Big\}, \quad a_1 < x < a_2,$$

where $\alpha_i = (a_i - \mu)/\sigma$, $i = 1,\ 2$. Then

(i) $E(X) = \mu - \sigma\dfrac{\phi(\alpha_2) - \phi(\alpha_1)}{\Phi(\alpha_2) - \Phi(\alpha_1)}$.

(ii) $E(X^2) = \mu^2 + \sigma^2 - \sigma^2\dfrac{\alpha_2\phi(\alpha_2) - \alpha_1\phi(\alpha_1)}{\Phi(\alpha_2) - \Phi(\alpha_1)} - 2\mu\sigma\dfrac{\phi(\alpha_2) - \phi(\alpha_1)}{\Phi(\alpha_2) - \Phi(\alpha_1)}$.

By Lemma 2, we have

$$E(\tau_j|y_j) = \mu_{\tau_j} + \frac{\phi(\frac{\mu_{\tau_j}}{\sigma_\tau})}{\Phi(\frac{\mu_{\tau_j}}{\sigma_\tau})}\sigma_\tau \quad \text{and} \quad E(\tau_j^2|y_j) = \mu_{\tau_j}^2 + \sigma_\tau^2 + \frac{\phi(\frac{\mu_{\tau_j}}{\sigma_\tau})}{\Phi(\frac{\mu_{\tau_j}}{\sigma_\tau})}\mu_{\tau_j}\sigma_\tau.$$

The ECM algorithm is as follows.

E-step: Calculating the conditional expectation of (5) at the kth iteration yields

$$\hat{s}_{1j}^{(k)} = E_{\hat{\boldsymbol{\theta}}^{(k)}}(\tau_j|y_j) = \hat{\mu}_{\tau_j}^{(k)} + \frac{\phi\left\{\hat{\lambda}^{(k)}\left(\frac{y_j-\hat{\xi}^{(k)}}{\hat{\sigma}^{(k)}}\right)\right\}}{\Phi\left\{\hat{\lambda}^{(k)}\left(\frac{y_j-\hat{\xi}^{(k)}}{\hat{\sigma}^{(k)}}\right)\right\}}\hat{\sigma}_{\tau}^{(k)},$$

$$\hat{s}_{2j}^{(k)} = E_{\hat{\boldsymbol{\theta}}^{(k)}}(\tau_j^2|y_j) = \hat{\mu}_{\tau_j}^{(k)2} + \hat{\sigma}_{\tau}^{(k)2} + \frac{\phi\left\{\hat{\lambda}^{(k)}\left(\frac{y_j-\hat{\xi}^{(k)}}{\hat{\sigma}^{(k)}}\right)\right\}}{\Phi\left\{\hat{\lambda}^{(k)}\left(\frac{y_j-\hat{\xi}^{(k)}}{\hat{\sigma}^{(k)}}\right)\right\}}\hat{\mu}_{\tau_j}^{(k)}\hat{\sigma}_{\tau}^{(k)},$$

where $\hat{\mu}_{\tau_j}^{(k)}$, $\hat{\sigma}_{\tau}^{(k)}$ are μ_{τ_j} and σ_τ in (6) with ξ, σ and λ replaced by $\hat{\xi}^{(k)}$, $\hat{\sigma}^{(k)}$ and $\hat{\lambda}^{(k)}$, respectively.

CM-steps

CM-step 1: Update $\hat{\xi}^{(k)}$ by

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

CM-step 2: Update $\hat{\sigma}^{2^{(k)}}$ by

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

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

$$n\hat{\sigma}^{2^{(k+1)}}\delta(\lambda)\big(1-\delta^2(\lambda)\big) + \big(1+\delta^2(\lambda)\big)\sum_{j=1}^{n}(y_j-\hat{\xi}^{(k+1)})\hat{s}_{1j}^{(k)}$$
$$-\delta(\lambda)\sum_{j=1}^{n}\hat{s}_{2j}^{(k)} - \delta(\lambda)\sum_{j=1}^{n}(y_j-\hat{\xi}^{(k+1)})^2 = 0.$$

For the ECME algorithm, the E-step and the first two CM steps are the same as ECM, while the CM-Step 3 of ECM is modified as the following CML-step.

CML-step: Update $\hat{\lambda}^{(k)}$ by optimizing the constrained log-likelihood function, i.e.,

$$\hat{\lambda}^{(k+1)} = \underset{\lambda}{\mathrm{argmax}}\sum_{j=1}^{n}\log\Big\{\Phi\Big(\lambda\frac{y_j-\hat{\xi}^{(k+1)}}{\hat{\sigma}^{(k+1)}}\Big)\Big\}.$$

The maximization in the CML-step requires a one-dimensional search, which can be easily solved by the function "optim" embedded in the statistical package "R". As noted by Liu and Rubin (1994), the ECME has a faster convergence rate than the ECM algorithm.

Lemma 3. *If* $Z \sim SN(0,1,\lambda)$, *then*

(i) $\mathrm{E}\Big\{\frac{\phi(\lambda Z)}{\Phi(\lambda Z)}\Big\} = \sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{1+\lambda^2}}$.

(ii) $\mathrm{E}\Big\{Z^{2k+1}\frac{\phi(\lambda Z)}{\Phi(\lambda Z)}\Big\} = 0,\ k = 0,\ 1,\ 2,\ldots$.

(iii) $\mathrm{E}\Big\{Z^{2}\frac{\phi(\lambda Z)}{\Phi(\lambda Z)}\Big\} = \sqrt{\frac{2}{\pi}}\frac{\lambda}{\big(1+\lambda^2\big)^{\frac{3}{2}}}$.

The method of moments estimators in (3) can provide good initial values. Applying Lemma 3, the Fisher information $\boldsymbol{I}(\xi,\sigma,\lambda)$ can be easily obtained. The results are shown in Azzalini (1985, p.175). The standard errors of ML estimates can be computed by taking the square root of the corresponding diagonal elements of $\boldsymbol{I}^{-1}(\hat{\xi},\hat{\sigma},\hat{\lambda})$.

3. The Skew Normal Mixtures

3.1. The model

We consider a finite mixture model in which a set of independent data $Y_1,\ldots,Y_n$ are from a g-component mixture of skew normal densities

$$f(y_j \mid \boldsymbol{\Theta}) = \sum_{i=1}^{g} \omega_i \psi(y_j \mid \xi_i, \sigma_i^2, \lambda_i), \tag{7}$$

where $\boldsymbol{\omega} = (\omega_1,\ldots,\omega_g)$ are the mixing probabilities, constrained to be non-negative and sum to unity, and $\boldsymbol{\Theta} = (\boldsymbol{\theta}_i,\ldots,\boldsymbol{\theta}_g)$ with $\boldsymbol{\theta}_i = (\omega_i,\xi_i,\sigma_i^2,\lambda_i)$ being the specific parameters for component i.

We introduce a set of latent component-indicators $\boldsymbol{Z}_j = (Z_{1j},\ldots,Z_{gj})$, $j = 1,\ldots,n$, whose values are a set of binary variables with

$$Z_{kj} = \begin{cases} 1 & \text{if } \boldsymbol{Y_j} \text{ belongs to group } k, \\ 0 & \text{otherwise,} \end{cases}$$

and $\sum_{i=1}^{g} Z_{ij} = 1$. Given the mixing probabilities $\boldsymbol{\omega}$, the component-indicators $\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_j$ are independent, with multinomial densities

$$f(\boldsymbol{z}_j) = \omega_1^{z_{1j}}\omega_2^{z_{2j}}\cdots(1-\omega_1-\cdots-\omega_{g-1})^{z_{gj}}. \tag{8}$$

We write $\boldsymbol{Z}_j \sim \mathcal{M}(1;\ \omega_1,\ldots,\omega_g)$ to denote $\boldsymbol{Z}_j$ with density (8).

From (4), a hierarchical model for skew normal mixtures can be written as

$$\begin{aligned} &Y_j \mid \tau_j,\ Z_{ij} = 1 \sim N\Big(\xi_i + \delta(\lambda_i)\tau_j,\ \big(1-\delta^2(\lambda_i)\big)\sigma_i^2\Big), \\ &\tau_j \mid Z_{ij} = 1 \sim TN(0,\sigma_i^2)I(\tau_j > 0), \\ &\boldsymbol{Z}_j \sim \mathcal{M}(1;\ \omega_1,\ldots,\omega_g) \qquad (j = 1,\ldots,n). \end{aligned} \tag{9}$$

3.2. Maximum likelihood estimation

As in (6), we have $\tau_j \mid Y_j = y_j, Z_{ij} = 1 \sim TN(\mu_{\tau_{ij}}, \sigma^2_{\tau_i}) I\{\tau_j > 0\}$, where

$$\mu_{\tau_{ij}} = \delta(\lambda_i)(y_j - \xi_i), \quad \sigma_{\tau_i} = \sigma_i \sqrt{1 - \delta^2(\lambda_i)}. \tag{10}$$

From (9), the complete-data log-likelihood function is

$$\ell_c(\boldsymbol{\theta}) = \sum_{j=1}^{n} \sum_{i=1}^{g} Z_{ij} \Bigg\{ \log(\omega_i) - \log(\sigma_i^2) - \frac{1}{2} \log\Big(1 - \delta^2(\lambda_i)\Big) - \frac{\tau_j^2 - 2\delta(\lambda_i)\tau_j(y_j - \xi_i) + (y_j - \xi_i)^2}{2\sigma_i^2\Big(1 - \delta^2(\lambda_i)\Big)} \Bigg\}. \tag{11}$$

Letting $\hat{z}_{ij} = E_{\hat{\boldsymbol{\Theta}}^{(k)}}(Z_{ij} \mid \boldsymbol{Y})$, $\hat{s}_{1ij} = E_{\hat{\boldsymbol{\Theta}}^{(k)}}(Z_{ij}\tau_j \mid \boldsymbol{Y})$ and $\hat{s}_{2ij} = E_{\hat{\boldsymbol{\Theta}}^{(k)}}(Z_{ij}\tau_j^2 \mid \boldsymbol{Y})$ be the necessary conditional expectations of (11), we obtain

$$\hat{z}_{ij}^{(k)} = \frac{\hat{\omega}_i^{(k)} \psi(y_j \mid \hat{\xi}_i^{(k)}, \hat{\sigma}_i^{2(k)}, \hat{\lambda}_i^{(k)})}{\sum_{m=1}^{g} \hat{\omega}_m^{(k)} \psi(y_j \mid \hat{\xi}_m^{(k)}, \hat{\sigma}_m^{2(k)}, \hat{\lambda}_m^{(k)})}, \tag{12}$$

$$\hat{s}_{1ij}^{(k)} = \hat{z}_{ij}^{(k)} \left[\hat{\mu}_{\tau_{ij}}^{(k)} + \hat{\sigma}_{\tau_i}^{(k)} \frac{\phi\Big\{ \hat{\lambda}^{(k)} \Big(\frac{y_j - \hat{\xi}_i^{(k)}}{\hat{\sigma}_i^{(k)}} \Big) \Big\}}{\Phi\Big\{ \hat{\lambda}_i^{(k)} \Big(\frac{y_j - \hat{\xi}_i^{(k)}}{\hat{\sigma}_i^{(k)}} \Big) \Big\}} \right], \tag{13}$$

$$\hat{s}_{2ij}^{(k)} = \hat{z}_{ij}^{(k)} \left[\hat{\mu}_{\tau_{ij}}^{(k)2} + \hat{\sigma}_{\tau_i}^{(k)2} + \frac{\phi\Big\{ \hat{\lambda}^{(k)} \Big(\frac{y_j - \hat{\xi}_i^{(k)}}{\hat{\sigma}_i^{(k)}} \Big) \Big\}}{\Phi\Big\{ \hat{\lambda}_i^{(k)} \Big(\frac{y_j - \hat{\xi}_i^{(k)}}{\hat{\sigma}_i^{(k)}} \Big) \Big\}} \hat{\mu}_{\tau_{ij}}^{(k)} \hat{\sigma}_{\tau_i}^{(k)} \right], \tag{14}$$

where $\hat{\mu}_{\tau_{ij}}^{(k)}$, $\hat{\sigma}_{\tau_i}^{(k)}$ are $\mu_{\tau_{ij}}$ and σ_{τ_i} in (10) with ξ, σ and λ replaced by $\hat{\xi}^{(k)}$, $\hat{\sigma}^{(k)}$ and $\hat{\lambda}^{(k)}$, respectively.

The ECM algorithm is as follows.

E-step: Given $\boldsymbol{\Theta} = \hat{\boldsymbol{\Theta}}^{(k)}$, compute $\hat{z}_{ij}^{(k)}$, $\hat{s}_{1ij}^{(k)}$ and $\hat{s}_{2ij}^{(k)}$ for $i = 1, \ldots, g$ and $j = 1, \ldots, n$, using (12), (13) and (14).

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

CM-step 2: Calculate

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

CM-step 3: Calculate

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

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

$$\hat{\sigma}_i^{2(k+1)}\delta(\lambda_i)\big(1-\delta^2(\lambda_i)\big)\sum_{j=1}^n \hat{z}_{ij}^{(k)} + \big(1+\delta^2(\lambda_i)\big)\sum_{j=1}^n (y_j - \hat{\xi}_i^{(k+1)})\hat{s}_{1ij}^{(k)}$$
$$-\delta(\lambda_i)\sum_{j=1}^n \hat{s}_{2ij}^{(k)} - \delta(\lambda_i)\sum_{j=1}^n \hat{z}_{ij}^{(k)}(y_j - \hat{\xi}_i^{(k+1)})^2 = 0.$$

ECME is identical to ECM except for the CM-Step 4 of ECM, which can be modified by the following CML-Step.

CML-step: Let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_g)$, and update $\hat{\boldsymbol{\lambda}}^{(k)}$ to

$$\hat{\boldsymbol{\lambda}}^{(k+1)} = \underset{\lambda_1,\ldots,\lambda_g}{\text{argmax}} \sum_{j=1}^n \log\bigg(\sum_{i=1}^g \hat{\omega}_i^{(k+1)}\psi(y_j \mid \hat{\xi}_i^{(k+1)},\ \hat{\sigma}_i^{2(k+1)},\ \lambda_i)\bigg).$$

We remark here that if the skewness parameters $\lambda_1, \ldots, \lambda_g$ are assumed to be identical, we use ECME since it is more efficient than ECM. Otherwise, the CML-step becomes a non-trivial high dimensional optimization problem, while using the CM-step 4 can avoid the complication.

3.3. Standard errors

We let $\boldsymbol{I}_o(\boldsymbol{\Theta} \mid \boldsymbol{y}) = -\partial^2\ell(\boldsymbol{\Theta} \mid \boldsymbol{Y})/\partial\boldsymbol{\Theta}\partial\boldsymbol{\Theta}^{\mathrm{T}}$ be the observed information matrix for the mixture model (7). Under some regularity conditions, the covariance matrix of ML estimates $\hat{\boldsymbol{\Theta}}$ can be approximated by the inverse of $\boldsymbol{I}_o(\hat{\boldsymbol{\Theta}} \mid \boldsymbol{y})$. We follow Basford, Greenway, McLachlan and Peel (1997) to evaluate

$$\boldsymbol{I}_o(\hat{\boldsymbol{\Theta}} \mid \boldsymbol{y}) = \sum_{j=1}^n \hat{\boldsymbol{s}}_j \hat{\boldsymbol{s}}_j^{\mathrm{T}}, \tag{15}$$

where $\hat{\boldsymbol{s}}_j = \partial \log\Big\{\sum_{i=1}^g \omega_i\psi(y_j \mid \xi_i, \sigma_i^2, \lambda_i)\Big\}/\partial\boldsymbol{\Theta}\big|_{\boldsymbol{\Theta}=\hat{\boldsymbol{\Theta}}}$.

Corresponding to the vector of all $4g-1$ unknown parameters in $\boldsymbol{\Theta}$, we partition $\hat{\boldsymbol{s}}_j$ $(j = 1, \ldots, n)$ as

$$\hat{\boldsymbol{s}}_j = (\hat{s}_{j,\omega_1}, \ldots, \hat{s}_{j,\omega_{g-1}}, \hat{s}_{j,\xi_1}, \ldots, \hat{s}_{j,\xi_g}, \hat{s}_{j,\sigma_1}, \ldots, \hat{s}_{j,\sigma_g}, \hat{s}_{j,\lambda_1}, \ldots, \hat{s}_{j,\lambda_g})^{\mathrm{T}}.$$

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

$$\hat{s}_{j,\omega_r} = \frac{\psi(y_j \mid \hat{\xi}_r, \hat{\sigma}_r^2, \hat{\lambda}_r) - \psi(y_j \mid \hat{\xi}_g, \hat{\sigma}_g^2, \hat{\lambda}_g)}{\sum_{i=1}^g \hat{\omega}_i \psi(y_j \mid \hat{\xi}_i, \hat{\sigma}_i^2, \hat{\lambda}_i)} \quad (r = 1, \ldots, g-1),$$

$$\hat{s}_{j,\xi_r} = \frac{2\hat{\omega}_r \phi\{\frac{y_j - \hat{\xi}_r}{\hat{\sigma}_r}\}}{\hat{\sigma}_r^2 \sum_{i=1}^g \hat{\omega}_i \psi(y_j \mid \hat{\xi}_i, \hat{\sigma}_i^2, \hat{\lambda}_i)} \left\{ \left(\frac{y_j - \hat{\xi}_r}{\hat{\sigma}_r}\right) \Phi\left(\hat{\lambda}_r \frac{y_j - \hat{\xi}_r}{\hat{\sigma}_r}\right) - \hat{\lambda}_r \phi\left(\hat{\lambda}_r \frac{y_j - \hat{\xi}_r}{\hat{\sigma}_r}\right) \right\} \quad (r = 1, \ldots, g),$$

$$\hat{s}_{j,\sigma_r} = \frac{\hat{\omega}_r \psi(y_j \mid \hat{\xi}_r, \hat{\sigma}_r^2, \hat{\lambda}_r)}{\sum_{i=1}^g \hat{\omega}_i \psi(y_j \mid \hat{\xi}_i, \hat{\sigma}_i^2, \hat{\lambda}_i)} \left\{ -\frac{1}{\hat{\sigma}_r} + \frac{(y_j - \hat{\xi}_r)^2}{\hat{\sigma}_r^3} \right\} - \frac{2\hat{\omega}_r \hat{\lambda}_r (y_j - \hat{\xi}_r) \phi(\frac{y_j - \hat{\xi}_r}{\hat{\sigma}_r}) \phi(\frac{\hat{\lambda}_r (y_j - \hat{\xi}_r)}{\hat{\sigma}_r})}{\hat{\sigma}_r^3 \sum_{i=1}^g \hat{\omega}_i \psi(y_j \mid \hat{\xi}_i, \hat{\sigma}_i^2, \hat{\lambda}_i)} \quad (r = 1, \ldots, g),$$

$$\hat{s}_{j,\lambda_r} = \frac{\hat{\omega}_r \psi(y_j \mid \hat{\xi}_r, \hat{\sigma}_r^2, \hat{\lambda}_r)}{\sum_{i=1}^g \hat{\omega}_i \psi(y_j \mid \hat{\xi}_i, \hat{\sigma}_i^2, \hat{\lambda}_i)} \left(\frac{y_j - \hat{\xi}_r}{\hat{\sigma}_r}\right) \frac{\phi\left\{\frac{\hat{\lambda}_r (y_j - \hat{\xi}_r)}{\hat{\sigma}_r}\right\}}{\Phi\left\{\frac{\hat{\lambda}_r (y_j - \hat{\xi}_r)}{\hat{\sigma}_r}\right\}} \quad (r = 1, \ldots, g).$$

The information-based approximation (15) is asymptotically applicable. However, it may not be reliable unless the sample size is large. It is common in practice to perform the bootstrap approach (Efron and Tibshirani (1986)) for obtaining an alternative estimate of the covariance matrix for $\hat{\boldsymbol{\Theta}}$. The bootstrap method may provide more accurate standard error estimates than (15), but, it requires enormous computing power.

3.4. Notes on implementation

In the mixture context, the log-likelihood function may have multiple modes. A convenient way to circumvent such limitations is to try several EM iterations with a variety of starting values that are representatives of the parameter space. If there exist several modes, one can find the global mode by comparing their relative masses and log-likelihood values. In particular, the algorithm running with different starting values can be used to assess the stability of the resulting estimates.

Although the EM-type algorithm tends to be robust with respect to the choice of the starting values, it may not converge when initial values are far from optimum. The following outlines a simple procedure to achieve a set of reasonable initial values. (a) Randomly generate a set of B bootstrap resampling samples $\boldsymbol{y}_1^*, \ldots, \boldsymbol{y}_B^*$ from the original data $\boldsymbol{y}$. (b) For each bootstrap sample, partition them into g components using the K-means clustering algorithm and compute the initial values $\hat{w}_i^{(0)} = \sum_{j=1}^n Z_{ij}^{(0)}/n$. (c) For each partitioned component, compute the initial values $\hat{\xi}_i^{(0)}$, $\hat{\sigma}_i^{2(0)}$ and $\hat{\delta}_i^{(0)}(\lambda_i^{(0)})$ using the method of moments as in (3).

4. Bayesian Modelling For Skew Normal Mixtures

4.1. The prior distributions and posterior MCMC sampling

We consider a Bayesian approach to (7) in which $\mathbf{\Theta}$ is regarded as random with a prior distribution that reflects our degree of belief in different values of these quantities. Since fully non-informative prior distributions are not permissible in the mixture context, the prior distributions chosen are weakly informative subject to vague prior knowledge and this avoids nonintegrable posterior distributions. The prior distributions for model (7) takes

$$\begin{aligned} \xi_i &\sim N(\eta, \kappa^{-1}) && (i = 1, \ldots, g), \\ \sigma_i^{-2} \mid \beta &\sim \Gamma(\alpha, \beta) && (i = 1, \ldots, g), \\ \beta &\sim \Gamma(\nu_1, \nu_2), \\ \delta(\lambda_i) &\sim U(-1, 1) && (i = 1, \ldots, g), \\ \boldsymbol{\omega} &\sim D(h, \ldots, h), \end{aligned}$$

where β is an unknown hyperparameter, $(\eta, \kappa, \alpha, \nu_1, \nu_2, h)$ are known (data-dependent) constants, $\Gamma(\alpha, \beta)$ denotes the gamma distribution with mean α/β and variance α/β^2, $U(-1, 1)$ denotes the continuous uniform distribution on the interval $[-1, 1]$, and $D(h, \ldots, h)$ stands for the Dirichlet distribution with the density function

$$\frac{\Gamma(gh)}{\Gamma(h)^g} \omega_1^{h-1} \cdots \omega_{g-1}^{h-1} \Big(1 - \sum_{i=1}^{g-1} \omega_i\Big)^{h-1}.$$

For the values of $(\eta, \kappa, \alpha, \nu_1, \nu_2, h)$, we follow Richardson and Green (1997) in letting η equal the midpoint of the observed interval and $\kappa^{-1} = R^2$, where R is the range of the interval, and in setting $\alpha = 2$, $\nu_1 = 0.2$, $\nu_2 = 100\alpha/(\alpha R^2)$ and $h = 2$.

Given $\mathbf{\Theta} = \mathbf{\Theta}^{(k)}$, the MCMC sampling scheme at the $(k + 1)$st iteration consists of the following steps.

Step 1. Sample $\boldsymbol{Z}_j^{(k+1)}$ $(j = 1, \ldots, n)$ from $\mathcal{M}(1;\ \omega_1^*, \ldots, \omega_g^*)$, where

$$\omega_i^* = \frac{\psi(y_j \mid \xi_i^{(k)}, \sigma_i^{2(k)}, \lambda_i^{(k)})}{\sum_{m=1}^g \omega_m^{(k)} \psi(y_j \mid \xi_m^{(k)}, \sigma_m^{2(k)}, \lambda_m^{(k)})} \quad (i = 1, \ldots, g).$$

Step 2. Given $Z_{ij} = 1$, sample $\tau_j^{(k+1)}$ $(j = 1, \ldots, n)$ from

$$TN\Big(\delta(\lambda_i^{(k)})(y_j - \xi_i^{(k)}),\ \sigma_i^{2(k)}\big(1 - \delta^2(\lambda_i^{(k)})\big)\Big) I\{\tau_j > 0\}.$$

Step 3. Sample $\beta^{(k+1)}$ from $\Gamma(\nu_1 + g\alpha,\ \nu_2 + \sum_{i=1}^{g} \sigma_i^{-2^{(k)}})$.

Step 4. Sample $\boldsymbol{\omega}^{(k+1)}$ from $D(h + n_1^{(k+1)}, \ldots, h + n_g^{(k+1)})$, where $n_i^{(k+1)} = \sum_{j=1}^{n} Z_{ij}^{(k+1)}$.

Step 5. Given $Z_{ij} = 1$, sample $\xi_i^{(k+1)}$ from

$$\mathrm{N}\left(\mu_{\xi_i}^{(k+1)},\ \left\{\frac{n_i^{(k+1)}}{\sigma_i^{2^{(k)}}\big(1-\delta^2(\lambda_i^{(k)})\big)} + \kappa\right\}^{-1}\right),$$

where

$$\mu_{\xi_i}^{(k+1)} = \frac{\sum_{j=1}^{n} Z_{ij}^{(k+1)} y_j - \delta(\lambda_i^{(k)}) \sum_{j=1}^{n} Z_{ij}^{(k+1)} \tau_j^{(k+1)} + \kappa\eta\sigma_i^{2^{(k)}}\big(1-\delta^2(\lambda_i^{(k)})\big)}{n_i^{(k+1)} + \kappa\sigma_i^{2^{(k)}}\big(1-\delta^2(\lambda_i^{(k)})\big)}.$$

Step 6. Given $Z_{ij} = 1$, sample $\sigma_i^{-2^{(k+1)}}$ from $\Gamma\big(\alpha + n_i^{(k+1)},\ \beta^{(k+1)} + b\big)$, where

$$b = \frac{1}{2\big(1-\delta^2(\lambda_i^{(k)})\big)} \Big\{ \sum_{j=1}^{n} Z_{ij}^{(k+1)} \tau_j^{2^{(k+1)}} - 2\delta(\lambda_i^{(k)}) \sum_{j=1}^{n} Z_{ij}^{(k+1)} \tau_j^{(k+1)} (y_j - \xi_i^{(k+1)}) \\ + \sum_{j=1}^{n} Z_{ij}^{(k+1)} (y_j - \xi_i^{(k+1)})^2 \Big\}.$$

Step 7. Sample $\boldsymbol{\delta}^{(k+1)} = \big(\delta(\lambda_1^{(k+1)}), \ldots, \delta(\lambda_g^{(k+1)})\big)$ via the Metropolis Hastings (M-H) algorithm (Hastings (1970)) from

$$f(\boldsymbol{\delta}) \propto \prod_{i=1}^{g} \prod_{j=1}^{n} \Bigg[\big(1-\delta^2(\lambda_i)\big)^{-\frac{1}{2}} \\ \times \exp\left\{ \frac{\tau_j^{(k+1)^2} - 2\delta(\lambda_i)\tau_j^{(k+1)}(y_j - \xi_i^{(k+1)}) + (y_j - \xi_i^{(k+1)})^2}{2\sigma_i^{2^{(k+1)}}\big(1-\delta^2(\lambda_i)\big)} \right\} \Bigg]^{Z_{ij}^{(k+1)}}.$$

To elaborate on Step 7 of the above algorithm, we transform $\delta(\lambda_i)$ to $\delta^*(\lambda_i) = \log\big\{\big(1+\delta(\lambda_i)\big)/\big(1-\delta(\lambda_i)\big)\big\}$ and then apply the M-H algorithm to $g\big(\boldsymbol{\delta}^*\big) = f\big(\boldsymbol{\delta}(\boldsymbol{\delta}^*)\big) \prod_{i=1}^{g} J_{\delta^*(\lambda_i)}$, where $\boldsymbol{\delta}^* = \big(\delta^*(\lambda_1), \ldots, \delta^*(\lambda_g)\big)$, and $J_{\delta^*(\lambda_i)} = 2e^{\delta^*(\lambda_i)}/\big(1+e^{\delta^*(\lambda_i)}\big)^2$ is the Jacobian of transformation from $\delta(\lambda_i)$ to $\delta^*(\lambda_i)$. A g-dimensional multivariate normal distribution with mean $\boldsymbol{\delta}^{*^{(k)}}$ and covariance matrix $c^2\boldsymbol{\Sigma}_{\boldsymbol{\delta}^*}^{(k)}$ is chosen as the proposal distribution, where the scale $c \approx 2.4/\sqrt{g}$, as suggested in Gelman, Robert and Gilks (1996). The value of $\boldsymbol{\Sigma}_{\boldsymbol{\delta}^*}^{(k)}$ can be estimated by the inverted sample information matrix given $\boldsymbol{y}$ and $\boldsymbol{\Theta} = \boldsymbol{\Theta}^{(k)}$. Having obtained $\boldsymbol{\delta}^*$ from the M-H algorithm, we transform it back to $\boldsymbol{\delta}$ by $\delta(\lambda_i) =$

$(e^{\delta^*(\lambda_i)}-1)/(e^{\delta^*(\lambda_i)}+1)$ $(i=1,\ldots,g)$, and then transform $\delta(\lambda_i)$ back to λ_i by $\delta(\lambda_i)/\sqrt{1-\delta^2(\lambda_i)}$. To avoid the label-switching problem and slow stabilization of the Markov chain, our initial values $\boldsymbol{\Theta}^{(0)}$ are chosen to be dispersed around the ML estimates with the restriction $\xi_1^{(0)} < \cdots < \xi_g^{(0)}$.

4.2. Convergence assessment using multiple chains

Before conducting inference using MCMC samples, the output should be analyzed to determine the required run length of MCMC sequences. Gelman and Rubin (1992) proposed a convergence diagnostic $\hat{\mathcal{R}}$, the *potential scale reduction factor* (PSRF), obtained by running multiple chains with overdispersed starting values. However, the approach is essentially univariate. Recently, Brooks and Gelman (1998) provided a generalization of Gelman and Rubin's method that consider several parameters simultaneously.

Suppose there are I independent parallel chains and the length of each chain is $2n$. Let $\boldsymbol{\theta}$ denote a $p\times 1$ vector of parameters and $\boldsymbol{\theta}_i = (\boldsymbol{\theta}_i^{(1)},\ldots,\boldsymbol{\theta}_i^{(n)})$ denote the simulation sample of the ith chain $(i=1,\ldots,I)$, after discarding the first n iterations. Brooks and Gelman (1998) stated that the posterior variance-covariance matrix of $\boldsymbol{\theta}$ can be estimated by

$$\hat{\boldsymbol{V}} = \frac{n-1}{n}\boldsymbol{W} + \Big(1+\frac{1}{I}\Big)\frac{\boldsymbol{B}}{n},$$

where $\boldsymbol{W}$ and $\boldsymbol{B}/n$ denote the within and between-sequence sample covariance matrix estimates of $(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_I)$, respectively.

They then proposed the *multivariate potential scale reduction factor* (MPSRF), $\hat{\mathcal{R}}^p = (n-1)/n + (1+1/I)\lambda_1$, where λ_1 is the largest eigenvalue of $\boldsymbol{W}^{-1}\boldsymbol{B}/n$. Note that the multivariate measure $\hat{\mathcal{R}}^p$ bounds above the univariate $\hat{\mathcal{R}}$ values over all p variables.

Suppose the I parallel chains are mixing well within the model, $\hat{\mathcal{R}}^p$ will decline to 1 for reasonably large n. Meanwhile, if the I parallel chains are essentially overlapping, then the determinants of $\hat{\boldsymbol{V}}$ and $\boldsymbol{W}$ should stabilize over the iterations and be sufficiently close.

5. Examples

5.1. The enzyme data

We first carry out our methodology for the enzyme data set with $n = 245$ observations. The data were first analyzed by Bechtel, Bonaita-Pellieé, Poisson, Magnette and Bechtel (1993), who identified a mixture of skew distributions by the maximum likelihood techniques of Maclean, Morton, Elston and Yee (1976). Richardson and Green (1997) provided the reversible jump MCMC approach for the univariate normal mixture models with an unknown number of components and identified the most possible values of g to be between 3 and 5.

Table 1. Estimated parameter values and the corresponding standard errors (SE) for model (16) with the enzyme data.

	ω	ξ_1	ξ_2	σ_1	σ_2	λ_1	λ_2
Estimate	0.6240	0.0949	0.7802	0.1331	0.7150	3.2780	6.6684
SE	0.0310	0.0107	0.0516	0.0109	0.0607	0.9467	3.9640

We fit the following two-component SNMIX model to the data

$$f(y) = \omega\psi(y|\xi_1, \sigma_1^2, \lambda_1) + (1-\omega)\psi(y|\xi_2, \sigma_2^2, \lambda_2). \tag{16}$$

The ECM algorithm was run with 100 starting values and was checked for convergence. All EM iterations under different stating values converge to the same stationary point with log-likelihood -41.92. The resulting ML estimates and the corresponding standard errors are listed in Table 1. In this table, we found that the standard error for λ_2 is relatively large. This is due to the fact that the log-likelihood function can be fairly flat near the ML estimates of the shape parameter of the skew normal components. We have shown this by plotting the profile log-likelihood function of (λ_1, λ_2) in Figure 1.

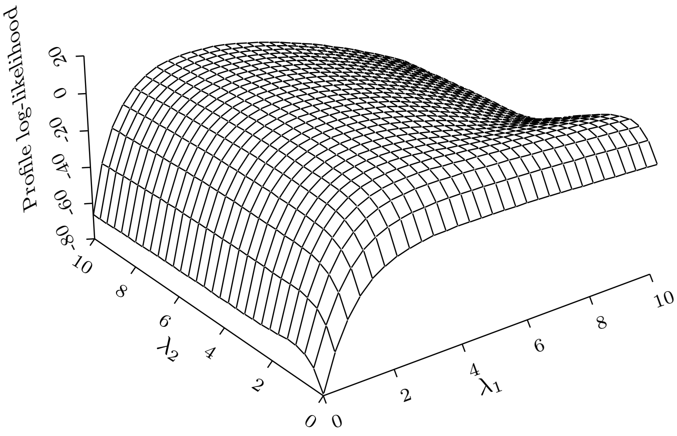

Figure 1. Plot of the profile log-likelihood for λ_1 and λ_2 for the enzyme data.

For comparison purposes, we also fit a NORMIX model ($\lambda_1 = \lambda_2 = 0$) with $g = 2 - 5$ components. The log-likelihood maximum and two information-based criteria, AIC (Akaike (1973)) and BIC (Schwarz (1978)), are displayed in the third to fifth columns of Table 2. Apparently, the fitted two-component SNMIX model is superior to the fitted NORMIX model, since it has the largest log-likelihood and the smallest AIC and BIC. The last two columns of this table present the required number of EM iterations and the associated rate of convergence, r, which is assessed in practice as

$$r = \lim_{t\to\infty} \frac{\|\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}\|}{\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t-1)}\|}.$$

A relative tolerance of 10^{-8} for the estimates of all parameters in the model was used as the convergence criterion. We note that the reported rate of convergence depends on the fraction of missing information and the greater the value of r implies the slower the convergence, see Meng (1994). In this example, we also note that the estimating procedure for fitting SNMIX model does not converge properly for $g \geq 3$.

Table 2. Comparison of log-likelihood maximum, AIC and BIC for fitted SNMIX and NORMIX models using the enzyme data. The number of parameters and the rate of convergence are denoted by m and r, respectively.

Model	g	m	log-likelihood	AIC†	BIC‡	Iterations	r
SNMIX	2	7	−41.92	97.84	122.35	175	0.82
NORMIX	2	5	−54.64	119.28	136.79	19	0.53
NORMIX	3	8	−47.83	111.66	139.67	170	0.81
NORMIX	4	11	−46.75	115.50	154.01	425	0.89
NORMIX	5	14	−46.26	120.52	169.54	562	0.93
NORMIX	≥ 6			> 123	> 185		

†AIC=-2(log-likelihood$-m$); ‡BIC=$-2\{$log-likelihood$-0.5m\log(n)\}$.

5.2. The faithful data

As another example, we consider the Old Faithful Geyser data taken from Silverman (1986). It consists of 272 eruption lengths (in minutes) of the Old Faithful Geyser in Yellowstone National Park, Wyoming, USA. The data appear to be bimodal with asymmetrical components. We fit a two-component SNMIX model (16) by analogy with the previous example. The ML estimates and the corresponding standard errors are reported in the second and third columns of Table 3, respectively. We carry out an MCMC simulation by running 10,000 iterations of ten independent parallel chains with different starting values for each chain *over-dispersed* around ±3 standard deviations of the ML estimates. The convergence of MCMC samplers is monitored by examining $\hat{\mathcal{R}}^p$ values as discussed in Section 4.2. The monitored values of $\hat{\mathcal{R}}^p$ and the determinants of $\hat{\boldsymbol{V}}$ and $\boldsymbol{W}$ are plotted in Figures 2(a) and 2(b), respectively. By examining both figures, convergence occurs around 4,000 iterations. Having obtained the remaining converged MCMC simulation samples, we computed the posterior mean, standard deviation, median and 95% posterior interval (2.5% and 97.5% posterior quantiles), which are listed in the 4th-8th columns of Table 3.

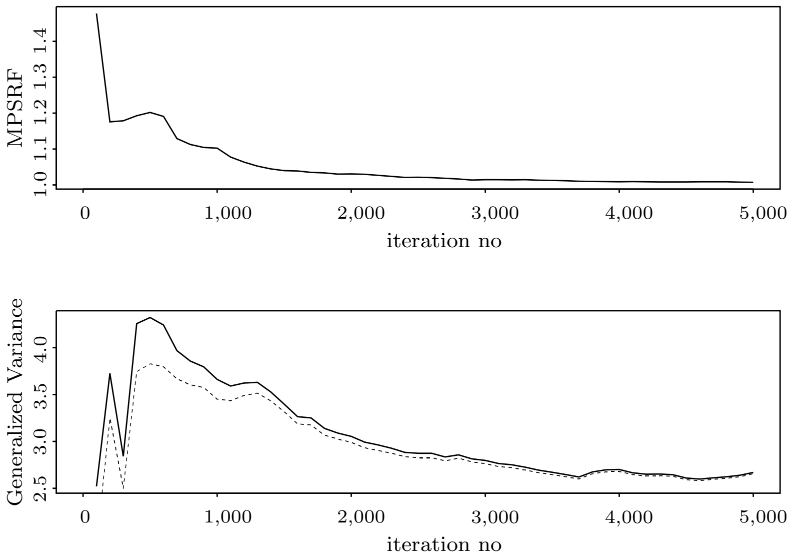

Figure 2. (a) Plot of MPSRF, $\hat{\mathfrak{R}}^p$; (b) Plot of the determinants ($\times 10^{13}$) of $\hat{\boldsymbol{V}}$ (solid) and $\boldsymbol{W}$ (dashed).

Table 3. ML estimation results and MCMC summary statistics for the parameters of model (16) with the faithful data.

Parameter	ML		MCMC				
	Estimate	SE	Mean	SE	Median	2.5%	97.5%
ω	0.3487	0.0294	0.3510	0.0294	0.3506	0.2948	0.4114
ξ_1	1.7267	0.0291	1.7225	0.0238	1.7232	1.6752	1.7690
ξ_2	4.8026	0.0511	4.7847	0.0660	4.7919	4.6427	4.8940
σ_1	0.3801	0.0415	0.3959	0.0418	0.3928	0.3211	0.4854
σ_2	0.6857	0.0621	0.6712	0.0675	0.6725	0.5381	0.8025
λ_1	5.8026	2.1436	6.2316	2.1176	5.8768	3.1025	11.2305
λ_2	-3.4951	1.1492	-3.4073	1.1704	-3.2700	-5.9843	-1.5502

Figure 3 displays the histograms of the posterior samples of the model parameters. It is evident that the shape of the posterior distribution of λ_1 is skewed to the right, while the shape of the posterior distribution of λ_2 is skewed to the left. It is interesting to note that the posterior distributions of the parameters (λ_1, λ_2), which regulate the skewness, are skewed as well.

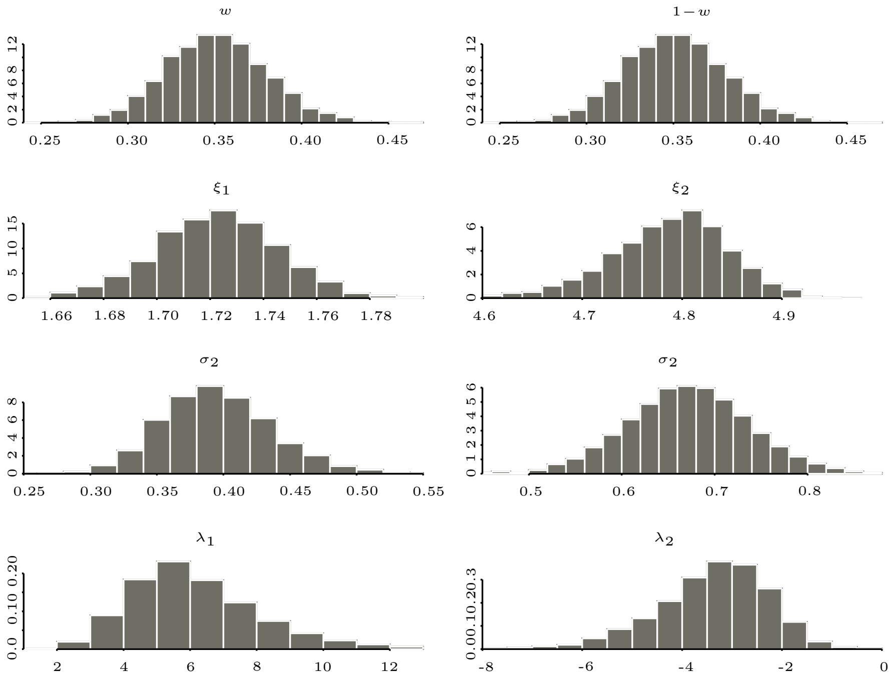

Figure 3. Histograms of the posterior sample of the SNMIX parameters for the faithful data.

Finally, it is interesting to compare the density estimation of NORMIX and SNMIX fitting results. The ML-fitted NORMIX and SNMIX densities, together with the Bayesian predictive SNMIX density, are superimposed in Figure 4(a). Subsequently, the fitted cumulative density functions (CDFs) and the empirical CDF are shown in Figure 4(b). Based on the graphical visualization, the resulting ML-fitted SNMIX density, as well as the Bayesian predictive SNMIX density, are more suitable than the ML-fitted NORMIX density for this data set. Furthermore, the fitted SNMIX CDFs more closely track the empirical CDF than does the fitted NORMIX CDF.

6. Concluding Remarks

In our examples, it is quite appealing that the skew normal mixtures can provide a more appropriate density estimation than normal mixtures based on information-based criteria and graphical visualization. There are a number of possible extensions of the current work. Mixture modelling using the multivariate skew normal distribution (e.g., Azzalini and Dalla Valle (1996), Shau, Dey and Branco (2003) and Gupta, González-Farías and Domínguez-Monila (2004)) is the most natural extension and will be reported in a follow-up paper. In addition, it would be a worthwhile task to model the number of components, g, and component parameters, $\boldsymbol{\Theta}$, jointly. For modelling both skewness and long tails in a mixture context, component densities using the skew t distribution (e.g., Jones and Faddy (2003), Azzalini and Capitaino (2003) and Lin, Lee and Hsieh (2007)) is a feasible choice and awaits further investigation.

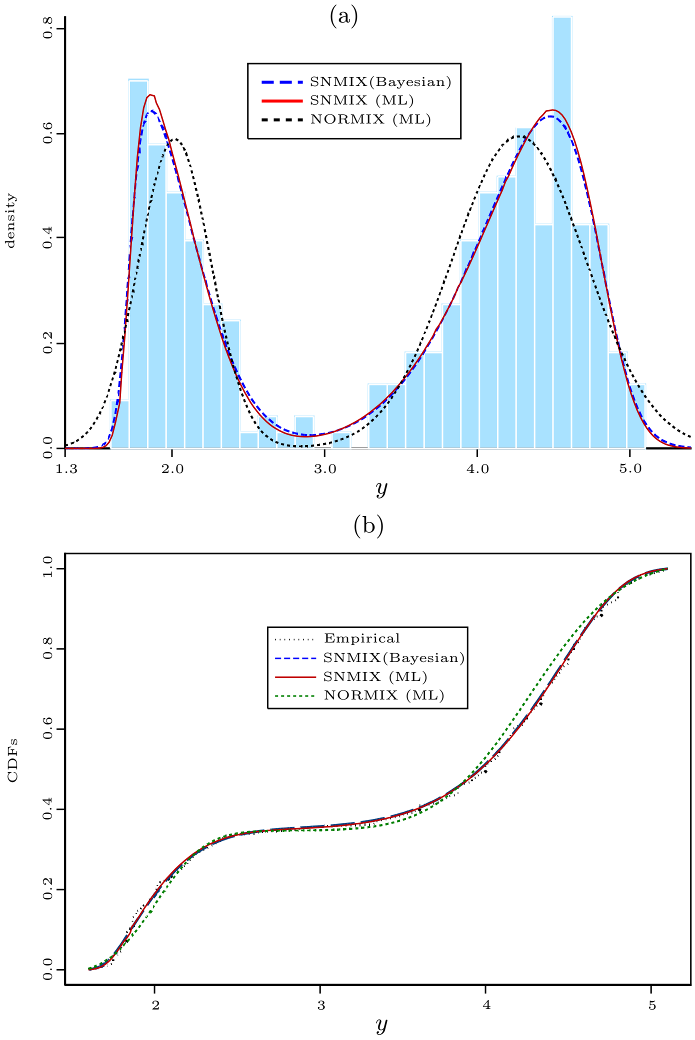

Figure 4. (a) Histogram of the faithful data overlaid with densities based on two fitted two-component SNMIX (ML and Bayesian), and a ML-fitted two-component NORMIX; (b) Empirical CDF of the faithful data overlaid with CDFs based on two fitted two-component SNMIX (ML and Bayesian) and a ML-fitted two-component NORMIX.

Acknowledgement

We gratefully acknowledge the Chair Co-Editor, an associate editor, and one referee for their valuable comments, which substantially improved the quality of the paper. This research was supported by the National Science Council of Taiwan.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In 2*nd Int. Symp. on Information Theory*, (Edited by B. N. Petrov and F. Csaki), 267-281. Akademiai Kiado, Budapest.

Arnold, B. C., Beaver, R. J., Groeneveld, R. A. and Meeker, W. Q. (1993). The nontruncated marginal of a truncated bivariate normal distribution. *Psychometrika* **58**, 471-488.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scand. J. Statist.* **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. (1999). Statistical applications of the multivariate skew-normal distribution. *J. Roy. Statist. Soc. Ser. B* **61**, 579-602.

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

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

Basord, K. E., Greenway D. R., McLachlan G. J. and Peel D. (1997). Standard errors of fitted means under normal mixture. *Comput. Statist.* **12**, 1-17.

Bechtel, Y. C., Bonaiti-Pellieé, C., Poisson, N., Magnette, J. and Bechtel, P. R. (1993). A population and family study of N-acetyltransferase using caffeine urinary metabolites. *Clin. Pharm. Therp.* **54**, 134-141.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Statist.* **7**, 434-455.

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

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *J. Roy. Statist. Soc. Ser. B* **56**, 363-375.

Efron B. and Tibshirani R. (1986). Bootstrap method for standard errors, confidence intervals, and other measures of statistical accuracy. *Statist. Sci.* **1**, 54-77.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90**, 577-588.

Gelman, A., Robert, G. and Gilks, W. (1996). Efficient Metropolis jumping rules. In *Bayesian Statistics* **5** (Edited by J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith). Oxford University Press, New York.

Gelman A. and Rubin D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist, Sci.* **7**, 457-472.

Genton, M. G. (2004). *Skew-Elliptical Distributions and Their Applications.* Chapman & Hall, New York.

Gupta, A. K., González-Farías G. and Domínguez-Monila, J. A. (2004). A multivariate skew normal distribution. *J. Multivariate Anal.* **89**, 181-190.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97-109.

Henze, N. (1986). A probabilistic representation of the "skew-normal" distribution. *Scand. J. Statist.* **13**, 271-275.

Jones, M. C. and Faddy, M. J. (2003). A skew extension of the t-distribution, with applications. *J. Roy. Statist. Soc. Ser. B* **65**, 159-174.

Lin, T. I., Lee, J. C. and Hsieh, W. J. (2007). Robust mixture modeling using the skew t distribution. *Statist. Comput.* **17**, 81-92.

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.

Maclean, C. J., Morton, N. E., Elston, R. C. and Yee, S. (1976). Skewness in commingled distributions. *Biometrics* **32**, 695-599.

McLachlan, G. J. and Basord, 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**, 267-278.

Meng, X. L. (1994). On the global and componentwise rates of convergence of the EM algorithm. *Lin. Alg. Applic.* **199**, 413-425.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. R. Statist. Soc.* **B 59**, 731-792.

Roberts, C. and Geisser, S. (1966). A necessary and sufficient condition for the square of a random variable to be gamma. *Biometrika* **53**, 275-278.

Sahu, S. K., Dey, D. K. and Branco, M. D. (2003). A new class of multivariate skew distributions with application to Bayesian regression models. *Canad. J. Statist.* **31**, 129-150.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman & Hall, London.

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

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

Department of Applied Mathematics, National Chung Hsing University, Taichung 402, Taiwan.

E-mail: tilin@amath.nchu.edu.tw

Institute of Statistics and Graduate Institute of Finance, National Chiao Tung University, Hsinchu 300, Taiwan.

E-mail: jclee@stat.nctu.edu.tw

Institute of Statistics, National Chiao Tung University, Hsinchu 300, Taiwan.

E-mail: kelly.st92g@nctu.edu.tw

(Received November 2004; accepted November 2005)