國立交通大學

統計學研究所

碩士論文

混合的偏斜常態分佈及其應用

On the mixture of skew normal distributions and its applications

研究生：顏淑儀
指導教授：李昭勝 博士
林宗儀 博士

中華民國九十四年六月

混合的偏斜常態分佈及其應用

On the mixture of skew normal distributions and its applications

研究生：顏淑儀　　Student：Shu-Yi Yen
指導教授：李昭勝　　Advisor：Dr. Jack C. Lee
林宗儀　　Dr. Tsung I. Lin

國立交通大學
統計學研究所
碩士論文

A Thesis
Submitted to Institute of Statistics
College of Science
National Chiao Tung University
in Partial Fulfillment of the Requirements
for the Degree of
Master
in
Statistics
June 2005

Hsinchu, Taiwan, Republic of China
中華民國九十四年六月

混合偏斜的常態分佈及其應用

研究生：顏淑儀　　　指導教授：李昭勝 博士
林宗儀 博士

國立交通大學統計學研究所

摘要

混合的常態分佈對於來自不同來自母體的異質性資料提供了一種的自然模型架構。近二十年來， 偏斜的常態分佈對於處理非對稱的資料問題已被驗證是一種很有用的工具。 本文我們提出了以概似函數與貝氏抽樣為基礎之方法去處理混合偏斜常態分佈的問題。 我們將利用”期望最大值型式”(EM-type)演算法求最大概似估計值。

對於所提出的先驗分佈及所推導之後驗分佈的結果, 我們也運用馬可夫鏈蒙地卡羅發展出貝氏的計算方法。 最後我們透過兩個實例來闡述所提出模型之應用。

致謝

研究所兩年的生涯，在快樂、充實、努力的時刻中，匆匆的過去了，在面臨口試通過即將要畢業的時刻，中心百感交集，期待與不捨的情緒錯綜複雜，當然此刻心情充滿了感激及感動。

論文得以順利的完成，首先要感謝的就是我的指導教授 李昭勝老師，老師豐富的學識涵養和對於學生的關心和照顧，都讓我受益良多、心存感激，身為李老師的學生讓我覺得很驕傲也很開心；另一位對於這篇論文貢獻最大的是林宗儀學長，經由學長的指導和用心，才能順利的完成，在這裡要感謝學長不辭辛苦的指導；接著要感謝口試委員盧鴻興老師和林淑惠老師，特地抽空參加還給予我寶貴的意見。

接著要感謝我們這間研究室408所有的朋友，包括了怡娟、婉菁、揚波、翊琳、秀仁和如美，因為有妳們這一群朋友，爲我研究生的生涯帶來了歡笑與淚水，讓我倍感溫馨與感動，因為有妳們的扶持與鼓勵，才能讓我快樂的度過，尤其要特別感謝怡娟，因為有妳的陪伴與支持，讓我在寫論文的這段期間，產生了有福同享、有難同當的患難真情，我會永遠珍惜的；接著要感謝大學的朋友玉華，感謝妳總是會在我最需要幫助的時候，給我打氣、加油，讓我對自己有更多的信心與動力，更有勇氣面對挫折與挑戰。

最後我要感謝我的家人-父親 顏荊州、母親 羅錦秀、哥哥允健和弟弟義松，感謝他們給我一個溫馨又快樂的家庭，在背後默默的支持我、鼓勵我，讓我無後顧之憂的為自己的理想努力，得以順利完成論文；還要感謝我最重要的朋友-靖泓，從高中、大學和研究所，感謝你一路走來對我的包容與照顧，你總是默默的支持我，給我最大的信心與鼓勵，適時的給予我溫暖與幫助，因為有你，讓我一路有所依靠，才能有現在的我；在鳳凰花開、驪歌輕唱之際，謹以本文獻給所有的好友與至親，與你們分享我的喜悅。

顏淑儀　　謹誌於
交通大學統計學研究所
中華民國九十四年六月二十日

Contents

Contents

List of Tables

List of Figures

On the mixture of skew normal distributions and its applications

Student: Shu-Yi Yen

Advisors: Dr. Jack C. Lee
Dr. Tsung I. Lin

Institute of Statistics
National Chiao Tung University
Hsinchu, Taiwan

Abstract

The normal mixture model provides a natural framework for modelling the heterogeneity of a population arising from several groups. In the last two decades, the skew normal distribution has been shown to be useful for modelling asymmetric data in many applied problems. In this thesis, we propose likelihood-based and Bayesian sampling-based approaches to address the problem of modelling data by a mixture of skew normal distributions. EM-type algorithms are implemented for computing the maximum likelihood estimates. The prior as well as the resulting posterior distributions are developed for Bayesian computation via Markov chain Monte Carlo methods. Applications are illustrated through two real examples.

Key words: EM-type algorithms; Fisher information; Markov chain Monte Carlo; maximum likelihood estimation; skew normal mixtures

1. Introduction

Finite mixture models have been broadly developed with applications to classification, density estimation and pattern recognition problems, as discussed by Titterington, Smith and Markov (1985), McLachlan and Basford (1988), McLachlan and Peel (2000), and the references therein. Due to the advances of computational methods, in particular for Markov chain Monte Carlo (MCMC), many authors are also devoted to Bayesian mixture modelling issues, including Diebolt and Robert (1994), Ecobar and West (1995), Richardson and Green (1997) and Stephens (2000), among others.

In many applied problems, the shape of normal mixtures may be distorted and inferences may be misleading when the data involves highly asymmetric observations. In particular, the normal mixture model tends to "overfit" in that additional components are included to capture the skewness. Sometimes, increasing the number of pseudo-components may lead to difficulties and inefficiency in computations. Instead, we consider using the skew normal distributions proposed by Azzalini (1985) for mixture modelling to overcome the potential weakness in 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 for mixture modelling. 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 skew normal mixture model. For simplicity, we treat the number of components as known and carry out the maximum likelihood (ML) inferences via EM-type algorithms. In addition, Bayesian methods for skew normal mixtures are considered as an alternative technique. The specification of the priors and hyperparameters are chosen as weakly informative to avoid nonidentifiability problems in the mixture context.

The rest of the thesis unfolds as follows. Section 2 briefly outlines some preliminaries of the skew normal distribution. Azzalini and Capitanio (1999) point out that the ML estimates can be optionally improved by a few EM iterations, but detailed expressions of the EM algorithm are not available in the literature. We thus present how to compute the ML estimates for the skew normal distribution by using the ECM and ECME algorithms. In Section 3 we show the hierarchical formulation for skew normal mixture models by comprising two latent variables. Based on the model, we derive 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 conduct Bayesian inferences. In Section 5 two real examples are illustrated, 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 Y has the following density function:

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

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal density function and cumulative distribution function, respectively; then, for brevity, we shall say that $Y \sim SN(\xi, \sigma^2, \lambda)$. Note that if $\lambda = 0$, then the density of Y will be reduced to $N(\xi, \sigma^2)$ density. Figure 1 shows the plots of standard skew normal densities ($\xi = 0,\ \sigma = 1$) for various λ.

Lemma 1 *If $Y \sim SN(\xi, \sigma^2, \lambda)$ and $X \sim N\left(\xi,\ \frac{\sigma^2}{1+\lambda^2}\right)$, we have*

(i) $E(X^{n+1}) = \xi E(X^n) + \dfrac{\sigma^2}{1+\lambda^2}\dfrac{d}{d\xi}E(X^n).$

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

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

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

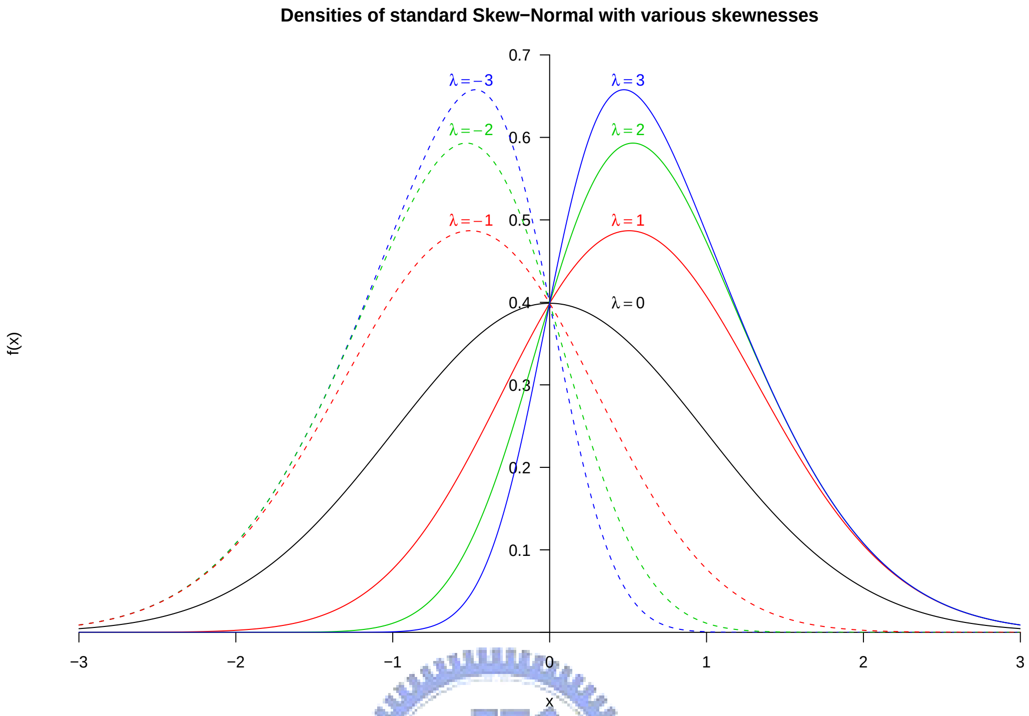

Figure 1: Standard skew normal densities for $\lambda = -3, -2, -1, 0, 1, 2, 3$.

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 lies in $(-0.9953,\ 0.9953)$ and κ_Y in $(3,\ 3.8692)$. Figure 2 displays γ_Y and κ_Y for different λ. Henze (1986) shows that the odd moments of the standard skew normal variable $Z = (Y - \xi)/\sigma$ have the following expressions:

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

while the even moments coincide with those of standard normal, as $Z^2 \sim \chi^2_1$.

From (2), Arnold, Beaver, Groeneveld and Meeker (1993) show the following

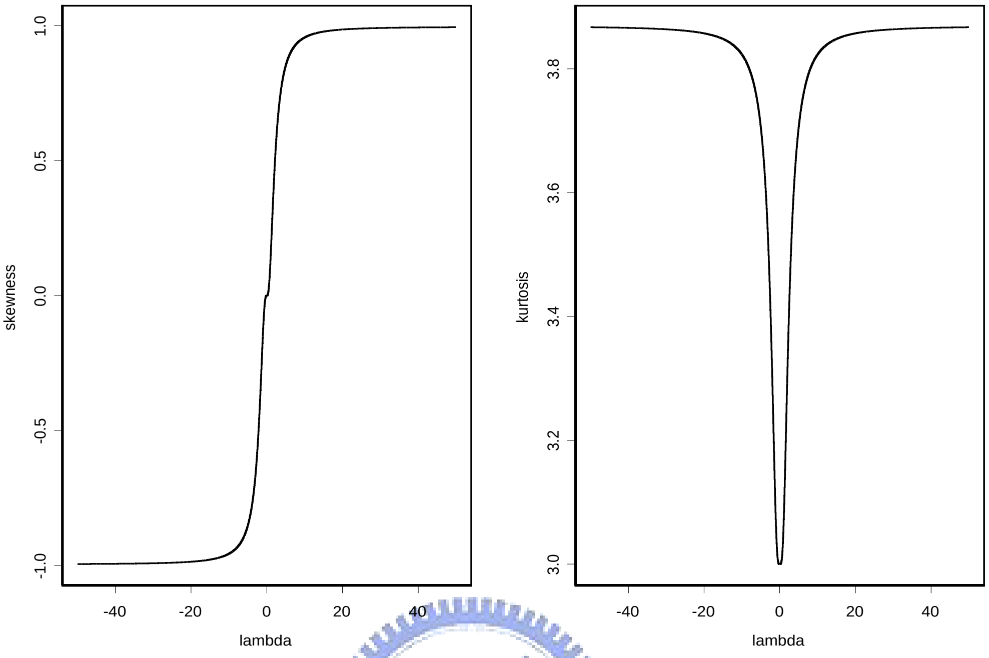

Figure 2: The skewness and kurtosis of the standard skew normal distribution.

method of moment estimators:

$$
\begin{aligned}
\tilde{\xi} &= m_1 - a_1 \left(\frac{m_3}{b_1}\right)^{1/3}, \\
\tilde{\sigma}^2 &= m_2 + a_1^2 \left(\frac{m_3}{b_1}\right)^{2/3}, \\
\tilde{\delta}(\lambda) &= \left\{a_1^2 + m_2 \left(\frac{b_1}{m_3}\right)^{2/3}\right\}^{-1/2}, \qquad (3)
\end{aligned}
$$

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

We show two faster 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 the ML estimation of the skew normal distribution. 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, Theorem 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

$$\tau_j \sim TN(0, \sigma^2)I\{\tau_j > 0\}, \quad U_j \sim N(0, \sigma^2),$$

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

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

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 following density function:*

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

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

(i) $M_X(t) = \exp\left(\mu t + \dfrac{\sigma^2 t^2}{2}\right)\left\{\dfrac{\Phi(\alpha_2 - \sigma t) - \Phi(\alpha_1 - \sigma t)}{\Phi(\alpha_2) - \Phi(\alpha_1)}\right\}.$

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

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

By Lemma 2, we have

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

Then we have the following ECM algorithm:

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

$$\begin{aligned}
\hat{s}_{1j}^{(k)} &= E_{\hat{\boldsymbol{\theta}}^{(k)}}(\tau_j|y_j) = \hat{\mu}_{\tau_j}^{(k)} + \frac{\phi\left\{\hat{\lambda}^{(k)}\left(\dfrac{y_j - \hat{\xi}^{(k)}}{\hat{\sigma}^{(k)}}\right)\right\}}{\Phi\left\{\hat{\lambda}^{(k)}\left(\dfrac{y_j - \hat{\xi}^{(k)}}{\hat{\sigma}^{(k)}}\right)\right\}}\hat{\sigma}_\tau^{(k)}, \\
\hat{s}_{2j}^{(k)} &= E_{\hat{\boldsymbol{\theta}}^{(k)}}(\tau_j^2|y_j) = \hat{\mu}_{\tau_j}^{(k)^2} + \hat{\sigma}_\tau^{(k)^2} + \frac{\phi\left\{\hat{\lambda}^{(k)}\left(\dfrac{y_j - \hat{\xi}^{(k)}}{\hat{\sigma}^{(k)}}\right)\right\}}{\Phi\left\{\hat{\lambda}^{(k)}\left(\dfrac{y_j - \hat{\xi}^{(k)}}{\hat{\sigma}^{(k)}}\right)\right\}}\hat{\mu}_{\tau_j}^{(k)}\hat{\sigma}_\tau^{(k)},
\end{aligned}$$

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

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

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 $\sigma^2 = \hat{\sigma}^{2^{(k+1)}}$, obtaining $\hat{\lambda}^{(k+1)}$ as the solution of the following equation:

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

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.,

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

The maximization in the CML-step needs 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) $\mathrm{E}\left\{ \dfrac{\phi(\lambda Z)}{\Phi(\lambda Z)} \right\} = \sqrt{\dfrac{2}{\pi}} \dfrac{1}{\sqrt{1+\lambda^2}}.$

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

(iii) $\mathrm{E}\left\{ Z^{2} \dfrac{\phi(\lambda Z)}{\Phi(\lambda Z)} \right\} = \sqrt{\dfrac{2}{\pi}} \dfrac{\lambda}{(1+\lambda^2)^{3/2}}.$

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 the following lemma. The standard errors of ML estimates can be computed by taking the square root of the corresponding diagonal elements of $\boldsymbol{I}^{-1}(\hat{\xi}, \hat{\sigma}, \hat{\lambda})$.

Lemma 4 *The Fisher information for $\boldsymbol{\theta} = (\xi, \sigma^2, \lambda)$ is*

$$\boldsymbol{I}(\boldsymbol{\theta}) = \mathrm{E}\left(-\frac{\partial^2 \ell(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\top}} \right) = \begin{bmatrix} I_{\xi\xi} & I_{\xi\sigma^2} & I_{\xi\lambda} \\ I_{\xi\sigma^2} & I_{\sigma^2\sigma^2} & I_{\sigma^2\lambda} \\ I_{\xi\lambda} & I_{\sigma^2\lambda} & I_{\lambda\lambda} \end{bmatrix}, \tag{7}$$

where

$$
\begin{aligned}
I_{\xi\xi} &= \mathrm{E}\left(-\frac{\partial^2 \ell(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \xi^2}\right) = \frac{n}{\sigma^2}(1+\lambda^2 a_0),\\
I_{\xi\sigma^2} &= \mathrm{E}\left(-\frac{\partial^2 \ell(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \xi \partial \sigma^2}\right) = \frac{n}{2\sigma^3}\left(\sqrt{\frac{2}{\pi}}\frac{\lambda+2\lambda^3}{(1+\lambda^2)^{3/2}} + \lambda^2 \mathrm{a}_1\right),\\
I_{\xi\lambda} &= \mathrm{E}\left(-\frac{\partial^2 \ell(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \xi \partial \lambda}\right) = \frac{n}{\sigma}\left(\frac{2}{\pi}\frac{1}{(1+\lambda^2)^{3/2}} - \lambda a_1\right),\\
I_{\sigma^2\sigma^2} &= \mathrm{E}\left(-\frac{\partial^2 \ell(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \sigma^2}\right) = \frac{n}{2\sigma^4}\left(1+\frac{1}{2}\lambda^2 a_2\right),\\
I_{\sigma^2\lambda} &= \mathrm{E}\left(-\frac{\partial^2 \ell(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \sigma^2 \partial \lambda}\right) = -\frac{n\lambda}{2\sigma^2}a_2,\\
I_{\lambda\lambda} &= \mathrm{E}\left(-\frac{\partial^2 \ell(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \lambda^2}\right) = n a_2.
\end{aligned}
$$

Note that the quantities $a_k = \mathrm{E}\Big(Z^k\big(\phi(\lambda Z)/\Phi(\lambda Z)\big)^2\Big)$, $(k = 0,\ 1,\ 2)$ need to be evaluated numerically. Based on the large sample theorem, the standard errors estimates for the ML estimates $\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = (\hat{\xi}_{\mathrm{ML}}, \hat{\sigma}^2_{\mathrm{ML}}, \hat{\lambda}_{\mathrm{ML}})$ can be computed by taking the square root of the corresponding diagonal elements of $\mathbf{J}^{-1}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}})$.

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

$$
f(y_j \mid \boldsymbol{\omega}, \boldsymbol{\Theta}) = \sum_{i=1}^{g} \omega_i\ \psi(y_j \mid \xi_i, \sigma_i^2, \lambda_i), \tag{8}
$$

where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_g)$ are the mixing probabilities which are constrained to be non-negative and sum to unity and $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g)$ 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 $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ are independent, with multinomial densities

$$f(\boldsymbol{z}_j) = \omega_1^{z_{1j}} \omega_2^{z_{2j}} \cdots (1 - \omega_1 - \cdots - \omega_g)^{z_{gj}}. \tag{9}$$

We shall write $\boldsymbol{Z}_j \sim MN(1;\ \omega_1, \ldots, \omega_g)$ to denote $\boldsymbol{Z}_j$ with density (9).

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

$$\begin{aligned} Y_j \mid \tau_j,\ Z_{ij} = 1 &\sim N\Big(\xi_i + \delta(\lambda_i)\tau_j,\ \big(1 - \delta^2(\lambda_i)\big)\sigma_i^2\Big), \\ \tau_j \mid Z_{ij} = 1 &\sim TN(0, \sigma_i^2) I(\tau_j > 0), \\ \boldsymbol{Z}_j &\sim MN(1;\ \omega_1, \ldots, \omega_g) \qquad (j = 1, \ldots, n). \end{aligned} \tag{10}$$

3.2. Maximum likelihood estimation

As in (6), we have

$$\tau_j \mid Y_j = y_j, Z_{ij} = 1 \sim TN(\mu_{\tau_{ij}}, \sigma_{\tau_i}^2) I\{\tau_j > 0\},$$

where

$$\mu_{\tau_{ij}} = \delta(\lambda_i)(y_j - \xi_i), \qquad \sigma_{\tau_i} = \sigma_i \sqrt{1 - \delta^2(\lambda_i)}. \tag{11}$$

From (10), the complete-data log-likelihood function is

$$\begin{aligned} \ell_c(\boldsymbol{\theta}) = \sum_{j=1}^{n} \sum_{i=1}^{g} Z_{ij} \Bigg\{ & \log(\omega_i) - \log(\sigma_i^2) - \frac{1}{2} \log\Big(1 - \delta^2(\lambda_i)\Big) \\ & - \frac{\tau_j^2 - 2\delta(\lambda_i)\tau_j(y_j - \xi_i) + (y_j - \xi_i)^2}{2\sigma_i^2\Big(1 - \delta^2(\lambda_i)\Big)} \Bigg\}. \end{aligned} \tag{12}$$

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 (12), we obtain

$$\hat{z}_{ij}^{(k)} = \frac{\omega_i^{(k)} \psi(y_j \mid \xi_i^{(k)}, \sigma_i^{2^{(k)}}, \lambda_i^{(k)})}{\sum_{m=1}^{g} \omega_m^{(k)} \psi(y_j \mid \xi_m^{(k)}, \sigma_m^{2^{(k)}}, \lambda_m^{(k)})}, \tag{13}$$

$$\hat{s}_{1ij}^{(k)} = \hat{z}_{ij}^{(k)} \left[\hat{\mu}_{\tau_{ij}}^{(k)} + \hat{\sigma}_{\tau_i}^{(k)} \frac{\phi\left\{ \hat{\lambda}^{(k)} \left(\frac{y_j - \hat{\xi}_i^{(k)}}{\hat{\sigma}_i^{(k)}} \right) \right\}}{\Phi\left\{ \hat{\lambda}_i^{(k)} \left(\frac{y_j - \hat{\xi}_i^{(k)}}{\hat{\sigma}_i^{(k)}} \right) \right\}} \right], \tag{14}$$

and

$$\hat{s}_{2ij}^{(k)} \quad = \quad \hat{z}_{ij}^{(k)} \left[\hat{\mu}_{\tau_{ij}}^{(k)^2} + \hat{\sigma}_{\tau_i}^{(k)^2} + \frac{\phi\left\{ \hat{\lambda}^{(k)} \left(\frac{y_j - \hat{\xi}_i^{(k)}}{\hat{\sigma}_i^{(k)}} \right) \right\}}{\Phi\left\{ \hat{\lambda}_i^{(k)} \left(\frac{y_j - \hat{\xi}_i^{(k)}}{\hat{\sigma}_i^{(k)}} \right) \right\}} \hat{\mu}_{\tau_{ij}}^{(k)} \hat{\sigma}_{\tau_i}^{(k)} \right], \tag{15}$$

where $\hat{\mu}_{\tau_{ij}}^{(k)}$, $\hat{\sigma}_{\tau_i}^{(k)}$ are $\mu_{\tau_{ij}}$ and σ_{τ_i} in (11) 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 $\mathbf{\Theta} = \hat{\mathbf{\Theta}}^{(k)}$, 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 (13), (14) and (15).

CM-step 1: Calculate

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

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 $\sigma_i^2 = \hat{\sigma}_i^{2^{(k+1)}}$, obtaining $\hat{\lambda}_i^{(k+1)}$ (i=1,. . . ,g) as the solution of the following equation:

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

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 $\hat{\boldsymbol{\lambda}}^{(k)}$ as

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

We remark here that if the skewness parameters $\lambda_1, \ldots, \lambda_g$ are assumed to be identical, we shall 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 $\boldsymbol{I}_o(\boldsymbol{\Theta} \mid \boldsymbol{y}) = -\partial^2 \ell(\boldsymbol{\Theta} \mid \boldsymbol{Y})/\partial\boldsymbol{\Theta}\partial\boldsymbol{\Theta}^\top$ be the observed information matrix for the mixture model (8). 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

$$\boldsymbol{I}_o(\hat{\boldsymbol{\Theta}} \mid \boldsymbol{y}) = \sum_{j=1}^{n} \hat{\mathbf{s}}_j \hat{\mathbf{s}}_j^\top, \tag{16}$$

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

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

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

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

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

$$\begin{aligned}\hat{s}_{j,\sigma_r} &= \frac{\hat{\omega}_r\psi(y_j \mid \hat{\xi}_r, \hat{\sigma}_r^2, \hat{\lambda}_r)}{\sum_{i=1}^{g}\hat{\omega}_i\psi(y_j \mid \hat{\xi}_i, \hat{\sigma}_i^2, \hat{\lambda}_i)}\Bigg\{ -\frac{1}{\hat{\sigma}_r} + \frac{(y_j-\hat{\xi}_r)^2}{\hat{\sigma}_r^3} \\ &\quad - \frac{2\hat{\omega}_r\hat{\lambda}_r(y_j-\hat{\xi}_r)\phi\big((y_j-\hat{\xi}_r)/\hat{\sigma}_r\big)\phi\big(\hat{\lambda}_r(y_j-\hat{\xi}_r)/\hat{\sigma}_r\big)}{\hat{\sigma}_r^3\sum_{i=1}^{g}\hat{\omega}_i\psi(y_j \mid \hat{\xi}_i, \hat{\sigma}_i^2, \hat{\lambda}_i)}\Bigg\} \quad (r=1,\ldots,g), \\ \hat{s}_{j,\lambda_r} &= \frac{\hat{\omega}_r\psi(y_j \mid \hat{\xi}_r, \hat{\sigma}_r^2, \hat{\lambda}_r)}{\sum_{i=1}^{g}\hat{\omega}_i\psi(y_j \mid \hat{\xi}_i, \hat{\sigma}_i^2, \hat{\lambda}_i)}\left(\frac{y_j-\hat{\xi}_r}{\hat{\sigma}_r}\right)\frac{\phi\Big\{\hat{\lambda}_r(y_j-\hat{\xi}_r)/\hat{\sigma}_r\Big\}}{\Phi\Big\{\hat{\lambda}_r(y_j-\hat{\xi}_r)/\hat{\sigma}_r\Big\}} \quad (r=1,\ldots,g).\end{aligned}$$

The information-based approximation (16) is asymptotically applicable. However, it may not be reliable unless the sample size is large. It is common in practice to use the bootstrap approach (Efron and Tibshirani, 1986) as an alternative Monte Carlo approximation of $\mathrm{cov}(\hat{\mathbf{\Theta}})$ via generating a sufficient number of bootstrap samples. The bootstrap method may provide more accurate stand error estimates than (16). However, it requires enormous computational burden.

4. Bayesian Modelling For Skew Normal Mixtures

We consider a Bayesian approach where $\mathbf{\Theta}$ 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 avoid causing nonintegrable posterior distributions. The prior distributions for model (8) are of the forms

$$\begin{aligned}\xi_i &\sim N(\eta, \kappa^{-1}) \quad (i=1,\ldots,g), \\ \sigma_i^{-2} \mid \beta &\sim Ga(\alpha, \beta) \quad (i=1,\ldots,g), \\ \beta &\sim Ga(\nu_1, \nu_2), \\ \delta(\lambda_i) &\sim U(-1,1) \quad (i=1,\ldots,g), \\ \boldsymbol{\omega} &\sim D(h,\ldots,h),\end{aligned} \tag{17}$$

with the restriction $\xi_1 < \cdots < \xi_g$. In (17), β is an unknown hyperparameter, $(\eta, \kappa, \alpha, \nu_1, \nu_2, h)$ are known (data-dependent) constants, $Ga(\alpha, \beta)$ 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) by letting η be equal to the midpoint of the observed interval and $\kappa^{-1}=R^2$, where R is the range and setting $\alpha=2$, $\nu_1=0.2$, $\nu_2=100\nu_1/(\alpha R^2)$ and $h=1$.

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

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

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

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 $Ga(\nu_1+g\alpha,\ \nu_2+\sum_{i=1}^g\sigma_i^{-2(k)})$.

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

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

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

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

$$\begin{aligned}b\ =\ &\Big\{\sum_{j=1}^nZ_{ij}^{(k+1)}\tau_j^{2(k+1)}-2\delta(\lambda_i^{(k)})\sum_{j=1}^nZ_{ij}^{(k+1)}\tau_j^{(k+1)}(y_j-\xi_i^{(k+1)})\\&+\sum_{j=1}^nZ_{ij}^{(k+1)}(y_j-\xi_i^{(k+1)})^2\Big\}/\Big\{2\big((1-\delta^2(\lambda_i^{(k)})\big)\Big\}.\end{aligned}$$

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

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

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 the following function:

$$g(\boldsymbol{\delta}^*) = f\big(\boldsymbol{\delta}(\boldsymbol{\delta}^*)\big)\prod_{i=1}^{g} J_{\delta^*(\lambda_i)},$$

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 Jacobin of transformation from $\delta(\lambda_i)$ to $\delta^*(\lambda_i)$. A g-dimensional multivariate normal distribution with mean $\boldsymbol{\delta}^{*(k)}$ and covariance matrix $c^2\boldsymbol{\Sigma}^{(k)}_{\boldsymbol{\delta}^*}$ is chosen as the proposal distribution, where the scale $c \approx 2.4/\sqrt{g}$, as suggested in Gelman, Roberts and Gilks (1995). The value of $\boldsymbol{\Sigma}^{(k)}_{\boldsymbol{\delta}^*}$ 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 $\xi_1^{(0)} < \cdots < \xi_g^{(0)}$.

5. Examples

5.1. The enzyme data

We first carry out our methodology for the enzyme data set with $n = 245$ observations. The data was 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) provide the reversible jump MCMC approach for the univariate normal mixture models with an unknown number of components and identify the most possible values of g to be between 3 and 5.

We fit the data to the two-component skew normal mixture model

$$f(y) = \omega\psi(y|\xi_1, \sigma_1^2, \lambda_1) + (1-\omega)\psi(y|\xi_2, \sigma_2^2, \lambda_2). \tag{18}$$

The ECM algorithm was run with various starting values and was checked for convergence. The resulting ML estimates and the corresponding standard errors are listed in Table 1. 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 3.

For comparison purposes, we also fit the data to the normal mixture models ($\lambda_1 = \lambda_2 = 0$) with $g = 2 - 5$ components. The log-likelihood maximum and two information-based criteria, AIC (Akaike, 1973) and BIC (Schwarz, 1978), are displayed in Table 2. As expected, the fitting of a skew normal mixture model is superior to normal mixtures since it has the largest log-likelihood with parsimonious parameters as well as the smallest AIC and BIC. A histogram of the data overlaid with various fitted mixture densities is displayed in Figure 4.

Table 1: Estimated parameter values and the corresponding standard errors (SE) for model (18) 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

Table 2: A comparison of log-likelihood maximum, AIC and BIC for the fitted skew normal mixture (SNMIX) model and normal mixture (NORMIX) model for the enzyme data. The number of parameters is denoted by m.

Model	g	m	log-likelihood	AIC[†]	BIC[‡]
SNMIX	2	7	−41.92	97.84	122.35
NORMIX	2	5	−54.64	119.28	136.79
NORMIX	3	8	−47.83	111.66	139.67
NORMIX	4	11	−46.75	115.50	154.01
NORMIX	5	14	−46.26	120.52	169.54
NORMIX	≥ 6			> 123	> 185

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

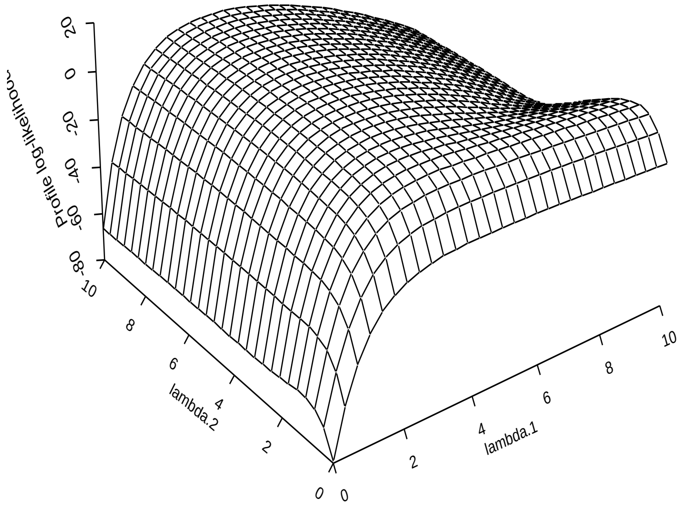

Figure 3: Plot of the profile log-likelihood for λ_1 and λ_2 for the enzyme data.

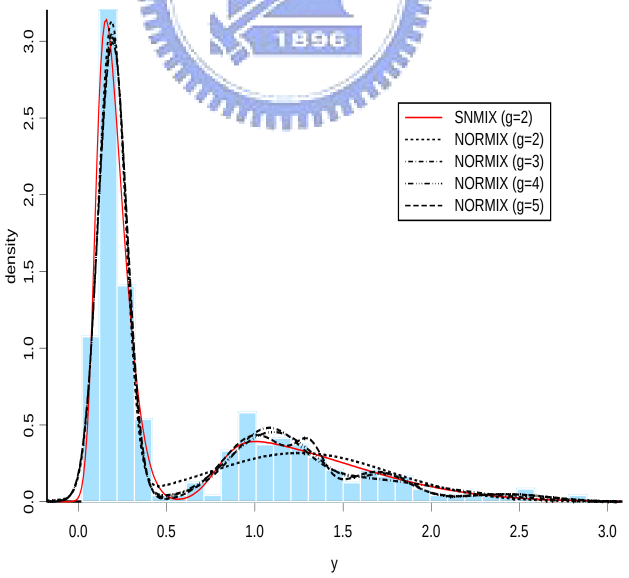

Figure 4: Histogram of the enzyme data overlaid with a ML-fitted two-component skew normal mixture (SNMIX) distribution and various ML-fitted g-component normal mixture (NORMIX) distributions ($g = 2 - 5$).

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 the Yellowstone National Park, Wyoming, USA. The data appear to be bimodal with asymmetrical components. We fit a two-component skew mixture normal model (18) 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.

To illustrate our Bayesian MCMC methodology described in Section 4, we ran 7 parallel chains of 10,000 iterations each with the starting values chosen dispersed around the ML estimates. After 5,000 iterations of "burn-in" for each chain, we monitor the convergence by examining the *multivariate potential scale reduction factor* (MPSRF) proposed by Brooks and Gelman (1998). The posterior mean, standard deviation, median and 95% HPD interval (2.5% and 97.5% posterior quantiles) of the converged MCMC simulation samples are listed in the 4-8th columns of Table 3.

Table 3: ML estimation results and MCMC summary statistics for the parameters of model (18) 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	5.8026	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 5 displays the convergence diagrams and histograms of the posterior samples of the 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.

Finally, we compare the ML-fitted normal mixture density with the fitted skew normal mixture densities via ML and Byesian methods based on graphical visualization. The ML density estimation for normal and skew normal mixtures densities together with Bayesian predictive density are shown in Figure 6(a). We have also plotted the comparison of fitted and empirical cumulative density functions (CDFs) in Figure 6(b). Obviously, the ML and Bayesian methods are quite comparable for the data. However, the appropriateness of density fitting for skew normal mixtures is clearly better than for normal mixtures. Furthermore, the CDFs of the skew normal mixtures track closer to the real data.

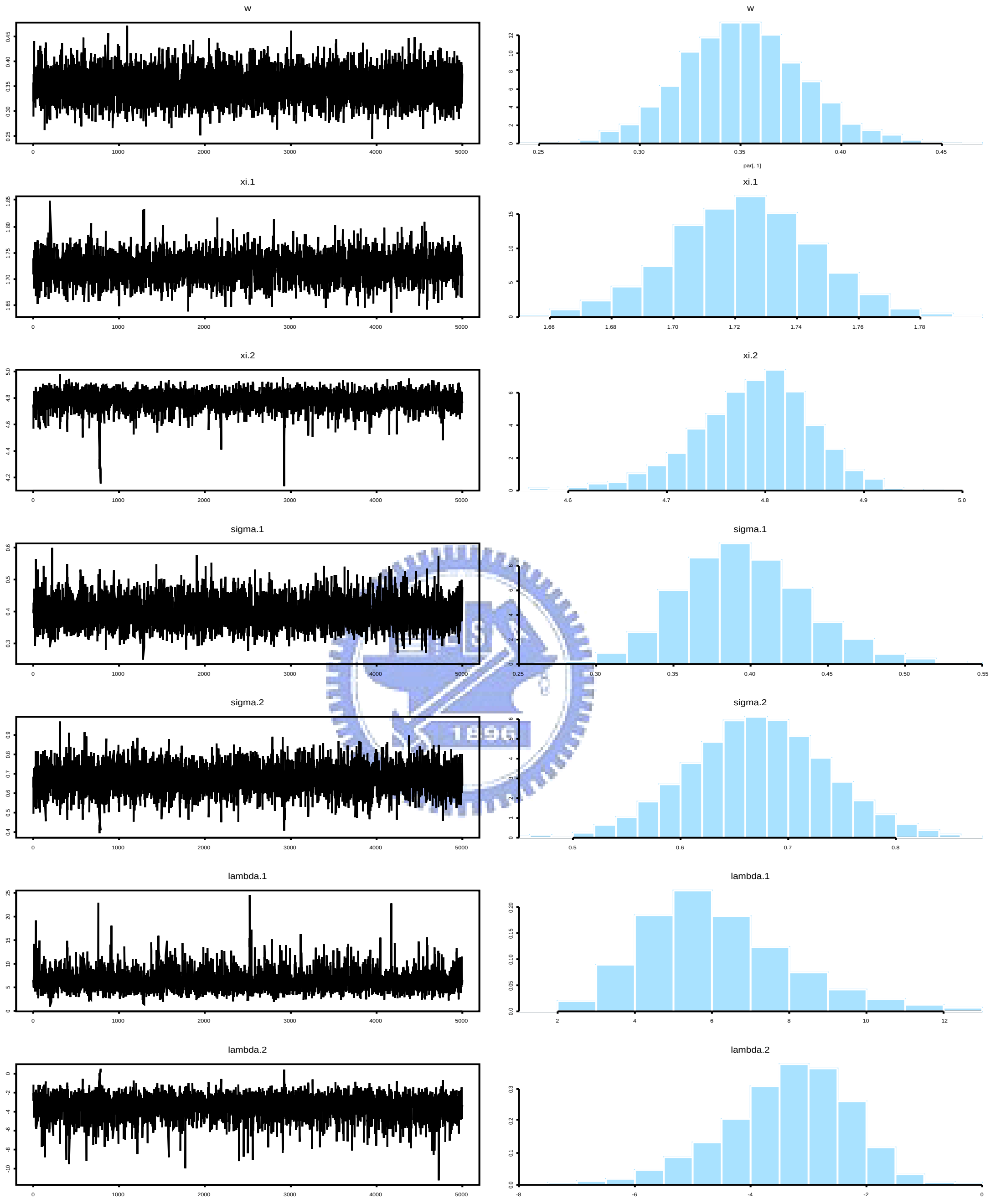

Figure 5: Convergence diagrams and histograms of the posterior sample of the parameters for model (18) with the faithful data.

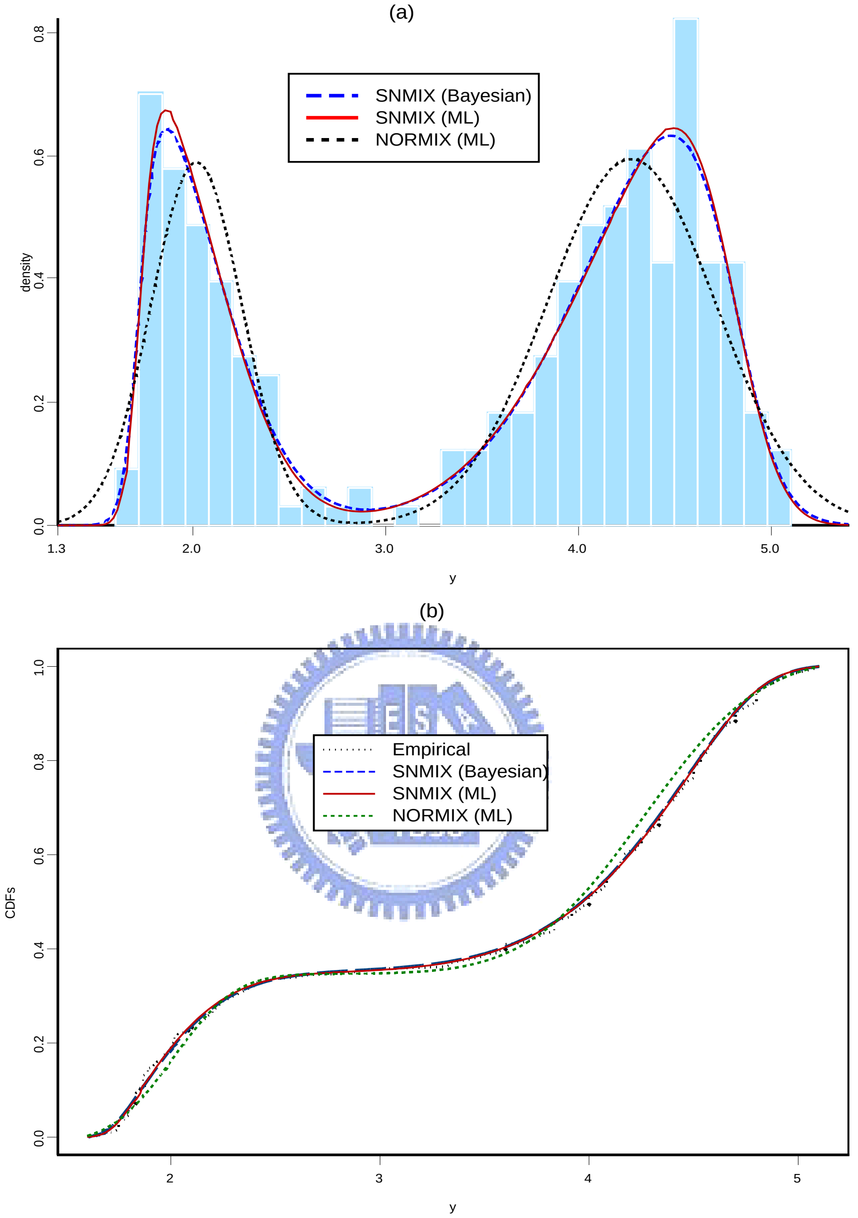

Figure 6: (a)Histogram of the faithful data overlaid with densities based on two fitted two-component skew normal mixture (SNMIX) distributions (ML and Bayesian), and a ML-fitted two-component normal mixture (NORMIX) distribution; (b)Empirical CDF of the faithful data overlaid with CDFs based on two fitted two-component SNMIX distributions (ML and Bayesian) and a ML-fitted two-component NORMIX distribution.

6. Conclusion

We have proposed and illustrated ML and Bayesian density estimations for finite mixture modelling using the skew normal distribution. The key contributions lie in the development of computational techniques for the hierarchical skew normal mixtures. We provide EM-type algorithms for calculating the ML estimates and a workable MCMC algorithm for sampling the posterior distributions in the Bayesian paradigm.

In our illustrated examples, it is quite appealing that the skew normal mixtures can provide more appropriate density estimation than normal mixtures based on information-based criteria and graphical visualization. Future work on extensions includes generalizations to mixture of multivariate skew normal distributions (e.g., Azzalini and Dalla Valle, 1996 and Gupta, González-Farías and Domínguez-Monila, 2004) and offering techniques to model the number of components and the mixture components parameters jointly.

Appendix

A. Proof of Lemma 1

(i)

$$\begin{aligned}
E(X^n) &= \int x^n \frac{\sqrt{1+\lambda^2}}{\sigma}\phi\big(\frac{\sqrt{1+\lambda^2}}{\sigma}(x-\xi)\big)dx, \\
\frac{d}{d\xi}E(X^n) &= \int \frac{1+\lambda^2}{\sigma^2}(x-\xi)x^n\frac{\sqrt{1+\lambda^2}}{\sigma}\phi\big(\frac{\sqrt{1+\lambda^2}}{\sigma}(x-\xi)\big)dx \\
&= \frac{1+\lambda^2}{\sigma^2}E(X^{n+1}) - \xi\frac{1+\lambda^2}{\sigma^2}E(X^n), \\
E(X^{n+1}) &= \xi E(X^n) + \frac{\sigma^2}{1+\lambda^2}\frac{d}{d\xi}E(X^n).
\end{aligned}$$

(ii)

$$E(Y^n) = \int y^n \frac{2}{\sigma}\phi\big(\frac{y-\xi}{\sigma}\big)\Phi\big(\lambda\frac{y-\xi}{\sigma}\big)dy,$$

$$
\begin{aligned}
\frac{d}{d\xi}E(Y^n) &= \int \frac{y-\xi}{\sigma^2}y^n\frac{2}{\sigma}\phi\big(\frac{y-\xi}{\sigma}\big)\Phi\big(\lambda\frac{y-\xi}{\sigma}\big)dy - \int \frac{\lambda}{\sigma}y^n\frac{2}{\sigma}\phi\big(\frac{y-\xi}{\sigma}\big)\phi\big(\lambda\frac{y-\xi}{\sigma}\big)dy \\
&= \frac{E(Y^{n+1})}{\sigma^2} - \frac{\xi E(Y^n)}{\sigma^2} - \frac{2\lambda}{\sigma^2}\int \frac{1}{\sqrt{2\pi}}y^n\phi\big(\sqrt{1+\lambda^2}\frac{y-\xi}{\sigma}\big)dy \\
&= \frac{E(Y^{n+1})}{\sigma^2} - \frac{\xi E(Y^n)}{\sigma^2} - \frac{2\lambda}{\sigma^2}\frac{1}{\sqrt{2\pi}}\frac{\sigma}{\sqrt{1+\lambda^2}}E(X^n).
\end{aligned}
$$

Hence,

$$
E(Y^{n+1}) = \xi E(Y^n) + \sigma^2\frac{d}{d\xi}E(Y^n) + \sqrt{\frac{2}{\pi}}\delta(\lambda)\sigma E(X).
$$

(iii)

$$
\begin{aligned}
E\big(Y-E(Y)\big)^n &= \int \big(y-E(y)\big)^n\frac{2}{\sigma}\phi\big(\frac{y-\xi}{\sigma}\big)\Phi\big(\lambda\frac{y-\xi}{\sigma}\big)dy, \\
\frac{d}{d\xi}E\big(Y-E(Y)\big)^n &= -\int n\big(Y-E(Y)\big)^{n-1}\big(\frac{d}{d\xi}E(Y)\big)\frac{2}{\sigma}\phi\big(\frac{y-\xi}{\sigma}\big)\Phi\big(\lambda\frac{y-\xi}{\sigma}\big)dy \\
&\quad + \int \frac{y-\xi}{\sigma^2}\big(Y-E(Y)\big)^n\frac{2}{\sigma}\phi\big(\frac{y-\xi}{\sigma}\big)\Phi\big(\lambda\frac{y-\xi}{\sigma}\big)dy \\
&\quad - \int \big(Y-E(Y)\big)^n\frac{\lambda}{\sigma}\frac{2}{\sigma}\phi\big(\frac{y-\xi}{\sigma}\big)\phi\big(\lambda\frac{y-\xi}{\sigma}\big)dy \\
&= -nE\big(Y-E(Y)\big)^{n-1} \\
&\quad + \int \frac{1}{\sigma^2}E\big(Y-E(Y)\big)^{n+1}\frac{2}{\sigma}\phi\big(\frac{y-\xi}{\sigma}\big)\Phi\big(\lambda\frac{y-\xi}{\sigma}\big)dy \\
&\quad + \int \big(Y-E(Y)\big)^n\frac{y-\xi}{\sigma^2}\frac{2}{\sigma}\phi\big(\frac{y-\xi}{\sigma}\big)\Phi\big(\lambda\frac{y-\xi}{\sigma}\big)dy \\
&\quad - \int \big(Y-E(Y)\big)^n\frac{\lambda}{\sigma}\frac{2}{\sigma}\frac{1}{\sqrt{2\pi}}\phi\big(\sqrt{1+\lambda^2}\frac{y-\xi}{\sigma}\big)dy \\
&= -nE\big(Y-E(Y)\big)^{n-1} + \frac{1}{\sigma^2}E\big(Y-E(Y)\big)^{n+1} \\
&\quad + \frac{1}{\sigma^2}\big(E(Y)-\xi\big)E\big(Y-E(Y)\big)^n - \sqrt{\frac{2}{\pi}}\delta(\lambda)\frac{1}{\sigma}E\big(X-E(Y)\big)^n.
\end{aligned}
$$

Hence,

$$
\begin{aligned}
E\big(Y-E(Y)\big)^{n+1} &= \sigma^2\frac{d}{d\xi}E\big(Y-E(Y)\big)^n + n\sigma^2E\big(Y-E(Y)\big)^{n-1} \\
&\quad - \big(E(Y)-\xi\big)E\big(Y-E(Y)\big)^n + \sqrt{\frac{2}{\pi}}\delta(\lambda)\sigma E\big(X-E(Y)\big)^n.
\end{aligned}
$$

B. Proof of Lemma 2

(i)

$$
\begin{aligned}
M_X(t) &= E\big(e^{tx}\big) = \int_{a1}^{a2} e^{tx} f(x \mid \mu, \sigma^2)dx \\
&= \frac{1}{\Phi(\alpha_2)-\Phi(\alpha_1)} \int_{a1}^{a2} e^{tx} \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\big(-\frac{1}{2\sigma^2}(x-\mu)^2\big)dx \\
&= \frac{1}{\Phi(\alpha_2)-\Phi(\alpha_1)} \exp\big(\mu t + \frac{\sigma^2 t^2}{2}\big) \int_{a1}^{a2} \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\big(-\frac{1}{2\sigma^2}[x-(\mu+\sigma^2 t)]^2\big)dx \\
&= \frac{1}{\Phi(\alpha_2)-\Phi(\alpha_1)} \exp\big(\mu t + \frac{\sigma^2 t^2}{2}\big) \int_{\alpha_1-\sigma t}^{\alpha_2-\sigma t} \frac{1}{\sqrt{2\pi}} \exp\big(-\frac{1}{2}z^2\big)dz \\
&= \exp\big(\mu t + \frac{\sigma^2 t^2}{2}\big)\frac{\Phi(\alpha_2-\sigma t)-\Phi(\alpha_1-\sigma t)}{\Phi(\alpha_2)-\Phi(\alpha_1)}.
\end{aligned}
$$

(ii)

$$
\begin{aligned}
M^{'}(t) &= (\mu+\sigma^2 t)\exp\big(\mu t + \frac{\sigma^2 t^2}{2}\big)\frac{\Phi(\alpha_2-\sigma t)-\Phi(\alpha_1-\sigma t)}{\Phi(\alpha_2)-\Phi(\alpha_1)} \\
&\quad -\exp\big(\mu t + \frac{\sigma^2 t^2}{2}\big)\sigma\frac{\phi(\alpha_2-\sigma t)-\phi(\alpha_1-\sigma t)}{\Phi(\alpha_2)-\Phi(\alpha_1)}, \\
E(X) &= M^{'}(0) = \mu - \sigma\frac{\phi(\alpha_2)-\phi(\alpha_1)}{\Phi(\alpha_2)-\Phi(\alpha_1)}.
\end{aligned}
$$

(iii)

$$
\begin{aligned}
M^{''}(t) &= \sigma^2 \exp\big(\mu t + \frac{\sigma^2 t^2}{2}\big)\frac{\Phi(\alpha_2-\sigma t)-\Phi(\alpha_1-\sigma t)}{\Phi(\alpha_2)-\Phi(\alpha_1)} \\
&\quad +(\mu+\sigma^2 t)^2 \exp\big(\mu t + \frac{\sigma^2 t^2}{2}\big)\frac{\Phi(\alpha_2-\sigma t)-\Phi(\alpha_1-\sigma t)}{\Phi(\alpha_2)-\Phi(\alpha_1)} \\
&\quad -(\mu+\sigma^2 t)\exp\big(\mu t + \frac{\sigma^2 t^2}{2}\big)\sigma\frac{\phi(\alpha_2-\sigma t)-\phi(\alpha_1-\sigma t)}{\Phi(\alpha_2)-\Phi(\alpha_1)} \\
&\quad -(\mu+\sigma^2 t)\exp\big(\mu t + \frac{\sigma^2 t^2}{2}\big)\sigma\frac{\phi(\alpha_2-\sigma t)-\phi(\alpha_1-\sigma t)}{\Phi(\alpha_2)-\Phi(\alpha_1)} \\
&\quad -\exp\big(\mu t + \frac{\sigma^2 t^2}{2}\big)\sigma\frac{(\alpha_2-\sigma t)\phi(\alpha_2-\sigma t)-(\alpha_1-\sigma t)\phi(\alpha_1-\sigma t)}{\Phi(\alpha_2)-\Phi(\alpha_1)}. \\
E(X^2) &= M^{''}(0) = \sigma^2 + \mu^2 - \mu\sigma\frac{\phi(\alpha_2)-\phi(\alpha_1)}{\Phi(\alpha_2)-\Phi(\alpha_1)} \\
&\quad -\mu\sigma\frac{\phi(\alpha_2)-\phi(\alpha_1)}{\Phi(\alpha_2)-\Phi(\alpha_1)} - \sigma\frac{\alpha_2\phi(\alpha_2)-\alpha_1\phi(\alpha_1)}{\Phi(\alpha_2)-\Phi(\alpha_1)} \\
&= \mu^2 + \sigma^2 - \sigma\frac{\alpha_2\phi(\alpha_2)-1\alpha_\phi(\alpha_1)}{\Phi(\alpha_2)-\Phi(\alpha_1)} - 2\mu\sigma\frac{\phi(\alpha_2)-\phi(\alpha_1)}{\Phi(\alpha_2)-\Phi(\alpha_1)}.
\end{aligned}
$$

C. Proof of Lemma 3

(i)

$$\begin{aligned}
E\big(\frac{\phi(\lambda Z)}{\Phi(\lambda Z)}\big) &= \int 2\phi(z)\Phi(\lambda z)\frac{\phi(\lambda z)}{\Phi(\lambda z)}dz \\
&= \int 2\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\lambda^2 z^2}dz \\
&= \frac{1}{\pi}\int e^{-\frac{1}{2}(1+\lambda^2)z^2}dz \\
&= \frac{1}{\pi}\sqrt{2\pi}\frac{1}{\sqrt{1+\lambda^2}} \\
&= \sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{1+\lambda^2}}.
\end{aligned}$$

(ii)

$$\begin{aligned}
E\big(Z^{2k+1}\frac{\phi(\lambda Z)}{\Phi(\lambda Z)}\big) &= \int 2\phi(z)\Phi(\lambda z)z^{2k+1}\frac{\phi(\lambda z)}{\Phi(\lambda z)}dz \\
&= \int 2z^{2k+1}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\lambda^2 z^2}dz \\
&= \frac{1}{\pi}\int z^{2k+1}e^{-\frac{1}{2}(1+\lambda^2)z^2}dz \\
&= 0 \quad (k = 0, 1, 2, \cdots).
\end{aligned}$$

(iii)

$$\begin{aligned}
E\big(Z^2\frac{\phi(\lambda Z)}{\Phi(\lambda Z)}\big) &= \int 2\phi(z)\Phi(\lambda z)z^2\frac{\phi(\lambda z)}{\Phi(\lambda z)}dz \\
&= \int 2z^2\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\lambda^2 z^2}dz \\
&= \frac{1}{\pi}\int z^2 e^{-\frac{1}{2}(1+\lambda^2)z^2}dz \\
&= \frac{1}{\pi}\sqrt{2\pi}\frac{1}{\sqrt{1+\lambda^2}}\frac{1}{1+\lambda^2} \\
&= \sqrt{\frac{2}{\pi}}\frac{\lambda}{(1+\lambda)^{3/2}}.
\end{aligned}$$

D. Proof of Lemma 4

The log-likelihood function is

$$\ell(\boldsymbol{\theta}|\boldsymbol{y}) \propto -n\log(\sigma) - \frac{1}{2}\sum_{i=1}^{n} z_i^2 + \sum_{i=1}^{n}\log(\Phi(\lambda z_i)),$$

where $\boldsymbol{\theta} = (\xi, \sigma^2, \lambda)$, $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ and

$$z_i = \frac{y_i - \xi}{\sigma}.$$

Let $\psi(\lambda z_i) = \phi(\lambda z_i)/\Phi(\lambda z_i)$, then

$$\frac{\partial \log(\Phi(\lambda z_i))}{\partial \theta_j} = \frac{\phi(\lambda z_i)}{\Phi(\lambda z_i)} \frac{\partial \lambda z_i}{\partial \theta_j} = \psi(\lambda z_i) \frac{\partial \lambda z_i}{\partial \theta_j}.$$

With the following algebraic operations, we have

$$\begin{aligned}
\frac{\partial \psi(\lambda z_i)}{\partial \theta_j} &= \frac{1}{\Phi^2(\lambda z_i)} \left(\Phi(\lambda z_i) \frac{\partial \phi(\lambda z_i)}{\partial \theta_j} - \phi^2(\lambda z_i) \frac{\partial \lambda z_i}{\partial \theta_j} \right) \\
&= \frac{1}{\Phi^2(\lambda z_i)} \left(\Phi(\lambda z_i) \left(-\lambda z_i \phi(\lambda z_i) \frac{\partial \lambda z_i}{\partial \theta_j} \right) - \phi^2(\lambda z_i) \frac{\partial \lambda z_i}{\partial \theta_j} \right) \\
&= -\psi(\lambda z_i) \Big(\lambda z_i + \psi(\lambda z_i) \Big) \frac{\partial \lambda z_i}{\partial \theta_j} \\
&= -\eta(\lambda z_i) \frac{\partial \lambda z_i}{\partial \theta_j},
\end{aligned}$$

where $\eta(\lambda z_i) = \psi(\lambda z_i) \Big(\lambda z_i + \psi(\lambda z_i) \Big)$.

Note that the 1st differentials are

$$\frac{\partial \lambda z_i}{\partial \xi} = -\frac{\lambda}{\sigma}, \quad \frac{\partial \lambda z_i}{\partial \sigma^2} = -\frac{\lambda z_i}{2\sigma^2} \text{ and } \quad \frac{\partial \lambda z_i}{\partial \lambda} = z_i.$$

The score vector and Hessian matrix are as follows:

$$\mathbf{s}(\boldsymbol{\theta}|\boldsymbol{Y}) = \begin{bmatrix} s_\xi \\ s_{\sigma^2} \\ s_\lambda \end{bmatrix}, \text{ and } \ \mathbf{H}(\boldsymbol{\theta}|\boldsymbol{Y}) = \begin{bmatrix} H_{\xi\xi} & H_{\xi\sigma^2} & H_{\xi\lambda} \\ H_{\xi\sigma^2} & H_{\sigma^2\sigma^2} & H_{\sigma^2\lambda} \\ H_{\xi\lambda} & H_{\sigma^2\lambda} & H_{\lambda\lambda} \end{bmatrix}.$$

The elements of the score vector $s(\boldsymbol{\theta}|\boldsymbol{Y})$ are

$$\begin{aligned}
s_\xi &= \frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \xi} = -\frac{1}{2} \sum_{i=1}^n 2 z_i \frac{-1}{\sigma} + \sum_{i=1}^n \psi(\lambda z_i) \lambda \frac{-1}{\sigma} \\
&= \frac{1}{\sigma} \sum_{i=1}^n z_i - \frac{\lambda}{\sigma} \sum_{i=1}^n \psi(\lambda z_i), \\
s_{\sigma^2} &= \frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \sigma^2} = -\frac{n}{2\sigma^2} - \frac{1}{2} \sum_{i=1}^n z_i \frac{-1(y_i - \xi)}{\sigma^4} + \sum_{i=1}^n \psi(\lambda z_i) \lambda \frac{-1\lambda(y_i - \xi)}{2\sigma^3} \\
&= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^2} \sum_{i=1}^n z_i^2 - \frac{\lambda}{2\sigma^2} \sum_{i=1}^n z_i \psi(\lambda z_i), \\
s_\lambda &= \frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \lambda} = \sum_{i=1}^n z_i \psi(\lambda z_i).
\end{aligned}$$

The elements of Hessian matrix $\mathbf{H}(\boldsymbol{\theta}|\boldsymbol{Y})$ are

$$\begin{aligned}
H_{\xi\xi} &= \frac{\partial}{\partial\xi}\frac{\partial\ell(\boldsymbol{\theta}|\boldsymbol{y})}{\partial\xi} = \frac{-n}{\sigma^2} - \frac{\lambda}{\sigma}\sum_{i=1}^{n} -\eta(\lambda z_i)\frac{-\lambda}{\sigma} = -\frac{n}{\sigma^2} - \frac{\lambda^2}{\sigma^2}\sum_{i=1}^{n}\eta(\lambda z_i), \\
H_{\xi\sigma^2} &= \frac{\partial}{\partial\sigma^2}\frac{\partial\ell(\boldsymbol{\theta}|\boldsymbol{y})}{\partial\xi} = \frac{-1}{2\sigma^3}\sum_{i=1}^{n} z_i + \frac{1}{\sigma}\sum_{i=1}^{n}\frac{-z_i}{2\sigma^2} - \frac{-\lambda}{2\sigma^3}\sum_{i=1}^{n}\psi(\lambda z_i) - \frac{\lambda}{\sigma}\sum_{i=1}^{n} -\eta(\lambda z_i)\frac{-\lambda z_i}{2\sigma^2} \\
&= -\frac{1}{\sigma^3}\sum_{i=1}^{n} z_i + \frac{\lambda}{2\sigma^2}\sum_{i=1}^{n}\psi(\lambda z_i) - \frac{\lambda^2}{2\sigma^3}\sum_{i=1}^{n} z_i\eta(\lambda z_i), \\
H_{\xi\lambda} &= \frac{\partial}{\partial\lambda}\frac{\partial\ell(\boldsymbol{\theta}|\boldsymbol{y})}{\partial\xi} = -\frac{1}{\sigma}\sum_{i=1}^{n}\psi(\lambda z_i) - \frac{\lambda}{\sigma}\sum_{i=1}^{n} -\eta(\lambda z_i) z_i \\
&= -\frac{1}{\sigma}\sum_{i=1}^{n}\psi(\lambda z_i) + \frac{\lambda}{\sigma}\sum_{i=1}^{n} z_i\eta(\lambda z_i), \\
H_{\sigma^2\sigma^2} &= \frac{\partial}{\partial\sigma^2}\frac{\partial\ell(\boldsymbol{\theta}|\boldsymbol{y})}{\partial\sigma^2} \\
&= \frac{n}{2\sigma^4} - \frac{1}{\sigma^4}\sum_{i=1}^{n} z_i^2 + \frac{3\lambda}{4\sigma^4}\sum_{i=1}^{n} z_i\psi(\lambda z_i) - \frac{\lambda}{2\sigma^2}\frac{-\lambda}{2\sigma^2}\sum_{i=1}^{n} z_i^2\left(-\eta(\lambda z_i)\frac{-\lambda z_i}{\sigma}\right) \\
&= \frac{n}{2\sigma^4} - \frac{1}{\sigma^4}\sum_{i=1}^{n} z_i^2 + \frac{3\lambda}{4\sigma^4}\sum_{i=1}^{n} z_i\psi(\lambda z_i) - \frac{\lambda^2}{4\sigma^4}\sum_{i=1}^{n} z_i^2\eta(\lambda z_i), \\
H_{\sigma^2\lambda} &= \frac{\partial}{\partial\lambda}\frac{\partial\ell(\boldsymbol{\theta}|\boldsymbol{y})}{\partial\sigma^2} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n} z_i\psi(\lambda z_i) + \frac{\lambda}{2\sigma^2}\sum_{i=1}^{n} z_i\eta(\lambda z_i) z_i \\
&= -\frac{1}{2\sigma^2}\sum_{i=1}^{n} z_i\psi(\lambda z_i) + \frac{\lambda}{2\sigma^2}\sum_{i=1}^{n} z_i^2\eta(\lambda z_i), \\
H_{\lambda\lambda} &= \frac{\partial}{\partial\lambda}\frac{\partial\ell(\boldsymbol{\theta}|\boldsymbol{y})}{\partial\lambda} = -\sum_{i=1}^{n} z_i\eta(\lambda z_i) z_i = -\sum_{i=1}^{n} z_i^2\eta(\lambda z_i).
\end{aligned}$$

In order to simplify symbols, let

$$a_k = E_Y\Big(Z^k\psi^2(\lambda Z)\Big), \quad b_k = E_Y\Big(Z^k\psi(\lambda Z)\Big), \quad c_k = E_Y\Big(Z^k\eta(\lambda Z)\Big),$$

where

$$Z = \frac{Y-\xi}{\sigma}.$$

Let $X \sim N\left(0, \frac{1}{1+\lambda^2}\right)$, then we have the relationship between $E_X(X^k)$ and b_k:

$$b_k = \frac{\alpha}{\sqrt{1+\lambda^2}}E_X(X^k), \quad \alpha = \sqrt{\frac{2}{\pi}}. \tag{19}$$

Since $\eta(\lambda Z) = \psi(\lambda Z)\big(\lambda Z + \psi(\lambda Z)\big)$, we then have

$$c_k = E_Y\big(Z^k\eta(\lambda Z)\big) = \lambda b_{k+1} + a_k = \alpha\delta\, E_X(X^{k+1}) + a_k, \quad \delta = \frac{\lambda}{\sqrt{1+\lambda^2}}.$$

The elements of the Fisher information are given as follows:

$$\begin{aligned}
I_{\xi\xi} &= E(-H_{\xi\xi}) = \frac{n}{\sigma^2}\left(1+\lambda^2 c_0\right) = \frac{n}{\sigma^2}\left(1+\lambda^2 a_0\right),\\
I_{\xi\sigma^2} &= E(-H_{\xi\sigma^2}) = \frac{n}{\sigma^3}\left(E_Z(Z) - \frac{1}{2}\lambda b_0 + \frac{1}{2}\lambda^2 c_1\right)\\
&= \frac{n}{2\sigma^3}\left(2\alpha\delta - \lambda\left(\frac{\alpha}{\sqrt{1+\lambda^2}} - \frac{\alpha\delta\lambda}{1+\lambda^2} - \lambda a_1\right)\right)\\
&= \frac{n}{2\sigma^3}\left(\alpha\delta + \frac{\alpha\delta\lambda^2}{1+\lambda^2} + \lambda^2 a_1\right) = \frac{n}{2\sigma^3}\left(\frac{\alpha\delta(1+2\lambda^2)}{1+\lambda^2} + \lambda^2 a_1\right),\\
I_{\xi\lambda} &= E(-H_{\xi\lambda}) = \frac{n}{\sigma}\left(b_0 - \lambda c_1\right) = \frac{n}{\sigma}\left(b_0 - \lambda^2 b_2 - \lambda a_1\right)\\
&= \frac{n}{\sigma}\left(\frac{\alpha}{\sqrt{1+\lambda^2}}\left(1 - \frac{\lambda^2}{1+\lambda^2}\right) - \lambda a_1\right) = \frac{n}{\sigma}\left(\frac{\alpha}{(1+\lambda^2)^{3/2}} - \lambda a_1\right),\\
I_{\sigma^2\sigma^2} &= E(-H_{\sigma^2\sigma^2}) = \frac{n}{\sigma^4}\left(-\frac{1}{2} + E_Z(Z^2) - \frac{3}{4}\lambda b_1 + \frac{1}{4}\lambda^2 c_2\right)\\
&= \frac{n}{\sigma^4}\left(-\frac{1}{2} + 1 + \frac{1}{4}\lambda^2 a_2\right) = \frac{n}{4\sigma^2}\left(2 + \lambda^2 a_2\right),\\
I_{\sigma^2\lambda} &= E(-H_{\sigma^2\lambda}) = \frac{n}{2\sigma^2}\left(b_1 - \lambda c_2\right) = -\frac{n\lambda}{2\sigma^2}a_2,\\
I_{\lambda\lambda} &= E(-H_{\lambda\lambda}) = nc_2 = na_2.
\end{aligned}$$

Therefore, the Fisher information can be reexpressed by

$$\begin{aligned}
\boldsymbol{I}(\boldsymbol{\theta}) &= E\Big(-\mathbf{H}(\boldsymbol{\theta}\,|\boldsymbol{Y})\Big) = \begin{bmatrix} I_{\xi\xi} & I_{\xi\sigma^2} & I_{\xi\lambda}\\ I_{\xi\sigma^2} & I_{\sigma^2\sigma^2} & I_{\sigma^2\lambda}\\ I_{\xi\lambda} & I_{\sigma^2\lambda} & I_{\lambda\lambda}\end{bmatrix}\\
&= n\begin{bmatrix} \left(1+\lambda^2 a_0\right)/\sigma^2 & \left(\frac{\alpha\delta(1+2\lambda^2)}{1+\lambda^2} + \lambda^2 a_1\right)/2\sigma^3 & \left(\frac{\alpha}{(1+\lambda^2)^{3/2}} - \lambda a_1\right)/\sigma\\ \left(\frac{\alpha\delta(1+2\lambda^2)}{1+\lambda^2} + \lambda^2 a_1\right)/2\sigma^3 & \left(1+\lambda^2 a_2/2\right)/2\sigma^4 & -\lambda a_2/2\sigma^2\\ \left(\frac{\alpha}{(1+\lambda^2)^{3/2}} - \lambda a_1\right)/\sigma & -\lambda a_2/2\sigma^2 & a_2\end{bmatrix}.
\end{aligned}$$

The proof of (19):

$$\begin{aligned}
b_k &= E_Y\Big(Z^k\psi(\lambda Z)\Big)\\
&= \int_{-\infty}^{\infty} z^k\,\frac{\phi(\lambda z)}{\Phi(\lambda z)}\,\frac{2}{\sigma}\,\phi(z)\,\Phi(\lambda z)\,dy\\
&= \int_{-\infty}^{\infty} z^k\,\phi(\lambda z)\,\frac{2}{\sigma}\,\phi(z)\,dy = 2\int_{-\infty}^{\infty} z^k\,\phi(\lambda z)\,\phi(z)\,dz\\
&= 2\int_{-\infty}^{\infty} z^k(2\pi)^{-1/2}(1+\lambda^2)^{-1/2}\,\phi_X(z)\,dz\\
&= \frac{\alpha}{\sqrt{1+\lambda^2}}\,E_X(X^k).
\end{aligned}$$

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In 2*nd Int. Symp. on Information Theory*, Ed. B. N. Petrov and F. Csaki, pp. 267-81. Budapest: Akademiai Kiado.

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**, 471-88.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scand. J. Statist.* **12**, 171-78.

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. R. Statist. Soc.* **B 65**, 367-89.

Azzalini, A. and Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika* **83**, 715-26.

Basord, K. E., Greenway D. R., McLachlan G. J. and Peel D. (1997). Standard errors of fitted means under normal mixture. *Comp. 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**, 134-41.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *J. Comp. Graph. Statist.* **7**, 434-55.

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

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *J. R. Statist. Soc.* **B 56**, 363-75.

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** 577-88.

Gelman, A., Robert, G. and Gilks, W. (1995). Efficient Metropolis jumping rules. In Bayesian Statistics 5, ed. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith. New York: Oxford University Press.

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

Gupta, A. K., González-Farías G. and Domínguez-Monila, J. A. (2004). A multivariate skew normal distribution. *J. Multivariate Anal.* **89**, 181-90.

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.** 271-75.

Liu, C. H. and Rubin, D. B. (1994). The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. *Biometrika* **81**, 633-48.

Maclean, C. J., Morton, N. E., Elston, R. C. and Yee, S. (1976). Skewness in commingled distributions. *Biometrics* **32**, 695-99.

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

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

Meng, X. L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80**, 267-78.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. R. Statist. Soc.* **B 59**, 731-92.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-4.

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

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.* New York: Wiely.