國 立 交 通 大 學

統計學研究所

碩 士 論 文

混合偏斜*t*分佈及其應用

**On the mixture of skew t distributons
and its applications**

研 究 生 ：謝宛茹
指導教授 ：李昭勝 博士
林宗儀 博士

中 華 民 國 九 十 五 年 六 月

混合偏斜t分佈及其應用

On the mixture of skew t distributions and its applications

研 究 生：謝宛茹		Student ：	Wan-Ju Hsieh
指導教授：李昭勝		Advisors：	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 2006

Hsinchu, Taiwan, Republic of China
中華民國九十五年六月

混合偏斜*t*分佈及其應用

研究生：謝宛茹　　　　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 t 分佈已被認為是混合常態分佈的一種具穩健性的延伸。近年來，處理具異質性並牽涉了具不對稱現象的資料問題中，混合偏斜常態分佈已經被驗證是一種很有效的工具。本文我們提出一種具穩健性的混合偏斜 t 分佈模型來有效地處理當資料同時具有厚尾、偏斜與多峰型式的現象。除此之外，混合常態分佈(NORMIX)、混合 t 分佈(TMIX)與混合偏斜常態分佈(SNMIX)模型皆可視為本篇論文所提出混合偏斜 t 分佈(STMIX)的特例。我們建立一些 EM-types 演算法，以遞迴的方式去求最大概似估計值。最後，我們也透過分析一組實例來闡述我們所提出來方法。

On the mixture of skew t distributions and its applications

Student：Wan-Ju Hsieh Advisors：Dr. Jack C. Lee
Dr. Tsung I. Lin

Institute of Statistic
National Chiao Tung University

ABSTRACT

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

誌謝

研究所兩年的生涯，這快樂、充實、努力的日子匆匆過去了。此時此刻，即將要畢業的我，回想著過去兩年的研究生活、生活點滴，期待與不捨的情緒百感交集。我的心情是充滿了感激與感動。

論文得以順利完成，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對於學生的關心和照顧，讓我如沐春風、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，經由學長的用心指導，才能順利的完成，在這裡要感謝學長不辭辛勞的指導；另外還要感謝所上老師們的用心授課，使我受益良多；接著還要感謝口試委員趙蓮菊老師和彭南夫老師，特地抽空參加並給予我寶貴的意見。謝謝你們。

接著要感謝我的好朋友，包括了婉文、沛君、孟樺、大宛、鶯筑、秀慧和小馬，因為有你們這一群朋友陪伴，為我的研究生生涯帶來了歡笑與淚水，因為有你們的扶持與鼓勵，使得這些日子充滿了溫馨與感動。另外要特別感謝我的室友婉文和沛君，總是在我面臨低潮時，為我加油、打氣，互相鼓舞著對方給彼此動力，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。接著還要感謝我最重要的朋友－泰彬，從大學以來，謝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，使我對自己有更的信心與動力，更有勇氣面對挫折與挑戰。

最後我要感謝我的家人－父親謝日春、母親鄭秀鳳、弟弟國鼎和妹妹蓓雅，感謝他們給我一個溫馨又快樂的家庭，總是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。在鳳凰花開、驪歌輕唱之際，謹以本文獻給所有的至親與好友，與你們分享我的喜悅。

謝宛茹　　謹誌于
國立交通大學統計學研究所
中華民國九十五年六月十日

Contents

List of Tables

List of Figures

On the mixture of skew t distributons and its applications

Student: Wan-Ju Hsieh Advisors: Dr. Jack C. Lee
Dr. Tsung I. Lin

Institute of Statistics
National Chiao Tung University
Hsinchu, Taiwan

Abstract

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

Key words: EM-type algorithms; Heterogeneity data; Maximum likelihood; Outlying observations; Skew t mixtures; Truncated normal.

1. INTRODUCTION

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

While NORMIX and TMIX models have been well recognized as useful in many practical applications, data with varying degrees of extreme skewness among subclasses may not be well modeled. In attempting to appropriately model a set of data arising from a class or several classes with asymmetric observations, Lin, Lee and Yen (2006) recently introduced a new mixture model with each unseen component following a skew normal distribution (Azzalini 1985, 1986). A skew normal mixture (SNMIX) model for a continuous random variable Y is of the form

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

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

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

is a skew normal density function with location parameter $\xi_i \in \mathbb{R}$, scale parameter $\sigma_i^2 > 0$ and skewness parameter $\lambda_i \in \mathbb{R}$. As described in Lin, Lee and Yen (2006), the SNMIX model (1) can be represented by a normal-truncated normal-multinomial hierarchial structure. Such representation leads to a convenient implementation for maximum likelihood (ML) estimation under a complete-data framework.

Although model (1) offers great flexibility in modeling data with varying asymmetric behaviors, it may suffer from a lack of robustness regarding extreme outlying observations. In general, the skewness parameters could be unduly affected by observations that are atypical within components in model (1) being fitted. This motivates us to develop a wider class of mixture distributions to accommodate asymmetry and long tails simultaneously. In this paper, we are devoted to the fitting of mixture of skew t distributions, introduced by Azzalini and Capitaino (2003), allowing heavy tails in addition to skewness as a natural extension of Lin, Lee and Yen (2006). With this skew t mixture (STMIX) model approach, the NORMIX, TMIX and SNMIX models can be treated as special cases in this family.

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

2. PRELIMINARIES

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

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

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

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

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

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

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

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

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

Note that as $\nu \to \infty$, $\tau \to 1$ with probability 1 and $Y = \xi + \sigma Z$. Figure 1 shows the plots of standard skew normal distributions superimposed the standard skew t distributions with $\nu = 5$ under $\lambda = 0, \pm 1, \pm 2, \pm 3$.

Standard algebraic manipulations yield the following:

$$E(Y) = \xi + \frac{\Gamma((\nu-1)/2)}{\Gamma(\nu/2)}\sqrt{\frac{\nu}{\pi}}\delta_\lambda \sigma. \tag{4}$$

$$\mathrm{var}(Y) = \sigma^2\nu\left\{\frac{1}{2}\frac{\Gamma((\nu-2)/2)}{\Gamma(\nu/2)} - \frac{\delta_\lambda^2}{\pi}\left(\frac{\Gamma((\nu-1)/2)}{\Gamma(\nu/2)}\right)^2\right\}. \tag{5}$$

$$\begin{aligned}\gamma_Y = {} & \frac{1}{2}\left\{\pi\delta_\lambda(3-\delta_\lambda^2)\Gamma\left(\frac{\nu-3}{2}\right)\Gamma\left(\frac{\nu}{2}\right)^2\right. \\ & \left. -3\pi\delta_\lambda\Gamma\left(\frac{\nu-2}{2}\right)\Gamma\left(\frac{\nu-1}{2}\right)\Gamma\left(\frac{\nu}{2}\right) + 4\delta_\lambda^3\Gamma\left(\frac{\nu-1}{2}\right)^3\right\} \\ & \times\left\{\frac{\pi}{2}\Gamma\left(\frac{\nu-2}{2}\right)\Gamma\left(\frac{\nu}{2}\right) - \delta_\lambda^2\Gamma\left(\frac{\nu-1}{2}\right)^2\right\}^{-3/2}.\end{aligned} \tag{6}$$

$$\begin{aligned}\kappa_Y = {} & \left\{3\pi^2\Gamma\left(\frac{\nu-4}{2}\right)\Gamma\left(\frac{\nu}{2}\right)^3 - 8\pi\delta_\lambda^2(3-\delta^2)\Gamma\left(\frac{\nu-3}{2}\right)\Gamma\left(\frac{\nu-1}{2}\right)\Gamma\left(\frac{\nu}{2}\right)^2\right. \\ & \left. +12\pi\delta_\lambda^2\Gamma\left(\frac{\nu-2}{2}\right)\Gamma\left(\frac{\nu-1}{2}\right)^2\Gamma\left(\frac{\nu}{2}\right) - 12\delta_\lambda^4\Gamma\left(\frac{\nu-1}{2}\right)^4\right\} \\ & \times\left\{\pi\Gamma\left(\frac{\nu-2}{2}\right)\Gamma\left(\frac{\nu}{2}\right) - 2\delta_\lambda^2\Gamma\left(\frac{\nu-1}{2}\right)^2\right\}^{-2}.\end{aligned} \tag{7}$$

where $\delta_\lambda = \lambda/\sqrt{1+\lambda^2}$, and γ_Y and κ_Y are the measures of skewness and kurtosis, respectively. Figure 2 displays the γ_Y and κ_Y as a function of λ for the standard skew t distribution with $\nu = 5$. The sketch of the proofs of Eqs (4)-(7) are given in Appendix A.

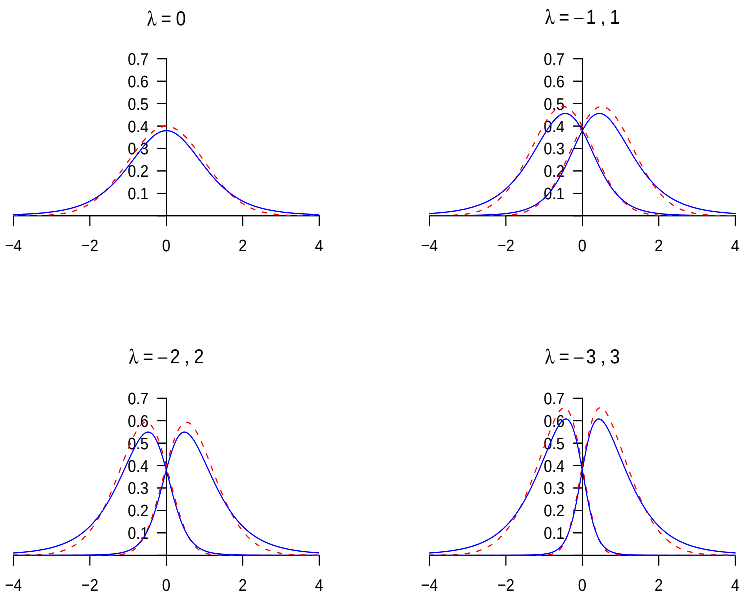

Figure 1: Plots of standard skew normal densities (dashed lines) and standard skew t densities (solid lines) with $\nu = 5$ under various λ.

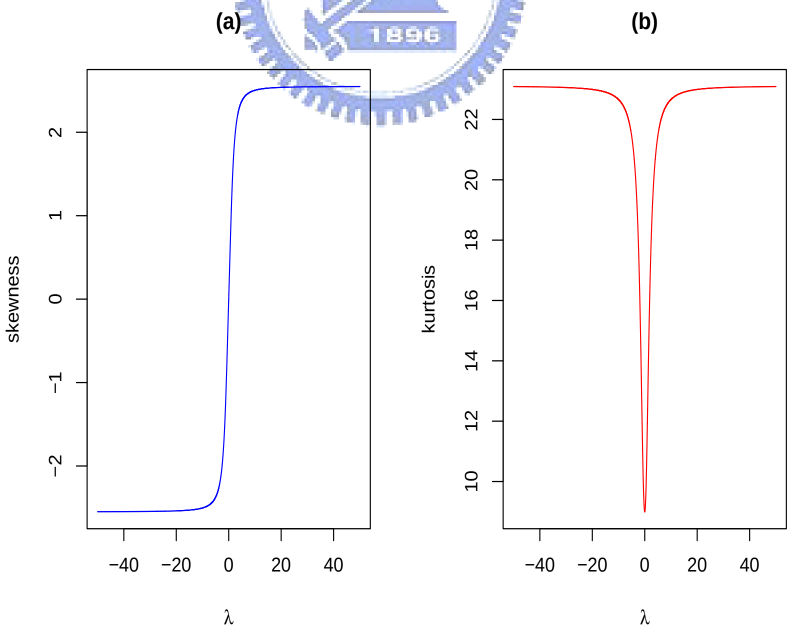

Figure 2: The skewness and kurtosis plots versus λ for the standard skew t distribution with $\nu = 5$.

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

$$\begin{aligned} Y \mid \gamma, \tau &\sim \mathcal{N}\left(\xi + \delta_\lambda \gamma, \frac{1-\delta_\lambda^2}{\tau}\sigma^2\right), \\ \gamma \mid \tau &\sim \mathcal{TN}\left(0, \frac{\sigma^2}{\tau}; [0,\infty)\right), \quad \tau \sim \Gamma(\nu/2, \nu/2), \end{aligned} \tag{8}$$

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

From (8), the joint pdf of Y, γ, τ is given by

$$\begin{aligned} f(\gamma, \tau, y) &= \frac{1}{\pi\sqrt{1-\delta_\lambda^2}\sigma^2}\frac{(\nu/2)^{\frac{\nu}{2}}}{\Gamma(\nu/2)}\tau^{\frac{\nu}{2}}\exp\Big(-\frac{\tau}{2(1-\delta_\lambda^2)}\eta^2\Big)\exp\Big(-\frac{\tau}{2}\nu\Big) \\ &\quad \times \exp\bigg(-\frac{\gamma^2\tau}{2(1-\delta_\lambda^2)\sigma^2} + \frac{\gamma\tau}{(1-\delta_\lambda^2)\sigma^2}\delta_\lambda(y-\xi)\bigg). \end{aligned} \tag{9}$$

Integrating out γ in (9), we get

$$f(\tau, y) = \sqrt{\frac{2}{\pi}}\frac{1}{\sigma}\tau^{\frac{\nu-1}{2}}\frac{(\nu/2)^{\frac{\nu}{2}}}{\Gamma(\nu/2)}\exp\Big(-\frac{\tau}{2}(\eta^2+\nu)\Big)\Phi\Big(\lambda\eta\sqrt{\tau}\Big). \tag{10}$$

Dividing (9) by (10) gives

$$f(\gamma \mid \tau, y) = \frac{1}{\sqrt{2\pi}}\frac{\sqrt{\tau}}{\sigma\sqrt{1-\delta_\lambda^2}}\exp\Big(-\frac{\tau\big(\gamma-(y-\xi)\delta_\lambda\big)^2}{2(1-\delta^2)\sigma^2}\Big)\Phi^{-1}\Big(\lambda\eta\sqrt{\tau}\Big). \tag{11}$$

It follows from (11) that the conditional distribution of γ given τ and Y is

$$\gamma \mid \tau, Y \sim \mathcal{TN}\left(\delta_\lambda(y-\xi), \frac{(1-\delta_\lambda^2)\sigma^2}{\tau}; (0,\infty)\right). \tag{12}$$

From (10), applying Proposition 1 yields the conditional density of τ given Y

$$f(\tau \mid y) = b\tau^{(\nu-1)/2}\exp\Big(-\frac{\tau}{2}(\eta^2+\nu)\Big)\Phi\Big(\lambda\eta\sqrt{\tau}\Big), \tag{13}$$

where

$$b = \left(\frac{\eta^2+\nu}{2}\right)^{(\nu+1)/2}\left\{\Gamma\left(\frac{\nu+1}{2}\right)T_{\nu+1}\left(\lambda\eta\sqrt{\frac{\nu+1}{\eta^2+\nu}}\right)\right\}^{-1} \tag{14}$$

is the normalizing constant.

Proposition 2. *Given the hierarchical representation (8), we have the following:*

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

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

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

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

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

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

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

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

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

and

$$\begin{aligned} g_\nu(x) &= \mathrm{DG}\left(\frac{\nu+2}{2}\right) - \mathrm{DG}\left(\frac{\nu+1}{2}\right) - \log\left(1+\frac{x^2}{\nu+1}\right) \\ &\quad + \frac{(\nu+1)x^2-\nu-1}{(\nu+1)(\nu+1+x^2)}, \end{aligned} \tag{15}$$

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

Proof. See Appendix B.

3. ML ESTIMATION OF THE SKEW t DISTRIBUTION

In this section, we demonstrate how to employ the EM-type algorithms for ML estimation of the skew t distribution, which can be viewed as a single component skew t mixture model that we shall discuss in the next section. From the representation (8), n independent observations from $\mathcal{ST}(\xi, \sigma^2, \lambda, \tau)$ can be expressed by

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

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

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

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

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

$$\begin{aligned} & Q(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(k)}) \\ = \; & -\frac{\nu}{2}\sum_{j=1}^n \hat{s}_{1j}^{(k)} - \sum_{j=1}^n \left(\frac{\eta_j^2 \hat{s}_{1j}^{(k)}}{2(1-\delta_\lambda^2)}\right) + \sum_{j=1}^n \left(\frac{\delta_\lambda \eta_j \hat{s}_{2j}^{(k)}}{(1-\delta_\lambda^2)\sigma}\right) - \sum_{j=1}^n \left(\frac{\hat{s}_{3j}^{(k)}}{2(1-\delta_\lambda^2)\sigma^2}\right) \\ & -n\log\sigma^2 - \frac{n}{2}\log(1-\delta_\lambda^2) + \frac{n\nu}{2}\log\left(\frac{\nu}{2}\right) - n\log\Gamma\left(\frac{\nu}{2}\right) + \frac{\nu}{2}\sum_{j=1}^n \hat{s}_{4j}^{(k)}, \end{aligned} \tag{16}$$

where

$$\hat{s}_{1j}^{(k)} = E(\tau_j|y_j,\hat{\boldsymbol{\theta}}^{(k)}) = \left(\frac{\hat{\nu}^{(k)}+1}{\hat{\eta}_j^{2(k)}+\hat{\nu}^{(k)}}\right)\frac{T_{\hat{\nu}^{(k)}+3}\left(\hat{M}_j^{(k)}\sqrt{\frac{\hat{\nu}^{(k)}+3}{\hat{\nu}^{(k)}+1}}\right)}{T_{\hat{\nu}^{(k)}+1}\left(\hat{M}_j^{(k)}\right)}, \tag{17}$$

$$\begin{aligned}\hat{s}_{2j}^{(k)} &= E(\gamma_j\tau_j|y_j,\hat{\boldsymbol{\theta}}^{(k)})\\ &= \hat{\delta}_\lambda^{(k)}(y_j-\hat{\xi}^{(k)})\hat{s}_{1j}^{(k)}+\frac{\sqrt{1-\hat{\delta}_\lambda^{2(k)}}}{\pi\hat{f}_{Y_j}^{(k)}(y_j)}\left(\frac{\hat{\eta}_j^{2(k)}}{\hat{\nu}^{(k)}(1-\hat{\delta}_\lambda^{2(k)})}+1\right)^{-(\hat{\nu}^{(k)}/2+1)},\end{aligned} \tag{18}$$

$$\begin{aligned}\hat{s}_{3j}^{(k)} &= E(\gamma_j^2\tau_j|y_j,\hat{\boldsymbol{\theta}}^{(k)})\\ &= \hat{\delta}_\lambda^{2(k)}(y_j-\hat{\xi}^{(k)})^2\hat{s}_{1j}^{(k)}+(1-\hat{\delta}_\lambda^{2(k)})\hat{\sigma}^{2(k)}\\ &\quad+\frac{\hat{\delta}_\lambda^{(k)}(y_j-\hat{\xi}^{(k)})\sqrt{1-\hat{\delta}_\lambda^{2(k)}}}{\pi\hat{f}_{Y_j}^{(k)}(y_j)}\left(\frac{\hat{\eta}_j^{2(k)}}{\hat{\nu}^{(k)}(1-\hat{\delta}_\lambda^{2(k)})}+1\right)^{-(\hat{\nu}^{(k)}/2+1)},\end{aligned} \tag{19}$$

and

$$\begin{aligned}\hat{s}_{4j}^{(k)} &= E(\log\tau_j|y_j,\hat{\boldsymbol{\theta}}^{(k)})\\ &= \mathrm{DG}\left(\frac{\hat{\nu}^{(k)}+1}{2}\right)+\frac{\hat{\nu}^{(k)}+1}{\hat{\eta}_j^{2(k)}+\hat{\nu}^{(k)}}\left(\frac{T_{\hat{\nu}^{(k)}+3}\left(\hat{M}_j^{(k)}\sqrt{\frac{\hat{\nu}^{(k)}+3}{\hat{\nu}^{(k)}+1}}\right)}{T_{\hat{\nu}^{(k)}+1}(\hat{M}_j^{(k)})}-1\right)\\ &\quad-\log\left(\frac{\hat{\eta}_j^{2(k)}+\hat{\nu}^{(k)}}{2}\right)+\frac{\hat{\lambda}^{(k)}\hat{\eta}_j^{(k)}(\hat{\eta}_j^{2(k)}-1)}{\sqrt{(\hat{\nu}^{(k)}+1)(\hat{\nu}^{(k)}+\hat{\eta}_j^{2(k)})^3}}\left(\frac{t_{\hat{\nu}^{(k)}+1}(\hat{M}_j^{(k)})}{T_{\hat{\nu}^{(k)}+1}(\hat{M}_j^{(k)})}\right)\\ &\quad+\frac{1}{T_{\hat{\nu}^{(k)}+1}(\hat{M}_j^{(k)})}\int_{-\infty}^{\hat{M}_j^{(k)}}g_{\hat{\nu}^{(k)}}(x)t_{\hat{\nu}^{(k)}+1}(x)dx,\end{aligned} \tag{20}$$

with

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

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

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

E-step: Given $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^{(k)}$, compute $\hat{s}_{1j}^{(k)}$, $\hat{s}_{2j}^{(k)}$, $\hat{s}_{3j}^{(k)}$ and $\hat{s}_{4j}^{(k)}$ in Eqs (17)-(20) for $j = 1, \ldots, n$.

CM-step 1: Update $\hat{\xi}^{(k)}$ by maximizing (16) over ξ, which leads to

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

CM-step 2: Fix $\xi = \hat{\xi}^{(k+1)}$, update $\hat{\sigma}^{2^{(k)}}$ by maximizing (16) over σ^2, which gives

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

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

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

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

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

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

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

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

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

PX.CM-step 2:

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

PX.CM-step 4:

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

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

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

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

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

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

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

4. THE SKEW t MIXTURE MODEL

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

$$\psi(y_j \mid \boldsymbol{\Theta}) = \sum_{i=1}^{g} w_i f(y_j \mid \xi_i, \sigma_i^2, \lambda_i, \nu_i), \quad w_i \geq 0, \quad \sum_{i=1}^{g} w_i = 1, \tag{21}$$

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

$\boldsymbol{Z}_j \sim M(1;\ w_1, \ldots, w_g)$. Note that the rth element $Z_{rj} = 1$ if Y_j arises from the component r. With the inclusion of indicator variables $Z_j's$, a hierarchical representation of (21) is given by

$$\begin{aligned} Y_j \mid \gamma_j, \tau_j, z_{ij} = 1 &\sim N\Big(\xi_i + \delta_{\lambda_i}\gamma_j, \frac{1 - \delta_{\lambda_i}{}^2}{\tau_j}\sigma_i^2\Big), \\ \gamma_j \mid \tau_j, z_{ij} = 1 &\sim TN\Big(0, \frac{\sigma_i^2}{\tau_j}; (0, \infty)\Big), \\ \tau_j \mid z_{ij} = 1 &\sim \Gamma(\nu_i/2, \nu_i/2), \\ \boldsymbol{Z}_j &\sim M(1; w_1, w_2, \ldots, w_g). \end{aligned} \tag{22}$$

It follows from the hierarchical structure (22) on the basis of the observed data $\boldsymbol{y}$ and latent variables $\boldsymbol{\gamma}$, $\boldsymbol{\tau}$ and $\boldsymbol{Z}_j$'s that the complete data log-likelihood function of $\boldsymbol{\Theta}$, ignoring constants, is

$$\begin{aligned} \ell_c(\boldsymbol{\Theta}) = &\sum_{j=1}^{n}\sum_{i=1}^{g} Z_{ij}\Big\{\log w_i - \frac{\nu_i\tau_j}{2} - \frac{\tau_j\eta_{ij}^2}{2(1-\delta_{\lambda_i}^2)} + \frac{\delta_{\lambda_i}\eta_{ij}\gamma_j\tau_j}{(1-\delta_{\lambda_i}^2)\sigma_i} - \frac{\gamma_j{}^2\tau_j}{2(1-\delta_{\lambda_i}^2)\sigma_i^2} \\ &-\frac{1}{2}\log(1-\delta_{\lambda_i}^2) - \log\sigma_i^2 + \frac{\nu_i}{2}\log\frac{\nu_i}{2} - \log\Big(\Gamma\Big(\frac{\nu_i}{2}\Big)\Big) + \frac{\nu_i}{2}\log\tau_j\Big\}, \end{aligned} \tag{23}$$

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

Let $\hat{z}_{ij}^{(k)} = E(Z_{ij}|y_j, \hat{\boldsymbol{\Theta}}^{(k)})$, $\hat{s}_{1ij}^{(k)} = E(Z_{ij}\tau_j|y_j, \hat{\boldsymbol{\Theta}}^{(k)})$, $\hat{s}_{2ij}^{(k)} = E(Z_{ij}\gamma_j\tau_j|y_j, \hat{\boldsymbol{\Theta}}^{(k)})$ $\hat{s}_{3ij}^{(k)} = E(Z_{ij}\gamma_j^2\tau_j|y_j, \hat{\boldsymbol{\Theta}}^{(k)})$ and $\hat{s}_{4ij}^{(k)} = E(Z_{ij}\log(\tau_j)|y_j, \hat{\boldsymbol{\Theta}}^{(k)})$ be the necessary conditional expectations of (23) for obtaining the Q-function at the kth iteration. These expressions, for $i = 1, \ldots, g$ and $j = 1, \ldots, n$, are given by

$$\hat{z}_{ij}^{(k)} = \frac{w_i^{(k)} f(y_j \mid \xi_i^{(k)}, \sigma_i^{2(k)}, \lambda_i^{(k)}, \nu_i^{(k)})}{\psi(y_j|\hat{\boldsymbol{\Theta}}^{(k)})}, \tag{24}$$

$$\hat{s}_{1ij}^{(k)} = \hat{z}_{ij}^{(k)}\left(\frac{\hat{\nu}_i^{(k)} + 1}{\hat{\eta}_{ij}^{2(k)} + \hat{\nu}_i^{(k)}}\right)\frac{T_{\hat{\nu}_i^{(k)}+3}\left(\hat{M}_{ij}^{(k)}\sqrt{\frac{\hat{\nu}_i^{(k)}+3}{\hat{\nu}_i^{(k)}+1}}\right)}{T_{\hat{\nu}_i^{(k)}+1}\left(\hat{M}_{ij}^{(k)}\right)}, \tag{25}$$

$$\hat{s}_{2ij}^{(k)} = \hat{\delta}_{\lambda_i}^{(k)}(y_j - \hat{\xi}_i^{(k)})\hat{s}_{1ij}^{(k)} + \hat{z}_{ij}^{(k)} \left\{ \frac{\sqrt{1-\hat{\delta}_{\lambda_i}^{2(k)}}}{\pi f_{Y_j|\hat{\Theta}_k}(y_j|\hat{\Theta}_k)} \left(\frac{\hat{\eta}_{ij}^{2(k)}}{\hat{\nu}_i^{(k)}(1-\delta_{\lambda_i}^{2(k)})} + 1 \right)^{-(\hat{\nu}_i^{(k)}/2+1)} \right\}, \quad (26)$$

$$\begin{aligned} \hat{s}_{3ij}^{(k)} &= \hat{\delta}_{\lambda_i}^{2(k)}(y_j - \hat{\xi}_i^{(k)})^2\hat{s}_{1ij}^{(k)} + \hat{z}_{ij}^{(k)} \Bigg\{ (1-\hat{\delta}_{\lambda_i}^{2(k)})\hat{\sigma}_i^{2(k)} \\ &\quad + \frac{\hat{\delta}_{\lambda_i}^{(k)}(y_j-\hat{\xi}_i^{(k)})\sqrt{1-\hat{\delta}_{\lambda_i}^{2(k)}}}{\pi f_{Y_j|\hat{\Theta}_k}(y_j|\hat{\Theta}_k)} \left(\frac{\hat{\eta}_{ij}^{2(k)}}{\hat{\nu}_i^{(k)}(1-\hat{\delta}_{\lambda_i}^{2(k)})} + 1 \right)^{-(\hat{\nu}_i^{(k)}/2+1)} \Bigg\}, \end{aligned} \quad (27)$$

and

$$\begin{aligned} \hat{s}_{4ij}^{(k)} &= \hat{z}_{ij}^{(k)} \Bigg\{ \mathrm{DG}\left(\frac{\hat{\nu}_i^{(k)}+1}{2}\right) + \frac{\hat{\nu}_i^{(k)}+1}{\hat{\eta}_{ij}^{2(k)}+\hat{\nu}_i^{(k)}} \left(\frac{T_{\hat{\nu}_i^{(k)}+3}\left(\hat{M}_{ij}^{(k)}\sqrt{\frac{\hat{\nu}_i^{(k)}+3}{\hat{\nu}_i^{(k)}+1}}\right)}{T_{\hat{\nu}_i^{(k)}+1}(\hat{M}_{ij}^{(k)})} - 1 \right) \\ &\quad - \log\left(\frac{\hat{\eta}_{ij}^{2(k)}+\hat{\nu}_i^{(k)}}{2}\right) + \frac{\hat{\lambda}_i^{(k)}\hat{\eta}_{ij}^{(k)}(\hat{\eta}_{ij}^{2(k)}-1)}{\sqrt{(\hat{\nu}_i^{(k)}+1)(\hat{\nu}_i^{(k)}+\hat{\eta}_{ij}^{2(k)})^3}} \frac{t_{\hat{\nu}_i^{(k)}+1}(\hat{M}_{ij}^{(k)})}{T_{\hat{\nu}_i^{(k)}+1}(\hat{M}_{ij}^{(k)})} \\ &\quad + \frac{1}{T_{\hat{\nu}_i^{(k)}+1}(\hat{M}_{ij}^{(k)})} \int_{-\infty}^{\hat{M}_{ij}^{(k)}} g_{\hat{\nu}_i^{(k)}}(x) t_{\hat{\nu}_i^{(k)}+1}(x)dx \Bigg\} \end{aligned} \quad (28)$$

with

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

$\psi(y_j|\hat{\mathbf{\Theta}}^{(k)})$ is $\psi(y_j|\mathbf{\Theta})$ in (21) with $\mathbf{\Theta}$ replaced by $\hat{\mathbf{\Theta}}^{(k)}$ and $g_{\hat{\nu}_i^{(k)}}(x)$ is $g_\nu(x)$ in (15) with ν replaced by $\hat{\nu}_i^{(k)}$. The ECM algorithm for the skew t mixture model is as follows:

E-step: Given $\mathbf{\Theta} = \hat{\mathbf{\Theta}}^{(k)}$, compute $\hat{z}_{ij}^{(k)}$, $\hat{s}_{1ij}^{(k)}$, $\hat{s}_{2ij}^{(k)}$, $\hat{s}_{3ij}^{(k)}$ and $\hat{s}_{4ij}^{(k)}$ in Eqs (24)-(28) for $i = 1, \ldots, g$ and $j = 1, \ldots, n$.

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

CM-step 2: Calculate

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

CM-step 3: Calculate

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

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

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

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

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

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

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

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

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

PX.CM-step 3:

$$\hat{\sigma_i}^{2(k+1)} = \frac{\sum_{j=1}^n \hat{s}_{1ij}^{(k)}(y_j - \hat{\xi_i}^{(k+1)})^2 - 2\hat{\delta_i}^{(k)}\sum_{j=1}^n \hat{s}_{2ij}^{(k)}(y_j - \hat{\xi_i}^{(k+1)}) + \sum_{j=1}^n \hat{s}_{3ij}^{(k)}}{2(1-\hat{\delta_i}^{2(k)})\sum_{j=1}^n \hat{s}_{1ij}^{(k)}}.$$

PX.CM-step 5:

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

Besides being simple in implementation while maintaining the simplicity and stability properties of the EM algorithm, the PX-EM algorithm is desirable since its convergence is always faster and often much faster than the original algorithm. Some additional remarks and explanations regarding the PX-EM algorithm are given in Appendix C.

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

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

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

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

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

$$\hat{u}_{j,w_r} = \frac{\hat{z}_{rj}}{\hat{w}_r} - \frac{\hat{z}_{gj}}{\hat{w}_g},$$

$$\hat{u}_{j,\xi_r} = \frac{\hat{z}_{rj}}{\hat{\sigma}_r}\left(\frac{\hat{\nu}_r+1}{\hat{\eta}_{rj}^2+\hat{\nu}_r}\right)\left[\hat{\eta}_{rj} - \frac{\hat{\lambda}_r\hat{\nu}_r}{\sqrt{(\hat{\nu}_r+1)(\hat{\eta}_{rj}^2+\hat{\nu}_r)}}\frac{t_{\hat{\nu}_r+1}\left(\hat{M}_{rj}\right)}{T_{\hat{\nu}_r+1}\left(\hat{M}_{rj}\right)}\right],$$

$$\hat{u}_{j,\sigma_r} = \frac{\hat{z}_{rj}}{\hat{\sigma}_r}\left[\frac{\hat{\nu}_r(\hat{\eta}_{rj}^2-1)}{\hat{\eta}_{rj}^2+\hat{\nu}_r} - \hat{\eta}_{rj}\frac{\hat{\lambda}_r\hat{\nu}_r}{\hat{\sigma}_r}\sqrt{\frac{\hat{\nu}_r+1}{(\hat{\eta}_{rj}^2+\hat{\nu}_r)^3}}\frac{t_{\hat{\nu}_r+1}\left(\hat{M}_{rj}\right)}{T_{\hat{\nu}_r+1}\left(\hat{M}_{rj}\right)}\right],$$

$$\hat{u}_{j,\lambda_r} = \hat{z}_{rj}\hat{\eta}_{rj}\sqrt{\frac{\hat{\nu}_r+1}{\hat{\eta}_{rj}^2+\hat{\nu}_r}}\frac{t_{\hat{\nu}_r+1}\left(\hat{M}_{rj}\right)}{T_{\hat{\nu}_r+1}\left(\hat{M}_{rj}\right)},$$

$$\begin{aligned}\hat{u}_{j,\nu_r} = {} & \frac{\hat{z}_{rj}}{2}\Bigg[\mathrm{DG}\left(\frac{\hat{\nu}_r+1}{2}\right) - \mathrm{DG}\left(\frac{\hat{\nu}_r}{2}\right) - \log\left(\frac{\hat{\nu}_r+\hat{\eta}_{rj}^2}{\hat{\nu}_r}\right) + \frac{\hat{\eta}_{rj}^2-1}{\hat{\eta}_{rj}^2+\hat{\nu}_r} \\ & + \frac{\hat{\lambda}_r\hat{\eta}_{rj}(\hat{\eta}_{rj}^2-1)}{\sqrt{(\hat{\nu}_r+1)(\hat{\eta}_{rj}^2+\hat{\nu}_r)^3}}\frac{t_{\hat{\nu}_r+1}\left(\hat{M}_{rj}\right)}{T_{\hat{\nu}_r+1}\left(\hat{M}_{rj}\right)} + \frac{1}{T_{\hat{\nu}_r+1}(\hat{M}_{rj})}\int_{-\infty}^{\hat{M}_{rj}} g_{\hat{\nu}_r}(x_j)t_{\hat{\nu}_r+1}(x_j)dx_j\Bigg],\end{aligned}$$

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

5. AN ILLUSTRATIVE EXAMPLE

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

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

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

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

$$\psi(y|\mathbf{\Theta}) = wf(y|\xi_1, \sigma_1^2, \lambda_1, \nu) + (1-\omega)f(y|\xi_2, \sigma_2^2, \lambda_2, \nu). \tag{29}$$

Of course, model (29) will include NORMIX ($\lambda_1 = \lambda_2 = 0;\ \nu = \infty$), TMIX ($\lambda_1 =$

$\lambda_2 = 0$), and SNMIX ($\nu = \infty$) as special cases.

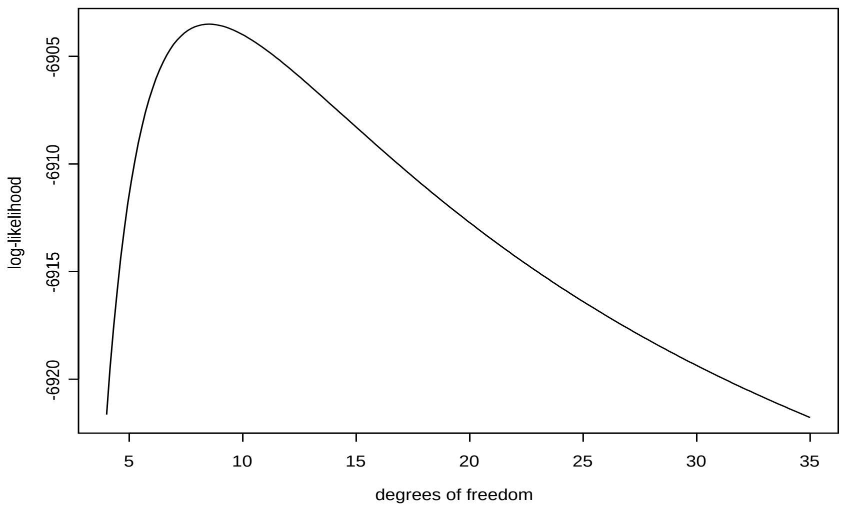

Figure 3: Plot of the profile log-likelihood of the degrees of freedom ν for fitting the bmimen data with a two component STMIX model with equal degrees of freedom ($\nu_1 = \nu_2 = \nu$).

For comparing the fitting results, the ML estimates and the associated information-based standard errors together with the log-likelihood, and AIC and BIC values for NORMIX, TMIX, SNMIX and STMIX models are summarized in Table 1. When comparing these fitted models, we notice that the smaller the AIC and BIC values, the better the fit. It is evidently seen that the STMIX model has the best fitting result. Comparing STMIX with SNMIX, we see that using a heavy-tailed t distribution will reduce the skewness effects. In Figure 3, we plot the profile log-likelihood of the degrees of freedom ν for the STMIX model to illustrate that the SNMIX model is not favorable for this data set since the profile log-likelihood has a significant drop at the peak value of 8.5.

We further consider density estimations using the fitting results of four nested

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

Parameter	NORMIX		TMIX		SNMIX		STMIX	
	mle	se	mle	se	mle	se	mle	se
w	0.397	0.0188	0.438	0.017	0.531	0.013	0.539	0.017
ξ_1	21.443	0.0465	21.591	0.089	19.567	0.036	19.672	0.330
ξ_2	32.565	0.1845	33.030	0.264	28.760	0.009	29.173	0.182
σ_1	2.021	0.0866	1.956	0.083	3.731	0.288	3.482	0.350
σ_2	6.422	0.1584	5.006	0.242	7.960	0.159	6.679	0.232
λ_1	—	—	—	—	1.834	0.344	1.782	0.257
λ_2	—	—	—	—	10.184	2.615	5.912	1.400
ν	—	—	7.075	1.314	—	—	8.502	1.441
m	5		6		7		8	
$\ell(\hat{\boldsymbol{\Theta}})$	-6958.37		-6934.69		-6916.26		-6903.51	
AIC	13926.74		13881.38		13846.52		13823.02	
BIC	13955.04		13915.34		13886.14		13868.30	

AIC=$-2(\ell(\hat{\boldsymbol{\Theta}}) - m)$; BIC=$-2(\ell(\hat{\boldsymbol{\Theta}}) - 0.5m\log(n))$, and m is number of parameters.

mixture models and display them on a single set of coordinate axes in Figure 4. Based on the graphical visualization, we found that the STMIX fitted density is best followed by the SNMIX fitted density. Both NORMIX and TMIX densities do not fit this data set adequately. For further comparison between the two best models, we display the fitted cdfs for both models along with the empirical cdf of the data set in Figure 5. Again, STMIX provides a closer fit to the data since the fitted STMIX cdf tracks the empirical cdf more closely than does the fitted SNMIX.

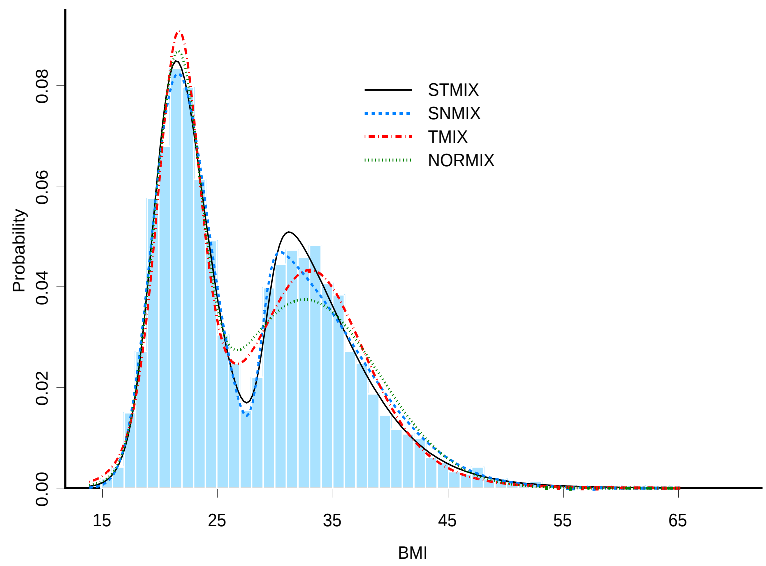

Figure 4: Histogram of the bmimen data with overlaid four ML-fitted two component mixture densities (normal, Student's t, skew normal and skew t).

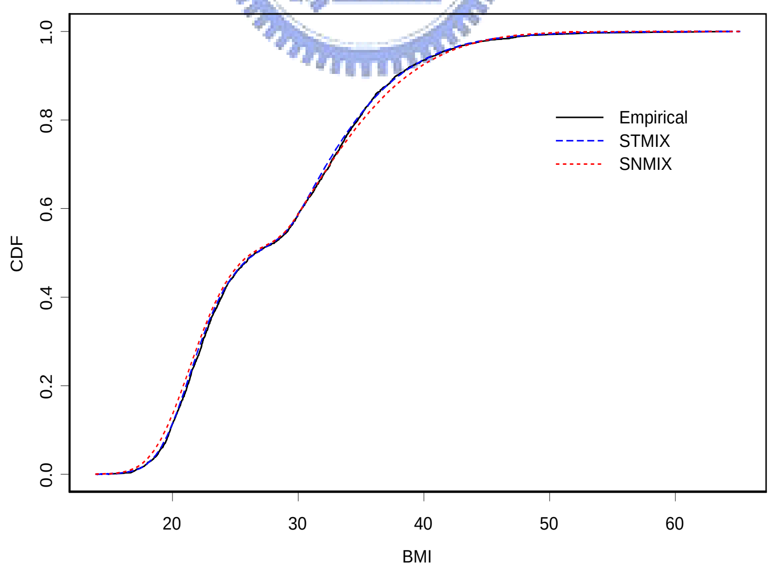

Figure 5: Empirical cdf of the bmimen data together with two superimposed cdfs from the ML-fitted two component SNMIX and STMIX models.

6. CONCLUDING REMARKS

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

Due to recent advances in computational technology, it is worthwhile to carry out Bayesian treatments via Markov chain Monte Carlo (MCMC) sampling methods in the context of STMIX model. The basic idea is to explore the joint posterior distributions of the model parameters together with latent variables $\boldsymbol{\gamma}$ and $\boldsymbol{\tau}$, and allocation variables $\boldsymbol{Z}$ when informative priors are employed. Other extensions of the current work include, for example, a generalization of STMIX to multivariate settings (Azzalini and Capitanio 2003; Jones and Faddy 2003) and determination of the number of components in skew t mixtures via reversible jump MCMC (Green 1995; Richardson and Green 1997; Zhang et al. 2004).

APPENDIX

A. Proofs of Eqs. (4), (5), (6) and (7)

Suppose Y$\sim \mathcal{ST}(\xi, \sigma^2, \lambda, \nu)$, where $Z \sim \mathcal{SN}(\lambda)$, it has following representation:

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

The condition distribution of Y given τ is

$$Y|\tau \sim \mathcal{SN}(\xi, \sigma^2/\tau, \lambda).$$

We then have the following result:

$$\begin{aligned} E(\tau^n) &= \int_0^\infty \tau^n \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} \tau^{\nu/2-1} e^{-\nu/2\tau} d\tau \\ &= \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} \int_0^\infty \tau^{(\nu+2n)/2-1} e^{-\nu/2\tau} d\tau \\ &= \frac{\Gamma((\nu+2n)/2)}{\Gamma(\nu/2)} \left(\frac{\nu}{2}\right)^{-n}. \end{aligned} \tag{A.1}$$

The first four moments of Z are

$$\begin{aligned} &E(Z) = \sqrt{\frac{2}{\pi}}\delta_\lambda, \quad E(Z^2) = 1, \\ &E(Z^3) = \sqrt{\frac{2}{\pi}}\delta_\lambda(3 - \delta_\lambda^2), \quad E(Z^4) = 3. \end{aligned} \tag{A.2}$$

Applying the double expectation trick, in conjunction of (A.1) and (A.2), we have

$$\begin{aligned} E(Y) &= E\Big(E(Y|\tau)\Big) \\ &= E(\xi + \sqrt{\frac{2}{\pi}}\delta_\lambda \frac{\sigma}{\sqrt{\tau}}) \\ &= \xi + \frac{\Gamma((\nu-1)/2)}{\Gamma(\nu/2)} \sqrt{\frac{\nu}{\pi}}\delta_\lambda \sigma. \end{aligned}$$

It is easy to verify

$$
\begin{aligned}
\mathrm{var}(Y) &= E(Y-EY)^2 \\
&= E\left(\xi + \frac{\sigma}{\sqrt{\tau}}Z - \xi - \frac{\Gamma((\nu-1)/2)}{\Gamma(\nu/2)}\sqrt{\frac{\nu}{\pi}}\delta_\lambda\sigma\right)^2 \\
&= E\left(\sigma^2\frac{Z^2}{\tau} - 2\delta_\lambda\sigma^2\sqrt{\frac{\nu}{\pi}}\frac{\Gamma((\nu-1)/2)}{\Gamma(\nu/2)}\frac{Z}{\sqrt{\tau}} + \delta_\lambda^2\sigma^2\frac{\nu}{\pi}(\frac{\Gamma((\nu-1)/2)}{\Gamma(\nu/2)})^2\right) \\
&= \sigma^2\nu\left(\frac{1}{2}\frac{\Gamma((\nu-2)/2)}{\Gamma(\nu/2)} - \frac{\delta_\lambda^2}{\pi}\Big(\frac{\Gamma((\nu-1)/2)}{\Gamma(\nu/2)}\Big)^2\right).
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
E(Y-EY)^3 &= \sigma^3\sqrt{\frac{\nu^3}{\pi}}\left(\frac{1}{2}\delta_\lambda(3-\delta_\lambda^2)\frac{\Gamma((\nu-3)/2)}{\Gamma(\nu/2)}\right. \\
&\quad \left. -\frac{3}{2}\delta_\lambda\frac{\Gamma((\nu-1)/2)\Gamma((\nu-2)/2)}{\Gamma(\nu/2)^2} + \frac{2}{\pi}\delta_\lambda^3\Big(\frac{\Gamma[(\nu-1)/2]}{\Gamma[\nu/2]}\Big)^3\right),
\end{aligned}
$$

and

$$
\begin{aligned}
E(Y-EY)^4 &= \sigma^4\nu^2\left(\frac{3}{4}\frac{\Gamma((\nu-4)/2)}{\Gamma(\nu/2)} - \frac{2}{\pi}\delta_\lambda(3-\delta_\lambda^2)\frac{\Gamma((\nu-1)/2)\Gamma((\nu-3)/2)}{\Gamma(\nu/2)^2}\right. \\
&\quad \left. +\frac{3}{\pi}\delta_\lambda^2\frac{\Gamma((\nu-1)/2)^2\Gamma((\nu-2)/2)}{\Gamma(\nu/2)^3} - \frac{3}{\pi^2}\delta_\lambda^4\Big(\frac{\Gamma((\nu-1)/2)}{\Gamma(\nu/2)}\Big)^4\right).
\end{aligned}
$$

Let γ_Y and κ_Y denote the skewness and kurtosis, respectively. We have

$$
\begin{aligned}
\gamma_Y &= \frac{E(Y-EY)^3}{\big(E(Y-EY)^2\big)^{3/2}} \\
&= \frac{1}{2}\left\{\pi\delta_\lambda(3-\delta_\lambda^2)\Gamma\left(\frac{\nu-3}{2}\right)\Gamma\left(\frac{\nu}{2}\right)^2\right. \\
&\quad \left. -3\pi\delta_\lambda\Gamma\left(\frac{\nu-2}{2}\right)\Gamma\left(\frac{\nu-1}{2}\right)\Gamma\left(\frac{\nu}{2}\right) + 4\delta_\lambda^3\Gamma\left(\frac{\nu-1}{2}\right)^3\right\} \\
&\quad \times\left\{\frac{\pi}{2}\Gamma\left(\frac{\nu-2}{2}\right)\Gamma\left(\frac{\nu}{2}\right) - \delta_\lambda^2\Gamma\left(\frac{\nu-1}{2}\right)^2\right\}^{-3/2},
\end{aligned}
$$

and

$$
\begin{aligned}
\kappa_Y &= \frac{E(Y-EY)^4}{\left(E(Y-EY)^2\right)^2} \\
&= \left\{ 3\pi^2\Gamma\left(\frac{\nu-4}{2}\right)\Gamma\left(\frac{\nu}{2}\right)^3 - 8\pi\delta_\lambda^2(3-\delta^2)\Gamma\left(\frac{\nu-3}{2}\right)\Gamma\left(\frac{\nu-1}{2}\right)\Gamma\left(\frac{\nu}{2}\right)^2 \right. \\
&\quad \left. +12\pi\delta_\lambda^2\Gamma\left(\frac{\nu-2}{2}\right)\Gamma\left(\frac{\nu-1}{2}\right)^2\Gamma\left(\frac{\nu}{2}\right) - 12\delta_\lambda^4\Gamma\left(\frac{\nu-1}{2}\right)^4 \right\} \\
&\quad \times \left\{ \pi\Gamma\left(\frac{\nu-2}{2}\right)\Gamma\left(\frac{\nu}{2}\right) - 2\delta_\lambda^2\Gamma\left(\frac{\nu-1}{2}\right)^2 \right\}^{-2}.
\end{aligned}
$$

B. Proof of Proposition 2

(a) Standard calculation of conditional expectation yields

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

where $\gamma(\cdot|\alpha,\beta)$ denotes the density of $\Gamma(\alpha,\beta)$ and b is given in (14).

By Proposition 1, it suffices to show

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

(b) We first need to show the following:

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

From (12), the expectation of a truncated normal distribution is given by

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

Applying the double expectation trick and using (B.1) and (B.2), we get

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

(c) Similarly, it is easy to verify that

$$E(\gamma^2 \mid y, \tau) = \delta_\lambda^2 (y-\xi)^2 + \frac{(1-\delta_\lambda^2)\sigma^2}{\tau} + \sigma\delta_\lambda(y-\xi)\sqrt{\frac{1-\delta_\lambda^2}{\tau}}\frac{\phi\big(\lambda\eta\sqrt{\tau}\big)}{\Phi\big(\lambda\eta\sqrt{\tau}\big)}. \tag{B.3}$$

Using (B.1) and (B.3), and the double expectation trick as before gives

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

(d) From (13), it is true that

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

By Leibnitz's rule, we can get

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

Hence

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

C. The PX-EM Algorithm

The method of parameter-expansion EM, PX-EM, introduced by Liu, Rubin and Wu (1998), shares the simplicity and stability of ordinary EM, but has a faster rate of convergence. PX-EM algorithm accelerates EM algorithm since its E-step execute a more efficient analysis. PX-EM is to perform a covariance adjustment to correct the analysis of the M step, capitalizing on extra information captured in the imputed complete data.

PX-EM expands the complete data model $f(y_{\text{com}}|\boldsymbol{\theta})$ to a larger model, $f_X(y_{\text{com}}|\boldsymbol{\Theta})$, with $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_*, \alpha\}$, and α is an auxiliary scale parameter whose value is fixed at α_0 in the original model. If the auxiliary parameter α equal to 1, $\{\boldsymbol{\theta}\} = \{\boldsymbol{\theta}_*\}$.

And then, we want to compare ECM algorithm with PX-EM algorithm for ML estimation of skew t distribution.

Model O:

$$\begin{aligned} Y \mid \gamma, \tau &\sim \mathcal{N}\left(\xi + \delta_\lambda \gamma, \frac{1-\delta_\lambda^2}{\tau}\sigma^2\right), \\ \gamma \mid \tau &\sim \mathcal{TN}\left(0, \frac{\sigma^2}{\tau}; [0, \infty)\right), \quad \tau \sim \Gamma(\nu/2, \nu/2), \end{aligned}$$

and $\boldsymbol{\theta} = (\xi, \sigma^2, \lambda, \nu)$ is the parameter of the stew t distribution in ECM algorithm. The results of ECM algorithm are referred to Section 3.

We now derive this modified ECM using PX-EM, and want to adjust current estimates by expanding the parameter:

Model X:

$$\begin{aligned} Y \mid \gamma, \tau &\sim \mathcal{N}\left(\xi_* + \delta_\lambda \gamma, \frac{1-\delta_\lambda^2}{\tau}\sigma_*^2\right), \\ \gamma \mid \tau &\sim \mathcal{TN}\left(0, \frac{\sigma_*^2}{\tau}; [0, \infty)\right), \quad \tau = \alpha\frac{\chi_\nu^2}{\nu} \sim \Gamma(\nu/2, \nu/2), \end{aligned}$$

and $\boldsymbol{\Theta} = (\xi_*, \sigma_*^2, \lambda, \nu, \alpha)$ is the parameter of stew t distribution in PX-EM algorithm.

And

$$(\xi, \sigma^2, \lambda, \nu) = R\{(\xi_*, \sigma_*^2, \lambda, \nu, \alpha)\} = (\xi_*, \sigma_*^2/\alpha, \lambda, \nu)$$

where R is the reduction function from the expanded parameter space to the original parameter space.

Applying routine algebraic manipulations leads to the following CM-step for updating α

$$\hat{\alpha}^{(k+1)} = n^{-1}\sum_{j=1}^{n} \hat{s}_{1j}^{(k)}$$

the application of the reduction function in the PX-EM algorithm leads to adjustments in the estimates of $\sigma^2$and ν, which can be obtained by replacing the CM-steps 2 and 4 in the previous EM algorithm with the following two PX.CM steps:

PX.CM-step2:

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

PX.CM-step4:

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

In the same way, under stew t mixture model, applying routine algebraic manipulations leads to the following CM-step for updating α_i

$$\hat{\alpha}_i^{(k+1)} = \sum_{j=1}^{n}\hat{s}_{1ij}^{(k)} / \sum_{j=1}^{n}\hat{z}_{ij}^{(k)}$$

the application of the reduction function in the PX-EM algorithm leads to adjustments in the estimates of $\sigma_i^2$and ν_i, which can be obtained by replacing the CM-step 3 and 5 in the previous EM algorithm with the following two PX.CM step:

PX.CM-step3:

$$\hat{\sigma_i}^{2(k+1)} = \frac{\sum_{j=1}^{n}\hat{s}_{1ij}^{(k)}(y_j - \hat{\xi_i}^{(k+1)})^2 - 2\hat{\delta_i}^{(k)}\sum_{j=1}^{n}\hat{s}_{2ij}^{(k)}(y_j - \hat{\xi_i}^{(k+1)}) + \sum_{j=1}^{n}\hat{s}_{3ij}^{(k)}}{2(1-\hat{\delta_i}^{2(k)})\sum_{j=1}^{n}\hat{s}_{1ij}^{(k)}}.$$

PX.CM-step5:

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

REFERENCES

Azzalini, A. (1985), "A Class of Distributions Which Includes the Normal Ones," *Scandinavian Journal of Statistics,* 12, 171-178.

Azzalini, A. (1986), "Further Results on a Class of Distributions Which Includes the Normal Ones," *Statistica* 46, 199-208.

Azzalini, A., and Capitaino, A. (2003), "Distributions Generated by Perturbation of Symmetry With Emphasis on a Multivariate Skew t-Distribution," *Journal of the Royal Statistical Society.* Ser. B, 65, 367-389.

Basord, K. E., Greenway D. R., McLachlan G. J., and Peel D. (1997), "Standard Errors of Fitted Means Under Normal Mixture," *Computational Statistics,* 12, 1-17.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion)," *Journal of the Royal Statistical Society.* Ser. B, 39, 1-38.

Flegal, K. M., Carroll, M. D., Ogden, C. L., and Johnson, C. L. (2002), "Prevalence and Trends in Obesity among US Adults, 1999-2000," *Journal of the American Medical Association,* 288, 1723-1727.

Fraley, C, and Raftery, A. E. (2002), "Model-Based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Association,* 97, 611-631.

Green, P. J., (1995), "Reversible Jump MCMC Computation and Bayesian Model Determination," *Biometrika,* 82, 711-732.

Henze, N. (1986), "A Probabilistic Representation of the Skew-Normal Distribution," *Scandinavian Journal of Statistics,* 13, 271-275.

Jones, M. C., and Faddy, M. J. (2003), "A Skew Extension of the t-Distribution, With Applications," *Journal of the Royal Statistical Society.* Ser. B, 65, 159-174.

Lin, T. I., Lee, J. C., and Ni, H. F. (2004), "Bayesian Analysis of Mixture Modelling Using the Multivariate t Distribution," *Statistics and Computing*, 14, 119-130.

Lin, T. I., Lee, J.C., and Yen, S. Y. (2006), "Finite Mixture Modelling Using the Skew Normal Distribution," *Statistica Sinica* (To appear).

Liu, C. H., and Rubin, D. B. (1994), "The ECME Algorithm: a Simple Extension of EM and ECM With Faster Monotone Convergence," *Biometrika*, 81, 633-648.

Liu, C. H., Rubin, D. B., and Wu, Y. (1998). "Parameter Expansion to Accelerate EM: the PX-EM Algorithm," *Biometrika*, 85, 755-770.

McLachlan, G. J. and 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, 267-78.

Peel, D., and McLachlan, G.J. (2000), “Robust Mixture Modeling Using the t Distribution,” *Statistics and Computing* 10, 339-348.

Richardson, S., and Green, P. J. (1997), “On Bayesian Analysis of Mixtures With an Unknown Number of Components (with discussion),” *Journal of the Royal Statistical Society.* Ser. B, 59, 731-792.

Shoham, S. (2002). “Robust Clustering by Deterministic Agglomeration EM of Mixtures of Multivariate t-Distributions,” *Pattern Recognition*, 35, 1127-1142.

Shoham, S., Fellows, M. R., and Normann R. A. (2003), “Robust, Automatic Spike Sorting Using Mixtures of Multivariate t-Distributions,” *Journal of Neuroscience Methods,* 127, 111-122.

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

Wang, H. X., Zhang, Q. B., Luo, B., and Wei, S. (2004), “Robust Mixture Modelling Using Multivariate t Distribution With Missing Information”, *Pattern Recognition Letter,* 25, 701-710.

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

Zhang, Z., Chan, K. L., Wu, Y., and Cen, C. B. (2004), “Learning a Multivariate Gaussian Mixture Model With the Reversible Jump MCMC Algorithm,” *Statistics and Computing,* 14, 343-355.