Revista Colombiana de Estadística
January 2017, Volume 40, Issue 1, pp. 45 to 64
DOI: http://dx.doi.org/10.15446/rce.v40n1.53580

Robust Mixture Regression Based on the Skew t Distribution

Mixtura robusta de modelos de regresión basada en la distribución t asimétrica

Fatma Zehra Doğru[1,a], Olcay Arslan[2,b]

[1]Department of Statistics, Faculty of Arts and Sciences, Giresun University, Giresun, Turkey

[2]Department of Statistics, Faculty of Science, Ankara University, Ankara, Turkey

Abstract

In this study, we explore a robust mixture regression procedure based on the skew t distribution in order to model heavy-tailed and/or skewed errors in a mixture regression setting. We present an EM-type algorithm to compute the maximum likelihood estimators for the parameters of interest using the scale mixture representation of the skew t distribution. The performance of proposed estimators is demonstrated by a simulation study and a real data example.

***Key words*:** EM Algorithm, Maximum Likelihood, Mixture Regression Model, Skew t Distribution.

Resumen

En este estudio se explora una mixtura robusta de modelos de regresión basada en la distribución t asimétrica, con el propósito de modelar colas pesadas o asimétricas en los errores, en un escenario de mixtura de regresiones. Se usa un algoritmo EM para obtener los estimadores máximo verosímiles empleando una mixtura de escala de la distribución t asimétrica. El comportamiento de los estimadores propuestos se ilustra a través de une estudio de simulación y de un ejemplo con datos reales.

***Palabras clave*:** Algoritmo EM, máxima verosimilitud, mixtura de regresiones, distribución t asimétrica.

[a]PhD. E-mail: fatma.dogru@giresun.edu.tr
[b]PhD. E-mail: oarslan@ankara.edu.tr

1. Introduction

Mixture regression models are used to investigate the relationship between variables that come from some unknown latent groups. These models were first introduced by Quandt (1972) and Quandt & Ramsey (1978) as switching regression models and are widely used in areas such as engineering, genetics, biology, econometrics and marketing. The parameter estimation of a mixture regression model is usually based on the normality assumption. It is well-known that the estimators that are based on the normality assumption perform well when the error distribution is normal, but they are very sensitive to departures from normality (outliers, heavy-tailedness, skewness). To deal with the departures from normality, robust mixture regression procedures have been proposed. Some of these works can be summarized as follows: Markatou (2000) and Shen, Yang & Wang (2004) used a weight function to estimate the parameters robustly in the mixture regression models. Bashir & Carter (2012) used the S-estimation method for the mixture linear regression model. Bai (2010) and Bai, Yao & Boyer (2012) proposed a robust estimation procedure based on M-regression estimation to estimate the parameters of the mixture regression model. Wei (2012) and Yao, Wei & Yu (2014) explored the mixture regression model based on t distribution, which is an extension of the mixtures of t distributions studied by Peel & McLachlan (2000). Furthermore, Zhang (2013) studied the robust mixture regression model using the Pearson Type VII distribution, and Song, Yao & Xing (2014) proposed a robust estimation procedure for mixture regression model using the mixtures of Laplace distributions. As it is pointed out by these authors, the robust mixture regression estimation procedure based on the Laplace distribution can be regarded as the application of the least absolute deviation (LAD) regression estimation to the mixture regression models. Liu & Lin (2014) proposed mixture regression model based on the skew normal distribution. Also, Pereira, Marques & da Costa (2012) studied the performance of the estimates procedure for the mixtures of skew normal distributions.

In this paper, we examine a robust mixture regression procedure based on the skew t distribution to efficiently deal with heavy-tailedness and skewness in the mixture regression model setting. This is an extension of the mixtures of skew t distributions proposed by Lin, Lee & Hsieh (2007) to the mixture regression models. We will use the skew t distribution results from the scale mixture of the skew normal distribution that was introduced by Gupta, Chang & Huang (2002), Gupta (2003) and Azzalini & Capitaino (2003). The scale mixture representation of the skew t distribution enables to easily implement an Expectation-Maximization (EM) algorithm to obtain the maximum likelihood (ML) estimators for the parameters of interest in the mixture regression model. For the mixture regression model based on the skew t distribution, refer to the works by Doğru & Arslan (2014) and Doğru (2015).

Recently, Zeller, Cabral & Lachos (2016) have proposed in robust mixture regression model based on scale mixtures of skew normal distributions. They consider the problem in general for the scale mixtures of skew normal distributions and compare the performance of the skew normal, skew t, and skew slash distri-

butions via simulation studies and real data example. In their simulation study and real data example, they assumed equal variance, which differs from our extensive simulation study and real data example. Also, in our paper, we consider the performance of the estimators for the outlier case, apart from heavy-tailedness, (see simulation study Case the V) which is not considered in their paper. We also explore the outlier case in real data example by adding ten extra outliers to the data to illustrate the performance of the estimators that are considered in this study. Furthermore, we compute the standard errors using Fisher information based theory in real data example.

The paper is organized as follows: In Section 2, we give the basic definition of the mixture regression model. In Section 3, we present the robust mixture regression results based on the skew t distribution. In Sections 4 and 5, we give a simulation study and a real data example to compare the performance of the proposed estimation procedure with the other estimation procedures obtained from normal, t (Yao et al. 2014), and skew normal (Liu & Lin 2014) distributions. The paper ends with a conclusion section.

2. Mixture Regression Model

The model setting for a general mixture of linear regression model can be defined as follows. Let $\mathbf{x}$ be a p-dimensional vector of observed values of the explanatory variables, Y be the response variable, and Z be a latent class variable independent of $\mathbf{x}$. Suppose that given $Z = i$, the response variable Y depends on the explanatory variable $\mathbf{x}$ in a linear way

$$Y = \mathbf{x}'\boldsymbol{\beta}_i + \epsilon_i, i = 1, 2, \ldots, g, \tag{1}$$

where $\boldsymbol{\beta}_i = (\beta_{i1}, \beta_{i2}, \ldots, \beta_{ip})'$ is the regression parameters, ϵ_i is the error term, and g is the number of components in the mixture regression model. It is assumed that ϵ_i and $\mathbf{x}$ are independent and $\mathbf{x}$ includes both predictors and constant 1. In the literature, it is often assumed that the random errors (ϵ_i) have distributions from the location-scale family with zero means and σ_i scale parameters. Suppose that $P(Z = i \mid \mathbf{x}) = w_i, i = 1, 2, \ldots, g$ denote the mixing probabilities with $\sum_{i=1}^{g} w_i = 1$, then the conditional density function of Y given $\mathbf{x}$ is

$$f(y \mid \mathbf{x}, \boldsymbol{\Theta}) = \sum_{i=1}^{g} w_i f_i(y; \mathbf{x}'\boldsymbol{\beta}_i, \sigma_i), \tag{2}$$

where $f_i(y; \mathbf{x}'\boldsymbol{\beta}_i, \sigma_i)$ is the probability density function (pdf) of the ith component with some shape parameters (e.g. degrees of freedom for t distribution), and $\boldsymbol{\Theta} = (w_1, \ldots, w_g, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_g, \sigma_1, \ldots, \sigma_g)'$ is the unknown parameter vector. This model is called as a g-component mixture regression model. The ML estimation method is used to estimate the unknown parameter vector $\boldsymbol{\Theta}$ in model (2). Let $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$ be a given sample. Then, the ML estimator of $\boldsymbol{\Theta}$ is obtained by maximizing the following log-likelihood function with respect to $\boldsymbol{\Theta}$

$$\ell(\boldsymbol{\Theta}) = \sum_{j=1}^{n} \log \left(\sum_{i=1}^{g} w_i f_i(y_j; \mathbf{x}_j' \boldsymbol{\beta}_i, \sigma_i) \right). \tag{3}$$

However, it should be noted that the ML estimators cannot be explicitly obtained. The EM algorithm (Dempster, Laird & Rubin 1977) is used to find the ML estimates.

3. Robust Mixture Regression Based on the Skew t Distribution

In this section, we will use the skew t distribution in order to model possible skewed and heavy-tailed errors in the mixture regression model. By doing so, we will obtain more robust estimators for the mixture regression model parameters. We will use the Azzalini type skew t distribution (Gupta et al. 2002, Gupta 2003, Azzalini & Capitaino 2003) with the pdf

$$f(\epsilon; \sigma^2, \lambda, \nu) = \frac{2}{\sigma} t_\nu(\eta) T_{\nu+1}\left(\lambda \eta \sqrt{\frac{\nu+1}{\eta^2+\nu}} \right), \eta = \frac{\epsilon}{\sigma}, \epsilon \in \mathbb{R}, \tag{4}$$

where $\lambda \in \mathbb{R}$ is the skewness parameter, $t_\nu(\cdot)$ is the pdf of the t distribution with $\nu \in (0, \infty)$ degrees of freedom, and $T_{\nu+1}(\cdot)$ is the cumulative density function (cdf) of the t distribution with $\nu + 1$ degrees of freedom.

In the mixture regression model given in (2), assume that the errors have a skew t distribution with zero location, and σ_i^2, λ_i and ν_i scale, skewness, and degrees of freedom parameters, respectively. In contrast to the symmetric case, the mean $E(\epsilon_i) \neq 0$. For the skew t distribution, $E(\epsilon_i) = \sigma_i \delta_{\lambda_i} \sqrt{\frac{\nu_i}{\pi}} \frac{\Gamma(\frac{\nu_i - 1}{2})}{\Gamma(\frac{\nu_i}{2})}$ when $\nu_i > 1$, where $\delta_{\lambda_i} = \lambda_i / \sqrt{(1 + \lambda_i^2)}$. Thus, $E(y_j) = \mathbf{x}_j' \boldsymbol{\beta}_i + E(\epsilon_i)$, which only affects the intercept. Thus, when we estimate the intercept, we will take this into account and correct $\hat{\beta}_0$ by using $\widehat{E(\epsilon_i)}$. In order to estimate the unknown parameters, we should maximize the following log-likelihood function

$$\ell(\boldsymbol{\Theta}) = \sum_{j=1}^{n} \log \left(\sum_{i=1}^{g} w_i f_i(y_j; \mathbf{x}_j' \boldsymbol{\beta}_i, \sigma_i^2, \lambda_i, \nu_i) \right), \tag{5}$$

where $\boldsymbol{\Theta} = (w_1, \ldots, w_g, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_g, \sigma_1^2, \ldots, \sigma_g^2, \lambda_1, \ldots, \lambda_g, \nu_1, \ldots, \nu_g)'$. However, the maximizer of the above log-likelihood function cannot be explicitly obtained, so an EM-type algorithm should be used to estimate $\boldsymbol{\Theta}$. Here, we will use the following EM algorithm to obtain the estimators.

Let $\mathbf{Z}_j = (Z_{1j}, \ldots, Z_{gj})'$ be the latent variables such that

$$Z_{ij} = \begin{cases} 1, \text{ if } j^{th} \text{ observation is from } i^{th} \text{ component} \\ 0, \text{ otherwise,} \end{cases}$$

where $j = 1, \ldots, n$ and $i = 1, \ldots, g$. To simplify the EM algorithm's steps, we will use the stochastic representation of the skew t distribution given by Azzalini & Capitaino (2003) (see Appendix for more detailed explanations). This stochastic representation yields the following hierarchical formulation in terms of the conditional distributions

$$\begin{aligned} Y_j \mid \gamma_j, \tau_j &\sim \mathrm{N}\left(\mathbf{x}_j'\boldsymbol{\beta}_i + \alpha_i\gamma_j, \frac{\kappa_i^2}{\tau_j}\right), \\ \gamma_j \mid \tau_j &\sim \mathrm{TN}\left(0, \frac{1}{\tau_j}; (0,\infty)\right), \\ \tau_j &\sim \mathrm{Gamma}(\nu_i/2, \nu_i/2), \end{aligned}$$

where $\mathrm{TN}(\cdot)$ denotes the truncated normal distribution, $\alpha_i = \sigma_i\delta_{\lambda_i}$, and $\kappa_i^2 = \sigma_i^2(1-\delta_{\lambda_i}^2)$. Then, considering $(\boldsymbol{\gamma}, \boldsymbol{\tau})$ and $\mathbf{z}_j$ are missing data, the complete data log likelihood function for $(\mathbf{y}, \boldsymbol{\gamma}, \boldsymbol{\tau}, \mathbf{z}_j)$ given $\mathbf{X}$ can be written as

$$\begin{aligned} \ell_c(\boldsymbol{\Theta}; \mathbf{y}, \boldsymbol{\gamma}, \boldsymbol{\tau}, \mathbf{z}) = \sum_{j=1}^{n}\sum_{i=1}^{g} z_{ij}\Bigg(&\log w_i - \log\pi - \frac{\log\kappa_i^2}{2} \\ &+ \frac{\nu_i}{2}\log\left(\frac{\nu_i}{2}\right) + \frac{\nu_i}{2}\log\tau_j \\ &- \log\left(\Gamma\left(\frac{\nu_i}{2}\right)\right) - \frac{\nu_i\tau_j}{2} - \frac{(y_j - \mathbf{x}_j^{'}\boldsymbol{\beta}_i - \alpha_i\gamma_j)^2}{2\kappa_i^2/\tau_j} - \frac{\tau_j\gamma_j^2}{2}\Bigg), \end{aligned} \tag{6}$$

where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)', \mathbf{y} = (y_1, \ldots, y_n), \boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)$ and $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n)$. Moreover, based on the theory of the EM algorithm, the conditional expectation of the complete data log-likelihood function, given the observed data and the current parameter estimate $\hat{\boldsymbol{\Theta}}^{(k)}$, should be calculated. That is, we have to find the following conditional expectation

$$\begin{aligned} E\left(\ell_c(\boldsymbol{\Theta}; \mathbf{y}, \boldsymbol{\gamma}, \boldsymbol{\tau}, \mathbf{z}_j) \mid y_j\right) = \sum_{j=1}^{n}\sum_{i=1}^{g} E(Z_{ij} \mid y_j)\Bigg(&\log w_i - \frac{\log\kappa_i^2}{2} + \frac{\nu_i}{2}\log\left(\frac{\nu_i}{2}\right) \\ &- \log\Gamma\left(\frac{\nu_i}{2}\right) + \frac{\nu_i}{2}E(\log\tau_j \mid y_j) - \frac{\nu_i}{2}E(\tau_j \mid y_j) \\ &- \frac{E(\tau_j \mid y_j)\left(y_j - \mathbf{x}_j^{'}\boldsymbol{\beta}_i\right)^2}{2\kappa_i^2} - \frac{\alpha_i^2 E(\tau_j\gamma_j^2 \mid y_j)}{2\kappa_i^2} \\ &+ \frac{\alpha_i E(\tau_j\gamma_j \mid y_j)(y_j - \mathbf{x}_j^{'}\boldsymbol{\beta}_i)}{\kappa_i^2}\Bigg). \end{aligned} \tag{7}$$

The conditional expectations $E(\tau_j \mid y_j)$, $E(\tau_j\gamma_j \mid y_j)$, $E(\tau_j\gamma_j^2 \mid y_j)$, and $E(\log\tau_j \mid y_j)$ can be calculated using the conditional expectations presented in the Appendix. The conditional expectation $E(Z_{ij} \mid y_j)$ can be computed using the classical theory of mixture modeling. Then, the steps of the EM algorithm can be given as follows.

EM algorithm:

1. Take initial parameter estimate $\boldsymbol{\Theta}^{(0)}$ and a stopping rule Δ.

2. E step: Compute the conditional expectations $z_{ij}^{(k)}, s_{1ij}^{(k)}, s_{2ij}^{(k)}, s_{3ij}^{(k)}$ and $s_{4ij}^{(k)}$ for $k = 0, 1, 2, \ldots$ using the following equations for $k = 0, 1, 2, \ldots$ iteration

$$\hat{z}_{ij}^{(k)} = E\left(Z_{ij} \mid y_j, \hat{\boldsymbol{\Theta}}^{(k)}\right) = \frac{\hat{w}_i^{(k)} f_i\left(y_j; \mathbf{x}_j^{'}\hat{\boldsymbol{\beta}}_i^{(k)}, \hat{\sigma}_i^{2(k)}, \hat{\lambda}_i^{(k)}, \hat{\nu}_i^{(k)}\right)}{f\left(y_j; \mathbf{x}_j, \hat{\boldsymbol{\Theta}}^{(k)}\right)}, \tag{8}$$

$$\hat{s}_{1ij} = E\left(Z_{ij}\tau_j \mid y_j, \hat{\boldsymbol{\Theta}}^{(k)}\right) = \hat{z}_{ij}^{(k)}\left(\frac{\hat{\nu}_i^k + 1}{\hat{\eta}_{ij}^{2(k)} + \hat{\nu}_i^{(k)}}\right)\frac{T_{\hat{\nu}_i^{(k)}+3}\left(\hat{M}_{ij}^{(k)}\sqrt{\frac{\hat{\nu}_i^{(k)}+3}{\hat{\nu}_i^{(k)}+1}}\right)}{T_{\hat{\nu}_i^{(k)}+1}\left(\hat{M}_{ij}^{(k)}\right)}, \tag{9}$$

$$\begin{aligned}
\hat{s}_{2j} &= E\left(Z_{ij}\gamma_j\tau_j \mid y_j, \hat{\boldsymbol{\Theta}}^{(k)}\right) \\
&= \frac{\hat{\delta}_{\lambda_i}^{(k)}\left(y_j - \mathbf{x}_j^{'}\hat{\boldsymbol{\beta}}_i^{(k)}\right)\hat{s}_{1ij}^{(k)}}{\hat{\sigma}_i^{(k)}} \\
&+ \frac{\hat{z}_{ij}^{(k)}\sqrt{1-\hat{\delta}_{\lambda_i}^{2(k)}}}{\pi\hat{\sigma}_i^{(k)}\hat{f}(y_j)}\left(\frac{\hat{\eta}_{ij}^{2(k)}}{\hat{\nu}_i^{(k)}\left(1-\hat{\delta}_{\lambda_i}^{2(k)}\right)} + 1\right)^{-\left(\frac{\hat{\nu}_i^{(k)}}{2}+1\right)},
\end{aligned} \tag{10}$$

$$\begin{aligned}
\hat{s}_{3ij}^{(k)} &= E\left(Z_{ij}\gamma_j^2\tau_j \mid y_j, \hat{\boldsymbol{\Theta}}^{(k)}\right) \\
&= \hat{\delta}_{\lambda_i}^{2(k)}\left(\frac{y_j - \mathbf{x}_j^{'}\hat{\boldsymbol{\beta}}_i^{(k)}}{\hat{\sigma}_i^{(k)}}\right)^2\hat{s}_{1ij}^{(k)} + \hat{z}_{ij}^{(k)}\left\{(1-\hat{\delta}_{\lambda_i}^{2(k)})\right. \\
&\left. + \frac{\hat{\delta}_{\lambda_i}^{(k)}\left(y_j - \mathbf{x}_j^{'}\hat{\boldsymbol{\beta}}_i^{(k)}\right)\sqrt{1-\hat{\delta}_{\lambda_i}^{2(k)}}}{\pi\hat{\sigma}_i^{2(k)}\hat{f}(y_j)^{(k)}}\left(\frac{\hat{\eta}_{ij}^{2(k)}}{\hat{\nu}_i^{(k)}(1-\hat{\delta}_{\lambda_i}^{2(k)})} + 1\right)^{-\left(\frac{\hat{\nu}_i^{(k)}}{2}+1\right)}\right\},
\end{aligned} \tag{11}$$

$$
\begin{aligned}
\hat{s}_{4ij}^{(k)} &= E\left(Z_{ij}\log(\tau_j) \mid y_j, \hat{\boldsymbol{\Theta}}^{(k)}\right) \\
&= \hat{z}_{ij}\Bigg\{DG\left(\frac{\hat{\nu}_i^{(k)}+1}{2}\right) - \log\left(\frac{\hat{\eta}_{ij}^{2(k)}+\hat{\nu}_i^{(k)}}{2}\right) \\
&+ \left(\frac{\hat{\nu}_i^{(k)}+1}{\hat{\eta}_{ij}^{2(k)}+\hat{\nu}_i^{(k)}}\right)\left(\frac{T_{\hat{\nu}_i^{(k)}+3}\left(\hat{\lambda}_i^{(k)}\hat{\eta}_{ij}^{(k)}\sqrt{\frac{\hat{\nu}_i^{(k)}+3}{\hat{\nu}_i^{(k)}+\hat{\eta}_{ij}^{2(k)}}}\right)}{T_{\hat{\nu}_i^{(k)}+1}\left(\hat{\lambda}_i^{(k)}\hat{\eta}_{ij}^{(k)}\sqrt{\frac{\hat{\nu}_i^{(k)}+1}{\hat{\nu}_i^{(k)}+\hat{\eta}_{ij}^{2(k)}}}\right)} - 1\right) \\
&+ \frac{\hat{\lambda}_i^{(k)}\hat{\eta}_{ij}^{(k)}(\hat{\eta}_{ij}^{2(k)}-1)}{\sqrt{(\hat{\nu}_i^{(k)}+1)(\hat{\nu}_i^{(k)}+\hat{\eta}_{ij}^{2(k)})^3}}\frac{t_{\hat{\nu}_i^{(k)}+1}\left(\hat{\lambda}_i^{(k)}\hat{\eta}_{ij}^{(k)}\sqrt{\frac{\hat{\nu}_i^{(k)}+1}{\hat{\nu}_i^{(k)}+\hat{\eta}_{ij}^{2(k)}}}\right)}{T_{\hat{\nu}_i^{(k)}+1}\left(\hat{\lambda}_i^{(k)}\hat{\eta}_{ij}^{(k)}\sqrt{\frac{\hat{\nu}_i^{(k)}+1}{\hat{\nu}_i^{(k)}+\hat{\eta}_{ij}^{2(k)}}}\right)} \\
&+ \frac{1}{T_{\hat{\nu}_i^{(k)}+1}\left(\hat{\lambda}_i^{(k)}\hat{\eta}_{ij}^{(k)}\sqrt{\frac{\hat{\nu}_i^{(k)}+1}{\hat{\nu}_i^{(k)}+\hat{\eta}_{ij}^{2(k)}}}\right)}\int_{-\infty}^{\hat{M}_{ij}^{(k)}} g_{\hat{\nu}_i^{(k)}}(x)t_{\hat{\nu}_i^{(k)}+1}(x)dx\Bigg\},
\end{aligned}
\tag{12}
$$

where $DG(\cdot) = \frac{\Gamma'(\cdot)}{\Gamma(\cdot)}$ is the digamma function and

$$
\hat{\eta}_{ij}^{(k)} = \frac{(y_j - \mathbf{x}_j'\hat{\boldsymbol{\beta}}_i^{(k)})}{\hat{\sigma}_i^{(k)}}, \hat{\delta}_{\lambda_i^{(k)}} = \frac{\hat{\lambda}_i^{(k)}}{\sqrt{1+\hat{\lambda}_i^{2(k)}}}, \hat{M}_{ij}^{(k)} = \hat{\lambda}_i^{(k)}\hat{\eta}_{ij}^{(k)}\sqrt{\frac{\hat{\nu}_i^{(k)}}{\hat{\nu}_i^{(k)}+\hat{\eta}_{ij}^{2(k)}}},
$$

$$
g_{\hat{\nu}}(x) = DG(\frac{\hat{\nu}+2}{2}) - DG(\frac{\hat{\nu}+1}{2}) - \log\left(1+\frac{x^2}{\hat{\nu}+1}\right) + \frac{x^2(\hat{\nu}+1)-\hat{\nu}-1}{(\hat{\nu}+1)(\hat{\nu}+1+x^2)},
$$

$$
\hat{f}(y_j)^{(k)} = \sum_{i=1}^{g}\hat{w}_i^{(k)}\frac{2}{\hat{\sigma}_i^{(k)}}t_{\hat{\nu}_i^{(k)}}(\hat{\eta}_{ij}^{(k)})T_{\hat{\nu}_i^{(k)}+1}(\hat{M}_{ij}^{(k)}).
$$

Then, we form the following objective function $Q\left(\boldsymbol{\Theta}; \hat{\boldsymbol{\Theta}}^{(k)}\right)$

$$
\begin{aligned}
Q\left(\boldsymbol{\Theta}; \hat{\boldsymbol{\Theta}}^{(k)}\right) = \sum_{j=1}^{n}\sum_{i=1}^{g}\hat{z}_{ij}^{(k)}\Bigg(&\log w_i - \frac{1}{2}\log(\kappa_i^2) + \frac{\nu_i}{2}\log\left(\frac{\nu_i}{2}\right) \\
&- \log\left(\Gamma\left(\frac{\nu_i}{2}\right)\right)\Bigg) \\
&- \frac{\nu_i\hat{s}_{1ij}^{(k)}}{2} + \frac{\nu_i\hat{s}_{4ij}^{(k)}}{2} - \frac{\hat{s}_{1ij}^{(k)}(y_j - \mathbf{x}_j^{'}\boldsymbol{\beta}_i)^2}{2\kappa_i^2} \\
&+ \frac{\alpha_i\hat{s}_{2ij}^{(k)}(y_j - \mathbf{x}_j^{'}\boldsymbol{\beta}_i)}{\kappa_i^2} - \frac{\alpha_i^2\hat{s}_{3ij}^{(k)}}{2\kappa_i^2}.
\end{aligned}
\tag{13}
$$

3. M step 1: Maximize the $Q\left(\boldsymbol{\Theta};\hat{\boldsymbol{\Theta}}^{(k)}\right)$ with respect to the unknown parameters $(w_i,\boldsymbol{\beta}_i,\sigma_i^2)$, assuming that (λ_i,ν_i) are fixed, in order to obtain $(k+1)$th values for the parameters $(w_i,\boldsymbol{\beta}_i,\sigma_i^2)$. This maximization yields

$$\hat{w}_i^{(k+1)}=\frac{\sum\limits_{j=1}^{n}\hat{z}_{ij}^{(k)}}{n}, \tag{14}$$

$$\hat{\boldsymbol{\beta}}_i^{(k+1)}=\left(\sum_{j=1}^{n}\hat{s}_{1ij}^{(k)}\mathbf{x}_j\mathbf{x}_j^{'}\right)^{-1}\left(\sum_{j=1}^{n}\left(y_j\hat{s}_{1ij}^{(k)}-\hat{\delta}_{\lambda_i}^{(k)}\hat{s}_{2ij}^{(k)}\right)\mathbf{x}_j\right), \tag{15}$$

$$\hat{\alpha}_i^{(k+1)}=\frac{\sum\limits_{j=1}^{n}\hat{s}_{2ij}^{(k)}(y_j-\mathbf{x}_j^{'}\hat{\boldsymbol{\beta}}_i^{(k)})}{\sum\limits_{j=1}^{n}\hat{s}_{3ij}^{(k)}}, \tag{16}$$

$$\hat{\kappa}_i^{2(k+1)}=\frac{\sum\limits_{j=1}^{n}\hat{s}_{1ij}^{(k)}(y_j-\mathbf{x}_j^{'}\hat{\boldsymbol{\beta}}_i^{(k)})^2-2\hat{\alpha}_i^{(k)}\hat{s}_{2ij}^{(k)}(y_j-\mathbf{x}_j^{'}\hat{\boldsymbol{\beta}}_i^{(k)})+\hat{\alpha}_i^{2(k)}\hat{s}_{2ij}^{(k)}}{\sum\limits_{j=1}^{n}\hat{z}_{ij}^{(k)}}, \tag{17}$$

$$\hat{\sigma}_i^{2(k+1)}=\hat{\kappa}_i^{2(k+1)}+\hat{\alpha}_i^{2(k+1)}. \tag{18}$$

4. M step 2: Using the new values for $(w_i,\boldsymbol{\beta}_i,\sigma_i^2)$ that were obtained in M step 1, the following equations are solved to obtain new estimates for the parameters (λ_i,ν_i)

$$\begin{aligned}\delta_{\lambda_i}(1-\delta_{\lambda_i}^2)\sum_{j=1}^{n}\hat{z}_{ij}^{(k)}-\delta_{\lambda_i}&\left(\sum_{j=1}^{n}\hat{s}_{1ij}^{(k)}\frac{(y_j-\mathbf{x}_j'\hat{\boldsymbol{\beta}}_i^{(k+1)})^2}{\hat{\sigma}_i^{2(k+1)}}+\sum_{j=1}^{n}\hat{s}_{3ij}^{(k)}\right)\\&+(1+\delta_{\lambda_i}^2)\sum_{j=1}^{n}\hat{s}_{2ij}^{(k)}\frac{(y_j-\mathbf{x}_j'\hat{\boldsymbol{\beta}}_i^{(k+1)})}{\hat{\sigma}_i^{(k+1)}}=0,\end{aligned} \tag{19}$$

$$\log\left(\frac{\nu_i}{2}\right)+1-DG\left(\frac{\nu_i}{2}\right)+\frac{\sum_{j=1}^{n}\left(\hat{s}_{4ij}^{(k)}-\hat{s}_{1ij}^{(k)}\right)}{\sum_{j=1}^{n}\hat{z}_{ij}^{(k)}}=0. \tag{20}$$

Also the $(k+1)$th estimate of λ_i can be obtained by using the following equation

$$\hat{\lambda}_i^{(k+1)}=\hat{\delta}_{\lambda_i}^{(k+1)}\,/\,\sqrt{1-\hat{\delta}_{\lambda_i}^{2(k+1)}}, \tag{21}$$

where $\hat{\delta}_{\lambda_i}^{(k+1)}=\hat{\alpha}_i^{(k+1)}\,/\,\hat{\sigma}_i^{(k+1)}$.

5. Repeat E and M steps until the convergence criteria $\| \hat{\boldsymbol{\Theta}}^{(k+1)} - \hat{\boldsymbol{\Theta}}^{(k)} \| < \Delta$ is satisfied. Alternatively, the absolute difference of the actual log-likelihood $\| \ell(\hat{\boldsymbol{\Theta}}^{(k+1)}) - \ell(\hat{\boldsymbol{\Theta}}^{(k)}) \| < \Delta$ or $\| \ell(\hat{\boldsymbol{\Theta}}^{(k+1)}) \,/\, \ell(\hat{\boldsymbol{\Theta}}^{(k)}) \| < \Delta$ can be used (see Dias & Wedel 2004).

Note that the equation given in (20) can be used to estimate the degrees of freedom of the skew t distribution. However, for the sake of robustness, we will assume that the degrees of freedom are fixed throughout this paper. We take all the degrees of freedom as 2. This is suggested by Lange, Little & Taylor (1989). Also, it is pointed out by Lucas (1997) that the estimators based on the t distribution are not locally robust when the degrees of freedom are estimated.

4. Simulation Study

In this section, we present a simulation study to demonstrate the performance of the proposed mixture regression model based on skew t distribution (MixregST) over the mixture regression model based on normal distribution (MixregN), mixture regression model based on t distribution (Mixregt) and mixture regression model based on skew normal distribution (MixregSN) in terms of bias and mean squared error (MSE). The formulas of bias and MSE are given

$$\begin{aligned} \widehat{bias}(\hat{\theta}) &= \bar{\theta} - \theta, \\ \widehat{MSE}(\hat{\theta}) &= \frac{1}{N} \sum_{i=1}^{N} (\hat{\theta}_i - \theta)^2, \end{aligned}$$

where θ is the true parameter value, $\hat{\theta}_i$ is the estimate of θ for the ith simulated data, $\bar{\theta} = \frac{1}{N} \sum_{i=1}^{N} \hat{\theta}_i$, and $N = 500$ is the number of replicates. In the simulation study, the sample sizes are taken as 200 and 400. The simulation study and real data example are conducted using $MATLAB$ $R2013a$. For all numerical calculations, the stopping rule Δ is taken as 10^{-6}.

We generate the data $\{(x_{1j}, x_{2j}, y_j), j = 1, \ldots, n\}$ from the following two component mixture regression models (Bai et al. 2012)

$$Y = \begin{cases} 0 + X_1 + X_2 + \epsilon_1, Z = 1, \\ 0 - X_1 - X_2 + \epsilon_2, Z = 2, \end{cases}$$

where $P(Z = 1) = 0.25 = w_1, X_1 \sim \mathrm{N}(0, 1)$ and $X_2 \sim \mathrm{N}(0, 1)$. Furthermore, the model coefficients are $\boldsymbol{\beta}_1 = (\beta_{10}, \beta_{11}, \beta_{12})^{'} = (0, 1, 1)^{'}$ and $\boldsymbol{\beta}_2 = (\beta_{20}, \beta_{21}, \beta_{22})^{'} = (0, -1, -1)^{'}$.

We take the following error distributions:
Case I: $\epsilon_1, \epsilon_2 \sim \mathrm{N}(0, 1)$, standard normal distribution.
Case II: $\epsilon_1, \epsilon_2 \sim t_3(0, 1)$, t distribution with the degrees of freedom 3.
Case III: $\epsilon_1, \epsilon_2 \sim 0.95\mathrm{N}(0, 1) + 0.05\mathrm{N}(0, 25)$, contaminated normal distribution.
Case IV: $\epsilon_1, \epsilon_2 \sim \mathrm{ST}(0, 1, 0.5, 3)$, skew t distribution.

Case V: $\epsilon_1, \epsilon_2 \sim \mathrm{N}(0,1)$, standard normal distribution with %5 outliers, $X_1 = 20, X_2 = 20$ and $Y = 100$.

We use Case I to compare the estimators with the traditional MLE (MixregN) when the error terms have the normal distribution and there are no outliers. Case II is the example for the heavy-tailed error distribution case. The distribution given in Case III is to create outliers. This distribution is often considered in literature as an outlier model. Case IV is to examine the behavior of the estimators when the error term is skewed and heavy-tailed. Case V is considered to test the performances of the estimators to deal with the high leverage points. In this case %5 of the observations are replaced by $X_1 = 20, X_2 = 20$ and $Y = 100$.

Table 1 and 2 show the simulation results for the sample sizes 200 and 400. The tables include the MSEs, and the biases of the parameter estimates, and the true parameter values. We can observe from the simulation study results that the MixregN has the best result in Case I. Moreover, the other estimators obtained from Mixregt, MixregSN, and MixregST have similar performances when the errors have a normal distribution. In Case II, Mixregt performs best, as expected. Also, MixregST has a lower bias and MSE values compared to the MixregN and MixregSN for almost all cases. For Case III, MixregN and MixregSN are drastically affected by the contamination. However, Mixregt and MixregST perform better than the other estimators and Mixregt is comparable with the MixregST. Similarly, MixregN and MixregSN have the worst performance and Mixregt, and MixregST have similar performance in Case IV. Finally, in the outlier case, all estimators are affected by the outliers. However, Mixregt and MixregST have the lowest bias and MSE values in almost all cases. In summary, concerning all the estimators, the Mixregt and MixregST are resistant to the skewness and the heavy tailedness in the data, and they behave better than MixregN and MixregSN in the case of outliers in x direction.

TABLE 1: MSE (bias) values of estimates for $n = 200$.

	MixregN	Mixregt	MixregSN	MixregST
	Case I:$\epsilon_1, \epsilon_2 \sim \mathrm{N}(0,1)$			
β_{10}:0	0.0456 (0.0150)	0.0587 (0.0134)	0.1726 (−0.3560)	0.1317 (0.2306)
β_{20}:0	0.0090 (0.0019)	0.0098 (0.0039)	0.1447 (−0.3678)	0.0575 (−0.2084)
β_{11}:1	0.0348 (−0.0013)	0.0495 (−0.0064)	0.0349 (−0.0016)	0.0546 (−0.0036)
β_{21}:−1	0.0085 (−0.0004)	0.0103 (0.0031)	0.0085 (−0.0004)	0.0118 (0.0212)
β_{12}:1	0.0401 (−0.0243)	0.0483 (−0.0308)	0.0401 (−0.0242)	0.0617 (−0.0296)
β_{22}:−1	0.0089 (−0.0062)	0.0107 (0.0024)	0.0089 (−0.0062)	0.0125 (0.0201)
w:0.25	0.0021 (0.0079)	0.0023 (0.0059)	0.0021 (0.0079)	0.0035 (−0.0063)
	Case II:$\epsilon_1, \epsilon_2 \sim t_3(0,1)$			
β_{10}:0	11.5674 (−0.2939)	0.0930 (−0.0121)	11.6586 (−0.9305)	0.3151 (0.2406)
β_{20}:0	1.2217 (0.0796)	0.0136 (−0.0050)	1.3914 (−0.5527)	0.1397 (−0.3327)
β_{11}:1	7.6108 (0.4273)	0.0959 (−0.0180)	7.6526 (0.3704)	0.1415 (0.0036)
β_{21}:−1	1.2984 (−0.0331)	0.0145 (−0.0064)	1.2011 (0.0192)	0.0171 (0.0259)
β_{12}:1	8.2789 (0.1660)	0.0981 (0.0027)	8.2956 (0.2624)	0.1678 (0.0282)
β_{22}:−1	1.9409 (0.1250)	0.0137 (−0.0031)	1.6075 (0.0762)	0.0167 (0.0283)
w:0.25	0.0226 (−0.0372)	0.0033 (0.0112)	0.0214 (−0.0352)	0.0055 (−0.0067)
	Case III:$\epsilon_1, \epsilon_2 \sim 0.95\mathrm{N}(0,1) + 0.05\mathrm{N}(0,25)$			
β_{10}:0	6.0158 (−0.0052)	0.0634 (−0.0062)	6.1249 (−0.6206)	0.1517 (0.2053)
β_{20}:0	0.6299 (0.0054)	0.0118 (−0.0080)	0.6282 (−0.5670)	0.0911 (−0.2711)
β_{11}:1	4.5781 (0.2371)	0.0599 (0.0078)	4.8849 (0.2067)	0.0727 (0.0119)
β_{21}:−1	0.2236 (0.0418)	0.0106 (−0.0068)	0.1302 (0.0649)	0.0124 (0.0155)
β_{12}:1	2.9126 (−0.0271)	0.0620 (0.0021)	2.7706 (0.0830)	0.0774 (0.0192)
β_{22}:−1	0.1607 (0.0628)	0.0090 (0.0033)	0.0614 (0.0778)	0.0108 (0.0250)
w:0.25	0.0167 (−0.0472)	0.0026 (0.0039)	0.0136 (−0.0526)	0.0034 (−0.0098)
	Case IV:$\epsilon_1, \epsilon_2 \sim \mathrm{ST}(0,1,0.5,3)$			
β_{10}:0	8.4499 (1.0601)	0.2783 (0.4422)	6.1264 (0.3167)	0.9691 (0.7550)
β_{20}:0	0.3472 (0.4787)	0.1524 (0.3759)	0.1323 (−0.0886)	0.0231 (0.0590)
β_{11}:1	2.9291 (0.2448)	0.0851 (−0.0296)	2.7053 (0.2225)	0.1605 (−0.0107)
β_{21}:−1	0.0600 (0.0432)	0.0120 (−0.0133)	0.0540 (0.0381)	0.0146 (0.0230)
β_{12}:1	5.9774 (−0.1412)	0.0862 (−0.0195)	5.6460 (−0.0863)	0.1911 (0.0005)
β_{22}:−1	0.0789 (0.0798)	0.0115 (−0.0029)	0.0731 (0.0715)	0.0154 (0.0336)
w:0.25	0.0125 (−0.0296)	0.0033 (0.0118)	0.0116 (−0.0260)	0.0050 (−0.0156)
	Case V:$\epsilon_1, \epsilon_2 \sim \mathrm{N}(0,1)$ (% 5 outliers)			
β_{10}:0	2.2247 (0.1553)	1.3245 (0.1820)	2.5926 (−0.4879)	5.9114 (2.1745)
β_{20}:0	0.0146 (0.0111)	0.0106 (0.0072)	0.2401 (−0.4728)	0.0392 (−0.1678)
β_{11}:1	3.2773 (1.5211)	2.8341 (1.5030)	3.3162 (1.5107)	2.6095 (1.4250)
β_{21}:−1	0.0833 (0.2528)	0.0234 (0.1077)	0.0826 (0.2519)	0.0296 (0.1283)
β_{12}:1	3.1162 (1.4674)	2.7897 (1.4869)	3.2436 (1.4870)	2.7237 (1.4655)
β_{22}:−1	0.0798 (0.2482)	0.0225 (0.1055)	0.0786 (0.2472)	0.0281 (0.1244)
w:0.25	0.0093 (−0.0937)	0.0061 (−0.0751)	0.0094 (−0.0939)	0.0112 (−0.1029)

Note: Value in parentheses indicates the bias.

Table 2: MSE (bias) values of estimates for $n = 400$.

	MixregN	Mixregt	MixregSN	MixregST
	Case I:$\epsilon_1, \epsilon_2 \sim \mathrm{N}(0,1)$			
β_{10}:0	0.0203 (0.0088)	0.0265 (0.0114)	0.1564 (−0.3687)	0.0782 (0.2081)
β_{20}:0	0.0043 (0.0044)	0.0050 (0.0058)	0.1427 (−0.3716)	0.0565 (−0.2211)
β_{11}:1	0.0149 (0.0008)	0.0192 (−0.0019)	0.0149 (0.0009)	0.0227 (0.0034)
β_{21}:−1	0.0040 (−0.0028)	0.0048 (0.0016)	0.0040 (−0.0027)	0.0057 (0.0197)
β_{12}:1	0.0160 (−0.0100)	0.0213 (−0.0185)	0.0161 (−0.0100)	0.0245 (−0.0110)
β_{22}:−1	0.0044 (0.0009)	0.0053 (0.0070)	0.0044 (0.0010)	0.0065 (0.0247)
w:0.25	0.0012 (0.0035)	0.0013 (0.0006)	0.0012 (0.0035)	0.0018 (−0.0123)
	Case II:$\epsilon_1, \epsilon_2 \sim t_3(0,1)$			
β_{10}:0	14.3296 (−0.3312)	0.0365 (−0.0137)	14.2254 (−0.9701)	0.0830 (0.1669)
β_{20}:0	0.6052 (0.0125)	0.0066 (−0.0066)	0.6330 (−0.6752)	0.1411 (−0.3601)
β_{11}:1	10.6597 (0.4839)	0.0321 (−0.0054)	10.1135 (0.4010)	0.0427 (0.0239)
β_{21}:−1	0.5987 (0.0527)	0.0068 (−0.0052)	0.1809 (0.0921)	0.0083 (0.0272)
β_{12}:1	12.1779 (0.3384)	0.0334 (0.0052)	11.8293 (0.5888)	0.0421 (0.0288)
β_{22}:−1	1.5732 (0.0903)	0.0062 (−0.0041)	0.9058 (0.0454)	0.0078 (0.0273)
w:0.25	0.0161 (−0.0602)	0.0014 (0.0049)	0.0143 (−0.0591)	0.0020 (−0.0134)
	Case III:$\epsilon_1, \epsilon_2 \sim 0.95\mathrm{N}(0,1) + 0.05\mathrm{N}(0,25)$			
β_{10}:0	4.6683 (−0.0431)	0.0287 (0.0004)	5.1651 (−0.7278)	0.0729 (0.1830)
β_{20}:0	0.0088 (0.0037)	0.0056 (−0.0012)	0.3555 (−0.5817)	0.0848 (−0.2769)
β_{11}:1	4.2093 (0.1003)	0.0229 (0.0038)	4.2202 (0.0962)	0.0278 (0.0214)
β_{21}:−1	0.0313 (0.0872)	0.0053 (0.0024)	0.0319 (0.0875)	0.0066 (0.0243)
β_{12}:1	3.2445 (0.1817)	0.0251 (0.0166)	3.1090 (0.1611)	0.0327 (0.0303)
β_{22}:−1	0.0328 (0.0886)	0.0054 (0.0064)	0.0325 (0.0878)	0.0069 (0.0292)
w:0.25	0.0093 (−0.0572)	0.0014 (−0.0020)	0.0093 (−0.0570)	0.0019 (−0.0160)
	Case IV:$\epsilon_1, \epsilon_2 \sim \mathrm{ST}(0,1,0.5,3)$			
β_{10}:0	7.8868 (0.9770)	0.2082 (0.4344)	5.1906 (0.0754)	0.4395 (0.6371)
β_{20}:0	0.2110 (0.4604)	0.1461 (0.3853)	0.0373 (−0.1476)	0.0105 (0.0455)
β_{11}:1	5.0109 (0.1370)	0.0247 (−0.0192)	5.6461 (0.1695)	0.0400 (0.0140)
β_{21}:−1	0.0280 (0.0717)	0.0053 (−0.0065)	0.0259 (0.0686)	0.0066 (0.0263)
β_{12}:1	6.6126 (0.3814)	0.0301 (−0.0120)	7.0245 (0.3604)	0.0485 (0.0140)
β_{22}:−1	0.0308 (0.0723)	0.0049 (−0.0044)	0.0276 (0.0691)	0.0069 (0.0290)
w:0.25	0.0081 (−0.0459)	0.0014 (0.0040)	0.0073 (−0.0436)	0.0021 (−0.0168)
	Case V:$\epsilon_1, \epsilon_2 \sim \mathrm{N}(0,1)$ (% 5 outliers)			
β_{10}:0	1.5208 (0.2485)	1.0975 (0.2305)	1.6056 (−0.3105)	6.9413 (2.5194)
β_{20}:0	0.0094 (0.0158)	0.0059 (0.0056)	0.2483 (−0.4883)	0.0419 (−0.1880)
β_{11}:1	2.6872 (1.4449)	2.4663 (1.4533)	2.6444 (1.4307)	2.3970 (1.4530)
β_{21}:−1	0.0783 (0.2591)	0.0175 (0.1066)	0.0770 (0.2572)	0.0239 (0.1284)
β_{12}:1	2.9720 (1.5383)	2.7078 (1.5341)	3.0209 (1.5560)	2.3044 (1.4204)
β_{22}:−1	0.0813 (0.2646)	0.0176 (0.1072)	0.0810 (0.2639)	0.0230 (0.1279)
w:0.25	0.0098 (−0.0974)	0.0069 (−0.0814)	0.0098 (−0.0976)	0.0138 (−0.1159)

Note: Value in parentheses indicates the bias.

5. Real Data Example

In this section, we will analyze the tone perception data set (Cohen 1984) in order to further illustrate the performance of the mixture regression model based on the skew t distribution on a real data set. In Cohen (1984) tone perception experiment, a pure fundamental tone was played to a trained musician. Also, electronically obtained overtones were added, which were determined by a stretching ratio. This ratio is between the adjusted tone and the fundamental tone. In the experiment, 150 trials were performed by the same musicians. The aim of this experiment was to find out how the tuning ratio affects the perception of the tone and to decide if either of two musical perception theories was reasonable (see Cohen (1984) for more detailed explanations). This data set can be accessed by using a `fpc` package (Henning 2013) in R. The variable perceived tone ratio is taken as the response variable and the actual tone ratio variable is taken as the explanatory variable. This data set has also been analyzed by Yao et al. (2014) and Song et al. (2014) to test the performance of the mixture regression estimates based on the t and Laplace distributions, respectively. Figure 1 shows the scatter plot and the histogram of the perceived tone ratio. From these plots, it is clear that there are two groups in the data; non-normality is also shows.

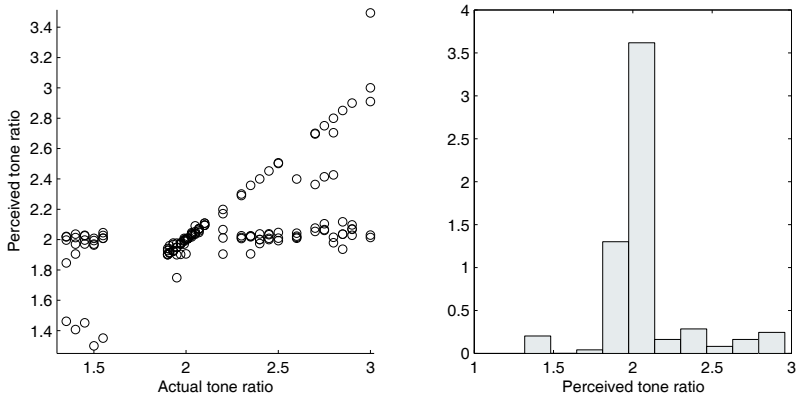

FIGURE 1: (a) The scatter plot of the data. (b) Histogram of the perceived tone ratio.

We use this data set to compare the estimator's performances both with and without outliers cases. For the comparison of the mixture regression models, we use the Akaike information criterion (AIC) (Akaike 1973), consistent AIC (CAIC) (Bozdogan 1993), and the Bayesian information criterion (BIC) (Schwarz 1978) values, which have the following form

$$-2\ell(\hat{\boldsymbol{\Theta}}) + mc_n,$$

where $\ell(\cdot)$ is the maximized log-likelihood, m is the number of parameters to be estimated, and c_n is the penalty term. We use $c_n = 2$ for AIC, $c_n = \log(n)$ for BIC and $\log(n) + 1$ for CAIC. We present the scatter plots with the fitted regression lines obtained from the MixregN, Mixregt, MixregSN, and MixregST procedures in

Figure 2 for the tone perception data set. Also, we summarize the ML estimates, standard errors (SE) of estimates, and some information criteria in Table 3. For all mixture regression models, the standard errors of estimates are calculated using the Fisher information-based method (see Basford, Greenway, McLachlan & Peel 1997). Note that in real data example, we set the normal mixture regression estimates as initial values for the mixing probability and regression coefficients. We also take small value from skewness parameters and we assume that in both groups the degrees of freedom are equal to 2. We try other values of degrees of freedom and get similar results. We observe that MixregST has the best fit than the other mixture regression models in terms of AIC, CAIC, and BIC criterion values.

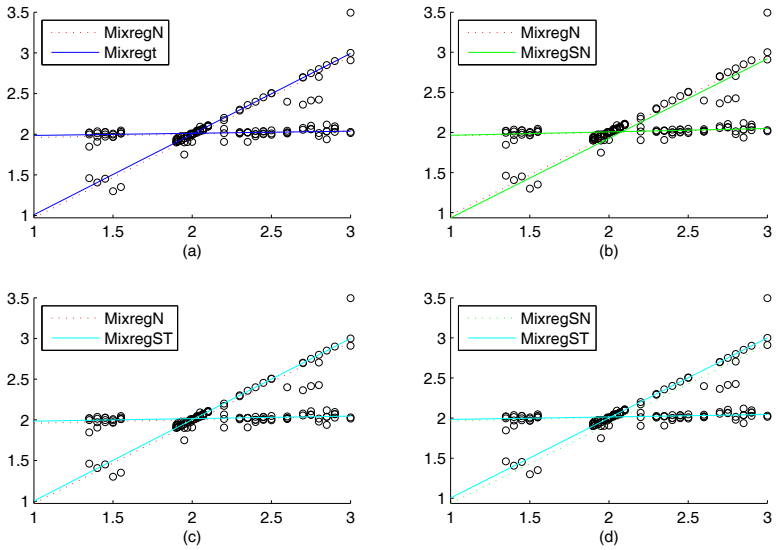

FIGURE 2: Fitted mixture regression lines for the tone perception data set. (a): dashed line- MixregN, solid line-Mixregt, (b): dashed line- MixregN, solid line-MixregSN, (c): dashed line- MixregN, solid line-MixregST, (d): dashed line-MixregSN, solid line-MixregST.

Next, we add ten pairs of outliers at (0,5). These points are shown in Figure 3 by an asterisk. These outliers can be considered as high leverage points. By adding these points, we want to see the performance of the estimators against the high leverage points. Figure 3 displays the scatter plots of the data set with the fitted regression lines obtained from MixregN, Mixregt, MixregSN, and MixregST procedures. The Table 4 presents the ML estimation results. We can see that MixregN and MixregSN are drastically affected by the high leverage points. On the other hand, the estimators based on the t and the skew t distributions (Mixregt and MixregST) give fits to the majority of the data without influencing from the high leverage points. Also, MixregST gives the best results in terms of information criteria. Note that the estimates, including the estimates for skewness parameters with and without outliers, are very similar (see Tables 3 and 4).

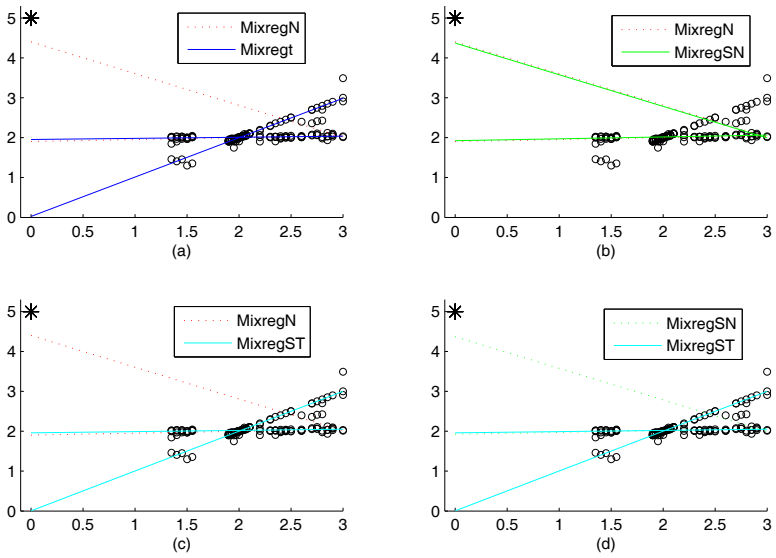

FIGURE 3: Fitted mixture regression lines with ten outliers at (0,5). (a): dashed line-MixregN, solid line-Mixregt; (b): dashed line- MixregN, solid line-MixregSN; (c): dashed line- MixregN, solid line-MixregST; (d): dashed line-MixregSN, solid line-MixregST.

TABLE 3: ML estimates, SE of estimates and some information criteria for fitting mixture regression models to the tone perception data set.

	MixregN		Mixregt		MixregSN		MixregST	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
$\hat{\beta}_{10}$	1.91637	0.02259	1.95857	0.01755	1.92157	1.08982	1.95232	0.05878
$\hat{\beta}_{11}$	0.04254	0.01044	0.02642	0.00782	0.04254	0.01053	0.03094	0.02299
$\hat{\beta}_{20}$	−0.01927	0.12571	0.01776	0.03975	−0.05439	0.49340	0.00544	0.01118
$\hat{\beta}_{21}$	0.99229	0.04721	0.99181	0.01835	0.99096	0.69809	0.99815	0.00331
$\hat{\sigma}_1$	0.04619	0.00382	0.02805	0.00389	0.04648	0.11867	0.03903	0.00192
$\hat{\sigma}_2$	0.13283	0.00836	0.02096	0.00328	0.13817	0.13592	0.00327	0.00335
$\hat{\lambda}_1$	-	-	-	-	−0.4106	29.89308	−0.22245	0.00003
$\hat{\lambda}_2$	-	-	-	-	0.36875	5.71451	0.44809	0.28316
$\hat{w}_1$	0.69772	0.04724	0.55182	0.05352	0.69775	0.06488	0.64037	0.43502
$\ell(\hat{\boldsymbol{\Theta}})$	141.19840		190.81770		141.24909		**211.65935**	
AIC	−268.39680		−367.63541		−264.49818		**−405.31870**	
$CAIC$	−240.32235		−339.56096		−228.40247		**−369.22298**	
BIC	−247.32235		−346.56096		−237.40247		**−378.22298**	

Note: Bold value indicates the smallest values of AIC, CAIC and BIC.

TABLE 4: ML estimates, SE of estimates and some information criteria for fitting mixture regression models to the tone perception data set with ten outliers at (0,5).

	MixregN		Mixregt		MixregSN		MixregST	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
$\hat{\beta}_{10}$	1.90577	0.02686	1.95289	0.02635	1.92316	1.56013	1.96201	0.03431
$\hat{\beta}_{11}$	0.04707	0.01294	0.02877	0.01177	0.04722	0.01745	0.02960	0.01683
$\hat{\beta}_{20}$	4.40096	0.41131	0.02512	0.04855	4.37160	4.75098	0.00567	0.00324
$\hat{\beta}_{21}$	-0.79538	0.17740	0.98808	0.02157	-0.79334	0.24530	0.99810	0.00189
$\hat{\sigma}_1$	0.05060	0.00391	0.03999	0.00426	0.05946	0.15277	0.05356	0.00504
$\hat{\sigma}_2$	0.85912	0.14761	0.02795	0.00496	1.06486	2.30277	0.00306	0.00002
$\hat{\lambda}_1$	-	-	-	-	0.12257	33.27402	-0.27627	0.19334
$\hat{\lambda}_2$	-	-	-	-	0.57263	7.37487	0.45105	0.48747
$\hat{w}_1$	0.73677	0.03747	0.60833	0.05431	0.73624	0.07519	0.67532	0.05843
$\ell(\hat{\boldsymbol{\Theta}})$	54.09971		77.57685		39.83069		**108.07958**	
AIC	−94.19942		−141.15371		−61.66139		**−198.15916**	
$CAIC$	−65.67320		−112.62749		−24.98483		**−161.48260**	
BIC	−72.67320		−119.62749		−33.98483		**−170.48260**	

Note: Bold value indicates the smallest values of AIC, CAIC and BIC.

6. Conclusions

In this paper, we have explored a robust mixture regression procedure based on the skew t distribution. For the proposed mixture regression model, we have given an EM-type algorithm to compute the estimates. We have presented a simulation study to compare the performance of the estimators based on the skew t distribution with the performance of estimators obtained from mixture regression model based on normal, t, and skew normal distributions. The simulation results confirm that when heavy-tailedness and skewness are present, the proposed estimators behave better than their counterparts. We have also given a real data example to further illustrate the capability of the proposed estimators in dealing with the outliers and/or high leverage points in the data. Likewise, for the real data, our proposed estimators show superiority over the estimators based on normal, t and, skew normal.

Acknowledgements

We would like to thank two anonymous referees and the Associate Editor for their valuable comments and suggestions that have greatly improved the paper.

[Received: April 2015 — Accepted: February 2016]

References

Akaike, H. (1973), Information theory and an extension of the maximum likelihood principle, *in* B. N. Petrov & F. Caski, eds, 'Proceeding of the Second International Symposium on Information Theory', Akademiai Kiado, Budapest, pp. 267–281.

Azzalini, A. (1986), 'Further results on a class of distributions which includes the normal ones', *Statistica* **46**, 199–208.

Azzalini, A. & Capitaino, A. (2003), 'Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution', *Journal of the Royal Statistical Society: Series B* **65**, 367–389.

Bai, X. (2010), Robust mixture of regression models, Master's thesis, Kansas State University.

Bai, X., Yao, W. & Boyer, J. E. (2012), 'Robust fitting of mixture regression models', *Computational Statistics and Data Analysis* **56**, 2347–2359.

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

Bashir, S. & Carter, E. (2012), 'Robust mixture of linear regression models', *Communications in Statistics-Theory and Methods* **41**, 3371–3388.

Bozdogan, H. (1993), Choosing the number of component clusters in the mixture model using a new informational complexity criterion of the inverse-fisher information matrix, *in* 'Information and Classification', Springer Berlin Heidelberg, pp. 40–54.

Cohen, A. C. (1984), 'Some effects of inharmonic partials on interval perception', *Music Perception* **1**, 323–349.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the E-M algorithm', *Journal of the Royal Statistical Society: Series B* **39**, 1–38.

Dias, J. G. & Wedel, M. (2004), 'An empirical comparison of em, sem and mcmc performance for problematic gaussian mixture likelihoods', *Statistics and Computing* **14**, 323–332.

Doğru, F. Z. (2015), Robust Parameter Estimation in Mixture Regression Models, PhD thesis, Ankara University.

Doğru, F. Z. & Arslan, O. (2014), Robust mixture regression modelling based on the skew t distribution, *in* 'International Conference on Robust Statistics (ICORS14)', Martin-Luther-University Halle-Wittenberg/Germany.

Gupta, A. (2003), 'Multivariate skew t distribution', *Statistics* **37**, 359–363.

Gupta, A., Chang, F. & Huang, W. (2002), 'Some skew symmetric models', *Random Operators Stochastic Equations* **10**, 133–140.

Henning, C. (2013), *fpc: Flexible procedure for clustering*. R Package Version 2.1-5.

Henze, N. (1986), 'A probabilistic representation of the skew-normal distribution', *Scandinavian Journal of Statistics* **13**, 271–275.

Lange, K. L., Little, J. A. & Taylor, M. G. J. (1989), 'Robust statistical modeling using the t distribution', *Journal of the American Statistical Association* **84**, 881–896.

Lin, T. I., Lee, J. C. & Hsieh, W. J. (2007), 'Robust mixture modeling using the skew t distribution', *Statistics and Computing* **17**, 81–92.

Liu, M. & Lin, T. I. (2014), 'A skew-normal mixture regression model', *Educational and Psychological Measurement* **74**(1), 139–162.

Lucas, A. (1997), 'Robustness of the student t based m-estimator', *Communications in Statistics: Theory and Methods* **26**, 1165–1182.

Markatou, M. (2000), 'Mixture models, robustness, and the weighted likelihood methodology', *Biometrics* **56**, 483–486.

Peel, D. & McLachlan, G. J. (2000), 'Robust mixture modelling using the t distribution', *Statistics and Computing* **10**(4), 339–348.

Pereira, J. R., Marques, L. A. & da Costa, J. M. (2012), 'An empirical comparison of EM initialization methods and model choice criteria for mixtures of skew normal distributions', *Revista Colombiana de Estadistica* **35**(3), 457–478.

Quandt, R. E. (1972), 'A new approach to estimating switching regressions', *Journal of the American Statistical Association* **67**, 306–310.

Quandt, R. E. & Ramsey, J. B. (1978), 'Estimating mixtures of normal distributions and switching regressions', *Journal of the American Statistical Association* **73**, 730–752.

Schwarz, G. (1978), 'Estimating the dimension of a model', *Annals of Statistics* **6**(2), 461–464.

Shen, H., Yang, J. & Wang, S. (2004), Outlier detecting in fuzzy switching regression models, *in* 'International Conference on Artificial Intelligence: Methodology, Systems, and Applications', Springer, pp. 208–215.

Song, W., Yao, W. & Xing, Y. (2014), 'Robust mixture regression model fitting by laplace distribution', *Computational Statistics and Data Analysis* **71**, 128–137.

Wei, Y. (2012), Robust mixture regression models using t-distribution, Master's thesis, Kansas State University.

Yao, W., Wei, Y. & Yu, C. (2014), 'Robust mixture regression using the t-distribution', *Computational Statistics and Data Analysis* **71**, 116–127.

Zeller, C. B., Cabral, C. R. B. & Lachos, V. H. (2016), 'Robust mixture regression modeling based on scale mixtures of skew normal distributions', *Test* **25**, 375–396.

Zhang, J. (2013), Robust mixture regression modeling with pearson type vii distribution, Master's thesis, Kansas State University.

Appendix

If a random variable Y has the skew t distribution ($\mathrm{ST}(\xi, \sigma^2, \lambda, \nu)$) with the location parameter $\xi \in \mathbb{R}$, scale parameter $\sigma^2 \in (0, \infty)$, skewness parameter $\lambda \in \mathbb{R}$ and degrees of freedom ν, it has the following stochastic representation (Azzalini & Capitaino 2003)

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

where Z and τ are independent and SN shows the skew normal distribution, respectively. Also, we can further give the following stochastic representation for Z, which has already given by (Azzalini 1986, p. 201) and (Henze 1986, Theorem 1)

$$Z = \delta_\lambda \mid U_1 \mid + \sqrt{1 - \delta_\lambda^2} U_2,$$

where U_1 and U_2 are independent standard normal random variables and $\mid U_1 \mid$ will have truncated normal distribution. This stochastic representation can be used to obtain the following conditional distributions

$$Y \mid \gamma, \tau \sim \mathrm{N}(\xi + \sigma\delta_\lambda\gamma, \frac{1 - \delta_\lambda^2}{\tau}\sigma^2),$$
$$\gamma \mid \tau \sim \mathrm{TN}(0, \frac{1}{\tau}; (0, \infty)).$$

These conditional distributions will help us to undertake the steps of the EM algorithm. According to Proposition 2 of Lin et al. (2007), we can obtain the following conditional expectations for $\tau, \gamma\tau, \gamma^2\tau$, and $\log(\tau)$ given $Y = y$

$$E(\tau \mid y) = \left(\frac{\nu + 1}{\eta^2 + \nu}\right) \frac{T_{\nu+3}\left(M\sqrt{\frac{\nu+3}{\nu+1}}\right)}{T_{\nu+1}(M)},$$
$$E(\gamma\tau \mid y) = \delta_\lambda \frac{(y - \xi)}{\sigma} E(\tau \mid y) + \frac{\sqrt{1 - \delta_\lambda^2}}{\pi\sigma f(y)} \left(\frac{\eta^2}{\nu(1 - \delta_\lambda^2)} + 1\right)^{-(\frac{\nu}{2}+1)},$$

$$
\begin{aligned}
E(\gamma^2\tau \mid y) &= \delta_\lambda^2 \frac{(y-\xi)^2}{\sigma^2} E(\tau \mid y) + (1-\delta_\lambda^2) \\
&\quad + \frac{\delta_\lambda (y-\xi)\sqrt{1-\delta_\lambda^2}}{\pi\sigma^2 f(y)} \left(\frac{\eta^2}{\nu(1-\delta_\lambda^2)} + 1 \right)^{-\frac{\nu}{2}+1}, \\
E(\log\tau \mid y) &= DG\left(\frac{\nu+1}{2}\right) - \log\left(\frac{\eta^2+\nu}{2}\right) \\
&\quad + \left(\frac{\nu+1}{\eta^2+\nu}\right)\left(\frac{T_{\nu+3}\left(\lambda\eta\sqrt{\frac{\nu+3}{\eta^2+\nu}}\right)}{T_{\nu+1}\left(\lambda\eta\sqrt{\frac{\nu+1}{\eta^2+\nu}}\right)} - 1\right) \\
&\quad + \frac{\lambda\eta(\eta^2-1)}{\sqrt{(\nu+1)(\eta^2+\nu)^3}} \frac{t_{\nu+1}\left(\lambda\eta\sqrt{\frac{\nu+1}{\eta^2+\nu}}\right)}{T_{\nu+1}\left(\lambda\eta\sqrt{\frac{\nu+1}{\eta^2+\nu}}\right)} \\
&\quad + \frac{1}{T_{\nu+1}\left(\lambda\eta\sqrt{\frac{\nu+1}{\eta^2+\nu}}\right)} \int_{-\infty}^{M} g_\nu(x) t_{\nu+1}(x) dx,
\end{aligned}
$$

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

$$
g_\nu(x) = DG\left(\frac{\nu+2}{2}\right) - DG\left(\frac{\nu+1}{2}\right) - \log\left(1+\frac{x^2}{\nu+1}\right) + \frac{x^2(\nu+1)-\nu-1}{(\nu+1)(\nu+1+x^2)}.
$$

Note that these conditional expectations will be used in the EM algorithm presented in Section 3.