

Robust Mixture Regression Using Mixture of Different Distributions

Fatma Zehra Dođru and Olcay Arslan

1 Introduction

Mixture regression models were first introduced by Quandt (1972), Quandt and Ramsey (1978) as switching regression models, which are used to explore the relationship between variables that come from some unknown latent groups. These models are widely applied in areas such as engineering, genetics, biology, econometrics and marketing. These mixture regression models are used to model data sets which contain heterogeneous groups. Figure 1 shows the scatter plots of this type of real data sets used in literature. A pure fundamental tone electronically obtained overtones added was played to a trained musician in the tone perception data which given by Cohen (1984) in Fig. 1a. The overtones were determined by a stretching ratio which is between the adjusted tone and the fundamental tone. 150 trials were performed by the same musicians in this experiment. This experiment was to reveal how the tuning ratio affects the perception of the tone and to choose if either of two musical perception theories was reasonable (see Cohen 1984 for more detail). The other data contains a number of green peach aphids which were released at various times over 51 small tobacco plants (used as surrogates for potato plants) and the number of infected plants was recorded after each release given in Fig. 1b (see Turner 2000 for more detailed explanations). From these figures, we can observe that there are two groups in both examples. Therefore, these data sets should be modeled by using the mixture regression models.

In general, the parameters of a mixture regression model are estimated under normality assumption. Since the estimators based on normal distribution are sensitive to the outliers, robust mixture regression models have been proposed by Bai (2010) and Bai et al. (2012) to estimate the parameters of mixture regression using the M-estimation method. Wei (2012), Yao et al. (2014) proposed the mixture regression

F.Z. Dođru (✉)

Faculty of Arts and Sciences, Department of Statistics, Giresun University,
28100 Giresun, Turkey
e-mail: fatma.dogru@giresun.edu.tr

O. Arslan

Faculty of Science, Department of Statistics, Ankara University, 06100 Ankara, Turkey
e-mail: oarslan@ankara.edu.tr

© Springer India 2016

C. Agostinelli et al. (eds.), *Recent Advances in Robust Statistics: Theory and Applications*, DOI 10.1007/978-81-322-3643-6_4

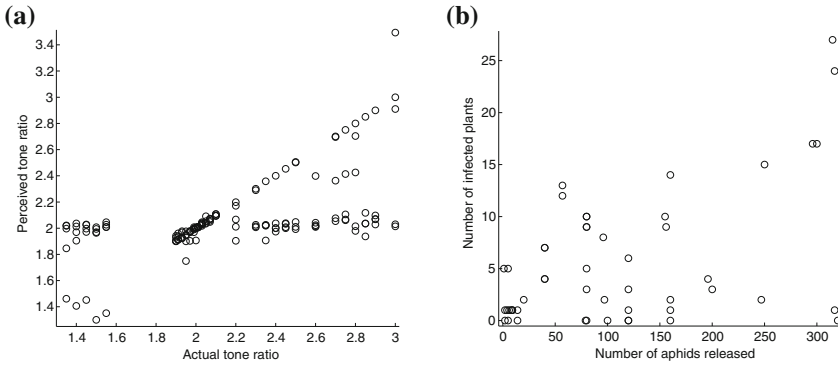


Fig. 1 **a** The scatter plot of the tone perception data. **b** The scatter plot of the aphids data

model based on the mixture of t distribution. Liu and Lin (2014) studied the mixture regression model based on the skew normal (Azzalini 1985, 1986) distribution. Dođru (2015), Dođru and Arslan (2016) propose a robust mixture regression procedure using the mixture of skew t distribution (Azzalini and Capitanio 2003) to model skewness and heavy-tailedness in the data with the groups.

Up to now mixture regression models are considered using the finite mixture of the same type of distributions such as mixture of normal or mixture of t distributions. The purpose of this work is to deal with the mixture regression model using the mixture of different type of distributions. This is due to the fact that the subclasses of data may not have same type of behavior. For example some of them may be heavy-tailed, skew or heavy-tailed skew. Using the same type of distributions to model such heterogeneous data may not produce efficient estimators. To accurately model this type of data we may need a mixture of distributions with different type of components. For example, it is clear that in the tone perception data (Fig. 1) two groups should have different type of error distributions. This is due to the fact that the observations around each line has differently scattered.

The rest of the paper is organized as follows. In Sect. 2, we give the mixture regression estimation based on mixture of different distributions. We consider two different mixtures. First, we consider the mixture of symmetric distributions. In particular, we take the mixture of normal and t distribution to estimate the regression parameters in a mixture regression model. Second model will be the mixture of skew distributions. In this context, we study the mixture of skew t and skew normal distribution to estimate the parameters of the mixture regression model. In both cases we give the EM algorithms in details. In Sect. 3, we provide a simulation study to demonstrate the performances of the proposed mixture regression estimators over the counterparts. In Sect. 4, we explore two real data examples to see the capability of the proposed estimators for real data sets. The paper is finalized with a conclusion section.

2 Mixture Regression Model Using the Mixture of Different Type of Distributions

In this section, we will carry out the mixture regression procedure based on the mixture of different distributions. We will only consider the mixture of two distributions, but mixture of more than two different types of distributions can be easily done using the methodology given in this paper.

2.1 Mixture Regression Estimation Based on the Mixture of Normal and t Distributions

A two-component mixture regression model can be defined as follows. Let Z be a latent class variable which is independent of explanatory variable \mathbf{x} . Then given $Z = i$, the response variable y and the p -dimensional explanatory variable \mathbf{x} have the following linear model

$$y_j = \mathbf{x}'_j \boldsymbol{\beta}_i + \epsilon_i, i = 1, 2, \quad (1)$$

where \mathbf{x}_j contains both the predictors and constant 1. Let $w_i = P(Z = i|\mathbf{x})$, $i = 1, 2$, be the mixing probability with $\sum_{i=1}^2 w_i = 1$. The conditional density of y given \mathbf{x} has the following form

$$f(y_j; \mathbf{x}_j, \boldsymbol{\Theta}) = w\phi(y_j; \mathbf{x}'_j \boldsymbol{\beta}_1, \sigma_1^2) + (1 - w)f_t(y_j; \mathbf{x}'_j \boldsymbol{\beta}_2, \sigma_2^2, \nu), \quad (2)$$

where Z is not observed. This implies that the distribution of the first error term is a normal distribution with 0 mean and the variance σ_1^2 and the distribution of the second error term is a t distribution with 0 mean, the scale parameter σ_2^2 and the degrees of freedom ν . Let $\boldsymbol{\Theta} = (w, \boldsymbol{\beta}_1, \sigma_1^2, \boldsymbol{\beta}_2, \sigma_2^2, \nu)'$ be the vector of all the unknown parameters in the model (2).

The ML estimator of the unknown parameter $\boldsymbol{\Theta}$ is obtained by maximizing the following log-likelihood function

$$\ell(\boldsymbol{\Theta}) = \sum_{j=1}^n \log(w\phi(y_j; \mathbf{x}'_j \boldsymbol{\beta}_1, \sigma_1^2) + (1 - w)f_t(y_j; \mathbf{x}'_j \boldsymbol{\beta}_2, \sigma_2^2, \nu)). \quad (3)$$

However, the maximizer of the log-likelihood function does not have an explicit solution. Therefore, the numerical methods should be used to obtain the estimators for the parameters of interest. Because of the mixture structure of the model the EM algorithm (Dempster et al. 1977) will be the convenient numerical method to obtain the estimators for the parameters.

Let z_j be the latent variable with

$$z_j = \begin{cases} 1, & \text{if } j^{\text{th}} \text{ observation is from 1}^{\text{th}} \text{ component} \\ 0, & \text{otherwise,} \end{cases} \quad (4)$$

for $j = 1, \dots, n$. The joint density function of two-component mixture regression model is

$$f(y_j, z_j) = (w\phi(y_j; \mathbf{x}'_j\boldsymbol{\beta}_1, \sigma_1^2))^{z_j} ((1-w)f_t(y_j; \mathbf{x}'_j\boldsymbol{\beta}_2, \sigma_2^2, \nu))^{(1-z_j)}. \quad (5)$$

To further simplify the steps of the EM algorithm, we will use the scale mixture representation of the t distribution. Let the random variable u has a gamma distribution with the parameters $(\nu/2, \nu/2)$. Then, the conditional distribution of ϵ_2 given u will be $N(0, \sigma^2/u)$. With the scale mixture representation of the t distribution this joint density can be further simplified as

$$f(y_j, u_j, z_j) = \left(w \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(y_j - \mathbf{x}'_j\boldsymbol{\beta}_1)^2}{2\sigma_1^2}} \right)^{z_j} \left((1-w) \frac{(\nu/2)^{\nu/2} u_j^{\frac{\nu}{2}-1} e^{-\frac{\nu}{2}u_j}}{\Gamma(\frac{\nu}{2}) \sqrt{2\pi\sigma_2^2/u_j}} e^{-\frac{(y_j - \mathbf{x}'_j\boldsymbol{\beta}_2)^2}{2\sigma_2^2/u_j}} \right)^{1-z_j}. \quad (6)$$

In this model, (\mathbf{z}, \mathbf{u}) are regarded as missing data and y is taken as observed data, where $\mathbf{y} = (y_1, \dots, y_n)$, $\mathbf{u} = (u_1, \dots, u_n)$ and $\mathbf{z} = (z_1, \dots, z_n)$. Equation (6) is the joint density function of the complete data $(\mathbf{y}, \mathbf{u}, \mathbf{z})$. Using this joint density function the complete data log-likelihood function for $\boldsymbol{\Theta}$ can be written as follows

$$\begin{aligned} \ell(\boldsymbol{\Theta}; \mathbf{y}, \mathbf{u}, \mathbf{z}) &= \sum_{j=1}^n z_j \left(\log w - \frac{\log 2\pi}{2} - \frac{\log \sigma_1^2}{2} - \frac{(y_j - \mathbf{x}'_j\boldsymbol{\beta}_1)^2}{2\sigma_1^2} \right) \\ &+ (1 - z_j) \left(\log(1-w) - \frac{\log 2\pi}{2} - \frac{\log \sigma_2^2}{2} + \frac{\log u_j}{2} - \frac{\nu}{2} u_j \right. \\ &\left. - \frac{(y_j - \mathbf{x}'_j\boldsymbol{\beta}_2)^2}{2\sigma_2^2/u_j} - \log \Gamma\left(\frac{\nu}{2}\right) + \frac{\nu}{2} \log\left(\frac{\nu}{2}\right) + \left(\frac{\nu}{2} - 1\right) \log u_j \right). \quad (7) \end{aligned}$$

Since u_j and z_j for $j = 1, \dots, n$, are taken as missing observations this log-likelihood function cannot be directly used to obtain the estimator for $\boldsymbol{\Theta}$. To overcome this latency problem we have to take the conditional expectation of the complete data log-likelihood function given y_j . This will be the E-step of the EM algorithm:

E-step:

$$\begin{aligned} E(\ell(\boldsymbol{\Theta}; \mathbf{y}, \mathbf{u}, \mathbf{z}) | y_j) &= \sum_{j=1}^n E(z_j | y_j) \left(\log w - \frac{\log 2\pi}{2} - \frac{\log \sigma_1^2}{2} - \frac{(y_j - \mathbf{x}'_j\boldsymbol{\beta}_1)^2}{2\sigma_1^2} \right) \\ &+ (1 - E(z_j | y_j)) \left(\log(1-w) - \frac{\log 2\pi}{2} - \frac{\log \sigma_2^2}{2} \right. \end{aligned}$$

$$\begin{aligned}
& + \frac{1}{2} E(\log u_j | y_j) - \frac{(y_j - \mathbf{x}'_j \boldsymbol{\beta}_2)^2 E(u_j | y_j)}{2\sigma_2^2} - \frac{\nu}{2} E(u_j | y_j) \\
& - \log \Gamma\left(\frac{\nu}{2}\right) + \frac{\nu}{2} \log\left(\frac{\nu}{2}\right) + \left(\frac{\nu}{2} - 1\right) E(\log u_j | y_j). \quad (8)
\end{aligned}$$

To obtain this conditional expectation of the complete data log-likelihood function we have to find $\hat{z}_j = E(z_j | y_j, \hat{\boldsymbol{\Theta}})$, $\hat{u}_{1j} = E(u_{1j} | y_j, \hat{\boldsymbol{\Theta}})$ and $\hat{u}_{2j} = E(u_{2j} | y_j, \hat{\boldsymbol{\Theta}})$ given in (36), (37) and (38), where $\hat{\boldsymbol{\Theta}}$ is the current estimate for $\boldsymbol{\Theta}$.

The M-step of the EM algorithm will be as follows.

M-step: Maximize the following function with respect to $\boldsymbol{\Theta}$

$$\begin{aligned}
Q(\boldsymbol{\Theta}; \hat{\boldsymbol{\Theta}}) = & \sum_{j=1}^n \hat{z}_j \left(\log w - \frac{\log 2\pi}{2} - \frac{\log \sigma_1^2}{2} - \frac{(y_j - \mathbf{x}'_j \boldsymbol{\beta}_1)^2}{2\sigma_1^2} \right) \\
& + (1 - \hat{z}_j) \left(\log(1 - w) - \frac{\log 2\pi}{2} - \frac{\log \sigma_2^2}{2} + \frac{\hat{u}_{2j}}{2} - \frac{\nu}{2} \hat{u}_{1j} \right. \\
& \left. - \frac{(y_j - \mathbf{x}'_j \boldsymbol{\beta}_2)^2 \hat{u}_{1j}}{2\sigma_2^2} - \log \Gamma\left(\frac{\nu}{2}\right) + \frac{\nu}{2} \log\left(\frac{\nu}{2}\right) + \left(\frac{\nu}{2} - 1\right) \hat{u}_{2j} \right). \quad (9)
\end{aligned}$$

Then, E- and M-steps of the EM algorithm will form the following iteratively reweighting algorithm.

Iteratively reweighting algorithm (EM algorithm)

1. Set initial parameter estimate $\boldsymbol{\Theta}^{(0)}$ and a stopping rule Δ .
2. Calculate the conditional expectations $\hat{z}_j^{(k)}$, $\hat{u}_{1j}^{(k)}$ and $\hat{u}_{2j}^{(k)}$ for the $(k+1)$ th for $k = 0, 1, 2, \dots$ iteration using the Eqs. (36), (37) and (38) given in appendix.
3. Insert the current values $\hat{z}_j^{(k)}$, $\hat{u}_{1j}^{(k)}$, $\hat{u}_{2j}^{(k)}$ and $\hat{\boldsymbol{\Theta}}^{(k)}$ in $Q(\boldsymbol{\Theta}; \hat{\boldsymbol{\Theta}})$ to form $Q(\boldsymbol{\Theta}; \hat{\boldsymbol{\Theta}}^{(k)})$ and maximize $Q(\boldsymbol{\Theta}; \hat{\boldsymbol{\Theta}}^{(k)})$ with respect to the parameters $(w, \boldsymbol{\beta}_1, \sigma_1^2, \boldsymbol{\beta}_2, \sigma_2^2, \nu)$ to get new estimates for the parameters. This maximization will give the following updating equations:

$$\hat{w}^{(k+1)} = \frac{\sum_{j=1}^n \hat{z}_j^{(k)}}{n}, \quad (10)$$

$$\hat{\boldsymbol{\beta}}_1^{(k+1)} = \left(\sum_{j=1}^n \hat{z}_j^{(k)} \mathbf{x}_j \mathbf{x}'_j \right)^{-1} \left(\sum_{j=1}^n \hat{z}_j^{(k)} \mathbf{x}_j y_j \right), \quad (11)$$

$$\hat{\sigma}_1^{2(k+1)} = \frac{\sum_{j=1}^n \hat{z}_j^{(k)} \left(y_j - \mathbf{x}'_j \hat{\boldsymbol{\beta}}_1^{(k)} \right)^2}{\sum_{j=1}^n \hat{z}_j^{(k)}}, \quad (12)$$

$$\hat{\boldsymbol{\beta}}_2^{(k+1)} = \left(\sum_{j=1}^n \left(1 - \hat{z}_j^{(k)} \right) \hat{u}_{1j}^{(k)} \mathbf{x}_j \mathbf{x}'_j \right)^{-1} \left(\sum_{j=1}^n \left(1 - \hat{z}_j^{(k)} \right) \hat{u}_{1j}^{(k)} \mathbf{x}_j y_j \right), \quad (13)$$

$$\hat{\sigma}_2^{2(k+1)} = \frac{\sum_{j=1}^n \left(1 - \hat{z}_j^{(k)} \right) \hat{u}_{1j}^{(k)} \left(y_j - \mathbf{x}'_j \hat{\boldsymbol{\beta}}_2^{(k)} \right)^2}{\sum_{j=1}^n \left(1 - \hat{z}_j^{(k)} \right)}. \quad (14)$$

4. To obtain the $\hat{\nu}^{(k+1)}$ solve the following equation

$$\sum_{j=1}^n \left(1 - \hat{z}_j^{(k)} \right) \left(DG \left(\frac{\nu}{2} \right) - \log \left(\frac{\nu}{2} \right) - 1 - \hat{u}_{2j}^{(k)} + \hat{u}_{1j}^{(k)} \right) = 0, \quad (15)$$

where $DG \left(\frac{\nu}{2} \right) = \frac{\Gamma' \left(\frac{\nu}{2} \right)}{\Gamma \left(\frac{\nu}{2} \right)}$ is the digamma function.

5. Repeat E and M steps until the convergence rule $\|\hat{\boldsymbol{\Theta}}^{(k+1)} - \hat{\boldsymbol{\Theta}}^{(k)}\| < \Delta$ is satisfied.

Note that the Eq. (15) can be solved by using some numerical methods.

2.2 Mixture Regression Estimation Based on the Mixture of Skew t (ST) and Skew Normal (SN) Distributions

Next we will consider the parameter estimation for the mixture regression model assuming that the error terms have mixture of skew t and skew normal distributions. By taking this mixture of two different skew distributions we attempt to model skewness, as well as, the heavy-tailedness in the sub groups of the data.

For two-component mixture regression model given in (1), the conditional density of y given \mathbf{x} is

$$f(y_j; \mathbf{x}_j, \boldsymbol{\Theta}) = w f_{ST} \left(y_j; \mathbf{x}'_j \boldsymbol{\beta}_1, \sigma_1^2, \lambda_1, \nu \right) + (1 - w) f_{SN} \left(y_j; \mathbf{x}'_j \boldsymbol{\beta}_2, \sigma_2^2, \lambda_2 \right), \quad (16)$$

where $f_{ST}(\cdot)$ is the density function of the skew t distribution proposed by Azzalini and Capitanio (2003) with the parameters $(\sigma_1^2, \lambda_1, \nu)$ and $f_{SN}(\cdot)$ is the density function of the skew normal distribution proposed by Azzalini (1985, 1986) with the parameters (σ_2^2, λ_2) . Note that the skew t is the distribution of ϵ_1 and the skew nor-

mal is the distribution of ϵ_2 . Let $\Theta = (w, \beta_1, \sigma_1^2, \lambda_1, \nu, \beta_2, \sigma_2^2, \lambda_2)'$ be the unknown parameter vector for this model. Notice that we have extra two skewness parameters to be estimated compare to the model given in Sect. 2.1. In this mixture regression model we note that different from the symmetric case $E(\epsilon) \neq 0$. Therefore, when we estimate the intercept we take into consideration $\widehat{E(\epsilon)}$.

To find the ML estimator of the unknown parameter Θ we should maximize the following log-likelihood function

$$\ell(\Theta) = \sum_{j=1}^n \log (w f_{ST}(y_j; \mathbf{x}'_j \beta_1, \sigma_1^2, \lambda_1, \nu) + (1 - w) f_{SN}(y_j; \mathbf{x}'_j \beta_2, \sigma_2^2, \lambda_2)). \quad (17)$$

Since the log-likelihood function does not have an explicit maximizer, the estimates for the unknown parameter vector Θ can be again obtained by using the EM algorithm.

Let z_j define as in Eq. (4), for $j = 1, \dots, n$. The joint density function of two-component mixture regression model is

$$f(y_j, z_j) = (w f_{ST}(y_j; \mathbf{x}'_j \beta_1, \sigma_1^2, \lambda_1, \nu))^{z_j} ((1 - w) f_{SN}(y_j; \mathbf{x}'_j \beta_2, \sigma_2^2, \lambda_2))^{1-z_j}. \quad (18)$$

To represent this joint density function in terms of the normal distribution, we will use the stochastic representation of the skew t and the skew normal distributions. By doing this we will simplify the steps of the EM algorithm. One can see the papers proposed by Azzalini and Capitanio (2003), Azzalini (1986, p. 201), Henze (1986, Theorem 1) to get more details for the stochastic representation of the skew t and the skew normal distributions. Using the scale mixture representation of the skew t distribution and the stochastic representation of the skew t and the skew normal distribution following conditional distributions can be given (Lin et al. 2007; Liu and Lin 2014). Let γ and τ be the latent variables. Then, we have

$$\begin{aligned} y_j | \gamma_j, \tau_j &\sim N \left(\mathbf{x}'_j \beta_1 + \alpha_1 \gamma_j, \frac{\kappa_1^2}{\tau_j} \right), \\ y_j | \tau_j &\sim TN \left(0, \frac{1}{\tau_j}; (0, \infty) \right), \tau_j \sim \text{Gamma} \left(\frac{\nu}{2}, \frac{\nu}{2} \right), \\ y_j | \gamma_j &\sim N \left(\mathbf{x}'_j \beta_2 + \alpha_2 \gamma_j, \kappa_2^2 \right), \gamma_j \sim TN (0, 1; (0, \infty)), \end{aligned}$$

where $\delta_{\lambda_1} = \lambda_1 / \sqrt{1 + \lambda_1^2}$, $\delta_{\lambda_2} = \lambda_2 / \sqrt{1 + \lambda_2^2}$, $\alpha_1 = \sigma_1 \delta_{\lambda_1}$, $\alpha_2 = \sigma_2 \delta_{\lambda_2}$, $\kappa_1^2 = \sigma_1^2 (1 - \delta_{\lambda_1}^2)$, $\kappa_2^2 = \sigma_2^2 (1 - \delta_{\lambda_2}^2)$ and $TN(\cdot)$ shows the truncated normal distribution.

Using the conditional distributions given above the joint density function given in (18) can be rewritten as

$$f(y_j, \gamma_j, \tau_j, z_j) = \left(w \frac{(v/2)^{v/2} \tau_j^{v/2}}{\pi \Gamma(\frac{v}{2}) \sqrt{\kappa_1^2}} e^{-\frac{v\tau_j}{2} - \frac{\tau_j(y_j - \mathbf{x}'_j \boldsymbol{\beta}_1 - \alpha_1 \gamma_j)^2}{2\kappa_1^2} - \frac{\tau_j \gamma_j^2}{2}} \right)^{z_j} \left(\frac{(1-w)}{\pi \sqrt{\kappa_2^2}} e^{-\frac{(y_j - \mathbf{x}'_j \boldsymbol{\beta}_2 - \alpha_2 \gamma_j)^2}{2\kappa_2^2} - \frac{\gamma_j^2}{2}} \right)^{1-z_j}. \quad (19)$$

Note that in this model $(\boldsymbol{\gamma}, \boldsymbol{\tau}, \mathbf{z})$ will be regarded as the missing and \mathbf{y} will be the observed data, where $\mathbf{y} = (y_1, \dots, y_n)$, $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_n)$, $\boldsymbol{\tau} = (\tau_1, \dots, \tau_n)$ and $\mathbf{z} = (z_1, \dots, z_n)$. Let $(\mathbf{y}, \boldsymbol{\gamma}, \boldsymbol{\tau}, \mathbf{z})$ be the complete data. Then, using the complete data joint density function given in (19), the complete data log-likelihood function can be obtained as follows

$$\begin{aligned} \ell_c(\boldsymbol{\Theta}; \mathbf{y}, \boldsymbol{\gamma}, \boldsymbol{\tau}, \mathbf{z}) &= \sum_{j=1}^n z_j \left(\log w - \log \pi - \frac{\log \kappa_1^2}{2} + \frac{v}{2} \log \left(\frac{v}{2} \right) - \log \left(\Gamma \left(\frac{v}{2} \right) \right) \right. \\ &\quad \left. + \frac{v}{2} \log \tau_j - \frac{v\tau_j}{2} - \frac{(y_j - \mathbf{x}'_j \boldsymbol{\beta}_1 - \alpha_1 \gamma_j)^2}{2\kappa_1^2 / \tau_j} - \frac{\tau_j \gamma_j^2}{2} \right) + (1 - z_j) \\ &\quad \left(\log(1 - w) - \log \pi - \frac{\log \kappa_2^2}{2} - \frac{(y_j - \mathbf{x}'_j \boldsymbol{\beta}_2 - \alpha_2 \gamma_j)^2}{2\kappa_2^2} - \frac{\gamma_j^2}{2} \right). \quad (20) \end{aligned}$$

Since we cannot be able to observe the missing data $(\boldsymbol{\gamma}, \boldsymbol{\tau}, \mathbf{z})$ this complete data log-likelihood function cannot be used to obtain the estimator for $\boldsymbol{\Theta}$. To overcome this problem we have to take the conditional expectation of the complete data log-likelihood function given the observed data \mathbf{y} . This will be the E-step of the EM algorithm

E-step

$$\begin{aligned} E(\ell_c(\boldsymbol{\Theta}; \mathbf{y}, \boldsymbol{\gamma}, \boldsymbol{\tau}, \mathbf{z}) | y_j) &= \sum_{j=1}^n E(z_j | y_j) \left(\log w - \frac{\log \kappa_1^2}{2} + \frac{v}{2} \log \left(\frac{v}{2} \right) - \log \Gamma \left(\frac{v}{2} \right) \right) \\ &\quad + \frac{v E(z_j \log \tau_j | y_j)}{2} - \frac{v E(Z_j \tau_j | y_j)}{2} - \frac{E(Z_j \tau_j | y_j) (y_j - \mathbf{x}'_j \boldsymbol{\beta}_1)^2}{2\kappa_1^2} \\ &\quad - \frac{\alpha_1^2 E(z_j \tau_j \gamma_j^2 | y_j)}{2\kappa_1^2} + \frac{\alpha_1 E(Z_j \tau_j \gamma_j | y_j) (y_j - \mathbf{x}'_j \boldsymbol{\beta}_1)}{\kappa_1^2} \\ &\quad + (1 - E(z_j | y_j)) \left(\log(1 - w) - \frac{\log \kappa_2^2}{2} - \frac{(y_j - \mathbf{x}'_j \boldsymbol{\beta}_2)^2}{2\kappa_2^2} \right. \\ &\quad \left. + \frac{\alpha_2 E(\gamma_j | y_j) (y_j - \mathbf{x}'_j \boldsymbol{\beta}_2)}{\kappa_2^2} - \frac{\alpha_2^2 E(\gamma_j^2 | y_j)}{2\kappa_2^2} \right). \quad (21) \end{aligned}$$

To obtain the conditional expectation of the complete data log-likelihood function we have to find $\hat{z}_j = E(z_j | y_j, \hat{\boldsymbol{\Theta}})$, $\hat{s}_{1j} = E(z_j \tau_j | y_j, \hat{\boldsymbol{\Theta}})$, $\hat{s}_{2j} = E(z_j \gamma_j \tau_j | y_j, \hat{\boldsymbol{\Theta}})$,

$\hat{s}_{3j} = E(z_j \gamma_j^2 \tau_j | y_j, \hat{\Theta})$, $\hat{s}_{4j} = E(z_j \log(\tau_j) | y_j, \hat{\Theta})$, $\hat{t}_{1j} = E(\gamma_j | y_j, \hat{\Theta})$ and $\hat{t}_{2j} = E(\gamma_j^2 | y_j, \hat{\Theta})$ given in (39)–(45).

M-step: For the M step of the EM algorithm, the expected complete data log-likelihood function will be maximized with respect to the parameter Θ

$$\begin{aligned} Q(\Theta; \hat{\Theta}) &= \sum_{j=1}^n \hat{z}_j \left(\log w - \frac{1}{2} \log(\kappa_1^2) + \frac{\nu}{2} \log\left(\frac{\nu}{2}\right) - \log\left(\Gamma\left(\frac{\nu}{2}\right)\right) \right) - \frac{\nu \hat{s}_{1j}}{2} \\ &+ \frac{\nu \hat{s}_{4j}}{2} - \frac{\hat{s}_{1j}(y_j - \mathbf{x}'_j \boldsymbol{\beta}_1)^2}{2\kappa_1^2} + \frac{\alpha_1 \hat{s}_{2j}(y_j - \mathbf{x}'_j \boldsymbol{\beta}_1)}{\kappa_1^2} - \frac{\alpha_1^2 \hat{s}_{3j}}{2\kappa_1^2} + (1 - \hat{z}_j) \\ &\left(\log(1 - w) - \frac{\log \kappa_2^2}{2} - \frac{(y_j - \mathbf{x}'_j \boldsymbol{\beta}_2)^2 - 2\alpha_2 \hat{t}_{1j}(y_j - \mathbf{x}'_j \boldsymbol{\beta}_2) + \alpha_2^2 \hat{t}_{2j}}{2\kappa_2^2} \right). \end{aligned} \quad (22)$$

Similar to the iteratively reweighting algorithm given in Sect. 2.1, we can give the following algorithm based on the steps of the EM algorithm for the two-component mixture regression model obtained from the skew t and skew normal distributions.

Iteratively reweighting algorithm (EM algorithm)

1. Set an initial parameter estimates $\Theta^{(0)}$ and stopping rule Δ .
2. Use $\hat{\Theta}^{(k)}$ to compute the conditional expectations $\hat{z}_j^{(k)}$, $\hat{s}_{1j}^{(k)}$, $\hat{s}_{2j}^{(k)}$, $\hat{s}_{3j}^{(k)}$, $\hat{s}_{4j}^{(k)}$, $\hat{t}_{1j}^{(k)}$, $\hat{t}_{2j}^{(k)}$ for $k = 0, 1, 2, \dots$ from the Eqs. (39)–(45) given in appendix.
3. Insert $\hat{z}_j^{(k)}$, $\hat{s}_{1j}^{(k)}$, $\hat{s}_{2j}^{(k)}$, $\hat{s}_{3j}^{(k)}$, $\hat{s}_{4j}^{(k)}$, $\hat{t}_{1j}^{(k)}$, $\hat{t}_{2j}^{(k)}$ and $\hat{\Theta}^{(k)}$ in $Q(\Theta; \hat{\Theta})$ to form $Q(\Theta; \hat{\Theta}^{(k)})$. Maximize the function $Q(\Theta; \hat{\Theta}^{(k)})$ given in (22) with respect to the parameters $(w, \boldsymbol{\beta}_1, \sigma_1^2, \lambda_1, \boldsymbol{\beta}_2, \sigma_2^2, \lambda_2)$ to get $(k + 1)$ iterated values

$$\hat{w}^{(k+1)} = \frac{\sum_{j=1}^n \hat{z}_j^{(k)}}{n}, \quad (23)$$

$$\hat{\boldsymbol{\beta}}_1^{(k+1)} = \left(\sum_{j=1}^n \hat{s}_{1j}^{(k)} \mathbf{x}_j \mathbf{x}'_j \right)^{-1} \left(\sum_{j=1}^n (y_j \hat{s}_{1j}^{(k)} - \hat{\delta}_{\lambda_1} \hat{s}_{2j}^{(k)}) \mathbf{x}_j \right), \quad (24)$$

$$\hat{\alpha}_1^{(k+1)} = \frac{\sum_{j=1}^n \hat{s}_{2j}^{(k)} (y_j - \mathbf{x}'_j \hat{\boldsymbol{\beta}}_1^{(k)})}{\sum_{j=1}^n \hat{s}_{3j}^{(k)}}, \quad (25)$$

$$\hat{\kappa}_1^{2(k+1)} = \frac{\sum_{j=1}^n \hat{s}_{1j}^{(k)} (y_j - \mathbf{x}'_j \hat{\boldsymbol{\beta}}_1^{(k)})^2 - 2\hat{\alpha}_1^{(k)} \hat{s}_{2j}^{(k)} (y_j - \mathbf{x}'_j \hat{\boldsymbol{\beta}}_1^{(k)}) + \hat{\alpha}_1^{2(k)} \hat{s}_{2j}^{(k)}}{\sum_{j=1}^n \hat{z}_j^{(k)}}, \quad (26)$$

$$\hat{\boldsymbol{\beta}}_2^{(k+1)} = \left(\sum_{j=1}^n (1 - \hat{z}_j^{(k)}) \mathbf{x}_j \mathbf{x}_j' \right)^{-1} \left(\sum_{j=1}^n (1 - \hat{z}_j^{(k)}) (y_j - \hat{\alpha}_2^{(k)} \hat{t}_{1j}^{(k)}) \mathbf{x}_j \right), \quad (27)$$

$$\hat{\alpha}_2^{(k+1)} = \frac{\sum_{j=1}^n (1 - \hat{z}_j^{(k)}) \hat{t}_{1j}^{(k)} (y_j - \mathbf{x}_j' \hat{\boldsymbol{\beta}}_2^{(k)})}{\sum_{j=1}^n (1 - \hat{z}_j^{(k)}) \hat{t}_{2j}^{(k)}}, \quad (28)$$

$$\begin{aligned} \hat{\kappa}_2^{2(k+1)} &= \frac{1}{\sum_{j=1}^n (1 - \hat{z}_j^{(k)})} \sum_{j=1}^n (1 - \hat{z}_j^{(k)}) \left((y_j - \mathbf{x}_j' \hat{\boldsymbol{\beta}}_2^{(k)})^2 \right. \\ &\quad \left. - 2\hat{\alpha}_2^{(k)} \hat{t}_{1j}^{(k)} (y_j - \mathbf{x}_j' \hat{\boldsymbol{\beta}}_2^{(k)}) + \hat{\alpha}_2^{2(k)} \hat{t}_{2j}^{(k)} \right). \end{aligned} \quad (29)$$

Then, we obtain the $\hat{\sigma}_1^{2(k+1)}$, $\hat{\lambda}_1^{(k+1)}$, $\hat{\sigma}_2^{2(k+1)}$ and $\hat{\lambda}_2^{(k+1)}$ parameter estimates

$$\hat{\sigma}_1^{2(k+1)} = \hat{\kappa}_1^{2(k+1)} + \hat{\alpha}_1^{2(k+1)}, \quad (30)$$

$$\hat{\lambda}_1^{(k+1)} = \hat{\delta}_{\lambda_1}^{(k+1)} \left(1 - \hat{\delta}_{\lambda_1}^{2(k+1)} \right)^{-1/2}, \quad (31)$$

$$\hat{\sigma}_2^{2(k+1)} = \hat{\kappa}_2^{2(k+1)} + \hat{\alpha}_2^{2(k+1)}, \quad (32)$$

$$\hat{\lambda}_2^{(k+1)} = \hat{\delta}_{\lambda_2}^{(k+1)} \left(1 - \hat{\delta}_{\lambda_2}^{2(k+1)} \right)^{-1/2}, \quad (33)$$

where $\hat{\delta}_{\lambda_1}^{(k+1)} = \hat{\alpha}_1^{(k+1)} / \hat{\sigma}_1^{(k+1)}$ and $\hat{\delta}_{\lambda_2}^{(k+1)} = \hat{\alpha}_2^{(k+1)} / \hat{\sigma}_2^{(k+1)}$.

4. Also $(k + 1)$ th value of λ_1 can be found by solving following equation

$$\begin{aligned} \delta_{\lambda_1} (1 - \delta_{\lambda_1}^2) \sum_{j=1}^n \hat{z}_j^{(k)} - \delta_{\lambda_1} \left(\sum_{j=1}^n \hat{s}_{1j}^{(k)} \frac{(y_j - \mathbf{x}_j' \hat{\boldsymbol{\beta}}_1^{(k)})^2}{\hat{\sigma}_1^{2(k)}} + \sum_{j=1}^n \hat{s}_{3j}^{(k)} \right) \\ + (1 + \delta_{\lambda_1}^2) \sum_{j=1}^n \hat{s}_{2j}^{(k)} \frac{(y_j - \mathbf{x}_j' \hat{\boldsymbol{\beta}}_1^{(k)})}{\hat{\sigma}_1^{(k)}} = 0. \end{aligned} \quad (34)$$

The $(k + 1)$ th values of ν can be calculated solving the following equation

$$\log \left(\frac{\nu}{2} \right) + 1 - DG \left(\frac{\nu}{2} \right) + \frac{\sum_{j=1}^n (\hat{s}_{4j}^{(k)} - \hat{s}_{1j}^{(k)})}{\sum_{j=1}^n \hat{z}_j^{(k)}} = 0. \quad (35)$$

5. Repeat E and M steps until the convergence rule $\|\hat{\Theta}^{(k+1)} - \hat{\Theta}^{(k)}\| < \Delta$ is satisfied.

Note that we can solve the Eqs. (34) and (35) using some numerical algorithms.

3 Simulation Study

In this section we will give a simulation study to assess and compare the performances of the mixture regression estimators proposed in this paper with the existing mixture regression estimators in the literature. We specifically compare the mixture regression estimators obtained from normal and t (MixregNt) distributions with the estimators obtained from normal (MixregN) and t (Mixregt) distributions for the two-component mixture regression models for the symmetric case. For the skew case, we compare the mixture regression estimators obtained from the skew t and the skew normal (MixregSTSN) distributions with the estimators obtained from skew normal (MixregSN) and skew t (MixregST) distributions for the two-component mixture regression models. The comparison will be done in terms of bias and mean square error (MSE) which are given the following formulas

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

where θ is the true parameter value, $\hat{\theta}_i$ is the i th simulated parameter estimate, $\bar{\theta} = \frac{1}{N} \sum_{i=1}^N \hat{\theta}_i$ and $N = 500$ is the replication number. For the sample sizes, we take $n = 200$ and $n = 400$. The simulation is conducted using *MATLAB R2013a*. The *MATLAB* codes can be obtained upon request.

Alternatively, the MSE for the $\hat{\Theta}$ which can be defined as $\|\hat{\Theta} - \Theta_0\|^2$, where Θ_0 is the true parameter, can be also used to illustrate the performance of the parameter vector as is suggested by one of the referee. However, to see the performance of each parameter we prefer computing the MSE for each parameter separately. We compare the both the MSE values and observe the similar behavior.

The data $\{(x_{1j}, x_{2j}, y_j), j = 1, \dots, n\}$ are generated from the following two-component mixture regression model (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 $P(Z = 1) = 0.25 = w_1$, $X_1 \sim N(0, 1)$ and $X_2 \sim N(0, 1)$. The values of the regression coefficients are $\beta_1 = (\beta_{10}, \beta_{11}, \beta_{12})' = (0, 1, 1)'$ and $\beta_2 = (\beta_{20}, \beta_{21}, \beta_{22})' = (0, -1, -1)'$, respectively.

We consider the following error distributions for the symmetric (i) and skew (ii) cases.

(i) Case I: $\epsilon_1, \epsilon_2 \sim N(0, 1)$, the standard normal distribution.

Case II: $\epsilon_1, \epsilon_2 \sim t_3(0, 1)$, the t distribution with 3 degrees of freedom.

Case III: $\epsilon_1 \sim N(0, 1)$ and $\epsilon_2 \sim t_3(0, 1)$.

Case IV: $\epsilon_1, \epsilon_2 \sim N(0, 1)$ and we also added %5 outliers at $X_1 = 20, X_2 = 20$ and $Y = 100$.

(ii) Case I: $\epsilon_1, \epsilon_2 \sim SN(0, 1, 0.5)$, the skew normal distribution.

Case II: $\epsilon_1, \epsilon_2 \sim ST(0, 1, 0.5, 3)$, the skew t distribution with 3 degrees of freedom.

Case III: $\epsilon_1 \sim ST(0, 1, 0.5, 3)$ and $\epsilon_2 \sim SN(0, 1, 0.5)$.

Case IV: $\epsilon_1, \epsilon_2 \sim N(0, 1)$ and we also added %5 outliers at $X_1 = 20, X_2 = 20$ and $Y = 100$.

The simulation results are summarized in Tables 1, 2, 3 and 4. Tables 1 and 2 show the simulation results for the estimators based on MixregNt with the error distributions given in case (i). For the Case I the best result is obtained from the estimators based on MixregN. For this case, the estimators based on Mixregt and the estimators based on MixregNt have similar behavior. For the error distribution given in Case II the best behavior is obtained, as expected, from Mixregt. In this case, the estimators based on MixregN are drastically affected. The proposed estimators (MixregNt) again have similar behavior with the estimators obtained from Mixregt which shows that it tolerates the heavy-tailedness. The estimators obtained from MixregNt perform the best for the error distribution given in Case III. In this case the estimator obtained from MixregN again has the worst performance. On the other hand, the performance of the estimators based on Mixregt is comparable with the estimators based on MixregNt. Finally, for the outlier case (Case IV) the behavior of the estimators based on MixregN and Mixregt is very similar. In both cases the worst performance is obtained for small groups. That is, they fail to find the regression line for the smaller group. In contrast, the estimators based on MixregNt can be able to accommodate the regression lines for both groups. This can be seen from the smaller bias and the MSE values. In summary, for all the cases considered in this part of the simulation the behavior of the proposed estimators is comparable with the counterparts.

In Tables 3 and 4 we summarize the simulation results obtained from the skew distributions with the error distributions given in case (ii). From this table we can observe that when the error distribution is the mixture of skew normal distribution, the estimators obtained from MixregSN behave better than the other cases. The same behavior can be noticed for the skew t distribution as well. When the error distribution is the mixture of the skew t and the skew normal the estimators obtained from MixregSTSN outperform the counterparts in terms of the MSE values. In this case, the estimators based on MixregSN have the worst performance. When we add

Table 1 MSE (bias) values of estimates for $n = 200$

	MixregN	Mixregt	MixregNt
<i>Case I: $\epsilon_1, \epsilon_2 \sim N(0, 1)$</i>			
$\beta_{10}:0$	0.0468 (-0.0039)	0.0547 (-0.0038)	0.0468 (-0.0036)
$\beta_{20}:0$	0.0099 (0.0028)	0.0112 (0.0056)	0.0111 (0.0048)
$\beta_{11}:1$	0.0457 (-0.0273)	0.0483 (-0.0364)	0.0529 (-0.0573)
$\beta_{21}:-1$	0.0147 (0.0055)	0.0101 (0.0090)	0.0096 (-0.0025)
$\beta_{12}:1$	0.0414 (-0.0015)	0.0463 (-0.0046)	0.0545 (-0.0353)
$\beta_{22}:-1$	0.0216 (0.0026)	0.0108 (0.0013)	0.0109 (-0.0090)
$w:0.25$	0.0029 (0.0071)	0.0022 (0.0048)	0.0036 (0.0335)
<i>Case II: $\epsilon_1, \epsilon_2 \sim t_3(0, 1)$</i>			
$\beta_{10}:0$	12.7323 (0.2016)	0.0846 (0.0245)	0.1378 (0.0265)
$\beta_{20}:0$	1.9712 (-0.0228)	0.0146 (-0.0048)	0.0140 (-0.0047)
$\beta_{11}:1$	10.6545 (0.3274)	0.1037 (-0.0479)	0.1417 (0.0332)
$\beta_{21}:-1$	1.5718 (-0.0867)	0.0128 (-0.0110)	0.0136 (0.0148)
$\beta_{12}:1$	8.2033 (0.0267)	0.0806 (0.0024)	0.1288 (0.0403)
$\beta_{22}:-1$	2.8037 (0.2205)	0.0143 (-0.0149)	0.0156 (0.0126)
$w:0.25$	0.0250 (-0.0347)	0.0030 (0.0062)	0.0045 (-0.0374)
<i>Case III: $\epsilon_1 \sim N(0, 1)$ and $\epsilon_2 \sim t_3(0, 1)$</i>			
$\beta_{10}:0$	6.5822 (0.0716)	0.0608 (-0.0015)	0.0564 (0.0039)
$\beta_{20}:0$	0.3372 (-0.0249)	0.0144 (0.0036)	0.0139 (0.0030)
$\beta_{11}:1$	5.0332 (0.0836)	0.0494 (-0.0387)	0.0459 (-0.0170)
$\beta_{21}:-1$	0.2761 (0.0073)	0.0138 (-0.0185)	0.0133 (-0.0105)
$\beta_{12}:1$	5.1167 (0.1571)	0.0673 (-0.0496)	0.0647 (-0.0287)
$\beta_{22}:-1$	0.5948 (0.0062)	0.0136 (-0.0213)	0.0134 (-0.0145)
$w:0.25$	0.0120 (0.002)	0.0039 (0.0352)	0.0029 (0.0174)
<i>Case IV: $\epsilon_1, \epsilon_2 \sim N(0, 1)$ (% 5 outliers)</i>			
$\beta_{10}:0$	2.3909 (0.1047)	1.4186 (0.0618)	0.0506 (0.0003)
$\beta_{20}:0$	0.0149 (0.0029)	0.0110 (0.0022)	0.0173 (0.0036)
$\beta_{11}:1$	3.1304 (1.4785)	2.7396 (1.4819)	0.1498 (-0.0521)
$\beta_{21}:-1$	0.0799 (0.2483)	0.0241 (0.1065)	0.2815 (0.1407)
$\beta_{12}:1$	3.2398 (1.5125)	2.8212 (1.5088)	0.1734 (-0.0633)
$\beta_{22}:-1$	0.0834 (0.2492)	0.0246 (0.1078)	0.1636 (0.1285)
$w:0.25$	0.0095 (-0.0943)	0.0062 (-0.0753)	0.0081 (-0.0208)

Note Value in parentheses indicates the bias

the leverage point (the error distribution given in Case IV) the behavior of all the estimators are similarly worse. However, the estimators obtained from MixregST and MixregSTSN give comparable results which have smaller bias and MSE than MixregSN.

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

	MixregN	Mixregt	MixregNt
<i>Case I: $\epsilon_1, \epsilon_2 \sim N(0, 1)$</i>			
$\beta_{10}:0$	0.0202 (-0.0021)	0.0260 (-0.0004)	0.0217 (-0.0003)
$\beta_{20}:0$	0.0041 (0.0018)	0.0052 (0.0036)	0.0050 (0.0031)
$\beta_{11}:1$	0.0160 (0.0033)	0.0206 (-0.0062)	0.0188 (-0.0199)
$\beta_{21}:-1$	0.0045 (-0.0035)	0.0053 (0.0018)	0.0051 (-0.0069)
$\beta_{12}:1$	0.0177 (-0.0019)	0.0254 (-0.0099)	0.0210 (-0.0213)
$\beta_{22}:-1$	0.0038 (-0.0038)	0.0049 (0.0035)	0.0046 (-0.0055)
w:0.25	0.0010 (0.0037)	0.0011 (0.0007)	0.0018 (0.0241)
<i>Case II: $\epsilon_1, \epsilon_2 \sim t_3(0, 1)$</i>			
$\beta_{10}:0$	13.4210 (-0.1318)	0.0376 (-0.0066)	0.0473 (-0.0154)
$\beta_{20}:0$	1.4586 (0.0306)	0.0068 (-0.0021)	0.0066 (-0.0026)
$\beta_{11}:1$	9.2787 (0.5967)	0.0335 (0.0019)	0.0449 (0.0632)
$\beta_{21}:-1$	1.8295 (0.0194)	0.0063 (0.0015)	0.0079 (0.0314)
$\beta_{12}:1$	11.8714 (0.4395)	0.0388 (-0.0008)	0.0533 (0.0722)
$\beta_{22}:-1$	0.7543 (0.01106)	0.0064 (0.0024)	0.0082 (0.0308)
w:0.25	0.0171 (-0.0596)	0.0014 (0.0040)	0.0037 (-0.0454)
<i>Case III: $\epsilon_1 \sim N(0, 1)$ and $\epsilon_2 \sim t_3(0, 1)$</i>			
$\beta_{10}:0$	7.7436 (0.0036)	0.0247 (0.0066)	0.0240 (0.0060)
$\beta_{20}:0$	0.3984 (0.0218)	0.0070 (0.0050)	0.0068 (0.0050)
$\beta_{11}:1$	5.8227 (0.1401)	0.0251 (-0.0372)	0.0206 (-0.0124)
$\beta_{21}:-1$	0.4005 (-0.0086)	0.0062 (-0.0126)	0.0060 (-0.0039)
$\beta_{12}:1$	6.6747 (0.2501)	0.0244 (-0.0365)	0.0213 (-0.0125)
$\beta_{22}:-1$	0.3341 (-0.0023)	0.0064 (-0.0143)	0.0063 (-0.0060)
w:0.25	0.0090 (-0.0070)	0.0021 (0.0289)	0.0015 (0.0077)
<i>Case IV: $\epsilon_1, \epsilon_2 \sim N(0, 1)$ (% 10 outliers)</i>			
$\beta_{10}:0$	1.4233 (0.1705)	1.0180 (0.2007)	0.0236 (0.0089)
$\beta_{20}:0$	0.0079 (0.0102)	0.0048 (0.0045)	0.0045 (0.0042)
$\beta_{11}:1$	2.9197 (1.5268)	2.5645 (1.4732)	0.0280 (-0.0003)
$\beta_{21}:-1$	0.0774 (0.2583)	0.0162 (0.1023)	0.0450 (0.0842)
$\beta_{12}:1$	2.7131 (1.4602)	2.6908 (1.5157)	0.0277 (0.0093)
$\beta_{22}:-1$	0.0759 (0.2536)	0.0150 (0.0961)	0.0264 (0.0733)
w:0.25	0.0099 (-0.0981)	0.0069 (-0.0817)	0.0036 (-0.0412)

Note Value in parentheses indicates the bias

Note that from the computational point of view, computing the estimators based on MixregSTSN is less intensive than the estimators obtained from MixregST. Therefore, even they show similar behavior MixregSTSN should be preferred.

Table 3 MSE (bias) values of estimates for $n = 200$

	MixregSN	MixregST	MixregSTSN
<i>Case I: $\epsilon_1, \epsilon_2 \sim SN(0, 1, 0.5)$</i>			
$\beta_{10}:0$	0.0396 (0.0145)	0.3796 (0.5558)	0.2090 (0.3916)
$\beta_{20}:0$	0.0083 (0.0124)	0.0388 (0.1602)	0.0095 (0.0228)
$\beta_{11}:1$	0.0322 (-0.0016)	0.0481 (-0.0041)	0.0445 (-0.0325)
$\beta_{21}:-1$	0.0080 (-0.0025)	0.0107 (0.0166)	0.0083 (-0.0082)
$\beta_{12}:1$	0.0366 (-0.0051)	0.0509 (-0.0150)	0.0491 (-0.0357)
$\beta_{22}:-1$	0.0080 (0.0019)	0.0104 (0.0212)	0.0083 (-0.0039)
$w:0.25$	0.0021 (0.0044)	0.0029 (-0.0087)	0.0028 (0.0162)
<i>Case II: $\epsilon_1, \epsilon_2 \sim ST(0, 1, 0.5, 3)$</i>			
$\beta_{10}:0$	5.5600 (0.3072)	0.8540 (0.7143)	0.4782 (0.2833)
$\beta_{20}:0$	1.7060 (-0.0335)	0.0239 (0.0805)	0.0421 (-0.0715)
$\beta_{11}:1$	5.1535 (0.2173)	0.1600 (0.0246)	0.2281 (-0.1303)
$\beta_{21}:-1$	0.9447 (0.0240)	0.0154 (0.0272)	0.0198 (-0.0147)
$\beta_{12}:1$	2.9528 (0.0176)	0.1445 (0.0178)	0.2302 (-0.1455)
$\beta_{22}:-1$	3.6893 (-0.0872)	0.0159 (0.0190)	0.0206 (-0.0084)
$w:0.25$	0.0175 (-0.0268)	0.0041 (-0.0128)	0.0084 (0.0444)
<i>Case III $\epsilon_1 \sim ST(0, 1, 0.5, 3)$ and $\epsilon_2 \sim SN(0, 1, 0.5)$</i>			
$\beta_{10}:0$	2.7217 (0.2228)	0.6258 (0.6722)	0.2932 (0.4207)
$\beta_{20}:0$	0.0555 (-0.0775)	0.0250 (0.1057)	0.0102 (-0.0123)
$\beta_{11}:1$	2.3168 (0.2017)	0.1033 (0.0327)	0.0921 (-0.0165)
$\beta_{21}:-1$	0.1975 (0.0372)	0.0108 (0.0305)	0.0088 (0.0059)
$\beta_{12}:1$	2.2086 (0.0959)	0.1158 (0.0228)	0.0958 (-0.0324)
$\beta_{22}:-1$	0.0244 (0.0598)	0.0111 (0.0314)	0.0090 (0.0079)
$w:0.25$	0.0079 (-0.0450)	0.0049 (-0.0369)	0.0031 (-0.0026)
<i>Case IV: $\epsilon_1, \epsilon_2 \sim N(0, 1)$ (% 5 outliers)</i>			
$\beta_{10}:0$	2.8415 (-0.5539)	5.7397 (2.1180)	1.7247 (0.5967)
$\beta_{20}:0$	0.2470 (-0.4804)	0.0437 (-0.1774)	0.1687 (-0.3970)
$\beta_{11}:1$	3.4886 (1.5398)	2.9568 (1.5342)	3.0075 (1.5224)
$\beta_{21}:-1$	0.0703 (0.2303)	0.0263 (0.1172)	0.0310 (0.1347)
$\beta_{12}:1$	3.2631 (1.4610)	2.4560 (1.3592)	2.7780 (1.4447)
$\beta_{22}:-1$	0.0784 (0.2423)	0.0278 (0.1211)	0.0347 (0.1435)
$w:0.25$	0.0093 (-0.0928)	0.0113 (-0.1031)	0.0049 (-0.0632)

Note Value in parentheses indicates the bias

Table 4 MSE (bias) values of estimates for $n = 400$

	MixregSN	MixregST	MixregSTSN
<i>Case I: $\epsilon_1, \epsilon_2 \sim SN(0, 1, 0.5)$</i>			
$\beta_{10}:0$	0.0172 (0.0055)	0.3194 (0.5421)	0.1741 (0.3920)
$\beta_{20}:0$	0.0035 (0.0040)	0.0258 (0.1422)	0.0038 (0.0147)
$\beta_{11}:1$	0.0145 (-0.0061)	0.0229 (-0.0045)	0.0197 (-0.0334)
$\beta_{21}:-1$	0.0035 (-0.0066)	0.0045 (0.0165)	0.0039 (-0.0122)
$\beta_{12}:1$	0.0145 (0.0012)	0.0220 (-0.0019)	0.0181 (-0.0235)
$\beta_{22}:-1$	0.0035 (-0.0004)	0.0048 (0.0194)	0.0036 (-0.0065)
$w:0.25$	0.0010 (0.0010)	0.0015 (-0.0147)	0.0013 (0.0126)
<i>Case II: $\epsilon_1, \epsilon_2 \sim ST(0, 1, 0.5, 3)$</i>			
$\beta_{10}:0$	9.1023 (0.5460)	0.6374 (0.6651)	0.2101 (0.3311)
$\beta_{20}:0$	0.7382 (-0.1361)	0.0105 (0.0402)	0.0141 (-0.0717)
$\beta_{11}:1$	7.2165 (0.2491)	0.0540 (0.0211)	0.0783 (-0.0925)
$\beta_{21}:-1$	0.7558 (0.0446)	0.0078 (0.0307)	0.0091 (-0.0188)
$\beta_{12}:1$	7.3711 (0.2215)	0.1714 (0.0671)	0.0785 (-0.0720)
$\beta_{22}:-1$	1.9910 (-0.0097)	0.0078 (0.0331)	0.0098 (-0.0157)
$w:0.25$	0.0144 (-0.0534)	0.0025 (-0.0217)	0.0044 (0.0441)
<i>Case III: $\epsilon_1 \sim ST(0, 1, 0.5, 3)$ and $\epsilon_2 \sim SN(0, 1, 0.5)$</i>			
$\beta_{10}:0$	1.7316 (0.1760)	0.4115 (0.6026)	0.2035 (0.4153)
$\beta_{20}:0$	0.0162 (-0.0883)	0.0142 (0.0890)	0.0042 (-0.0102)
$\beta_{11}:1$	1.9504 (0.1686)	0.0400 (0.0425)	0.0272 (0.0063)
$\beta_{21}:-1$	0.0140 (0.0589)	0.0057 (0.0318)	0.0038 (0.0055)
$\beta_{12}:1$	1.1483 (0.1871)	0.0426 (0.0430)	0.0303 (0.0114)
$\beta_{22}:-1$	0.0157 (0.0639)	0.0062 (0.0364)	0.0039 (0.0076)
$w:0.25$	0.0053 (-0.0483)	0.0032 (-0.0405)	0.0012 (-0.0037)
<i>Case IV: $\epsilon_1, \epsilon_2 \sim N(0, 1)$ (% 5 outliers)</i>			
$\beta_{10}:0$	1.4831 (-0.4157)	7.0063 (2.5017)	1.2509 (0.6932)
$\beta_{20}:0$	0.2537 (-0.4938)	0.0412 (-0.1874)	0.1635 (-0.3971)
$\beta_{11}:1$	2.9504 (1.4961)	2.4430 (1.4533)	2.5928 (1.4990)
$\beta_{21}:-1$	0.0792 (0.2631)	0.0245 (0.1325)	0.0311 (0.1548)
$\beta_{12}:1$	2.9498 (1.4958)	2.3507 (1.4210)	2.4870 (1.4631)
$\beta_{22}:-1$	0.0760 (0.2565)	0.0222 (0.1247)	0.0286 (0.1467)
$w:0.25$	0.0099 (-0.0978)	0.0139 (-0.1165)	0.0051 (-0.0681)

Note Value in parentheses indicates the bias

4 Real Data Examples

In this section, we will analyze two real data examples to show the performances of the proposed estimators over the estimators given in literature for the cases with and without outliers.

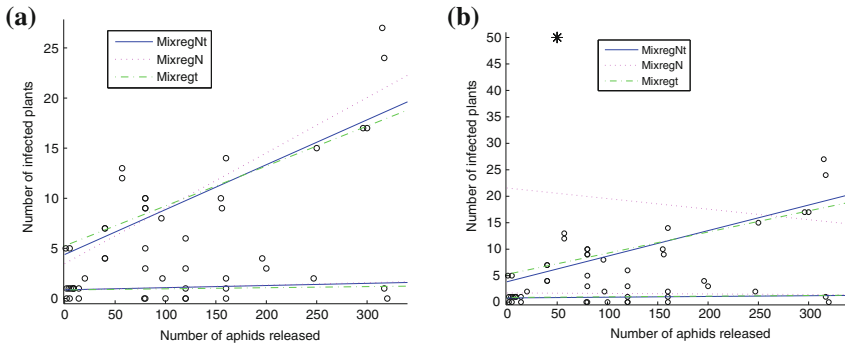


Fig. 2 **a** Fitted mixture regression lines without outlier. **b** Fitted mixture regression lines with outliers at (50, 50)

Table 5 ML estimates and some information criteria for fitting mixture regression models to the aphids data

	MixregN	Mixregt	MixregNt
$\hat{\beta}_{10}$	0.8586	0.8522	0.8648
$\hat{\beta}_{20}$	0.0024	0.0011	0.0022
$\hat{\beta}_{11}$	3.4745	5.2578	4.3813
$\hat{\beta}_{21}$	0.0553	0.0398	0.0448
$\hat{\sigma}_1$	1.1249	0.9127	1.0892
$\hat{\sigma}_2$	3.1153	1.0946	2.2073
\hat{w}_1	0.4984	0.5821	0.5128
$\ell(\hat{\Theta})$	-132.0651	-137.4716	-133.7615
AIC	278.1302	288.9432	281.5230
CAIC	298.6530	309.4660	302.0458
BIC	291.6530	302.4660	295.0458

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

Example 1 In this example, we use the aphids data introduced in Sect. 1 which can be accessed by using mixreg package (Turner 2000) in R. We first fit the lines using the estimates based on MixregN, Mixregt and MixregNt. These fitted lines along with the scatter plot of the data are shown in Fig. 2a. We can see that all methods successfully find the groups and give the correct fitted lines. Also, we summarize the ML estimates and the values of some information criteria in Table 5. Note that for the t distribution we assume that $\nu = 2$. We observe that MixregN has the best fit than the other mixture regression models in terms of the Akaike information criterion (AIC) (Akaike 1973), consistent AIC (CAIC) (Bozdogan 1993) and the Bayesian information criterion (BIC) (Schwarz 1978) values.

To see the performances of our estimators when there are outliers in the data, we add five pairs of high leverage outliers at point (50, 50). These points are shown in Fig. 2b by asterisk. Also, the fitted lines and the scatter plot of the data are displayed

Table 6 ML estimates and some information criteria for fitting mixture regression models to the aphids data with five outliers at (50, 50)

	MixregN	Mixregt	MixregNt
$\hat{\beta}_{10}$	1.7334	0.8694	0.7843
$\hat{\beta}_{20}$	-0.0010	0.0013	0.0014
$\hat{\beta}_{11}$	21.5569	5.2484	3.8322
$\hat{\beta}_{21}$	-0.0199	0.0400	0.0484
$\hat{\sigma}_1$	1.6859	1.0449	0.9362
$\hat{\sigma}_2$	16.3307	1.3244	4.0712
\hat{w}_1	0.5539	0.5547	0.41212
$\ell(\hat{\Theta})$	-194.0987	-192.1452	-182.4878
AIC	402.1974	398.2904	378.9756
CAIC	416.3749	412.4678	393.1531
BIC	423.3749	419.4678	400.1531

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

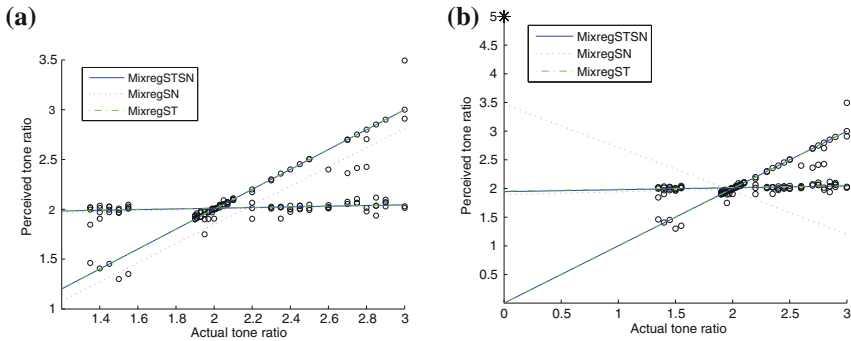


Fig. 3 a Fitted mixture regression lines without outlier. b Fitted mixture regression lines with outliers at (0, 5)

in Fig. 2b. We give the ML estimates in Table 6. We can see that the fitted lines obtained from MixregN are drastically affected by the outliers. On the other hand, the estimators obtained from Mixregt and MixregNt correctly identifies the groups and fit the regression lines. However, when we compare all methods MixregNt provides the best model in terms of the values of the information criteria.

Example 2. In this example, we use the tone perception data described in Sect. 1 which is given in fpc package (Hennig 2013) in R. This data analyzed by Bai et al. (2012) to model robust mixture regression model. Also, Yao et al. (2014) and Song et al. (2014) used the same data to test performances of the mixture regression estimators based on t and Laplace distributions. The results of these papers show that there should be two groups in the data. We fit mixture of skew normal, mixture of skew t and mixture of skew t and skew normal to check the performances of estimators based on these finite mixture models. We first consider this data without outlier and obtain the

Table 7 ML estimates and some information criteria for fitting mixture regression models to the tone perception data

	MixregSN	MixregST	MixregSTSN
$\hat{\beta}_{10}$	1.9171	1.9491	1.9430
$\hat{\beta}_{20}$	0.0424	0.0318	0.0339
$\hat{\beta}_{11}$	-0.0717	0.0054	0.0030
$\hat{\beta}_{21}$	0.9604	0.9982	0.9988
$\hat{\sigma}_1$	0.0463	0.0393	0.0419
$\hat{\sigma}_2$	0.1883	0.0033	0.0043
$\hat{\lambda}_1$	-0.0100	-0.1666	-0.1692
$\hat{\lambda}_2$	1.7534	0.4465	0.1297
\hat{w}_1	0.7006	0.6410	0.6534
$\ell(\hat{\Theta})$	140.5585	211.7766	215.4246
<i>AIC</i>	-263.1171	-405.5532	-412.8491
<i>CAIC</i>	-227.0213	-369.4574	-376.7534
<i>BIC</i>	-236.0213	-378.4574	-385.7534

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

Table 8 ML estimates and some information criteria for fitting mixture regression models to the tone perception data with ten outliers at (0, 5)

	MixregSN	MixregST	MixregSTSN
$\hat{\beta}_{10}$	1.8948	1.9553	1.9450
$\hat{\beta}_{20}$	0.0478	0.0313	0.0339
$\hat{\beta}_{11}$	3.4734	0.0057	0.0031
$\hat{\beta}_{21}$	-0.7579	0.9981	0.9987
$\hat{\sigma}_1$	0.0612	0.0542	0.0562
$\hat{\sigma}_2$	1.2593	0.0031	0.0043
$\hat{\lambda}_1$	-0.2667	-0.2030	-0.1971
$\hat{\lambda}_2$	1.6770	0.4493	0.1297
\hat{w}_1	0.7382	0.6759	0.6752
$\ell(\hat{\Theta})$	40.7933	109.3612	115.0275
<i>AIC</i>	-63.5867	-200.7225	-212.0549
<i>CAIC</i>	-26.9101	-164.0459	-175.3783
<i>BIC</i>	-35.9101	-173.0459	-184.3783

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

fitted lines from the mixture models mentioned above. The fitted lines along with the scatter plot are displayed in Fig. 3a. This figure shows that all the models give similar fits. Also, we give the ML estimates and some values of the information criteria in Table 7. The value of the degrees of freedom of the skew *t* distribution is taken as 2. We see that MixregSTSN gives the best fit than the other mixture regression models in terms of the AIC, CAIC and the BIC values.

To see the performances of the estimators when there are outliers in the data we added ten identical outliers at point (0, 5). The results for the data with outliers are shown in Fig. 3b. Note that the asterisk in this figure shows the location of outliers. It is clear from this figure that the outliers badly affect the estimators obtained from MixregSN. On the other hand, the estimators based on MixregST and MixregSTSN are not affected from the outliers. From the results of information criteria given in Table 8, MixregSTSN has the best fit to model to the tone perception data.

5 Conclusions

In this paper, we have proposed an alternative robust mixture regression model based on the mixture of different type of distributions. We have specifically considered two-component mixture regression based on mixture of t and normal distributions for the symmetric case, and the mixture of skew t and skew normal distributions for the skew case. We have given the EM algorithms for the mixture of different distributions. We have provided a simulation study and two real data examples. The simulation results and the real data examples have shown that the proposed method based on the mixture of different distributions is superior to or comparable with the method based on mixture of the same type of distributions such as mixture of (skew) normal and mixture of (skew) t distribution. If the groups in the data set have different tail behavior using the mixture of different type of distributions should be preferred. For example, in two group case if one of the groups has heavier tails but the other one is not then instead of using mixture of (skew) t distribution one can use mixture of (skew) t and (skew) normal and get the similar result. Using the mixture of t and normal will be computationally less intensive.

Acknowledgments The authors would like to thank referees and the editor for their constructive comments and suggestions that have considerably improved this work. The first author would like to thank the Higher Education Council of Turkey for providing financial support for Ph.D. study in Ankara University. The second author would like to thank the European Commission-JRC and the Indian Statistical Institute for providing financial support to attend the ICORS 2015 in Kolkata in India.

Appendix

To get the conditional expectation of the complete data log-likelihood function given in (8), the following conditional expectations should be calculated given y_j and the current parameter estimate $\hat{\Theta} = (\hat{\beta}_1, \hat{\sigma}_1^2, \hat{\beta}_2, \hat{\sigma}_2^2, \hat{\nu})$

$$\hat{z}_j = E(z_j | y_j, \hat{\Theta}) = \frac{\hat{w}\phi(y_j; \mathbf{x}'_j \hat{\beta}_1, \hat{\sigma}_1^2)}{\hat{w}\phi(y_j; \mathbf{x}'_j \hat{\beta}_1, \hat{\sigma}_1^2) + (1 - \hat{w})f_t(y_j; \mathbf{x}'_j \hat{\beta}_2, \hat{\sigma}_2, \hat{\nu})}, \quad (36)$$

$$\hat{u}_{1j} = E(u_j | y_j, \hat{\Theta}) = \frac{\hat{\nu} + 1}{\hat{\nu} + \left((y_j - \mathbf{x}'_j \hat{\beta}_2) / \hat{\sigma}_2 \right)^2}, \quad (37)$$

$$\hat{u}_{2j} = E(\log u_j | y_j, \hat{\Theta}) = DG \left(\frac{\hat{\nu} + 1}{2} \right) - \log \left(\frac{\hat{\nu}}{2} + \frac{(y_j - \mathbf{x}'_j \hat{\beta}_2)^2}{2\hat{\sigma}_2^2} \right). \quad (38)$$

These conditional expectations will be used in EM algorithm given in Sect. 2.1. Similarly, to obtain the conditional expectation of the complete data log-likelihood function given in (21) the following expectations should be computed given y_j and the current parameter estimate $\hat{\Theta} = (\hat{\beta}_1, \hat{\sigma}_1^2, \hat{\lambda}_1, \hat{\nu}, \hat{\beta}_2, \hat{\sigma}_2^2, \hat{\lambda}_2)$

$$\hat{z}_j = E(z_j | y_j, \hat{\Theta}) = \frac{\hat{w}_{fST}(y_j; \mathbf{x}'_j \hat{\beta}_1, \hat{\sigma}_1^2, \hat{\lambda}_1, \hat{\nu})}{\hat{w}_{fST}(y_j; \mathbf{x}'_j \hat{\beta}_1, \hat{\sigma}_1^2, \hat{\lambda}_1, \hat{\nu}) + (1 - \hat{w}) f_{SN}(y_j; \mathbf{x}'_j \hat{\beta}_1, \hat{\sigma}_1^2, \hat{\lambda}_1)}, \quad (39)$$

$$\hat{s}_{1j} = E(z_j \tau_j | y_j, \hat{\Theta}) = \hat{z}_j \left(\frac{\hat{\nu} + 1}{\hat{\eta}_{1j}^2 + \hat{\nu}} \right) \frac{T_{\hat{\nu}+3} \left(\hat{M}_j \sqrt{\frac{\hat{\nu}+3}{\hat{\nu}+1}} \right)}{T_{\hat{\nu}+1}(\hat{M}_j)}, \quad (40)$$

$$\hat{s}_{2j} = E(z_j \gamma_j \tau_j | y_j, \hat{\Theta}) = \frac{\hat{\delta}_{\lambda_1} (y_j - \mathbf{x}'_j \hat{\beta}_1) \hat{s}_{1j}}{\hat{\sigma}_1} + \frac{\hat{z}_j \sqrt{1 - \hat{\delta}_{\lambda_1}^2}}{\pi \hat{\sigma}_1 \hat{f}(y_j)} \left(\frac{\hat{\eta}_{1j}^2}{\hat{\nu}(1 - \hat{\delta}_{\lambda_1}^2)} + 1 \right)^{-(\frac{\hat{\nu}}{2}+1)}, \quad (41)$$

$$\begin{aligned} \hat{s}_{3j} = E(z_j \gamma_j^2 \tau_j | y_j, \hat{\Theta}) &= \hat{\delta}_{\lambda_1}^2 \left(\frac{y_j - \mathbf{x}'_j \hat{\beta}_1}{\hat{\sigma}_1} \right)^2 \hat{s}_{1j} + \hat{z}_j \left\{ (1 - \hat{\delta}_{\lambda_1}^2) \right. \\ &\left. + \frac{\hat{\delta}_{\lambda_1} (y_j - \mathbf{x}'_j \hat{\beta}_1) \sqrt{1 - \hat{\delta}_{\lambda_1}^2}}{\pi \hat{\sigma}_1^2 \hat{f}(y_j)} \left(\frac{\hat{\eta}_{1j}^2}{\hat{\nu}(1 - \hat{\delta}_{\lambda_1}^2)} + 1 \right)^{-(\frac{\hat{\nu}}{2}+1)} \right\}, \end{aligned} \quad (42)$$

$$\begin{aligned} \hat{s}_{4j} = E(z_j \log(\tau_j) | y_j, \hat{\Theta}) &= \hat{z}_j \left\{ DG \left(\frac{\hat{\nu} + 1}{2} \right) - \log \left(\frac{\hat{\eta}_{1j}^2 + \hat{\nu}}{2} \right) \right. \\ &+ \left(\frac{\hat{\nu} + 1}{\hat{\eta}_{1j}^2 + \hat{\nu}} \right) \left(\frac{T_{\hat{\nu}+3} \left(\hat{\lambda}_1 \hat{\eta}_{1j} \sqrt{\frac{\hat{\nu}+3}{\hat{\nu} + \hat{\eta}_{1j}^2}} \right)}{T_{\hat{\nu}+1} \left(\hat{\lambda}_1 \hat{\eta}_{1j} \sqrt{\frac{\hat{\nu}+1}{\hat{\nu} + \hat{\eta}_{1j}^2}} \right)} - 1 \right) \\ &+ \frac{\hat{\lambda}_1 \hat{\eta}_{1j} (\hat{\eta}_{1j}^2 - 1)}{\sqrt{(\hat{\nu} + 1)(\hat{\nu} + \hat{\eta}_{1j}^2)^3}} \frac{t_{\hat{\nu}+1} \left(\hat{\lambda}_1 \hat{\eta}_{1j} \sqrt{\frac{\hat{\nu}+1}{\hat{\nu} + \hat{\eta}_{1j}^2}} \right)}{T_{\hat{\nu}+1} \left(\hat{\lambda}_1 \hat{\eta}_{1j} \sqrt{\frac{\hat{\nu}+1}{\hat{\nu} + \hat{\eta}_{1j}^2}} \right)} \\ &\left. + \frac{1}{T_{\hat{\nu}+1} \left(\hat{\lambda}_1 \hat{\eta}_{1j} \sqrt{\frac{\hat{\nu}+1}{\hat{\nu} + \hat{\eta}_{1j}^2}} \right)} \int_{-\infty}^{\hat{M}_j} g_{\hat{\nu}}(x) t_{\hat{\nu}+1}(x) dx \right\}, \end{aligned} \quad (43)$$

$$\hat{t}_{1j} = E(\gamma_j | y_j, \hat{\Theta}) = \hat{\delta}_{\lambda_2} \hat{\eta}_{2j} + \sqrt{1 - \hat{\delta}_{\lambda_2}^2} \frac{\phi(\hat{\lambda}_2 \hat{\eta}_{2j})}{\Phi(\hat{\lambda}_2 \hat{\eta}_{2j})}, \quad (44)$$

$$\hat{t}_{2j} = E(\gamma_j^2 | y_j, \hat{\Theta}) = 1 - \hat{\delta}_{\lambda_2}^2 + \hat{\delta}_{\lambda_2} \hat{\eta}_{2j} \hat{t}_{1j}, \quad (45)$$

where

$$\begin{aligned} \hat{\eta}_{1j} &= \frac{(y_j - \mathbf{x}'_j \hat{\beta}_1)}{\hat{\sigma}_1}, \quad \hat{\delta}_{\lambda_1} = \frac{\hat{\lambda}_1}{\sqrt{1 + \hat{\lambda}_1^2}}, \\ \hat{\eta}_{2j} &= \frac{(y_j - \mathbf{x}'_j \hat{\beta}_2)}{\hat{\sigma}_2}, \quad \hat{\delta}_{\lambda_2} = \frac{\hat{\lambda}_2}{\sqrt{1 + \hat{\lambda}_2^2}}, \quad \hat{M}_j = \hat{\lambda}_1 \hat{\eta}_{1j} \sqrt{\frac{\hat{\nu}}{\hat{\nu} + \hat{\eta}_{1j}^2}}, \\ g_{\hat{\nu}}(x) &= DG\left(\frac{\hat{\nu} + 2}{2}\right) - DG\left(\frac{\hat{\nu} + 1}{2}\right) - \log\left(1 + \frac{x^2}{\hat{\nu} + 1}\right) + \frac{x^2(\hat{\nu} + 1) - \hat{\nu} - 1}{(\hat{\nu} + 1)(\hat{\nu} + 1 + x^2)}, \\ \hat{f}(y_j) &= \hat{w}_1 \frac{2}{\hat{\sigma}_1} t_{\hat{\nu}}(\hat{\eta}_{1j}) T_{\hat{\nu}+1}(\hat{M}_j) + (1 - \hat{w}_1) \frac{2}{\hat{\sigma}_2} \phi(\hat{\eta}_{2j}) \Phi(\hat{\lambda}_2 \hat{\eta}_{2j}). \end{aligned}$$

These conditional expectations will be used in EM algorithm given in Sect. 2.2.

References

- Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Caski F (eds) Proceeding of the second international symposium on information theory. Akademiai Kiado, Budapest, pp 267–281
- Azzalini A (1985) A class of distributions which includes the normal ones. *Scand J Stat* 12:171–178
- Azzalini A (1986) Further results on a class of distributions which includes the normal ones. *Statistica* 46:199–208
- Azzalini A, Capitanio A (2003) Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution. *J R Statist Soc B* 65:367–389
- Bai X (2010) Robust mixture of regression models. Master's thesis, Kansas State University
- Bai X, Yao W, Boyer JE (2012) Robust fitting of mixture regression models. *Comput Stat Data An* 56:2347–2359
- Bozdoğan H (1993) Choosing the number of component clusters in the mixture model using a new informational complexity criterion of the inverse-fisher information matrix. *Information and classification*. Springer, Berlin, pp 40–54
- Cohen AC (1984) Some effects of inharmonic partials on interval perception. *Music Percept* 1:323–349
- Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the E-M algorithm. *J Roy Stat Soc B Met* 39:1–38
- Dođru FZ (2015) Robust parameter estimation in mixture regression models. PhD thesis, Ankara University
- Dođru FZ, Arslan O (2016) Robust mixture regression based on the skew t distribution. *Revista Colombiana de Estadística*, accepted
- Hennig C (2013) fpc: Flexible procedures for clustering. R Package Version 2.1-5
- Henze N (1986) A probabilistic representation of the skew-normal distribution. *Scand J Stat* 13:271–275
- Lin TI, Lee JC, Hsieh WJ (2007) Robust mixture modeling using the skew t distribution. *Stat Comput* 17:81–92
- Liu M, Lin TI (2014) A skew-normal mixture regression model. *Educ Psychol Meas* 74(1):139–162

- Quandt RE (1972) A new approach to estimating switching regressions. *J Am Statist Assoc* 67:306–310
- Quandt RE, Ramsey JB (1978) Estimating mixtures of normal distributions and switching regressions. *J Am Stat Assoc* 73:730–752
- Schwarz G (1978) Estimating the dimension of a model. *Ann Stat* 6(2):461–464
- Song W, Yao W, Xing Y (2014) Robust mixture regression model fitting by laplace distribution. *Comput Stat Data An* 71:128–137
- Turner TR (2000) Estimating the propagation rate of a viral infection of potato plants via mixtures of regressions. *J Roy Statist Soc Ser C* 49:371–384
- 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. *Comput Stat Data An* 71:116–127