Robust mixture regression model fitting by Laplace distribution

Weixing Song, Weixin Yao, Yanru Xing

How to cite this manuscript

If you make reference to this version of the manuscript, use the following information:

Song, W., Yao, W., & Xing, Y. (2014). Robust mixture regression model fitting by Laplace distribution. Retrieved from http://krex.ksu.edu

Published Version Information

Citation: Song, W., Yao, W., & Xing, Y. (2014). Robust mixture regression model fitting by Laplace distribution. Computational Statistics and Data Analysis, 71, 128-137.

Digital Object Identifier (DOI): doi:10.1016/j.csda.2013.06.022

Publisher's Link: http://www.sciencedirect.com/science/article/pii/S0167947313002442

Robust Mixture Regression Model Fitting By Laplace Distribution

Weixing Song, Weixin Yao, Yanru Xing

Kansas State University

Abstract

A robust estimation procedure for mixture linear regression models is proposed by assuming that the error terms follow a Laplace distribution. The estimation procedure is implemented by an EM algorithm based on the fact that the Laplace distribution is a scale mixture of a normal distribution. Finite sample performance of the proposed algorithm is evaluated by numerical simulation studies. The superiority of the proposed method is illustrated by some comparison studies with other existing procedures in the literature. A real data example is also included to illustrate the application of the proposed method.

MSC: primary 62F35; secondary 62F10

Key words and phrases: Least Absolute Deviation; EM Algorithm; Mixture Regression Model; Normal Mixture; Laplace Distribution

1 Introduction

Least absolute deviation (LAD) regression has been widely used in practice if robust estimates are desired. The research on its computation and theoretical properties is abundant in the literature. A detailed survey on this topic can be found in Deilman (1984, 2005). In this paper, LAD will be applied to a class of mixture linear regression models to obtain robust estimates for the regression coefficients.

To be specific, let X be a p-dimensional vector of explanatory variables, Y be a scalar response variable. The relationship between Y and X is often investigated through a linear regression model. In the mixture linear regression setup, we assume that with probability π_i, $i = 1, 2, \ldots, g$, (X', Y) comes from one of the following $g \geq 2$ linear regression models

$$Y = X'\beta_i + \sigma_i\varepsilon_i, \quad i = 1, 2, \ldots, g, \tag{1.1}$$

where $\sum_{i=1}^{g} \pi_i = 1$, β_i's are unknown p-dimensional vectors of regression coefficients, σ_i's are unknown positive scalars. The random error ε_i's are assumed to be independent of X_i's. It is commonly assumed that the density functions of ε_i's are members in a location-scale family with mean 0 and variances 1. In this paper the design variable X is assumed to be random, but the proposed estimation procedure also works for the fixed design.

If $g = 1$, LAD estimate of β is the minimizer of the target function $Q(\beta) = \sum_{j=1}^{n} |Y_j - X_j'\beta|$, where $(X_j', Y_j)_{j=1}^{n}$ is a sample from model (1.1). However, if $g > 1$ the formulation of LAD target function is not straightforward since for a sample, we simply do not know which regression model an observation is from. Our formulation of the LAD target function is motivated by the fact that the maximum likelihood estimate of the regression coefficients given double exponentially distributed random error is indeed the LAD estimator for $g = 1$. Therefore, for $g \geq 2$ case, we assume that ε_i follows a double exponential distribution with location 0 and scale parameter $1/\sqrt{2}$, which makes the variance of ε_i being 1, $i = 1, 2, \ldots, g$. Then it is easily seen that for a sample $\mathbf{S} = \{(X_j', Y_j), j = 1, 2, \ldots, n\}$ from the model (1.1), the log-likelihood function of $\theta = (\beta_1, \sigma_1^2, \pi_1, \beta_2, \sigma_2^2, \pi_2, \ldots, \beta_g, \sigma_g^2, \pi_g)$ can be written as

$$L(\theta; \mathbf{S}) = \sum_{j=1}^{n} \log \sum_{i=1}^{g} \frac{\pi_i}{\sqrt{2}\sigma_i} \exp\left(-\frac{\sqrt{2}|Y_j - X_j'\beta_i|}{\sigma_i} \right) \tag{1.2}$$

and thus the maximum likelihood estimate of θ can be obtained by maximizing $L(\theta; \mathbf{S})$ with respect to θ. Usually no explicit solution can be obtained, and some numerical method will be needed.

If $g = 1$, many algorithms are developed in the literature to tackle the minimization problem $\hat{\beta} = \text{argmin}_{\beta} Q(\beta)$, such as the linear programming, least angle regression, the modified maximum likelihood method by Li and Arce (2004), among others. An often adopted but ad-hoc scheme for finding the solution β is to formally take the derivative of $Q(\beta)$ with respect to β, and set it equal to 0. Here σ^2 is treated as a nuisance parameter. By doing this, we obtain

$$\frac{\partial Q(\beta)}{\partial \beta} = -\sum_{j=1}^{n} X_j \ \text{sgn}(Y_j - X_j'\beta) = 0, \tag{1.3}$$

where $\text{sgn}(\cdot)$ is the sign function which takes $-1, 0, 1$ if the argument is negative, 0, and positive, respectively. Let $w_j = 1/|Y_j - X_j'\beta|$, and rewrite the equation (1.3) as $\sum_{j=1}^{n} w_j X_j (Y_j - X_j'\beta) = 0$. Thus by supplying an initial value β_0 for β, the updated value β can be found by the weighted least square solution

$$\beta_1 = \left(\sum_{j=1}^{n} w_j X_j X_j' \right)^{-1} \sum_{j=1}^{n} w_j X_j Y_j. \tag{1.4}$$

By iterating the above procedure, one can eventually find an approximate solution to $\text{argmin}_{\theta} Q(\theta)$.

A very interesting connection between the iterated weighted least square procedure stated above and an EM algorithm in conjunction with the Laplace distribution is found in Phillips (2002). For the sake of completeness, we briefly describe here Phillips (2002)' procedure.

Andrews and Mallows (1974) showed that a Laplace distribution in fact can be expressed as a mixture of a normal distribution and another distribution related to exponential distribution. To be specific, let Z and V be two random variables, V has a distribution with density function

$v^{-3}\exp(-(2v^2)^{-1})$, $v > 0$, and given $V = v$, the conditional distribution of Z is normal with mean 0 and variance $\sigma^2/(2v^2)$. Denote $f(z, v)$ the joint density function of Z and V, that is

$$f(z, v) = \frac{v}{\sqrt{\pi}\sigma} \exp\left(-\frac{v^2 z^2}{\sigma^2}\right) \frac{1}{v^3} \exp\left(-\frac{1}{2v^2}\right).$$

Then the marginal distribution of Z will be a Laplace distribution with density function $h_\varepsilon(z) = \exp(-\sqrt{2}|z|/\sigma)/(\sqrt{2}\sigma)$. Based on this finding, Phillips (2002) developed an EM algorithm to search for the minimizer of $Q(\beta)$.

Consider V as a latent variable. If V could be observed, then it is easy to see that the complete log-likelihood function of $\theta = (\beta, \sigma^2)$, based on the sample $\mathbf{P} = (X_j, Y_j, V_j)_{j=1}^n$, is

$$L(\theta; \mathbf{P}) = -\frac{1}{2}\log \pi\sigma^2 - \frac{1}{\sigma^2}\sum_{j=1}^n V_j^2 (Y_j - X_j'\beta)^2 - \sum_{j=1}^n \log V_j^2 - \frac{1}{2}\sum_{j=1}^n \frac{1}{V_j^2}.$$

Following the two steps in EM algorithm procedure, assume that $\theta^{(k)} = (\beta^{(k)}, \sigma^{2(k)})$ is the value in the kth iteration, then in the $(k+1)$th iteration, we have to first calculate the conditional expectation of the complete log likelihood function $L(\theta; \mathbf{P})$, given the observed data set $(Y_j, X_j)_{j=1}^n$ and $\theta = \theta^{(k)}$, which has the following form

$$\begin{aligned} E[L(\theta; \mathbf{P})|\mathbf{S}] = &-\frac{n}{2}\log \pi\sigma^2 - \frac{\sum_{j=1}^n E[V_j^2|\theta^{(k)}, (X_j, Y_j)_{j=1}^n](Y_j - X_j'\beta)^2}{\sigma^2} \\ &- \sum_{j=1}^n E[\log V_j^2|\theta^{(k)}, (X_j, Y_j)_{j=1}^n] - \frac{1}{2}\sum_{j=1}^n E\left[\frac{1}{V_j^2}\Big|\theta^{(k)}, (X_j, Y_j)_{j=1}^n\right]. \end{aligned}$$

In the second step, the conditional expectation will be maximized with respect to θ. Denote $w_j = E[V_j^2|\theta^{(k)}, (X_j, Y_j)_{j=1}^n]$, and notice that the third and fourth term on the right hand side do not involve the unknown regression parameters, so to maximize the above conditional expectation is equivalent to maximizing the following terms with respect to θ,

$$-\frac{n}{2}\log\sigma^2 - \frac{\sum_{j=1}^n w_j (Y_j - X_j'\beta)^2}{\sigma^2}.$$

Interestingly, Phillips (2002) showed $w_j = E[V_j^2|\theta^{(k)}, (X_j, Y_j)_{j=1}^n] = \sigma^{(k)}/(\sqrt{2}|Y_j - X_j'\beta^{(k)}|)$, this implies that the solution of $\beta^{(k+1)}$ indeed is the same as the one based on (1.4). It is also easy to see that $\sigma^{2(k+1)}$ can be estimated by $\sigma^{2(k+1)} = 2\sum_{j=1}^n w_j (Y_j - X_j'\beta^{(k+1)})^2/n$. In the next section, the above methodology will be extended to the mixture regression setting.

Yao and Wei (2012) proposed a robust estimation procedure for the mixture linear regression models based on t distribution by extending Peel and McLachlan (2000)'s work. The research conducted in this paper deals with the same questions as in Yao and Wei (2012), but the LAD technique, or the Laplace distribution, instead of the less commonly used t-distribution, is used for achieving robustness. In addition, the implementation of Yao and Wei (2012)'s procedure needs to specify the degrees of freedom in the t-distribution, our method does not need such tuning parameters.

The paper is organized as follows. The EM algorithm is developed in Section 2, together with some discussion on how to control the outliers in x-direction. Section 3 conducts some numerical simulations to evaluate the finite performance of the proposed method, comparison with some other existing methods will be also made. Finally, the proposed method will be applied on a real data example in Section 3.

2 EM Algorithm for Robust Mixture Regression

Assume that ε_i's follow a Laplace distribution with mean 0 and scale parameter $\sigma_i/\sqrt{2}$. For $i = 1, 2, \ldots, g$, $j = 1, 2, \ldots, n$, denote G_{ij} as latent Bernoulli variables such that

$$G_{ij} = \begin{cases} 1, & \text{if } j\text{th observation } (X_j, Y_j) \text{ is from } i\text{th component;} \\ 0, & \text{otherwise.} \end{cases}$$

Arguing as above, if the full data set $\mathbf{T} = \{(X_j, Y_j, G_{ij})\}_{i=1,2,\ldots,g;j=1,2,\ldots,n}$ are observable, then the full log likelihood function of $\theta = (\beta_1, \sigma_1^2, \pi_1, \beta_2, \sigma_2^2, \pi_2, \ldots, \beta_g, \sigma_g^2, \pi_g)$ can be written as

$$L(\theta; \mathbf{T}) = \sum_{j=1}^{n} \sum_{i=1}^{g} G_{ij} \log \frac{\pi_i}{\sqrt{2}\sigma_i} \exp\left(-\frac{\sqrt{2}|Y_j - X_j'\beta_i|}{\sigma_i}\right). \tag{2.1}$$

From Andrews and Mallows (1974), we know that a Laplace distributed random variable is a scale mixture of a normal random variable and another variable related to exponential distribution. Also see Section 1 for the detail. Denote V_j, coupled with (X_j, Y_j), as the latent scale variable, $j = 1, 2, \ldots$, then the full log-likelihood function of θ, based on $\mathbf{D} = \{X_j, Y_j, V_j, G_{ij}\}_{i=1,2,\ldots,g;j=1,2,\ldots,n}$, has the form

$$\begin{aligned} L(\theta; \mathbf{D}) &= \sum_{j=1}^{n} \sum_{i=1}^{g} G_{ij} \log \pi_i \frac{V_j}{\sqrt{\pi}\sigma_i} \exp\left(-\frac{V_j^2 (Y_j - X_j'\beta_i)^2}{\sigma_i^2}\right) \frac{1}{V_j^3} \exp\left(-\frac{1}{2V_j^2}\right) \\ &= \sum_{j=1}^{n} \sum_{i=1}^{g} G_{ij} \log \pi_i - \frac{1}{2} \sum_{j=1}^{n} \sum_{i=1}^{g} G_{ij} \log \pi\sigma_i^2 - \sum_{j=1}^{n} \sum_{i=1}^{g} \frac{G_{ij} V_j^2 (Y_j - X_j'\beta_i)^2}{\sigma_i^2} \\ &\quad - \sum_{j=1}^{n} \sum_{i=1}^{g} G_{ij} \log V_j^2 - \frac{1}{2} \sum_{j=1}^{n} \sum_{i=1}^{g} \frac{G_{ij}}{V_j^2}. \end{aligned} \tag{2.2}$$

Based on EM algorithm principle, in E-step, we have to calculate the condition expectation $E[L(\theta; \mathbf{D})|\mathbf{S}, \theta^{(0)}]$, where $\mathbf{S} = \{(X_j, Y_j)\}_{j=1}^n$ and $\theta^{(0)} = (\beta_1^{(0)}, \sigma_1^{2(0)}, \pi_1^{(0)}, \ldots, \beta_g^{(0)}, \sigma_g^{2(0)}, \pi_g^{(0)})$ is a proper initial value for θ. Since the last two terms in (2.2) do not involve the unknown regression parameters, we can simply drop them from the analysis. Thus, to find $E[L(\theta; \mathbf{D})|\mathbf{S}, \theta^{(0)}]$, we only have to calculate the following three terms

$$\tau_{ij} = E[G_{ij}|\mathbf{S}, \theta^{(0)}], \quad \delta_{ij} = E[V_j^2|\mathbf{S}, \theta^{(0)}, G_{ij} = 1].$$

One can show that

$$\tau_{ij} = \frac{\pi_i^{(0)} \sigma_i^{-1(0)} \exp(-|Y_j - X_j' \beta_i^{(0)}|/\sigma_i^{(0)})}{\sum_{m=1}^{g} \pi_m^{(0)} \sigma_m^{-1(0)} \exp(-|Y_j - X_j' \beta_m^{(0)}|/\sigma_m^{(0)})}, \quad \delta_{ij} = \frac{\sigma_i^{(0)}}{\sqrt{2}|Y_j - X_j' \beta_i^{(0)}|}. \tag{2.3}$$

The calculation for δ_{ij} follows the same thread as in Phillips (2002). In M-step, the following expression will be maximized with respect to π_i's, β_i's and σ_i^2's,

$$\sum_{j=1}^{n} \sum_{i=1}^{g} \tau_{ij} \log \pi_i - \frac{1}{2} \sum_{j=1}^{n} \sum_{i=1}^{g} \tau_{ij} \log \sigma_i^2 - \sum_{j=1}^{n} \sum_{i=1}^{g} \frac{\tau_{ij} \delta_{ij} (Y_j - X_j' \beta_i)^2}{\sigma_i^2}, \tag{2.4}$$

and the maximizer will be used for the next iteration.

In summary, we propose the following EM algorithm to maximize (1.2).

EM Algorithm:

(1). Choose an initial value for $\theta = (\beta_1, \sigma_1^2, \pi_1, \ldots, \beta_g, \sigma_g^2, \pi_g)$,

(2). E-Step: at the $(k+1)$-th iteration, calculate $\tau_{ij}^{(k+1)}$ and $\delta_{ij}^{(k+1)}$ from equation (2.3) with (0) replaced by (k).

(3). M-Step: at the $(k+1)$-th iteration, use the following formulas to calculate the maximizer of (2.4):

$$\pi_i^{(k+1)} = \frac{1}{n} \sum_{j=1}^{n} \tau_{ij}^{(k)},$$

$$\beta_i^{(k+1)} = \left(\sum_{j=1}^{n} \tau_{ij}^{(k+1)} \delta_{ij}^{(k+1)} X_j X_j' \right)^{-1} \left(\sum_{j=1}^{n} \tau_{ij}^{(k+1)} \delta_{ij}^{(k+1)} X_j Y_j \right),$$

$$\sigma_i^{2(k+1)} = \frac{2 \sum_{j=1}^{n} \tau_{ij}^{(k+1)} \delta_{ij}^{(k+1)} (Y_j - X_j' \beta_i^{(k+1)})^2}{\sum_{j=1}^{n} \tau_{ij}^{(k+1)}}.$$

(4). Repeat steps (2), (3) until the convergence is obtained.

If we further assume that all σ_i^2 are equal, then in the above EM algorithm, a common initial value for σ_i^2 should be used, and σ^2 can be updated in M-step by

$$\hat{\sigma}^{2(k+1)} = \frac{2 \sum_{j=1}^{n} \sum_{i=1}^{g} \tau_{ij}^{(k+1)} \delta_{ij}^{(k+1)} (Y_j - X_j' \beta_i^{(k+1)})^2}{n}.$$

The robustness of the above EM procedure is resulted from the adoption of LAD regression, it is also obvious from the formulae of the updated β_i's in each iteration. Note that the factor δ_{ij}^{k+1}

is reversely related to the term $|Y_j - X_j'\beta_i^{(k)}|$, meaning that larger residuals give smaller values of δ_{ij}^{k+1}, hence downweight the corresponding observations when calculating the estimates.

It is easy to see that when estimating β_i^{k+1}, the weight δ_{ij} can be simplified to $\delta_{ij} = 1/|Y_j - X_j'\beta_i^{(k)}|$. After estimating $\hat{\beta}_i^{k+1}$'s in the $k+1$-th iteration, similar to Philips (2002) in one population case, we can estimate σ_i^2 using formula

$$\sigma_i^{2(k+1)} = \frac{\sqrt{2}\sum_{j=1}^n \tau_{ij}^{(k+1)} \delta_{ij}^{(k+1)} |Y_j - X_j'\beta_i^{(k+1)}|}{\sum_{j=1}^n \tau_{ij}^{(k+1)}}.$$

Accordingly, when all σ_i^2's are assumed to be equal, then one can estimate the common variance by

$$\hat{\sigma}^{2(k+1)} = \frac{\sqrt{2}\sum_{j=1}^n \sum_{j=1}^g \tau_{ij}^{(k+1)} \delta_{ij}^{(k+1)} |Y_j - X_j'\beta_i^{(k+1)}|}{n}.$$

The EM algorithm proposed above for calculating $\hat{\beta}$ indeed is an iterated reweighted least square (IRLS) procedure, as the one proposed in Schlossmacher (1973) for one population case and the weights are given by $\tau_{ij}^{(k+1)}\delta_{ij}^{(k+1)}$ in the $k+1$-th iteration. Extra attention should be paid when programming the proposed EM algorithm. In the case of $g = 1$, Schlossmacher (1973) warned that if a perfect LAD fit occurs, i.e., $Y_j - X_j'\hat{\beta}_i = 0$ for some i, j, then the algorithm will eventually gives $Y_j - X_j'\beta_i^k \approx 0$ when iteration proceeds. As a result, δ_{ij}^{k+1} which is reciprocally related to $|Y_j - X_j'\beta_i^k|$ will be very large, and numerical instability would follow. Although Philips (2002) noticed that this problem rarely arises in the case of $g = 1$, this does occur often in our case, which is not out of expectation, simply because more than one regression models provide more chance for a perfect LAD fitting. But simply adopting Schlossmacher (1973)'s weight scheme by setting $\delta_{ij}^{k+1} = 0$ whenever $|Y_j - X_j'\beta_i^k| < e$ for a pre-assigned $e > 0$ is not quite reasonable. It makes much sense to allocate big weights for small residuals and small weights for big residuals. A cogent arguments on this issue is provided in Philips (2002). In our simulation study, we simply adopt a hard threshold rule to control the extremely small LAD residuals in each iteration steps. Under this rule, $\delta_{ij}^{(k+1)}$ will be assigned a value of 10^6 for any perfect LAD fit. We also tried other threshold values, such as $10^8, 10^{10}$ in the simulation, all these choices generate almost identical results. For the sake of brevity, we only report the simulation results by using 10^6 as the threshold value.

It is well known that in IRLS procedure, numerical instability could occur if the weights are very small. A common way to deal with this issue is to impose a hard threshold on τ_{ij}^{k+1} obtained in the $k+1$-th iteration. Namely, for a pre-specified value e say, if $\tau_{ij}^{k+1} > e$, then τ_{ij}^{k+1} itself will be used for the next iteration; otherwise, e will be used as the weight for the next iteration. Same technique is used in Yao and Wei (2012). In our simulation study, $e = 10e - 6$ is adopted.

Similar to the traditional M-estimate for linear regression and Yao and Wei (2012)'s mixture regression by t-distribution, the above EM algorithm based on Laplace distribution is robust against outliers along y-direction, but not in x-direction, which is also confirmed by our real data analysis conducted in Section 3. As a consequence, if there are any high leverage points in the data sets potentially being not from the model under discussion, which we intend to throw away from further analysis, or simply we do not want these observation exerting too much influence on the estimation, then the proposed EM algorithm might fail our expectation, and certain modification would be necessary. An obvious modification is first to identify these high leverage points, then just exclude them from further analysis. A commonly used method is to calculate the leverage value for each observation using formula $h_{jj} = n^{-1} + (n-1)^{-1}MD_j$, where $MD_j = (X_j - \bar{X})' S^{-1}(X_j - \bar{X})$, $\bar{X}$, S are the sample mean and sample covariance matrix ofX_j's, respectively. The j-th observation will be identified as a high leverage point if $h_{jj} > 2p/n$, where p is the dimension of X. To avoid the masking effect caused by using $\bar{X}$ and S in detecting the high leverage points, some robust estimation of the population mean and covariance matrix of X can be used instead of the sample mean and sample covariance. Yao and Wei (2012) adopted the minimum covariance determinant (MCD) estimators for the population mean and covariance matrix, which is implemented by the Fast MCD algorithm developed in Rousseeuw and Van Driessen (1999). Certainly, other robust estimates of the population mean and covariance matrix could be also used for this purpose, for example, the Stahel-Donoho (SD) estimator from Stahel (1981) and Donoho (1982). The j-th observation will be considered as a high leverage point if the resulting MD_j exceeds the threshold $\chi^2_{p-1,0.975}$. This threshold is proposed by Pison et al. (2002). In this paper, we propose to implement the proposed EM algorithm based on Laplace distribution after removing the observations with $MD_j > \chi^2_{p-1,0.975}$ using both MCD estimator and SD estimator to calculate MD_j.

3 Numerical Studies

To see the finite sample performance of the proposed robust estimation procedure, an extensive simulation study is conducted in this section. It is well known that the label switching issue is always an issue when evaluating different estimation methods in mixture models, and there are no widely accepted labeling standard. In our simulation, similar to Yao and Wei (2012), we simply choose the labels by minimizing the distance to the true parameter values. The effects of labeling schemes on comparison different estimation procedures deserves an independent research in the future.

3.1 Simulation Studies

In the simulation study, we choose equal variance for all components. The reason for doing this has two folds. Firstly, the log-likelihood function (2.1) is unbounded and goes to infinity if one observation exactly lies on one component line and the corresponding variance goes to 0, which makes the simulation very unstable. Secondly, choosing the same variances for all components can shorten the computation time, in particular, when the number of components is big.

To compare our method with some existing estimation procedures, we generate sample data $(X_{j1}, X_{j2}, Y_j)_{j=1}^n$ from the following two-component mixture regression models which are also used in Yao and Wei (2012):

$$Y = \begin{cases} 0 + X_1 + X_2 + \varepsilon_1, & \text{if } Z = 1, \\ 0 - X_1 - X_2 + \varepsilon_2, & \text{if } Z = 2, \end{cases}$$

where Z is the component indicator. That is, the data are generated from a two-component mixture linear regression models with $\beta_1 = (\beta_{10}, \beta_{11}, \beta_{12})' = (0, 1, 1)'$, and $\beta_2 = (\beta_{20}, \beta_{21}, \beta_{22})' = (0, -1, -1)'$. The predictors $X_1 \sim N(0, 1)$ and $X_2 \sim N(0, 1)$ are independent. The random error ε_1 and ε_2 are independent and has the same distribution as ε. To see the effects of different distributions of ε and the high leverage outliers in x-direction on various estimation methods, we consider the following six cases: (1), $\varepsilon \sim N(0, 1)$; (2), $\varepsilon \sim$Laplace distribution with mean 0 and variance 1; (3). $\varepsilon \sim t_1$, t-distribution with degrees of freedom 1 or the Cauchy distribution; (4). $\varepsilon \sim t_3$, t-distribution with degrees of freedom 3. (5). $\varepsilon \sim 0.95N(0, 1) + 0.05N(0, 25)$, a mixture of two normal distributions; (6), $\varepsilon \sim N(0, 1)$ with 5% high leverage outliers being $X_1 = X_2 = 20$ and $Y = 100$.

Case 1 is often used to evaluate the efficiency of different estimation methods compared to the traditional MLE when the error is exactly normally distributed and there are no outliers. For Case 2, the estimation methods proposed in the paper will provide the MLE of unknown parameters, which, as in the first case, would serve a reference line to evaluate the performance of other estimation procedures. Both Case 3 and 4 are heavy tailed distributions and often used in literature to mimic the outlier situations. Case 5 would produce 5% data likely to be low leverage outliers, and in Case 6, 5% observations are replicated serving as the high leverage outliers, which will be used to check the robustness of estimation procedures against the high leverage outliers.

Nine estimation methods will be compared in the simulation study: (1), maximum likelihood method based on normality assumption (MLE); (2), Trimmed likelihood estimator (TLE) proposed by Neykov et al. (2007); (3), the robust modified EM algorithm based on bisquare (Bisquare) proposed by Bai et al. (2012); (4), the robust mixture regression based on t-distribution (Mixregt) proposed by Yao and Wei (2012); (5), the trimmed mixture regression based on t-distribution (MixregtTrim), with MCD trimming method; (6), the trimmed mixture regression based on t-distribution (MixregtTrim), with SD trimming method; (7), the proposed robust EM mixture regression based on Laplace-distribution (MixregL); (8), the trimed mixture regression based on Laplace-distribution (MixregLTrim), with MCD trimming method, and (9), the trimed mixture regression based on Laplace-distribution (MixregLTrim), with SD trimming method.

From the simulation studies, we can see that if the true distribution of ε is normal, the MSEs of MLE procedure are slightly bigger than our proposed method for the first regression when the sample size is 100, but the superiority of MLE over all other methods becomes clear when the sample size gets bigger. But for other cases when the distribution of ε has a heavier tail, contaminated by some outliers, or there are high leverage outliers in the data set, then MLE fails to provide reasonable estimates.

The performance of TLE and Bisquare is satisfying when ε has a lighter tail, see the simulation results for all cases except Case III, where ε has a t-distribution with degrees of freedom 1. The overall performance of the Mixregt proposed by Yao and Wei (2012) is also satisfying when sample size gets bigger except for the Case VI when high leverage points present in the data set, but this disadvantage is remedied by the modified procedure Mixregt-MCD.

The simulation results clearly show that the proposed method in the paper outperforms or at least is comparable to any other methods. It is rather unexpected that our proposed method performs better than the Mixregt and Mixregt-MCD procedures even when ε has a t-distribution. The bigger MSEs in the later two procedures might be resulted from the extra step involved in the algorithm, the selection of v, which is the degrees of freedom of the t-distribution.

MCD estimator is used in Mixregt-MCD and MixregL-MCD to remove the high leverage outliers. In the simulation study, the SD estimator is also used to remove the high leverage outliers. The simulation results are similar to those from MIxregt-MCD and MixregL-MCD, hence omitted here for the sake of brevity.

3.2 Real Data Example

A typical real data set suitable for mixture regression modeling is the tone data collected in a tone perception experiment of Cohen (1984). In the experiment, a pure fundamental tone was played to a trained musician and electronically generated overtones were added, determined by a stretching ratio (stretchratio). A value of 2 for the stretch ratio corresponds to the harmonic pattern usually heard in traditional definite pitched instruments. Then the musician was asked to tune an adjustable tone to the octave above the fundamental tone, and a measurement called "tuned" gives the ratio of the adjusted tone to the fundamental. 150 pairs of (tuned, stretchratio) values are obtained with the same musician. The variable "strechratio" is treated as response variable and "tuned" as predictor. To see the impact of different types of outliers on various procedures, we first add 5 identical pairs, $(3, 4.5)$, to the original data set as outliers in y-direction. Here and after, the circles in the plots denote the original data points, and the star denotes the outliers. The right plots in all the figures below have the same y-scales as in the left plots.

	MLE	TLE	Bisquare	Mixregt	Mixregt-MCD	MixregL	MixregL-MCD
	Case I: $\varepsilon \sim N(0,1)$						
β_{10}	0.130(0.011)	0.139(0.033)	0.143(0.011)	0.124(0.021)	0.163(0.029)	0.093(0.079)	0.090(0.069)
β_{11}	0.160(-0.025)	0.212(-0.195)	0.157(-0.022)	0.130(-0.032)	0.175(-0.115)	0.094(-0.015)	0.113(-0.103)
β_{12}	0.135(-0.034)	0.248(-0.195)	0.171(-0.048)	0.123(-0.004)	0.247(-0.031)	0.088(0.008)	0.165(-0.039)
β_{20}	0.018(-0.003)	0.038(-0.004)	0.021(-0.001)	0.022(-0.012)	0.022(0.008)	0.028(-0.026)	0.027(-0.001)
β_{21}	0.021(-0.016)	0.030(0.011)	0.023(-0.017)	0.021(-0.006)	0.029(-0.011)	0.027(-0.001)	0.035(-0.021)
β_{22}	0.018(0.009)	0.024(0.034)	0.019(0.014)	0.021(-0.010)	0.030(-0.020)	0.026(-0.010)	0.042(-0.017)
π_1	0.005(0.003)	0.007(0.025)	0.005(0.005)	0.005(0.013)	0.007(0.016)	0.005(0.017)	0.007(0.022)
	Case II: $\varepsilon \sim$ Laplace(1)						
β_{10}	0.177(-0.006)	0.075(-0.007)	0.137(-0.016)	0.085(0.012)	0.123(-0.001)	0.058(0.022)	0.060(0.020)
β_{11}	0.145(-0.040)	0.097(-0.107)	0.142(-0.054)	0.084(-0.029)	0.150(-0.033)	0.050(-0.024)	0.080(-0.033)
β_{12}	0.152(0.009)	0.084(-0.077)	0.126(0.000)	0.080(-0.021)	0.150(-0.026)	0.055(-0.006)	0.063(-0.020)
β_{20}	0.016(-0.002)	0.013(0.004)	0.013(0.002)	0.011(-0.007)	0.016(-0.019)	0.010(-0.010)	0.015(-0.026)
β_{21}	0.021(-0.017)	0.013(0.007)	0.014(-0.019)	0.012(-0.008)	0.018(-0.030)	0.011(-0.004)	0.019(-0.020)
β_{22}	0.016(-0.006)	0.013(0.019)	0.013(-0.002)	0.012(-0.002)	0.020(0.009)	0.012(0.003)	0.026(0.018)
π_1	0.004(-0.004)	0.004(0.019)	0.004(0.016)	0.004(0.015)	0.005(0.012)	0.003(0.013)	0.005(0.009)
	Case III: $\varepsilon \sim t_1$						
β_{10}	242.992(-0.120)	3.200(-0.150)	1.683(-0.116)	1.708(-0.026)	0.945(-0.075)	0.163(0.061)	0.122(0.034)
β_{11}	174.666(-1.568)	1.886(-0.170)	1.571(-0.347)	1.990(-0.252)	1.621(-0.535)	0.521(-0.377)	0.561(-0.430)
β_{12}	148.108(-1.770)	1.797(-0.033)	1.642(-0.306)	2.410(-0.447)	1.538(-0.360)	0.548(-0.412)	0.418(-0.405)
β_{20}	244.822(0.172)	1.526(0.065)	0.910(0.024)	0.113(-0.020)	3.237(-0.173)	0.032(-0.024)	0.025(-0.038)
β_{21}	175.583(-1.080)	0.774(-0.129)	0.489(-0.088)	0.079(-0.041)	0.949(-0.102)	0.032(0.052)	0.047(0.081)
β_{22}	142.861(-0.454)	0.773(-0.065)	0.580(-0.116)	0.112(-0.049)	0.968(-0.028)	0.037(0.052)	0.048(0.054)
π_1	0.084(0.213)	0.039(0.060)	0.047(0.105)	0.023(0.093)	0.028(0.108)	0.022(0.070)	0.023(0.083)
	Case IV: $\varepsilon \sim t_3$						
β_{10}	1.568(-0.129)	0.238(0.007)	0.460(0.006)	0.529(0.031)	0.475(0.126)	0.131(0.065)	0.130(0.108)
β_{11}	0.997(-0.234)	0.264(-0.135)	0.341(-0.041)	0.361(0.010)	0.772(-0.109)	0.176(-0.021)	0.183(-0.041)
β_{12}	1.240(-0.024)	0.239(-0.096)	0.375(-0.058)	0.394(-0.010)	0.804(-0.040)	0.132(0.013)	0.186(-0.046)
β_{20}	0.723(-0.029)	0.038(-0.008)	0.063(0.013)	0.034(0.002)	0.077(-0.018)	0.032(-0.005)	0.030(-0.009)
β_{21}	0.188(0.028)	0.034(0.010)	0.085(-0.034)	0.037(-0.005)	0.062(-0.014)	0.042(0.004)	0.052(-0.018)
β_{22}	0.115(0.031)	0.026(0.010)	0.041(-0.013)	0.029(-0.018)	0.166(-0.027)	0.035(-0.015)	0.048(0.003)
π_1	0.028(0.025)	0.007(0.037)	0.009(0.030)	0.006(0.011)	0.014(0.035)	0.007(0.012)	0.007(0.021)
	Case V: $\varepsilon \sim 0.95N(0,1) + 0.05N(0,25)$						
β_{10}	2.243(-0.020)	0.124(0.046)	0.202(0.042)	0.152(0.015)	0.350(0.037)	0.097(0.034)	0.098(0.042)
β_{11}	1.366(0.054)	0.282(-0.209)	0.225(-0.037)	0.153(-0.029)	0.528(-0.106)	0.100(-0.008)	0.160(-0.056)
β_{12}	2.117(-0.113)	0.221(-0.190)	0.217(-0.056)	0.163(-0.050)	0.705(0.094)	0.099(-0.030)	0.175(0.023)
β_{20}	1.767(0.159)	0.030(0.013)	0.021(0.011)	0.026(0.020)	0.028(-0.004)	0.029(0.008)	0.035(-0.003)
β_{21}	1.277(-0.122)	0.034(0.001)	0.028(-0.023)	0.022(-0.009)	0.035(0.010)	0.026(-0.005)	0.040(0.008)
β_{22}	0.284(0.006)	0.027(0.011)	0.029(-0.009)	0.120(-0.036)	0.038(-0.017)	0.027(-0.006)	0.044(-0.020)
π_1	0.040(0.015)	0.010(0.034)	0.008(0.020)	0.007(0.015)	0.009(0.012)	0.005(0.006)	0.009(0.013)
	Case VI: $\varepsilon \sim N(0,1)$ with 5% high leverage outliers						
β_{10}	18.364(-2.878)	0.173(0.002)	0.152(0.015)	2.456(0.169)	0.175(-0.032)	0.036(0.080)	0.111(0.092)
β_{11}	5.876(1.422)	0.248(-0.209)	0.200(-0.068)	3.444(1.473)	0.219(-0.055)	0.056(-0.037)	0.133(-0.012)
β_{12}	6.520(1.641)	0.219(-0.168)	0.227(-0.091)	3.589(1.517)	0.262(0.006)	0.042(-0.014)	0.153(-0.046)
β_{20}	11.938(2.451)	0.036(-0.002)	0.023(-0.011)	0.023(0.002)	0.027(0.019)	0.015(-0.058)	0.032(0.011)
β_{21}	12.578(3.316)	0.028(0.000)	0.025(-0.014)	0.053(0.139)	0.027(0.010)	0.013(0.033)	0.042(0.000)
β_{22}	12.561(3.315)	0.022(0.025)	0.020(0.019)	0.053(0.136)	0.023(-0.017)	0.012(0.021)	0.046(0.004)
π_1	0.113(0.165)	0.007(0.017)	0.007(0.003)	0.007(-0.074)	0.006(0.005)	0.005(0.030)	0.006(0.011)

Table 1: MSE(Bias) of Point Estimates for $n = 100$

	MLE	TLE	Bisquare	Mixregt	Mixregt-MCD	MixregL	MixregL-MCD
	Case I: $\varepsilon \sim N(0,1)$						
β_{10}	0.043(-0.010)	0.073(-0.002)	0.044(-0.010)	0.047(-0.022)	0.052(-0.030)	0.053(0.008)	0.039(0.022)
β_{11}	0.041(-0.007)	0.129(-0.162)	0.044(-0.007)	0.035(0.008)	0.064(-0.005)	0.044(-0.023)	0.070(-0.032)
β_{12}	0.040(-0.020)	0.174(-0.199)	0.044(-0.018)	0.044(-0.007)	0.057(-0.030)	0.051(-0.015)	0.067(-0.037)
β_{20}	0.009(-0.006)	0.018(-0.016)	0.009(-0.007)	0.008(-0.010)	0.009(0.015)	0.013(-0.025)	0.013(0.005)
β_{21}	0.008(-0.006)	0.012(0.020)	0.008(-0.007)	0.009(-0.001)	0.013(-0.009)	0.014(0.013)	0.018(-0.002)
β_{22}	0.010(-0.014)	0.017(0.000)	0.012(-0.015)	0.008(-0.006)	0.013(-0.011)	0.012(0.004)	0.020(-0.001)
π_1	0.002(0.010)	0.004(0.019)	0.003(0.012)	0.002(0.007)	0.002(0.005)	0.002(0.008)	0.002(0.006)
	Case II: $\varepsilon \sim$ Laplace(1)						
β_{10}	0.046(0.006)	0.039(0.003)	0.033(0.015)	0.027(-0.002)	0.030(-0.005)	0.026(0.015)	0.022(0.013)
β_{11}	0.048(0.039)	0.033(-0.064)	0.034(0.017)	0.026(-0.021)	0.032(0.000)	0.020(-0.020)	0.024(-0.007)
β_{12}	0.043(0.009)	0.028(-0.058)	0.030(-0.004)	0.033(0.018)	0.036(-0.012)	0.020(0.002)	0.024(-0.011)
β_{20}	0.009(-0.007)	0.007(-0.007)	0.007(-0.010)	0.006(-0.004)	0.005(-0.001)	0.005(-0.012)	0.005(-0.006)
β_{21}	0.008(-0.020)	0.007(0.007)	0.007(-0.019)	0.005(-0.005)	0.007(-0.010)	0.004(-0.004)	0.007(-0.010)
β_{22}	0.009(-0.006)	0.006(0.007)	0.006(-0.009)	0.005(0.006)	0.009(-0.009)	0.005(0.007)	0.009(-0.004)
π_1	0.002(0.004)	0.002(0.019)	0.002(0.023)	0.002(0.006)	0.002(0.005)	0.002(0.005)	0.002(0.003)
	Case III: $\varepsilon \sim t_1$						
β_{10}	286.806(1.711)	1.026(-0.123)	1.256(-0.042)	0.326(0.029)	0.411(0.049)	0.067(0.049)	0.048(0.074)
β_{11}	36.053(-0.902)	0.906(0.103)	0.981(-0.222)	0.612(-0.362)	0.808(-0.471)	0.268(-0.362)	0.406(-0.471)
β_{12}	85.816(-0.726)	0.904(-0.024)	1.031(-0.222)	0.807(-0.392)	0.810(-0.485)	0.289(-0.349)	0.434(-0.506)
β_{20}	283.651(1.486)	0.774(0.128)	0.587(-0.018)	0.036(-0.006)	0.060(-0.065)	0.013(-0.052)	0.013(-0.042)
β_{21}	30.042(1.056)	0.201(0.052)	0.273(0.012)	0.030(0.004)	0.063(-0.009)	0.017(0.078)	0.028(0.109)
β_{22}	49.441(0.368)	0.253(-0.032)	0.281(0.019)	0.039(-0.011)	0.047(-0.004)	0.019(0.074)	0.028(0.101)
π_1	0.067(0.240)	0.020(0.031)	0.033(0.094)	0.013(0.057)	0.025(0.087)	0.025(0.076)	0.033(0.106)
	Case IV: $\varepsilon \sim t_3$						
β_{10}	0.600(-0.069)	0.080(0.020)	0.121(-0.030)	0.084(-0.015)	0.155(0.036)	0.064(0.017)	0.078(0.022)
β_{11}	0.486(-0.167)	0.082(-0.082)	0.096(0.019)	0.101(0.021)	0.181(-0.041)	0.072(0.005)	0.112(-0.049)
β_{12}	0.778(-0.050)	0.082(-0.095)	0.078(-0.005)	0.098(-0.040)	0.194(-0.020)	0.071(-0.034)	0.103(-0.017)
β_{20}	3.107(-0.153)	0.019(-0.008)	0.016(0.002)	0.015(-0.006)	0.015(-0.007)	0.015(-0.015)	0.017(-0.015)
β_{21}	0.459(-0.026)	0.017(0.014)	0.016(-0.021)	0.012(-0.004)	0.020(-0.014)	0.014(0.006)	0.021(-0.013)
β_{22}	0.227(0.046)	0.014(-0.009)	0.016(-0.043)	0.016(-0.003)	0.018(-0.020)	0.017(0.002)	0.019(-0.017)
π_1	0.029(0.018)	0.004(0.035)	0.004(0.031)	0.003(0.007)	0.004(0.012)	0.003(0.004)	0.004(0.010)
	Case V: $\varepsilon \sim 0.95N(0,1)+0.05N(0,25)$						
β_{10}	1.077(-0.002)	0.072(0.029)	0.051(0.006)	0.074(-0.028)	0.100(0.024)	0.056(0.004)	0.064(0.023)
β_{11}	0.834(0.031)	0.095(-0.142)	0.046(0.014)	0.059(-0.021)	0.113(0.000)	0.054(-0.029)	0.075(-0.024)
β_{12}	0.675(-0.121)	0.097(-0.125)	0.055(0.006)	0.062(0.004)	0.096(0.007)	0.060(0.008)	0.068(0.013)
β_{20}	0.348(0.062)	0.013(0.007)	0.010(0.012)	0.012(-0.007)	0.013(-0.019)	0.014(-0.020)	0.017(-0.026)
β_{21}	0.042(0.072)	0.011(0.014)	0.009(0.001)	0.012(0.002)	0.014(0.010)	0.016(0.007)	0.017(0.009)
β_{22}	0.036(0.067)	0.012(0.016)	0.010(-0.005)	0.012(0.005)	0.015(-0.023)	0.014(0.008)	0.018(-0.019)
π_1	0.016(-0.023)	0.003(0.020)	0.003(0.014)	0.002(0.002)	0.003(0.003)	0.002(-0.001)	0.003(0.000)
	Case VI: $\varepsilon \sim N(0,1)$ with 5% high leverage outliers						
β_{10}	12.459(-2.191)	0.061(0.006)	0.044(-0.008)	1.773(-0.009)	0.054(-0.015)	0.021(0.057)	0.050(-0.004)
β_{11}	4.875(1.543)	0.078(-0.093)	0.060(-0.021)	3.168(1.552)	0.065(-0.031)	0.025(-0.041)	0.064(-0.043)
β_{12}	4.678(1.468)	0.087(-0.132)	0.056(-0.033)	2.853(1.447)	0.067(0.013)	0.031(-0.037)	0.067(-0.028)
β_{20}	15.169(2.671)	0.012(-0.013)	0.010(-0.007)	0.010(-0.009)	0.010(0.000)	0.009(-0.063)	0.016(-0.023)
β_{21}	12.212(3.243)	0.010(0.015)	0.008(0.007)	0.031(0.134)	0.013(0.006)	0.009(0.037)	0.015(-0.006)
β_{22}	13.057(3.364)	0.016(-0.004)	0.012(-0.002)	0.027(0.133)	0.014(-0.022)	0.007(0.027)	0.017(-0.006)
π_1	0.147(0.221)	0.003(0.013)	0.003(0.004)	0.008(-0.085)	0.002(0.007)	0.005(0.027)	0.003(0.003)

Table 2: MSE(Bias) of Point Estimates for $n = 200$

	MLE	TLE	Bisquare	Mixregt	Mixregt-MCD	MixregL	MixregL-MCD
Case I: $\varepsilon \sim N(0,1)$							
β_{10}	0.018(-0.006)	0.041(0.012)	0.020(-0.005)	0.019(0.004)	0.027(0.008)	0.025(0.018)	0.031(0.014)
β_{11}	0.020(0.002)	0.108(-0.178)	0.021(-0.001)	0.018(-0.014)	0.028(-0.014)	0.024(-0.028)	0.034(-0.029)
β_{12}	0.018(-0.006)	0.096(-0.171)	0.020(0.000)	0.016(0.008)	0.031(0.012)	0.029(-0.001)	0.042(-0.012)
β_{20}	0.004(0.003)	0.009(0.002)	0.004(0.002)	0.005(-0.006)	0.005(0.012)	0.008(-0.010)	0.008(0.014)
β_{21}	0.004(0.004)	0.007(0.020)	0.004(0.002)	0.004(-0.009)	0.006(-0.002)	0.006(-0.005)	0.009(0.002)
β_{22}	0.004(-0.005)	0.006(0.013)	0.004(-0.006)	0.005(-0.004)	0.006(0.000)	0.007(0.003)	0.008(0.009)
π_1	0.001(0.000)	0.002(-0.001)	0.001(0.002)	0.001(0.001)	0.002(0.005)	0.001(0.000)	0.002(0.006)
Case II: $\varepsilon \sim$ Laplace(1)							
β_{10}	0.022(-0.005)	0.012(0.012)	0.015(-0.003)	0.012(-0.004)	0.013(0.003)	0.010(0.007)	0.012(0.010)
β_{11}	0.014(0.008)	0.013(-0.041)	0.010(0.005)	0.012(0.003)	0.018(-0.013)	0.011(0.005)	0.017(-0.007)
β_{12}	0.016(-0.006)	0.017(-0.050)	0.012(-0.004)	0.011(-0.013)	0.016(0.000)	0.008(-0.007)	0.014(0.005)
β_{20}	0.004(-0.003)	0.003(-0.003)	0.003(-0.003)	0.002(0.001)	0.002(0.000)	0.002(0.002)	0.002(-0.001)
β_{21}	0.004(-0.013)	0.003(0.005)	0.003(-0.015)	0.003(-0.009)	0.004(-0.003)	0.003(-0.004)	0.004(-0.001)
β_{22}	0.004(-0.011)	0.004(0.012)	0.003(-0.009)	0.003(-0.003)	0.004(-0.006)	0.002(-0.003)	0.003(-0.003)
π_1	0.001(0.002)	0.001(0.016)	0.001(0.022)	0.001(0.004)	0.001(0.006)	0.001(0.001)	0.001(0.004)
Case III: $\varepsilon \sim t_1$							
β_{10}	313.757(-0.917)	0.735(-0.040)	0.631(-0.083)	0.147(0.019)	0.154(0.002)	0.016(0.073)	0.017(0.076)
β_{11}	278.219(-3.135)	0.398(0.097)	0.607(-0.187)	0.458(-0.191)	0.485(-0.257)	0.194(-0.352)	0.322(-0.454)
β_{12}	455.172(-1.369)	0.399(0.059)	0.716(-0.146)	0.351(-0.177)	0.484(-0.200)	0.197(-0.361)	0.351(-0.462)
β_{20}	313.757(-0.917)	0.021(-0.001)	0.514(-0.052)	0.023(-0.008)	0.021(-0.002)	0.008(-0.061)	0.008(-0.067)
β_{21}	269.680(-1.135)	0.032(0.003)	0.047(0.034)	0.014(0.006)	0.022(-0.003)	0.011(0.092)	0.015(0.099)
β_{22}	453.695(0.630)	0.093(-0.009)	0.083(0.014)	0.017(0.009)	0.020(-0.002)	0.012(0.094)	0.016(0.102)
π_1	0.061(0.247)	0.008(0.003)	0.016(0.062)	0.009(0.031)	0.008(0.037)	0.037(0.160)	0.038(0.161)
Case IV: $\varepsilon \sim t_3$							
β_{10}	0.301(0.020)	0.037(-0.008)	0.038(-0.010)	0.039(-0.014)	0.059(-0.016)	0.033(0.002)	0.044(0.005)
β_{11}	0.210(-0.046)	0.039(-0.070)	0.044(0.049)	0.034(-0.013)	0.071(-0.008)	0.028(-0.019)	0.049(-0.033)
β_{12}	0.227(-0.049)	0.037(-0.081)	0.034(0.021)	0.046(0.000)	0.045(0.009)	0.031(0.008)	0.048(-0.043)
β_{20}	0.066(0.018)	0.008(-0.017)	0.007(-0.007)	0.006(-0.007)	0.006(0.011)	0.008(-0.011)	0.006(0.008)
β_{21}	0.069(0.055)	0.007(0.001)	0.006(-0.025)	0.007(-0.005)	0.009(-0.008)	0.007(0.003)	0.010(0.005)
β_{22}	0.069(0.055)	0.009(0.006)	0.008(-0.025)	0.008(0.009)	0.010(-0.001)	0.008(0.011)	0.012(0.003)
π_1	0.010(-0.017)	0.002(0.023)	0.002(0.023)	0.002(0.004)	0.003(0.007)	0.002(-0.001)	0.003(0.003)
Case V: $\varepsilon \sim 0.95N(0,1) + 0.05N(0,25)$							
β_{10}	0.098(0.000)	0.041(0.005)	0.024(0.004)	0.029(-0.007)	0.038(0.015)	0.034(0.009)	0.042(0.028)
β_{11}	0.394(0.028)	0.048(-0.095)	0.021(0.027)	0.022(0.011)	0.044(-0.012)	0.025(0.003)	0.040(-0.012)
β_{12}	0.081(-0.050)	0.051(-0.119)	0.022(0.014)	0.026(0.001)	0.045(0.012)	0.032(0.000)	0.048(-0.001)
β_{20}	0.041(0.015)	0.006(0.003)	0.005(0.002)	0.006(-0.002)	0.006(0.006)	0.008(-0.006)	0.008(0.003)
β_{21}	0.088(0.046)	0.006(0.010)	0.005(-0.008)	0.006(0.006)	0.009(0.004)	0.008(0.009)	0.011(0.009)
β_{22}	0.135(0.041)	0.007(0.024)	0.004(0.000)	0.005(0.002)	0.008(0.000)	0.007(0.008)	0.011(0.007)
π_1	0.007(-0.033)	0.001(0.003)	0.001(0.006)	0.001(0.000)	0.002(-0.002)	0.002(-0.003)	0.002(-0.007)
Case VI: $\varepsilon \sim N(0,1)$ with 5% high leverage outliers							
β_{10}	9.355(-1.688)	0.033(0.010)	0.020(-0.010)	1.375(0.246)	0.021(-0.014)	0.013(0.065)	0.029(0.002)
β_{11}	5.188(1.667)	0.049(-0.102)	0.023(-0.011)	2.505(1.479)	0.027(-0.002)	0.014(-0.049)	0.033(-0.037)
β_{12}	4.187(1.307)	0.039(-0.098)	0.021(-0.007)	2.594(1.507)	0.029(0.007)	0.017(-0.034)	0.031(-0.015)
β_{20}	11.697(2.305)	0.005(0.002)	0.004(0.003)	0.005(0.005)	0.005(0.004)	0.007(-0.047)	0.007(-0.002)
β_{21}	11.586(3.309)	0.006(0.011)	0.005(0.012)	0.021(0.125)	0.006(0.004)	0.004(0.026)	0.009(0.005)
β_{22}	12.442(3.437)	0.006(0.003)	0.005(0.003)	0.020(0.122)	0.006(-0.005)	0.005(0.028)	0.010(0.000)
π_1	0.140(0.204)	0.002(0.004)	0.001(-0.006)	0.008(-0.089)	0.001(0.005)	0.004(0.020)	0.001(0.002)

Table 3: MSE(Bias) of Point Estimates for $n = 400$

The left plot in Figure 1 clearly shows that the fitting by MixregL and Bisquare are almost identical, and Mixregt also provides a very good fit. For comparison, The Bisquare fit is also drawn in the right plot in Figure 1, it is quite obvious that the TLE and MLE are affected severely by the outliers. Then we add 10 identical pairs, $(0, 3)$, to the original data set as high leverage outliers.

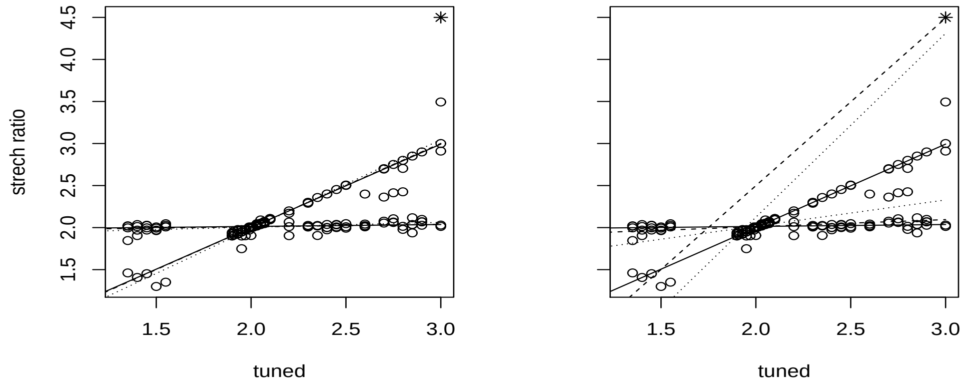

Figure 1. Mixture Linear Fitting with Outlier $(3, 4.5)$
Left panel: solid line – Bisquare, dashed line – MixregL, dotted line – Mixregt,
Right panel: solid line – Bisquare, dashed – TLE, dotted line – MLE

The left plot in Figure 2 shows that both Bisquare and MixregL gives a reasonable fit, but the Mixregt performs less satisfying. From the right plot in Figure 2, we see that MLE has inferior performance against the outliers, and TLE works better.

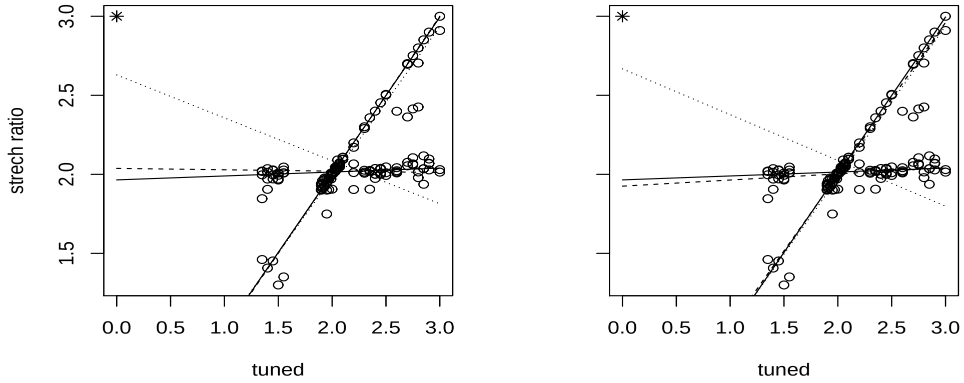

Figure 2. Mixture Linear Fitting with Outlier $(0, 3)$
Left panel: solid line – Bisquare, dashed line – MixregL, dotted line – Mixregt,
Right panel: solid line – Bisquare, dashed – TLE, dotted line – MLE

Finally 10 identical pairs $(0, 4)$ are added to the original data set as both outliers in x and y-direction. The left plot in Figure 3 shows that Bisquare continues to provide a robust fit, MixregL barely keeps a vague two-line structure, and Mixregt is affected severely by the outliers. The right plot in Figure 3 shows that MLE is still the worst, and TLE works fine.

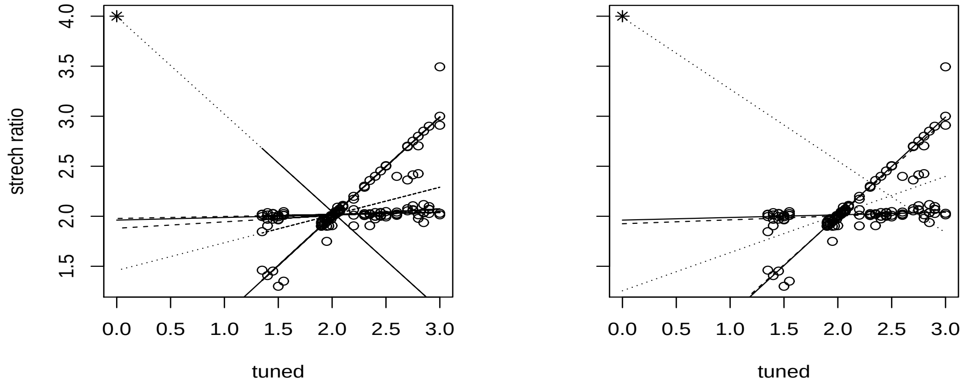

Figure 3. Mixture Linear Fitting with Outlier $(0, 4)$
Left panel: solid line – Bisquare, dashed line – MixregL, dotted line – Mixregt,
Right panel: solid line – Bisquare, dashed – TLE, dotted line – MLE

In all the scenarios, the Bisquare performs uniformly better than other fitting procedures, although the simulation studies show that Bisquare is less satisfying in some cases, such as when $\varepsilon \sim t$-distributions. Generally MixregL performs better than Mixregt, but both procedures are not quite robust to the high leverage outliers. We also applied Mixregt-MCD and MixregL-MCD to the data set, both procedures can successfully remove the high leverage outliers and give similar results to the Bisquare.

4 Conclusion

In this paper, we propose a new robust estimation procedure tailored to the mixture linear regression models by assuming the random error has a Laplace distribution. The robustness is achieved essentially by LAD procedure, and implemented by the EM algorithm. The efficiency and effectiveness of the proposed EM algorithm depends upon the fact that the Laplace distribution indeed is a scale mixture of a normal distribution and a distribution of a function of exponentially distributed random variable. The simulation study shows that the proposed method is superior to and comparable to existing robust estimation procedures in all simulation setups. However, the real data example shows that when the high leverage outliers exist, then the trimmed version of the proposed procedure should be used.

References

[1] Andrews, D.F. and Mallows, C.L. (1974). Scale mixtures of normal distributions. *J.R.S.S. (B)*, **36**(1), 99-102.

[2] Bai, X., Yao, W., and Boyer, J. E. (2012). Robust fitting of mixture regression models. *Computational Statistics and Data Analysis*, **56**, 2347-2359.

[3] Cohen, E. (1984). Some effects of inharmonic partials on interval perception. *Music Perception*, **1**, 323-349.

[4] Dielman, T.E. (1984). Leasl absolute value estimation in regression models: An annotated bibliography. *Communications in Statistics - Theory and Methods*, **4**, 513-541.

[5] Dielman, T.E. (2005). Least absolute value regression: recent contributions. *Journal of Statistical Computation and Simulation*, **75**(4), 263-286.

[6] Donoho, D. L. (1982). Breakdown properties of multivariate location estimators. Qualifying paper, Harvard University, Boston.

[7] Li, Y. and Arce, G.R. (2004). A maximum likelihood approach to least absolute deviation regression. *Jouranl of applied signal processing*, **12**, 1762-1769.

[8] McLachlan, G.J., Peel, D., (2000). Finite mixture models. Wiley, New York.

[9] Neykov, N., Filzmoser, P., Dimova, R., and Neytchev, P. (2007). Robust fitting of mixtures using the trimmed likelihood estimator. *Computational Statistics and Data Analysis*, **52**, 299-308.

[10] Phillips, R.F. (2002). Least absolute deviations estimation via the EM algorithm. *Statistics and Computing*, **12**, 281-285.

[11] Pison, G., Van Aelst, S. and Willems, G. (2002). Small sample corrections for LTS and MCD. *Metrika*, **55**, 111-123.

[12] Rousseeuw, P.J., Van Driessen, K., (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, **41**, 212-223.

[13] Schlossmacher, E.J. (1973). An iterative technique for absolute deviations curve fitting. *JASA*, **68**, 857859.

[14] Stahel, W. A. (1981). Robuste Schätzungen: Infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen. Ph.D. thesis, ETH Zürich.