Robust mixture regression using the $t-$distribution

Weixin Yao,* Yan Wei, and Chun Yu

Kansas State University

Abstract

The traditional estimation of mixture regression models is based on the normal assumption of component errors and thus is sensitive to outliers or heavy-tailed errors. A robust mixture regression model based on the $t-$distribution by extending the mixture of $t-$distributions to the regression setting is proposed. However, this proposed new mixture regression model is still not robust to high leverage outliers. In order to overcome this, a modified version of the proposed method, which fits the mixture regression based on the $t-$distribution to the data after adaptively trimming high leverage points, is also proposed. Furthermore, it is proposed to adaptively choose the degrees of freedom for the $t-$distribution using profile likelihood. The proposed robust mixture regression estimate has high efficiency due to the adaptive choice of degrees of freedom.

Key words: EM algorithm; Mixture regression models; Outliers; Robust regression; $t-$distribution.

*Corresponding Author. Weixin Yao is Associate Professor, Department of Statistics, Kansas State University, Manhattan, Kansas 66506, U.S.A. Email: wxyao@ksu.edu.

1 Introduction

Mixture regression models are well known as switching regression models in econometrics literature, which were introduced by Goldfeld and Quandt (1973). These models have been widely used to investigate the relationship between variables coming from several unknown latent homogeneous groups and applied in many fields, such as business, marketing, and social sciences (Jiang and Tanner, 1999; Böhning, 1999; Wedel and Kamakura, 2000; McLachlan and Peel, 2000; Skrondal and Rabe-Hesketh, 2004; Frühwirth-Schnatter, 2006).

Let Z be a latent class variable such that given $Z = j$, the response y depends on the $p-$dimensional predictor $\mathbf{x}$ in a linear way

$$y = \mathbf{x}^T\boldsymbol{\beta}_j + \epsilon_j, j = 1, 2, \cdots, m, \tag{1.1}$$

where m is the number of homogeneous groups (also called components in mixture models) in the population and $\epsilon_j \sim N(0, \sigma_j^2)$ is independent of $\mathbf{x}$. Suppose $P(Z = j) = \pi_j, j = 1, 2, \cdots, m$, and Z is independent of $\mathbf{x}$, then the conditional density of Y given $\mathbf{x}$, without observing Z, is

$$f(y|\mathbf{x}, \boldsymbol{\theta}) = \sum_{j=1}^{m} \pi_j \phi(y; \mathbf{x}^T\boldsymbol{\beta}_j, \sigma_j^2), \tag{1.2}$$

where $\phi(\cdot; \mu, \sigma^2)$ is the density function of $N(\mu, \sigma^2)$ and $\boldsymbol{\theta} = (\pi_1, \boldsymbol{\beta}_1, \sigma_1, \ldots, \pi_m, \boldsymbol{\beta}_m, \sigma_m)^T$. The model (1.2) is the so called *mixture of regression models*. Hennig (2000) proved identifiability of model (1.2) under some general conditions for the covariates. In general, the model (1.2) is identifiable if the number of components, m, is smaller than the number of distinct $(p-1)-$dimensional hyperplanes that one needs to cover the covariates of each cluster. The above conditions are usually satisfied if the domain of $\mathbf{x}$ contains an

open set in $\mathbb{R}^p$

The unknown parameter $\boldsymbol{\theta}$ in (1.2), given observations $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, is traditionally estimated by the maximum likelihood estimate (MLE):

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log \left[\sum_{j=1}^{m} \pi_j \phi(y_i; \mathbf{x}_i^T \boldsymbol{\beta}_j, \sigma_j^2) \right]. \tag{1.3}$$

Note that the maximizer of (1.3) does not have an explicit solution and is usually estimated by the EM algorithm (Dempster et al., 1977).

It is well known that the log-likelihood function (1.3) is unbounded and goes to infinity if one or more observations lie exactly on one component hyperplane and the corresponding component variance goes to zero. When running the EM algorithm, some initial values might converge to the boundary point with small variance and very large log-likelihood. In such situations, our objective is to find a local maximum of (1.3) in the interior of parameter space (Kiefer, 1978; Peters and Walker, 1978). However, the challenge is to find this interior local maximum. Hathaway (1985, 1986) proposed putting some constraints on the parameter space such that the component variance has some low limit. Yao (2010) proposed using the profile likelihood and a graphical method to locate the interior local maximum. Practically, the interior local maximum can usually be found by starting from some "good" initial values such as the K-means (MacQueen, 1967) and the moment method estimator (Lindsay and Basak, 1993). Chen et al. (2008) also proposed using a penalized likelihood method to avoid the unboundedness of mixture likelihood. In this article, for simplicity of computation and comparison, we assume equal variance for all components.

The MLE $\hat{\boldsymbol{\theta}}$ in (1.3) works well when the error distribution is normal. However, the normality based MLE is sensitive to outliers or heavy-tailed error distributions. There is little research about how to estimate the mixture regression parameters robustly.

Markatou (2000) and Shen et al. (2004) proposed using a weight factor for each data point to robustify the estimation procedure for mixture regression models. Neykov et al. (2007) proposed robust fitting of mixtures using the trimmed likelihood estimator (TLE). Bai et al. (2012) proposed a modified EM algorithm to robustly estimate the mixture regression parameters by replacing the least squares criterion in M step with a robust criterion. Bashir and Carter (2012) extended the idea of the S-estimator to mixture of linear regression. There are also some related robust methods for linear clustering (Hennig, 2002, 2003; Mueller and Garlipp, 2005; García-Escudero et al., 2009; García-Escudero et al., 2010).

In this article, we propose a new robust mixture regression model by extending the mixture of $t-$distributions proposed by Peel and McLachlan (2000) to the regression setting. Similar to the traditional M-estimate for linear regression (Maronna et al., 2006), the proposed estimate is expected to be sensitive to high leverage outliers. To overcome this problem, we also propose a modified version of the new method by fitting the new model to the data after adaptively trimming high leverage points. Compared to the TLE, the proportion of trimming of our new method is data adaptive instead of a fixed value. In addition, we propose to use the profile likelihood to adaptively choose the degrees of freedom for the $t-$distribution. The proposed estimate has high efficiency, i.e., comparable performance to the traditional MLE when the error is normal, due to the adaptive choice of degrees of freedom. Using a simulation study and real data application, we compare the new method to some existing methods, and demonstrate the effectiveness of the proposed method.

The rest of this article is organized as follows. In Section 2, we introduce our new robust mixture linear regression models based on the $t-$distribution. In Section 3, we propose to further improve the robustness of the proposed method against high leverage outliers by adaptively trimming high leverage points. In Section 4, we introduce how

to adaptively choose the degrees of freedom for the $t-$distribution. In Section 5, we compare the proposed method to the traditional MLE and some other robust methods by using a simulation study and real data application. Section 6 contains a discussion of possible future work.

2 Robust Mixture Regression Using the $t-$distribution

In order to more robustly estimate the mixture regression parameters in (1.2), we assume that the error density $f_j(\epsilon)$ is a $t-$distribution with degrees of freedom ν_j and scale parameter σ:

$$f(\epsilon;\sigma,\nu) = \frac{\Gamma(\frac{\nu+1}{2})\sigma^{-1}}{(\pi\nu)^{\frac{1}{2}}\Gamma(\frac{\nu}{2})\left\{1+\frac{\epsilon^2}{\sigma^2\nu}\right\}^{\frac{1}{2}(\nu+1)}}. \tag{2.1}$$

We first assume that ν_js are known. We will discuss about how to adaptively choose ν_js in Section 4. The unknown parameter $\boldsymbol{\theta}$ in (1.2) can be estimated by maximizing the log-likelihood

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\left\{\sum_{j=1}^{m} \pi_j f(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_j;\sigma,\nu_j)\right\}. \tag{2.2}$$

Note, however, the above log-likelihood does not have an explicit maximizer. Here, we also propose to use an EM algorithm to simplify the computation. Let

$$z_{ij} = \begin{cases} 1, & \text{if the } i\text{th observation is from the } j\text{th component;} \\ 0, & \text{otherwise.} \end{cases},$$

where $i = 1,\cdots,n, j = 1,\cdots,m$. Then the complete likelihood for $(\mathbf{y},\mathbf{z})$ given $\mathbf{X}$ is

$$\ell_c(\boldsymbol{\theta};\mathbf{y},\mathbf{z}) = \sum_{i=1}^{n}\sum_{j=1}^{m} z_{ij}\log\{\pi_j f(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_j;\sigma,\nu_j)\},$$

where $\mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_n)^T, \mathbf{y} = (y_1,\ldots,y_n)$, and $\mathbf{z} = (z_{11},\ldots,z_{nm})$. Based on the theory

of the EM algorithm, in E step, given the current estimate $\boldsymbol{\theta}^{(k)}$ at the kth step, we calculate $\mathrm{E}(\ell_c(\boldsymbol{\theta}; \mathbf{y}, \mathbf{z}) \mid \mathbf{X}, \mathbf{y}, \boldsymbol{\theta}^{(k)})$ which simplifies to the calculation of $\mathrm{E}(z_{ij} \mid \mathbf{X}, \mathbf{y}, \boldsymbol{\theta}^{(k)})$. Then in M step, we find the maximizer of

$$\mathrm{E}(\ell_c(\boldsymbol{\theta}; \mathbf{y}, \mathbf{z}) \mid \mathbf{X}, \mathbf{y}, \boldsymbol{\theta}^{(k)}) = \sum_{i=1}^{n} \sum_{j=1}^{m} \mathrm{E}(z_{ij} \mid \mathbf{X}, \mathbf{y}, \boldsymbol{\theta}^{(k)}) \log\{\pi_j f(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_j; \sigma, \nu_j)\}.$$

Note that the above maximizer does not have explicit solutions for $\boldsymbol{\beta}_j$ and σ.

The computation can be further simplified based on the fact that the $t-$distribution can be considered a scale mixture of normal distributions. Let u be the latent variable such that

$$\epsilon | u \sim N(0, \sigma^2/u), \ u \sim \text{gamma}(\frac{1}{2}\nu, \frac{1}{2}\nu), \tag{2.3}$$

where $\text{gamma}(\alpha, \gamma)$ has density

$$f(u; \alpha, \gamma) = \frac{1}{\Gamma(\alpha)} \gamma^\alpha u^{\alpha-1} e^{-\gamma u}, \ u > 0.$$

Then, marginally ϵ has a $t-$distribution with degrees of freedom ν and scale parameter σ. Therefore, we can simplify the computation of M step of the proposed EM algorithm by introducing another latent variable u.

Note that the complete likelihood for $(\mathbf{y}, \mathbf{u}, \mathbf{z})$ given $\mathbf{X}$ is

$$\begin{aligned}
\ell_c(\boldsymbol{\theta}; \mathbf{y}, \mathbf{u}, \mathbf{z}) =& \sum_{i=1}^{n} \sum_{j=1}^{m} z_{ij} \log\{\pi_j \phi(y_i; \mathbf{x}_i^T \boldsymbol{\beta}_j, \sigma^2/u_i) f(u_i; \frac{1}{2}\nu_j, \frac{1}{2}\nu_j)\}, \\
=& \sum_{i=1}^{n} \sum_{j=1}^{m} z_{ij} \log(\pi_j) + \sum_{i=1}^{n} \sum_{j=1}^{m} z_{ij} \log\{f(u_i; \frac{1}{2}\nu_j, \frac{1}{2}\nu_j)\}, \\
&+ \sum_{i=1}^{n} \sum_{j=1}^{m} z_{ij} \left\{ -\frac{1}{2} \log(2\pi\sigma^2) + \frac{1}{2} \log(u_i) - \frac{u_i}{2\sigma^2} (y_i - \mathbf{x}_i^T \boldsymbol{\beta}_j)^2 \right\},
\end{aligned} \tag{2.4}$$

where $\mathbf{u} = (u_1, \ldots, u_n)$ is independent of $\mathbf{z}$. Since the above second term does not involve unknown parameters, based on the theory of the EM algorithm, in E step, given the current estimate $\boldsymbol{\theta}^{(k)}$ at the kth step, the calculation of $\mathrm{E}(\ell_c(\boldsymbol{\theta}; \mathbf{y}, \mathbf{u}, \mathbf{z}) \mid \mathbf{X}, \mathbf{y}, \boldsymbol{\theta}^{(k)})$ simplifies to the calculation of $\mathrm{E}(z_{ij} \mid \mathbf{X}, \mathbf{y}, \boldsymbol{\theta}^{(k)})$ and $\mathrm{E}(u_i \mid \mathbf{X}, \mathbf{y}, \boldsymbol{\theta}^{(k)}, z_{ij} = 1)$. Then in M step, we find the maximizer of

$$\begin{aligned}
&\mathrm{E}(\ell_c(\boldsymbol{\theta}; \mathbf{y}, \mathbf{u}, \mathbf{z}) \mid \mathbf{X}, \mathbf{y}, \boldsymbol{\theta}^{(k)}) \\
&\propto \sum_{i=1}^{n} \sum_{j=1}^{m} \mathrm{E}(z_{ij} \mid \mathbf{X}, \mathbf{y}, \boldsymbol{\theta}^{(k)}) \left[\log(\pi_j) - \frac{1}{2} \log(2\pi\sigma^2) - \frac{\mathrm{E}(u_i \mid \mathbf{X}, \mathbf{y}, \boldsymbol{\theta}^{(k)}, z_{ij} = 1)}{2\sigma^2} (y_i - \mathbf{x}_i^T \boldsymbol{\beta}_j)^2 \right]
\end{aligned} \tag{2.5}$$

which has an explicit solution for $\boldsymbol{\theta}$.

Based on the above, we propose the following EM algorithm to maximize (2.2).

Algorithm 2.1. *Given the initial parameter estimate $\boldsymbol{\theta}^{(0)}$, at the $(k+1)$th iteration, we calculate the following two steps:*

E step: *Calculate*

$$p_{ij}^{(k+1)} = E(z_{ij} \mid \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta}^{(k)}) = \frac{\pi_j^{(k)} f(y_i - \boldsymbol{x}_i^T \beta_j^{(k)}; \sigma^{(k)}, \nu_j)}{\sum_{l=1}^{m} \pi_l^{(k)} f(y_i - \boldsymbol{x}_i^T \beta_l^{(k)}; \sigma^{(k)}, \nu_l)} \tag{2.6}$$

and

$$u_{ij}^{(k+1)} = E(u_i \mid \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta}^{(k)}, z_{ij} = 1) = \frac{\nu + 1}{\nu + \left\{ (y_i - \boldsymbol{x}_i^T \beta_j^{(k)}) / \sigma^{(k)} \right\}^2}, \tag{2.7}$$

where $f(\epsilon; \sigma, \nu)$ is defined in (2.1).

M step: *Update parameter estimates:*

$$\pi_j^{(k+1)} = \sum_{i=1}^{n} p_{ij}^{(k+1)}/n, \tag{2.8}$$

$$\boldsymbol{\beta}_j^{(k+1)} = \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_j^T p_{ij}^{(k+1)} u_{ij}^{(k+1)}\right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{x}_i y_i p_{ij}^{(k+1)} u_{ij}^{(k+1)}\right), \tag{2.9}$$

and

$$\sigma^{(k+1)} = \left\{ \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}^{(k+1)} u_{ij}^{(k+1)} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_j^{(k+1)})^2}{n} \right\}^{1/2}. \tag{2.10}$$

Theorem 2.1. *Each iteration of the E step and M step of Algorithm 2.1 monotonically non-decreases the objective function (2.2), i.e.,* $\ell(\boldsymbol{\theta}^{(k+1)}) \geq \ell(\boldsymbol{\theta}^{(k)})$*, for all* $k \geq 0$.

The proof of the above theorem is simple and omitted here. Based on (2.9) in M step, the regression parameters can be considered a weighted least squares estimate with the weights depending on $u_{ij}^{(k+1)}$. Based on (2.7) in E step, the weights $u_{ij}^{(k+1)}$ decrease if the standardized residuals increase. Therefore, the weights $u_{ij}^{(k+1)}$ reduce the effects of the outliers and provide a robust estimate for the mixture regression parameters. In addition, based on (2.10) in M step, the larger residuals also have smaller effects on $\sigma_j^{(k+1)}$ due to the weights $u_{ij}^{(k+1)}$.

Hennig (2004) showed that the mixture of $t-$distributions proposed by Peel and McLachlan (2000) has a low breakdown point. We expect similar results from the proposed mixture regression models based on the $t-$distribution. However, Hennig (2004) mentioned that only very extreme outliers can lead to the breakdown of mixture of $t-$distributions. Our real data application in Section 5 further confirms this finding. Therefore, we believe that the $t-$distribution can still be used as an alternative tool to provide a robust estimation for the mixture model against modest outliers.

3 Adaptively Trimmed Version

Similar to the traditional M-estimate for linear regression (Maronna et al., 2006), the proposed mixture regression model based on the $t-$distribution is sensitive to high leverage outliers. To overcome this problem, we then propose a trimmed version of the new method by fitting the new model to the data after adaptively trimming high leverage points. In addition, unlike TLE (Neykov et al., 2007), the proportion of trimming of the new method is data adaptive instead of a fixed value.

Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ and h_{ii} be the ith diagonal of H, where $H = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. Then, h_{ii} is called the leverage for the ith predictor $\mathbf{x}_i$ and $\mathbf{x}_i$ is considered a high leverage point if h_{ii} is large.

Note however

$$h_{ii} = n^{-1} + (n-1)^{-1}\mathrm{MD}_i, \tag{3.1}$$

where

$$\mathrm{MD}_i = (\mathbf{x}_i - \bar{\mathbf{x}})^T\mathbf{S}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})$$

is the Mahalanobis distance, $\bar{\mathbf{x}}$ is the sample mean of $\mathbf{x}_i$s , and $\mathbf{S}$ is the sample covariance of $\mathbf{x}_i$s (without the intercept 1). It is well known that $\bar{\mathbf{x}}$ and $\mathbf{S}$ are not resistant to outliers and might create the *masking effect* (Rousseeuw and van Zomeren, 1990), i.e., some high leverage points might not be identified due to the influence of other high-leverage points. In order to overcome this, a modified Mahalanobis distance is proposed

$$\mathrm{MD}_i = (\mathbf{x}_i - \mathbf{m}(\mathbf{X}))^T\mathbf{C}(\mathbf{X})^{-1}(\mathbf{x}_i - \mathbf{m}(\mathbf{X})),$$

where $\mathbf{m}(\mathbf{X})$ and $\mathbf{C}(\mathbf{X})$ are robust estimates of location and scatter for $\mathbf{X}$ (after removing the first column 1s).

We propose to use the minimum covariance determinant (MCD, Rosseeuw, 1984) es-

timators for $\mathbf{m}(\mathbf{X})$ and $\mathbf{C}(\mathbf{X})$ and implement it by the Fast MCD algorithm of Rousseeuw and Van Driessen (1999). Note that the resulting robust estimate MD_i is the same as the robust distance proposed by Rousseeuw and Leroy (1987). After getting the robust estimate MD_i, we propose to trim the data based on the cut point $\chi^2_{p-1,0.975}$, which is proposed by Pison et al. (2002) to improve the finite-sample efficiency for the raw MCD estimator using a one-step weighted estimate. Therefore, to make the proposed method also robust against high leverage outliers, we propose to implement the proposed mixture of regression based on the $t-$distribution after trimming the observations with $\mathrm{MD}_i > \chi^2_{p-1,0.975}$.

We might also utilize some other robust estimates for $\mathbf{m}(\mathbf{X})$ and $\mathbf{C}(\mathbf{X})$. There have been many robust estimators proposed for multivariate location and scatter, such as the Stahel-Donoho estimator (Stahel, 1981; Donoho, 1982), minimum volume ellipsoid (MVE) estimator (Rousseeuw, 1984), S-estimator (Rousseeuw and Leroy, 1987; Davies, 1987), and the depth based estimator (Donoho and Gasko, 1992; Liu et al., 1999; Zuo and Serfling, 2000; Zuo et al., 2004).

4 Adaptive Choice of the Degrees of Freedom for the $t-$distribution

In previous sections, we assume that the degrees of freedom ν_js for the $t-$distribution are known. In this section, we discuss how to adaptively choose ν. We first consider the case where $\nu_1 = \nu_2 = \cdots = \nu_m = \nu$. We will further discuss the case where ν_js are different later.

When ν is unknown, we typically estimate ν and the mixture regression parameter $\boldsymbol{\theta}$ by maximizing the log-likelihood (2.2) over both ν and $\boldsymbol{\theta}$. In order to maximize the

log-likelihood (2.2), we define the *profile likelihood* for ν:

$$L(\nu) = \max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{m} \pi_j f(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_j; \sigma, \nu) \right\}. \tag{4.1}$$

For each fixed ν, we can easily find $L(\nu)$ based on Algorithm 2.1. Then we propose to estimate ν by

$$\hat{\nu} = \arg\max_{\nu} L(\nu).$$

In practice, we can calculate $L(\nu)$ in a set of grid points of ν, say $\nu = 1, \ldots, \nu_{\max}$.

Note that the above proposed profile method can be also applied to the case where ν_js are different, however, the computation will be intensive when m is large, since we need to compute $L(\nu)$ for $\nu_{\max}^m$ times.

Similar to Peel and McLachlan (2000), we can also incorporate the estimation of ν_js in the EM Algorithm 2.1. Based on the complete likelihood (2.4), at the $(k+1)$th iteration of M step given the current estimate $\boldsymbol{\theta}^{(k)}$, we can update ν_j by

$$\nu_j^{(k+1)} = \arg\max_{\nu_j} \mathrm{E} \left[\sum_{i=1}^{n} z_{ij} \log\{ f(u_i; \frac{1}{2}\nu_j, \frac{1}{2}\nu_j) \} \mid \mathbf{X}, \mathbf{y}, \boldsymbol{\theta}^{(k)} \right]. \tag{4.2}$$

Note that

$$\mathrm{E}\left[\log(u_i) \mid \mathbf{X}, \mathbf{y}, \boldsymbol{\theta}^{(k)}, z_{ij} = 1 \right] = \psi\left(\frac{\nu_j^{(k)} + 1}{2} \right) - \log\left(\frac{\nu_j^{(k)} + \left\{ (y_i - \mathbf{x}_i^T \beta_j^{(k)})/\sigma^{(k)} \right\}^2}{2} \right) \triangleq v_{ij}^{(k+1)},$$

where $\psi(t) = \partial \log(\Gamma(t))/\partial t$ is the Digamma function. Therefore, (4.2) is equivalent to

$$\nu_j^{(k+1)} = \arg\max_{\nu_j} \sum_{i=1}^{n} p_{ij}^{(k+1)} \left[-\log\Gamma(0.5\nu_j) + 0.5\nu_j \log(0.5\nu_j) + 0.5\nu_j \left\{ v_{ij}^{(k+1)} - u_{ij}^{(k+1)} \right\} - v_{ij}^{(k+1)} \right]. \tag{4.3}$$

Note that (4.3) does not have an explicit formula. We might use some numerical algo-

rithms to solve it or simply use grid search for $\nu_j = 1 \ldots, \nu_{\max}$.

5 Examples

5.1 Simulation studies

In this section, we use a simulation study to demonstrate the effectiveness of the proposed method and compare the following five methods:

1. traditional MLE assuming the error has a normal density (MLE),
2. trimmed likelihood estimator (TLE) proposed by Neykov et al. (2007) with the percentage of trimmed data α set to 0.1, (The choice of α plays an important role for the TLE. If α is too large, the TLE will lose much efficiency. If α is too small and the percentage of outliers is more than α, then the TLE will fail. In our simulation study, the proportion of outliers is never greater than 0.1.)
3. the robust modified EM algorithm based on bisquare (MEM-bisquare) proposed by Bai et al. (2012),
4. the proposed robust mixture regression model based on the $t-$distribution (Mixregt),
5. the proposed trimmed version of Mixregt (Mixregt-trim).

To compare the different methods, we report the mean squared errors (MSE) and the bias of the parameter estimates for each estimation method. However, under mixture models, there are well known label switching issues (Celeux, et al., 2000; Stephens, 2000; Yao and Lindsay, 2009; Yao, 2012) when performing comparisons using simulation studies. There are no generally accepted labeling methods. In our simulation study, we choose the labels by minimizing the distance to the true parameter values. However, more research comparing different labeling methods is needed.

Example 1. Suppose the independently and identically distributed samples $\{(x_{1i}, x_{2i}, y_i), i = 1, \ldots, n\}$ are sampled from the model

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

where Z is a component indicator of Y with $P(Z = 1) = 0.25$, $X_1 \sim N(0, 1), X_2 \sim N(0, 1)$, and ϵ_1 and ϵ_2 have the same distribution as ϵ. We consider the following five cases for the error density of ϵ:

Case I: $\epsilon \sim N(0, 1)$ – standard normal distribution,

Case II: $\epsilon \sim t_3$ – the $t-$distribution with degrees of freedom 3,

Case III: $\epsilon \sim t_1$ – the $t-$distribution with degree of freedom 1 (Cauchy distribution),

Case IV: $\epsilon \sim 0.95N(0, 1) + 0.05N(0, 5^2)$ – contaminated normal mixture,

Case V: $\epsilon \sim N(0, 1)$ with 5% of high leverage outliers being $X_1 = 20, X_2 = 20$, and $Y = 100$.

Case I is used to test the efficiency of different estimation methods compared to the traditional MLE when the error is exactly normally distributed and there are no outliers. Case II is a heavy-tailed distribution. The t-distributions with degrees of freedom from 3 to 5 are often used to represent the heavy-tailed distributions. Case III is a Cauchy distribution which has extreme heavy tails. The contaminated normal mixture model in Case IV is often used to mimic the situation with outliers. The 5% data from $N(0, 5^2)$ are likely to be low leverage outliers. In Case V, 95% of the observations have the error distribution $N(0, 1)$, but 5% of the observations are identical high leverage outliers with $X_1 = 20, X_2 = 20$, and $Y = 100$.

In this example, we check the performance of the proposed profile likelihood method assuming all ν_js are equal. Note that when ν is large enough, the $t-$distribution is close to a normal distribution. Therefore, in practice, $\nu_{\max}$ does not need to be large. In this example, we set $\nu_{\max} = 15$. However, one might choose a larger $\nu_{\max}$ for real data application.

Tables 1 and 2 report the mean squared errors (MSE) and the *absolute* bias (Bias) of the parameter estimates for each estimation method for sample size $n = 200$ and $n = 400$, respectively. The number of replicates is 200. As shown in Table 1 and 2, in Case I through IV, Mixregt and Mixregt-trim performed at a level that is better or equal to the other three methods. In case V where there are high leverage outliers, Mixregt-trim also outperformed the other four methods. Specifically, we have the following findings:

1. MLE worked best in Case I ($\epsilon \sim N(0, 1)$), but failed to provide reasonable estimates in Case II to V.

2. Mixregt and Mixregt-trim performed better than MEM-bisquare in Case I, II, and IV when $n = 200$, but performed comparably to MEM-bisquare when $n = 400$. In addition, Mixregt and Mixregt-trim also performed better than MEM-bisquare in Case III when $n = 400$.

3. Mixregt, Mixregt-trim, and MEM-bisquare performed better than TLE in Case I to IV.

4. In Case V, where there are high leverage outliers, Mixregt-trim worded best. In addition, TLE and MEM-bisquare also worked better than Mixregt and MLE.

In order to check the performance of the proposed profile likelihood for the selection of degrees of freedom for $t-$distribution, in Table 4, we report the mean and median of estimated degrees of freedom for Mixregt and Mixregt-trim. The degrees of freedom were chosen based on the grid points from $[1, v_{\max}]$, where $v_{\max} = 15$ was used in our

simulation study. Therefore, for Case I−normal distribution, the "optimal" solution is $v_{\max} = 15$. Based on the results of Case I, II, and III in Table 4, the proposed profile likelihood adaptively estimated the degrees of freedom for $t-$distribution. In Case IV, although the true error density is not a $t-$distribution, both Mixregt and Mixregt-trim were able to use a heavy-tailed $t-$distribution to approximate the contaminated normal mixture to produce a robust estimate for mixture regression parameters. In Case V, the estimated degrees of freedom for Mixregt-trim are close to $v_{\max} = 15$. Therefore, Mixregt-trim successfully trimmed the high leverage outliers and recovered the original normal error density.

Example 2. In this example, we consider a case where the number of components is larger than two and the components are close. We generate the independent and identically distributed (i.i.d.) data $\{(x_i, y_i), i = 1, \ldots, n\}$ from the model

$$Y = \begin{cases} 1 + X + \epsilon_1, & \text{if } Z = 1; \\ 2 + 2X + \epsilon_2, & \text{if } Z = 2; \\ 3 + 5X + \epsilon_3, & \text{if } Z = 3; \end{cases},$$

where Z is a component indicator of Y with $P(Z = 1) = P(Z = 2) = 0.3, P(Z = 3) = 0.4$, and $X \sim N(0, 1)$. Note that in this case all three components have the same sign of the slopes and the first two components are very close. We consider the following four cases for component error densities:

Case I: ϵ_1, ϵ_2, and ϵ_3 have the same distribution from $N(0, 1)$,

Case II: $\epsilon_1 \sim t_9$, $\epsilon_2 \sim t_6$ and $\epsilon_3 \sim t_3$,

Case III: $\epsilon_1 \sim N(0, 1)$, $\epsilon_2 \sim N(0, 1)$, and $\epsilon_3 \sim t_3$,

Case VI: ϵ_1, ϵ_2, and ϵ_3 have the same distribution from $N(0, 1)$ with 5% of high leverage outliers being $X = 20$ and $Y = 200$.

In this example, we compare the performance of the proposed method, when all ν_js are assumed to be unequal, to the other methods. Tables 4 and 5 report the mean squared errors (MSE) and the *absolute* bias (Bias) of the parameter estimates for each estimation method for sample size $n = 200$ and $n = 400$, respectively. In Case I where the error is normal, all five methods worked well and MLE, MEM-bisquare, Mixregt, and Mixregt-trim worked better than TLE. In Case II and III where the errors have heavy tails, all four robust methods performed well but MLE failed. In Case IV where there are high leverage outliers, TLE, MEM-bisquare, and Mixregt-trim still worked well, but MLE and Mixregt failed.

5.2 Real data application

We further apply the proposed robust procedure to tone perception data (Cohen, 1984). In the tone perception experiment of Cohen (1984), a pure fundamental tone with electronically generated overtones added was played to a trained musician. The experiment recorded 150 trials from the same musician. The overtones were determined by a stretching ratio, which is the ratio between adjusted tone and the fundamental tone. The purpose of this experiment was to see how this tuning ratio affects the perception of the tone and to determine whether either of two musical perception theories was reasonable.

To better illustrate the robustness of the proposed estimation procedure, we added ten identical outliers $(1.5, 5)$ to the original data set, and fit the data with both MLE and Mixregt. Figure 1 shows the scatter plot of the data with the estimated regression lines generated by the traditional MLE (dashed lines) and the proposed Mixregt (solid line) for the data augmented by the outliers (stars). As shown in Figure 1, the MLE for one of the components fit the line through the outliers and the MLE for the other component fit the line using the rest of data. In this example, the ten outliers had a significant impact on the fitted regression lines by MLE. In addition, note that

the proposed Mixregt well recovered the two regression lines and thus was robust to the added outliers. Additionally, TLE, MEM-Bisquare, and Mixregt-trim all provided similar results to Mixregt.

Similar to Hennig (2004), in order to see how large the outliers can lead to the breakdown of at least one component estimate, we further applied Mixregt by adding ten identical outliers $(1.5, a)$ to the original data set using different a values. We found that Mixregt still worked well when $a = 4700$ but failed when $a = 4800$. However, such extreme outliers can usually easily be deleted.

6 Discussion

In this article, we proposed a new robust estimation method for mixture of regression based on the $t-$distribution. In order to make the new method work against high leverage outliers, we further proposed a trimmed version of the proposed method by fitting the new model to the data after adaptively trimming high leverage points. The simulation study demonstrated the effectiveness of the proposed new method.

In the trimmed version of the new method, we use the same weights as Pison et al. (2002), i.e, delete the high leverage points based on the cut point $\chi^2_{p-1,0.975}$. However, some high leverage points might have small residuals and thus can also provide valuable information to regression parameters. More research is needed on how to incorporate information from data with high leverage points and small residuals. One possible way is to borrow the ideas from GM-estimators (Krasker and Welsch, 1982; Maronna and Yohai, 1981) and one-step GM-estimators (Coakley and Hettmansperger, 1993; Simpson and Yohai, 1998).

It is also interesting to investigate the sample breakdown points for the proposed method and some of the other robust mixture regression models. However, we should

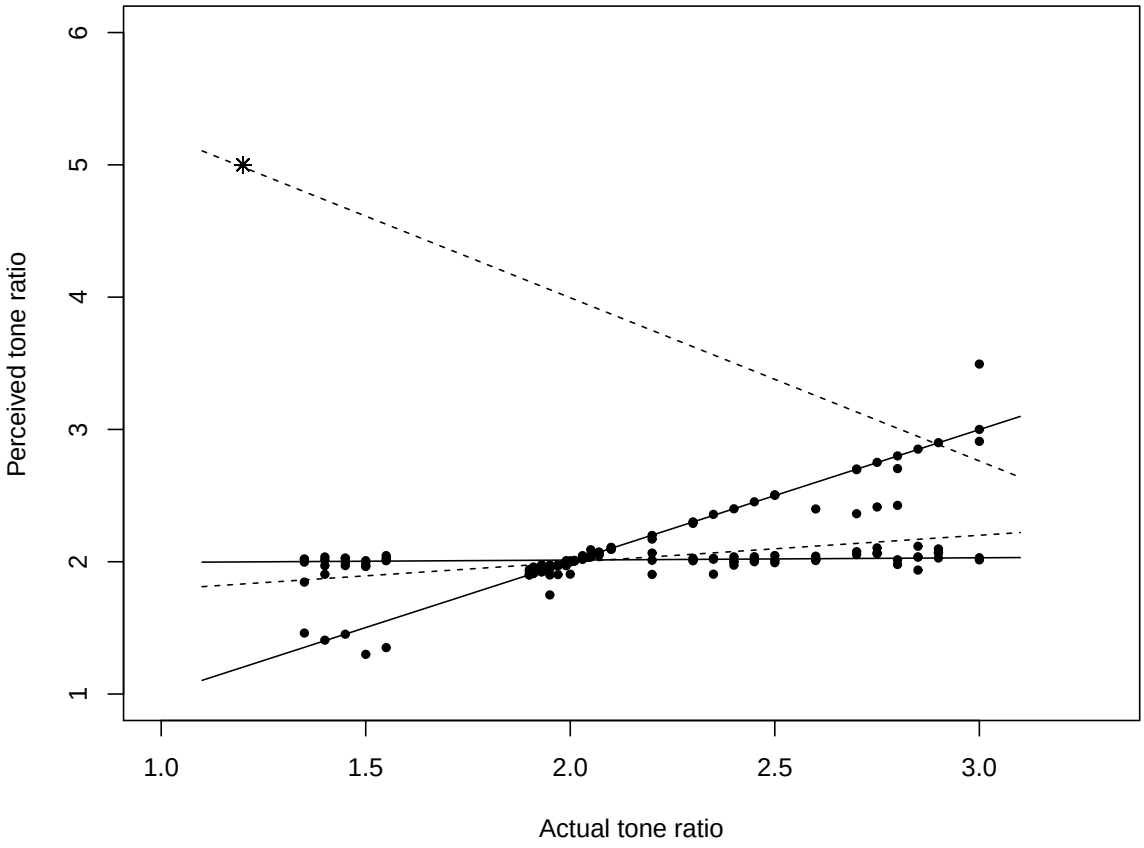

Figure 1: The scatter plot of the tone perception data and the fitted mixture regression lines with ten added identical outliers $(1.5, 5)$ (denoted by stars at the upper left corner). The predictor is the actual tone ratio and the response is the perceived tone ratio by a trained musician. The solid lines represent the fit by the proposed Mixregt and the dashed lines represent the fit by the traditional MLE.

note that the analysis of breakdown point for traditional linear regression cannot be directly applied to mixture regression. For example, the breakdown point of TLE for traditional linear regression does not apply to the mixture regression. García-Escudero et al. (2010) also stated that the traditional definition of breakdown point is not the right one to quantify the robustness of clustering regression procedures to outliers, since the robustness of these procedures is not only data dependent but also cluster dependent.

Hennig (2004) provided a new definition of breakdown points for mixture model based on the breakdown of at least one of the mixture components. Based on this new definition, the mixture of $t-$distributions has a very small breakdown point (Hennig, 2004). However, Hennig (2004) mentioned that only very extreme outliers could lead to the breakdown of mixture of $t-$distributions, especially when the degrees of freedom were small. Therefore, we believe the $t-$distribution can still be used as a robust estimation method for mixture models with the exception of extreme outliers.

Note that model (1.2) assumes that component proportions π_js are constant and do not depend on $\mathbf{x}$. This might be unrealistic in some situations. The ideas that allow the proportions to depend on the covariates in a mixture model can be found in literature, e.g., the hierarchical mixtures of experts model (Jordan and Jacobs, 1994) in machine learning. Young and Hunter (2010) used kernel regression to model covariate-dependent proportions for mixture of linear regression models. Huang and Yao (2012) proposed a semiparametric mixture regression model by allowing π_j to depend on $\mathbf{x}$ nonparametrically. It will be interesting to understand how to apply the proposed robust method based on the $t-$distribution to the above models when the proportions also depend on $\mathbf{x}$. This will be the topic of our future research.

Acknowledgements

The authors are grateful to the editors and the referees for their insightful comments and suggestions, which greatly improved this article.

Table 1: MSE (Bias) of point estimates for $n = 200$ in Example 1

TRUE	MLE	TLE	MEM-bisquare	Mixregt	Mixregt-trim
Case I: $\epsilon \sim N(0,1)$					
$\beta_{10} : 0$	0.046 (0.001)	0.305 (0.033)	0.066 (0.008)	0.046 (0.003)	0.048 (0.008)
$\beta_{20} : 0$	0.010 (0.014)	0.069 (0.015)	0.010 (0.012)	0.010 (0.014)	0.010 (0.015)
$\beta_{11} : 1$	0.032 (0.013)	0.938 (0.618)	0.052 (0.006)	0.032 (0.011)	0.040 (0.001)
$\beta_{21} : -1$	0.009 (0.001)	0.018 (0.013)	0.010 (0.001)	0.009 (0.001)	0.011 (0.002)
$\beta_{12} : 1$	0.042 (0.007)	0.910 (0.648)	0.087 (0.030)	0.041 (0.006)	0.050 (0.012)
$\beta_{22} : -1$	0.009 (0.000)	0.015 (0.005)	0.010 (0.000)	0.009 (0.000)	0.011 (0.002)
$\pi_1 : 0.25$	0.002 (0.004)	0.009 (0.049)	0.002 (0.007)	0.002 (0.004)	0.002 (0.006)
Case II: $\epsilon \sim t_3$					
$\beta_{10} : 0$	38.42 (0.205)	0.253 (0.021)	0.205 (0.033)	0.141 (0.014)	0.153 (0.020)
$\beta_{20} : 0$	16.73 (0.117)	0.029 (0.010)	0.148 (0.020)	0.015 (0.002)	0.106 (0.008)
$\beta_{11} : 1$	12.59 (0.148)	0.380 (0.331)	0.217 (0.095)	0.151 (0.064)	0.169 (0.081)
$\beta_{21} : -1$	5.235 (0.365)	0.022 (0.015)	0.032 (0.029)	0.014 (0.012)	0.052 (0.035)
$\beta_{12} : 1$	19.57 (0.576)	0.350 (0.282)	0.200 (0.048)	0.143 (0.035)	0.189 (0.071)
$\beta_{22} : -1$	5.236 (0.278)	0.023 (0.017)	0.149 (0.054)	0.015 (0.008)	0.020 (0.010)
$\pi_1 : 0.25$	0.098 (0.076)	0.007 (0.041)	0.012 (0.042)	0.003 (0.008)	0.008 (0.017)
Case III: $\epsilon \sim t_1$					
$\beta_{10} : 0$	4.7e+4 (8.158)	3.242 (0.082)	0.985 (0.006)	0.305 (0.025)	0.429 (0.016)
$\beta_{20} : 0$	4.2e+6 (147.0)	4.871 (0.070)	0.083 (0.017)	0.061 (0.013)	0.072 (0.012)
$\beta_{11} : 1$	2.2e+4 (38.27)	3.850 (0.018)	0.764 (0.125)	0.691 (0.343)	1.025 (0.402)
$\beta_{21} : -1$	3.6e+6 (241.3)	1.770 (0.182)	0.085 (0.001)	0.053 (0.069)	0.059 (0.012)
$\beta_{12} : 1$	2.7e+4 (35.81)	2.301 (0.448)	0.669 (0.207)	0.634 (0.353)	0.837 (0.398)
$\beta_{22} : -1$	1.7e+5 (44.15)	1.429 (0.189)	0.193 (0.076)	0.056 (0.095)	0.154 (0.038)
$\pi_1 : 0.25$	0.305 (0.272)	0.084 (0.106)	0.025 (0.103)	0.019 (0.068)	0.022 (0.080)
Case IV: $\epsilon \sim 0.95N(0,1) + 0.05N(0,5^2)$					
$\beta_{10} : 0$	5.372(0.020)	0.183(0.024)	0.056(0.008)	0.057(0.013)	0.065(0.015)
$\beta_{20} : 0$	7.378(0.235)	0.039(0.000)	0.014(0.010)	0.011(0.010)	0.011(0.008)
$\beta_{11} : 1$	3.979(0.096)	0.470(0.382)	0.126(0.036)	0.057(0.002)	0.078(0.009)
$\beta_{21} : -1$	1.763(0.131)	0.016(0.007)	0.013(0.016)	0.013(0.013)	0.014(0.010)
$\beta_{12} : 1$	4.217(0.138)	0.568(0.415)	0.117(0.044)	0.063(0.008)	0.081(0.018)
$\beta_{22} : -1$	2.300(0.244)	0.017(0.003)	0.013(0.012)	0.013(0.001)	0.015(0.007)
$\pi_1 : 0.25$	0.088(0.067)	0.006(0.032)	0.006(0.028)	0.003(0.006)	0.003(0.008)
Case V: $\epsilon \sim N(0,1)$ with 5% of high leverage outliers					
$\beta_{10} : 0$	2.099 (0.059)	0.163 (0.054)	0.508 (0.092)	1.508 (0.240)	0.016 (0.015)
$\beta_{20} : 0$	0.014 (0.000)	0.022 (0.007)	0.010 (0.001)	0.034 (0.013)	0.010 (0.001)
$\beta_{11} : 1$	3.443 (1.534)	0.487 (0.129)	1.152 (0.532)	3.055 (1.561)	0.054 (0.008)
$\beta_{21} : -1$	0.076 (0.235)	0.063 (0.020)	0.011 (0.023)	0.089 (0.138)	0.010 (0.003)
$\beta_{12} : 1$	3.233 (1.459)	0.426 (0.139)	0.747 (0.364)	2.663 (1.425)	0.042 (0.004)
$\beta_{22} : -1$	0.070 (0.227)	0.086 (0.021)	0.012 (0.018)	0.082 (0.132)	0.011 (0.015)
$\pi_1 : 0.25$	0.009 (0.092)	0.004 (0.010)	0.004 (0.015)	0.007 (0.080)	0.003 (0.005)

Table 2: MSE (Bias) of point estimates for $n = 400$ in Example 1

TRUE	MLE	TLE	MEM-bisquare	Mixregt	Mixregt-trim
Case I: $\epsilon \sim N(0,1)$					
$\beta_{10} : 0$	0.020 (0.003)	0.144 (0.037)	0.021 (0.003)	0.020 (0.003)	0.023 (0.008)
$\beta_{20} : 0$	0.004 (0.000)	0.037 (0.027)	0.004 (0.001)	0.004 (0.001)	0.004 (0.004)
$\beta_{11} : 1$	0.021 (0.006)	0.579 (0.455)	0.023 (0.009)	0.021 (0.005)	0.019 (0.003)
$\beta_{21} : -1$	0.004 (0.003)	0.012 (0.014)	0.004 (0.003)	0.004 (0.003)	0.005 (0.003)
$\beta_{12} : 1$	0.017 (0.002)	0.625 (0.471)	0.019 (0.000)	0.017 (0.002)	0.025 (0.001)
$\beta_{22} : -1$	0.004 (0.005)	0.011 (0.003)	0.004 (0.008)	0.004 (0.005)	0.005 (0.002)
$\pi_1 : 0.25$	0.001 (0.004)	0.009 (0.028)	0.001 (0.006)	0.001 (0.004)	0.001 (0.000)
Case II: $\epsilon \sim t_3$					
$\beta_{10} : 0$	22.41 (0.078)	0.092 (0.030)	0.044 (0.008)	0.040 (0.007)	0.042 (0.006)
$\beta_{20} : 0$	12.13 (0.012)	0.011 (0.003)	0.008 (0.000)	0.006 (0.001)	0.006 (0.000)
$\beta_{11} : 1$	16.13 (0.482)	0.107 (0.162)	0.039 (0.024)	0.035 (0.005)	0.037 (0.003)
$\beta_{21} : -1$	21.65 (0.638)	0.007 (0.008)	0.007 (0.026)	0.006 (0.006)	0.007 (0.004)
$\beta_{12} : 1$	23.00 (0.245)	0.094 (0.181)	0.040 (0.022)	0.038 (0.007)	0.039 (0.005)
$\beta_{22} : -1$	11.33 (0.467)	0.007 (0.004)	0.008 (0.028)	0.006 (0.007)	0.007 (0.008)
$\pi_1 : 0.25$	0.087 (0.059)	0.002 (0.021)	0.002 (0.021)	0.001 (0.001)	0.002 (0.001)
Case III: $\epsilon \sim t_1$					
$\beta_{10} : 0$	5.2e+6 (210)	2.515 (0.079)	0.205 (0.002)	0.017 (0.012)	0.124 (0.030)
$\beta_{20} : 0$	9.1e+5 (71.5)	1.919 (0.131)	0.063 (0.013)	0.010 (0.002)	0.025 (0.006)
$\beta_{11} : 1$	1.2e+7 (330)	0.951 (0.157)	0.417 (0.202)	0.255 (0.013)	0.313 (0.171)
$\beta_{21} : -1$	9.4e+5 (184)	0.634 (0.047)	0.118 (0.068)	0.009 (0.016)	0.037 (0.017)
$\beta_{12} : 1$	1.8e+6 (109)	1.318 (0.083)	0.418 (0.134)	0.198 (0.032)	0.233 (0.171)
$\beta_{22} : -1$	2.11e+5 (74)	0.667 (0.064)	0.085 (0.059)	0.008 (0.004)	0.025 (0.010)
$\pi_1 : 0.25$	0.303 (0.253)	0.049 (0.054)	0.025 (0.107)	0.008 (0.014)	0.010 (0.033)
Case IV: $\epsilon \sim 0.95N(0,1) + 0.05N(0,5^2)$					
$\beta_{10} : 0$	3.509(0.178)	0.117(0.058)	0.023(0.016)	0.025(0.012)	0.030(0.004)
$\beta_{20} : 0$	4.298(0.194)	0.021(0.013)	0.005(0.000)	0.005(0.001)	0.005(0.006)
$\beta_{11} : 1$	2.057(0.137)	0.340(0.307)	0.025(0.026)	0.027(0.020)	0.033(0.008)
$\beta_{21} : -1$	2.889(0.341)	0.007(0.011)	0.004(0.002)	0.005(0.011)	0.007(0.012)
$\beta_{12} : 1$	2.436(0.122)	0.303(0.301)	0.019(0.013)	0.021(0.011)	0.032(0.012)
$\beta_{22} : -1$	2.422(0.134)	0.007(0.011)	0.005(0.004)	0.005(0.007)	0.005(0.002)
$\pi_1 : 0.25$	0.059(0.030)	0.004(0.011)	0.001(0.009)	0.001(0.007)	0.001(0.004)
Case V: $\epsilon \sim N(0,1)$ with 5% of high leverage outliers					
$\beta_{10} : 0$	1.708 (0.129)	0.116 (0.029)	0.264 (0.040)	1.141 (0.203)	0.020 (0.007)
$\beta_{20} : 0$	0.008 (0.013)	0.035 (0.015)	0.005 (0.007)	0.005 (0.011)	0.005 (0.005)
$\beta_{11} : 1$	2.814 (1.473)	0.195 (0.016)	0.600 (0.333)	2.714 (1.498)	0.020 (0.008)
$\beta_{21} : -1$	0.074 (0.252)	0.078 (0.033)	0.007 (0.028)	0.024 (0.135)	0.005 (0.002)
$\beta_{12} : 1$	2.940 (1.516)	0.276 (0.005)	0.672 (0.341)	2.691 (1.490)	0.024 (0.015)
$\beta_{22} : -1$	0.073 (0.251)	0.052 (0.018)	0.006 (0.021)	0.021 (0.128)	0.004 (0.003)
$\pi_1 : 0.25$	0.009 (0.095)	0.002 (0.003)	0.002 (0.016)	0.008 (0.087)	0.001 (0.001)

Table 3: The mean (median) of estimated degrees of freedom by Mixregt and Mixregt-trim based on the grid points from $[1, 15]$ for Example 1

Case	n	Mixregt	Mixregt-trim
I: $\epsilon \sim N(0,1)$	200	14.5 (15)	14.4 (15)
	400	14.7 (15)	14.8 (15)
II: $\epsilon \sim t_3$	200	3.33 (3)	3.39 (3)
	400	3.18 (3)	3.18 (3)
III: $\epsilon \sim t_1$	200	1 (1)	1 (1)
	400	1 (1)	1 (1)
IV: $\epsilon \sim 0.95N(0,1) + 0.05N(0,5^2)$	200	3.52(3)	3.45 (3)
	400	3.91(3)	3.92 (3)
V: $\epsilon \sim N(0,1)$ with 5% high leverage outliers	200	4.62 (4)	13.8 (15)
	400	4.26 (4)	14.7 (15)

Table 4: MSE (Bias) of point estimates for $n = 200$ in Example 2

TRUE	MLE	TLE	MEM-bisquare	Mixregt	Mixregt-trim
Case I: $\epsilon_1 \sim N(0,1), \epsilon_2 \sim N(0,1)$, and $\epsilon_3 \sim N(0,1)$					
$\beta_{10} : 1$	0.075(0.025)	0.113(0.068)	0.063(0.036)	0.068(0.043)	0.071(0.044)
$\beta_{20} : 2$	0.189(0.089)	0.211(0.222)	0.145(0.073)	0.149(0.071)	0.200(0.113)
$\beta_{30} : 3$	0.021(0.006)	0.057(0.006)	0.022(0.004)	0.021(0.005)	0.027(0.012)
$\beta_{11} : 1$	0.086(0.034)	0.205(0.306)	0.062(0.010)	0.060(0.029)	0.073(0.046)
$\beta_{21} : 2$	0.186(0.078)	0.511(0.013)	0.171(0.065)	0.150(0.036)	0.191(0.066)
$\beta_{31} : 5$	0.020(0.032)	0.047(0.030)	0.018(0.028)	0.020(0.029)	0.023(0.027)
$\pi_1 : 0.3$	0.009(0.015)	0.006(0.037)	0.008(0.011)	0.008(0.011)	0.009(0.017)
$\pi_2 : 0.3$	0.008(0.002)	0.004(0.008)	0.007(0.001)	0.006(0.003)	0.008(0.002)
Case II: $\epsilon_1 \sim t_9$, $\epsilon_2 \sim t_6$, and $\epsilon_3 \sim t_3$					
$\beta_{10} : 1$	25.31(0.589)	0.155(0.126)	0.175(0.072)	0.123(0.023)	0.143(0.016)
$\beta_{20} : 2$	7.065(0.832)	0.290(0.273)	0.276(0.060)	0.201(0.020)	0.238(0.007)
$\beta_{30} : 3$	13.88(0.835)	0.066(0.032)	0.034(0.042)	0.033(0.044)	0.034(0.047)
$\beta_{11} : 1$	15.09(0.164)	0.183(0.256)	0.086(0.032)	0.075(0.035)	0.108(0.035)
$\beta_{21} : 2$	5.927(0.869)	0.456(0.103)	0.299(0.136)	0.311(0.161)	0.310(0.172)
$\beta_{31} : 5$	12.82(1.469)	0.051(0.021)	0.029(0.042)	0.029(0.046)	0.039(0.065)
$\pi_1 : 0.3$	0.056(0.106)	0.009(0.043)	0.010(0.020)	0.012(0.016)	0.014(0.017)
$\pi_2 : 0.3$	0.029(0.042)	0.006(0.015)	0.010(0.002)	0.012(0.006)	0.014(0.007)
Case III: $\epsilon_1 \sim N(0,1)$, $\epsilon_2 \sim N(0,1)$, and $\epsilon_3 \sim t_3$					
$\beta_{10} : 1$	5.111(0.045)	0.127(0.112)	0.094(0.008)	0.091(0.061)	0.095(0.057)
$\beta_{20} : 2$	8.283(0.928)	0.219(0.241)	0.253(0.069)	0.205(0.037)	0.221(0.068)
$\beta_{30} : 3$	9.160(0.442)	0.044(0.012)	0.083(0.037)	0.081(0.037)	0.081(0.044)
$\beta_{11} : 1$	3.508(0.165)	0.172(0.257)	0.065(0.035)	0.062(0.058)	0.070(0.070)
$\beta_{21} : 2$	5.687(1.084)	0.202(0.053)	0.347(0.167)	0.330(0.193)	0.405(0.236)
$\beta_{31} : 5$	11.32(1.492)	0.046(0.028)	0.050(0.062)	0.053(0.072)	0.065(0.085)
$\pi_1 : 0.3$	0.064(0.146)	0.007(0.043)	0.011(0.029)	0.012(0.037)	0.014(0.045)
$\pi_2 : 0.3$	0.029(0.020)	0.005(0.022)	0.008(0.003)	0.010(0.001)	0.009(0.004)
Case IV: $\epsilon_1, \epsilon_2, \epsilon_3, \sim N(0,1)$ with 5% of high leverage outliers					
$\beta_{10} : 1$	0.240(0.467)	0.117(0.111)	0.088(0.005)	0.128(0.125)	0.143(0.094)
$\beta_{20} : 2$	0.917(0.936)	0.224(0.216)	0.180(0.027)	0.380(0.351)	0.218(0.132)
$\beta_{30} : 3$	16.39(2.228)	0.039(0.020)	0.022(0.017)	3.231(0.562)	0.025(0.014)
$\beta_{11} : 1$	0.242(0.495)	0.126(0.188)	0.069(0.032)	0.121(0.180)	0.113(0.097)
$\beta_{21} : 2$	8.576(2.907)	0.261(0.007)	0.245(0.080)	3.005(1.037)	0.217(0.017)
$\beta_{31} : 5$	24.41(4.913)	0.030(0.012)	0.022(0.001)	8.058(1.643)	0.026(0.009)
$\pi_1 : 0.3$	0.060(0.236)	0.006(0.018)	0.010(0.006)	0.023(0.079)	0.017(0.013)
$\pi_2 : 0.3$	0.008(0.078)	0.006(0.004)	0.010(0.001)	0.009(0.039)	0.018(0.008)

Table 5: MSE (Bias) of point estimates for $n = 400$ in Example 2

TRUE	MLE	TLE	MEM-bisquare	Mixregt	Mixregt-trim
Case I: $\epsilon_1 \sim N(0,1), \epsilon_2 \sim N(0,1)$, and $\epsilon_3 \sim N(0,1)$					
$\beta_{10} : 1$	0.042(0.030)	0.077(0.117)	0.045(0.019)	0.039(0.009)	0.049(0.027)
$\beta_{20} : 2$	0.066(0.051)	0.108(0.214)	0.088(0.087)	0.078(0.075)	0.098(0.077)
$\beta_{30} : 3$	0.129(0.042)	0.032(0.006)	0.014(0.007)	0.013(0.008)	0.012(0.005)
$\beta_{11} : 1$	0.053(0.002)	0.136(0.267)	0.040(0.022)	0.039(0.023)	0.048(0.021)
$\beta_{21} : 2$	0.123(0.001)	0.443(0.057)	0.261(0.074)	0.244(0.072)	0.102(0.014)
$\beta_{31} : 5$	0.061(0.034)	0.021(0.043)	0.015(0.020)	0.011(0.016)	0.011(0.013)
$\pi_1 : 0.3$	0.006(0.005)	0.004(0.021)	0.005(0.015)	0.005(0.014)	0.005(0.011)
$\pi_2 : 0.3$	0.004(0.007)	0.003(0.007)	0.004(0.010)	0.003(0.011)	0.005(0.013)
Case II: $\epsilon_1 \sim t_9$, $\epsilon_2 \sim t_6$, and $\epsilon_3 \sim t_3$					
$\beta_{10} : 1$	7.735(0.157)	0.094(0.116)	0.108(0.110)	0.082(0.045)	0.063(0.028)
$\beta_{20} : 2$	3.897(0.431)	0.163(0.234)	0.150(0.117)	0.093(0.027)	0.111(0.016)
$\beta_{30} : 3$	3.772(0.270)	0.024(0.022)	0.020(0.014)	0.019(0.009)	0.021(0.008)
$\beta_{11} : 1$	6.219(0.031)	0.124(0.233)	0.060(0.043)	0.050(0.018)	0.056(0.006)
$\beta_{21} : 2$	2.077(0.251)	0.077(0.091)	0.146(0.015)	0.140(0.037)	0.141(0.049)
$\beta_{31} : 5$	3.055(0.460)	0.020(0.020)	0.015(0.027)	0.016(0.027)	0.017(0.028)
$\pi_1 : 0.3$	0.032(0.022)	0.004(0.026)	0.006(0.016)	0.006(0.001)	0.008(0.003)
$\pi_2 : 0.3$	0.025(0.056)	0.004(0.018)	0.006(0.001)	0.007(0.017)	0.008(0.016)
Case III: $\epsilon_1 \sim N(0,1)$, $\epsilon_2 \sim N(0,1)$, and $\epsilon_3 \sim t_3$					
$\beta_{10} : 1$	25.00(0.632)	0.062(0.071)	0.047(0.016)	0.049(0.015)	0.056(0.007)
$\beta_{20} : 2$	3.977(0.495)	0.125(0.214)	0.064(0.041)	0.048(0.009)	0.070(0.033)
$\beta_{30} : 3$	56.16(0.722)	0.022(0.014)	0.015(0.020)	0.015(0.022)	0.014(0.022)
$\beta_{11} : 1$	5.088(0.034)	0.123(0.232)	0.030(0.032)	0.029(0.039)	0.044(0.046)
$\beta_{21} : 2$	2.322(0.315)	0.077(0.107)	0.063(0.032)	0.063(0.047)	0.081(0.047)
$\beta_{31} : 5$	57.05(1.247)	0.017(0.026)	0.014(0.024)	0.015(0.028)	0.020(0.035)
$\pi_1 : 0.3$	0.031(0.062)	0.004(0.030)	0.005(0.014)	0.006(0.010)	0.007(0.013)
$\pi_2 : 0.3$	0.019(0.016)	0.003(0.023)	0.005(0.003)	0.006(0.013)	0.007(0.013)
Case IV: $\epsilon_1, \epsilon_2, \epsilon_3, \sim N(0,1)$ with 5% of high leverage outliers					
$\beta_{10} : 1$	0.224(0.459)	0.071(0.097)	0.044(0.040)	0.096(0.114)	0.146(0.096)
$\beta_{20} : 2$	0.928(0.989)	0.137(0.207)	0.058(0.049)	0.342(0.320)	0.129(0.052)
$\beta_{30} : 3$	12.57(2.632)	0.015(0.012)	0.008(0.009)	1.828(0.461)	0.009(0.004)
$\beta_{11} : 1$	0.226(0.467)	0.101(0.212)	0.025(0.014)	0.097(0.208)	0.108(0.088)
$\beta_{21} : 2$	8.583(2.928)	0.059(0.068)	0.042(0.026)	2.404(0.807)	0.092(0.015)
$\beta_{31} : 5$	24.83(4.981)	0.015(0.023)	0.008(0.015)	6.451(1.414)	0.011(0.011)
$\pi_1 : 0.3$	0.058(0.247)	0.003(0.025)	0.006(0.001)	0.018(0.070)	0.018(0.002)
$\pi_2 : 0.3$	0.006(0.071)	0.003(0.008)	0.006(0.003)	0.005(0.020)	0.018(0.004)

References

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

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

Böhning, D. (1999). *Computer-Assisted Analysis of Mixtures and Applications.* Boca Raton, FL: Chapman and Hall/CRC.

Celeux, G., Hurn, M., and Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95, 957-970.

Chen, J., Tan, X., and Zhang, R. (2008). Inference for normal mixture in mean and variance. *Statistica Sincia*, 18, 443-465.

Coakley, C. W. and Hettmansperger, T. P. (1993). A bounded influence, high breakdown, efficient regression estimator. *Journal of the American Statistical Association*, 88, 872-880.

Davies, L. (1987). Asymptotic behavior of S-estimators of multivariate location parameters and dispersion matrices. *Annals of Statistics*, 15, 1269-1292.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of Royal Statistical Society, Ser B.*, 39, 1-38.

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

Donoho, D. L. and Gasko, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Annals of Statistics*, 20, 1803-1827.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer.

García-Escudero, L. A., Gordaliza, A., Mayo-Iscara, A., and San Martín, R. (2010). Robust clusterwise linear regression through trimming. *Computational Statistics & Data Analysis*, 54, 3057-3069.

García-Escudero, L. A., Gordaliza, A., San Martín, R., Van Aelst, S., and Zamar, R. (2009). Robust linear clustering. *Journal of The Royal Statistical Society Series*, B71, 301-318.

Goldfeld, S. M. and Quandt, R. E. (1973). A Markov model for switching regression. *Journal of Econometrics*, 1, 3-15.

Hathaway, R. J. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *Annals of Statistics*, 13, 795-800.

Hathaway, R. J. (1986). A constrained EM algorithm for univariate mixtures. *Journal of Statistical Computation and Simulation*, 23, 211-230.

Hennig, C. (2000). Identifiability of models for clusterwise linear regression. *Journal of Classification*. 17, 273-296.

Hennig, C. (2002). Fixed point clusters for linear regression: computation and comparison. *Journal of Classification*, 19, 249-276

Hennig, C. (2003). Clusters, outliers, and regression: Fixed point clusters. *Journal of Multivariate Analysis*, 86, 183-212.

Hennig, C. (2004). Breakdown points for maximum likelihood-estimators of location-scale mixtures. *Annals of Statistics*, 32, 1313-1340.

Huang, M. and Yao, W. (2012). Mixture of regression models with varying mixing proportions: A semiparametric approach. *Journal of the American Statistical Association,* 107, 711-724.

Jiang, W. and Tanner, M. A. (1999). Hierarchical mixtures-of-experts for exponential family regression models: Approximation and maximum likelihood estimation. *The Annals of Statistics*, 27, 987-1011.

Jordan, M. I. and Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*. 6, 181C214.

Kiefer, N. M. (1978). Discrete parameter variation: efficient estimation of a switching regression model. *Econometrica*, 46, 427-434.

Krasker, W. S. and Welsch, R. E. (1982). Efficient bounded influence regression estimation. *Journal of the American Statistical Association*, 77, 595-604.

Lindsay, B. G. and Basak, P. (1993). Multivariate normal mixtures: a fast consistent method of moments. *Journal of American Statistical Association*, 88, 468-475.

Liu, R. Y., Parelius, J. M. and Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *Annals of Statistics*, 27, 783-840.

MacQueen, J. (1967). Some methods for classfication and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 281-297.

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

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. Wiley, New York.

Maronna, R. A. and Yohai, V. J. (1981). Asymptotic behavior of general M-estimators for regression and scale with random carriers. *Probability Theory and Related Fields*, 58, 7-20.

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

Mueller, C. H. and Garlipp, T. (2005). Simple consistent cluster methods based on redescending M-estimators with an application to edge identification in images. *Journal of Multivariate Analysis* 92, 359-385.

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.

Peel, D. and McLachlan, G. J. (2000). Robust mixture modelling using the t distribution *Statistics and Computing*, 10, 339-348.

Peters, B. C. and Walker, H. F. (1978). An iterative procedure for obtaining maximum likelihood estimators of the parameters for a mixture of normal distributions. *SIAM Journal on Applied Mathematics, 35, 362-378*.

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

Rosseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79, 871-880.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. Wiley-Interscience, New York.

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

Rosseeuw, P. J. and van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverate points. *Journal of the American Statistical Association*, 85, 633-639.

Shen, H., Yang, J., and Wang, S. (2004). Outlier detecting in fuzzy switching regression models. *Artificial Intelligence: Methodology, Systems, and Applications Lecture Notes in Computer Science*, 2004, Vol. 3192/2004, 208-215.

Simpson, D. G. and Yohai, V. J. (1998). Functional stability of one-step estimators in approximately linear regression. *Annals of Statistics*, 26, 1147-1169.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models.* Boca Raton. Chapman and Hall/CRC.

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

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of Royal Statistical Society, Ser B.*, 62, 795-809.

Wedel, M. and Kamakura, W. A. (2000). *Market Segmentation: Conceptual and Methodological Foundations.* 2nd edition, Norwell, MA: Kluwer Academic Publishers. Journal of Classification. Springer, New York.

Yao, W. (2010). A profile likelihood method for normal mixture with unequal variance. *Journal of Statistical Planning and Inference*, 140, 2089-2098.

Yao, W. (2012). Model based labeling for mixture models. *Statistics and Computing*, 22, 337-347.

Yao, W. and Lindsay, B. G. (2009). Bayesian mixture labeling by highest posterior density. *Journal of American Statistical Association*, 104, 758-767.

Young, D. S. and Hunter, D. R. (2010). Mixtures of regressions with predictor-dependent mixing proportions. *Computational Statistics and Data Analysis*. 54, 2253-2266.

Zuo, Y., Cui, H. and He, X. (2004). On the Stahel-Donoho estimator and depth-weighted means of multivariate data. *Annals of Statistics*, 32, 167-188.

Zuo, Y. and Serfling, R. (2000). General notions of statistical depth function. *Annals of Statistics*, 28, 461-482.