Hacettepe Journal of Mathematics and Statistics
Volume 45 (2) (2016), 549 – 559

Robust variable selection for mixture linear regression models

Yunlu Jiang[*][†]

Abstract

In this paper, we propose a robust variable selection to estimate and select relevant covariates for the finite mixture of linear regression models by assuming that the error terms follow a Laplace distribution to the data after trimming the high leverage points. We introduce a revised Expectation-maximization (EM) algorithm for numerical computation. Simulation studies indicate that the proposed method is robust to both the high leverage points and outliers in the y-direction, and can obtain a consistent variable selection in the case of outliers or heavy-tail error distribution. Finally, we apply the proposed methodology to analyze a real data.

Keywords: Finite mixture of linear regression models, Robustness, EM-algorithm.

2000 AMS Classification: 62G35, 62G35, 62H12.

Received : 03.11.2014 *Accepted* : 02.03.2015 *Doi* : 10.15672/HJMS.2015549560

1. Introduction

Finite mixture of linear regression (FMLR) models provide a very important statistical tool to fit the unobserved heterogeneous relationships. They are extensively used in many research fields, e.g., marketing and social sciences [29, 25], machine learning [12, 13]. A comprehensive review of finite mixture models was given in [20]. It is well-known that the traditional maximum likelihood estimator (MLE) for mixture linear regression models works well when the error term follows a normal distribution. However, the normality based MLE is not robust to outliers in the datasets.

Many robust methodologies were proposed and widely studied for mixture linear regression models in the literature. For instance, [18] and [24] introduced the weighted MLE. [21] proposed the trimmed likelihood estimator. [1] proposed a modified Expectation-maximization (EM)-algorithm by replacing the least squares criterion with a robust

[*] Department of Statistics, College of Economics, Jinan University, Guangzhou, 510632, China, Email: `tjiangyl@jnu.edu.cn`

[†] Corresponding Author.

criteria in the M step. [30] and [26] proposed a robust estimation procedure based a t-distribution and a Laplace distribution, respectively.

In many practical applications, there are many covariates involved in the FMLR models. Nevertheless, the number of important ones is usually relatively small. In fact, the problem of variable selection in a FMLR model has received much attention recently. For example, [28] used Akaike information criterion (AIC) and Bayesian information criterion (BIC) to study model choice issues for a class of Poisson mixture models. [15] introduced a penalized likelihood approach for variable selection in FMLR models based on some well-known families such as Gaussian, Poisson, and Binomial distributions, and developed an EM algorithm for numerical computations. [17] proposed a mixture regression LASSO (MR-LASSO) method to penalize both regression coefficients and mixture components simultaneously. [14] gave an overview of the new feature selection methods in FMLR models. [16] studied the issue of variable selection in FMLR models when the number of parameters in the model can increase with the sample size. [5] proposed a penalized likelihood approach to simultaneously select important fixed and random effects in the finite mixtures of linear mixed-effects models. It is very important to note that many of those methods are closely related to the traditional MLE method.

To the best of our knowledge, the robust feature selection for FMLR models has not been well studied. In the linear regression models, the least absolute deviation (LAD)estimator is very important when the error terms follow a heavy-tailed distribution, and has the desired robust properties. In fact, the maximum-likelihood estimator of the regression parameters given a Laplace distributed regression errors is LAD estimator. [26] applied the LAD estimator to a class of FMLR models. In this article, we propose a robust variable selection procedure based on the LAD estimator for FMLR models, and introduce a revised EM-type algorithm for numerical computation. Simulation studies show that the proposed method is robust and can obtain a consistent variable selection when there are outliers in the datasets or the error term follows a heavy-tailed distribution. In addition, the proposed robust variable selection approach works comparably to the traditional penalized likelihood-based method when there are no outliers and the error is normal.

The rest of this paper is organized as follows. In Section 2, we propose a robust variable selection for FMLR models, and introduce a revised EM-algorithm for numerical computation. In Section 3, numerical simulations and a real data analysis are conducted to compare the performance of the proposed method with the existing method. We conclude with some remarks in Section 4.

2. Methodology

Let Z be a latent class variable with $P(Z=i|\mathbf{X}=\mathbf{x})=\pi_i, i=1,\cdots,m$, where $\mathbf{x}$ is a q-dimensional vector. Given $Z=i$, suppose that the response Y depends on $\mathbf{X}$ in a linear way

$$Y=\mathbf{X}^T\boldsymbol{\beta}_i+\sigma_i\epsilon_i,$$

where $\boldsymbol{\beta}_i$ is an unknown q-dimensional vectors of regression parameters, σ_i is an unknown positive scalar, and ϵ_i is a random error with density $f_i(\cdot)$ and mean 0, and is independent of $\mathbf{X}$. Then, the density Y given $\mathbf{X}$ is

$$g(y|\mathbf{x},\boldsymbol{\theta})=\sum_{i=1}^{m}\pi_i\frac{1}{\sigma_i}f_i\left(\frac{y-\mathbf{x}^T\boldsymbol{\beta}_i}{\sigma_i}\right), \tag{2.1}$$

where $\boldsymbol{\theta}=(\pi_1,\boldsymbol{\beta}_1,\sigma_1,\cdots,\pi_m,\boldsymbol{\beta}_m,\sigma_m)^T$.

Suppose that $\mathbf{D}_n = \{(\mathbf{X}_1, Y_1), \cdots, (\mathbf{X}_n, Y_n)\}$ are random observations from the model (2.1). The log-likelihood function is

$$\ell_n(\boldsymbol{\theta}) = \sum_{j=1}^{n} \log \left[\sum_{i=1}^{m} \pi_i \frac{1}{\sigma_i} f_i \left(\frac{Y_j - \mathbf{X}_j^T \boldsymbol{\beta}_i}{\sigma_i} \right) \right].$$

The MLE of $\boldsymbol{\theta}$ is obtained by maximizing the log likelihood function $\ell_n(\boldsymbol{\theta})$.

To simultaneously estimate and select relevant covariates, [6] proposed a unified approach via penalized likelihood. A penalized log-likelihood function is defined as follows:

$$\tilde{\ell}_n(\boldsymbol{\theta}) = \ell_n(\boldsymbol{\theta}) - \sum_{i=1}^{m} \pi_i \left\{ \sum_{k=1}^{q} p_{ni}(\beta_{ik}) \right\}, \tag{2.2}$$

where $p_{ni}(\beta_{ik})$ is nonnegative and nondecreasing functions in $|\beta_{ik}|$. Although there are many methods to deal with the problem of feature selection in finite mixture of linear regression models in the literature, many of those methods are closely related to the least squares method. It is well-known that the least squares estimator is very sensitive to the outliers in the dataset. Next, we will study the robust variable selection for finite mixture of regression models. Similar to the idea proposed by [26], we consider the density function f_i of error term follows a Laplace density function with mean 0 and scale parameter $1/\sqrt{2}$. Then, (2.2) can be written as

$$\hat{\ell}_n(\boldsymbol{\theta}) = \sum_{j=1}^{n} \log \left[\sum_{i=1}^{m} \frac{\pi_i}{\sqrt{2}\sigma_i} \exp\left(-\frac{\sqrt{2}|Y_j - \mathbf{X}_j^T \boldsymbol{\beta}_i|}{\sigma_i} \right) \right] - \sum_{i=1}^{m} \pi_i \left\{ \sum_{k=1}^{q} p_{ni}(\beta_{ik}) \right\}. \tag{2.3}$$

[26] pointed out that the EM algorithm based on the Laplace distribution is robust against outliers along the y-direction, but not for the high leverage points. Therefore, in order to obtain a robust variable selection for both the high leverage points and outliers in the y-direction, we consider a trimmed version of the new method by fitting the new model to the data after trimming the high leverage points. Let $\mathbf{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_n)^T$. For each covariate $\mathbf{X}_j$, we first compute a robust Mahalanobis distance

$$MD_j = (\mathbf{X}_j - m(\mathbf{X}))C(\mathbf{X})^T(\mathbf{X}_j - m(\mathbf{X})),$$

where $m(\mathbf{X})$ and $C(\mathbf{X})$ are robust estimates of location and scatter for $\mathbf{X}$, respectively.

In the literature, there are many robust location and scatter estimators. Those estimators include M-estimator [19], Stahel-Donoho (SD) estimators [27, 4], minimum volume ellipsoid (MVE) [22], S-estimators [3], and minimum covariance determinant (MCD) estimators [2]. Due to the availability of fast MCD algorithm [23], we employ MCD estimators to calculate a robust Mahalanobis distance in this paper. Denote

$$\omega_j = \begin{cases} 1, & \text{if } MD_j \le \chi^2_{q,0.975}, \\ 0, & \text{otherwise.} \end{cases}$$

With such a weight function, the high leverage points are discarded. Then, by taking an adaptive LASSO for the penalty function, the proposed robust variable selection estimator is defined by maximizing the following objective function

$$\bar{\ell}_n(\boldsymbol{\theta}) = \sum_{j=1}^{n} \omega_j \log \left[\sum_{i=1}^{m} \frac{\pi_i}{\sqrt{2}\sigma_i} \exp\left(-\frac{\sqrt{2}|Y_j - \mathbf{X}_j^T \boldsymbol{\beta}_i|}{\sigma_i} \right) \right] - \sum_{i=1}^{m} \pi_i \left\{ \sum_{k=1}^{q} \lambda_{ik} \frac{|\beta_{ik}|}{|\hat{\beta}_{ik}|} \right\}, \tag{2.4}$$

where $\hat{\beta}_{ik}$ is the unpenalized estimator for $\boldsymbol{\beta}$ in (2.4).

2.1. The revised EM algorithm for robust variable selection. If j-th observation $(\mathbf{X}_j, Y_j)$ is from i-th component, we denote $R_{ij} = 1, i = 1, \cdots, m, j = 1, \cdots, n$, otherwise, $R_{ij} = 0$. Assume the complete data set $\{(\mathbf{X}_j, Y_j, R_{ij}), i = 1, \cdots, m, j = 1, \cdots, n\}$ is observed, then, (2.4) can be written as

$$\begin{aligned}
&\bar{\ell}_n(\boldsymbol{\theta}) \\
&= \sum_{j=1}^{n} \omega_j \sum_{i=1}^{m} R_{ij} \log \left[\frac{\pi_i}{\sqrt{2}\sigma_i} \exp \left(- \frac{\sqrt{2}|Y_j - \mathbf{X}_j^T \boldsymbol{\beta}_i|}{\sigma_i} \right) \right] - \sum_{i=1}^{m} \pi_i \left\{ \sum_{k=1}^{q} \lambda_{ik} \frac{|\beta_{ik}|}{|\hat{\beta}_{ik}|} \right\} \\
&= \sum_{j=1}^{n} \sum_{i=1}^{m} \omega_j R_{ij} \log \pi_i - \sum_{j=1}^{n} \sum_{i=1}^{m} \omega_j R_{ij} \log(\sqrt{2}\sigma_i) - \sum_{j=1}^{n} \sum_{i=1}^{m} \frac{\omega_j R_{ij} \sqrt{2}|Y_j - \mathbf{X}_j^T \boldsymbol{\beta}_i|}{\sigma_i} \\
&- \sum_{i=1}^{m} \pi_i \left\{ \sum_{k=1}^{q} \lambda_{ik} \frac{|\beta_{ik}|}{|\hat{\beta}_{ik}|} \right\}.
\end{aligned}$$

In the following, we introduce the revised EM algorithm to maximize $\bar{\ell}_n(\boldsymbol{\theta})$ iteratively.

(1) Choose an initial value for $\boldsymbol{\theta}$, denote $\boldsymbol{\theta}^{(0)}$.

(2) E-Step. Given the data $\mathbf{D}_n$ and $\boldsymbol{\theta}^{(k)}$, we compute the conditional expectation of the function $\bar{\ell}_n(\boldsymbol{\theta})$ with respect to R_{ij}. The conditional expectation is given as follows:

$$\begin{aligned}
Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}) &= \sum_{j=1}^{n} \omega_j \sum_{i=1}^{m} \kappa_{ij}^{(k)} \log \pi_i - \sum_{j=1}^{n} \omega_j \sum_{i=1}^{m} \kappa_{ij}^{(k)} \log(\sqrt{2}\sigma_i) \\
&- \sum_{j=1}^{n} \omega_j \sum_{i=1}^{m} \kappa_{ij}^{(k)} \frac{\sqrt{2}|Y_j - \mathbf{X}_j^T \boldsymbol{\beta}_i|}{\sigma_i} - \sum_{i=1}^{m} \pi_i \left\{ \sum_{k=1}^{q} \lambda_{ik} \frac{|\beta_{ik}|}{|\hat{\beta}_{ik}|} \right\}.
\end{aligned} \tag{2.5}$$

where

$$\kappa_{ij}^{(k)} = E[R_{ij} | \mathbf{D}_n, \boldsymbol{\theta}^{(k)}] = \frac{\pi_i^{(k)} \sigma_i^{(k)-1} \exp\{-|Y_j - \mathbf{X}_j^T \boldsymbol{\beta}_i^{(k)}|/\sigma_i^{(k)}\}}{\sum_{i=1}^{m} \pi_i^{(k)} \sigma_i^{(k)-1} \exp\{-|Y_j - \mathbf{X}_j^T \boldsymbol{\beta}_i^{(k)}|/\sigma_i^{(k)}\}}.$$

(3) M-step. The M step on the $(k+1)$-th iteration maximizes $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$ with respect to $\boldsymbol{\theta}$. In the usual EM algorithm, the mixing proportions are updated by

$$\pi_i^{(k+1)} = \frac{\sum_{j=1}^{n} \omega_j \kappa_{ij}^{(k)}}{\sum_{j=1}^{n} \omega_j}, i = 1, \cdots, m,$$

which maximize the leading term of $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$. This works well in our simulations.

In the following, we consider that the π_k are constant in $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$, and maximize $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$ with respect to the other parameters. Since the objective function $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$ is not smooth, we maximize the following objective function by the local quadratic approximation [6, 10],

$$\begin{aligned}
&\sum_{j=1}^{n} \omega_j \sum_{i=1}^{m} \kappa_{ij}^{(k)} \log \pi_i - \frac{1}{2} \sum_{j=1}^{n} \omega_j \sum_{i=1}^{m} \kappa_{ij}^{(k)} \log(2\sigma_i^2) \\
&- \sum_{j=1}^{n} \omega_j \sum_{i=1}^{m} \kappa_{ij}^{(k)} \frac{\sqrt{2}(Y_j - \mathbf{X}_j^T \boldsymbol{\beta}_i)^2}{\sigma_i^2} \frac{\sigma_i^{(k)}}{|Y_j - \mathbf{X}_j^T \boldsymbol{\beta}_i^{(k)}|} \\
&- \sum_{i=1}^{m} \pi_i \left\{ \sum_{k=1}^{q} \lambda_{ik} \frac{\beta_{ik}^2}{|\hat{\beta}_{ik}||\beta_{ik}^{(k)}|} \right\}.
\end{aligned} \tag{2.6}$$

Then, the regression coefficients are updated by solving the following equations

$$\begin{aligned}&\sum_{j=1}^{n}\omega_j\kappa_{ij}^{(k)}\frac{\partial}{\partial\beta_{it}}\left[\frac{\sqrt{2}(Y_j-\mathbf{X}_j^T\boldsymbol{\beta}_i)^2}{(\sigma_i^{(k)})^2}\frac{\sigma_i^{(k)}}{|Y_j-\mathbf{X}_j^T\boldsymbol{\beta}_i^{(k)}|}\right]\\ &+\frac{\partial}{\partial\beta_{it}}\left[\pi_i\left\{\lambda_{it}\frac{\beta_{it}^2}{|\hat{\beta}_{it}||\beta_{it}^{(k)}|}\right\}\right]=0,\end{aligned}\tag{2.7}$$

where $i=1,\cdots,m$, and $t=1,\cdots,q$.

The dispersion parameters are updated by the following expression

$$\sigma_i^{2(k+1)}=\frac{2}{\sum_{j=1}^{n}\omega_j\kappa_{ij}^{(k)}}\sum_{j=1}^{n}\omega_j\kappa_{ij}^{(k)}\frac{\sqrt{2}(Y_j-\mathbf{X}_j^T\boldsymbol{\beta}_i^{(k+1)})^2\sigma_i^{(k)}}{|Y_j-\mathbf{X}_j^T\boldsymbol{\beta}_i^{(k)}|},i=1,\cdots,m.\tag{2.8}$$

(4) Repeat steps 2, 3 until convergence.

Remark 2.1 The above proposed revised EM-algorithm involves in an initial estimator, we select a robust estimation proposed by [26] for the unpenalized FMLR models as an initial estimator, that is, by maximizing the following objective function,

$$\sum_{j=1}^{n}\log\left[\sum_{i=1}^{m}\frac{\pi_i}{\sqrt{2}\sigma_i}\exp\left(-\frac{\sqrt{2}|Y_j-\mathbf{X}_j^T\boldsymbol{\beta}_i|}{\sigma_i}\right)\right].$$

Remark 2.2 To avoid numerical instability of the proposed algorithm due to very small values in the denominator of (2.7) and (2.8), as suggested by [10], we replace $|\beta_{it}^{(k)}|$ and $|Y_j-\mathbf{X}_j^T\boldsymbol{\beta}_i^{(k)}|$ by $|\beta_{it}^{(k)}|+\epsilon$ and $|Y_j-\mathbf{X}_j^T\boldsymbol{\beta}_i^{(k)}|+\epsilon$ for a given small value $\epsilon>0$. In this paper, we take $\epsilon=10^{-6}$.

3. Simulation and Application

3.1. Simulation study. In this section, we will evaluate the finite sample performance of proposed method via simulation studies. To compare the proposed approach with some existing methods, we generate the sample data $(\mathbf{X}_1,Y_1),\cdots,(\mathbf{X}_n,Y_n)$ from the following two-component mixture regression models,

$$Y_i=\left\{\begin{array}{ll}\mathbf{X}_i^T\boldsymbol{\beta}_1+\epsilon_1, & \text{if } Z=1,\\ \mathbf{X}_i^T\boldsymbol{\beta}_2+\epsilon_2, & \text{if } Z=2,\end{array}\right.\quad i=1,\cdots,n\tag{3.1}$$

with $P(Z=1)=\alpha$, $P(Z=2)=1-\alpha$, $\alpha=0.4$. We also simulate $\alpha=0.25$; the outcomes are similar, and thus we do not report the corresponding results here. The sparse regression parameters are

$$\begin{aligned}&\boldsymbol{\beta}_1=(0,\cdots,0,-2.5,-1.5)^T,\\ &\boldsymbol{\beta}_2=(0,\cdots,0,-2.5,1.5)^T,\end{aligned}$$

where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ have eight zero elements. Covariate $\mathbf{X}_i$ follows a multi-normal distribution $N(\mathbf{0},\Sigma)$, where the (i,j)-th element of Σ is $\rho^{|i-j|}$, $\rho=0.5$. The error terms ϵ_1 and ϵ_2 are independent and identically distributed random variables. To study the robustness of proposed method, we consider the following four settings:

(1) The error terms follow a standard normal distribution, $N(0,1)$;

(2) The error terms follow a Student's t-distribution with 2 degrees of freedom, t_2;

(3) The error terms follow a 5% contaminated normal distribution, $CN_{0.05}=0.95N(0,1)+0.05N(10,20^2)$;

(4) The error terms follow a standard normal distribution with 5% high leverage outliers being $\mathbf{X}_1=(50,\cdots,50)^T$, and $Y=100$.

For each setting, we simulate 200 data sets from model (3.1) with sample sizes of $n = 200, 400$, and compare the performance of proposed method (MixregL-MCD) with the penalized likelihood approach (MixregL-ALASSO) [15] and the oracle estimator based on the Laplace error to the data after trimming the high leverage points based on a robust Mahalanobis distance with the MCD estimators. To measure the finite sample performance, we report the proportions of correctly estimated zero coefficients (specificity: S_1) and correctly estimated non-zero coefficients (sensitivity: S_2), and the component-wise median empirical mean squared errors (MEMSE) of the estimators $\hat{\boldsymbol{\beta}}_k, k = 1, 2$. According to [15] and [17], we consider the tuning parameter $\lambda_{ik} = \log(n) \times \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$. In simulation studies, the finite sample performance of $\lambda_{ik} = \log(n) \times 0.2$ is slightly better than that of others. Therefore, we take $\lambda_{ik} = \log(n) \times 0.2$ in all simulation studies and real data applications. Clearly, the choice of tuning parameter is a very important issue, however, we shall not address the problem of how to find the optimal tuning parameter, and will consider the choice of tuning parameter as future work. The simulation results are given in Table 1-4.

From Table 1, we find that when the true distribution of error term is normal and there are no outliers in the dataset, both S_1 and S_2 are around 1 for all three methods. The MEMSE of both methods is close to that of oracle estimator. When there are outliers in the datasets or the error term follows a heavy-tailed distribution, the simulation results clearly show from Table 2 to Table 4 that the proposed method works much better than the MixregL-ALASSO. S_1 and S_2 of the proposed method are higher than those of the MixregL-ALASSO, and our proposed approach has smaller MEMSE than the MixregL-ALASSO. In addition, the performance of proposed method is closer to that of the oracle estimator as the sample size n increases.

Based on the above findings, the proposed method is not sensitive to outliers in the dataset, and has the overall best performance. Thus, we recommend the use of proposed method in practical applications.

In the above simulations, we assume that the number of mixture components is known. However, the order m needs to be estimated based on the dataset in some applications. There are many methods to choose the order m in the literature, e.g., cross-validation (CV), generalized cross-validation (GCV), Akaike information criterion (AIC), and bayesian information criterion (BIC). In this paper, we select the order m by minimizing a following BIC-score

$$BIC(m) = -2l_n(\bar{\boldsymbol{\theta}}_m) + S\log(n),$$

where $\bar{\boldsymbol{\theta}}_m$ is the maximizer of the proposed objective function for a mixture regression model with the order m, and S is the number of nonzero of the estimator $\bar{\boldsymbol{\theta}}_m$.

In the following, we will use simulation studies to illustrate how to select the order. A total of 300 data sets with sample sizes $n = 400$ are generated according to the second setting with true $m = 2$. The simulation result is shown in Figure 1. We can see from Figure 1 that the BIC performs well to select the true order.

3.2. Real data application. In this section, we will apply the proposed methodology to analyze the baseball salaries dataset, which can be downloaded from

`www.amstat.org/publications/jse`.

This dataset contains 337 observations. Of interest are to study the relationships between the salary (measured in thousands of dollars) and the following 16 covariates: batting average (X_1), on-base percentage (X_2), runs (X_3), hits (X_4), doubles (X_5), triples (X_6), home runs (X_7), runs batted in (X_8), walks (X_9), strikeouts (X_{10}), stolen bases(X_{11}) , errors (X_{12}), free agency eligibility (X_{13}), free agent in 1991/2 (X_{14}), arbitration eligibility (X_{15}), and arbitration in 1991/2 (X_{16}). $X_{13}, X_{14}, X_{15}, X_{16}$ are indicators.

Table 1. Simulation results for the first setting

n		Method	S_1	S_2	MEMSE
200	$\boldsymbol{\beta}_1$	MixregL-ALASSO	0.9850	1.0000	0.0087
		MixregL-MCD	1.0000	0.9950	0.0054
		Oracle	1.0000	1.0000	0.0052
	$\boldsymbol{\beta}_2$	MixregL-ALASSO	0.9863	1.0000	0.0027
		MixregL-MCD	1.0000	1.0000	0.0043
		Oracle	1.0000	1.0000	0.0038
400	$\boldsymbol{\beta}_1$	MixregL-ALASSO	0.9950	1.0000	0.0047
		MixregL-MCD	1.0000	1.0000	0.0038
		Oracle	1.0000	1.0000	0.0033
	$\boldsymbol{\beta}_2$	MixregL-ALASSO	1.0000	1.0000	0.0016
		MixregL-MCD	1.0000	1.0000	0.0021
		Oracle	1.0000	1.0000	0.0016

Table 2. Simulation results for the second setting

n		Method	S_1	S_2	MEMSE
200	$\boldsymbol{\beta}_1$	MixregL-ALASSO	0.6925	0.7750	0.2260
		MixregL-MCD	1.0000	0.9850	0.0096
		Oracle	1.0000	1.0000	0.0075
	$\boldsymbol{\beta}_2$	MixregL-ALASSO	0.6775	0.9450	0.0398
		MixregL-MCD	1.0000	0.9900	0.0043
		Oracle	1.0000	1.0000	0.0038
400	$\boldsymbol{\beta}_1$	MixregL-ALASSO	0.8288	0.7100	0.2317
		MixregL-MCD	1.0000	1.0000	0.0055
		Oracle	1.0000	1.0000	0.0039
	$\boldsymbol{\beta}_2$	MixregL-ALASSO	0.8363	0.9500	0.1001
		MixregL-MCD	1.0000	1.0000	0.0026
		Oracle	1.0000	1.0000	0.0019

Table 3. Simulation results for the third setting

n		Method	S_1	S_2	MEMSE
200	$\boldsymbol{\beta}_1$	MixregL-ALASSO	0.8550	0.7600	0.3095
		MixregL-MCD	1.0000	0.9200	0.0097
		Oracle	1.0000	1.0000	0.0065
	$\boldsymbol{\beta}_2$	MixregL-ALASSO	0.8838	0.8750	0.1749
		MixregL-MCD	1.0000	0.9750	0.0064
		Oracle	1.0000	1.0000	0.0039
400	$\boldsymbol{\beta}_1$	MixregL-ALASSO	0.7512	0.7850	0.3382
		MixregL-MCD	1.0000	0.9450	0.0051
		Oracle	1.0000	1.0000	0.0048
	$\boldsymbol{\beta}_2$	MixregL-ALASSO	0.8275	0.9300	0.1486
		MixregL-MCD	1.0000	0.9750	0.0052
		Oracle	1.0000	1.0000	0.0048

We plot a histogram of home runs and stolen bases in Figure 2. Figure 2 indicates that there are unusual points in the dataset. According to the suggestion proposed by [15], we apply the MixregL-ALASSO and MixregL-MCD with $m = 2$ to deal with this dataset. The results are summarized in Table 5. From Table 5, we find that the MixregL-ALASSO

Table 4. Simulation results for the fourth setting

n		Method	S_1	S_2	MEMSE
200	$\boldsymbol{\beta}_1$	MixregL-ALASSO	0.2350	0.9450	0.7180
		MixregL-MCD	1.0000	0.9900	0.0055
		Oracle	1.0000	1.0000	0.0041
	$\boldsymbol{\beta}_2$	MixregL-ALASSO	0.3188	1.0000	0.1766
		MixregL-MCD	1.0000	1.0000	0.0030
		Oracle	1.0000	1.0000	0.0021
400	$\boldsymbol{\beta}_1$	MixregL-ALASSO	0.1075	0.9750	0.6932
		MixregL-MCD	1.0000	1.0000	0.0033
		Oracle	1.0000	1.0000	0.0031
	$\boldsymbol{\beta}_2$	MixregL-ALASSO	0.1862	1.0000	0.1420
		MixregL-MCD	1.0000	1.0000	0.0014
		Oracle	1.0000	1.0000	0.0011

obtains more significant explanatory variables than the MixregL-MCD. However, our proposed method should give the more reasonable model when there are outliers in the dataset.

Table 5. Estimated regression coefficients from the baseball salaries dataset

	Method			
	MixregL-ALASSO		MixregL-MCD	
Variable	Component 1	Component 2	Component 1	Component 2
X_1	6.5864	0.0011	0	0
X_2	10.677	15.388	16.742	19.816
X_3	0	0	0	0
X_4	0	0.0026	0	0
X_5	0	0	0	0
X_6	0	0	0	0
X_7	0	-0.0005	0	0
X_8	0.0063	0.0058	0	0
X_9	-0.0067	-0.0104	0	0
X_{10}	0.0053	0.0066	0	0
X_{11}	0	0	0	0
X_{12}	0	0.0002	0	0
X_{13}	1.7160	1.6827	2.1480	0
X_{14}	-0.0084	-0.0012	0	0
X_{15}	1.4961	1.4833	1.6340	0
X_{16}	0	-0.0001	0	0

4. Discussion

In this article, we proposed a robust variable selection by assuming that the error terms follow a Laplace distribution for FMLR models. We used the revised EM-algorithm to solve the proposed optimization problem. The merits of proposed methodology were illustrated via the simulation studies. According to our simulation studies, the proposed method was robust and possessed a consistent variable selection when there were outliers or the error distribution was heavy-tail.

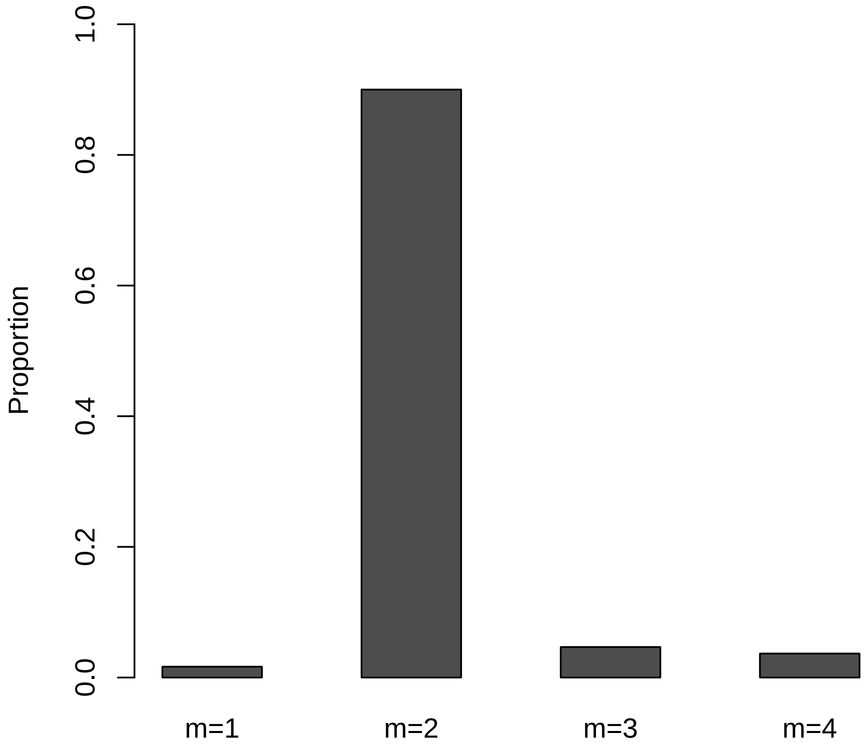
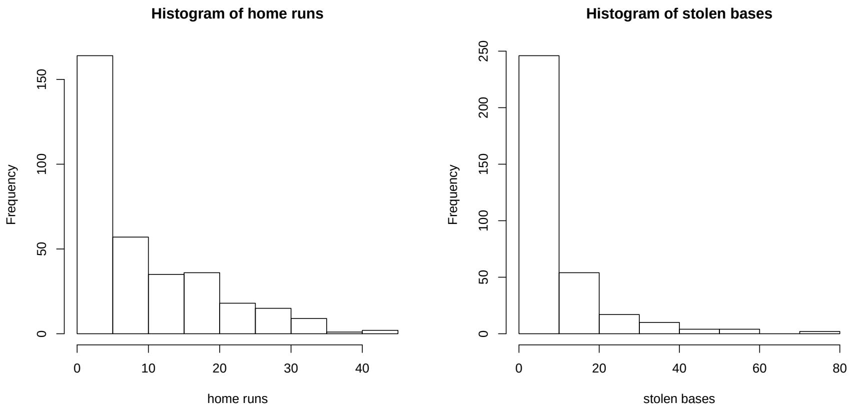


Figure 1. Order selection results based on BIC for the FMLR models with true order $m = 2$

As a variable selection procedure, it is very desirable to enjoy the oracle properties. Therefore, it warrants further effort to investigate the asymptotic properties for the proposed method. Meanwhile, it is very interesting to extent our methodology to nonparametric mixture of regression models [8], a class of semiparametric mixtures of regression models [11, 9], and mixture of gaussian processes [7].

Acknowledgements

Jiang's research is partially supported by the National Natural Science Foundation of China (No. 11301221), and is supported by the Fundamental Research Funds for the Central Universities (No. 11615455).



(a) (b)

Figure 2. Histogram of home runs (a) and stolen bases (b).

References

[1] Bai, X., Yao, W., and Boyer, J. E. *Robust fitting of mixture regression models*, Comput. Stat. Data. An. **56** (7), 2347-2359, 2012.

[2] Croux, C., and Haesbroeck, G. *Influence function and efficiency of the mini- mum covariance determinant scatter matrix estimator*, J. Multivariate Anal. **71** (2), 161-190, 1999.

[3] Davies, P. *Asymptotic Behaviour of S-Estimates of Multivariate Location Param- eters and Dispersion Matrices*, Ann. Statist. **15** (3), 1269-1292, 1987.

[4] Donoho, D. L. *Breakdown properties of multivariate location estimators,, Technical report, Technical report, Harvard University, Boston. URL http://www-stat. stanford. edu/ donoho/Reports/Oldies/BPMLE. pdf*, 1982.

[5] Du, Y., Khalili, A., Neslehova, J. G., and Steele, R. J. *Simultaneous fixed and random effects selection in finite mixture of linear mixed-effects models*, Can. J. Stat. **41** (4), 596-616, 2013.

[6] Fan, J., and Li, R. *Variable selection via nonconcave penalized likelihood and its oracle properties*, J. Amer. Statist. Assoc. **96** (456), 1348-1360, 2001.

[7] Huang, M., Li, R., Wang, H., and Yao, W. *Estimating Mixture of Gaussian Pro- cesses by Kernel Smoothing*, J. Bus. Econ. Stat. **32** (2), 259-270, 2014.

[8] Huang, M., Li, R., and Wang, S. *Nonparametric mixture of regression models*, J. Amer. Statist. Assoc. **108** (503), 929-941, 2013.

[9] Huang, M., and Yao, W. *Mixture of regression models with varying mixing pro- portions: a semiparametric approach*, J. Amer. Statist. Assoc. **107** (498), 711-724, 2012.

[10] Hunter, D. R., and Li, R. *Variable selection using MM algorithms*, Ann. Statist. **33** (4), 1617-1642, 2005.

[11] Hunter, D. R., and Young, D. S. *Semiparametric mixtures of regressions*, J. Nonparametr. Stat. **24** (1), 19-38, 2012.

[12] Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. *Adaptive mixtures of local experts*, Neural. Comput. **3** (1), 79-87, 1991.

[13] Jiang, W., Tanner, M. A. et al. *Hierarchical mixtures-of-experts for exponential fam- ily regression models: approximation and maximum likelihood estimation*, Ann. Statist. **27** (3), 987-1011, 1997.

[14] Khalili, A. *An Overview of the New Feature Selection Methods in Finite Mixture of Regression Models*, J. Iran. Stat. Soc. **10** (2), 201-235, 2011.

[15] Khalili, A., and Chen, J. *Variable Selection in Finite Mixture of Regression Models*, J. Amer. Statist. Assoc. **102** (479), 1025-1038, 2007.

[16] Khalili, A., and Lin, S. *Regularization in finite mixture of regression models with diverging number of parameters*, Biometrics **69** (2), 436-446, 2013.

[17] Luo, R., Wang, H., and Tsai, C.-L.*On Mixture Regression Shrinkage and Selection Via the MR-Lasso*, Int. J. Pure. Ap. Mat. **46**, 403-414, 2008.

[18] Markatou, M. *Mixture models, robustness, and the weighted likelihood methodol- ogy*, Biometrics **56** (2), 483-486, 2000.

[19] Maronna, R. A. et al. *Robust M-Estimators of Multivariate Location and Scatter*, Ann. Statist. **4** (1), 51-67, 1976.

[20] McLachlan, G., and Peel, D. *Finite mixture models*(John Wiley & Sons, 2004).

[21] Neykov, N., Filzmoser, P., Dimova, R., and Neytchev, P. *Robust fitting of mix- tures using the trimmed likelihood estimator*, Comput. Stat. Data. An. **52** (1), 299-308, 2007.

[22] Rousseeuw, P. *Multivariate estimation with high breakdown point*, status: published, 1985.

[23] Rousseeuw, P. J., and Driessen, K. V. *A fast algorithm for the minimum covariance determinant estimator*, Technometrics **41** (3), 212-223, 1990.

[24] Shen, H.-b., Yang, J., and Wang, S.-t. *Outlier detecting in fuzzy switching regres- sion models*, in Artitificial Intelligence: Methodology, Systems, and Applications Springer, 208-215, 2004.

[25] Skrondal, A., and Rabe-Hesketh, S. *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models* (CRC Press, 2004).

[26] Song, W., Yao, W., and Xing, Y. *Robust mixture regression model fitting by Laplace distribution*, Comput. Stat. Data. An. **71**, 128-137, 2014.

[27] Stahel, W. *Robust estimation: Infinitesimal optimality and covariance matrix esti- mators*, Unpublished doctoral dissertation, ETH, Zurich, Switzerland, 1981.

[28] Wang, P., Puterman, M. L., Cockburn, I., and Le, N. *Mixed Poisson regression models with covariate dependent rates*, Biometrics **52** (2), 381-400, 1996.

[29] Wedel, M. *Market segmentation: Conceptual and methodological foundations*(Springer, 2000).

[30] Wei, Y. *Robust mixture regression models using t-distribution*, Master's thesis, 2012.