Scientific Research Publishing

Advances in Pure Mathematics, 2020, 10, 39-55
https://www.scirp.org/journal/apm
ISSN Online: 2160-0384
ISSN Print: 2160-0368

Adaptive Sparse Group Variable Selection for a Robust Mixture Regression Model Based on Laplace Distribution

Jiangtao Wang, Wanzhou Ye

Department of Mathematics, College of Science, Shanghai University, Shanghai, China
Email: jiangtao2107@163.com, wzhy@shu.edu.cn

How to cite this paper: Wang, J.T. and Ye, W.Z. (2020) Adaptive Sparse Group Variable Selection for a Robust Mixture Regression Model Based on Laplace Distribution. *Advances in Pure Mathematics*, **10**, 39-55.
https://doi.org/10.4236/apm.2020.101004

Received: December 25, 2019
Accepted: January 16, 2020
Published: January 19, 2020

Copyright © 2020 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/

Open Access

Abstract

The traditional estimation of Gaussian mixture model is sensitive to heavy-tailed errors; thus we propose a robust mixture regression model by assuming that the error terms follow a Laplace distribution in this article. And for the variable selection problem in our new robust mixture regression model, we introduce the adaptive sparse group Lasso penalty to achieve sparsity at both the group-level and within-group-level. As numerical experiments show, compared with other alternative methods, our method has better performances in variable selection and parameter estimation. Finally, we apply our proposed method to analyze NBA salary data during the period from 2018 to 2019.

Keywords

Robust Mixture Regression, Laplace Distribution, Adaptive Sparse Group Lasso

1. Introduction

The mixture regression model is a powerful tool to explain the relationships between the response variable and the covariates when the population is heterogeneous and consists of several homogeneous components, and the early research can trace back to [1]. In 1977, EM algorithm was first proposed by [2]; it greatly simplified the solution procedure of the mixture regression model. Then the mixture regression model attracted a lot of interest from statisticians; it was widely applied in many fields, such as business, marketing and social sciences. See [3] [4] [5] for example.

Recently, the research about the mixture regression model is becoming more and more detailed. On the one hand, statisticians paid attention to improving

the robustness of mixture regression model, [6] used the *t*-distribution for overcoming the influence of outliers and [7] introduced a robust mixture regression model by assuming the error terms follow a Laplace distribution. Further, Wu *et al.* [8] dropped any parametric assumption about the error densities and proposed the mixture of quantile regressions model. On the other hand, variable selection became a research hotspot in mixture regression modeling. Khalili and Chen [9] considered a class of penalization methods, including L_1-norm penalty, SCAD penalty and Hard penalty. Furthermore, the adaptive Lasso was introduced as a penalty function in [10], and [11] suggested the use of the Lasso-penalized mixture regression model as a screening mechanism in a two-step procedure.

However, the above regularization methods are more focused on individual variable selection. They all ignore the grouping structures which describe the inherent interconnections among predictors. It may lead to inefficient models. In order to achieve both the robustness of mixture regression model and correct identification of group structures, we assume random errors follow a Laplace distribution and consider a situation that covariates have natural grouping structures, where those in the same group are correlated. In this case, variable selection should be conducted at both the group-level and within-group-level; thus we use the adaptive sparse group Lasso [12] as a penalty function of our proposed mixture regression model and adopt EM algorithm to estimate the mixture regression parameters. Moreover, under some mild conditions, we can prove that the maximum penalized log-likelihood estimators are both sparse and $\sqrt{n}$ -consistent simultaneously.

The rest of this article is organized as follows. In Section 2, we introduce the robust mixture regression model based on Laplace distribution and adopt the adaptive sparse group Lasso for variable selection. In Section 3, we prove some asymptotic properties for our proposed method. In Section 4, we solve the problem of tuning parameters and components selection. Section 5 conducts a numerical simulation to evaluate the performance of our method. In Section 6, we apply our proposed method to NBA salary data. Finally, the conclusion of this paper is given in Section 7.

2. Model Overview

2.1. Robust Mixture Regression with Laplace Distribution

Let $\{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)\}$ be a random sample of observations from the population $(\boldsymbol{x}, y)$, where $\boldsymbol{x} = (x_1, \cdots, x_p)^{\mathrm{T}} \in \mathbb{R}^p$ is a p-dimensional covariate vector, and $y \in \mathbb{R}$ is the response variable which is dependent on corresponding $\boldsymbol{x}$. Furthermore, for g mixture components, we can say that $(\boldsymbol{x}, y)$ follows a mixture linear regression model based on a normal distribution if the conditional density of y given $\boldsymbol{x}$ is

$$f(y \mid \boldsymbol{x}, \boldsymbol{\Psi}) = \sum_{j=1}^{g} \frac{\pi_j}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{\left(y - \alpha_j - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_j\right)^2}{2\sigma_j^2} \right), \tag{1}$$

where the mixing probabilities satisfy $\sum_{j=1}^{g}\pi_j = 1$, $\pi_j > 0$ and the parameter $\boldsymbol{\Psi} = \left(\pi_1, \alpha_1, \boldsymbol{\beta}_1, \sigma_1^2, \cdots, \pi_g, \alpha_g, \boldsymbol{\beta}_g, \sigma_g^2\right)$. For the jth mixture component, there are intercept $\alpha_j \in \mathbb{R}$, regression coefficients $\boldsymbol{\beta}_j = \left(\beta_{j1}, \cdots, \beta_{jp}\right)^{\mathrm{T}}$ and variance $\sigma_j^2\left(\sigma_j > 0\right)$.

It is known that the mixture linear regression model is sensitive to outliers or heavy-tailed error distributions, and outliers impact more heavily on the mixture linear regression model than on the usual linear regression model, since outliers not only affect the estimation of the regression parameters, but also possibly totally blur the mixture structure. In order to improve the robustness of the estimation procedure, we introduce a robust mixture regression model with a Laplace distribution

$$f\left(y \mid \boldsymbol{x}, \boldsymbol{\Psi}\right) = \sum_{j=1}^{g}\frac{\pi_j}{\sqrt{2\sigma_j^2}}\exp\left(-\frac{\sqrt{2}\left|y - \alpha_j - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_j\right|}{\sigma_j}\right). \tag{2}$$

Then we can estimate the unknown parameter $\boldsymbol{\Psi}$ by maximizing the log-likelihood function

$$L_n\left(\boldsymbol{\Psi}\right) = \sum_{i=1}^{n}\log\sum_{j=1}^{g}\frac{\pi_j}{\sqrt{2\sigma_j^2}}\exp\left(-\frac{\sqrt{2}\left|y_i - \alpha_j - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j\right|}{\sigma_j}\right). \tag{3}$$

2.2. Adaptive Sparse Group Lasso for Variable Selection

Now, we consider a situation that covariates have natural grouping structures and can be divided into K groups as $\boldsymbol{x} = \left(\boldsymbol{x}^{[1]\mathrm{T}}, \cdots, \boldsymbol{x}^{[k]\mathrm{T}}, \cdots, \boldsymbol{x}^{[K]\mathrm{T}}\right)^{\mathrm{T}} \in \mathbb{R}^p$ by some known rules, where $\boldsymbol{x}^{[k]} = \left(x_1^{[k]}, \cdots, x_{p_k}^{[k]}\right)^{\mathrm{T}}$ is a group which contains p_k variables and $\sum_{k=1}^{K} p_k = p$. Then, the log-likelihood function can be written as

$$L_n\left(\boldsymbol{\Psi}\right) = \sum_{i=1}^{n}\log\sum_{j=1}^{g}\frac{\pi_j}{\sqrt{2\sigma_j^2}}\exp\left(-\frac{\sqrt{2}\left|y_i - \alpha_j - \sum_{k=1}^{K}\boldsymbol{x}_i^{[k]\mathrm{T}}\boldsymbol{\beta}_j^{[k]}\right|}{\sigma_j}\right), \tag{4}$$

where $\boldsymbol{\beta}_j^{[k]} = \left(\beta_{j1}^{[k]}, \cdots, \beta_{jp_k}^{[k]}\right)^{\mathrm{T}}$.

In order to exploit the grouping structures of covariates, we apply the adaptive sparse group Lasso (adaSGL) to the robust mixture regression model, the penalized log-likelihood function

$$F_n\left(\boldsymbol{\Psi}\right) = L_n\left(\boldsymbol{\Psi}\right) - P_n\left(\boldsymbol{\Psi}\right). \tag{5}$$

Here

$$P_n\left(\boldsymbol{\Psi}\right) = n\sum_{j=1}^{g}\pi_j\sum_{t=1}^{p}\lambda_{jt1}\left|\beta_{jt}\right| + n\sum_{j=1}^{g}\pi_j\sum_{k=1}^{K}\lambda_{jk2}\left\|\boldsymbol{\beta}_j^{[k]}\right\|, \tag{6}$$

where $\|.\|$ represents the Euclidean norm, $\lambda_{jt1} = \lambda_{j1}\omega_{jt}$ and $\lambda_{jk2} = \lambda_{j2}\xi_{jk}$. Moreover, the weights are defined based on the maximum penalized log-likelihood estimator $\tilde{\boldsymbol{\Psi}}$ when $P_n\left(\boldsymbol{\Psi}\right)$ is a Lasso penalty,

$$\omega_{jt} = \left|\tilde{\beta}_{jt}\right|^{-1}, \quad \xi_{jk} = \left\|\tilde{\boldsymbol{\beta}}_j^{[k]}\right\|^{-1}. \tag{7}$$

Next, we follow the approach of Hunter and Li [13] and consider to maximize the ε-approximate penalized log-likelihood function

$$F_{n,\varepsilon}(\boldsymbol{\Psi}) = L_n(\boldsymbol{\Psi}) - P_{n,\varepsilon}(\boldsymbol{\Psi}). \tag{8}$$

Here

$$P_{n,\varepsilon}(\boldsymbol{\Psi}) = n\sum_{j=1}^{g}\pi_j\sum_{t=1}^{p}\lambda_{jt1}\sqrt{\beta_{jt}^2+\varepsilon^2} + n\sum_{j=1}^{g}\pi_j\sum_{k=1}^{K}\lambda_{jk2}\sqrt{\sum_{t=1}^{p_k}\beta_{jt}^{2[k]}+\varepsilon^2} \tag{9}$$

for some small $\varepsilon > 0$, and the weights are

$$\omega_{jt} = \left(\tilde{\beta}_{jt}^2+\varepsilon^2\right)^{-1/2}, \quad \xi_{jk} = \left(\sum_{t=1}^{p_k}\tilde{\beta}_{jt}^{2[k]}+\varepsilon^2\right)^{-1/2}. \tag{10}$$

Following Hunter and Li [13], we can similarly show that $\left|F_{n,\varepsilon}(\boldsymbol{\Psi}) - F_n(\boldsymbol{\Psi})\right| \to 0$ uniformly as $\varepsilon \to 0$, over any compact subset of the parameter space.

2.3. EM Algorithm for Robust Mixture Regression

However, the above penalized log-likelihood does not have an explicit maximizer. We introduce an EM algorithm to simplify the computation and denote Z_{ij} as a latent Bernoulli variable such that

$$Z_{ij} = \begin{cases} 1, & \text{if } i\text{th observation } (\boldsymbol{x}_i, y_i) \text{ is from } j\text{th component;} \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

If the complete data set $\boldsymbol{T} = (\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ is observable, the complete log-likelihood function is

$$L_n(\boldsymbol{\Psi};\boldsymbol{T}) = \sum_{i=1}^{n}\sum_{j=1}^{g} Z_{ij}\log\left[\frac{\pi_j}{\sqrt{2\sigma_j^2}}\exp\left(-\frac{\sqrt{2}\left|y_i - \alpha_j - \sum_{k=1}^{K}\boldsymbol{x}_i^{[k]\mathrm{T}}\boldsymbol{\beta}_j^{[k]}\right|}{\sigma_j}\right)\right], \tag{12}$$

where $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)^{\mathrm{T}}$, $\boldsymbol{Y} = (y_1, \cdots, y_n)$ and $\boldsymbol{Z} = (Z_{11}, \cdots, Z_{ng})$.

According to Andrews and Mallows [14], we know that a Laplace distribution can be expressed as a mixture of a normal distribution and another distribution related to the exponential distribution. To be specific, there are latent scale variables $\boldsymbol{V} = (V_1, \cdots, V_n)$ such that we have the complete log-likelihood function

$$\begin{aligned} L_n(\boldsymbol{\Psi};\boldsymbol{D}) = {} & \sum_{i=1}^{n}\sum_{j=1}^{g} Z_{ij}\log\left\{\frac{\pi_j V_i}{\sqrt{\pi\sigma_j^2}}\exp\left[-\frac{V_i^2\left(y_i - \alpha_j - \sum_{k=1}^{K}\boldsymbol{x}_i^{[k]\mathrm{T}}\boldsymbol{\beta}_j^{[k]}\right)^2}{\sigma_j^2}\right]\right\} \\ & + \sum_{i=1}^{n}\sum_{j=1}^{g} Z_{ij}\log\left[\frac{1}{V_i^3}\exp\left(-\frac{1}{2V_i^2}\right)\right], \end{aligned} \tag{13}$$

where $\boldsymbol{D} = (\boldsymbol{T}, \boldsymbol{V})$. Naturally, we can obtain the penalized complete log-likelihood $F_{n,\varepsilon}(\boldsymbol{\Psi};\boldsymbol{D}) = L_n(\boldsymbol{\Psi};\boldsymbol{D}) - P_{n,\varepsilon}(\boldsymbol{\Psi})$.

Suppose that $\boldsymbol{\Psi}^{(r)}$ is a parameter estimate for the rth iteration. In E step of EM algorithm, we can get $E\left[F_{n,\varepsilon}(\boldsymbol{\Psi};\boldsymbol{D}) \mid \boldsymbol{S}, \boldsymbol{\Psi}^{(r)}\right]$ by calculating

$$\tau_{ij}^{(r)} = E\left[Z_{ij} \mid \boldsymbol{S}, \boldsymbol{\Psi}^{(r)}\right], \quad \delta_{ij}^{(r)} = E\left[V_i^2 \mid \boldsymbol{S}, \boldsymbol{\Psi}^{(r)}, Z_{ij} = 1\right], \tag{14}$$

where $\boldsymbol{S}=(\boldsymbol{X},\boldsymbol{Y})$. And we can show that

$$\tau_{ij}^{(r)}=\frac{\pi_j^{(r)}\sigma_j^{-1(r)}\exp\left(-\sqrt{2}\left|y_i-\alpha_j^{(r)}-\sum_{k=1}^{K}\boldsymbol{x}_i^{[k]\mathrm{T}}\boldsymbol{\beta}_j^{[k](r)}\right|\sigma_j^{-1(r)}\right)}{\sum_{j=1}^{g}\pi_j^{(r)}\sigma_j^{-1(r)}\exp\left(-\sqrt{2}\left|y_i-\alpha_j^{(r)}-\sum_{k=1}^{K}\boldsymbol{x}_i^{[k]\mathrm{T}}\boldsymbol{\beta}_j^{[k](r)}\right|\sigma_j^{-1(r)}\right)} \tag{15}$$

and

$$\delta_{ij}^{(r)}=\frac{\sigma_j^{(r)}}{\sqrt{2}\left|y_i-\alpha_j^{(r)}-\sum_{k=1}^{K}\boldsymbol{x}_i^{[k]\mathrm{T}}\boldsymbol{\beta}_j^{[k](r)}\right|}. \tag{16}$$

The calculation for $\delta_{ij}^{(r)}$ follows the same argument as in Phillips [15].

In M step, we will maximize $E\left[F_{n,\varepsilon}(\boldsymbol{\Psi};\boldsymbol{D})\,|\,\boldsymbol{S},\boldsymbol{\Psi}^{(r)}\right]$ for updating $\boldsymbol{\Psi}$. Now, we follow the tactic of [16] and find a local quadratic approximation of $\sqrt{\psi^2+\varepsilon^2}$,

$$\sqrt{\psi^2+\varepsilon^2}\simeq\sqrt{\psi_0^2+\varepsilon^2}+\frac{\psi^2-\psi_0^2}{2\sqrt{\psi_0^2+\varepsilon^2}} \tag{17}$$

in a neighborhood of ψ_0. Then, we can replace the penalty function $P_{n,\varepsilon}(\boldsymbol{\Psi})$ in $(r+1)$th iteration by

$$\begin{aligned}\tilde{P}_{n,\varepsilon}\left(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(r)}\right)=&\,n\sum_{j=1}^{g}\pi_j\sum_{t=1}^{p}\lambda_{jt1}\left[\eta_{jt}^{(r)}+\frac{\beta_{jt}^2-\beta_{jt}^{2(r)}}{2\eta_{jt}^{(r)}}\right]\\&+n\sum_{j=1}^{g}\pi_j\sum_{k=1}^{K}\lambda_{jk2}\left[\gamma_{jk}^{(r)}+\frac{\sum_{t=1}^{p_k}\left(\beta_{jt}^{2[k]}-\beta_{jt}^{2[k](r)}\right)}{2\gamma_{jk}^{(r)}}\right],\end{aligned} \tag{18}$$

where $\eta_{jt}^{(r)}=\sqrt{\beta_{jt}^{2(r)}+\varepsilon^2}$ and $\gamma_{jk}^{(r)}=\sqrt{\sum_{t=1}^{p_k}\beta_{jt}^{2[k](r)}+\varepsilon^2}$. Similarly, from Lange [17], we have

$$\left(y-\alpha-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\psi}\right)^2\simeq\frac{1}{p}\sum_{t=1}^{p}\left[y-\alpha-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\psi}_0-px_t\left(\psi_t-\psi_{0t}\right)\right]^2 \tag{19}$$

in a neighborhood of $\boldsymbol{\psi}_0=\left(\psi_{01},\cdots,\psi_{0p}\right)^{\mathrm{T}}$, where p is the dimensionality of $\boldsymbol{\psi}$. And we apply (19) to $E\left[L_{n,\varepsilon}(\boldsymbol{\Psi};\boldsymbol{D})\,|\,\boldsymbol{S},\boldsymbol{\Psi}^{(r)}\right]$, there is a local approximation $Q\left(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(r)}\right)$ of $E\left[F_{n,\varepsilon}(\boldsymbol{\Psi};\boldsymbol{D})\,|\,\boldsymbol{S},\boldsymbol{\Psi}^{(r)}\right]$

$$\begin{aligned}&Q\left(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(r)}\right)\\&=-\sum_{i=1}^{n}\sum_{j=1}^{g}\tau_{ij}^{(r)}\frac{\delta_{ij}^{(r)}}{p\sigma_j^2}\sum_{t=1}^{p}\left[y_i-\alpha_j-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j^{(r)}-px_{it}\left(\beta_{jt}-\beta_{jt}^{(r)}\right)\right]^2\\&\quad-\frac{n}{2}\sum_{j=1}^{g}\pi_j\left(\sum_{t=1}^{p}\frac{\lambda_{jt1}}{\eta_{jt}^{(r)}}\beta_{jt}^2+\sum_{k=1}^{K}\frac{\lambda_{jk2}}{\gamma_{jk}^{(r)}}\sum_{t=1}^{p_k}\beta_{jt}^{2[k]}\right)+\sum_{i=1}^{n}\sum_{j=1}^{g}\tau_{ij}^{(r)}\left(\log\pi_j-\frac{1}{2}\log\sigma_j^2\right)\\&\quad-n\sum_{j=1}^{g}\pi_j\left[\sum_{t=1}^{p}\lambda_{jt1}\left(\eta_{jt}^{(r)}-\frac{\beta_{jt}^{2(r)}}{2\eta_{jt}^{(r)}}\right)+\sum_{k=1}^{K}\lambda_{jk2}\left(\gamma_{jk}^{(r)}-\frac{\sum_{t=1}^{p_k}\beta_{jt}^{2[k](r)}}{2\gamma_{jk}^{(r)}}\right)\right]\\&\quad-\sum_{i=1}^{n}\sum_{j=1}^{g}\tau_{ij}^{(r)}\left(\frac{1}{2}\log\pi+\log\delta_{ij}^{(r)}+\frac{1}{2\delta_{ij}^{(r)}}\right)\end{aligned} \tag{20}$$

in a neighborhood of $\boldsymbol{\Psi}^{(r)}$. Note that (20) can be block-wise maximized in the

coordinates of the parameter components $\boldsymbol{\pi}$, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$. Here, $\boldsymbol{\pi}=(\pi_1,\cdots,\pi_g)$, $\boldsymbol{\alpha}=(\alpha_1,\cdots,\alpha_g)$, $\boldsymbol{\beta}=(\beta_{11},\cdots,\beta_{gp})$ and $\boldsymbol{\sigma}^2=(\sigma_1^2,\cdots,\sigma_g^2)$.

Under the constraints that $\sum_{j=1}^{g}\pi_j=1$ and $\pi_j>0$, we adopt Lagrangian multiplier to update $\boldsymbol{\pi}$ by solving

$$\nabla Q\left(\boldsymbol{\pi},\boldsymbol{\alpha}^{(r)},\boldsymbol{\beta}^{(r)},\boldsymbol{\sigma}^{2(r)};\boldsymbol{\Psi}^{(r)}\right)=\mathbf{0}, \tag{21}$$

where ∇ is the gradient operator, ζ is a positive scalar and $\mathbf{0}$ is a zero vector. Then we have the set of simultaneous equations

$$a_j/\pi_j-b_j-\zeta=0, \tag{22}$$

where $a_j=\sum_{i=1}^{n}\tau_{ij}^{(r)}$ and $b_j=n\sum_{t=1}^{p}\lambda_{jt1}\eta_{jt}^{(r)}+n\sum_{k=1}^{K}\lambda_{jk2}\gamma_{jk}^{(r)}$, for each *j*.

According to $\pi_j=a_j/\left(b_j+\zeta\right)$ and $\sum_{j=1}^{g}\pi_j=1$, we can obtain the unique root ζ^* by solving the equation

$$\sum_{j=1}^{g}\frac{a_j}{b_j+\zeta}-1=0. \tag{23}$$

Therefore, the $(r+1)$th iterate

$$\pi_j^{(r+1)}=\frac{a_j}{b_j+\zeta^*}. \tag{24}$$

Furthermore, by solving $\nabla Q\left(\boldsymbol{\pi}^{(r+1)},\boldsymbol{\alpha},\boldsymbol{\beta}^{(r)},\boldsymbol{\sigma}^2;\boldsymbol{\Psi}^{(r)}\right)=\mathbf{0}$, we have the updates

$$\alpha_j^{(r+1)}=\frac{\sum_{i=1}^{n}\tau_{ij}^{(r)}\delta_{ij}^{(r)}\left(y_i-\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j^{(r)}\right)}{\sum_{i=1}^{n}\tau_{ij}^{(r)}\delta_{ij}^{(r)}} \tag{25}$$

and

$$\sigma_j^{2(r+1)}=\frac{2\sum_{i=1}^{n}\tau_{ij}^{(r)}\delta_{ij}^{(r)}\left(y_i-\alpha_j^{(r+1)}-\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j^{(r)}\right)^2}{\sum_{i=1}^{n}\tau_{ij}^{(r)}}. \tag{26}$$

Similarly, for the parameter β_{jt} in *k*th group, we obtain the updated formula

$$\beta_{jt}^{(r+1)}=\frac{\sum_{i=1}^{n}2\tau_{ij}^{(r)}\delta_{ij}^{(r)}x_{it}\left(y_i-\alpha_j^{(r+1)}-\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j^{(r)}+px_{it}\beta_{jt}^{(r)}\right)}{n\pi_j^{(r+1)}\sigma_j^{2(r+1)}\left(\lambda_{jt1}/\eta_{jt}^{(r)}+\lambda_{jk2}/\gamma_{jk}^{(r)}\right)+2p\sum_{i=1}^{n}\tau_{ij}^{(r)}\delta_{ij}^{(r)}x_{it}^2} \tag{27}$$

by solving $\nabla Q\left(\boldsymbol{\pi}^{(r+1)},\boldsymbol{\alpha}^{(r+1)},\boldsymbol{\beta},\boldsymbol{\sigma}^{2(r+1)};\boldsymbol{\Psi}^{(r)}\right)=\mathbf{0}$.

Based on the above, we propose the following *EM* algorithm.

1) Choose an initial value $\boldsymbol{\Psi}^{(0)}=\left(\pi_1^{(0)},\alpha_1^{(0)},\boldsymbol{\beta}_1^{(0)},\sigma_1^{2(0)},\cdots,\pi_g^{(0)},\alpha_g^{(0)},\boldsymbol{\beta}_g^{(0)},\sigma_g^{2(0)}\right)$.

2) *E*-Step: at the $(r+1)$th iteration, calculate $\tau_{ij}^{(r)}$ and $\delta_{ij}^{(r)}$ by (15) and (16).

3) *M*-Step: at the $(r+1)$th iteration, update π_j, α_j, σ_j^2 and β_{jt} by (24), (25), (26) and (27).

4) Repeat *E*-Step and *M*-Step until convergence is obtained.

Note that if a perfect least absolute deviation (LAD) fit occurs in *EM* algorithm, *i.e.* $y_i-\alpha_j^{(r)}-\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j^{(r)}\approx 0$ for some *i*, *j* and *r*. As a result, $\delta_{ij}^{(r+1)}$ will become very large and numerical instability. In this article, we simply introduce a hard threshold to control the extremely small LAD residuals, $\delta_{ij}^{(r+1)}$ will be assigned a value of 10^6 when the perfect LAD fit occurs.

2.4. Convergence Analysis

The *EM* algorithm is iterated until some convergence criterion is met. Let *tol* be a small tolerance constant and M be the maximum iterations for the proposed algorithm, we believe the algorithm has converged to an ideal state when

$$\left|F_{n,\varepsilon}\left(\boldsymbol{\Psi}^{(r+1)}\right)-F_{n,\varepsilon}\left(\boldsymbol{\Psi}^{(r)}\right)\right|<tol, \tag{28}$$

or the iterations over the maximum iterations M. See [17] for details regarding the relative merits of convergence criteria.

According to Dempster *et al.* [2], each iteration of the E step and M step of EM algorithm monotonically non-decreases the objective function (8), *i.e.* $F_{n,\varepsilon}\left(\boldsymbol{\Psi}^{(r+1)}\right)-F_{n,\varepsilon}\left(\boldsymbol{\Psi}^{(r)}\right)\geq 0$, for all $r\geq 0$. Moreover, Wu [18] proved that if the EM sequence $\left\{\boldsymbol{\Psi}^{(r)}\right\}$ coverges to some point $\boldsymbol{\Psi}^{*}$, $\boldsymbol{\Psi}^{*}$ is a stationary point of (8) under some general conditions for $F_{n,\varepsilon}\left(\boldsymbol{\Psi}\right)$ and $E\left[F_{n,\varepsilon}\left(\boldsymbol{\Psi};\boldsymbol{D}\right)|\,\boldsymbol{S},\boldsymbol{\Psi}^{(r)}\right]$. Given the facts above, in this article, we run multiple times from different initializations $\boldsymbol{\Psi}^{(0)}$ in order to obtain an appropriate limit point.

3. Asymptotic Properties

For the regression coefficient vector $\boldsymbol{\beta}_j$ in jth component, we can separate it into $\boldsymbol{\beta}_j=\left(\boldsymbol{\beta}_{1j}^{\mathrm{T}},\boldsymbol{\beta}_{2j}^{\mathrm{T}}\right)^{\mathrm{T}}$, where $\boldsymbol{\beta}_{1j}$ is the set of non-zero effects and $\boldsymbol{\beta}_{2j}$ is the set of zero effects. Naturally, we decompose the parameter $\boldsymbol{\Psi}=\left(\boldsymbol{\Psi}_1,\boldsymbol{\Psi}_2\right)$ such that $\boldsymbol{\Psi}_2$ contains all zero effects, namely $\boldsymbol{\beta}_{2j}$, $j=1,\cdots,g$. The true parameter is denoted as $\boldsymbol{\Psi}_0$ and the elements of $\boldsymbol{\Psi}_0$ are denoted with a subscript, such as β_{0jt}.

For the purpose of easy discussion, we define $a_n=\max\limits_{j,t}\left\{\lambda_{jt1}:\beta_{0jt}\neq 0\right\}$, $a_n^{*}=\max\limits_{j,k}\left\{\lambda_{jk2}:\boldsymbol{\beta}_{0j}^{[k]}\neq\mathbf{0}\right\}$, $b_n=\min\limits_{j,t}\left\{\lambda_{jt1}:\beta_{0jt}=0\right\}$, $b_n^{*}=\min\limits_{j,k}\left\{\lambda_{jk2}:\boldsymbol{\beta}_{0j}^{[k]}=\mathbf{0}\right\}$. Furthermore, we let $f\left(\boldsymbol{z};\boldsymbol{\Psi}\right)$ be the joint density function of $\boldsymbol{z}=\left(\boldsymbol{x},y\right)$ and $\boldsymbol{\Omega}$ be an open parameter space. In order to prove the asymptotic properties of the proposed algorithm, some regularity conditions on the joint distribution of $\boldsymbol{z}$ are also required.

A1. The density $f\left(\boldsymbol{z};\boldsymbol{\Psi}\right)$ has common support in $\boldsymbol{z}$ for all $\boldsymbol{\Psi}\in\boldsymbol{\Omega}$ and $f\left(\boldsymbol{z};\boldsymbol{\Psi}\right)$ is identifiable in $\boldsymbol{\Psi}$ up to a permutation of the components of the mixture.

A2. For each $\boldsymbol{\Psi}\in\boldsymbol{\Omega}$, the density $f\left(\boldsymbol{z};\boldsymbol{\Psi}\right)$ admits third partial derivatives with respect to $\boldsymbol{\Psi}$ for almost all $\boldsymbol{z}$.

A3. For each $\boldsymbol{\Psi}_0\in\boldsymbol{\Omega}$, there are functions $M_1\left(\boldsymbol{z}\right)$ and $M_2\left(\boldsymbol{z}\right)$ (possibly depending on $\boldsymbol{\Psi}_0$) such that for $\boldsymbol{\Psi}$ in a neighborhood of $N\left(\boldsymbol{\Psi}_0\right)$,

$$\left|\frac{\partial f\left(\boldsymbol{z};\boldsymbol{\Psi}\right)}{\partial\Psi_u}\right|\leq M_1\left(\boldsymbol{z}\right),\quad \left|\frac{\partial^2 f\left(\boldsymbol{z};\boldsymbol{\Psi}\right)}{\partial\Psi_u\partial\Psi_v}\right|\leq M_1\left(\boldsymbol{z}\right),\quad \left|\frac{\partial^3 \log f\left(\boldsymbol{z};\boldsymbol{\Psi}\right)}{\partial\Psi_u\partial\Psi_v\partial\Psi_w}\right|\leq M_2\left(\boldsymbol{z}\right)$$

such that $\int M_1\left(\boldsymbol{z}\right)\mathrm{d}\boldsymbol{z}<\infty$ and $\int M_2\left(\boldsymbol{z}\right)f\left(\boldsymbol{z};\boldsymbol{\Psi}\right)\mathrm{d}\boldsymbol{z}<\infty$.

A4. The Fisher information matrix

$$I(\Psi)=E\left\{\left[\frac{\partial}{\partial\Psi}\log f(\mathbf{z};\Psi)\right]\left[\frac{\partial}{\partial\Psi}\log f(\mathbf{z};\Psi)\right]^{\mathrm{T}}\right\}$$

is finite and positive definite for each $\Psi\in\Omega$.

Theorem 1. *Let* $\mathbf{z}_i=(\mathbf{x}_i,y_i)$, $i=1,\cdots,n$, *be a random sample from the joint density function* $f(\mathbf{z};\Psi)$ *that satisfies the regularity conditions A*1*-A*4. *Suppose that* $\sqrt{n}a_n\to^p 0$ *and* $\sqrt{n}a_n^*\to^p 0$, *as* $n\to 0$, *then there is a local maximizer* $\hat{\Psi}_n$ *of the model* (5) *for which*

$$\left\|\hat{\Psi}_n-\Psi_0\right\|=O_p\left\{n^{-1/2}\right\},$$

where $\to^p$ represents convergence in probability.

Proof. Let $r_n=n^{-1/2}$. It suffices that for any given $\varepsilon>0$, there is a constant M_ε such that

$$\lim_{n\to\infty}P\left\{\sup_{\|\boldsymbol{\mu}\|=M_\varepsilon}F_n(\Psi_0+r_n\boldsymbol{\mu})<F_n(\Psi_0)\right\}\geq 1-\varepsilon. \tag{29}$$

Now, there is a local maximum in $\left\{\Psi_0+r_n\boldsymbol{\mu}:\|\boldsymbol{\mu}\|\leq M_\varepsilon\right\}$ with large probability, and this local maximizer $\hat{\Psi}_n$ satisfies $\left\|\hat{\Psi}_n-\Psi_0\right\|=O_p(r_n)$. Then we let

$$\begin{aligned}\Delta_n(\boldsymbol{\mu})&=F_n(\Psi_0+r_n\boldsymbol{\mu})-F_n(\Psi_0)\\&=\left[L_n(\Psi_0+r_n\boldsymbol{\mu})-L_n(\Psi_0)\right]-\left[P_n(\Psi_0+r_n\boldsymbol{\mu})-P_n(\Psi_0)\right].\end{aligned} \tag{30}$$

Without loss of generality, we assume that the first d_j coefficients of $\boldsymbol{\beta}_{0j}$ are non-zero and the first K_j groups contain all non-zero effects of $\boldsymbol{\beta}_{0j}$, where $\boldsymbol{\beta}_{0j}$ is the true regression coefficient vector in the jth component of the mixture regression model. Hence, we have

$$P_n(\Psi_0)=n\sum_{j=1}^{g}\pi_j\sum_{t=1}^{d_j}\lambda_{jt1}\left|\beta_{0jt}\right|+n\sum_{j=1}^{g}\pi_j\sum_{k=1}^{K_j}\lambda_{jk2}\left\|\boldsymbol{\beta}_{0j}^{[k]}\right\|. \tag{31}$$

Since $P_n(\Psi_0+r_n\boldsymbol{\mu})$ is a sum of non-negative terms, removing terms corresponding to zero effects makes it smaller,

$$\begin{aligned}\Delta_n(\boldsymbol{\mu})\leq{}&L_n(\Psi_0+r_n\boldsymbol{\mu})-L_n(\Psi_0)\\&-n\sum_{j=1}^{g}\pi_j\sum_{t=1}^{d_j}\lambda_{jt1}\left[\left|\beta_{0jt}+r_n\mu_{jt}\right|-\left|\beta_{0jt}\right|\right]\\&-n\sum_{j=1}^{g}\pi_j\sum_{k=1}^{K_j}\lambda_{jk2}\left[\left\|\boldsymbol{\beta}_{0j}^{[k]}+r_n\boldsymbol{\mu}_j^{[k]}\right\|-\left\|\boldsymbol{\beta}_{0j}^{[k]}\right\|\right].\end{aligned} \tag{32}$$

By Taylor's expansion, triangular inequality and arithmetic-geometric mean inequality,

$$\begin{aligned}&L_n(\Psi_0+r_n\boldsymbol{\mu})-L_n(\Psi_0)=n^{-1/2}L_n'(\Psi_0)^{\mathrm{T}}\boldsymbol{\mu}-\frac{1}{2}\left[\boldsymbol{\mu}^{\mathrm{T}}I(\Psi_0)\boldsymbol{\mu}\right]\left\{1+o_p(1)\right\},\\&\left|n\sum_{j=1}^{g}\pi_j\sum_{t=1}^{d_j}\lambda_{jt1}\left[\left|\beta_{0jt}+r_n\mu_{jt}\right|-\left|\beta_{0jt}\right|\right]\right|\leq\sqrt{n}a_n\sqrt{\sum_{j=1}^{g}d_j}\,\|\boldsymbol{\mu}\|,\\&\left|n\sum_{j=1}^{g}\pi_j\sum_{k=1}^{K_j}\lambda_{jk2}\left[\left\|\boldsymbol{\beta}_{0j}^{[k]}+r_n\boldsymbol{\mu}_j^{[k]}\right\|-\left\|\boldsymbol{\beta}_{0j}^{[k]}\right\|\right]\right|\leq\sqrt{n}a_n^*\sqrt{\sum_{j=1}^{g}K_j}\,\|\boldsymbol{\mu}\|.\end{aligned} \tag{33}$$

Regularity conditions indicate that $L_n'(\Psi_0)=O_p\left(n^{1/2}\right)$ and $I(\Psi_0)$ is posi-

tive definite, and it is not difficult to find that the sign of $\Delta_n(\boldsymbol{\mu})$ is completely determined by $-\frac{1}{2}\left[\boldsymbol{\mu}^{\mathrm{T}} I(\boldsymbol{\Psi}_0)\boldsymbol{\mu}\right]\{1+o_p(1)\}$. Therefore, for any given $\varepsilon>0$, there is a sufficiently large M_ε such that

$$\lim_{n\to\infty} P\left\{\sup_{\|\boldsymbol{\mu}\|=M_\varepsilon} \Delta_n(\boldsymbol{\mu})<0\right\}>1-\varepsilon, \tag{34}$$

which implies (29), this completes the proof.

Theorem 2. *Suppose that the conditions given in Theorem* 1 *and g is known*, $\sqrt{n}a_n\to^p 0$, $\sqrt{n}a_n^*\to^p 0$, $\sqrt{n}b_n\to^p \infty$ *and* $\sqrt{n}b_n^*\to^p \infty$, *as* $n\to 0$. *Then, for any* $\sqrt{n}$ *-consistent maximum penalized log-likelihood estimator* $\hat{\boldsymbol{\Psi}}_n$, *we have the following*:

1) Sparsity: As $n\to\infty$, $P\left(\hat{\boldsymbol{\beta}}_{2j}=\mathbf{0}\right)\to 0$, $j=1,\cdots,g$.

2) Asymptotic normality:

$$\sqrt{n}\left\{\left[\frac{P_n''(\boldsymbol{\Psi}_{01})}{n}+I_1(\boldsymbol{\Psi}_{01})\right]\left(\hat{\boldsymbol{\Psi}}_1-\boldsymbol{\Psi}_{01}\right)+\frac{P_n'(\boldsymbol{\Psi}_{01})}{n}\right\}\to^d N\left(\mathbf{0}, I_1(\boldsymbol{\Psi}_{01})\right),$$

where $\to^d$ denotes convergence in distribution and $I_1(\boldsymbol{\Psi}_{01})$ is a Fisher information when all zero effects are removed.

Proof. In order to prove the sparsity of Theorem 2, we consider the partition $\boldsymbol{\Psi}=(\boldsymbol{\Psi}_1,\boldsymbol{\Psi}_2)$ and let $\left(\hat{\boldsymbol{\Psi}}_1,\mathbf{0}\right)$ is the maximizer of the penalized log-likelihood function $F_n(\boldsymbol{\Psi}_1,\mathbf{0})$, which is regarded as a function of $\boldsymbol{\Psi}_1$. It suffices to show that in the neighborhood $\|\boldsymbol{\Psi}-\boldsymbol{\Psi}_0\|=O\left(n^{-1/2}\right)$, there is the probability $P\left(F_n(\boldsymbol{\Psi}_1,\boldsymbol{\Psi}_2)-F_n\left(\hat{\boldsymbol{\Psi}}_1,\mathbf{0}\right)<0\right)\to 1$ as $n\to\infty$. Then we have

$$\begin{aligned} F_n(\boldsymbol{\Psi}_1,\boldsymbol{\Psi}_2)-F_n\left(\hat{\boldsymbol{\Psi}}_1,\mathbf{0}\right) &= \left[F_n(\boldsymbol{\Psi}_1,\boldsymbol{\Psi}_2)-F_n(\boldsymbol{\Psi}_1,\mathbf{0})\right]+\left[F_n(\boldsymbol{\Psi}_1,\mathbf{0})-F_n\left(\hat{\boldsymbol{\Psi}}_1,\mathbf{0}\right)\right] \\ &\le \left[F_n(\boldsymbol{\Psi}_1,\boldsymbol{\Psi}_2)-F_n(\boldsymbol{\Psi}_1,\mathbf{0})\right]. \end{aligned} \tag{35}$$

On the other hand,

$$F_n(\boldsymbol{\Psi}_1,\boldsymbol{\Psi}_2)-F_n(\boldsymbol{\Psi}_1,\mathbf{0})=\left[L_n(\boldsymbol{\Psi}_1,\boldsymbol{\Psi}_2)-L_n(\boldsymbol{\Psi}_1,\mathbf{0})\right]-\left[P_n(\boldsymbol{\Psi}_1,\boldsymbol{\Psi}_2)-P_n(\boldsymbol{\Psi}_1,\mathbf{0})\right]. \tag{36}$$

By the mean value theorem,

$$L_n(\boldsymbol{\Psi}_1,\boldsymbol{\Psi}_2)-L_n(\boldsymbol{\Psi}_1,\mathbf{0})=\left[\frac{\partial L_n(\boldsymbol{\Psi}_1,\boldsymbol{\xi})}{\partial\boldsymbol{\Psi}_2}\right]^{\mathrm{T}}\boldsymbol{\Psi}_2 \tag{37}$$

for some $\|\boldsymbol{\xi}\|\le\|\boldsymbol{\Psi}_2\|=O\left(n^{-1/2}\right)$. By the mean value theorem and regularity condition *A*3, we can get

$$\begin{aligned} &\left\|\frac{\partial L_n(\boldsymbol{\Psi}_1,\boldsymbol{\xi})}{\partial\boldsymbol{\Psi}_2}-\frac{\partial L_n(\boldsymbol{\Psi}_{01},\mathbf{0})}{\partial\boldsymbol{\Psi}_2}\right\| \\ &\le\left\|\frac{\partial L_n(\boldsymbol{\Psi}_1,\boldsymbol{\xi})}{\partial\boldsymbol{\Psi}_2}-\frac{\partial L_n(\boldsymbol{\Psi}_1,\mathbf{0})}{\partial\boldsymbol{\Psi}_2}\right\|+\left\|\frac{\partial L_n(\boldsymbol{\Psi}_1,\mathbf{0})}{\partial\boldsymbol{\Psi}_2}-\frac{\partial L_n(\boldsymbol{\Psi}_{01},\mathbf{0})}{\partial\boldsymbol{\Psi}_2}\right\| \\ &\le\left[\sum_{i=1}^n M_1(\mathbf{z}_i)\right]\|\boldsymbol{\xi}\|+\left[\sum_{i=1}^n M_1(\mathbf{z}_i)\right]\|\boldsymbol{\Psi}_1-\boldsymbol{\Psi}_{01}\|=\{\|\boldsymbol{\xi}\|+\|\boldsymbol{\Psi}_1-\boldsymbol{\Psi}_{01}\|\}O_p(n)=O_p\left(n^{1/2}\right). \end{aligned} \tag{38}$$

Here $\boldsymbol{\Psi}_{01}$ is a subvector of $\boldsymbol{\Psi}_0$ with all zero regression coefficients removed.

Regularity conditions imply that $\partial L_n\left(\boldsymbol{\Psi}_{01},\mathbf{0}\right)/\partial\boldsymbol{\Psi}_2 = O_p\left(n^{1/2}\right)$, therefore $\partial L_n\left(\boldsymbol{\Psi}_1,\boldsymbol{\xi}\right)/\partial\boldsymbol{\Psi}_2 = O_p\left(n^{1/2}\right)$. In this case, we have

$$L_n\left(\boldsymbol{\Psi}_1,\boldsymbol{\Psi}_2\right) - L_n\left(\boldsymbol{\Psi}_1,\mathbf{0}\right) = O_p\left(n^{1/2}\right)\sum_{j=1}^{g}\sum_{t=d_j+1}^{p}\left|\beta_{jt}\right| \tag{39}$$

for large *n*. And for the penalized function $P_n\left(\boldsymbol{\Psi}\right)$,

$$\begin{aligned} &P_n\left(\boldsymbol{\Psi}_1,\boldsymbol{\Psi}_2\right) - P_n\left(\boldsymbol{\Psi}_1,\mathbf{0}\right) \\ &\geq \sum_{j=1}^{g}\pi_j\sum_{t=d_j+1}^{p}\sqrt{n}b_n\sqrt{n}\left|\beta_{jt}\right| + \sum_{j=1}^{g}\pi_j\sum_{k=K_j+1}^{K}\sqrt{n}b_n^*\sqrt{n}\left\|\boldsymbol{\beta}_j^{[k]}\right\|. \end{aligned} \tag{40}$$

Since $\sqrt{n}b_n \to \infty$ and $\sqrt{n}b_n^* \to \infty$ as $n \to \infty$, we have

$$\left[L_n\left(\boldsymbol{\Psi}_1,\boldsymbol{\Psi}_2\right) - L_n\left(\boldsymbol{\Psi}_1,\mathbf{0}\right)\right] - \left[P_n\left(\boldsymbol{\Psi}_1,\boldsymbol{\Psi}_2\right) - P_n\left(\boldsymbol{\Psi}_1,\mathbf{0}\right)\right] < 0 \tag{41}$$

with probability to one as $n \to \infty$. This completes the proof of the sparsity.

For the asymptotic normality of Theorem 2, we still use the same argument as in Theorem 1 and consider $F_n\left(\boldsymbol{\Psi}_1,\mathbf{0}\right)$ is a function of $\boldsymbol{\Psi}_1$, there is a $\sqrt{n}$ -consistent local maximizer of this function, say $\hat{\boldsymbol{\Psi}}_1$, that satisfies

$$\frac{\partial F_n\left(\hat{\boldsymbol{\Psi}}_n\right)}{\partial\boldsymbol{\Psi}_1} = \left\{\frac{\partial L_n\left(\boldsymbol{\Psi}\right)}{\partial\boldsymbol{\Psi}_1} - \frac{\partial P_n\left(\boldsymbol{\Psi}\right)}{\partial\boldsymbol{\Psi}_1}\right\}_{\hat{\boldsymbol{\Psi}}_n=\left(\hat{\boldsymbol{\Psi}}_1,\mathbf{0}\right)} = \mathbf{0}. \tag{42}$$

By the Taylor's expansion,

$$\begin{aligned} \left.\frac{\partial L_n\left(\boldsymbol{\Psi}\right)}{\partial\boldsymbol{\Psi}_1}\right|_{\hat{\boldsymbol{\Psi}}_n=\left(\hat{\boldsymbol{\Psi}}_1,\mathbf{0}\right)} &= \frac{\partial L_n\left(\boldsymbol{\Psi}_{01}\right)}{\partial\boldsymbol{\Psi}_1} + \left\{\frac{\partial^2 L_n\left(\boldsymbol{\Psi}_{01}\right)}{\partial\boldsymbol{\Psi}_1\partial\boldsymbol{\Psi}_1^{\mathrm{T}}} + o_p\left(n\right)\right\}\left(\hat{\boldsymbol{\Psi}}_1 - \boldsymbol{\Psi}_{01}\right), \\ \left.\frac{\partial P_n\left(\boldsymbol{\Psi}\right)}{\partial\boldsymbol{\Psi}_1}\right|_{\hat{\boldsymbol{\Psi}}_n=\left(\hat{\boldsymbol{\Psi}}_1,\mathbf{0}\right)} &= P_n'\left(\boldsymbol{\Psi}_{01}\right) + \left\{P_n''\left(\boldsymbol{\Psi}_{01}\right) + o_p\left(n\right)\right\}\left(\hat{\boldsymbol{\Psi}}_1 - \boldsymbol{\Psi}_{01}\right). \end{aligned} \tag{43}$$

Substituting into (42), we have

$$\left\{P_n''\left(\boldsymbol{\Psi}_{01}\right) - \frac{\partial^2 L_n\left(\boldsymbol{\Psi}_{01}\right)}{\partial\boldsymbol{\Psi}_1\partial\boldsymbol{\Psi}_1^{\mathrm{T}}} + o_p\left(n\right)\right\}\left(\hat{\boldsymbol{\Psi}}_1 - \boldsymbol{\Psi}_{01}\right) = \frac{\partial L_n\left(\boldsymbol{\Psi}_{01}\right)}{\partial\boldsymbol{\Psi}_1} - P_n'\left(\boldsymbol{\Psi}_{01}\right). \tag{44}$$

In addition, regularity conditions imply that

$$-\frac{1}{n}\frac{\partial^2 L_n\left(\boldsymbol{\Psi}_{01}\right)}{\partial\boldsymbol{\Psi}_1\partial\boldsymbol{\Psi}_1^{\mathrm{T}}} = I_1\left(\boldsymbol{\Psi}_{01}\right) + o_p\left(\mathbf{1}\right), \quad \frac{1}{\sqrt{n}}\frac{\partial L_n\left(\boldsymbol{\Psi}_{01}\right)}{\partial\boldsymbol{\Psi}_1} \to^d N\left(\mathbf{0}, I_1\left(\boldsymbol{\Psi}_{01}\right)\right). \tag{45}$$

Finally, we can get

$$\sqrt{n}\left\{\left[\frac{P_n''\left(\boldsymbol{\Psi}_{01}\right)}{n} + I_1\left(\boldsymbol{\Psi}_{01}\right)\right]\left(\hat{\boldsymbol{\Psi}}_1 - \boldsymbol{\Psi}_{01}\right) + \frac{P_n'\left(\boldsymbol{\Psi}_{01}\right)}{n}\right\} \to^d N\left(\mathbf{0}, I_1\left(\boldsymbol{\Psi}_{01}\right)\right). \tag{46}$$

by the Slutsky's theorem. This completes the proof of the asymptotic normality.

Now, we know that, as long as the conditions $\sqrt{n}a_n \to^p 0$, $\sqrt{n}a_n^* \to^p 0$, $\sqrt{n}b_n \to^p \infty$ and $\sqrt{n}b_n^* \to^p \infty$ are satisfied when $n \to \infty$, the conclusions of theorem 1 and theorem 2 are tenable. Since the estimator $\tilde{\boldsymbol{\beta}}$ based on the Lasso penalty, it can be $\sqrt{n}$ -consistent. Then, for any $j \in \left(1,2,\cdots,g\right)$, we have $\left|\tilde{\beta}_{jt}\right| \to^p \left|\beta_{0jt}\right|$ for $t \leq d_j$ and $\left|\tilde{\beta}_{jt}\right| = O_p\left(1/\sqrt{n}\right)$ for $t > d_j$. Based on the fact

$\lambda_{jt1} = \lambda_{j1}\omega_{jt}$, we can take $\lambda_{j1} \in \left(n^{-1}, n^{-1/2}\right)$ which satisfies the $\sqrt{n}a_n \to^p 0$ and $\sqrt{n}b_n \to^p \infty$. Similarly, for $\lambda_{jk2} = \lambda_{j2}\xi_{jk}$ with $\left\|\tilde{\boldsymbol{\beta}}_j^{[k]}\right\| \to^p \left\|\boldsymbol{\beta}_{0j}^{[k]}\right\|$ for $k \le K_j$ and $\left\|\tilde{\boldsymbol{\beta}}_j^{[k]}\right\| = O_p\left(1/\sqrt{n}\right)$ for $k > K_j$, we also take $\lambda_{j2} \in \left(n^{-1}, n^{-1/2}\right)$ to satisfy the $\sqrt{n}a_n^* \to^p 0$ and $\sqrt{n}b_n^* \to^p \infty$.

4. Tuning Parameters and Components Selection

In this section, we need to solve two problems. One concerns the number of components g and the other problem is the selection of the tuning parameters $\boldsymbol{\lambda} = \left(\lambda_{11}, \lambda_{12}, \cdots, \lambda_{g1}, \lambda_{g2}\right)$. Until now, there is little theoretical support for the selection of these hyper parameters. In former literatures, the cross validation [19] and the generalized cross validation [20] provided some effective guidances for these problems. Grün and Leisch [21] and Nguyen and McLachlan [22] indicated that the Bayesian information criterion (BIC) has a good performance in solving these problems. In this paper, we still use the BIC,

$$\mathrm{BIC} = -2L_n\left(\hat{\boldsymbol{\Psi}}_n\right) + \left(3g - 1 + \sum_{j=1}^{g} d_j\right)\log(n), \tag{47}$$

where d_j the number of non-zero regression coefficients in the jth regression model.

Suppose that there is a set of parameter combinations $\left\{(g_1, \boldsymbol{\lambda}_1), \cdots, (g_S, \boldsymbol{\lambda}_S)\right\}$. For each parameter combination $(g_s, \boldsymbol{\lambda}_s)$, $s = 1, \cdots, S$, we can obtain the parameter estimate $\hat{\boldsymbol{\Psi}}_{n,s}$ by the proposed algorithm, and there is a BIC_s which depends on corresponding $\hat{\boldsymbol{\Psi}}_{n,s}$. Finally, we set the best parameter combination $(g, \boldsymbol{\lambda}) = \left(g_{s^*}, \boldsymbol{\lambda}_{s^*}\right)$ for our robust mixture model, where $s^* = \arg\min_s \mathrm{BIC}_s$.

5. Numerical Simulation

To quantify the performance of our proposed robust mixture regression model based on adaptive sparse group Lasso (adaSGL-RMR), we design a numerical simulation and generate sample data $\left(\boldsymbol{x}_i, y_i\right)_{i=1}^n$ from the following mixture regression model

$$y = \begin{cases} \alpha_1 + \sum_{k=1}^{K} \boldsymbol{x}^{[k]\mathrm{T}} \boldsymbol{\beta}_1^{[k]} + \varepsilon_1, & \text{if } Z = 1; \\ \alpha_2 + \sum_{k=1}^{K} \boldsymbol{x}^{[k]\mathrm{T}} \boldsymbol{\beta}_2^{[k]} + \varepsilon_2, & \text{if } Z = 2, \end{cases} \tag{48}$$

where Z is a component indicator. There are $K = 6$ groups and each group consists of 5 covariates, covariates within the same group are correlated, whereas those in different groups are uncorrelated. In order to generate the covariates $X_1, \cdots, X_{30}$, we first generate 30 random variables, $R_1, \cdots, R_{30}$, independently from $N(0,1)$, then obtain C_k, $k = 1, \cdots, 6$, from a multivariate normal distribution with mean zero and $cov\left(C_{k,t_1}, C_{k,t_2}\right) = 0.8^{|t_1 - t_2|}$. The generation of the covariates $X_1, \cdots, X_{30}$ as following:

$$X_{5(k-1)+t} = \frac{C_{k,t} + R_{5(k-1)+t}}{\sqrt{2}}, \quad 1 \le k \le 6,\ 1 \le t \le 5. \tag{49}$$

The model parameters include $\pi_1 = \pi_2 = 0.5$, $\alpha_1 = -3$, $\alpha_2 = 3$, $\boldsymbol{\beta}_1^{[1]} = (2.5, -2, -1.5, 0, 0)^{\mathrm{T}}$, $\boldsymbol{\beta}_1^{[2]} = (1.5, 2.5, -1, 0, 0)^{\mathrm{T}}$, $\boldsymbol{\beta}_2^{[1]} = (0, 0, 2, -1, 1.5)^{\mathrm{T}}$, $\boldsymbol{\beta}_2^{[2]} = (0, 0, 2, 1, -2.5)^{\mathrm{T}}$, $\boldsymbol{\beta}_1^{[3]} = \cdots = \boldsymbol{\beta}_2^{[6]} = (0, \cdots, 0)^{\mathrm{T}}$, $\sigma_1^2 = \sigma_2^2 = 1$. The random error ε_1 and ε_2 are independent and we consider the follow four cases: 1) $\varepsilon_1, \varepsilon_2 \sim N(0,1)$; 2) $\varepsilon_1, \varepsilon_2 \sim$ Laplace distribution with mean 0 and variance 1; 3) $\varepsilon_1, \varepsilon_2 \sim t_3$, t-distribution with 3 degrees of freedom; 4) a mixture normal distribution $\varepsilon_1, \varepsilon_2 \sim 0.95N(0,1) + 0.05N(0, 5^2)$.

We use three methods for comparing. The Gaussian mixture model (GMM) based on Lasso penalty (Lasso-GMM), the GMM model based on adaptive Lasso penalty (adaL-GMM) and the adaSGL-RMR model. The fmr package of the R programming language is used to compute the parameter estimates of Lasso-GMM and adaL-GMM.

In this article, the algorithm is terminated when the change in the penalized complete log-likelihood function is less than 10^{-5} or meets the maximum iterations 10^5. Furthermore, we adopt a threshold value for $\hat{\boldsymbol{\beta}}$ in the consideration of practical situation. To be specific, $\hat{\beta}_{jt}$ will be assigned a value of 0 if $\left|\hat{\beta}_{jt}\right| < 10^{-5}$, for some j and t. To evaluate the performances of variable selection and data fitting, we introduce the average number of selected non-zero variables (nvars) without intercepts, average number of selected non-zero groups (ngroups), frequency of correct identification of group sparsity structures (cgroups), false negative rate (FNR) of missing important predictors, false positive rate (FPR) of selecting unimportant predictors and average value of root mean square errors (RMSE). Here,

$$\mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{TP} + \mathrm{FN}}, \ \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{50}$$

and

$$\mathrm{RMSE} = \sqrt{\frac{1}{g(p+1)}\left(\left\|\boldsymbol{\alpha}_0 - \hat{\boldsymbol{\alpha}}\right\|^2 + \left\|\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}\right\|^2\right)}, \tag{51}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives and FN is the number of false negatives for each fitted model.

As shown in **Table 1**, in case (1) through case (4), Lasso-GMM fails to identify non-zero groups and selects too many unimportant predictors, adaL-GMM is inclined to select less unimportant predictors and achieves higher cgroups at the cost of ignoring some important predictors. As a contrast, adaSGL-RMR has a better performance in variable selection and RMSE indicates that the parameter estimates of our algorithm are closer to the true parameters of mixture regression model. Moreover, the simulation results clearly show that adaSGL-RMR still maintain its superiority in identifying the group sparsity structures when the distribution of random errors has a heavy tail. Therefore, no matter from the model complexity or the goodness of fit to the data, our proposed method is more competitive than other methods.

Table 1. Results of Lasso-GMM, adaL-GMM and adaSGL-RMR on 100 replicates, $n = 300$.

$\varepsilon_1, \varepsilon_2$	Model	nvars	ngroups	cgroups	FNR	FPR	RMSE
	True model	12	4	100	0	0	0
$N(0, 1)$	Lasso-GMM	49.71	11.97	2.08	0.00	0.79	0.09
	adaL-GMM	20.38	8.66	51.00	0.01	0.17	0.06
	adaSGL-RMR	13.66	4.00	88.00	0.00	0.03	0.06
Laplace (1)	Lasso-GMM	46.19	11.97	3.00	0.00	0.71	0.12
	adaL-GMM	26.15	9.95	34.83	0.01	0.29	0.10
	adaSGL-RMR	13.47	4.05	89.25	0.00	0.03	0.09
t_3	Lasso-GMM	40.84	11.81	6.83	0.00	0.63	0.15
	adaL-GMM	29.34	10.65	24.67	0.01	0.36	0.12
	adaSGL-RMR	13.67	4.09	88.42	0.00	0.04	0.11
$0.95N(0,1)+0.05N(0,5^2)$	Lasso-GMM	50.26	11.99	1.83	0.00	0.79	0.09
	adaL-GMM	21.02	8.93	46.75	0.01	0.19	0.07
	adaSGL-RMR	13.69	4.00	88.00	0.00	0.04	0.06

6. Real Data Analysis

In this section, we will analyze how NBA players' performances of regular season affect their salaries. We gather salaries for all players in the NBA from the website https://hoopshype.com/salaries/players/2018-2019, during the period from 2018 to 2019. Performance measures for individuals are gathered from the website https://www.foxsports.com/nba/stats?season=2018 in the 2018-2019 regular season, which include scoring, rebounding, assists and defense statistics. By liminating missing data, we obtain a complete dataset, which contains salaries for 248 NBA players and 27 measures of performance.

These performance measures are divided into five groups and covariates in the same group are correlated. Score: points per game (PPG), points per 48 minutes (PTS/48), field goals made per game (FGM/G), field goal attempts per game (FGA/G), 3 point FG made per game (3FGM/G), 3 point FG attempts per game (3FGA/G), free throws made per game (FTM/G), free throw attempts per game (FTA/G), high point total in a single game (HIGH), points per shot (PPS). Rebound: offensive rebounds per game (ORPG), defensive rebounds per game (DRPG), rebounds per 48 minutes (RPG/48), offensive rebound% (OFF REB%), defensive rebound% (DEF REB%), rebound% (REB%). Assist: assists per game (APG), assists per 48 minutes (AST/48), assist% (AST%), turnovers per game (TPG), turnover% (TO%). Steal: steals per game (SPG), steals per 48 minutes (STL/48), steal% (STL%). Block: blocks per game (BPG), blocks per 48 minutes (BLK/48), block% (BLK%).

The matrix $\boldsymbol{X}$ should be column standardized to have mean 0 and variance 1 for avoiding a poor fitting result. Then we use the stepAIC function from *R* package MASS to realize variable selection of the standard linear model via the

BIC, the predicted logged salaries from the stepwise-BIC linear model shows a mean square error (MSE) of 0.60 and adjusted R^2 of 0.42, these terrible results motivate us to conduct further research for this problem. The logged salaries histogram shows multi-modality from Figure 1, it is acceptable to use the mixture regression model for predicting the logged salaries.

For comparison, we run multiple analyses that include three sets of starting parameters for each of $g=(2,3,4)$ models. The predicted results from the $g=2$ adaSGL-RMR model (BIC = 625) have a MSE of 0.11 and adjusted R^2 of 0.90. The predicted results from the $g=3$ adaSGL-RMR model (BIC = 598) have a MSE of 0.05 and adjusted R^2 of 0.95. The predicted results from the $g=4$ adaSGL-RMR model (BIC = 517) have a MSE of 0.04 and adjusted R^2 of 0.96. See Table 2 for more details. These results suggest that the $g=4$ adaSGL-RMR model has the smallest MSE and explains the largest proportion of variance for the logged salaries from the 2018/19 NBA regular season. Moreover, from Figure 2, the predicted densities show a good characterization of the multi-modality in the logged salaries for the adaSGL-RMR models, with the stepwise-BIC linear model not being able to model this.

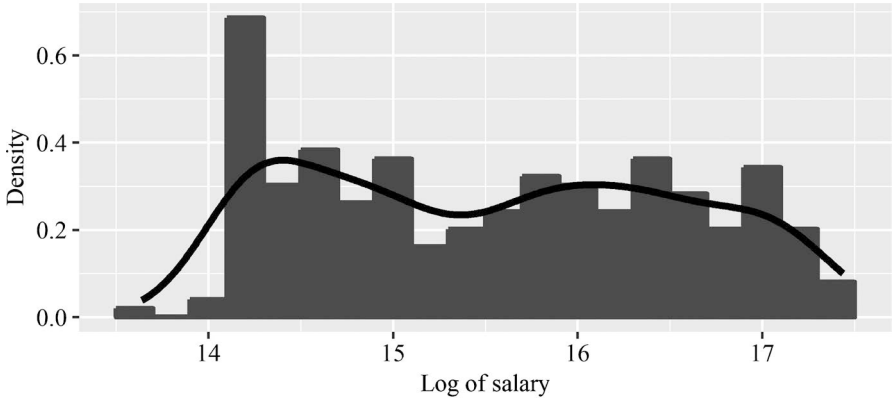

Figure 1. Histogram and density estimate for logged salaries

Table 2. Parameter estimates for NBA salary data.

Covariates	$g=2$		$g=3$			$g=4$			
	comp.1	comp.2	comp.1	comp.2	comp.3	comp.1	comp.2	comp.3	comp.4
π	0.56	0.44	0.17	0.41	0.42	0.19	0.36	0.41	0.04
Intercept	14.93	16.39	14.39	15.18	16.43	14.39	15.21	16.43	17.12
PPG	0.44		0.19	0.50	0.23			0.13	
PTS/48									
FGM/G		0.23					0.18		
FGA/G						0.11	0.49		
3FGM/G								−0.03	
3FGA/G									
FTM/G						0.09			

Continued

FTA/G	0.09					
HIGH						
PPS						
ORPG			0.06			
DRPG						
RPG/48						
OFF REB%		−0.20		−0.07		
DEF REB%					0.07	
REB%		0.31	0.06	0.23		
APG			0.11			
AST/48	0.08					0.06
AST%						
TPG						
TO%						
SPG						
STL/48						
STL%						
BPG						
BLK/48						
BLK%	0.06					

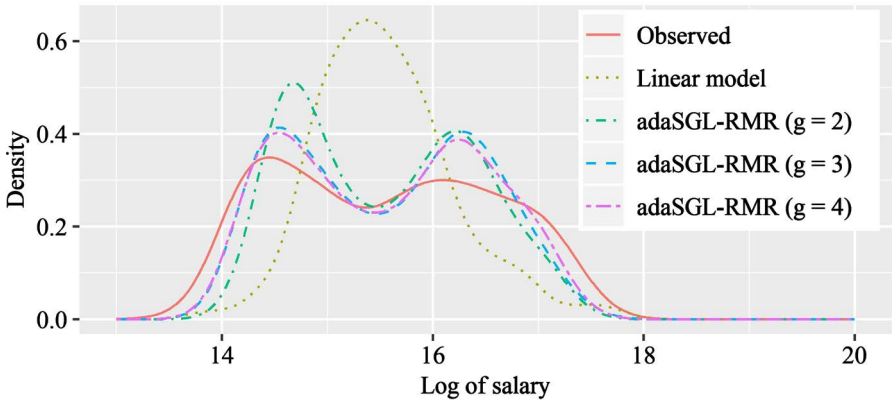

Figure 2. Summary of densities for predicted and observed logged salaries.

7. Conclusion

In this paper, we propose a robust mixture regression model based on a Laplace distribution and consider the adaptive sparse group Lasso for variable selection. Its oracle properties are proved completely in Section 3. In addition, the numerical simulation and real data application show that our method has better performance in parameter estimation and variable selection than other methods. A limitation of this study is that we only consider the mixture regression model

with K no-overlapping groups and ignore the case when there are some overlaps between different groups. In our future work, we will pay more attention to this problem.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Quandt, R.E. (1972) A New Approach to Estimating Switching Regressions. *Journal of the American Statistical Association*, **67**, 306-310. https://doi.org/10.1080/01621459.1972.10482378

[2] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*: *Series B*, **39**, 1-22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

[3] Jiang, W.X. 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. https://doi.org/10.1214/aos/1018031265

[4] McLachlan, G.J. and Peel, D. (2000) Finite Mixture Models. Wiley, New York. https://doi.org/10.1002/0471721182

[5] Frühwirth-Schnatter, S. (2006) Finite Mixture and Markov Switching Models. Springer, New York.

[6] Yao, W.X., Wei, Y. and Yu, C. (2014) Robust Mixture Regression Using the t-Distribution. *Computational Statistics and Data Analysis*, **71**, 116-127. https://doi.org/10.1016/j.csda.2013.07.019

[7] Song, W.X., Yao, W.X. and Xing, Y.R. (2014) Robust Mixture Regression Model Fitting by Laplace Distribution. *Computational Statistics and Data Analysis*, **71**, 128-137. https://doi.org/10.1016/j.csda.2013.06.022

[8] Wu, Q. and Yao, W.X. (2016) Mixtures of Quantile Regressions. *Computational Statistics and Data Analysis*, **93**, 162-176. https://doi.org/10.1016/j.csda.2014.04.014

[9] Khalili, A. and Chen, J.H. (2007) Variable Selection in Finite Mixture of Regression Models. *Journal of the American Statistical Association*, **102**, 1025-1038. https://doi.org/10.1198/016214507000000590

[10] Städler, N., Bühlmann, P. and van de Geer, S. (2010) L1-Penalization for Mixture Regression Models. *TEST*, **19**, 209-256. https://doi.org/10.1007/s11749-010-0197-z

[11] Lloyd-Jones, L.R., Nguyen, H.D. and McLachlan, G.J. (2018) A Globally Convergent Algorithm for Lasso-Penalized Mixture of Linear Regression Models. *Computational Statistics and Data Analysis*, **119**, 19-38. https://doi.org/10.1016/j.csda.2017.09.003

[12] Fang, K.N., Wang, X.Y., Zhang, S.W., Zhu, J.P. and Ma, S.G. (2015) Bi-Level Variable Selection via Adaptive Sparse Group Lasso. *Journal of Statistical Computation and Simulation*, **85**, 2750-2760. https://doi.org/10.1080/00949655.2014.938241

[13] Hunter, D.R. and Li, R.Z. (2005) Variable Selection Using MM Algorithms. *The Annals of Statistics*, **33**, 1617-1642. https://doi.org/10.1214/009053605000000200

[14] Andrews, D.F. and Mallows, C.L. (1974) Scale Mixtures of Normal Distributions.

Journal of the Royal Statistical: *Series B*, **36**, 99-102.
https://doi.org/10.1111/j.2517-6161.1974.tb00989.x

[15] Phillips, R.F. (2002) Least Absolute Deviations Estimation via the EM Algorithm. *Statistics and Computing*, **12**, 281-285. https://doi.org/10.1023/A:1020759012226

[16] Fan, J.Q. and Li, R.Z. (2001) Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, **96**, 1348-1360. https://doi.org/10.1198/016214501753382273

[17] Lange, K. (2013) Optimization. Springer, New York.
https://doi.org/10.1007/978-1-4614-5838-8

[18] Wu, C.F.J. (1983) On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, **11**, 95-103. https://doi.org/10.1214/aos/1176346060

[19] Stone, M. (1974) Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society*: *Series B*, **36**, 111-133.
https://doi.org/10.1111/j.2517-6161.1974.tb00994.x

[20] Craven, P. and Wahba, G. (1978) Smoothing Noisy Data with Spline Functions. *Numerische Mathematik*, **31**, 377-403. https://doi.org/10.1007/BF01404567

[21] Grün, B. and Leisch, F. (2007) Fitting Finite Mixtures of Generalized Linear Regressions in R. *Computational Statistics and Data Analysis*, **51**, 5247-5252.
https://doi.org/10.1016/j.csda.2006.08.014

[22] Nguyen, H.D. and McLachlan, G.J. (2016) Laplace Mixture of Linear Experts. *Computational Statistics and Data Analysis*, **93**, 177-191.
https://doi.org/10.1016/j.csda.2014.10.016