Scientific Research Publishing

Advances in Pure Mathematics, 2022, 12, 109-124
https://www.scirp.org/journal/apm
ISSN Online: 2160-0384
ISSN Print: 2160-0368

Variable Selection for Robust Mixture Regression Model with Skew Scale Mixtures of Normal Distributions

Tingzhu Chen, Wanzhou Ye

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

How to cite this paper: Chen, T.Z. and Ye, W.Z. (2022) Variable Selection for Robust Mixture Regression Model with Skew Scale Mixtures of Normal Distributions. *Advances in Pure Mathematics*, **12**, 109-124.
https://doi.org/10.4236/apm.2022.123010

Received: January 29, 2022
Accepted: March 4, 2022
Published: March 7, 2022

Abstract

In this paper, we propose a robust mixture regression model based on the skew scale mixtures of normal distributions (RMR-SSMN) which can accommodate asymmetric, heavy-tailed and contaminated data better. For the variable selection problem, the penalized likelihood approach with a new combined penalty function which balances the SCAD and l_2 penalty is proposed. The adjusted EM algorithm is presented to get parameter estimates of RMR-SSMN models at a faster convergence rate. As simulations show, our mixture models are more robust than general FMR models and the new combined penalty function outperforms SCAD for variable selection. Finally, the proposed methodology and algorithm are applied to a real data set and achieve reasonable results.

Keywords

Robust Mixture Regression Model, Skew Scale Mixtures of Normal Distributions, EM Algorithm, SCAD Penalty

1. Introduction

In applied statistics, the arc-sine laws for the Wiener process and the skew Brownian motion [1] are widely used in finance if the market is homogeneous. However, the problem of modeling heterogeneous data has been extensively studied in recent years and the Finite Mixture of Regression (FMR) model is an important tool for heterogeneous cases. A large number of applications associate a random response variable Y with covariates $\boldsymbol{x}$ through FMR models and the assumption is that for each observation data point $(\boldsymbol{x}_1, Y_1), \cdots, (\boldsymbol{x}_n, Y_n)$, the regression coefficients are not the same. More details about the FMR model can be found in

[2].

The Gaussian FMR model is the most common FMR model, which assumes that the random error of each subgroup follows the normal distribution. It is well known that using the normal distribution to model data with asymmetric and heavy-tailed behaviors is unsuitable, and the parameter estimates are sensitive to outliers. To overcome the potential shortcomings of Gaussian mixture models, McLachlan *et al.* [3] proposed to replace the mixtures of normal with mixtures of t-distribution which results in a more robust mixture model. Basso *et al.* [4] studied a finite mixture model based on scale mixtures of skew-normal distributions, and Franczak *et al.* [5] proposed a mixture model using shifted asymmetric Laplace distributions which parameterize the skewness as well as the location and the scale.

The problem of variable selection in FMR models has been widely discussed recently. There are generally two types of variable selection methods. One is the optimal subset selection method and the discontinuous penalty method based on the information criterion, including stepwise regression, best subset regression, BIC criterion, AIC criterion and so on. The other is the continuous penalty method. By imposing penalties on the parameters of the objective function, one can select significant variables and obtain the parameter estimates simultaneously. The Least Absolute Shrinkage and Selection Operator (LASSO), elastic net regularization [6], MCP penalty [7] and SCAD penalty [8] are penalty functions for variable selection. We utilize the SCAD and a new penalty function proposed in this paper which balance the SCAD and l_2 penalty to perform variable selection on a robust mixture regression model based on Skew Scale Mixtures of Normal (SSMN) distributions [9] and this robust model can accommodate asymmetric and heavy-tailed data better.

The paper is organized as follows. In Section 2, a robust mixture regression model using the skew scale mixtures of normal distributions (RMR-SSMN) is introduced. Then, variable selection methods with SCAD penalty function and a newly proposed penalty function are presented in Section 3. Section 4 outlines the adjusted EM algorithm for estimating and a BIC method for selecting turning parameters and components. In Section 5, we carry out simulation studies to compare the performances between FMR models and RMR-SSMN models, and show the effect of variable selection with penalty functions. An application to a real data set of the method is discussed in Section 6 and some conclusions are obtained in Section 7.

2. Robust Mixture Regression Model with SSMN Distributions

It is known that the FMR model can model heterogeneous data and the Skew Scale Mixtures of Normal (SSMN) distributions [9] cover both asymmetric and heavy-tailed distributions. Therefore, we propose a robust mixture regression model whose regression errors of components follow SSMN distributions. Unsurpris-

ingly, this model is more robust than general FMR models for heterogeneous cases.

2.1. Skew Scale Mixtures of Normal Distributions

If a random variable Y follows a skew-normal (SN) distribution with location parameter $\mu \in \mathbb{R}$, scale parameter σ^2, and skewness parameter λ, denoted $Y \sim SN\left(\mu, \sigma^2, \lambda\right)$, then its density function is given as follows:

$$f\left(y\right) = 2\phi\left(y; \mu, \sigma^2\right)\Phi\left(\lambda \frac{y-\mu}{\sigma}\right). \tag{1}$$

where $\phi\left(y; \mu, \sigma^2\right)$ and $\Phi\left(y; \mu, \sigma^2\right)$ are the probability density function and the cumulative distribution function of $N\left(\mu, \sigma^2\right)$ calculated at y, respectively.

Note that when $\lambda = 0$, the $SN\left(\mu, \sigma^2, \lambda\right)$ reduces to the $N\left(\mu, \sigma^2\right)$, and as given in [9], the SN distribution's marginal stochastic representation is presented by:

$$Y = \mu + \sigma\left(\delta\left|T_0\right| + \left(1-\delta^2\right)^{\frac{1}{2}} T_1\right), \tag{2}$$

where $\delta = \lambda \Big/ \left(1+\lambda^2\right)^{1/2}$, $T_0 \sim N\left(0,1\right)$ and $T_1 \sim N\left(0,1\right)$ are independent.

Furthermore, if a random variable Y follows a SSMN distribution [9] with location parameter $\mu \in \mathbb{R}$, scale parameter σ^2, and skewness parameter λ, then its probability density function is given by:

$$f\left(y\right) = 2\int_0^\infty \phi\left(y; \mu, \ell\left(u\right)\sigma^2\right) \mathrm{d}H\left(u; \boldsymbol{\tau}\right)\Phi\left(\lambda \frac{y-\mu}{\sigma}\right), \tag{3}$$

where $H\left(u; \boldsymbol{\tau}\right)$ is the cumulative distribution function of U who derived from the parameter vector $\boldsymbol{\tau}$, and U is a positive random variable, and $\ell\left(u\right)$ is a strictly positive function. If the probability density function of the random variable Y is shown as the Equation (3), it can be denoted as:
$Y \sim SSMN\left(\mu, \sigma^2, \lambda, H; \ell\right)$.

For $Y \sim SSMN\left(\mu, \sigma^2, \lambda, H; \ell\right)$, its hierarchical representation has the form as follows:

$$Y \left| U = u \right. \sim SN\left(\mu, \ell\left(u\right)\sigma^2, \lambda\ell^{\frac{1}{2}}\left(u\right)\right),\ U \sim H\left(\boldsymbol{\tau}\right). \tag{4}$$

This paper will consider the following distributions in the SSMN distributions family:

- The skew Student-t-normal distribution (STN) [10] with $U \sim Gamma\left(\nu/2, \nu/2\right)$, $\nu > 0$ and $\ell\left(u\right) = 1/u$, which follows probability density:

$$f\left(y\right) = 2\frac{1}{\sigma\sqrt{\nu\pi}}\frac{\Gamma\left(\left(\nu+1\right)/2\right)}{\Gamma\left(\nu/2\right)}\left(1+\frac{d}{\nu}\right)^{-\left(\frac{\nu+1}{2}\right)}\Phi\left(\lambda\frac{y-\mu}{\sigma}\right), \tag{5}$$

where $d = \left(y-\mu\right)^2 \Big/ \sigma^2$ and $\Gamma\left(.\right)$ is the gamma function. We can obtain that

$U \mid Y = y \sim Gamma\left((v+1)/2,(v+d)/2\right)$.

- The skew contaminated normal distribution (SCN). U is a discrete random variable taking one of two values and $\ell(u)=1/u$. Given the parameter vector $\boldsymbol{\tau}=(v,\gamma)^{\mathrm{T}}$, $0<v<1$, $0<\gamma<1$, the density function of U is $h(u;\boldsymbol{\tau})=v\mathbb{I}_{(u=\gamma)}+(1-v)\mathbb{I}_{(u=1)}$. Naturally get as follows:

$$f(y)=2\left\{v\phi\left(y;\mu,\sigma^2/\gamma\right)+(1-v)\phi\left(y;\mu,\sigma^2\right)\right\}\Phi\left(\lambda\frac{y-\mu}{\sigma}\right). \tag{6}$$

Therefore, the conditional distribution $U \mid Y = y$ can be obtained as:

$$f(u \mid Y=y)=\frac{1}{f_0(y)}\left\{v\phi\left(y;\mu,\sigma^2/\gamma\right)\mathbb{I}_{(u=\gamma)}+(1-v)\phi\left(y;\mu,\sigma^2\right)\mathbb{I}_{(u=1)}\right\}, \tag{7}$$

where $f_0(y)=v\phi\left(y;\mu,\sigma^2/\gamma\right)\mathbb{I}_{(u=\gamma)}+(1-v)\phi\left(y;\mu,\sigma^2\right)\mathbb{I}_{(u=1)}$.

- The skew power-exponential distribution (SPE) has following probability density:

$$f(y)=2\frac{v}{2^{1/2v}\sigma\Gamma(1/2v)}\mathrm{e}^{-d^v/2}\Phi\left(\lambda\frac{y-\mu}{\sigma}\right),\ 0.5<v\le 1, \tag{8}$$

Ferreira *et al.* [9] have proved that $E\left[\ell^{-1}(U)\mid Y=y\right]=vd^{v-1}$.

2.2. Robust Mixture Regression Model with SSMN Distributions

Suppose we have n independent random variables $y_1,y_2,\cdots,y_n$, which are taken from a mixture of SSMN distributions. The conditional density function of the robust mixture regression model with SSMN distributions (RMR-SSMN) which has K components is given by:

$$f(y_i;\boldsymbol{x}_i,\boldsymbol{\Psi})=\sum_{k=1}^{K}\omega_k SSMN\left(y_i;\alpha_k+\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_k,\sigma_k^2,\lambda_k,\boldsymbol{\tau}_k;\ell\right), \tag{9}$$

with covariate vector $\boldsymbol{x}_i\in\mathbb{R}^q$ and q-dimensional unknown regression coefficients vector $\boldsymbol{\beta}_k, k=1,\cdots,K$. $\omega_k, k=1,\cdots,K$ denote the mixing proportions satisfying $\omega_k\ge 0$, $\sum_{k=1}^{K}\omega_k=1$.

$\boldsymbol{\Psi}=\left(\omega_1,\cdots,\omega_{K-1},\alpha_1,\cdots,\alpha_K,\boldsymbol{\beta}_1^{\mathrm{T}},\cdots,\boldsymbol{\beta}_K^{\mathrm{T}},\sigma_1^2,\cdots,\sigma_K^2,\lambda_1,\cdots,\lambda_K,\boldsymbol{\tau}_1^{\mathrm{T}},\cdots,\boldsymbol{\tau}_K^{\mathrm{T}}\right)^{\mathrm{T}}$ is the parameter vector of the model. For convenience, let $\boldsymbol{\omega}=(\omega_1,\cdots,\omega_{K-1})^{\mathrm{T}}$, $\boldsymbol{\alpha}=(\alpha_1,\cdots,\alpha_K)^{\mathrm{T}}$, $\boldsymbol{\beta}=\left(\boldsymbol{\beta}_1^{\mathrm{T}},\cdots,\boldsymbol{\beta}_K^{\mathrm{T}}\right)^{\mathrm{T}}$, $\boldsymbol{\sigma}^2=\left(\sigma_1^2,\cdots,\sigma_K^2\right)^{\mathrm{T}}$, $\boldsymbol{\lambda}=(\lambda_1,\cdots,\lambda_K)^{\mathrm{T}}$, and $\boldsymbol{\tau}^*=\left(\boldsymbol{\tau}_1^{\mathrm{T}},\cdots,\boldsymbol{\tau}_K^{\mathrm{T}}\right)^{\mathrm{T}}$. In this paper, RMR-SSMN models contain the robust mixture regression model with STN distribution (RMR-STN), SCN distribution (RMR-SCN), SPE distribution (RMR-SPE) and SN distribution (RMR-SN).

3. Variable Selection Method

If a component in the q-dimensional explanatory variable $\boldsymbol{x}$ has no significant effect on the response variable y, the regression coefficient of this component estimated by the maximum likelihood method will close to 0 rather than 0. Thus, this covariate is not excluded from the model and makes the model unstable. To avoid this problem, we use a penalized likelihood approach [11] for selecting variables and estimating parameters simultaneously. Let $\{(y_i,\boldsymbol{x}_i);i=1,\cdots,n\}$ be sam-

ple observations from RMR-SSMN models. The log-likelihood function of Ψ is given by:

$$l(\Psi)=\sum_{i=1}^{n}\log\sum_{k=1}^{K}\omega_k SSMN\left(y_i;\alpha_k+\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_k,\sigma_k^2,\lambda_k,\boldsymbol{\tau}_k;\ell\right). \tag{10}$$

Following the idea in [11], we can get the estimates of Ψ by maximizing the penalized log-likelihood function which is defined as:

$$L(\Psi)=l(\Psi)-p(\Psi), \tag{11}$$

with the penalty function is given by:

$$p(\Psi)=n\sum_{k=1}^{K}\omega_k\sum_{j=1}^{q}p_{a_k}\left(\left|\beta_{kj}\right|\right), \tag{12}$$

where $p_{a_k}(.)$ is a nonnegative and non-decreasing function in $\left|\beta_{kj}\right|$ with the turning parameter $a_k, k=1,\cdots,K$. The turning parameter controls the intensity of the penalty for the regression coefficients.

The SCAD penalty has a type of oracle property as discussed in [8]. In this work, we complete the variable selection procedure using the following SCAD penalty function:

$$p_{a_k}\left(\left|\beta_{kj}\right|\right)=\begin{cases}a_k\left|\beta_{kj}\right|, & \left|\beta_{kj}\right|\le a_k\\ -\dfrac{\left(\beta_{kj}^2-2ca_k\left|\beta_{kj}\right|+a_k^2\right)}{2(c-1)}, & a_k<\left|\beta_{kj}\right|\le ca_k\\ \dfrac{a_k^2(c+1)}{2}, & \left|\beta_{kj}\right|>ca_k\end{cases} \tag{13}$$

Meanwhile, inspired by [12], we propose a combined penalty function which balance the SCAD penalty and l_2 penalty. This penalty function by introducing a connection parameter b is more effective in variable selection than directly mixing SCAD and l_2, and the specific form is given by:

$$p_{a_k}\left(\left|\beta_{kj}\right|\right)=\begin{cases}a_k\left[b\left|\beta_{kj}\right|+(1-b)\beta_{kj}^2\right], & \left|\beta_{kj}\right|\le a_k\\ -\dfrac{b\left(\beta_{kj}^2-2ca_k\left|\beta_{kj}\right|+a_k^2\right)}{2(c-1)}+a_k(1-b)\beta_{kj}^2, & a_k<\left|\beta_{kj}\right|\le ca_k\\ \dfrac{a_k^2b(c+1)}{2}+a_k(1-b)\beta_{kj}^2, & \left|\beta_{kj}\right|>ca_k\end{cases} \tag{14}$$

We call this new penalty function as MIXL2-SCAD. Some asymptotic properties of the penalty function are showed in [12], and the constant $a_k>0$ and $c>2$. Following the idea of [8], let $c=3.7$. In particular, the constant b, $0\le b\le 1$ and a_k in MIXL2-SCAD jointly control the speed of contraction of β_{kj}, and when $b=1$, MIXL2-SCAD penalty reduces to the SCAD penalty.

4. Numeric Solutions

The expectation-maximization (EM) algorithm can be applied to mixture regression models based on SSMN distributions for maximizing the penalized log-like-

lihood function. When the M-step of EM is analytically intractable for SSMN distributions, it can be replaced with a sequence of conditional maximization (CM) steps which is derived from ECM algorithm [13]. Furthermore, we also maximize the constrained actual marginal log-likelihood function which called CML steps [14] for simplicity.

4.1. Maximization of the Penalized Log-Likelihood Function

Let us introduce the latent vector $\boldsymbol{Z}_i = (z_{i1}, \cdots, z_{iK})^{\mathrm{T}}$ with the component indicator variable z_{ik} which has the following form:

$$z_{ik} = \begin{cases} 1, & \text{the } i\text{th sample comes from the latent } k\text{th component,} \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

Using the Equations (2) and (4), we can get the following hierarchical representation for the mixture of SSMN distributions.

$$\begin{aligned} &Y_i \mid (T_i = t_i, U_i = u_i, z_{ik} = 1) \sim N\left(\alpha_k + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_k + \frac{\ell(u_i)\sigma_k\lambda_k}{\left(1+\lambda_k^2\ell(u_i)\right)^{1/2}}t_i, \frac{\sigma_k^2\ell(u_i)}{1+\lambda_k^2\ell(u_i)}\right), \\ &U_i \mid z_{ik} = 1 \sim H(\boldsymbol{\tau}_k), \\ &T_i \mid z_{ik} = 1 \sim TN(0,1;(0,\infty)), \\ &Z_i \sim M(1;\omega_1,\cdots,\omega_K). \end{aligned} \tag{16}$$

$TN(0,1;(0,\infty))$ denotes the truncated normal distribution. Let $\boldsymbol{t} = (t_1, \cdots, t_n)^{\mathrm{T}}$, $\boldsymbol{u} = (u_1, \cdots, u_n)^{\mathrm{T}}$, $\boldsymbol{Y} = (y_1, \cdots, y_n)^{\mathrm{T}}$ and $\boldsymbol{Z}^* = (\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n)^{\mathrm{T}}$. Among them, $\boldsymbol{t}$ and $\boldsymbol{u}$ are also regarded as latent vectors. Then the complete log-likelihood function with complete-data $\boldsymbol{Y}_c = \left(\boldsymbol{Y}^{\mathrm{T}}, \boldsymbol{u}^{\mathrm{T}}, \boldsymbol{t}^{\mathrm{T}}, \boldsymbol{Z}^{*\mathrm{T}}\right)^{\mathrm{T}}$ is given by:

$$l_c(\boldsymbol{\Psi}) = l_c(\boldsymbol{\omega}) + l_c(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \boldsymbol{\lambda}, \boldsymbol{\tau}^*), \tag{17}$$

with:

$$l_c(\boldsymbol{\omega}) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log\omega_k, \tag{18}$$

$$\begin{aligned} l_c(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \boldsymbol{\lambda}, \boldsymbol{\tau}^*) = &\sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\left[C - \log\sigma_k^2 - \frac{t_i^2}{2\sigma_k^2} + \frac{t_i\lambda_k}{\sigma_k^2}\left(y_i - \alpha_k - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_k\right)\right] \\ &- \sum_{i=1}^{n}\sum_{k=1}^{K}\frac{z_{ik}}{2\sigma_k^2}\left\{\left[\ell^{-1}(u_i) + \lambda_k^2\right]\left(y_i - \alpha_k - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_k\right)^2\right\} \\ &+ \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log h(u_i;\boldsymbol{\tau}_k). \end{aligned} \tag{19}$$

C is a constant that does not depend on any unknown parameter, and $h(u_i;\boldsymbol{\tau}_k)$ is the density function of the latent variable u_i.

Replacing $l(\boldsymbol{\Psi})$ with $l_c(\boldsymbol{\Psi})$ in the penalized log-likelihood function, the complete penalized log-likelihood function is given by:

$$L_c(\boldsymbol{\Psi}) = l_c(\boldsymbol{\Psi}) - p(\boldsymbol{\Psi}). \tag{20}$$

Refer to the method of Fan and Li [8], given the initial parameter value $\boldsymbol{\Psi}^{(0)}$, $p(\boldsymbol{\Psi})$ can be replaced by the following local quadratic function:

$$p(\boldsymbol{\Psi}) \approx n\sum_{k=1}^{K}\omega_k \sum_{j=1}^{q}\left[p_{a_k}\left(\left|\beta_{kj}^{(0)}\right|\right)+\frac{p'_{a_k}\left(\left|\beta_{kj}^{(0)}\right|\right)}{2\left|\beta_{kj}^{(0)}\right|}\left(\beta_{kj}^{2}-\beta_{kj}^{(0)^2}\right)\right] \tag{21}$$

This approximation will be applied in the CM-step of the algorithm at each iteration. The adjusted EM algorithm proceeds with the following three steps. The E-step calculates the conditional expectation of the complete penalized log-likelihood function, the CM-step and CML-step obtain the closed-form of parameter estimates.

- The E-step. Given the current estimates $\hat{\boldsymbol{\Psi}}^{(m)}$, calculate the Q-function, $Q\left(\boldsymbol{\Psi}\mid\hat{\boldsymbol{\Psi}}^{(m)}\right)=E\left[L_c(\boldsymbol{\Psi})\mid \boldsymbol{Y},\hat{\boldsymbol{\Psi}}^{(m)}\right]$, obtained as:

$$\begin{aligned} Q\left(\boldsymbol{\Psi}\mid\hat{\boldsymbol{\Psi}}^{(m)}\right) &= Q_1\left(\boldsymbol{\omega}\mid\hat{\boldsymbol{\Psi}}^{(m)}\right)+Q_2\left(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\sigma}^2,\boldsymbol{\lambda}\mid\hat{\boldsymbol{\Psi}}^{(m)}\right) \\ &\quad +Q_3\left(\boldsymbol{\tau}^*\mid\hat{\boldsymbol{\Psi}}^{(m)}\right)-p\left(\boldsymbol{\Psi}\mid\hat{\boldsymbol{\Psi}}^{(m)}\right), \end{aligned} \tag{22}$$

with:

$$Q_1\left(\boldsymbol{\omega}\mid\hat{\boldsymbol{\Psi}}^{(m)}\right)=\sum_{i=1}^{n}\sum_{k=1}^{K}\hat{z}_{ik}^{(m)}\log\omega_k, \tag{23}$$

$$\begin{aligned} &Q_2\left(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\sigma}^2,\boldsymbol{\lambda}\mid\hat{\boldsymbol{\Psi}}^{(m)}\right) \\ &=\sum_{i=1}^{n}\sum_{k=1}^{K}\hat{z}_{ik}^{(m)}\left[-\log\sigma_k^2-\frac{\hat{t}_{ik}^{2(m)}}{2\sigma_k^2}+\frac{\hat{t}_{ik}^{(m)}\lambda_k}{\sigma_k^2}\left(y_i-\alpha_k-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_k\right)\right] \\ &\quad -\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{\hat{z}_{ik}^{(m)}}{2\sigma_k^2}\left[\left(\hat{\ell}_{ik}^{-1(m)}+\lambda_k^2\right)\left(y_i-\alpha_k-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_k\right)^2\right], \end{aligned} \tag{24}$$

$$Q_3\left(\boldsymbol{\tau}^*\mid\hat{\boldsymbol{\Psi}}^{(m)}\right)=E\left[\sum_{i=1}^{n}\sum_{k=1}^{K}z_{ik}\log h\left(u_i;\boldsymbol{\tau}_k\right)\mid\boldsymbol{Y},\hat{\boldsymbol{\Psi}}^{(m)}\right]. \tag{25}$$

The required expressions are $\hat{z}_{ik}^{(m)}$, $\hat{t}_{ik}^{2(m)}$, $\hat{t}_{ik}^{(m)}$ and $\hat{\ell}_{ik}^{-1(m)}$.

First, the conditional expectation $\hat{z}_{ik}^{(m)}=E\left[z_{ik}\mid y_i,\hat{\boldsymbol{\Psi}}^{(m)}\right]$ is given by:

$$\hat{z}_{ik}^{(m)}=\frac{\hat{\omega}_k^{(m)}SSMN\left(y_i;\hat{\alpha}_k^{(m)}+\boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_k^{(m)},\hat{\sigma}_k^{2(m)},\hat{\lambda}_k^{(m)},\hat{\boldsymbol{\tau}}_k^{(m)};\ell\right)}{\sum_{k=1}^{K}\hat{\omega}_k^{(m)}SSMN\left(y_i;\hat{\alpha}_k^{(m)}+\boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_k^{(m)},\hat{\sigma}_k^{2(m)},\hat{\lambda}_k^{(m)},\hat{\boldsymbol{\tau}}_k^{(m)};\ell\right)}. \tag{26}$$

Then, refer to [15], $\hat{t}_{ik}^{(m)}=E\left[t_i\mid y_i,\hat{\boldsymbol{\Psi}}^{(m)},z_{ik}=1\right]$ and $\hat{t}_{ik}^{2(m)}=E\left[t_i^2\mid y_i,\hat{\boldsymbol{\Psi}}^{(m)},z_{ik}=1\right]$ can be evaluated by:

$$\hat{t}_{ik}^{(m)}=\hat{\lambda}_k^{(m)}\hat{e}_{ik}^{(m)}+\hat{\sigma}_k^{(m)}W_{\Phi}\left(\frac{\hat{\lambda}_k^{(m)}\hat{e}_{ik}^{(m)}}{\hat{\sigma}_k^{(m)}}\right), \tag{27}$$

$$\hat{t}_{ik}^{2(m)}=\left(\hat{\lambda}_k^{(m)}\hat{e}_{ik}^{(m)}\right)^2+\hat{\sigma}_k^{2(m)}+\hat{\lambda}_k^{(m)}\hat{\sigma}_k^{(m)}\hat{e}_{ik}^{(m)}W_{\Phi}\left(\frac{\hat{\lambda}_k^{(m)}\hat{e}_{ik}^{(m)}}{\hat{\sigma}_k^{(m)}}\right). \tag{28}$$

with $W_{\Phi}(u)=\phi(u)/\Phi(u)$ and $\hat{e}_{ik}^{(m)}=y_i-\hat{\alpha}_k^{(m)}-\boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_k^{(m)}$.

Further, $\hat{\ell}_{ik}^{-1(m)}$ has different expressions for RMR-SSMN models with different distributions in the SSMN family, obtained as:

$$\hat{\ell}_{ik}^{-1(m)}=\begin{cases}1, & \text{for RMR-SN model}\\ \dfrac{\hat{v}_k^{(m)}+1}{\hat{v}_k^{(m)}+d_{ik}}, & \text{for RMR-STN model}\\ \dfrac{1-\hat{v}_k^{(m)}+\hat{v}_k^{(m)}\hat{\gamma}_k^{(m)^{\frac{3}{2}}}\exp\left[\left(1-\hat{\gamma}_k^{(m)}\right)d_{ik}/2\right]}{1-\hat{v}_k^{(m)}+\hat{v}_k^{(m)}\hat{\gamma}_k^{(m)^{\frac{1}{2}}}\exp\left[\left(1-\hat{\gamma}_k^{(m)}\right)d_{ik}/2\right]}, & \text{for RMR-SCN model}\\ \hat{v}_k^{(m)}d_{ik}^{\hat{v}_k^{(m)}-1}, & \text{for RMR-SPE model}\end{cases} \tag{29}$$

with $d_{ik}=\left(y_i-\hat{\alpha}_k^{(m)}-\boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_k^{(m)}\right)^2\Big/\hat{\sigma}_k^{2(m)}$.

- The CM-step. Maximize $Q\left(\boldsymbol{\Psi}\mid\hat{\boldsymbol{\Psi}}^{(m)}\right)$ with respect to $\boldsymbol{\Psi}$ on the $(m+1)$th iteration. As in [11], the mixing proportions are updated by:

$$\hat{\omega}_k^{(m+1)}=\frac{1}{n}\sum_{i=1}^{n}\hat{z}_{ik}^{(m)}, \tag{30}$$

which are the approximate iterated values. Maximizing $Q_1\left(\boldsymbol{\omega}\mid\hat{\boldsymbol{\Psi}}^{(m)}\right)$ with respect to the $\boldsymbol{\omega}$ instead of maximizing $Q\left(\boldsymbol{\Psi}\mid\hat{\boldsymbol{\Psi}}^{(m)}\right)$ will simplify the computation of $\hat{\omega}_k^{(m+1)}$ and this updating scheme works well in our simulations.

We now consider that $\boldsymbol{\omega}$ is constant, and maximize $Q\left(\boldsymbol{\Psi}\mid\hat{\boldsymbol{\Psi}}^{(m)}\right)$ with respect to the rest parameters in $\boldsymbol{\Psi}$. The updates of $\left(\alpha_k,\sigma_k^2,\lambda_k,\boldsymbol{\beta}_k\right)^{\mathrm{T}}$ are given by:

$$\hat{\alpha}_k^{(m+1)}=\frac{\sum_{i=1}^{n}\hat{z}_{ik}^{(m)}\left[-\hat{t}_{ik}^{(m)}\hat{\lambda}_k^{(m)}+\left(\hat{\ell}_{ik}^{-1(m)}+\hat{\lambda}_k^{(m)2}\right)\left(y_i-\boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_k^{(m)}\right)\right]}{\sum_{i=1}^{n}\hat{z}_{ik}^{(m)}\left(\hat{\ell}_{ik}^{-1(m)}+\hat{\lambda}_k^{(m)2}\right)}, \tag{31}$$

$$\hat{\sigma}_k^{2(m+1)}=\frac{\sum_{i=1}^{n}\hat{z}_{ik}^{(m)}\left[\hat{t}_{ik}^{2(m)}-2\hat{t}_{ik}^{(m)}\hat{\lambda}_k^{(m)}\hat{e}_{ik}^{(m)}+\left(\hat{\ell}_{ik}^{-1(m)}+\hat{\lambda}_k^{(m)2}\right)\hat{e}_{ik}^{(m)2}\right]}{\sum_{i=1}^{n}2\hat{z}_{ik}^{(m)}}, \tag{32}$$

$$\hat{\lambda}_k^{(m+1)}=\sum_{i=1}^{n}\hat{z}_{ik}^{(m)}\hat{t}_{ik}^{(m)}\hat{e}_{ik}^{(m)}\Big/\sum_{i=1}^{n}\hat{z}_{ik}^{(m)}\hat{e}_{ik}^{(m)2}, \tag{33}$$

$$\hat{\boldsymbol{\beta}}_k^{(m+1)}=\left[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}_k\boldsymbol{X}+n\hat{\sigma}_k^{2(m)}\hat{\omega}_k^{(m)}\Delta_a\left(\hat{\boldsymbol{\beta}}_k^{(m)}\right)\right]^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}_k\boldsymbol{B}_k, \tag{34}$$

with:

$$\boldsymbol{A}_k=\left[\operatorname{diag}(\hat{\ell}_{1k}^{-1(m)},\cdots,\hat{\ell}_{nk}^{-1(m)})+\hat{\lambda}_k^{(m)2}\mathbb{I}_n\right]\operatorname{diag}\left(\hat{z}_{1k}^{(m)},\cdots,\hat{z}_{nk}^{(m)}\right),$$

$$\boldsymbol{B}_k=\left(\hat{b}_{1k}^{(m)},\cdots,\hat{b}_{nk}^{(m)}\right)^{\mathrm{T}},\ \hat{b}_{ik}^{(m)}=y_i-\hat{\alpha}_k^{(m)}-\frac{\hat{t}_{ik}^{(m)}\hat{\lambda}_k^{(m)}}{\hat{\ell}_{ik}^{-1(m)}+\hat{\lambda}_k^{(m)2}},$$

$$\Delta_a\left(\hat{\boldsymbol{\beta}}_k^{(m)}\right)=\operatorname{diag}\left(\frac{p'_{a_k}\left(\left|\hat{\beta}_{k1}^{(m)}\right|\right)}{\left|\hat{\beta}_{k1}^{(m)}\right|},\frac{p'_{a_k}\left(\left|\hat{\beta}_{k2}^{(m)}\right|\right)}{\left|\hat{\beta}_{k2}^{(m)}\right|},\cdots,\frac{p'_{a_k}\left(\left|\hat{\beta}_{kq}^{(m)}\right|\right)}{\left|\hat{\beta}_{kq}^{(m)}\right|}\right),$$

and $\mathbb{I}_n$ is an identity matrix of order n, and $\boldsymbol{X}=\left(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n\right)^{\mathrm{T}}$ is a matrix of order $n\times q$.

- The CML-step. Fix $\hat{\boldsymbol{\Psi}}_p^{(m+1)} = \left(\hat{\alpha}_k^{(m+1)}, \hat{\boldsymbol{\beta}}_k^{(m+1)}, \hat{\sigma}_k^{2(m+1)}, \hat{\lambda}_k^{(m+1)}\right)^{\mathrm{T}}$ and $\hat{\omega}_k^{(m+1)}$, update $\boldsymbol{\tau}^*$ to get $\hat{\boldsymbol{\tau}}^{*(m+1)} = \left(\hat{\boldsymbol{\tau}}_1^{(m+1)}, \cdots, \hat{\boldsymbol{\tau}}_K^{(m+1)}\right)^{\mathrm{T}}$ by optimizing the constrained log-likelihood function:

$$\hat{\boldsymbol{\tau}}^{*(m+1)} = \operatorname{argmax}_{\boldsymbol{\tau}_1,\cdots,\boldsymbol{\tau}_K} \sum_{i=1}^{n} \log \sum_{k=1}^{K} \hat{\omega}_k^{(m+1)} SSMN\left(y_i; \hat{\boldsymbol{\Psi}}_p^{(m+1)}, \boldsymbol{\tau}_k; \ell\right). \tag{35}$$

The above iterations are repeated alternately until the maximum number of iterations is reached or a suitable stopping rule is met. In this work, the iterations will be completed when $\left\|\hat{\boldsymbol{\Psi}}^{(m+1)} - \hat{\boldsymbol{\Psi}}^{(m)}\right\|$ is sufficiently small, such as 10^{-5}.

4.2. Selection of Turning Parameters and Components

When using the methods proposed in this paper, we also need to consider how to determine the components K and the size of the turning parameters in the penalty function. Cross-Validation (CV), Generalized Cross-Validation (GCV), AIC and BIC are commonly used criteria for the selection of turning parameters.

As showed in [12], the final selected model will be overfitting if the turning parameter selected by GCV and they use the BIC to choose. In this paper, we also propose a proper BIC criterion for RMR-SSMN models to select turning parameters $\boldsymbol{a} = (a_1, \cdots, a_K)^{\mathrm{T}}$, the constant b and the components K.

Let $\boldsymbol{\theta} = (\boldsymbol{a}, b, K)^{\mathrm{T}}$, we should take a set of $\boldsymbol{\theta}$ at a time over a suitable range and use the proposed adjusted EM algorithm to obtain the corresponding parameter estimates $\hat{\boldsymbol{\Psi}}$. The optimal set of $\boldsymbol{\theta}$ is selected by minimizing the following BIC criterion:

$$BIC(\boldsymbol{\theta}) = -2l\left(\hat{\boldsymbol{\Psi}}\right) + \left(\tilde{p}K - 1 + \sum_{k=1}^{K} \eta_k\right) \times \log(n). \tag{36}$$

where η_k represents the number of non-zero regression coefficients of $\boldsymbol{\beta}_k$ and $\tilde{p}$ is either equal to 4 (RMR-SN model), 5 (RMR-STN and RMR-SPE models) or 6 (RMR-SCN model).

5. Simulation Studies

We perform Monte Carlo simulations to evaluate the performance of the proposed robust mixture model and adjusted EM algorithm. To evaluate the effect of variable selection and the accuracy of parameter estimates, we use the correctly estimated zero coefficients (S1), correctly estimated non-zero coefficients (S2), the mean estimate over all falsely identified non-zero predictors (M_{NZ}) [16] of $\boldsymbol{\beta}$ and the mean squared error (MSE) of regression coefficients ($\mathrm{MSE}\left(\hat{\boldsymbol{\beta}}\right)$),

$$\mathrm{MSE}\left(\hat{\boldsymbol{\beta}}\right) = E\left(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right)^{\mathrm{T}}\left(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right).$$

5.1. Simulation 1

The first simulation uses the SCAD penalty function to select significant variables for RMR-STN, RMR-SPE and RMR-SCN models, and compare the simula-

tion results with the Gaussian FMR model and RMR-SN model.

We set $K=2$ for the simulation so that the sample data set $\{(y_i, \boldsymbol{x}_i); i=1,\cdots,n\}$ for the mixture regression model is derived from the following model:

$$y=\begin{cases}\alpha_1+\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_1+\varepsilon_1, & Z=1\\ \alpha_2+\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_2+\varepsilon_2, & Z=2\end{cases} \tag{37}$$

where Z is used to identify the subgroup that the sample belongs to. $\alpha_1=2$, $\boldsymbol{\beta}_1=(4,1,-2,0,0,0,0,0,0,0)^{\mathrm{T}}$, $\alpha_2=-2$, $\boldsymbol{\beta}_2=(1,-3,0,2,3,0,0,0,0,0)^{\mathrm{T}}$, $\omega_1=0.6$ and $\omega_2=0.4$.

The covariate $\boldsymbol{x}$ is generated from a multivariate normal with mean 0, variance 1, and two correlation structures: $\rho_{ij}=0.5^{|i-j|}$, $1\le i, j\le n$. The simulation considers the following three error distributions cases: 1) the random errors ε_1 and ε_2 follow the t-distribution with 3 degrees of freedom ($t(3)$); 2) the random errors ε_1 and ε_2 follow the chi-square distribution with 3 degrees of freedom ($\chi^2(3)$); 3) The random errors ε_1 and ε_2 follow the mixture distribution of normal $0.9N(0,1)+0.1N(0,5^2)$. So that there are 15 sets of combination, and for each combination, we respectively performed 100 repetitions for the simulation with $n=300$.

From Table 1, the value of S1 in Com1 and Com2 from RMR-STN, RMR-SPE and RMR-SCN are all bigger than the value in FMR model for all three cases, respectively. In case (1), the S1 in Com2 from RMR-SPE is biggest (S1 = 0.9533), however, the S1 in Com2 from FMR is smallest (S1 = 0.8533). In case (2), the RMR-SCN model has the biggest S1 (S1 = 0.9033) in Com2 while the S1 in Com2 from FMR is 0.8167. In case (3), when the S2 in Com1 and Com2 from FMR are 0.9933 and 0.9950, respectively, the values of S2 in both components from RMR-STN are 1.00.

Furthermore, the value of $\mathrm{MSE}(\hat{\boldsymbol{\beta}})$ in Com1 and Com2 from RMR-STN, RMR-SPE and RMR-SCN are much smaller than the value in FMR model. When errors follow $\chi^2(3)$ distribution, RMR-SN performs well with the smallest $\mathrm{MSE}(\hat{\boldsymbol{\beta}})$ and S2 = 1.00 in both Com1 and Com2 that indicates the non-zero coefficients are all identified correctly. Overall, RMR-SSMN models are more robust than FMR for variable selection when the data set is asymmetric ($\chi^2(3)$), heavy-tailed ($t(3)$) and contaminated ($0.9N(0,1)+0.1N(0,5^2)$).

5.2. Simulation 2

Simulation 2 uses the MIXL2-SCAD penalty function to select significant variables for RMR-STN, RMR-SPE and RMR-SCN models. By comparing the results of Simulation 1 and Simulation 2, the effects of SCAD and MIXL2-SCAD penalty function on variable selection are analyzed. In addition, the generation of the sample data set and the distributions of random errors in this simulation are the same as in Simulation 1, and both $n=300$ and $n=500$ cases are considered.

Table 1. Results of FMR, RMR-SN, RMR-STN, RMR-SPE and RMR-SCN using SCAD penalty function on 100 replicates.

Error	Model	n	Component	S1	S2	M_{NZ}	MSE$\left(\hat{\boldsymbol{\beta}}\right)$
$t(3)$	FMR	300	Com1	0.9143	1.0000	−0.0357	0.0838
			Com2	0.8533	0.9800	0.0382	0.8976
	RMR-SN	300	Com1	0.9057	1.0000	−0.0029	0.9638
			Com2	0.8900	0.9850	−0.0339	1.2244
	RMR-STN	300	Com1	0.9343	1.0000	−0.0004	0.0619
			Com2	0.9467	1.0000	0.0181	0.1349
	RMR-SPE	300	Com1	0.9800	1.0000	−0.0048	0.0521
			Com2	0.9533	1.0000	0.0375	0.1560
	RMR-SCN	300	Com1	0.9714	1.0000	−0.0151	0.0496
			Com2	0.9400	1.0000	0.0195	0.1426
$\chi^2(3)$	FMR	300	Com1	0.9400	0.9000	0.0387	0.6687
			Com2	0.8167	0.9900	0.0568	0.5838
	RMR-SN	300	Com1	0.9771	1.0000	0.0138	0.0636
			Com2	0.8867	1.0000	−0.0342	0.1978
	RMR-STN	300	Com1	0.9543	0.9867	0.0053	0.1283
			Com2	0.8767	1.0000	−0.0491	0.2273
	RMR-SPE	300	Com1	0.9686	0.9933	0.0150	0.0909
			Com2	0.8967	1.0000	−0.0003	0.2287
	RMR-SCN	300	Com1	0.9543	0.9800	−0.0257	0.1555
			Com2	0.9033	0.9950	−0.0511	0.2516
$0.9N(0,1)+0.1N(0,5^2)$	FMR	300	Com1	0.9371	0.9933	−0.0098	0.1500
			Com2	0.8000	0.9950	−0.0305	0.4810
	RMR-SN	300	Com1	0.9314	1.0000	0.0023	0.1133
			Com2	0.8767	1.0000	−0.0306	0.3358
	RMR-STN	300	Com1	0.9686	1.0000	−0.0019	0.0461
			Com2	0.9333	1.0000	−0.0361	0.1000
	RMR-SPE	300	Com1	0.9714	0.9933	0.0065	0.0802
			Com2	0.9433	1.0000	−0.0287	0.1149
	RMR-SCN	300	Com1	0.9857	0.9933	0.0071	0.0691
			Com2	0.9433	1.0000	−0.0361	0.0908

From Table 2, we can know that as the sample size n increases, the values of S1 and S2 in Com1 and Com2 are getting closer and closer to 1, and the value of MSE$\left(\hat{\boldsymbol{\beta}}\right)$ is getting smaller and smaller, indicating the asymptotic property of parameter estimates. When $n=500$ and errors follow $t(3)$ distribution, the values of S1 and S2 in Com1 from RMR-SPE model are equal to 1.00, which indicates that the MIXL2-SCAD penalty ensures the non-zero and zero coefficients

Table 2. Results of RMR-STN, RMR-SPE and RMR-SCN using MIXL2-SCAD penalty function on 100 replicates.

Error	Model	n	Component	S1	S2	M_{NZ}	MSE$\left(\hat{\boldsymbol{\beta}}\right)$
$t(3)$	RMR-STN	300	Com1	0.9800	1.0000	−0.0050	0.0486
			Com2	0.9633	1.0000	0.0129	0.1251
		500	Com1	0.9914	1.0000	0.0066	0.0244
			Com2	0.9700	1.0000	−0.0025	0.0652
	RMR-SPE	300	Com1	0.9943	1.0000	−0.0045	0.0504
			Com2	0.9533	1.0000	0.0135	0.1504
		500	Com1	1.0000	1.0000	0.0000	0.0256
			Com2	0.9933	1.0000	−0.0049	0.0545
	RMR-SCN	300	Com1	0.9771	1.0000	−0.0025	0.0509
			Com2	0.9567	1.0000	0.0223	0.1368
		500	Com1	0.9971	1.0000	−0.0019	0.0272
			Com2	0.9933	1.0000	−0.0009	0.0653
$\chi^2(3)$	RMR-STN	300	Com1	0.9543	0.9933	−0.0091	0.1136
			Com2	0.8933	1.0000	−0.0584	0.2597
		500	Com1	0.9886	1.0000	0.0053	0.0417
			Com2	0.9867	1.0000	0.0065	0.0923
	RMR-SPE	300	Com1	0.9743	1.0000	−0.0020	0.0641
			Com2	0.9433	1.0000	−0.0116	0.1765
		500	Com1	0.9943	1.0000	0.0020	0.0321
			Com2	0.9900	1.0000	0.0009	0.1047
	RMR-SCN	300	Com1	0.9571	0.9933	−0.0210	0.0992
			Com2	0.9433	1.0000	−0.0179	0.2297
		500	Com1	0.9857	1.0000	0.0010	0.0491
			Com2	0.9933	1.0000	0.0017	0.1436
$0.9N(0,1)+0.1N(0,5^2)$	RMR-STN	300	Com1	0.9743	1.0000	−0.0013	0.0490
			Com2	0.9667	1.0000	−0.0233	0.0933
		500	Com1	0.9886	1.0000	−0.0014	0.0281
			Com2	0.9800	1.0000	0.0005	0.0616
	RMR-SPE	300	Com1	0.9800	1.0000	0.0064	0.0489
			Com2	0.9800	1.0000	−0.0197	0.0971
		500	Com1	0.9943	1.0000	−0.0009	0.0327
			Com2	1.0000	1.0000	0.0000	0.0968
	RMR-SCN	300	Com1	0.9857	1.0000	0.0011	0.0481
			Com2	0.9600	1.0000	−0.0096	0.1005
		500	Com1	0.9971	1.0000	0.0005	0.0227
			Com2	1.0000	1.0000	0.0000	0.0576

can be identified completely. When $n = 500$ and errors follow $0.9N(0,1) + 0.1N(0,5^2)$, the same result appears in Com2 from RMR-SPE and RMR-SCN model. The absolute values of mean estimate over all falsely identified non-zero predictors (M_{NZ}) are smaller than 0.01 from MIXL2-SCAD when $n = 500$.

By comparing Table 1 and Table 2, we can see that the values of S1 and S2 in Com1 and Com2 from MIXL2-SCAD are all greater than or equal to the values from SCAD penalty for all cases when $n = 300$. It is worth noting that in case (3), when n = 300, the values of S2 in Com1 and Com2 from RMR-STN, RMR-SPE and RMR-SCN using MIXL2-SCAD penalty are all 1.00, however, the values of S2 in Com1 from RMR-SPE and RMR-SCN using SCAD penalty are 0.9933. From these comparisons of experimental data, we can know that MIXL2-SCAD performs better than SCAD penalty in variable selection.

6. Real Data Analysis

In this section, we obtain the Seoul bike sharing demand data set from the website http://archive.ics.uci.edu/ml/datasets.php. From this dataset, we screen out the total number of bikes rented from 10:00 am to 11:00 am every functional day of bike rental system in Seoul from December 1, 2017 to November 30, 2018 with 12 features that may affect the demand of rental bikes. There are 353 observations in total. The 12 features are: temperature (x_1), humidity (x_2), wind-speed (x_3), visibility (x_4), dew point temperature (x_5), solar radiation (x_6), rainfall (x_7), snowfall (x_8), holiday (holiday = 1, else = 0; x_9), spring (spring = 1, else = 0; x_{10}), summer (summer = 1, else = 0; x_{11}) and autumn (autumn = 1, else = 0; x_{12}). x_9 - x_{12} are dummy variables and x_{10} - x_{12} indicate different seasons. Considering that there may be further differential effects between seasons and holiday, we continue to introduce 3 interaction terms between dummy variables, namely $x_9 * x_{10}$, $x_9 * x_{11}$, $x_9 * x_{12}$. This leads to a set of 15 potential covariates affecting rented bike count (RBC) from 10:00 am to 11:00 am.

Let $Y = \mathrm{RBC}/sd(\mathrm{RBC})$ be the response variable, where $sd(\mathrm{RBC})$ is the standard deviation of RBC. Figure 1 shows the histogram and density estimate of Y, we can see that the data set has obvious heterogeneity, so that the RMR-STN model is applicable. We also apply RMR-SPE and RMR-SCN models to this real data set, the outcomes are worse than RMR-STN's result, thus we do not report the results here.

The parameter estimates under FMR, RMR-STN ($K = 2$) and RMR-STN ($K = 3$) with BIC method and MIXL2-SCAD penalty function are given in Table 3. The $K = 3$ RMR-STN model has the lowest BIC (542.5) and the $K = 2$ RMR-STN model ranks second (BIC = 544.7) when FMR model has the biggest BIC (562.8). Furthermore, the predicted rented bike count from the $K = 3$ RMR-STN model has the smallest MSE of 0.09 and the biggest regression $\tilde{R}^2$ of 0.90.

From Table 3, the bike rented demand can be divided into three categories: "low", "medium" and "high" during the time period from 10:00 am to 11:00 am

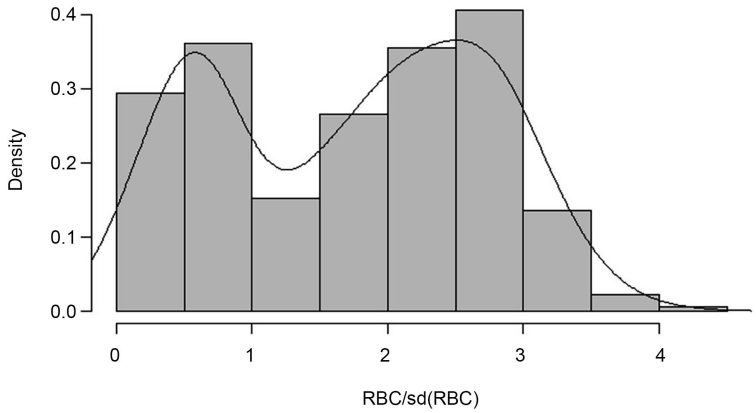

Figure 1. Histogram and density estimate for $Y = \mathrm{RBC}/sd(\mathrm{RBC})$.

Table 3. Summary of FMR, RMR-STN ($K=2$) and RMR-STN ($K=3$) model with BIC method and MIXL2-SCAD penalty for Seoul bike sharing demand data set.

Covariates	FMR		RMR-STN ($K=2$)		RMR-STN ($K=3$)		
	Com1	Com2	Com1	Com2	Com1	Com2	Com3
ω	0.44	0.56	0.53	0.47	0.52	0.04	0.44
Intercept	1.82	1.26	1.85	1.36	1.37	0.41	2.03
x_1	0.82	0.41		0.43	0.44		
x_2	−0.24	−0.04	−0.46	−0.08	−0.07	−0.05	−0.67
x_3				−0.03	−0.01		
x_4			0.04			0.01	
x_5			0.93				1.18
x_6	0.20	0.15	0.26	0.18	0.17		0.12
x_7	−0.28	−0.10	−0.32	−0.10	−0.10		−0.29
x_8		−0.03	−0.03	−0.04	−0.03		
x_9		−0.49		−0.64	−0.62		
x_{10}		0.54		0.51	0.50		
x_{11}			0.24				
x_{12}		1.18	0.63	1.02	1.03		0.49
$x_9 * x_{10}$			0.87				
$x_9 * x_{11}$							
$x_9 * x_{12}$			0.48				

with $K=3$ RMR-STN model. Humidity is a negative factor for all three types of demand. When the bike rented demand is "medium", warmer temperature and increased solar radiation help increase bike demand, while rainfall, snowfall, and holidays reduce the demand. In contrast, when bike rented demand is "high", the positive effect of dew point temperature on bike demand is greatest, while the negative effects of holidays and snowfall disappear. In addition, we can also find

that the rented bike count has a strong seasonality and the rented count will be more in other seasons than in winter.

7. Conclusion

In this paper, we mainly propose a robust mixture regression model based on the skew scale mixtures of normal distributions (RMR-SSMN) which can avoid the potential limitation of normal mixtures. A new penalty function (MIXL2-SCAD) which combines SCAD and l_2 penalties is presented for variable selection. Through simulations, we find that the RMR-SSMN models are more robust than general FMR models for heterogeneous data with asymmetry and heavy-tailed properties, and outliers. Furthermore, the capability of MIXL2-SCAD to select the most parsimonious FMR model is obviously better than SCAD. The proposed methodology is applied to a real data set and achieves reasonable results. However, this paper only focuses on the mixture of the simple linear model, and further research can focus on the mixture of the semiparametric model or nonparametric model.

Conflicts of Interest

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

References

[1] Krykun, I. (2018) The Arc-Sine Laws for the Skew Brownian Motion and Their Interpretation. *Journal of Applied Mathematics and Physics*, **6**, 347-357. https://doi.org/10.4236/jamp.2018.62033

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

[3] McLachlan, G.J. and Peel, D. (1998) Robust Cluster Analysis via Mixtures of Multivariate T-Distributions. *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition* (*SPR*) *and Structural and Syntactic Pattern Recognition* (*SSPR*), Sydney, 11-13 August 1998, 658-666. https://doi.org/10.1007/BFb0033290

[4] Basso, R.M., Lachos, V.H., Cabral, C.R.B. and Ghosh, P. (2011) Robust Mixture Modeling based on Scale Mixtures of Skew-Normal Distributions. *Computational Statistics & Data Analysis*, **54**, 2926-2941. https://doi.org/10.1016/j.csda.2009.09.031

[5] Franczak, B.C., Browne, R.P. and McNicholas, P.D. (2014) Mixtures of Shifted Asymmetric Laplace Distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **36**, 1149-1157. https://doi.org/10.1109/TPAMI.2013.216

[6] Zou, H. and Hastie, T. (2005) Regularization and Variable Selection via Elastic Net. *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), **67**, 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

[7] Zhang, C.H. (2010) Nearly Unbiased Variable Selection under Minimax Concave Penalty. *The Annals of Statistics*, **38**, 894-942. https://doi.org/10.1214/09-AOS729

[8] Fan, J. and Li, R. (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

[9] Ferreira, C.S., Bolfarine, H. and Lachos, V.H. (2011) Skew Scale Mixtures of Normal

Distributions: Properties and Estimation. *Statistical Methodology*, **8**, 154-171. https://doi.org/10.1016/j.stamet.2010.09.001

[10] Gomez, H.W., Venegas, O. and Bolfarine, H. (2007) Skew-Symmetric Distributions Generated by the Normal Distribution Function. *Environmetrics*, **18**, 395-407. https://doi.org/10.1002/env.817

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

[12] Khalili, A. (2010) New Estimation and Feature Selection Methods in Mixture-of-Experts Models. *Canadian Journal of Statistics*, **38**, 519-539. https://doi.org/10.1002/cjs.10083

[13] Meng, X.L. and Rubin, D.B. (1993) Maximum Likelihood Estimation via the ECM Algorithm: A General Framework. *Biometrika*, **80**, 267-278. https://doi.org/10.1093/biomet/80.2.267

[14] Liu, C. and Rubin, D.B. (1994) A Simple Extension of EM and ECM with Faster Monotone Convergence. *Biometrika*, **81**, 633-648. https://doi.org/10.1093/biomet/81.4.633

[15] Ferreira, C.S. and Lachos, V.H. (2016) Nonlinear Regression Models under Skew Scale Mixtures of Normal Distributions. *Statistical Methodology*, **33**, 131-146. https://doi.org/10.1016/j.stamet.2016.08.004

[16] 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