entropy

MDPI

Article

Model Selection for Exponential Power Mixture Regression Models

Yunlu Jiang, Jiangchuan Liu, Hang Zou and Xiaowen Huang *

Department of Statistics and Data Science, College of Economics, Jinan University, Guangzhou 510632, China; tjiangyl@jnu.edu.cn (Y.J.); biubiuyun@outlook.com (J.L.); zouhang19980304@163.com (H.Z.)
* Correspondence: xiaowen@stu2023.jnu.edu.cn

Abstract: Finite mixture of linear regression (FMLR) models are among the most exemplary statistical tools to deal with various heterogeneous data. In this paper, we introduce a new procedure to simultaneously determine the number of components and perform variable selection for the different regressions for FMLR models via an exponential power error distribution, which includes normal distributions and Laplace distributions as special cases. Under some regularity conditions, the consistency of order selection and the consistency of variable selection are established, and the asymptotic normality for the estimators of non-zero parameters is investigated. In addition, an efficient modified expectation-maximization (EM) algorithm and a majorization-maximization (MM) algorithm are proposed to implement the proposed optimization problem. Furthermore, we use the numerical simulations to demonstrate the finite sample performance of the proposed methodology. Finally, we apply the proposed approach to analyze a baseball salary data set. Results indicate that our proposed method obtains a smaller BIC value than the existing method.

Keywords: finite mixture of linear regression models; variable selection; exponential power distribution; modified EM algorithm

Citation: Jiang, Y.; Liu, J.; Zou, H.; Huang, X. Model Selection for Exponential Power Mixture Regression Models. *Entropy* **2024**, *26*, 422. https://doi.org/10.3390/e26050422

Academic Editor: Geert Verdoolaege

Received: 27 February 2024
Revised: 24 April 2024
Accepted: 14 May 2024
Published: 15 May 2024

1. Introduction

FMLR models are among the most exemplary statistical tools to deal with various heterogeneous data. Since FMLR models were first introduced by [1,2], they are widely applied in many research fields, e.g., machine learning [3], social sciences [4], and business [5]. For more references to FMLR models, see [6–8].

There are two important statistical problems in FMLR models: order selection and variable selection for the different regressions. However, order selection should be the first discussed issue in FMLR models. There exists a lot of literature to deal with this problem. For example, Ref. [9] introduced a penalized likelihood method for mixtures of univariate location distributions. Ref. [10] proposed a penalized likelihood method to select the number of mixing components for the finite multivariate Gaussian mixture models. For variable selection problems for each regression component, Ref. [11] applied subsect selection, REDapproaches such as Akaike information criterion (AIC) and Bayesian information criterion (BIC) to perform a variable selection for each component in a finite mixture of Poisson regression models. To avoid the drawbacks of subsect selection, Ref. [12] introduced a penalized likelihood method for variable selection in FMLR models. Ref. [13] proposed a robust variable selection procedure to estimate and select relevant covariates for FMLR models.

The above-proposed methods do not jointly select the order selection and significant variables in FMLR models. In fact, it is a challenging issue, although some literature exists to solve this problem. Ref. [14] introduced MR-Lasso for FMLR models to simultaneously identify the order selection and significant variables. However, they do not study the large sample properties of the proposed method. Ref. [15] proposed a robust mixture regression

estimator via an asymmetric exponential power distribution and [16] studied component selection for exponential power mixture models, while they did not consider the variable selection procedure. Ref. [17] applied the penalized method on the number of components and regression coefficients to conduct model selection for FMLR models, but the error followed a normal distribution. Therefore, the proposed method is very sensitive to the heavy-tailed distribution.

In this paper, motivated by [10,18], we propose a new model selection procedure for the FMLR models via an exponential power distribution, which includes normal distributions and Laplace distributions as special cases. Under some regularity conditions, we investigate the asymptotic properties of the proposed method. In addition, we introduce an expectation-maximization (EM) algorithm [19] and a majorization-maximization (MM) algorithm [20] to solve the proposed optimization problem. The finite sample performance of the proposed method is illustrated via some numerical simulations. Results indicate that the proposed method is more robust to the heavy-tailed distributions than the existing method.

The rest of this paper is organized as follows. In Section 2, we present the finite mixture of regression models with an exponential power distribution and a penalized likelihood-based model selection approach. The asymptotic properties of the resulting estimates are investigated. In Section 3, a modified EM algorithm and an MM algorithm are developed to maximize the penalized likelihood. In Section 4, we propose a data-driven procedure to select the tuning parameters. In Section 5, simulation studies are conducted to evaluate the finite sample performance of the proposed method. In Section 6, a real data set is analyzed to compare the proposed test with some existing methods. We conclude with some remarks in Section 7. Technical conditions and proofs are given in the Appendix A.

2. Methodology

The density function of an exponential power (EP) distribution is defined as follows:

$$f_p(x;0,\sigma) = \frac{p}{\Gamma(\frac{1}{p})2^{1+\frac{1}{p}}\sigma}\exp\left(-\frac{1}{2}|\frac{x}{\sigma}|^p\right),$$

where $p > 0, \sigma > 0$ is the scale parameter, and $\Gamma(\cdot)$ is the Gamma function. When $0 < p < 2$, the EP distribution is heavy-tailed, which indicates that it can provide protection against outliers. The EP density function is a flexible and general density function class, and includes some important statistical density functions as its special cases, e.g., Gaussian density function ($p = 2$), and Laplace density function ($p = 1$). Meanwhile, the EP distribution has a wide range of applications, particularly in the area of business applications [21].

Based on the EP density function, we study the FMLR models. Let Z be a latent class variable with $P(Z = j|x) = \pi_j$ for $j = 1, 2, \cdots, m$, where $\mathbf{X}$ is a p-dimensional vector. Given $Z = j$, suppose that the response Y depends on $\mathbf{X}$ in a linear way

$$Y = \mathbf{X}^T\boldsymbol{\beta}_j + \epsilon_j,$$

where $\boldsymbol{\beta}_j$ is a p-dimensional vector, and ϵ_j is a random error with an EP density function $f_{p_j}(x;0,\sigma_j)$. Then the conditional density of Y given $\mathbf{X}$ can be written as

$$f(y|x) = \sum_{j=1}^{m}\pi_j f_{p_j}(Y - \mathbf{X}^T\boldsymbol{\beta}_j;0,\sigma_j). \tag{1}$$

Let $\{(\mathbf{X}_1, Y_1), \cdots, (\mathbf{X}_n, Y_n)\}$ be a random sample from (1). Then, the log-likelihood function for observations $\{(\mathbf{X}_1, Y_1), \cdots, (\mathbf{X}_n, Y_n)\}$ is given by

$$Q_n(\boldsymbol{\theta}) = \sum_{i=1}^{n}\log\left[\sum_{j=1}^{m}\pi_j f_{p_j}(Y_i - \mathbf{X}_i^T\boldsymbol{\beta}_j;0,\sigma_j)\right],$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}_{11}, \cdots, \boldsymbol{\beta}_{1p}, \cdots, \boldsymbol{\beta}_{m1}, \cdots, \boldsymbol{\beta}_{mp}, \sigma_1, \cdots, \sigma_m, p_1, \cdots, p_m, \pi_1, \cdots, \pi_{m-1})$.

To deal with the model selection problem, according to [10], we consider the following objective function,

$$\tilde{Q}_n(\boldsymbol{\theta}) = Q_n(\boldsymbol{\theta}) - P_n^1(\boldsymbol{\theta}) - P_n^2(\boldsymbol{\theta}) \tag{2}$$

with the penalty function

$$P_n^1(\boldsymbol{\theta}) = n \sum_{j=1}^{m} \left\{ \sum_{t=1}^{p} p_{\lambda_1}(|\boldsymbol{\beta}_{jt}|) \right\},$$

$$P_n^2(\boldsymbol{\theta}) = n\lambda_2 \sum_{j=1}^{m} \left[\log(\epsilon + p_{\lambda_2}(\pi_j)) - \log(\epsilon) \right],$$

where $p_\lambda(\cdot)$ is a non-negative and non-decreasing function, and $\lambda_1 > 0$ and $\lambda_2 > 0$ are two penalized parameters. Thus, we can obtain the estimators $\hat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}$ as follows

$$\hat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta}} \tilde{Q}_n(\boldsymbol{\theta}). \tag{3}$$

To derive some theoretical properties of the estimators $\hat{\boldsymbol{\theta}}_n$, we first define

$$a_n = \max_{j,t} \{ p'_{\lambda_1}(\boldsymbol{\beta}_{jt}^0)/\sqrt{n}, p'_{\lambda_2}(\pi_j^0)/\sqrt{n} : \boldsymbol{\beta}_{jt}^0 \neq 0, \pi_j^0 \neq 0 \},$$

$$b_n = \max_{j,t} \{ p''_{\lambda_1}(\boldsymbol{\beta}_{jt}^0)/n, p''_{\lambda_2}(\pi_j^0)/n : \boldsymbol{\beta}_{jt}^0 \neq 0, \pi_j^0 \neq 0 \},$$

where $p'_\lambda(h)$ and $p''_\lambda(h)$ are the first and second derivatives of the function $p_\lambda(h)$ with respect to h. To establish the asymptotic properties of the proposed estimators, we assume the following regularity conditions:

(C1) For any λ, $p_\lambda(0) = 0$, and $p_\lambda(\cdot)$ is non-negative and symmetric. Furthermore, it is non-decreasing and twice differentiable in $(0, \infty)$ with at most a few exceptions.

(C2) As $n \to \infty$, $b_n = o(1)$.

(C3)

$$\lim_{n\to\infty} \inf_{0<h\leq n^{-1/2}\log n} \sqrt{n} p'_\lambda(h) = \infty.$$

(C4) The joint density $f(\mathbf{z}, \boldsymbol{\theta})$ of $\mathbf{Z} = (\mathbf{X}, Y)$ have the third partial derivatives with respect to $\boldsymbol{\theta}$ for almost all $\mathbf{z}$.

(C5) For each $\boldsymbol{\theta}_0$, there exists $R_1(\mathbf{z})$ and $R_2(\mathbf{z})$ such that for $\boldsymbol{\theta}$ in a neighborhood $N(\boldsymbol{\theta}_0)$ of $\boldsymbol{\theta}_0$,

$$\left| \frac{\partial f(\mathbf{z};\boldsymbol{\theta})}{\partial \theta_i} \right| \leq R_1(\mathbf{z}), \left| \frac{\partial^2 f(\mathbf{z};\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right| \leq R_1(\mathbf{z}), \left| \frac{\partial^3 f(\mathbf{z};\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j \partial \theta_k} \right| \leq R_2(\mathbf{z}),$$

where $\boldsymbol{\theta}_0$ is the true parameter, $R_1(\mathbf{z})$ and $R_2(\mathbf{z})$ satisfy $\int R_1(\mathbf{z}) d\mathbf{z} < \infty$, and $\int R_2(\mathbf{z}) f(\mathbf{z};\boldsymbol{\theta}) d\mathbf{z} < \infty$.

(C6) The Fisher information matrix $Q(\boldsymbol{\theta})$ is finite and positive definite at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, where $Q(\boldsymbol{\theta})$ is defined as follows,

$$Q(\boldsymbol{\theta}) = E\left\{ \left[\frac{\partial}{\partial \boldsymbol{\theta}} \log(f(\mathbf{Z};\boldsymbol{\theta})) \right] \left[\frac{\partial}{\partial \boldsymbol{\theta}} \log(f(\mathbf{Z};\boldsymbol{\theta})) \right]^T \right\}.$$

(C7) $p_j > 1, j = 1, \cdots, m$.

(C8) $c_1 \leq \sigma_j^2 \leq c_2$, $\left\|\boldsymbol{\beta}_j\right\| \leq c_3, j = 1, \cdots, m$, where c_1 is some positive constant, c_2 and c_3 are some large constants.

Remark 1. *Conditions C1–C3 are the assumption about the penalty function, and assure that the variable selection of the proposed estimators is consistent. The similar conditions are also used in [22]. Condition (C5) ensures that the main term dominates the remainder in the Taylor expansion. Conditions (C4)–(C6) are used in [17]. Condition (C7) ensures the concavity of the likelihood function since the log likelihood function of random sample from EP distribution is concave if $p > 1$. Condition (C8) ensures the compactness of parameter space. Conditions (C7) and (C8) are similarly applied in Wang and Feng [16].*

In the following, we have two theorems with proofs given in the Appendix A.

Theorem 1. *Under the conditions (C1), (C2), (C4)–(C8), and if* $\sqrt{n}\min\{\lambda_1, \lambda_2\} \to \infty$, *and* $\min\{\lambda_1, \lambda_2\} \to 0$, *then there exists a local maximizer* $\hat{\boldsymbol{\theta}}_n$ *of the penalized log-likelihood function (2) such that*

$$\left|\left|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right|\right| = O_p(n^{-1/2}).$$

Theorem 2. *Under the conditions (C1)–(C8), and if* $\sqrt{n}\min\{\lambda_1, \lambda_2\} \to \infty$, *and* $\min\{\lambda_1, \lambda_2\} \to 0$. *Then, for any* $\sqrt{n}$*-consistent estimator* $\hat{\boldsymbol{\theta}}_n$ *of* $\boldsymbol{\theta}$, *we have*

(a) *Sparsity:* $P\{\hat{\pi}_k = 0\} \to 1$ *as* $n \to \infty$, *where* $k = m_0 + 1, \cdots, m$.

(b) *Sparsity:* $P\{\hat{\boldsymbol{\beta}}_{kj} = 0\} \to 1$ *as* $n \to \infty$, *where* $k = 1, \cdots, m_0$ *and* $j = 1, \cdots, t_k$.

(c) *Asymptotic normality:*

$$\sqrt{n}\left\{\left[Q_1(\boldsymbol{\theta}_{01}) - \frac{P_n^{1''}(\boldsymbol{\theta}_{01})}{n} - \frac{P_n^{2''}(\boldsymbol{\theta}_{01})}{n}\right](\hat{\boldsymbol{\theta}}_{n1} - \boldsymbol{\theta}_{01}) + \frac{P_n^{1'}(\boldsymbol{\theta}_{01})}{n} + \frac{P_n^{2'}(\boldsymbol{\theta}_{01})}{n}\right\} \xrightarrow{D} N(0, Q_1(\boldsymbol{\theta}_{01})),$$

where m_0 *is the number of true non-zero mixing weights,* $\boldsymbol{\theta}_{01}$ *and* $Q_1(\boldsymbol{\theta}_{01})$ *are the true parameter and the corresponding Fisher information when all zero effects are removed, respectively.*

3. Algorithm

In this section, we apply a modified EM algorithm and an MM algorithm to solve the proposed optimization problem (3). Let z_{ij} be the indicator variables that show if the i-th observation arises from the j-th component as missing data, and p_{ij} is the posterior probability that the i-th observation belongs to the j-th component. Therefore, the expected complete-data log-likelihood function is given as follows:

$$\sum_{i=1}^{n}\sum_{j=1}^{m} z_{ij}\log\left[\pi_j f_{p_j}(Y_i - \mathbf{X}_i^T\boldsymbol{\beta}_j; 0, \sigma_j)\right].$$

Then, the objective function (2) is rewritten as

$$\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}\log\left[\pi_j f_{p_j}(Y_i - \mathbf{X}_i^T\boldsymbol{\beta}_j; 0, \sigma_j)\right] - P_n^1(\boldsymbol{\theta}) - P_n^2(\boldsymbol{\theta}). \tag{4}$$

Next, we apply a modified EM algorithm to maximize the objective function (4). The detailed procedure is given as follows:

Step 1 Given the l-th approximation

$$\hat{\boldsymbol{\theta}}^{(l)} = (\hat{\boldsymbol{\beta}}_{11}^{(l)}, \cdots, \hat{\boldsymbol{\beta}}_{1p}^{(l)}, \cdots, \hat{\boldsymbol{\beta}}_{m1}^{(l)}, \cdots, \hat{\boldsymbol{\beta}}_{mp}^{(l)}, \hat{\sigma}_1^{(l)}, \cdots, \hat{\sigma}_m^{(l)}, \hat{p}_1^{(l)}, \cdots, \hat{p}_m^{(l)}, \hat{\pi}_1^{(l)}, \cdots, \hat{\pi}_{m-1}^{(l)}),$$

we can calculate the classification probabilities:

$$\hat{p}_{ij}^{(l+1)} = \frac{\hat{\pi}_j^{(l)} f_{\hat{p}_j^{(l)}}(Y_i - \mathbf{X}_i^T \hat{\boldsymbol{\beta}}_j^{(l)}; 0, \hat{\sigma}_j^{(l)})}{\sum_{j=1}^m \hat{\pi}_j^{(l)} f_{\hat{p}_j^{(l)}}(Y_i - \mathbf{X}_i^T \hat{\boldsymbol{\beta}}_j^{(l)}; 0, \hat{\sigma}_j^{(l)})}.$$

Step 2 We first update $\{\pi_1, \cdots, \pi_m\}$. We use a Lagrange multiplier δ to take into account for the constraint $\sum_{j=1}^m \pi_j = 1$, then we have

$$\frac{\partial}{\partial \pi_j} \left\{ \sum_{i=1}^n \sum_{j=1}^m \hat{p}_{ij}^{(l+1)} \log(\pi_j) - n\lambda_2 \sum_{j=1}^m \left[\log\left(\epsilon + p_{\lambda_2}(\pi_j)\right) \right] - \delta\left(\sum_{j=1}^m \pi_j - 1\right) \right\} = 0. \tag{5}$$

In (5), we apply the local linear approximation [23] to $\log\left(\epsilon + p_{\lambda_2}(\pi_j)\right)$,

$$\log\left(\epsilon + p_{\lambda_2}(\pi_j)\right) \approx \log\left(\epsilon + p_{\lambda_2}(\hat{\pi}_j^{(l)})\right) + \frac{p'_{\lambda_2}(\hat{\pi}_j^{(l)})}{\epsilon + p_{\lambda_2}(\hat{\pi}_j^{(l)})}(\pi_j - \hat{\pi}_j^{(l)}).$$

Then, π_j can be updated by straightforward calculations,

$$\hat{\pi}_j^{(l+1)} = \frac{1}{D_j} \sum_{i=1}^n \hat{p}_{ij}^{(l+1)},$$

where

$$D_j = n \left[1 - \lambda_2 \sum_{j=1}^m \frac{\hat{\pi}_j^{(l)} p'_{\lambda_2}(\hat{\pi}_j^{(l)})}{\epsilon + p_{\lambda_2}(\hat{\pi}_j^{(l)})} + \lambda_2 \frac{p'_{\lambda_2}(\hat{\pi}_j^{(l)})}{\epsilon + p_{\lambda_2}(\hat{\pi}_j^{(l)})} \right].$$

Next, we update $\{\boldsymbol{\beta}_{11}, \cdots, \boldsymbol{\beta}_{1p}, \cdots, \boldsymbol{\beta}_{m1}, \cdots, \boldsymbol{\beta}_{mp}, \sigma_1, \cdots, \sigma_m, p_1, \cdots, p_m\}$ by maximizing the following objective function,

$$\sum_{i=1}^n \sum_{j=1}^m \hat{p}_{ij}^{(l+1)} \log\left[\hat{\pi}_j^{(l+1)} f_{p_j}(Y_i - \mathbf{X}_i^T \boldsymbol{\beta}_j; 0, \sigma_j) \right] - n \sum_{j=1}^m \left\{ \sum_{t=1}^p p_{\lambda_1}(|\boldsymbol{\beta}_{jt}|) \right\}.$$

We first update $\{\sigma_1, \cdots, \sigma_m\}$. For each $\sigma_j, j = 1, 2, \cdots, m$, we only need to maximize

$$\sum_{i=1}^n \hat{p}_{ij}^{(l+1)} \left(-\log(\sigma_j) - \frac{1}{2} \left| \frac{Y_i - \mathbf{X}_i^T \hat{\boldsymbol{\beta}}_j^{(l)}}{\sigma_j} \right|^{\hat{p}_j^{(l)}} \right).$$

Then, the resulting estimator is given as follows:

$$\hat{\sigma}_j^{(l+1)} = \frac{\sum_{i=1}^n \hat{p}_{ij}^{(l+1)} |Y_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}_j^{(l)}|^{\hat{p}_j^{(l)}}}{2 \sum_{i=1}^n \hat{p}_{ij}^{(l+1)}}.$$

Next, we update $\{p_1, \cdots, p_m\}$. For each $p_j, j = 1, 2 \cdots, m$, according to the condition (C7), we have

$$\hat{p}_j^{(l+1)} = \arg\max_{p_j > 1} \sum_{i=1}^n \hat{p}_{ij}^{(l+1)} \left\{ \log(p_j) - log(\Gamma(\frac{1}{p_j}) - (1 + \frac{1}{p_j}) \log(2) - \frac{1}{2} \left| \frac{Y_i - \mathbf{X}_i^T \boldsymbol{\beta}_j^{(l)}}{\hat{\sigma}_j^{(l+1)}} \right|^{p_j} \right\}.$$

Finally, we update $\{\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_m\}$. By ignoring some terms which do not involve in $\boldsymbol{\beta}_j$, we have

$$L(\boldsymbol{\beta}_j) = -\sum_{i=1}^{n} \hat{p}_{ij}^{(l+1)} \frac{1}{2\hat{\sigma}_j^{(l+1)}} |Y_i - \mathbf{X}_i^T \boldsymbol{\beta}_j|^{\hat{p}_j^{(l+1)}} - n \sum_{t=1}^{p} p_{\lambda_1}(|\beta_{jt}|).$$

By using a MM algorithm for $L(\boldsymbol{\beta}_j)$'s the first term, we have

$$\left\{(Y_i - \mathbf{X}_i^T \boldsymbol{\beta}_j)^T (Y_i - \mathbf{X}_i^T \boldsymbol{\beta}_j)\right\}^{\frac{\hat{p}_j^{(l+1)}}{2}} \leq \left\{(Y_i - \mathbf{X}_i^T \hat{\boldsymbol{\beta}}_j^{(l)})^T (Y_i - \mathbf{X}_i^T \hat{\boldsymbol{\beta}}_j^{(l)})\right\}^{\frac{\hat{p}_j^{(l+1)}}{2}}$$
$$+ \frac{\hat{p}_j^{(l+1)}}{2} \left\{(Y_i - \mathbf{X}_i^T \hat{\boldsymbol{\beta}}_j^{(l)})^T (Y_i - \mathbf{X}_i^T \hat{\boldsymbol{\beta}}_j^{(l)})\right\}^{\frac{\hat{p}_j^{(l+1)}}{2} - 1} \left\{(Y_i - \mathbf{X}_i^T \boldsymbol{\beta}_j)^T (Y_i - \mathbf{X}_i^T \boldsymbol{\beta}_j) - (Y_i - \mathbf{X}_i^T \hat{\boldsymbol{\beta}}_j^{(l)})^T (Y_i - \mathbf{X}_i^T \hat{\boldsymbol{\beta}}_j^{(l)})\right\}.$$

For $p_{\lambda_1}(|\beta_{jt}|)$, we apply a local quadratic approximation [22], then we have

$$p_{\lambda_1}(|\beta_{jt}|) \approx p_{\lambda_1}(|\hat{\beta}_{jt}^{(l)}|) + \frac{p'_{\lambda_1}(|\hat{\beta}_{jt}^{(l)}|)}{2|\hat{\beta}_{jt}^{(l)}|}(\beta_{jt}^2 - \hat{\beta}_{jt}^{(l)2}).$$

Thus, for each $\boldsymbol{\beta}_j, j = 1, 2, \cdots, m$, we only need to solve the following minimization problem

$$\hat{\boldsymbol{\beta}}_j^{(l+1)} = \arg\min_{\boldsymbol{\beta}_j} \left[\sum_{i=1}^{n} \hat{p}_{ij}^{(l+1)} \frac{1}{2\hat{\sigma}_j^{(l+1)}} \hat{w}_{ij}^{(l)} (Y_i - \mathbf{X}_i^T \boldsymbol{\beta}_j)^T (Y_i - \mathbf{X}_i^T \boldsymbol{\beta}_j) + n \sum_{t=1}^{p} \beta_{jt}^2 \frac{p'_{\lambda_1}(|\hat{\beta}_{jt}^{(l)}|)}{2|\hat{\beta}_{jt}^{(l)}|} \right],$$

where $\hat{w}_{ij}^{(l)} = \frac{\hat{p}_j^{(l+1)}}{2} \left\{(Y_i - \mathbf{X}_i^T \hat{\boldsymbol{\beta}}_j^{(l)})^T (Y_i - \mathbf{X}_i^T \hat{\boldsymbol{\beta}}_j^{(l)})\right\}^{\frac{\hat{p}_j^{(l+1)}}{2} - 1}$.

Thus, we can update $\boldsymbol{\beta}_j$ as follows

$$\hat{\boldsymbol{\beta}}_j^{(l+1)} = (\mathbf{X} B \mathbf{X}^T + A)^{-1} \mathbf{X} B Y,$$

where

$$A = n * diag\left\{ \frac{p'_{\lambda_1}(|\hat{\boldsymbol{\beta}}_{j1}^{(l)}|)}{2|\hat{\boldsymbol{\beta}}_{j1}^{(l)}|}, \cdots, \frac{p'_{\lambda_1}(|\hat{\boldsymbol{\beta}}_{jp}^{(l)}|)}{2|\hat{\boldsymbol{\beta}}_{jp}^{(l)}|} \right\},$$
$$B = diag\left\{ \hat{p}_{1j}^{(l+1)} \frac{1}{2\hat{\sigma}_j^{(l+1)}} \hat{w}_{1j}^{(l)}, \hat{p}_{2j}^{(l+1)} \frac{1}{2\hat{\sigma}_j^{(l+1)}} \hat{w}_{2j}^{(l)}, \cdots, \hat{p}_{nj}^{(l+1)} \frac{1}{2\hat{\sigma}_j^{(l+1)}} \hat{w}_{nj}^{(l)} \right\}.$$

Step 3 Repeat **Step 1**, and **Step 2** until convergence.

4. Choice of the Tuning Parameters

The selection of tuning parameters is a vital part in the order selection and variable selection procedure. In order to guarantee that a true model can be chosen correctly, we should select the proper tuning parameters λ_1 and λ_2 in the process of practice. There are lots of methods to select λ_1 and λ_2, such as cross-validation (CV), generalized cross-validation (GCV), AIC, and BIC.

As suggested in [24], we introduce a data-driven procedure to choose the tuning parameters λ_1 and λ_2 by minimizing the following modified Bayesian information criterion,

$$MBIC(\lambda_1, \lambda_2) = -2 \sum_{i=1}^{n} \log\left\{ \sum_{j=1}^{\hat{m}} \hat{\pi}_j f_{\hat{p}_j}(Y_i - \mathbf{X}_i^T \hat{\boldsymbol{\beta}}_j; 0, \hat{\sigma}_j) \right\} + \log n * df, \tag{6}$$

where $\hat{m}$ denotes the estimate of the number of components, $df = 3\hat{m} - 1 + \hat{\mathcal{M}}_{\boldsymbol{\beta}}$, and

$$\hat{\mathcal{M}}_{\boldsymbol{\beta}} = \#\Big\{|\hat{\beta}_{jt}| > 10^{-3}, j = 1, \cdots, \hat{m}, t = 1, \ldots, p\Big\}.$$

5. Simulation

In this section, we use some numerical simulations to illustrate the finite sample performance of the proposed method. For the penalty function, we use the SCAD penalty [22], which is given as follows:

$$p_\lambda(t;a) = \begin{cases} \lambda|t|, & \text{if } |t| \le \lambda, \\ -(t^2 - 2a\lambda|t| + \lambda^2)/[2(a-1)], & \text{if } \lambda < |t| \le a\lambda, \\ (a+1)\lambda^2/2, & \text{otherwise,} \end{cases}$$

where λ is a tuning parameter and $a > 2$. According to the suggestion in Fan and Li [22], a is equal to 3.7 by minimizing the Bayes risks. The datasets are generated via a three-component FMLR model

$$f(y|\mathbf{x}) = \sum_{j=1}^{3} \pi_j f_{p_j}(y - \mathbf{x}^T\boldsymbol{\beta}_j; 0, \sigma_j), \tag{7}$$

where the components of $\mathbf{x}$ are generated independently from the 7-dimensional standard normal distribution. In detail, we generate random samples of each component from the following linear model

$$Y = \mathbf{X}^T\boldsymbol{\beta} + \epsilon.$$

We simulate 100 datasets from the FMLR model (7) with sample size of n=200, 600, 800, 1000. The datasets are generated by the following four scenarios:

Scenario 1. $\boldsymbol{\beta}_1 = (1,1,1,1,0,0,0)^T, \boldsymbol{\beta}_2 = (1,2,3,4,0,0,0)^T, \boldsymbol{\beta}_3 = (5,6,7,8,0,0,0)^T$ and $\pi_1 = 0.4, \pi_2 = 0.3, \pi_3 = 0.3$, and the random error $\epsilon \sim N(0,1)$;

Scenario 2. We use the same setting as in **Scenario 1**, except that the error term follows a t-distribution with freedom degree 2;

Scenario 3. We use the same setting as in **Scenario 1**, except that the error term follows a mixture t distribution: $\epsilon \sim 0.5t(1) + 0.5t(3)$;

Scenario 4. We use the same setting as in **Scenario 1**, except that the error term follows a mixture normal distribution: $\epsilon \sim 0.95N(0,1) + 0.05N(0.5^2)$.

We compare our proposed method with the method proposed by [17]. To assess the finite-sample performance, we consider four different measures:

(1) $RMSE_{\pi_j}$: the root mean square error of $\hat{\pi}_j$ when the order is corrected estimated, which is defined by

$$RMSE_{\pi_j} = \sqrt{\frac{1}{M^*}\sum_{m=1}^{M^*}(\hat{\pi}_j^m - \pi_j)^T(\hat{\pi}_j^m - \pi_j)}$$

where M^* is the number of simulations with correct estimation of the order.

(2) $RMSE_{\boldsymbol{\beta}_c}$: the root mean square error of $\hat{\boldsymbol{\beta}}_j$, which can be similarly calculated as $RMSE_{\pi_j}$.

(3) NCZ (the number of correct zeros): It denotes that the number of the true value of the parameter is zero and is correctly estimated as zero. NCZ can be calculated by

$$NCZ = \#\big\{t : \beta_t = 0 \wedge \hat{\beta}_t = 0\big\},$$

where $\#\{A\}$ denotes the number of elements within A.

(4) NIZ (the number of incorrect zeros): It indicates that the number of the true value of the parameter is non-zero and is incorrectly estimated as zero. NIZ is given as follows:

$$NIZ = \#\big\{t : \beta_t \neq 0 \wedge \hat{\beta}_t = 0\big\}.$$

In simulation studies, suppose we know that the data come from a mixture regression model with at most five components, but the true number of components should be estimated. For each scenario, the simulation is repeated 100 times. The corresponding results are shown in Tables 1–8. In these Tables, $M1$ and $M2$ denote the results by [17] and our proposed method, respectively.

Table 1. Order selection results in Scenario 1.

n	$M1$			$M2$		
	Underfitted	Correctly Fitted	Overfitted	Underfitted	Correctly Fitted	Overfitted
200	0.00	0.99	0.01	0.40	0.60	0.00
600	0.00	0.99	0.01	0.00	0.99	0.01
800	0.00	0.99	0.01	0.00	0.99	0.01
1000	0.00	1.00	0.00	0.00	0.99	0.01

Table 2. Variable selection and parameter estimation results in Scenario 1.

n	$M1$				$M2$			
	$RMSE_{\pi_c}$	$RMSE_{\beta_c}$	NCZ	NIZ	$RMSE_{\pi_c}$	$RMSE_{\beta_c}$	NCZ	NIZ
200	0.092	0.535	2.900	0.200	0.143	0.352	2.667	0.000
	0.048	0.596	2.700	0.000	0.141	0.207	2.333	0.000
	0.073	0.703	2.670	0.000	0.048	0.427	2.833	0.000
600	0.024	0.154	2.990	0.000	0.023	0.156	2.990	0.000
	0.025	0.153	2.980	0.000	0.025	0.151	2.990	0.000
	0.021	0.154	2.990	0.000	0.022	0.156	2.980	0.000
800	0.022	0.142	2.990	0.000	0.020	0.145	3.000	0.000
	0.020	0.138	2.980	0.000	0.021	0.153	3.000	0.000
	0.019	0.141	2.990	0.000	0.020	0.138	3.000	0.000
1000	0.014	0.123	2.990	0.000	0.015	0.130	3.000	0.000
	0.016	0.122	3.000	0.000	0.014	0.121	3.000	0.000
	0.014	0.121	3.000	0.000	0.014	0.122	3.000	0.000

Table 3. Order selection results in Scenario 2.

n	$M1$			$M2$		
	Underfitted	Correctly Fitted	Overfitted	Underfitted	Correctly Fitted	Overfitted
200	0.50	0.20	0.30	0.16	0.64	0.20
600	0.07	0.81	0.12	0.00	0.99	0.01
800	0.03	0.75	0.22	0.01	0.98	0.01
1000	0.11	0.84	0.05	0.00	0.99	0.01

Table 1 shows the simulation results of order selection. Columns labeled "Underfitted" are the proportion of the fitted model with less than three components in 100 simulations. Meanwhile, "Correctly fitted" and "Overfitted" can be similarly interpreted. From Table 1, we can find that the effects of the two models are very similar, and the accuracy rate of order selection can reach more than 98% for $M1$ and $M2$ when n is larger than or equal to 600. Table 2 presents the results of variable selection and parameter estimation for each component. From Table 2, we observe that the finite sample performances of the two models are very similar for $n \geq 600$. Therefore, when the error term follows a normal distribution, the two models have similar performance when the sample size is sufficiently large.

Tables 3 and 4 present the results of Scenario 2, which is a heavy-tailed scenario. We can observe from Table 3 that $M1$ can only estimate about 20% underfitted or overfitted model, while our method keeps robustness and continues to maintain 98% accuracy when $n \geq 600$. In Table 4, $M1$ has a poor performance in variable selection. $M1$ has many

non-zero NIZ, while our method's NIZ is all zero for $n \geq 600$. Meanwhile, the NCZ of our proposed method increases as n increases. In addition, our proposed method has a smaller RMSE than M1.

Table 4. Variable selection and parameter estimation results in Scenario 2.

n	*M*1				*M*2			
	$RMSE_{\pi_c}$	$RMSE_{\beta_c}$	*NCZ*	*NIZ*	$RMSE_{\pi_c}$	$RMSE_{\beta_c}$	*NCZ*	*NIZ*
200	0.285	8.381	2.500	0.000	0.088	0.964	2.722	0.056
	0.126	0.457	1.500	0.000	0.058	1.957	2.778	0.076
	0.223	2.876	2.000	2.000	0.090	0.852	2.772	0.000
600	0.057	0.671	2.893	0.000	0.048	0.261	2.963	0.000
	0.062	1.119	2.844	0.011	0.040	0.240	2.876	0.000
	0.055	1.264	2.872	0.034	0.044	0.264	2.896	0.000
800	0.065	0.715	2.897	0.013	0.034	0.223	2.845	0.000
	0.053	0.874	2.892	0.012	0.032	0.228	2.957	0.000
	0.047	1.241	2.887	0.000	0.033	0.193	2.929	0.000
1000	0.063	0.905	2.912	0.012	0.029	0.198	2.906	0.000
	0.056	0.926	2.923	0.011	0.031	0.191	2.946	0.000
	0.047	0.837	2.921	0.012	0.033	0.188	2.979	0.000

For Table 5, the performance of order selection for $M1$ is worse than that for $M2$. The ratio of the correctly fitted model remains above 98% with our method for $n \geq 600$, while $M1$ is easy to overfit the model's components. In Table 6, it can be seen that the NCZ value of $M1$ is a little better than that of $M2$. Compared with $RMSE_{\beta_c}$, we can find that our method is better than $M1$ consistently.

Table 5. Order selection results in Scenario 3.

n	*M*1			*M*2		
	Underfitted	**Correctly Fitted**	**Overfitted**	**Underfitted**	**Correctly Fitted**	**Overfitted**
200	0.60	0.25	0.15	0.32	0.66	0.02
600	0.00	0.74	0.26	0.00	0.98	0.02
800	0.03	0.73	0.24	0.00	0.99	0.01
1000	0.05	0.79	0.16	0.00	0.99	0.01

Table 6. Variable selection and parameter estimation results in Scenario 3.

n	*M*1				*M*2			
	$RMSE_{\pi_c}$	$RMSE_{\beta_c}$	*NCZ*	*NIZ*	$RMSE_{\pi_c}$	$RMSE_{\beta_c}$	*NCZ*	*NIZ*
200	0.236	2.312	2.000	0.000	0.039	0.766	3.000	0.500
	0.165	3.903	2.400	0.200	0.054	0.969	3.000	0.167
	0.060	1.732	2.600	1.400	0.049	0.929	3.000	0.667
600	0.025	0.164	2.887	0.000	0.023	0.122	2.874	0.000
	0.025	0.156	2.889	0.000	0.024	0.127	2.869	0.000
	0.027	0.162	2.896	0.000	0.025	0.124	2.877	0.000
800	0.024	0.154	2.893	0.000	0.018	0.114	2.878	0.000
	0.023	0.137	2.886	0.000	0.017	0.123	2.931	0.000
	0.019	0.134	2.897	0.000	0.017	0.123	2.931	0.000
1000	0.021	0.132	2.894	0.000	0.017	0.097	2.924	0.000
	0.020	0.138	2.924	0.000	0.017	0.097	2.924	0.000
	0.019	0.122	2.891	0.000	0.016	0.114	2.971	0.000

Tables 7 and 8 present the results of Scenario 4. $M1$ absolutely stays away from the right number of components. On the contrary, our method can select the correct number

of components with 98% accuracy for $n \geq 600$. In Table 8, $M1$ is better than $M2$ in NCZ, but $M1$ is unstable in NIZ. Comparing $RMSE_{\beta_c}$, we can find that $M1$ is larger than $M2$. In general, our model is better than $M1$ in both order selection and variable selection and parameter estimation.

Table 7. Order selection results in Scenario 4.

n	*M*1			*M*2		
	Underfitted	Correctly Fitted	Overfitted	Underfitted	Correctly Fitted	Overfitted
200	0.49	0.41	0.10	0.07	0.72	0.21
600	0.13	0.28	0.59	0.02	0.98	0.00
800	0.19	0.31	0.50	0.00	1.00	0.00
1000	0.10	0.39	0.51	0.01	0.99	0.00

Table 8. Variable selection and parameter estimation results in Scenario 4.

n	*M*1				*M*2			
	$RMSE_{\pi_c}$	$RMSE_{\beta_c}$	*NCZ*	*NIZ*	$RMSE_{\pi_c}$	$RMSE_{\beta_c}$	*NCZ*	*NIZ*
200	0.014	5.455	2.889	0.444	0.218	0.971	2.500	0.000
	0.180	1.337	2.889	0.222	0.228	1.669	2.000	0.000
	0.079	2.956	2.889	0.000	0.320	2.284	2.250	0.000
600	0.050	2.672	2.832	0.000	0.054	0.372	2.776	0.000
	0.053	1.256	2.717	0.000	0.055	0.273	2.724	0.000
	0.054	2.136	2.846	0.038	0.061	0.271	2.878	0.000
800	0.051	1.134	2.811	0.000	0.035	0.398	2.600	0.000
	0.045	1.535	2.623	0.000	0.039	0.163	2.600	0.000
	0.047	2.724	2.747	0.000	0.022	0.183	3.000	0.000
1000	0.074	1.217	2.942	0.000	0.035	0.347	2.973	0.000
	0.073	1.736	2.974	0.103	0.031	0.220	2.697	0.000
	0.047	3.734	2.772	0.000	0.025	0.129	2.949	0.000

6. Real Data Analysis

In this section, we apply the proposed methodology to analyze baseball salary data, which consists of information about major league baseball players. The response variable is their 1992 salaries (measured in thousands of dollars). In addition, there are 16 performance measures for 337 MLB players who participated in at least one game in both the 1991 and 1992 seasons. This data set has been analyzed by others, such as [12,17]. We want to study how the performance measures affect salaries using our method.

The performance measures are batting average (x_1), on-base percentage (x_2), runs (x_3), hits (x_4), doubles (x_5), triples (x_6), home runs (x_7), runs batted in (x_8), walks (x_9), strikeouts (x_{10}), stolen bases (x_{11}), and errors (x_{12}); and indicators of free agency eligibility (x_{13}), free agent in 1991/2 (x_{14}), arbitration eligibility (x_{15}), and arbitration in 1991/2 (x_{16}). The four (dummy) variables $x_{13} - x_{16}$ indicate how free each player was to move to another team. As suggested in [25], the interaction effects between (dummy) variables $x_{13} - x_{16}$ and the quantitative variables x_1, x_3, x_7, and x_8 should be added to the consideration. Therefore, we obtain a set of 32 potential covariates affecting each player's salary. Ref. [12] fitted a mixture of linear regression models with two or three components to depict the overlaid shape of the histogram of log(salary), and concluded that a two-component mixture regression model labeled MIXSCAD fitted the data well. [17] uses an FMLR model based on normal distribution, and the number of components is two.

As advocated by [12], we use log(salary) as the response variable. We first fit a linear model via stepwise regression, the results are shown in Table 9, denoted as $\hat{\boldsymbol{\beta}}_{ols}$. Based on [17], we consider the following four-component mixture model,

$$Y|X \sim \sum_{j=1}^{4} \pi_j f_{p_j}(Y - \mathbf{X}^T\boldsymbol{\beta}_j; 0, \sigma_j),$$

where Y = log(salary) and X is a 33×1 vector containing 32 covariates plus an intercept term. In order to implement the proposed modified EM algorithm, we set the initial values as follows

$$\boldsymbol{\pi}^0 = (0.4, 0.2, 0.2, 0.2)^T, \boldsymbol{\sigma}^0 = (10, 10, 10, 10)^T, \boldsymbol{p}^0 = (1, 1, 1, 1)^T, \boldsymbol{\beta}_j^0 = \hat{\boldsymbol{\beta}}_{ols} + \boldsymbol{\epsilon}_j$$

where $\boldsymbol{\epsilon}_j \sim N(0, \boldsymbol{I}), j = 1, 2, 3, 4$. The results are reported in Tables 9 and 10. From Table 9, we find that both $M1$ and $M2$ choose two components. Furthermore, we can observe from Table 10 that M2 has a smaller BIC value than M1, which indicates that our proposed method can better fit this dataset than M1.

Table 9. Parameter estimates for baseball salary data.

Covariates	Linear Model	M1		M2	
		Comp1	Comp2	Comp1	Comp2
x_0	5.48	4.81	5.66	4.70	4.67
x_1	-	-	-	-	-
x_2	-1.54	-	-	-	-
x_3	-	-	-	-	-
x_4	-	-	0.01	0.03	0.02
x_5	-	-	-	-	0.01
x_6	-	-	-	-	-
x_7	-	-	-	-	-
x_8	0.01	0.01	0.02	0.01	-
x_9	0.01	-	-	0.03	0.01
x_{10}	-0.01	-	-	-	-
x_{11}	-	0.03	-	-	0.01
x_{12}	-	-	-	-	-
x_{13}	1.52	2.04	-	3.13	2.16
x_{14}	-0.48	-	-	-	-
x_{15}	1.35	1.60	-	2.73	1.28
x_{16}	-	-	-	0.01	1.40
$x_1 * x_{13}$	-	-	-	-	-
$x_1 * x_{14}$	-	-	-	-	10.05
$x_1 * x_{15}$	-	-	-	0.01	-
$x_1 * x_{16}$	-4.38	-	-	-	-
$x_3 * x_{13}$	-	-	-	-	-
$x_3 * x_{14}$	-	-	-	-0.01	-0.02
$x_3 * x_{15}$	-	-	-	-	-
$x_3 * x_{16}$	-	-	-	0.01	-
$x_7 * x_{13}$	0.01	-	-	0.03	-
$x_7 * x_{14}$	0.03	-	-	-	0.02
$x_7 * x_{15}$	-	-	-	-	-
$x_7 * x_{16}$	-	-	-	-	-
$x_8 * x_{13}$	-	-	0.01	-	0.01
$x_8 * x_{14}$	-	0.01	-	0.01	-
$x_8 * x_{15}$	-	-	0.02	-	-
$x_8 * x_{16}$	0.02	-	-	-	0.02

Table 10. Model parameter estimates for baseball salary data.

Parameter	*M*1		*M*2	
$\hat{\pi}$	0.69	0.31	0.84	0.16
$\hat{p}$	2	2	1.05	1.49
$\hat{\sigma}$	-	-	2.27	7.13
λ_1	0.300		0.220	
λ_2	0.040		0.016	
MBIC	569.64		**547.25**	

Of interest is to explain how the performance measures affect salaries by interpreting the outcome of the fit, although it can be a source of controversy. Do not think about it; there should be many positive correlations between a baseball player and his salary. $M1$ and $M2$ have the same sign and approximate coefficients in x_0, x_{13}, x_{15}, and interactions of x_8 and x_{14}. Recall that x_1 and x_7 are individual performances, while x_{13}, x_{15}, and x_{16} are three dummy variables indicating how freely players change teams. For example, the effect of $x_1 * x_{16}$ implies that for most players having arbitration eligibility in 1991/2 enhances the individual ability (x_1) toward a lower salary, but not the value of their team contribution (x_8).

The main differences between the two models are interaction effects $x_1 * x_{14}$ and $x_1 * x_{15}$. This implies that $M1$ disregards $x_1 * x_{14}$'s effect, but $M2$ indicates that it is in two directions. And $M2$ attaches great importance to the interaction effect of $x_1 * x_{14}$.

7. Discussion

In this paper, we introduced the FMLR models via an exponential power distribution. Under some conditions, the asymptotic properties of the proposed estimators were established. Meanwhile, a modified EM algorithm and an MM algorithm were applied to solve the proposed optimization problem. Furthermore, the merits of our proposed methodology were illustrated through some numerical simulations and real data analysis. Simulation studies showed that the proposed method had better performance than the existing methods under difference errors. By analyzing a baseball salary dataset, our proposed method had a smaller BIC value than the method proposed [17].

Author Contributions: Methodology, Y.J. and J.L.; Formal analysis, H.Z. and X.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research is partially supported by NSFC (12171203), the Fundamental Research Funds for the Central Universities (23JNQMX21) and the Natural Science Foundation of Guangdong (2022A1515010045).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The data used in this study is publicly available. Code is available on request from the second authors.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Proof of Theorem 1. For any given $\epsilon > 0$, let $\|\mathbf{u}\| = M_\epsilon$. Denote

$$\Gamma_n(\mathbf{u}) = \tilde{Q}_n(\boldsymbol{\theta}_0 + \mathbf{u}/\sqrt{n}) - \tilde{Q}_n(\boldsymbol{\theta}_0).$$

According to (2), we have

$$\Gamma_n(\mathbf{u}) = [Q_n(\boldsymbol{\theta}_0 + \mathbf{u}/\sqrt{n}) - Q_n(\boldsymbol{\theta}_0)] - [P_n^1(\boldsymbol{\theta}_0 + \mathbf{u}/\sqrt{n}) - P_n^1(\boldsymbol{\theta}_0)] - [P_n^2(\boldsymbol{\theta}_0 + \mathbf{u}/\sqrt{n}) - P_n^2(\boldsymbol{\theta}_0)].$$

Under condition (C1), we have $p_\lambda(0) = 0$ for any λ. Therefore, $P_n^1(\boldsymbol{\theta}_0) = P_n^1(\boldsymbol{\theta}_{01})$ and $P_n^2(\boldsymbol{\theta}_0) = P_n^2(\boldsymbol{\theta}_{01})$. Since $P_n^1(\boldsymbol{\theta}_0 + \mathbf{u}/\sqrt{n})$ and $P_n^2(\boldsymbol{\theta}_0 + \mathbf{u}/\sqrt{n})$ are a sum of positive terms, we then have

$$\Gamma_n(\mathbf{u}) \leq [Q_n(\boldsymbol{\theta}_0 + \mathbf{u}/\sqrt{n}) - Q_n(\boldsymbol{\theta}_0)] - [P_n^1(\boldsymbol{\theta}_{01} + \mathbf{u}_1/\sqrt{n}) - P_n^1(\boldsymbol{\theta}_{01})] - [P_n^2(\boldsymbol{\theta}_{01} + \mathbf{u}_1/\sqrt{n}) - P_n^2(\boldsymbol{\theta}_{01})],$$

where $\mathbf{u}_1$ is a subvector of $\mathbf{u}$ with the corresponding non-zero coefficients.

By conditions (C4), (C5), (C7) and (C8), and the Taylor's expansion, we have

$$Q_n(\boldsymbol{\theta}_0 + \mathbf{u}/\sqrt{n}) - Q_n(\boldsymbol{\theta}_0) = n^{-1/2} Q_n'(\boldsymbol{\theta}_0)^T \mathbf{u} - \frac{1}{2}\Big(\mathbf{u}^T Q(\boldsymbol{\theta}_0)\mathbf{u}\Big)(1 + o_p(1)).$$

By condition (C1), the Taylor's expansion, triangular inequality, and Cauchy–Schwarz inequality, we have

$$\begin{aligned} & P_n^1(\boldsymbol{\theta}_{01} + \mathbf{u}_1/\sqrt{n}) - P_n^1(\boldsymbol{\theta}_{01}) \\ &= n \sum_{k=1}^{m_0} \left\{ \sum_{j=1}^{t_k} \Big[p_{\lambda_1}(|\boldsymbol{\beta}_{kj} + u_{kj}/\sqrt{n}|) - p_{\lambda_1}(|u_{kj}/\sqrt{n}|) \Big] \right\} \\ &= \sqrt{m_0 t}\, a_n \|\mathbf{u}\| + \frac{b_n}{2}\|\mathbf{u}\|^2 (1 + o(1)), \end{aligned}$$

where m_0 is the number of true non-zero mixing weights, and $t = \max_k \sqrt{t_k}$, and t_k is the number of true non-zero regression coefficients in the k-th component.

Since $\sqrt{n}\lambda_2 \to \infty$, and $\lambda_2 \to 0$, we have

$$|P_n^2(\boldsymbol{\theta}_{01} + \mathbf{u}_1/\sqrt{n}) - P_n^2(\boldsymbol{\theta}_{01})| = 0.$$

Regularity condition (C6) implies that $n^{-1/2} Q_n'(\boldsymbol{\theta}_0) = O_p(1)$. Since $\sqrt{n}\min\{\lambda_1, \lambda_2\} \to \infty$, and $\min\{\lambda_1, \lambda_2\} \to 0$, we have $a_n = 0$. By conditions (C2) and (C6), for any given $\epsilon > 0$, there exists a sufficiently large M_ϵ such that

$$\lim_{n\to\infty} P\left\{ \sup_{\|\mathbf{u}\| = M_\epsilon} \Gamma_n(\mathbf{u}) < 0 \right\} \geq 1 - \epsilon.$$

Therefore, with large probability, there is a local maximum in $\{\boldsymbol{\theta} + \mathbf{u}/\sqrt{n} : \|\mathbf{u}\| \leq M_\epsilon\}$. That is to say, this local maximizer $\hat{\boldsymbol{\theta}}_n$ satisfies $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| = O_p(1/\sqrt{n})$. This completes the proof of Theorem 1. □

Proof of Theorem 2. We first show that $\hat{\pi}_k = 0$ for $k = m_0 + 1, \cdots, m$. Since $\big|\big|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\big|\big| = O_p(n^{-1/2})$, we have $\hat{\pi}_k = O_p(1/\sqrt{n})$ for $k = m_0 + 1, \cdots, m$. To prove (a), it is sufficient to show with probability tending to 1 as $n \to \infty$ for any π_k satisfying $\hat{\pi}_k - \pi_k = O_p(1/\sqrt{n})$ and $k = m_0 + 1, \cdots, m$

$$\frac{\partial Q^*(\boldsymbol{\theta})}{\partial \hat{\pi}_k} < 0 \quad \text{for} \quad \hat{\pi}_k < C/\sqrt{n}, \tag{A1}$$

where C is a positive constant number,

$$Q^*(\boldsymbol{\theta}) = \tilde{Q}_n(\boldsymbol{\theta}) - \delta\left(\sum_{k=1}^{m} \pi_k - 1\right),$$

and δ is a Lagrange multiplier. Therefore, $\hat{\pi}_k, k = 1, \cdots, m$ should satisfy

$$\frac{\partial Q^*(\boldsymbol{\theta})}{\partial \hat{\pi}_k} = \sum_{i=1}^{n} \frac{f_{p_j}(Y_i - \mathbf{X}_i^T \boldsymbol{\beta}_j; 0, \sigma_j)}{\sum_{j=1}^{m} \hat{\pi}_j f_{p_j}(Y_i - \mathbf{X}_i^T \boldsymbol{\beta}_j; 0, \sigma_j)} - n\lambda_2 \frac{p'_{\lambda_2}(\hat{\pi}_k)}{C_0 + p_{\lambda_2}(\hat{\pi}_k)} - \delta = 0. \tag{A2}$$

We first consider $k \leq m_0$. By the law of large numbers, we have

$$\sum_{i=1}^{n} \frac{f_{p_j}(Y_i - \mathbf{X}_i^T \boldsymbol{\beta}_j; 0, \sigma_j)}{\sum_{j=1}^{m} \hat{\pi}_j f_{p_j}(Y_i - \mathbf{X}_i^T \boldsymbol{\beta}_j; 0, \sigma_j)} = O_p(n). \tag{A3}$$

For $k \leq m_0$, we have $\hat{\pi}_k = \pi_k^0 + O_p(1/\sqrt{n}) > \frac{1}{2}\min\{\pi_1^0, \cdots, \pi_{m_0}^0\}$. Since $n\lambda_2 = o_p(n)$, $p'_{\lambda_2}(\hat{\pi}_k) = o_p(1)$ and $p_{\lambda_2}(\hat{\pi}_k) = o_p(n)$, then we have

$$n\lambda_2 \frac{p'_{\lambda_2}(\hat{\pi}_k)}{C_0 + p_{\lambda_2}(\hat{\pi}_k)} = o_p(1). \tag{A4}$$

By (A2)–(A4), we have $\delta = O_p(n)$. For $k \geq m_0 + 1$ and $\hat{\pi}_k < C/\sqrt{n}$, we have $\hat{\pi}_k = O_p(1/\sqrt{n})$. By $\sqrt{n}\lambda_2 \to \infty$, C_0 is sufficient small and $p_\lambda(\cdot)$ is the SCAD penalty, we have

$$\left\{ n\lambda_2 \frac{p'_{\lambda_2}(\hat{\pi}_k)}{C_0 + p_{\lambda_2}(\hat{\pi}_k)} \right\} \Big/ n = \frac{\lambda_2^2}{C_0 + \lambda_2 \hat{\pi}_k} = O_p(\sqrt{n}\lambda_2) \to \infty.$$

Therefore, the first term and the third term in the Equation (A2) are dominated by the second term. Thus, we prove the Equation (A1). This completes the proof of (a).

To prove (b), for any $\boldsymbol{\theta}$ with m_0 components, we split $\boldsymbol{\theta}_{m_0} = (\boldsymbol{\theta}_{m_0}^1, \boldsymbol{\theta}_{m_0}^2)$ for any $\boldsymbol{\theta}_{m_0}$ in the neighborhood $||\boldsymbol{\theta}_{m_0} - \boldsymbol{\theta}_{m_0}^0|| = O_p(1/\sqrt{n})$ such that $\boldsymbol{\theta}_{m_0}^2$ contains all zero effects, e.g., $\boldsymbol{\beta}_{kj} = 0, k = 1, \cdots, m_0$ and $j = 1, \cdots, t_k$. By (2), we have

$$\begin{aligned}
&\tilde{Q}_n\{(\boldsymbol{\theta}_{m_0}^1, \boldsymbol{\theta}_{m_0}^2)\} - \tilde{Q}_n\{(\boldsymbol{\theta}_{m_0}^1, 0)\} \\
&= [Q_n\{(\boldsymbol{\theta}_{m_0}^1, \boldsymbol{\theta}_{m_0}^2)\} - Q_n\{(\boldsymbol{\theta}_{m_0}^1, 0)\}] - [P_n^1\{(\boldsymbol{\theta}_{m_0}^1, \boldsymbol{\theta}_{m_0}^2)\} - P_n^1\{(\boldsymbol{\theta}_{m_0}^1, 0)\}] \\
&= [Q_n\{(\boldsymbol{\theta}_{m_0}^1, \boldsymbol{\theta}_{m_0}^2)\} - Q_n\{(\boldsymbol{\theta}_{m_0}^1, 0)\}] - n \sum_{k=1}^{m_0} \sum_{j=t_k+1}^{p} p_{\lambda_1}(|\boldsymbol{\beta}_{kj}|).
\end{aligned}$$

According to the mean value theorem, we have

$$Q_n\{(\boldsymbol{\theta}_{m_0}^1, \boldsymbol{\theta}_{m_0}^2)\} - Q_n\{(\boldsymbol{\theta}_{m_0}^1, 0)\} = \left[\frac{\partial Q_n\{(\boldsymbol{\theta}_{m_0}^1, \boldsymbol{\gamma})\}}{\partial \boldsymbol{\theta}_{m_0}^2} \right]^T \boldsymbol{\theta}_{m_0}^2, \tag{A5}$$

where $||\boldsymbol{\gamma}|| \leq ||\boldsymbol{\theta}_{m_0}^2|| = O(n^{-1/2})$. Since

$$\begin{aligned}
&\left|\left| \frac{\partial Q_n\{(\boldsymbol{\theta}_{m_0}^1, \boldsymbol{\gamma})\}}{\partial \boldsymbol{\theta}_{m_0}^2} - \frac{\partial Q_n\{(\boldsymbol{\theta}_{m_0}^{01}, 0)\}}{\partial \boldsymbol{\theta}_{m_0}^2} \right|\right| \\
&\leq \left|\left| \frac{\partial Q_n\{(\boldsymbol{\theta}_{m_0}^1, \boldsymbol{\gamma})\}}{\partial \boldsymbol{\theta}_{m_0}^2} - \frac{\partial Q_n\{(\boldsymbol{\theta}_{m_0}^1, 0)\}}{\partial \boldsymbol{\theta}_{m_0}^2} \right|\right| + \left|\left| \frac{\partial Q_n\{(\boldsymbol{\theta}_{m_0}^1, 0)\}}{\partial \boldsymbol{\theta}_{m_0}^2} - \frac{\partial Q_n\{(\boldsymbol{\theta}_{m_0}^{01}, 0)\}}{\partial \boldsymbol{\theta}_{m_0}^2} \right|\right| \\
&\leq \left[\sum_{i=1}^{n} R_1(\mathbf{z}_i) \right] ||\gamma|| + \left[\sum_{i=1}^{n} R_1(\mathbf{z}_i) \right] ||\boldsymbol{\theta}_{m_0}^1 - \boldsymbol{\theta}_{m_0}^{01}|| \\
&= (||\gamma|| + ||\boldsymbol{\theta}_{m_0}^1 - \boldsymbol{\theta}_{m_0}^{01}||) O_p(n) = O_p(n^{1/2}),
\end{aligned}$$

and

$$\frac{\partial Q_n\{(\boldsymbol{\theta}_{m_0}^{01}, 0)\}}{\partial \boldsymbol{\theta}_{m_0}^2} = O_p(n^{1/2}),$$

we have

$$\frac{\partial Q_n\{(\boldsymbol{\theta}^1_{m_0}, \gamma)\}}{\partial \boldsymbol{\theta}^2_{m_0}} = O_p(n^{1/2}). \tag{A6}$$

By (A5) and (A6), we have

$$Q_n\{(\boldsymbol{\theta}^1_{m_0}, \boldsymbol{\theta}^2_{m_0})\} - Q_n\{(\boldsymbol{\theta}^1_{m_0}, 0)\} = O_p(n^{1/2}) \left[\sum_{k=1}^{m_0} \sum_{j=t_k+1}^{p} |\boldsymbol{\beta}_{kj}| \right].$$

Thus, we have

$$\tilde{Q}_n\{(\boldsymbol{\theta}^1_{m_0}, \boldsymbol{\theta}^2_{m_0})\} - \tilde{Q}_n\{(\boldsymbol{\theta}^1_{m_0}, 0)\} = \sum_{k=1}^{m_0} \sum_{j=t_k+1}^{p} \Big[O_p(n^{1/2}) |\boldsymbol{\beta}_{kj}| - n p_{\lambda_1}(|\boldsymbol{\beta}_{kj}|) \Big].$$

By condition (C3), for $|t| \leq n^{-1/2} \log n$, we have $O_p(n^{1/2})|t| < n p_{\lambda_1}(|t|)$. Therefore, we can obtain

$$\tilde{Q}_n\{(\boldsymbol{\theta}^1_{m_0}, \boldsymbol{\theta}^2_{m_0})\} - \tilde{Q}_n\{(\boldsymbol{\theta}^1_{m_0}, 0)\} < 0. \tag{A7}$$

By (A7), with probability tending to 1 as $n \to \infty$, we have

$$\begin{aligned} &\tilde{Q}_n\{(\boldsymbol{\theta}^1_{m_0}, \boldsymbol{\theta}^2_{m_0})\} - \tilde{Q}_n\{(\hat{\boldsymbol{\theta}}^1_{m_0}, 0)\} \\ &= [\tilde{Q}_n\{(\boldsymbol{\theta}^1_{m_0}, \boldsymbol{\theta}^2_{m_0})\} - \tilde{Q}_n\{(\boldsymbol{\theta}^1_{m_0}, 0)\}] + [\tilde{Q}_n\{(\boldsymbol{\theta}^1_{m_0}, 0)\} - \tilde{Q}_n\{(\hat{\boldsymbol{\theta}}^1_{m_0}, 0)\}] < 0. \end{aligned}$$

Thus, this completes the proof of part (b).

Using the result of Theorem 1, there exists a $\sqrt{n}$-consistent local maximizer $\hat{\boldsymbol{\theta}}_{n1}$ of $\tilde{Q}_n\{(\boldsymbol{\theta}_1, 0)\}$ such that $\hat{\boldsymbol{\theta}}_n = (\hat{\boldsymbol{\theta}}_{n1}, 0)$ satisfies

$$\frac{\partial \tilde{Q}_n(\hat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}_1} = \left\{ \frac{Q_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1} - \frac{\partial P^1_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1} - \frac{\partial P^2_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1} \right\}_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n} = 0. \tag{A8}$$

By the Taylor's expansion, we have

$$\left\{ \frac{\partial Q_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1} \right\}_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n} = \frac{\partial Q_n(\boldsymbol{\theta}_{01})}{\partial \boldsymbol{\theta}_1} + \left\{ \frac{\partial^2 Q_n(\boldsymbol{\theta}_{01})}{\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\theta}_1^T} + o_p(n) \right\} (\hat{\boldsymbol{\theta}}_{n1} - \boldsymbol{\theta}_{01}), \tag{A9}$$

$$\left\{ \frac{\partial P^1_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1} \right\}_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n} = P^{1'}_n(\boldsymbol{\theta}_{01}) + \left\{ P^{1''}_n(\boldsymbol{\theta}_{01}) + o_p(n) \right\} (\hat{\boldsymbol{\theta}}_{n1} - \boldsymbol{\theta}_{01}), \tag{A10}$$

$$\left\{ \frac{\partial P^2_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1} \right\}_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n} = P^{2'}_n(\boldsymbol{\theta}_{01}) + \left\{ P^{2''}_n(\boldsymbol{\theta}_{01}) + o_p(n) \right\} (\hat{\boldsymbol{\theta}}_{n1} - \boldsymbol{\theta}_{01}). \tag{A11}$$

By substituting Equations (A9)–(A11) into (A8), we have

$$\begin{aligned} &\left\{ \frac{\partial^2 Q_n(\boldsymbol{\theta}_{01})}{\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\theta}_1^T} - P^{1''}_n(\boldsymbol{\theta}_{01}) - P^{2''}_n(\boldsymbol{\theta}_{01}) + o_p(n) \right\} (\hat{\boldsymbol{\theta}}_{n1} - \boldsymbol{\theta}_{01}) \\ &= \frac{\partial Q_n(\boldsymbol{\theta}_{01})}{\partial \boldsymbol{\theta}_1} - P^{1'}_n(\boldsymbol{\theta}_{01}) - P^{2'}_n(\boldsymbol{\theta}_{01}). \end{aligned}$$

By the conditions (C4), (C5), and (C6), we have

$$\frac{1}{n} \frac{\partial^2 Q_n(\boldsymbol{\theta}_{01})}{\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\theta}_1^T} = Q_1(\boldsymbol{\theta}_{01}) + o_p(1),$$

$$\frac{1}{\sqrt{n}}\frac{\partial Q_n(\boldsymbol{\theta}_{01})}{\partial \boldsymbol{\theta}_1} \xrightarrow{D} N(0, Q_1(\boldsymbol{\theta}_{01})).$$

By Slutsky's theorem, we have

$$\sqrt{n}\left\{\left[Q_1(\boldsymbol{\theta}_{01}) - \frac{P_n^{1''}(\boldsymbol{\theta}_{01})}{n} - \frac{P_n^{2''}(\boldsymbol{\theta}_{01})}{n}\right](\hat{\boldsymbol{\theta}}_{n1} - \boldsymbol{\theta}_{01}) + \frac{P_n^{1'}(\boldsymbol{\theta}_{01})}{n} + \frac{P_n^{2'}(\boldsymbol{\theta}_{01})}{n}\right\} \xrightarrow{D} N(0, Q_1(\boldsymbol{\theta}_{01})).$$

This completes the proof of part (c). □

References

1. Quandt, R.E. A new approach to estimating switching regressions. *J. Am. Stat. Assoc.* **1972**, *67*, 306–310. [CrossRef]
2. Goldfeld, S.M.; Quandt, R.E. A Markov model for switching regressions. *J. Econom.* **1973**, *1*, 3–15. [CrossRef]
3. Jacobs, R.A.; Jordan, M.I.; Nowlan, S.J.; Hinton, G.E. Adaptive mixtures of local experts. *Neural Comput.* **1991**, *3*, 79–87. [CrossRef] [PubMed]
4. Wedel, M.; Kamakura, W.A. *Market Segmentation: Conceptual and Methodological Foundations*; Springer Science & Business Media: Berlin, Germany, 2000.
5. Skrondal, A.; Rabe-Hesketh, S. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2004.
6. Peel, D.; MacLahlan, G. *Finite Mixture Models*; John & Sons: Toronto, ON, Canada, 2000.
7. McLachlan, G.J.; Lee, S.X.; Rathnayake, S.I. Finite mixture models. *Annu. Rev. Stat. Its Appl.* **2019**, *6*, 355–378. [CrossRef]
8. Yu, C.; Yao, W.; Yang, G. A selective overview and comparison of robust mixture regression estimators. *Int. Stat. Rev.* **2020**, *88*, 176–202. [CrossRef]
9. Chen, J.; Khalili, A. Order selection in finite mixture models with a nonsmooth penalty. *J. Am. Stat. Assoc.* **2009**, *104*, 187–196. [CrossRef]
10. Peng, H.; Huang, T.; Zhang, K. Model Selection for Gaussian Mixture Models. *Stat. Sin.* **2017**, *27*, 147–169.
11. Wang, P.; Puterman, M.L.; Cockburn, I.; Le, N. Mixed Poisson regression models with covariate dependent rates. *Biometrics* **1996**, *52*, 381–400. [CrossRef]
12. Khalili, A.; Chen, J. Variable selection in finite mixture of regression models. *J. Am. Stat. Assoc.* **2007**, *102*, 1025–1038. [CrossRef]
13. Jiang, Y. Robust variable selection for mixture linear regression models. *Hacet. J. Math. Stat.* **2016**, *45*, 549–559. [CrossRef]
14. Luo, R.; Wang, H.; Tsai, C.L. On mixture regression shrinkage and selection via the MR-Lasso. *Int. J. Pure Appl. Math.* **2008**, *46*, 403–414.
15. Jiang, Y.; Huang, M.; Wei, X.; Tonghua, H.; Hang, Z. Robust mixture regression via an asymmetric exponential power distribution. *Commun. Stat.-Simul. Comput.* **2022**, 1–12. [CrossRef]
16. Wang, X.; Feng, Z. Component selection for exponential power mixture models. *J. Appl. Stat.* **2023**, *50*, 291–314. [CrossRef] [PubMed]
17. Yu, C.; Wang, X. A new model selection procedure for finite mixture regression models. *Commun. Stat.-Theory Methods* **2020**, *49*, 4347–4366. [CrossRef]
18. Chen, X. Robust mixture regression with Exponential Power distribution. *arXiv* **2020**, arXiv:2012.10637.
19. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B (Methodol.)* **1977**, *39*, 1–22. [CrossRef]
20. Hunter, D.R.; Lange, K. A tutorial on MM algorithms. *Am. Stat.* **2004**, *58*, 30–37. [CrossRef]
21. Kobayashi, G. Skew exponential power stochastic volatility model for analysis of skewness, non-normal tails, quantiles and expectiles. *Comput. Stat.* **2016**, *31*, 49–88. [CrossRef]
22. Fan, J.; Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360. [CrossRef]
23. Zou, H.; Li, R. One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Stat.* **2008**, *36*, 1509–1533.
24. Wang, H.; Li, R.; Tsai, C.L. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **2007**, *94*, 553–568. [CrossRef]
25. Watnik, M.R. Pay for play: Are baseball salaries based on performance? *J. Stat. Educ.* **1998**, *6*, 1–5. [CrossRef]