Appl. Math. J. Chinese Univ.
2021, 36(4): 475-491

Variable selection for skew-normal mixture of joint location and scale models

WU Liu-cang[1] YANG Song-qin[1] TAO Ye[2]

Abstract. Although there are many papers on variable selection methods based on mean model in the finite mixture of regression models, little work has been done on how to select significant explanatory variables in the modeling of the variance parameter. In this paper, we propose and study a novel class of models: a skew-normal mixture of joint location and scale models to analyze the heteroscedastic skew-normal data coming from a heterogeneous population. The problem of variable selection for the proposed models is considered. In particular, a modified Expectation-Maximization(EM) algorithm for estimating the model parameters is developed. The consistency and the oracle property of the penalized estimators is established. Simulation studies are conducted to investigate the finite sample performance of the proposed methodologies. An example is illustrated by the proposed methodologies.

§1 Introduction

Homoscedasticity of the scale parameter is a common assumption in many regression models. However, this may not be appropriate in some situations, such as heteroscedasticity in data. One suitable approach is to model the variance parameter. Particularly in the econometric area and industrial quality improvement experiments, model the variance to identify the source of variability in the observations will be of direct interest in its own right, such as Taguchitype experiments for robust design[1]. In addition, the efficient estimation of mean parameters in regression depends on correct modelling of the variance. The loss of efficiency in using constant variance models when the variance is varying may be substantial. Thus, modelling of the variance can be as important as that of the mean. Joint mean and variance models have received a lot of attention in recent years. Park [2] proposed a log-linear model for the variance parameter and described the Gaussian model using a two-stage process to estimate the parameters. Harvey [3] discussed the maximum-likelihood (ML) estimation of the mean and variance effects and the subsequent likelihood ratio test under general conditions. Aitkin [4] provided ML estimation for a joint mean and variance model and applied it to the commonly

Received: 2019-03-11. Revised: 2020-02-24.
MR Subject Classification: 62F10, 62H12.
Keywords: heterogeneous population, skew-normal(SN) distribution, mixture of joint location and scale models, variable selection, EM algorithm.
Digital Object Identifier(DOI): https://doi.org/10.1007/s11766-021-3774-x.
Supported by the National Natural Science Foundation of China(11861041).

cited Minitab tree data. Wu *et al.* [5] proposed a hybrid strategy, in which variable selection is employed to reduce the dimension of the explanatory variables in joint mean and variance models, and Box-Cox transformation is made to remedy the distribution of the response. Wu *et al.* [6] investigated the simultaneously variable selection in joint location and scale models of the skew-normal distribution. Wu *et al.* [7] propose a unified penalized likelihood method to simultaneously select significant variables and estimate unknown parameters in a joint location, scale and skewness model with a skew-t-normal (StN) distribution when outliers and asymmetrical outcomes are present.

Mixture of regression models are well known as switching regression models in econometrics literature, which were the first introduced by Goldfeld and Quandt [8]. Mixture of regression models have been applied in various fields including biology, medicine, economics, agriculture, animal sciences and so on, see [9–12]. Mixture of regression models can be easily applied to analyze data sets in which two or more subpopulations are mixed together. Due to its flexibility in modeling, mixture of regression models have enjoyed intensive attentions over the past years, from both practical and theoretical perspectives. In particular, on the variable selection, Khalili and Chen [13] proposed a new regularization method for variable selection in finite mixture of regression models. Khalili [14] investigated new estimation and feature selection methods in mixture-of-experts models. Khalili and Lin [15] developed a new regularization in finite mixture of regression models with diverging number of parameters. Khalili [16] gave an overview of the new feature selection methods in finite mixture of regression models. For other methods of variable selection for the mixture of regression models, we can refer to Du *et al.*[17], Ormoz and Eskandari[18], Lee *et al.*[19], Khalili *et al.*[20], Tang and Karunamun[21].

The existing studies on the mixture of regression models mainly focus on the normality assumption of response variables. This assumption may be inappropriate in some applications. For a set of data containing a group or groups of observations with asymmetric behavior, the use of normal component may be unsuitable and inferences can be misleading. In particular, the normal mixture model tends to overfit when additional components are included to capture the skewness. So we introduce the skew-normal distribution to overcome the potential weakness of normal mixtures. Liu and Lin [22] first developed a skew-normal mixture of regression model, but they only considered the mixture of location regression models using the univariate skew-normal distribution.

In the standard formulation of all the above mentioned work within the framework of the mixture of regression models, it is assumed that the equal variance for each component is constant across observations. However, in many practical situations, this assumption may be not hold. A more general model formulation allows for the variance to vary across observations, which has received little attention in the literature. When the variance is really varying, inference carried out under the assumption of fixed constant variance may be highly inaccurate. The common strategy in this situation is to augment the model by defining another regression structure for the variance parameter in a homogeneous population(see Aitkin [4]; Cepeda and Gamerman [23]; Taylor and Verbyla [24] ; Zhang and Wang [25] ; Wu and Li [26] ; Zhao and Zhang [27] ; Wu *et al.* [7]).

Similar to modelling of the variance parameter in a homogeneous population, we apply the idea of joint mean and variance models to the mixture of regression models and propose and study a novel class of models: a skew-normal mixture of joint location and scale mod-

el(SNMJLSM) to analyze the heteroscedastic skew-normal data coming from a heterogeneous population. The problem of variable selection for the proposed model is considered. In particular, a modified Expectation-Maximization(EM) algorithm for estimating the model parameters is developed. The consistency and the oracle property of the penalized estimators are established. Simulation studies are conducted to investigate the finite sample performance of the proposed methodologies. An example is illustrated by the proposed methodologies.

The outline of the article is as follows. In Section 2, we propose a skew-normal mixture of joint location and scale model(SNMJLSM). Then, in Section 3, we present variable selection for our proposed models via the penalized likelihood-based method. To do this in Section 3.1, penalized likelihood-based method is considered, and in Section 3.2 asymptotic properties of the proposed variable selection procedure are studied. Section 3.3 will show the EM algorithm and numerical solution for estimators. Choosing the tuning parameters will be considered in Section 3.4. In Section 4, we carry out simulation studies to assess the finite sample performance of the method. In section 5, a real data set on the *air quality index* (AQI) date is analyzed to demonstrate the proposed methods. Some concluding remarks are given in Section 6. Some technical proofs are put in the appendix.

§2 SN mixture of joint location and scale models

2.1 Skew-normal distribution

In this section, we introduce the skew-normal distribution, as developed by Azzalini[28]. The random variable Y is said to have a skew-normal distribution with location parameter μ, scale parameter σ and skewness parameter λ, denoted by $Y \sim SN(\mu, \sigma^2, \lambda)$, if it has the probability density function

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

where $\phi(.)$, $\Phi(.)$ are the density function and the cumulative distribution function of the standard normal distribution, respectively. If $\lambda = 0$, then the density of Y in (1) reduces to $N(\mu, \sigma^2)$.

2.2 SN mixture of joint location and scale models

Let $y_1, y_2, \cdots, y_n$ be a random sample of size n from a m-component skew-normal mixture model with unknown mixing probabilities $\pi_1, \pi_2, \cdots, \pi_m$, the probability density function of random variable Y has a m-component mixture form:

$$f(y) = \sum_{j=1}^{m} \pi_j SN(\mu_j, \sigma_j^2, \lambda_j), \tag{2}$$

where π_j represents the mixing probabilities and $\sum_{j=1}^{m} \pi_j = 1$. Where $\pi_j, \mu_j, \sigma_j^2, \lambda_j$ are unknown, and the parameter vector of the model is

$$\mathbf{\Psi} = (\pi_1, \cdots, \pi_m, \mu_1, \cdots, \mu_m, \sigma_1^2, \cdots, \sigma_m^2, \lambda_1, \cdots, \lambda_m)^T.$$

In this paper, we assume that the number of components m is fixed and known. Of course, in some practical applications, and m may be unknown and needs to be estimated along with

mixing probabilities and other parameters, but in this paper, to simplify, we only consider the case where m is known, and only estimate the unknown vector of parameters $\boldsymbol{\Psi}$.

Consider the following skew-normal mixture of joint location and scale model(SNMJLSM):

$$\begin{cases} y_i \sim \sum_{j=1}^{m} \pi_j SN(\mu_{ij}, \sigma_{ij}^2, \lambda_j), \\ \mu_{ij} = \boldsymbol{x}_i^T \boldsymbol{\beta}_j, \\ \log \sigma_{ij}^2 = \boldsymbol{h}_i^T \boldsymbol{\gamma}_j, \\ i = 1, 2, \cdots, n; j = 1, 2, \cdots, m. \end{cases} \tag{3}$$

In the model (3), y_i is an independent response variable, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T$ and $\boldsymbol{h}_i = (h_{i1}, h_{i2}, \cdots, h_{iq})^T$ are explanatory variables. In the jth subpopulation, $\boldsymbol{\beta}_j = (\beta_{j1}, \beta_{j2}, \cdots, \beta_{jp})^T$ is an unknown parameter of the location model , $\boldsymbol{\gamma}_j = (\gamma_{j1}, \gamma_{j2}, \cdots, \gamma_{jq})^T$ is an unknown parameter of the scale model and λ_j is skewness parameter. $\boldsymbol{x}_i, \boldsymbol{h}_i$ may be identical or completely different or part of the same, that is, the location model and the scale model may incorporate different covariates, or some of the same covariates, and may depend on common covariates in different ways. In this paper, we aim to remove the unnecessary explanatory variables from the model (3).

2.3 Identifiability

In models of finite mixtures of any class of distributions, it is important to consider the property of the identifiability, because procedures for estimation of parameters can be ill defined without such property. Otiniano et al [29]presents the proof of the identifiability of the classes of all finite mixtures of Skew-Normal and Skew-t distributions. The identifiability of some mixture models has been investigated by several authors, Teicher [30], Atienza et al [31], among others.

In this paper, we assume that the SNMJLSM under study are identifiable. Consider a SNMJLSM as (3). For a given design matrix, the SNMJLSM are said to be identifiable if for any two parameters

$$\sum_{j=1}^{m} \pi_j SN(\mu_j, \sigma_j^2, \lambda_j) = \sum_{j=1}^{m^*} \pi_j^* SN(\mu_j^*, \sigma_j^{*2}, \lambda_j^*),$$

for each $i = 1, 2, \cdots, n$ and all possible values of y, implies $m = m^*$ and $\boldsymbol{\Psi} = \boldsymbol{\Psi}^*$.

The identifiability implies that no two sets of different parameter values have the same density functions. Following Hennig [32] and Wang *et al.*[33] , we interpret $\boldsymbol{\Psi} = \boldsymbol{\Psi}^*$, in the above definition, up to a permutation .

§3 Estimation and variable selection method

3.1 Penalized likelihood

Many traditional variable selection criterias can be considered as a penalized likelihood which balances modeling biases and estimation variances [34]. They can enhance model interpretability with parsimonious representation and also improve the accuracy of estimation by efficiently identifying the significant variables.

Suppose that we have a random sample $y_i, \boldsymbol{x}_i, \boldsymbol{h}_i, i = 1, 2, \cdots, n$, from the model (3). Let $l(\boldsymbol{\Psi})$ denote the log-likelihood function conditional on $y_i, \boldsymbol{x}_i, \boldsymbol{h}_i$. Then the log-likelihood function

of $\mathbf{\Psi}$ is given by:

$$l_n(\mathbf{\Psi}) = \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{m} \pi_j SN(\mu_{ij}, \sigma_{ij}^2, \lambda_j) \right\}. \tag{4}$$

Similar to Khalili and Chen [13] , the penalty log-likelihood function is defined as

$$L_n(\mathbf{\Psi}) = l_n(\mathbf{\Psi}) - p_n(\mathbf{\Psi}), \tag{5}$$

where $l_n(\mathbf{\Psi})$ is a log-likelihood functions in (4), and the penality function $p_n(\mathbf{\Psi})$ is as follows:

$$p_n(\mathbf{\Psi}) = \sum_{j=1}^{m} \pi_j \sum_{t=1}^{p} p_n(\beta_{jt}; \tau_{1j}) + \sum_{j=1}^{m} \pi_j \sum_{t=1}^{q} p_n(\gamma_{jt}; \tau_{2j}). \tag{6}$$

Here $p_n(\mathbf{\Psi}; \tau_j)$ is a nonnegative function indexed by certain tuning parameters $\tau_j \geq 0$. The first part of $p_n(\mathbf{\Psi})$ penalizes the regression coefficients β_{jt} of the location model, and the second part is the penalty on the coefficients γ_{jt} of the scale model. General conditions on the proper choice of the penalty functions in (6) are given in the next subsection.

If some of the coefficients β_{jt} or γ_{jt} (or both) are small in the models(3), then in fitting the model to the data through maximization of the function $L_n(\mathbf{\Psi})$ the hope is that the penalty function $p_n(\mathbf{\Psi})$ will force the estimated values of those coefficients to zero. The method automatically performs variable selection, which makes it computationally much more efficient than all-subset selection methods.

In general, we should choose appropriate penalty functions to suit the need of the application, under the guidance of statistical theory. However, the following three penalty functions have been investigated in the literature in a number of contexts and will be used here to illustrate the theory:

LASSO:

$$p_{\tau_j}(|\mathbf{\Psi}_{jt}|) = \tau_j |\mathbf{\Psi}_{jt}|,$$

HARD:

$$p_{\tau_j}(|\mathbf{\Psi}_{jt}|) = {\tau_j}^2 - (|\mathbf{\Psi}_j| - \tau_j)^2 I(|\mathbf{\Psi}_{jt}| < \tau_j),$$

SCAD:

$$p^{'}_{\tau_j}(|\mathbf{\Psi}_{jt}|) = \tau_j \{ I(|\mathbf{\Psi}_{jt}| \leq \tau_j) + \frac{(a\tau_j - |\mathbf{\Psi}_{jt}|)+}{(a-1)\tau_j} I(|\mathbf{\Psi}_{jt}| > \tau_j) \}.$$

Following the convention in Fan and Li [34], we set a = 3.7 in our work. The penalty function of LASSO([35]) has a good property enables easy numerical computation. The SCAD penalty function gives good performance in selecting important variables without creating excessive biases. HARD([36]) should work more like SCAD, although less smoothly([13]).

3.2 Asymptotic properties

Without loss of generality, the coefficient vectors $\boldsymbol{\beta}_j$ and $\boldsymbol{\gamma}_j$ are decomposed into $\boldsymbol{\beta}_j^T = (\boldsymbol{\beta}_{1j}^T, \boldsymbol{\beta}_{2j}^T)$ and $\boldsymbol{\gamma}_j^T = (\boldsymbol{\gamma}_{1j}^T, \boldsymbol{\gamma}_{2j}^T)$ such that $\boldsymbol{\beta}_{2j}$ and $\boldsymbol{\gamma}_{_{2j}}$ contain the 0 effects. The parameter vector $\mathbf{\Psi}$ is also decomposed into $\mathbf{\Psi}^T = (\mathbf{\Psi}_1^T, \mathbf{\Psi}_2^T)$ such that $\mathbf{\Psi}_2^T$ contains all the parameters corresponding to zero effects, that is $\boldsymbol{\beta}_{2j}$ and $\boldsymbol{\gamma}_{2j}$ in the true model. The vector of parameters in the true model is $\mathbf{\Psi}_0$, and its components are denoted with a superscript, such as $\beta_{jt}^0 (1 \leq j \leq m, 1 \leq t \leq p)$ and $\gamma_{jt}^0 (1 \leq j \leq m, 1 \leq t \leq q)$. Denote

$$q_{1n} = \max_{j,t} \{ |p_n^{'}(\beta_{jt}^0; \tau_{1j})| / \sqrt{n} : \beta_{jt}^0 \neq 0 \}; q_{1n}^* = \max_{j,t} \{ |p_n^{'}(\gamma_{jt}^0; \tau_{2j})| / \sqrt{n} : \gamma_{jt}^0 \neq 0 \}$$

$$q_{2n} = \max_{j,t}\{|p_n^{''}(\beta_{jt}^0;\tau_{1j})|/\sqrt{n} : \beta_{jt}^0 \neq 0\}; q_{2n}^* = \max_{j,t}\{|p_n^{''}(\gamma_{jt}^0;\tau_{2j})|/\sqrt{n} : \gamma_{jt}^0 \neq 0\}$$

The $p_n^{'}(\mathbf{\Psi};\tau_j)$ and $p_n^{''}(\mathbf{\Psi};\tau_j)$ are the first and second derivatives of the function $p_n(\mathbf{\Psi};\tau_j)$ with respect to $\mathbf{\Psi}$, for the tuning parameters τ_j and γ_j that depend on the simple size n. Using these quantities, the function $p_n(\mathbf{\Psi};\tau_j)$ is required to satisfy the following conditions:

C_0: For all n and τ_j, $p_n(0;\tau_j) = 0$, and $p_n(\mathbf{\Psi};\tau_j)$ is symmetric and nonnegative. In addition, it is nondecreasing and twice differentiable for all $\mathbf{\Psi}$ in $(0,\infty)$ with at most a few exceptions.

C_1: As $n \to \infty$, $q_{2n} = o(1)$, $q_{2n}^* = o(1)$.

C_2: For $T_n = \{\mathbf{\Psi}; 0 < \mathbf{\Psi} \leq n^{-1/2}\log n\}$, $\lim\limits_{n\to\infty}\inf\limits_{\mathbf{\Psi}\in T_n}(p_n^{'}(\mathbf{\Psi};\tau_j))/\sqrt{n} = \infty$.

Conditions $C_0 - C_2$ guarantee $\sqrt{n}$-consistency of the estimators of the true nonzero regression coefficients, and also consistency of the method in variable selection.

Theorem 1 Let $V_i = (x_i, Y_i)$, $i = 1, 2, \cdots, n$, be a random sample from a density function $f(v, \mathbf{\Psi})$ that satisfies the regularity conditions $R_1 - R_5$ in the Appendix. Assume that the function $p_n(\mathbf{\Psi};\tau_j)$ satisfies the regularity conditions C_0 and C_1, and also suppose $\tau_j/\sqrt{n} \to 0$ as $n \to \infty$. Then there exists a local minimizer $\hat{\mathbf{\Psi}}_n$ of the regularized log-likelihood function $L_n(\mathbf{\Psi})$ for which

$$\|\hat{\mathbf{\Psi}}_n - \hat{\mathbf{\Psi}}_0\| = O_p\{n^{-1/2}(1 + q_{1n}^* + q_{1n})\}.$$

The estimator $\hat{\mathbf{\Psi}}_n$ is $\sqrt{n}$ consistent if $q_{1n} = o(1)$ and $q_{1n}^* = o(1)$. The rate is achievable by proper choice of the tuning parameters τ_j.

Consistency of $\hat{\mathbf{\Psi}}_n$ in variable selection, that is: $\hat{\boldsymbol{\beta}}_{2j} = 0, j = 1, 2, \cdots, m$, and $\hat{\boldsymbol{\gamma}}_{2j} = 0, j = 1, 2, \cdots, m$, with probability tending to 1, is shown in Theorem 2.

Theorem 2 Assume that the conditions in Theorem 1 are fulfilled, and the function $p_n(\mathbf{\Psi};\tau_j)$ satisfies conditions $C_0 - C_2$, and also suppose $\tau_j/\sqrt{n} \to 0$ as $n \to \infty$. Then for any $\sqrt{n}$-consistent maximum regularized likelihood estimator $\hat{\mathbf{\Psi}}_n$ of $\mathbf{\Psi}$, as $n \to \infty$. We have

(a) Consistency in the variable selection: $P(\hat{\boldsymbol{\beta}}_{2j} = 0, \hat{\boldsymbol{\gamma}}_{2j} = 0) \to 1, j = 1, 2, \cdots, m$.

(b) Asymptotic normality:

$$\sqrt{n}\{[I_1(\mathbf{\Psi}_{01}) + \frac{p_n^{''}(\mathbf{\Psi}_{01})}{n}](\hat{\mathbf{\Psi}}_1 - \mathbf{\Psi}_{01}) + \frac{p_n^{''}(\mathbf{\Psi}_{01})}{n}\} \xrightarrow{d} N(0, I_1(\mathbf{\Psi}_{01})),$$

where $I_1(\mathbf{\Psi}_{01})$ is the Fisher information matrix under the true model with all zero effects removed.

Brief proofs of the theorems are in the Appendix.

3.3 The EM algorithm

In the context of finite mixture models, because of the existence of latent variable z_{ij}, the classical MLE is not applicable. The expectation-maximization (EM) algorithm of Dempster *et al.* [37]provides a convenient approach to the estimation of parameters. However, due to condition C_0, the function $p_n(\mathbf{\Psi};\tau_j)$ are not differentiable at $\mathbf{\Psi} = 0$. Then, the Newton-Raphson algorithm can not be used directly. We follow the suggestion of Fan and Li [34], and approximate $p_n(\mathbf{\Psi};\tau_j)$ in a neighbourhood of $\mathbf{\Psi}_0$ by the local quadratic function

$$p_n(\mathbf{\Psi};\tau_j) \cong p_n(\mathbf{\Psi}_0;\tau_j) + \frac{p_n^{'}(\mathbf{\Psi}_0;\tau_j)}{2\mathbf{\Psi}_0}(\mathbf{\Psi}^2 - \mathbf{\Psi}_0^2). \tag{7}$$

This function increase to infinity when $n \to \infty$, which is more suitable for our application than the simple Taylors expansion. Let $\mathbf{\Psi}^{(k)}$ be the parameter value after the kth iteration.

We replace $p_n(\boldsymbol{\Psi})$ in the penalized log-likelihood function in (5) by the following function:

$$\tilde{p}_n(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)}) = \sum_{j=1}^{m}\pi_j\sum_{t=1}^{p}\{p_{nj}(\beta_{jt}^{(k)}) + \frac{p_n^{'}(\beta_{jt}^{(k)})}{2\beta_{jt}^{(k)}}(\beta_{jt}^2 - (\beta_{jt}^{(k)})^2)\}$$
$$+\sum_{j=1}^{m}\pi_j\sum_{t=1}^{q}\{p_{nj}(\gamma_{jt}^{(k)}) + \frac{p_n^{'}(\gamma_{jt}^{(k)})}{2\gamma_{jt}^{(k)}}(\gamma_{jt}^2 - (\gamma_{jt}^{(k)})^2)\}.$$

The revised EM algorithm is as follows. Let the complete log-likelihood function be

$$l_n^c(\boldsymbol{\Psi}) = \sum_{i=1}^{n}\sum_{j=1}^{m} z_{ij}\{\log SN(y_i;\boldsymbol{x}_i^T\boldsymbol{\beta}_j, \exp(\boldsymbol{h}_i^T\boldsymbol{\gamma}_j), \lambda_j)\},$$

where z_{ij} is indicator variable showing the component membership of the ith observation in the SNMJLSM and is an unobserved imaginary variable. The penalized complete log-likelihood function is, then, given by $L_n^c(\boldsymbol{\Psi}) = l_n^c(\boldsymbol{\Psi}) - p_n(\boldsymbol{\Psi})$. After kth iteration, the model parameters are updated as follows:

E-Step: Let $\boldsymbol{\Psi}^{(k)}$ be the estimate of the parameters after kth iteration. The E-step computes the conditional expectation of the $L_n^c(\boldsymbol{\Psi})$ with the respect to z_{ij}, given the data $(\boldsymbol{x}_i, \boldsymbol{h}_i, y_i)$, and assuming that the values of the current estimate $\boldsymbol{\Psi}^{(k)}$ are the true parameters of the model. The conditional expectation is

$$Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)}) = \sum_{i=1}^{n}\sum_{j=1}^{m}\omega_{ij}^{(k)}\{\log[\pi_j SN(y_i;\boldsymbol{x}_i^T\boldsymbol{\beta}_j, \exp(\boldsymbol{h}_i^T\boldsymbol{\gamma}_j), \lambda_j)]\} - p_n(\boldsymbol{\Psi}),$$

where the conditional expectation of the missing labels z_{ij} is:

$$\omega_{ij}^{(k)} = \frac{\pi_j^{(k)} SN(y_i;\boldsymbol{x}_i^T\boldsymbol{\beta}_j^{(k)}, \exp(\boldsymbol{h}_i^T\boldsymbol{\gamma}_j^{(k)}), \lambda_j^{(k)})}{\sum_{j=1}^{m}\pi_j^{(k)} SN(y_i;\boldsymbol{x}_i^T\boldsymbol{\beta}_j^{(k)}, \exp(\boldsymbol{h}_i^T\boldsymbol{\gamma}_j^{(k)}), \lambda_j^{(k)})}. \tag{8}$$

M-Step: The $M-Step$ on the $(k+1)th$ iteration maximizes the function $Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)})$ with respect to $\boldsymbol{\Psi}$. In the usual EM algorithm, the mixing probabilities are updated by

$$\pi_j^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n}\omega_{ij}^{(k)}, j = 1, 2, \cdots, m. \tag{9}$$

which maximizes leading term of $Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)})$. Maximizing $Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{r})$ itself with respect to π_j will be more complex. For simplicity, we use updating scheme (8) nevertheless; this performed well in our simulations.

We now consider the π_j, as constants in $Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)})$, and maximize $Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)})$ with respect to the other parameters in $\boldsymbol{\Psi}$. By replacing $p_n(\boldsymbol{\Psi})$ by $\tilde{p}_n(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)})$ in $Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)})$, the regression coefficients are updated by solving

$$\sum_{i=1}^{n}\omega_{ij}^{(k)}\frac{\partial}{\partial\beta_{jt}}\{\log SN(y_i;\boldsymbol{x}_i^T\boldsymbol{\beta}_j, \exp(\boldsymbol{h}_i^T\boldsymbol{\gamma}_j), \lambda_j)\} - \pi_j\{\frac{\partial}{\partial\beta_{jt}}\tilde{p}_{nt}(\beta_{jt})\} = 0,$$
$$\sum_{i=1}^{n}\omega_{ij}^{(k)}\frac{\partial}{\partial\gamma_{jt}}\{\log SN(y_i;\boldsymbol{x}_i^T\boldsymbol{\beta}_j, \exp(\boldsymbol{h}_i^T\boldsymbol{\gamma}_j), \lambda_j)\} - \pi_j\{\frac{\partial}{\partial\gamma_{jt}}\tilde{p}_{nt}(\gamma_{jt})\} = 0,$$
$$\sum_{i=1}^{n}\omega_{ij}^{(k)}\frac{\partial}{\partial\lambda_j}\{\log SN(y_i;\boldsymbol{x}_i^T\boldsymbol{\beta}_j, \exp(\boldsymbol{h}_i^T\boldsymbol{\gamma}_j), \lambda_j)\} = 0,$$

where $\tilde{p}_{nt}(\beta_{jt})$ and $\tilde{p}_{nt}(\gamma_{jt})$ are the corresponding term in $\tilde{p}_n(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)})$, for $j = 1, 2, \cdots, m; t = 1, 2, \cdots, p(q)$.

Starting from an initial value $\mathbf{\Psi}^{(0)}$, we iterate between the E and M steps until the Euclidean norm $\|\mathbf{\Psi}^{(k+1)} - \mathbf{\Psi}^{(k)}\|$ is smaller than some threshold value.

3.4 Choosing the tuning parameters

When using the methods proposed in this paper, one needs to choose the tuning parameter τ_j. Our current theory provides only some guidance on the order of the tuning parameter τ_j to ensure the consistency of the method in variable selection. Fan and Li [34], Khalili and Chen [13] showed that the tuning parameter selected by GCV leads to a nonignorable overfitting effect in the final selected model. They used the BIC for choosing the tuning parameter, and the method was shown to be consistent in selecting the true sparse model. We also propose using a BIC-type approach.

Consider the maximizer $\mathbf{\Psi}_n$ of the log-likelihood function (3) which is based on the full SNMJLSM. The estimator $\mathbf{\Psi}_n$ is used to calculate the weights ω_{ij} in (8). The weights remain fixed throughout the tuning parameter selection process. For a given value of the τ_j, let $(\hat{\boldsymbol{\beta}}_j, \hat{\boldsymbol{\gamma}}_j, \hat{\lambda}_j)$ be the maximum regularized likelihood estimates of the parameters in the jth component of the SNMJLSM by fixing the remaining elements of $\mathbf{\Psi}$ at $\mathbf{\Psi}_n$. Denote the likelihood-based deviance statistics, evaluated at $(\hat{\boldsymbol{\beta}}_j, \hat{\boldsymbol{\gamma}}_j, \hat{\lambda}_j)$, corresponding to the jth component of SNMJLSM as

$$D_j(\hat{\boldsymbol{\beta}}_j, \hat{\boldsymbol{\gamma}}_j, \hat{\lambda}_j) = \sum_{i=1}^{n} \omega_{ij} \log SN(y_i; \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_j, \exp(\boldsymbol{h}_i^T \hat{\boldsymbol{\gamma}}_j), \hat{\lambda}_j).$$

We define

$$BIC(\tau_j) = 2D_j(\hat{\boldsymbol{\beta}}_j, \hat{\boldsymbol{\gamma}}_j, \hat{\lambda}_j) + N(\tau_j) \log n_j, j = 1, 2, \cdots, m. \tag{10}$$

where $n_j = \sum_{i=1}^{n} \omega_{ij}$ is expected sample size from the nonzero element of the SNMJLSM , and $N(\tau_j)$ is the number of nonzero elements of $\hat{\boldsymbol{\beta}}_j$ and $\hat{\boldsymbol{\gamma}}_j$, respectively.

Similar to Wu *et al.* [6] , we suggest

$$(i)\tau_{1j} = \frac{\tau_j}{|\hat{\boldsymbol{\beta}}_j^0|}, j = 1, 2, \cdots, m.$$
$$(ii)\tau_{2j} = \frac{\tau_j}{|\hat{\boldsymbol{\gamma}}_j^0|}, j = 1, 2, \cdots, m.$$

where $\hat{\boldsymbol{\beta}}_j^0$ and $\hat{\boldsymbol{\gamma}}_j^0$ are initial estimators of $\boldsymbol{\beta}_j$ and $\boldsymbol{\gamma}_j$, respectively. The tuning parameter τ_j can be obtained as

$$\hat{\tau_j} = arg \min_{\tau_j} BIC(\tau_j).$$

From our simulation study, we found that this method works well.

§4 Simulation study

To evaluate the finite sample performance of the proposed penalized likelihood method, we conduct some Monte Carlo simulations. The performance of estimators $\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_n, \hat{\lambda}_n$ will be assessed by using the mean square error(MSE), defined as

$$\text{MSE}(\hat{\boldsymbol{\beta}}_n) = \text{E}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)^T(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0),$$
$$\text{MSE}(\hat{\boldsymbol{\gamma}}_n) = \text{E}(\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0)^T(\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0),$$
$$\text{MSE}(\hat{\lambda}_n) = \text{E}(\hat{\lambda}_n - \lambda_0)^2.$$

To save space we have reported only the results for SNMJLSM with $m = 2$. We simulate data from the model (3):

$$\begin{cases} y_i \sim \pi_1 SN(\mu_{i1}, \sigma_{i1}^2, \lambda_1) + \pi_2 SN(\mu_{i2}, \sigma_{i2}^2, \lambda_2), \\ \mu_{ij} = \boldsymbol{x}_i^T \boldsymbol{\beta}_j, \\ \log \sigma_{ij}^2 = \boldsymbol{h}_i^T \boldsymbol{\gamma}_j, \\ i = 1, 2, ..., n; j = 1, 2. \end{cases} \tag{11}$$

To perform this simulation, we take $\boldsymbol{\beta}_1 = (0.8, 1.8, 0, 0, 2.8, 0, 0, 0)^T$, $\boldsymbol{\beta}_2 = (0, 0, 0.8, 1.8, 0, 0, 2.8, 3.8)^T$, $\boldsymbol{\gamma}_1 = (0.5, 1.5, 0, 0, 2.5, 0, 0, 0)^T$,$\boldsymbol{\gamma}_2 = (0, 0, 0.5, 1.5, 0, 0, 2.5, 3.5)^T$, respectively. The covariates $x_i \sim U(-1, 1)$, $h_i \sim U(-1, 1)$, $\pi_1 = 0.35, 0.5$, $\pi_2 = 1 - \pi_1$, $\lambda_1 = 0.5, \lambda_2 = -0.5$, $n = 150, 250, 500$. y_i is generated according to the model (11).

For the sake of comparison, we carry out simulations with three penalties as described above. The performance of proposed method for variable selection is investigated via simulations.

The simulation results are reported using the following two quantities:

C: average number of zero regression coefficients that are correctly estimated as zero.

IC: average number of nonzero regression coefficients that are correctly estimated as zero.

Note: according to $\boldsymbol{\beta}_1 = (0.8, 1.8, 0, 0, 2.8, 0, 0, 0)^T$,$\boldsymbol{\beta}_2 = (0, 0, 0.8, 1.8, 0, 0, 2.8, 3.8)^T$, $\boldsymbol{\gamma}_1 = (0.5, 1.5, 0, 0, 2.5, 0, 0, 0)^T$,$\boldsymbol{\gamma}_2 = (0, 0, 0.5, 1.5, 0, 0, 2.5, 3.5)^T$, we know that component 1 has 5 really zero regression coefficients in the location model and scale model, respectively. Component 2 has 4 really zero regression coefficients in the location model and scale model, respectively.

Table 1. Simulation results of location model (Parameter $\boldsymbol{\beta}$).

			π_1=0.35			π_1=0.5		
Method	Component	n	C	IC	MSE	C	IC	MSE
SCAD	Component 1	150	4.3500	0.0100	0.2064	4.7600	0	0.0631
		250	4.8700	0	0.0428	4.8700	0	0.0258
		500	4.9200	0	0.0149	4.8800	0	0.0116
	Component 2	150	3.8100	0	0.0186	3.8600	0	0.0264
		250	3.9400	0	0.0054	3.8800	0	0.0094
		500	3.9500	0	0.0025	3.9700	0	0.0039
LASSO	Component 1	150	3.7700	0.0100	0.1960	4.2300	0	0.0696
		250	4.3400	0	0.0484	4.3100	0	0.0313
		500	4.5300	0	0.0181	4.5600	0	0.0137
	Component 2	150	3.3200	0	0.0236	3.2800	0	0.0316
		250	3.6300	0	0.0065	3.3800	0	0.0111
		500	3.6600	0	0.0029	3.6200	0	0.0044
HARD	Component 1	150	4.5800	0.0100	0.1953	4.9500	0	0.0538
		250	4.9500	0	0.0383	4.9200	0	0.0270
		500	4.9700	0	0.0154	4.9900	0	0.0109
	Component 2	150	3.9500	0	0.0172	3.9100	0	0.0271
		250	3.9600	0	0.0052	3.9600	0	0.0091
		500	3.9800	0	0.0026	3.9900	0	0.0040

Table 2. Simulation results of scale model (Parameter $\boldsymbol{\gamma}$).

			π_1=0.35			π_1=0.5		
Method	Component	n	C	IC	MSE	C	IC	MSE
SCAD	Component 1	150	4.1900	0.6900	2.9315	4.6600	0.7000	0.6776
		250	4.7200	0.6300	0.5574	4.9000	0.5800	0.3333
		500	4.9200	0.5300	0.2432	4.9200	0.3600	0.1534
	Component 2	150	3.8200	0.5900	0.5856	3.7600	0.6400	0.7729
		250	3.9000	0.4400	0.3206	3.7900	0.5300	0.4150
		500	3.9300	0.2200	0.1418	3.8800	0.3300	0.2061
LASSO	Component 1	150	4.0500	0.6500	2.1286	4.5100	0.6500	0.5599
		250	4.5600	0.5200	0.4648	4.7200	0.4500	0.3224
		500	4.8200	0.3000	0.2446	4.8200	0.1800	0.1665
	Component 2	150	3.6500	0.4400	0.6206	3.5400	0.5500	0.7138
		250	3.7900	0.2800	0.3104	3.7200	0.4100	0.3923
		500	3.8200	0.0600	0.1450	3.7900	0.2200	0.2117
HARD	Component 1	150	3.2900	0.5200	3.1657	4.1400	0.5500	0.8497
		250	4.2900	0.4400	0.6389	4.7100	0.4100	0.3628
		500	4.8200	0.3600	0.2270	4.9200	0.2000	0.1179
	Component 2	150	3.6600	0.4000	0.5994	3.1300	0.4700	0.9940
		250	3.8500	0.3100	0.3029	3.7700	0.4300	0.3907
		500	3.9200	0.0900	0.1103	3.9000	0.2500	0.1836

Table 3. Simulation results of skewness parameter λ.

		π_1=0.35		π_1=0.5	
Component	n	Estimate	MSE	Estimate	MSE
Component 1	150	0.7768	0.2510	0.6908	0.0828
	250	0.5756	0.0354	0.5733	0.0334
	500	0.5316	0.0148	0.5464	0.0102
Component 2	150	-0.6153	0.0436	-0.6616	0.0808
	250	-0.5617	0.0191	-0.5627	0.0243
	500	-0.5243	0.0071	-0.5364	0.0106

From Tables 1–3 , we have the following observations:

(1) For a given penalty, as expected, the performance of variable selection for components 1 and 2 become better and better as the sample size n increases. The MSEs of estimators $\hat{\boldsymbol{\beta}}_j$, $\hat{\boldsymbol{\gamma}}_j$ and $\hat{\lambda}_j (j = 1, 2)$ also become smaller as n increases, which indicates the convergence property of the maximum penalized likelihood estimator of the model.

(2) For a given sample size n, the performances of three variable selection procedures are similar in terms of model complexity. The performances of SCAD and HARD procedures are similar in terms of model error. Furthermore, the performances of SCAD and HARD are better than that of LASSO in terms of model error. However, this paper proposed method does not perform well for small sample sizes.

(3) For a given penalty function and sample size n, the performance of variable selection for components 1 and 2 in the location model is significantly better than that of the scale model in the sense of model error and model complexity. It is may be that the estimation of scale model parameters is not unbiased.

(4) For a given sample size n and the mixture proportion $\pi_1 = \pi_2 = 0.5$, as expected, the performances of three variable selection procedures in two subpopulation is similar in the

sense of model error and model complexity. However, in the case $\pi_1 = 0.35, \pi_2 = 0.65$, the performance of three variable selection procedures in the second subpopulation is significantly better than that of the first subpopulation in the sense of model error and model complexity.

§5 Application to real data

In this section, we illustrate the proposed variable selection procedure by using the *air quality index* (AQI) data. We collected the daily average value of the AQI data of Hangzhou city and Zhengzhou city in China from May 1, 2015 to March 31, 2017, totaling 670 data. This AQI data set consists of the response variable Y–AQI and seven predictors: X_1–fine particulate matter(PM2.5); X_2– inhalable particulate matter(PM10); X_3–Sulfur dioxide(SO_2); X_4–Carbon monoxide (CO); X_5–Nitrogen dioxide(NO_2); X_6–Ozone(O_3) and X_7–AQI day ranking. We are interested in establishing the relationship between the Y–AQI and the important predictors.

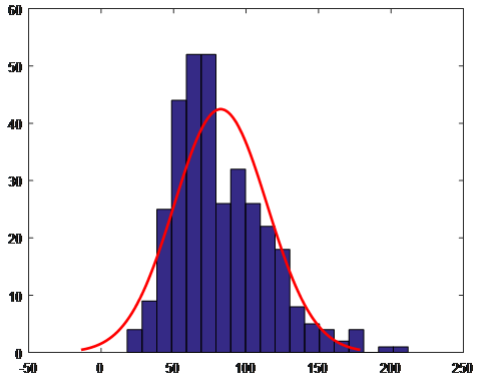

Figure 1. Histogram of AQI for hangzhou city.

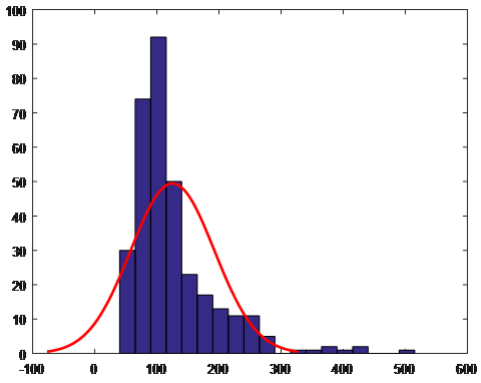

Figure 2. Histogram of AQI for zhengzhou city.

Figures 1 and 2 indicate that the AQI data of Hangzhou city and Zhengzhou city follow approximately a skew-normal distribution, respectively. AQI data in Hangzhou is concentrated in the 50–100, while the AQI data of Zhengzhou is mainly concentrated in 100–200. Thus, there

are some differences in the air quality index of two part. It may not be appropriate to describe it with the same model, so it is analyzed using a finite mixture of models.

According to the above analysis, Y, the AQI data of Hangzhou city and Zhengzhou city follow approximately a skew-normal distribution. So, we can consider the AQI data variable selection for the following skew-normal mixture joint location and scale models(SNMJLSM):

$$\begin{cases} y_i \sim 0.5SN(y_i;\mu_{i1},\sigma_{i1}^2,\lambda_1)+0.5SN(y_i;\mu_{i2},\sigma_{i2}^2,\lambda_2), \\ \mu_{ij}=\boldsymbol{x}_i^T\boldsymbol{\beta}_j, \\ \log\sigma_{ij}^2=\boldsymbol{h}_i^T\boldsymbol{\gamma}_j, \\ i=1,2,...,670; j=1,2. \end{cases}$$

We apply the variable selection procedure based on the SCAD, LASSO and HARD proposed in Section 2 to the above model. The results are displayed in Table 4, where H and Z denote Hangzhou and Zhengzhou, respectively.

Table 4. Variable selection for the *air quality index* (AQI) data.

			Constant	x_1	x_2	x_3	x_4	x_5	x_6	x_7
$\boldsymbol{\beta}$	SCAD	H	0.2019	0.0821	0	0	0.8820	1.1515	0	0.0047
		Z	-11.473	0.7792	0.1552	0	5.2534	0	0.0776	0.0912
	LASSO	H	0	0.0952	0	0	0.1938	1.1520	0	0.0056
		Z	-2.9686	0.7957	0.1587	0	2.8099	0	0.0637	0.0711
	HARD	H	0.1965	0.0816	0	0	0.8824	1.1522	0	0.0046
		Z	-16.373	0.7560	0.1664	0	5.8392	0	0.1007	0.1008
$\boldsymbol{\gamma}$	SCAD	H	2.8468	0.1011	0	0	0	-0.1280	0.0196	0.0078
		Z	3.6748	0.0067	0	0	0	0	0.0126	0
	LASSO	H	2.8843	0.0912	0	0	0	-0.1150	0.0188	0.0076
		Z	2.7953	0.0069	0	0	0	0	0.0138	0
	HARD	H	2.8591	0.1013	0	0	0	-0.1280	0.0196	0.0078
		Z	3.8429	0.0066	0	0	0	0	0.0110	0
λ	SCAD	H				5.6845				
		Z				1.1668				
	LASSO	H				5.6845				
		Z				1.1668				
	HARD	H				5.6845				
		Z				1.1668				

From Table 4, we can see that in this data example,

(1) The estimation and variable selection procedure based on the SCAD, LASSO and HARD method perform very similarly in terms of the selected variables.

(2) In the location model, β_3 is zero. This indicates $X_3(SO_2)$ has no significant impact the location of Y (AQI). For Hangzhou, $X_2(PM10)$ and $X_6(O_3)$ are unimportant variable impact the location of Y (AQI). For Zhengzhou, $X_5(NO_2)$ is an unimportant variable impact the location of Y (AQI). There may be differences between North and South air pollutants. The northern winter supply of heating needs to burn a large number of coal so that variable selection results of Hangzhou and Zhengzhou are different.

(3) In the scale model, γ_2, γ_3 and γ_4 are zero. This indicates the $X_2(PM10)$, $X_3(SO_2)$ and $X_4(CO)$ have little influence on the scale parameters of Y (AQI) of Hangzhou and Zhengzhou. For Zhengzhou, however, $X_5(NO_2)$ and X_7 (AQI day ranking) have no significant impact on the scale parameters of Y (AQI).

(4) In the skewness parameter, the Y (AQI) of Hangzhou is $\lambda_1 = 5.6845$, Zhengzhou is $\lambda_1 = 1.1668$. This indicates the AQI data of Hangzhou and Zhengzhou have positively skewed, respectively.

§6 Conclusions

In this paper, our chief interest is to consider a variable selection procedure based on the skew-normal distribution for mixture of joint location and scale models. On the basis of the traditional selection of the mean variables in the finite mixture of regression models, we further consider the variable selection in variance. The simulation studies show that the procedure is to be consistent in selecting the most parsimonious mixture of joint location and scale models. The proposed method is applied to a real data set, the results show that the proposed variable selection procedure can be used in practical situations.

In addition, we only consider variable selection in mixture of joint location and scale models based on the skew-normal distribution. A natural future extension of this work is to consider generalizations of the skew-normal distribution (e.g. skew-t-normal distribution and skew-t distributions), which may be more suitable in different contexts. Furthermore, one interesting future direction is to extend the proposed model to the mixture of experts model framework [38].

Appendix Regularity Conditions and Proofs

Let $f(\nu;\mathbf{\Psi})$ be the density function of $V = (x, Y)$, with the parameter space $\mathbf{\Psi} \in \Omega$. In the regularity conditions we write $\mathbf{\Psi} = (\mathbf{\Psi}_1, \mathbf{\Psi}_2, \cdots, \mathbf{\Psi}_s)$, so that s is the total number of parameters in the model.

R_1: The density $f(\nu;\mathbf{\Psi})$ has common support in ν for all $\mathbf{\Psi} \in \Omega$, and $f(\nu;\mathbf{\Psi})$ is identifiable with respect to $\mathbf{\Psi}$.

R_2: There exists an open subset $\Omega^* \in \Omega$, containing the true parameter $\mathbf{\Psi}_0$ such that for almost all ν, $f(\nu;\mathbf{\Psi})$ admits third partial derivatives with respect to $\mathbf{\Psi} \in \Omega^*$.

R_3: For all j, $l = 1, 2, \cdots, s$, the first and second derivatives of $f(\nu;\mathbf{\Psi})$ satisfy:

$$E_0[\frac{\partial}{\partial \mathbf{\Psi}_j} \log f(V;\mathbf{\Psi})] = 0,$$

$$E_0[\frac{\partial}{\partial \mathbf{\Psi}_j} \log f(V;\mathbf{\Psi}) \frac{\partial}{\partial \mathbf{\Psi}_i} \log f(V;\mathbf{\Psi})] = -E_0[\frac{\partial^2}{\partial \mathbf{\Psi}_j \partial \mathbf{\Psi}_i} \log f(V;\mathbf{\Psi})].$$

R_4: The Fisher information matrix is finite and positive definite at $\mathbf{\Psi} = \mathbf{\Psi}_0$:

$$I(\mathbf{\Psi}) = E\{[\frac{\partial}{\partial \mathbf{\Psi}} \log f(V;\mathbf{\Psi})][\frac{\partial}{\partial \mathbf{\Psi}} \log f(V;\mathbf{\Psi})]^T\}.$$

R_5: There exists an integrable function $B(\nu)$ such that:

$$|\frac{\partial f(V;\mathbf{\Psi})}{\partial \mathbf{\Psi}_j}| \le B(\nu), |\frac{\partial^2 f(V;\mathbf{\Psi})}{\partial \mathbf{\Psi}_j \partial \mathbf{\Psi}_i}| \le B(\nu), |\frac{\partial^3 \log f(V;\mathbf{\Psi})}{\partial \mathbf{\Psi}_j \partial \mathbf{\Psi}_i \partial_m}| \le B(\nu).$$

Proof of Theorem 1 Let $\xi_n = n^{-\frac{1}{2}}(1 + q_{1n}^* + q_{1n})$. It is sufficient to show that for any

given $\varepsilon > 0$, there exists a constant C such that

$$\lim_{n\to\infty} P\{\sup_{\|u\|=C} L_n(\mathbf{\Psi}_0 + \xi_n u) < L_n(\mathbf{\Psi}_0)\} \geq 1 - \varepsilon. \tag{A.1}$$

This implies that for large n, with probability at least $1-\varepsilon$, there is a local maximum in the ball $\{\mathbf{\Psi}_0 + \xi_n u; \|u\| \leq C\}$. This local maximizer, say $\widehat{\mathbf{\Psi}}_n$, satisfies $\|\widehat{\mathbf{\Psi}}_n - \mathbf{\Psi}_0\| = O_p(\xi_n)$.

We proceed as follows. Let $D_n(u) = L_n(\mathbf{\Psi}_0 + \xi_n u) - L_n(\mathbf{\Psi}_0)$. By definition of $L_n(\cdot)$,

$$D_n(u) = [l_n(\mathbf{\Psi}_0 + \xi_n u) - l_n(\mathbf{\Psi}_0)] - [p_n(\mathbf{\Psi}_0 + \xi_n u) - p_n(\mathbf{\Psi}_0)].$$

By $p_n(0;\tau_j) = 0$, and the definitions of $l_n(\mathbf{\Psi})$ and $p_n(\cdot)$,

$$\begin{aligned} l_n(\mathbf{\Psi}_0 + \xi_n u) - l_n(\mathbf{\Psi}_0) &= n^{-1/2}(1 + q_{1n}^* + q_{1n}) l_n^{'}(\mathbf{\Psi}_0)^T u \\ &\quad - \frac{(1 + q_{1n}^* + q_{1n})^2}{2}(u^\tau I(\mathbf{\Psi}_0) u)(1 + o_p(1)) \end{aligned}$$

and

$$\begin{aligned} \mid p_n(\mathbf{\Psi}_0 + \xi_n u) - p_n(\mathbf{\Psi}_0) \mid &\leq d(q_{1n}^* + q_{1n})\|u\| + \frac{c_n}{2}(1 + q_{1n}^* + q_{1n})^2\|u\|^2 \\ &\quad + \sqrt{m} a_n (1 + q_{1n}^* + q_{1n})\|u\|. \end{aligned}$$

where $d = \max_j \sqrt{d_j}$ and the d_j is the number of nonzero regression coefficients in the jth component of the FMR model. Regularity conditions imply that $l_n^{'}(\mathbf{\Psi}_0) = O_p(\sqrt{n})$ and $I(\mathbf{\Psi}_0)$ is positive definite. In addition, $c_n = o(1)$ and $a_n = o(1 + q_{1n}^* + q_{1n})$. The order comparison of the foregoing expression implies that for sufficiently large C, the quadratic function

$$-(1 + q_{1n}^* + q_{1n})^2 \{u^\tau I(\mathbf{\Psi}_0) u\}\{1 + o_p(1)\}/2,$$

and thus for any given $\varepsilon > 0$, we have $\lim_{n\to\infty} P\{\sup_{\|u\|=c} D_n(u) < 0\} > 1 - \varepsilon$ which is (A.1).

The results of the following Lemma is used to prove Theorem 2.

Lemma 1 Under the conditions of Theorem 2, for any $\mathbf{\Psi}$ in the neighborhood$\|\mathbf{\Psi} - \mathbf{\Psi}_0\| = O(n^{-1/2})$. By the definition of $L_n(\cdot)$, we have that

$$L_n(\mathbf{\Psi}_1, \mathbf{\Psi}_2) - L_n(\mathbf{\Psi}_1, 0) = [l_n(\mathbf{\Psi}_1, \mathbf{\Psi}_2) - l_n(\mathbf{\Psi}_0, 0)] - [p_n(\mathbf{\Psi}_1, \mathbf{\Psi}_2) - p_n(\mathbf{\Psi}_0, 0)].$$

By the mean value theorem,

$$l_n(\mathbf{\Psi}_1, \mathbf{\Psi}_2) - l_n(\mathbf{\Psi}_1, 0) = [\frac{\partial l_n^\tau(\mathbf{\Psi}_1, \xi)}{\partial \mathbf{\Psi}_2}]\mathbf{\Psi}_2 \tag{A.2}$$

Where $\|\xi\| \leq \|\mathbf{\Psi}_2\| = O(n^{-1/2})$. Also, by R_5 and the mean value theorem,

$$\begin{aligned} \|\frac{\partial l_n(\mathbf{\Psi}_1, \xi)}{\partial \mathbf{\Psi}_2} - \frac{\partial l_n(\mathbf{\Psi}_{01}, 0)}{\partial \mathbf{\Psi}_2}\| &\leq \|\frac{\partial l_n(\mathbf{\Psi}_1, \xi)}{\partial \mathbf{\Psi}_2} - \frac{\partial l_n(\mathbf{\Psi}_1, 0)}{\partial \mathbf{\Psi}_2}\| + \|\frac{\partial l_n(\mathbf{\Psi}_1, 0)}{\partial \mathbf{\Psi}_2} - \frac{\partial l_n(\mathbf{\Psi}_{01}, 0)}{\partial \mathbf{\Psi}_2}\| \\ &\leq [\sum_{i=1}^n B(Z_i)]\{\|\xi\| + \|\mathbf{\Psi}_1 - \mathbf{\Psi}_{01}\|\}. \end{aligned}$$

By the regularity conditions $R_1 - R_5$, $\partial l_n(\mathbf{\Psi}_{01}, 0)/\partial \mathbf{\Psi}_2 = O_p(n^{1/2})$, and thus $\partial l_n(\mathbf{\Psi}_1, 0)/\partial \mathbf{\Psi}_2 = O_p(n^{1/2})$. Applying this to (A.2), we get

$$l_n(\mathbf{\Psi}_1, \mathbf{\Psi}_2) - l_n(\mathbf{\Psi}_1, 0) = O_p(\sqrt{n})\{\sum_{j=1}^m \sum_{t=d_j+1}^p |\beta_{jt}| + \sum_{j=1}^m \sum_{t=d_j+1}^q |\gamma_{jt}|\}$$

for large n. On the other hand,

$$p_n(\mathbf{\Psi}_1, \mathbf{\Psi}_2) - p_n(\mathbf{\Psi}_1, 0) = \sum_{j=1}^m \sum_{t=d_j+1}^p p_n(\beta_{jt}; \tau_{1j}) + \sum_{j=1}^m \sum_{t=d_j+1}^q p_n(\gamma_{jt}; \tau_{2j}).$$

Therefore,

$$L_n(\mathbf{\Psi}_1, \mathbf{\Psi}_2) - L_n(\mathbf{\Psi}_1, 0) =$$
$$\sum_{j=1}^{m}\sum_{t=d_j+1}^{p}\{|\beta_{jt}|O_p(\sqrt{n}) - \frac{p_n(\beta_{jt};\tau_{1j})}{\sqrt{n}} + \sum_{j=1}^{m}\sum_{t=d_j+1}^{q}\{|\beta_{jt}|O_p(\sqrt{n}) - \frac{p_n(\gamma_{jt};\tau_{2j})}{\sqrt{n}}.$$

By conditions C_2 on the $p_n(\theta;\eta)$, the two double sums are negative, for β_{jt} and γ_{jt} in a shrinking neighborhood of 0. This completes the proof of the Lemma 1.

Proof of Theorem 2 Part (a). Consider the partition $\mathbf{\Psi} = (\mathbf{\Psi}_1, \mathbf{\Psi}_2)$, let $\mathbf{\Psi} = (\hat{\mathbf{\Psi}}_1, 0)$ be the maximizer of function $L_n(\mathbf{\Psi}_1, 0)$. It suffices to show that in the neighbourhood $\|\mathbf{\Psi} - \mathbf{\Psi}_0\| = O(n^{-1/2})$, as $n \to \infty$, with probability tending to one: $L_n(\mathbf{\Psi}_1, \mathbf{\Psi}_2) < L_n(\hat{\mathbf{\Psi}}_1, 0)$. The claim is proved as follows.

By the definition of $\hat{\mathbf{\Psi}}_1$, we have $L_n(\mathbf{\Psi}_1, 0) < L_n(\hat{\mathbf{\Psi}}_1, 0)$. Thus,

$$L_n(\mathbf{\Psi}_1, \mathbf{\Psi}_2) < L_n(\hat{\mathbf{\Psi}}_1, 0) \le L_n(\mathbf{\Psi}_1, \mathbf{\Psi}_2) < L_n(\mathbf{\Psi}_1, 0).$$

By Lemma 1, $L_n(\mathbf{\Psi}_1, \mathbf{\Psi}_2) < L_n(\mathbf{\Psi}_1, 0) < 0$, with probability tending to one, as $n \to \infty$.

Part (b). The regularized log-likelihood function $L_n(\mathbf{\Psi}_1, 0)$ is considered as a function of $\mathbf{\Psi}_1$. In light of Theorem 2, there exists a $\sqrt{n}$ consistent local maximizer of this function, say $\hat{\mathbf{\Psi}}_1$, such that:

$$\frac{\partial L_n(\widehat{\mathbf{\Psi}}_n)}{\partial \mathbf{\Psi}_1} = \frac{\partial \tilde{l}_n(\mathbf{\Psi})}{\partial \mathbf{\Psi}_1} - \frac{\partial R_n(\mathbf{\Psi})}{\partial \mathbf{\Psi}_1}|_{\hat{\mathbf{\Psi}}_n=(\hat{\mathbf{\Psi}}_1,0)} = 0.$$

By substituting the first order Taylors expansions of $\partial\tilde{l}_n(\mathbf{\Psi})/\partial\mathbf{\Psi}_1$ and $\partial R_n(\mathbf{\Psi})/\partial\mathbf{\Psi}_1$ into the above expression, we have

$$\{-l_n^{''}(\mathbf{\Psi}_{01}) + p_n^{''}(\mathbf{\Psi}_{01}) + o_p(n)\}(\hat{\mathbf{\Psi}}_1 - \mathbf{\Psi}_{01}) = l_n^{'}(\mathbf{\Psi}_{01}) - p_n^{'}(\mathbf{\Psi}_{01}),$$

where $p_n^{'}, l_n^{'}, p_n^{''}, l_n^{''}$ are the gradient and the matrix of the second derivatives of $p_n(\cdot)$ and $l_n(\cdot)$, respectively, and I is an identity matrix of the required dimension. Under the regularity conditions $R_1 - R_5$,

$$-l_n^{''}(\mathbf{\Psi}_{01})/n = I_1(\mathbf{\Psi}_{01}),$$
$$l_n^{'}(\mathbf{\Psi}_{01})/\sqrt{n} \xrightarrow{d} N(0, I_1(\mathbf{\Psi}_{01})).$$

Using the above facts and Slutskys Theorem, the results follows.

References

[1] J Engel, A F Huele. *A generalized linear modeling approach to robust design,* Technometrics, 1996, 38(4): 365-373.

[2] R E Park. *Estimation with heteroscedastic error terms,* Econometrica, 1966, 34: 888.

[3] A C Harvey. *Estimating regression models with multiplicative heteroscedasticity,* Econometrica, 1976, 44(3): 461-465.

[4] M Aitkin. *Modelling variance heterogeneity in normal regression using GLIM,* J R Statist Soc Ser C, 1987, 36(3): 332-339.

[5] L C Wu, Z Z Zhang, D K Xu. *Variable selection in joint mean and variance models of Box-Cox transformation,* J Appl Statist, 2012, 39(12): 2543-2555.

[6] L C Wu, Z Z Zhang, D K Xu. *Variable selection in joint location and scale models of the skew-normal distribution,* J Stat Comput Simul, 2013, 83(7): 1266-1278.

[7] L C Wu, G L Tian, Y Q Zhang, T Ma. *Variable selection in joint location, scale and skewness models with a skew-t-normal distribution,* Stat Interface, 2017, 10(2): 217-277.

[8] S M Goldfeld, R E Quandt. *A Markov model for switching regression,* J Econometrics, 1973, 1(1): 3-15.

[9] W S DeSarbo, W L Cron. *A maximum likelihood methodology for cluster wise linear regressions,* J Classification, 1988, 5(2): 249-282.

[10] P N Jones, G J McLachlan. *Fitting finite mixture models in a regression context,* Aust N Z J Stat, 1992, 34(2): 233-240.

[11] K Jedidi, V Ramaswamy, W S DeSarbo. *On estimating finite mixtures of multivariate regression and simultaneous equation models,* Struct Equ Model, 1996, 3(3): 266-289.

[12] G McLachlan, D Peel. *Finite Mixture Models,* Wiley, New York, 2000.

[13] A Khalili, J Chen. *Variable selection in finite mixture of regression models,* J Am Statist Assoc, 2007, 102(479): 1025-1038.

[14] A Khalili. *New estimation and feature selection methods in mixture-of-experts models,* Canad J Statist, 2010, 38(4): 519-539.

[15] A Khalili, S L Lin. *Regularization in finite mixture of regression models with diverging number of parameters,* Biometrics, 2011, 69(2): 436-446.

[16] A Khalili. *An overview of the new feature selection methods in finite mixture of regression models,* J Ital Stat Assoc, 2011, 10(2): 201-235.

[17] Y T Du, A Khalili, G N Johanna, J S Russell. *Simultaneous fixed and random effects selection in finite mixture of linear mixed-effects models,* Canad J Statist, 2013, 41(4): 596-616.

[18] E Ormoz, F Eskandari. *Variable selection in finite mixture of semiparametric regression models,* Comm Statist Theory Methods, 2016, 45(3): 695-711.

[19] K J Lee, R B Chen, Y N Wu. *Bayesian variable selection for finite mixture model of linear regressions,* Comput Statist Data Anal, 2016, 95: 1-16.

[20] A Khalili, J Chen, D A Stephens. *Regularization and selection in Gaussian mixture of autoregressive models,* Canad J Statist, 2017, 45(4): 356-374.

[21] Q G Tang, J Karunamun. *Robust variable selection for finite mixture regression models,* Ann Inst Statist Math, 2018, 70(3): 489-521.

[22] M Liu, T I Lin. *A skew-normal mixture regression model,* Educ Psychol Meas, 2014, 74(1): 139-162.

[23] E Cepeda, D Gamerman. *Bayesian modeling of variance heterogeneity in normal regression models,* Braz J Probab Stat, 2001, 14: 207-221.

[24] J T Taylor, A P Verbyla. *Joint modelling of location and scale parameters of the t distribution,* Statist Model, 2004, 4(2): 91-112.

[25] Z Z Zhang, D R Wang. *Simultaneous variable selection for heteroscedastic regression models,* Sci China Math, 2011, 54(3): 515-530.

[26] L C Wu, H Q Li. *Variable selection for joint mean and dispersion models of the inverse Gaussian distribution,* Metrika, 2012, 75(6): 795-808.

[27] W Zhao, R Zhang. *Variable selection of varying dispersion Student-t regression models,* J Syst Sci Complex, 2015, 28(4): 961-977.

[28] A Azzalini. *A class of distributions which includes the normal ones,* Scand J Stat, 1985, 12: 171-178.

[29] H Teicher. *Identifiability of finite mixtures,* Ann Math Stat, 1963, 34: 1265-1269.

[30] N Atienza, J Garcia-Heras, J M Munoz-Pichardo. *A new condition for identifiability of finite mixture distributions,* Metrika, 2006, 63(2): 215-221.

[31] C E G Otiniano, P N Rathie, L C S M Ozelim. *On the identifiability of finite mixture of Skew-Normal and Skew-t distributions,* Stat Probabil Lett, 2015, 106: 103-108.

[32] C Hennig. *Identifiability of models for clusterwise linear regression,* J Classification, 2000, 17(2): 273-296.

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

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

[35] R Tibshirani. *Regression shrinkage and selection via the lasso,* J R Statist Soc Ser B, 1996, 58(1): 267-288.

[36] A Antoniadis. *Wavelets in statistics: A Review,* J Ital Stat Assoc, 1997, 6(2): 97-130.

[37] A P Dempster, N M Laird and D B Rubin. *Maximum likelihood from incomplete data via the EM algorithm,* J R Statist Soc Ser B, 1977, 39(1): 1-22.

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

[1]Faculty of Science, Kunming University of Science and Technology, Kunming 650093, China.
Email: wuliucang@163.com, yangsongqin@163.com

[2]School of Economics and Statistics, Guangzhou University, Guangzhou 510006, China.
Email: taoye1357@163.com