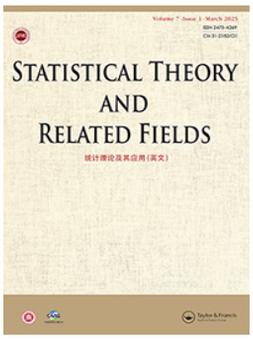

Statistical Theory and Related Fields

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tstf20

Variable selection in finite mixture of median regression models using skew-normal distribution

Xin Zeng, Yuanyuan Ju & Liucang Wu

To cite this article: Xin Zeng, Yuanyuan Ju & Liucang Wu (2023) Variable selection in finite mixture of median regression models using skew-normal distribution, Statistical Theory and Related Fields, 7:1, 30-48, DOI: 10.1080/24754269.2022.2107974

To link to this article: https://doi.org/10.1080/24754269.2022.2107974

Published online: 06 Aug 2022.

STATISTICAL THEORY AND RELATED FIELDS
2023, VOL. 7, NO. 1, 30–48
https://doi.org/10.1080/24754269.2022.2107974

Variable selection in finite mixture of median regression models using skew-normal distribution

Xin Zeng[a,b], Yuanyuan Ju[a] and Liucang Wu[a]

[a]Faculty of Science, Kunming University of Science and Technology, Kunming, People's Republic of China; [b]School of Economics, Xiamen University, Xiamen, People's Republic of China

ABSTRACT

A regression model with skew-normal errors provides a useful extension for traditional normal regression models when the data involve asymmetric outcomes. Moreover, data that arise from a heterogeneous population can be efficiently analysed by a finite mixture of regression models. These observations motivate us to propose a novel finite mixture of median regression model based on a mixture of the skew-normal distributions to explore asymmetrical data from several subpopulations. With the appropriate choice of the tuning parameters, we establish the theoretical properties of the proposed procedure, including consistency for variable selection method and the oracle property in estimation. A productive nonparametric clustering method is applied to select the number of components, and an efficient EM algorithm for numerical computations is developed. Simulation studies and a real data set are used to illustrate the performance of the proposed methodologies.

ARTICLE HISTORY
Received 18 April 2021
Revised 28 May 2022
Accepted 25 July 2022

KEYWORDS
Variable selection; mixture of median regression; skew-normal distribution; heterogeneous population; EM algorithm

1. Introduction

When the data involve asymmetrical outcomes, inference under the linear regression model with the skewed random errors can be viewed as an alternative procedure to the classical regression models with symmetric errors, since the use of a skewed distribution for the errors could reduce the influence of outliers and thus make statistical analysis more robust. Specifically, suppose that a response variable Y given a set of predictors $\boldsymbol{x}$ takes the form of

$$y = \boldsymbol{x}^\top \boldsymbol{\beta} + \epsilon, \tag{1}$$

where $\boldsymbol{\beta}$ represents a vector of the unknown regression coefficients and the conditional density of the error term ϵ given $\boldsymbol{x}$ follows an unknown distribution with the probability density function (pdf) $g(\epsilon \mid \boldsymbol{x})$. It is known that if $g(\epsilon \mid \boldsymbol{x})$ is symmetrical about 0, the estimation of β in (1) will be the same as the coefficients obtained by conventional mean linear regression. However, if $g(\epsilon \mid \boldsymbol{x})$ is skewed, the median regression provides a more reliable statistical analysis with adaptive robustness to outliers, since the median of a distribution is less susceptible to outliers, especially when the data involve asymmetrical outcomes. We here refer the interested readers to Kottas and Gelfand (2001), Zhou and Liu (2016) and Hu et al. (2019) for relevant research on the median regression of population distributions.

It is noteworthy to mention that the median regression has been widely used for studying the relationship between the response variable Y and a set of predictors $\boldsymbol{x}$ in symmetrical distribution, whereas such a median regression may not be suitable for analysing the data exhibiting asymmetrical behaviour or the data that arise from a heterogeneous population. To tackle this difficulty, mixture of regression models (known as switching regression models in econometrics), initially introduced by Goldfeld and Quandt (1973), may be employed as a flexible tool for studying the skewed data from two or more subpopulations. Since then, finite mixture of regression (FMR) models has been widely used in a variety of fields including but not limited to biology, medicine, economics, environmental science, sampling survey and engineering technology. The book by McLachlan and Peel (2004) contains a comprehensive review of FMR models. An FMR model is obtained when a response variable with a finite mixture distribution depends on a set of covariates, and FMR models have been discussed extensively when the normality is assumed for the regression error in each component.

However, it has been shown that the commonly used normal mixture model tends to be an over fitting model, since additional components are usually needed to capture the skewness of the data. To overcome the potential

CONTACT Liucang Wu wuliucang@163.com Faculty of Science, Kunming University of Science and Technology, Kunming 650031, People's Republic of China

inappropriateness of normal mixtures in some context, we may consider the use of the skew-normal distributions (Azzalini, 1985) as component densities of the errors; see, for example, Wu et al. (2013), Wu (2014), Tang and Tang (2015), and H. Li et al. (2016, 2017), to name just a few. These observations motivate us to develop a novel finite mixture of the median regression (FMMeR) model based on a mixture of the skew-normal distributions to explore asymmetrical data that arise from several subpopulations. There exist two barriers for the development of the FMMeR model. The first barrier is to deal with computational aspects of parameter estimation when fitting the FMMeR model with the skew-normal distribution for the errors. We tackle this barrier by utilizing the stochastic representation and hierarchical representation (see, for example, Liu & Lin, 2014) of skew-normal mixtures. A second technical barrier is to determine the number of components of the FMMeR model under consideration. Popularly, the log-likelihood maximum and two information-based criteria, AIC (Akaike, 1973) and BIC (Schwarz, 1978), can be used to select the number of components. Although some success has been shown using the model choice criteria, choosing the right number of components for a mixture model is known to be difficult. Thus, we consider a procedure of clustering to determine the number of components, which has been shown to be very effective via real-data example, and it is introduced in Subsection 5.3.

To enhance predictability and to give a concise model, it is reasonable to include only the significant covariates in the model. As a result, variable selection has also become increasingly important for FMR models and a rich literature has been generated in recent several decades. All-subset selection methods, such as the AIC and BIC, and their modifications, have been widely investigated in the context of FMR models; for instance, P. Wang et al. (1996) researched model selection in a finite mixture of Poisson regression models via AIC and BIC. However, all-subset selection methods for FMR models are computationally intensive. To improve computational efficiency, the least absolute shrinkage and selection operator (LASSO) of Tibshirani (1996) and the smoothly clipped absolute deviation (SCAD) method of Fan and Li (2001) are proposed as new methods for variable selection. The penalized likelihood for FMR models, the extension of penalized least square methods, were proposed by Khalili and Chen (2007). Recently, Wu et al. (2020) proposed an estimation and variable selection method for mixture of joint mean and variance models; Yin, Wu, and Dai (2020) proposed variable selection procedures in FMR models using the skew-normal distribution.

The remainder of this paper is organized as follows. In Section 2, we briefly introduce the skew-normal distribution and its median expression. In Section 3, we develop a variable selection method for FMMeR model via the penalized likelihood-based procedure for analysing asymmetrical data from several subpopulations. Section 4 studies asymptotic properties of the resulting estimators. In Section 5, a numerical algorithm, a productive nonparametric clustering method for determining the number of components and a data-adaptive method for choosing tuning parameters are discussed. In Section 6, we carry out simulation studies to investigate the finite sample performance of the proposed methodology. A real-data example is provided in Section 7 for illustrative purposes. Some concluding remarks are given in Section 8. Brief proofs of theorems and some technical derivations are given in Appendices 1 and 2.

2. The skew-normal mixture of median regression models

2.1. Skew-normal distribution

A random variable Y is said to follow a univariate skew-normal distribution with location parameter μ, scale parameter $\sigma \in (0, \infty)$ and skewness parameter $\lambda \in \mathbb{R}$, denoted by $Y \sim SN(\mu, \sigma^2, \lambda)$, if its pdf is given by

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

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the pdf and cumulative distribution function (cdf) of the standard normal distribution, respectively. It is worth noting that if $\lambda = 0$, the density of Y reduces to a normal density $N(\mu, \sigma^2)$ and that the distribution is positively skewed if $\lambda > 0$ and is negatively skewed if $\lambda < 0$.

We represent the skew-normal distribution in an incomplete data framework. Specifically, the stochastic representation for random variable $Y \sim SN(\mu, \sigma^2, \lambda)$ is given by

$$Y_i = \mu + \sigma\left(\delta(\lambda)R_i + \sqrt{1-\delta^2(\lambda)}V_i\right), \tag{3}$$

where $i = 1, \ldots, n$ with a sample size of n, $\delta(\lambda) = \lambda/\sqrt{1+\lambda^2}$. Here, $R_i \sim TN(0,1)I\{r_i > 0\}$ and $V_i \sim N(0,1)$, where R_i and V_i are independent. $R \sim TN(\mu, \sigma^2)I\{a_1 < r < a_2\}$ is a truncated normal distribution with the density

$$f_R(r \mid \mu, \sigma^2) = \left\{\Phi\left(\frac{a_2-\mu}{\sigma}\right) - \Phi\left(\frac{a_1-\mu}{\sigma}\right)\right\}^{-1} \times \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(r-\mu)^2\right\},$$

where $a_1 < r < a_2$ and $I\{\cdot\}$ represents an indicator function. For notational simplicity, let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ and $\boldsymbol{R} = (R_1, \ldots, R_n)^\top$. Furthermore, the skew-normal distribution can be decomposed into a normal distribution and a truncated normal distribution by a hierarchical representation given by

$$\begin{aligned} Y_i \mid R_i = r_i &\sim N\big(\mu + \sigma r_i \delta(\lambda), \sigma^2(1-\delta^2(\lambda))\big), \\ R_i &\sim TN\big(0,1\big)I\{r_i > 0\}. \end{aligned} \tag{4}$$

Azzalini and Capitanio (2013) adopted the moment-generating function to calculate the mean and variance for the skew-normal distribution in (2) and they are given by

$$E(Y) = \mu + \mu_0(\lambda)\sigma, \quad \mathrm{Var}(Y) = \sigma_0^2(\lambda)\sigma^2, \tag{5}$$

respectively, where $\mu_0(\lambda) \hat{=} \sqrt{2/\pi}\delta(\lambda)$ and $\sigma_0^2(\lambda) \hat{=} 1 - \mu_0^2(\lambda)$. Of particular note is that Lin et al. (2007) introduced a simple way of obtaining higher moments of the skew-normal distribution without the use of its moment-generating function. Letting $m_0(\lambda)$ be the mode of the distribution $SN(0,1,\lambda)$, a quite accurate approximation of $m_0(\lambda)$ evaluated by the numerical maximization method is given by

$$m_0(\lambda) \approx \mu_0(\lambda) - \frac{t_0(\lambda)\sigma_0(\lambda)}{2} - \frac{\mathrm{sign}(\lambda)}{2} \exp\left\{-\frac{2\pi}{|\lambda|}\right\},$$

where $\mathrm{sign}(\lambda)$ indicates the sign function for λ and

$$t_0(\lambda) \hat{=} \frac{4-\pi}{2} \frac{\mu_0^3(\lambda)}{\sigma_0^3(\lambda)}.$$

It deserves mentioning that the logarithm of the density for the skew-normal distribution is a concave function and that this property is not altered by a change of location and scale parameters. Thus, $m_0(\lambda)$ is unique and the mode of the skew-normal distribution in (2) can be reexpressed as $\mathrm{Mode}(Y) = \mu + m_0(\lambda)\sigma$. $\mathrm{Mean}(Y)$, $\mathrm{Mode}(Y)$ and $\mathrm{Median}(Y)$ have the quantitative relationship when the observations follow a skew-normal distribution: $\mathrm{Median}(Y) \approx [\mathrm{Mode}(Y) + 2\,\mathrm{Mean}(Y)]/3$, that is,

$$\mathrm{Meadian}(Y) \approx \mu + \frac{[m_0(\lambda) + 2\mu_0(\lambda)]\sigma}{3}, \tag{6}$$

which could facilitate the development of the median regression with the skew-normal mixtures discussed below.

2.2. Median regression for skew-normal mixtures

In this paper, we assume that the response variable Y_i follows a skew-normal distribution with location parameter μ_i, scale parameter σ and skewness parameter λ, denoted by $Y_i \sim SN(\mu_i, \sigma^2, \lambda)$ for $i = 1, \ldots, n$. A linear mode regression model with skew-normal errors can be expressed as

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \tag{7}$$

where $\mathrm{Median}(Y_i \mid \boldsymbol{X}) = \boldsymbol{x}_i^\top \boldsymbol{\beta} = \mu_i + [m_0(\lambda) + 2\mu_0(\lambda)]\sigma/3$ defined by (6). Here $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ is a $p \times n$ design matrix, such that each of its element $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ is the p-dimensional vector of predictors, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is a p-dimensional vector of the unknown regression coefficients, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$ stands for the n-dimensional vector of random errors such that $\epsilon_i \overset{iid}{\sim} SN(-[m_0(\lambda) + 2\mu_0(\lambda)]\sigma/3, \sigma^2, \lambda)$.

We consider the case where the data from heterogeneous populations. A finite mixture median regression (FMMeR) model with m-components of the skew-normal distributions is defined as

$$\begin{cases} f(y_i \mid \boldsymbol{\Psi}) = \sum_{j=1}^{m} \nu_j SN(y_i \mid \mu_{ij}, \sigma_j^2, \lambda_j), \quad i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, m, \\ \text{Median}(y_{ij}) = \boldsymbol{x}_i^\top \boldsymbol{\beta}_j, \end{cases} \tag{8}$$

where

$$SN(y_i \mid \mu_{ij}, \sigma_j^2, \lambda_j) = \frac{2}{\sigma_j} \phi\left(\frac{y_i - \mu_j}{\sigma_j}\right) \Phi\left(\lambda_j \frac{y_i - \mu_j}{\sigma_j}\right),$$

$\boldsymbol{\nu} = (\nu_1, \ldots, \nu_m)^\top$ are the mixing proportions which are constrained to be non-negative and sum to unity, $\boldsymbol{\beta_j} = (\beta_{j1}, \ldots, \beta_{jp})^\top$ and $\boldsymbol{\Psi} = (\nu_1, \ldots, \nu_{m-1}, \boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_m^\top, \sigma_1, \ldots, \sigma_m, \lambda_1, \ldots, \lambda_m)^\top$. It is obvious that

$$\mu_{ij} = \boldsymbol{x}_i^\top \boldsymbol{\beta}_j - \frac{\sigma_j}{3}\left[m_0(\lambda_j) + 2\sqrt{\frac{2}{\pi}}\delta(\lambda_j)\right], \tag{9}$$

which shows that the location in the FMMeR model is altered by a change of scale and skewness parameters.

2.3. Identifiability

An important part associated with statistical inference for FMR models is their identifiability. It is well known that mixture models are not absolutely identifiable in general. However, in some mixture model settings, it is possible to establish a weaker sense of identifiability. Titterington et al. (1985) have given relevant conclusions that the FMR models of continuous distribution are identifiable in most cases. Otiniano et al. (2015) introduced the identifiability of finite mixture of skew-normal distribution and gave detailed explanation. The cumulative distribution function of Y is denoted by F_Y. It is possible to define the skew-normal family as the set

$$\mathscr{F} = \left\{F : F_Y(y \mid \mu, \sigma^2, \lambda) = \int_{-\infty}^{y} f(t \mid \mu, \sigma^2, \lambda)\, \mathrm{d}t\right\}$$

and

$$\mathscr{H} = \left\{H : H(y \mid \boldsymbol{\Psi}) = \sum_{j=1}^{m} \nu_j F_j(y \mid \mu_{ij}, \sigma_j^2, \lambda_j); F_j(y \mid \mu_{ij}, \sigma_j^2, \lambda_j) \in \mathscr{F}\right\}$$

as the class of finite mixture of skew-normal distributions. The class $\mathscr{H}$ of all finite mixtures of $\mathscr{F}$ is identifiable if and only if for any $H, \bar{H} \in \mathscr{H}$,

$$H = \sum_{j=1}^{m} \nu_j F_j, \bar{H} = \sum_{j=1}^{\bar{m}} \bar{\nu}_j \bar{F}_j.$$

The equality $H = \bar{H}$ implies $m = \bar{m}$ and $(\nu_1, F_1), \ldots, (\nu_m, F_m)$ are a permutation of $(\bar{\nu}_1, \bar{F}_1), \ldots, (\bar{\nu}_m, \bar{F}_m)$. The following theorem given by Atienza et al. (2006) gives a sufficient condition for the identifiability of finite mixtures of distributions. A' denotes the accumulation set of A.

Theorem 2.1 (Atienza et al., 2006): *Let $\mathscr{F}$ be a family of distributions. Let M be a linear mapping which transforms any $F \in \mathscr{F}$ into a real function φ_F with domain $S_\varphi(F)$. Let $S_0(F) = \{k \in S_\varphi(F) : \varphi_F(k) \neq 0\}$. Suppose that there exists a total order $\prec$ on $\mathscr{F}$, such that for any $F \in \mathscr{F}$ there exists a point $k(F) \in S_0(F)'$ verifying*

(1) *if $F_1, F_2, \ldots, F_l \in \mathscr{F}$, with $F_1 \prec F_j$ for $2 \leq j \leq l$, then $k(F_1) \in [S_0(F_1) \cup [\cup_{j=2}^{l} S_\varphi(F_j)]]'$;*
(2) *if $F_1 \prec F_2$, then $\lim_{k \to k(F_1)} \frac{\varphi_{F_2}(k)}{\varphi_{F_1}(k)} = 0$.*

Then, the class $\mathscr{H}$ of all finite mixture distributions of $\mathscr{F}$ is identifiable.

3. The method for variable selection

Various classical variable selection criteria can be considered as tradeoffs based on the estimation variance and modelling biases of penalized likelihood. The density $f(\boldsymbol{x})$ is functionally independent of the parameters as an assumption in FMMeR model when $\boldsymbol{x}$ is random. Hence, the variable selection can be done based absolutely on the conditional density function specified in (2). Denote $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ as a sample of observations from FMMeR model specified in (8). The log-likelihood function of $\boldsymbol{\Psi}$ is given by

$$\ell(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \sum_{j=1}^{m} \nu_j SN(y_i \mid \mu_{ij}, \sigma_j^2, \lambda_j).$$

A maximum likelihood estimate (MLE) is obtained via maximizing $\ell(\boldsymbol{\Psi})$. The MLE is often close to, but not strictly equal to 0 when a component of $\boldsymbol{x}$ is not important. Thus, this covariate is not excluded from the model. To address this problem, according to Khalili and Chen (2007), we define a penalized log-likelihood function as

$$L(\boldsymbol{\Psi}) = \ell(\boldsymbol{\Psi}) - p(\boldsymbol{\Psi}), \tag{10}$$

with the penalty function

$$p(\boldsymbol{\Psi}) = n \sum_{j=1}^{m} \nu_j \sum_{t=1}^{p} p_{\tau_j}(|\beta_{jt}|),$$

where $p_{\tau_j}(\cdot)$ is a given penalty function with the tuning parameter $\tau_j \geq 0$ ($j = 1, 2, \ldots, m$), and the tuning parameters and the penalty functions are not necessarily the same for all the parameters. A data-driven criterion for determining tuning parameters is introduced in Subsection 5.2. By choosing appropriate tuning parameters and maximizing function $L(\boldsymbol{\Psi})$ in (10) to obtain penalized maximum likelihood estimator of $\boldsymbol{\Psi}$, denoted by $\widehat{\boldsymbol{\Psi}}$, the coefficients in the vicinity of 0 are compressed to 0 and automatically excluded. Thus, the procedure combines the parameter estimation and variable selection and reduces the computational burden substantially. We use the following three penalty functions to illustrate the theory that we develop for the FMMeR model:

$$\begin{aligned}
&\text{LASSO penalty} : p_{\tau_j}(|\beta_{jt}|) = \tau_j |\beta_{jt}|,\\
&\text{HARD penalty} : p_{\tau_j}(|\beta_{jt}|) = {\tau_j}^2 - (|\beta_{jt}| - \tau_j)^2 I(|\beta_{jt}| < \tau_j),\\
&\text{SCAD penalty} : p'_{\tau_j}(|\beta_{jt}|) = \tau_j \left\{ I(|\beta_{jt}| \leq \tau_j) + \frac{(a\tau_j - |\beta_{jt}|)}{(a-1)\tau_j} I(|\beta_{jt}| > \tau_j) \right\}.
\end{aligned}$$

Following the idea of Fan and Li (2001), we set $a = 3.7$ for application purposes in this article. The LASSO penalty has a good performance in numerical computation because of its convex property. The SCAD penalty gives a good performance in selecting important variables. HARD penalty should work more like SCAD, although less smoothly.

4. Asymptotic properties

In this section, we consider the consistency for variable selection method and the oracle property in estimation. Without loss of generality, the coefficient vector $\boldsymbol{\beta}_j (j = 1, \ldots, m)$ of the j-th component is decomposed into $\boldsymbol{\beta}_j^\top = (\boldsymbol{\beta}_{1j}^\top, \boldsymbol{\beta}_{2j}^\top)$, where $\boldsymbol{\beta}_{1j}$ and $\boldsymbol{\beta}_{2j}$ contain the nonzero effects and zero effects of $\boldsymbol{\beta}_j$, respectively. Naturally, we also split the parameter $\boldsymbol{\Psi}^\top = (\boldsymbol{\Psi}_1^\top, \boldsymbol{\Psi}_2^\top)$ such that $\boldsymbol{\Psi}_2$ contains all zero effects, that is, $\boldsymbol{\beta}_{2j}$ in the true model. The vector of true parameters is denoted as $\boldsymbol{\Psi}_0$. The components of $\boldsymbol{\Psi}_0$ are denoted with a superscript, namely $\boldsymbol{\Psi}_0 = (\nu_1^0, \ldots, \nu_{m-1}^0, \boldsymbol{\beta}_1^{0\top}, \ldots, \boldsymbol{\beta}_m^{0\top}, \sigma_1^0, \ldots, \sigma_m^0, \lambda_1^0, \ldots, \lambda_m^0)^\top$, where β_{jt}^0 is the t-th component of $\boldsymbol{\beta}_j^0$. Let d_j denote the number of nonzero elements β_{jt}^0 of the subvector $\boldsymbol{\beta}_{1j}^0$ for each j. Let

$$a_n = \max_{j,t}\{p'_{\tau_j}(\beta_{jt}^0; \beta_{jt}^0 \neq 0)\}, \quad b_n = \max_{j,t}\{p''_{\tau_j}(\beta_{jt}^0; \beta_{jt}^0 \neq 0)\},$$

where $1 \leq t \leq d_j$ and $1 \leq j \leq m$. $p'_{\tau_j}(\beta_{jt}^0)$ and $p''_{\tau_j}(\beta_{jt}^0)$ are the first and second derivatives of the penalty function $p_{\tau_j}(\beta_{jt}^0)$ with respect to β_{jt}^0, respectively. The asymptotic results obtained in this article are based on the three conditions on the penalty functions $p_{\tau_j}(\cdot)$.

C_0: For all j, $p_{\tau_j}(0) = 0$, and $p_{\tau_j}(\beta)$ is symmetric and non-negative. Furthermore, it is nondecreasing and twice differentiable for all $\beta \in (0, \infty)$ with at most a few exceptions.

C_1: As $n \to \infty$, $b_n = o(1)$.
C_2: For all j and $T_n = \{\beta; 0 < \beta \leq n^{-1/2} \log n\}$, $\lim_{n\to\infty} \inf_{\beta \in T_n} p'_{\tau_j}(\beta) = \infty$.

Condition C_1 is used to explain the asymptotic properties of the estimators of nonzero effects. Conditions C_0 and C_2 are required for sparsity.

Theorem 4.1 (Consistency): *Let $h_i = (\boldsymbol{x}_i, Y_i)$, $i = 1, 2, \ldots, n$, be a random sample from a density function $f(h, \boldsymbol{\Psi})$ that satisfies the regularity conditions R_1–R_4 in the Appendix* 1. *The penalty functions $p_{\tau_j}(\cdot)$ satisfy conditions C_0 and C_1 as a assumption. Then there exists a local maximizer $\widehat{\boldsymbol{\Psi}}$ of the penalized log-likelihood function $L(\boldsymbol{\Psi})$ for which*

$$\|\widehat{\boldsymbol{\Psi}} - \widehat{\boldsymbol{\Psi}}_0\| = O_p\{n^{-1/2}(1 + a_n)\},$$

where $\|\cdot\|$ represents the Euclidean norm.

Theorem 4.2 (Oracle property): *Assume that the conditions given in Theorem* 4.1 *are fulfilled. The penalty functions $p_{\tau_j}(\cdot)$ satisfy conditions C_0–C_2, and m is known in parts* (a) *and* (b). *We then have the following.*

(a) *For any $\boldsymbol{\Psi}$ such that $\|\boldsymbol{\Psi} - \boldsymbol{\Psi}_0\| = O_p(n^{-1/2})$, with probability tending to 1,*

$$L(\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2) - L(\boldsymbol{\Psi}_1, \mathbf{0}) < 0.$$

(b) *For any $\sqrt{n}$-consistent maximum penalized likelihood estimator $\widehat{\boldsymbol{\Psi}}$ of $\boldsymbol{\Psi}$,*
(1) *sparity: $P(\hat{\boldsymbol{\beta}}_{2j} = \mathbf{0}) \to 1$, $j = 1, 2, \ldots, m$ as $n \to \infty$;*
(2) *asymptotic normality:*

$$\sqrt{n}\left\{\left[I_1(\boldsymbol{\Psi}_{01}) - \frac{p''(\boldsymbol{\Psi}_{01})}{n}\right](\widehat{\boldsymbol{\Psi}}_1 - \boldsymbol{\Psi}_{01}) + \frac{p'(\boldsymbol{\Psi}_{01})}{n}\right\} \xrightarrow{d} N(\mathbf{0}, I_1(\boldsymbol{\Psi}_{01})),$$

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

Brief proofs of theorems are put in Appendix 1. Detailed proofs can be seen in the previous literature (see, for example, Fan & Li, 2001; Khalili & Chen, 2007; Yin, Wu, & Dai, 2020).

5. Numerical computations

5.1. Maximization algorithm

In general, due to the unboundedness of the likelihood function, the maximum likelihood estimator of the mixture distribution is often inconsistent in the context of finite mixture models. The alternative is to add a regular term that prevents the likelihood function from tending to infinity to get a consistent maximum penalty likelihood estimator, see, for example, Chen and Tan (2009), Chen (2017), including recent works, Chen et al. (2020), He and Chen (2022a, 2022b). McLachlan and Peel (2004) proposed that the EM algorithm can calculate the maximum likelihood estimation of arbitrary distribution in finite mixture model. We maximize the regularized log-likelihood function by the EM algorithm. We define the latent component-indicators $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$ with $\boldsymbol{Z}_i = (z_{i1}, \ldots, z_{im})^\top$, $i =, 1 \ldots, n$. Then $\boldsymbol{Z}_i$ is an m-dimensional indicator vector with its j-th element given by

$$z_{ij} = \begin{cases} 1, & \text{if } (\boldsymbol{x}_i, y_i) \text{ belongs to } j\text{-th component,} \\ 0, & \text{otherwise.} \end{cases}$$

Since an observation cannot simultaneously belong to both components, we have $\sum_{j=1}^{m} z_{ij} = 1$. By assuming the component-indicators $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ to be independent, we obtain a conditional density of the multinomial

distribution given the mixing probabilities

$$f(z_i \mid \boldsymbol{\nu}) = \nu_1^{z_{i1}} \nu_2^{z_{i2}} \cdots \nu_{m-1}^{z_{i,m-1}} \left(1 - \sum_{j=1}^{m} \nu_j\right)^{z_{im}}, \tag{11}$$

which is denoted as $\boldsymbol{Z}_i \sim \mathcal{M}(1; \nu_1, \ldots, \nu_m)$, and it will be used in combination with (3) to generate the following hierarchical representation for the skew-normal mixtures, such that

$$\begin{aligned} Y_i \mid (r_i, z_{ij} = 1) &\sim N\big(\mu_{ij} + \sigma_j r_i \delta(\lambda_k), \sigma_j^2(1 - \delta^2(\lambda_j))\big), \\ R_i \mid z_{ij} = 1 &\sim TN(0,1)I(\tau_i > 0), \\ \boldsymbol{Z}_i &\sim \mathcal{M}(1; \nu_1, \nu_2, \ldots, \nu_m). \end{aligned} \tag{12}$$

It deserves mentioning that the hierarchical representation of the finite skew-normal mixtures in (12) allows us to address computational barriers of the parameter estimation when fitting the FMMeR model. Let $Y_{\text{obs}} = \{y_i\}_{i=1}^n$ be the observed data. For each $Y_i = y_i$, we use the latent variables $\boldsymbol{Z}_i$ and R_i to form the complete data $Y_{\text{com}} = Y_{\text{obs}} \cup Y_{\text{mis}} = \{y_i, z_{ij}, r_i\}$, where Y_{mis} denotes the missing data. From hierarchical representation (12), the complete data log-likelihood function can be given by

$$\begin{aligned} \ell_c(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \sum_{j=1}^{m} z_{ij} \Bigg\{ &\log \nu_j - \frac{1}{2} \log\left(2\pi\sigma_j^2\right) - \frac{1}{2} \log(1 - \delta^2(\lambda_j)) \\ &- \frac{1}{2\sigma_j^2(1 - \delta^2(\lambda_j))} \left[e_{ij}^2 - 2\sigma_j e_{ij} \delta(\lambda_j) r_i + \sigma_j^2 r_i^2 \delta^2(\lambda_j) \right] \Bigg\}. \end{aligned} \tag{13}$$

Similar to the approach in Fan and Li (2001), $p(\Psi)$ is replaced by the following local quadratic function given the value $\boldsymbol{\Psi}^{(0)}$,

$$\begin{aligned} p(\boldsymbol{\Psi}) \approx \widetilde{p}(\boldsymbol{\Psi}) &= p(\boldsymbol{\Psi}^{(0)}) + \frac{p'(\boldsymbol{\Psi}^{(0)})}{2\boldsymbol{\Psi}^{(0)}} (\boldsymbol{\Psi}^2 - \boldsymbol{\Psi}^{(0)^2}) \\ &= n \sum_{j=1}^{m} \nu_j \sum_{t=1}^{p} \left[p_{\tau_j}(\beta_{jt}^{(0)}) + \frac{p'_{\tau_j}(\beta_{jt}^{(0)})}{2\beta_{jt}^{(0)}} (\beta_{jt}^2 - \beta_{jt}^{(0)^2}) \right]. \end{aligned}$$

This approximation is used in the M-step of the EM algorithm in each iteration. The complete penalized log-likelihood function of (10) can be given by

$$L_c(\boldsymbol{\Psi}) = \ell_c(\boldsymbol{\Psi}) - p(\boldsymbol{\Psi}). \tag{14}$$

• E-step. The E-step computes the conditional expectation of the function $L_c(\boldsymbol{\Psi})$ with respect to z_{ij}. Given the observed data $\{x_i, y_i\}_{i=1}^n$ from FMMeR model (8), $\boldsymbol{\Psi}^{(k)}$ is denoted as parameter estimation for k-th iteration. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma, \lambda)^\top$. Then the surrogate function can be constructed as

$$Q(\boldsymbol{\Psi} \mid \boldsymbol{\Psi}^{(k)}) = Q_1(\nu \mid \boldsymbol{\Psi}^{(k)}) + Q_2(\boldsymbol{\theta} \mid \boldsymbol{\Psi}^{(k)}) - p(\boldsymbol{\Psi} \mid \boldsymbol{\Psi}^{(k)}), \tag{15}$$

where

$$Q_1(\nu \mid \boldsymbol{\Psi}^{(k)}) = \sum_{i=1}^{n} \sum_{j=1}^{m} \omega_{ij}^{(k)} \log \nu_j,$$

$$Q_2(\boldsymbol{\theta} \mid \boldsymbol{\Psi}^{(k)}) = \sum_{i=1}^{n}\sum_{j=1}^{m}\omega_{ij}^{(k)}\left[-\frac{1}{2}\log\left(2\pi\sigma_j^2\right) - \frac{1}{2}\log(1-\delta^2(\lambda_j))\right.$$
$$\left. -\frac{1}{2\sigma_j^2(1-\delta^2(\lambda_j))}\left(e_{ij}^2 - 2\sigma_j e_{ij}\delta(\lambda_j)r_{1i}^{(k)} + \sigma_j^2\delta^2(\lambda_j)r_{2i}^{(k)}\right)\right].$$

The required conditional expectations are obtained as follows. First, the conditional expectation $E_{\boldsymbol{\Psi}^{(k)}}(z_{ij} \mid y_i, \boldsymbol{x}_i)$ is given by

$$\omega_{ij}^{(k)} = \frac{\nu_j^{(k)} SN(y_i; \mu_{ij}^{(k)}, \sigma_j^{2(k)}, \lambda_j^{(k)})}{\sum_{j=1}^{m}\nu_j^{(k)} SN(y_i; \mu_{ij}^{(k)}, \sigma_j^{2(k)}, \lambda_j^{(k)})}. \tag{16}$$

Then, it can be easily shown that

$$r_{1i}^{(k)} = E(R_i \mid y_i, \boldsymbol{x}_i, z_{ij} = 1, \boldsymbol{\Psi}^{(k)}) = \frac{e_{ij}^{(k)}\delta(\lambda_j^{(k)})}{\sigma_j^{(k)}} + \frac{\delta(\lambda_j^{(k)})}{\lambda_j^{(k)}}\frac{\phi(\gamma_{ij}^{(k)})}{\Phi(\gamma_{ij}^{(k)})},$$
$$r_{2i}^{(k)} = E(R_i^2 \mid y_i, \boldsymbol{x}_i, z_{ij} = 1, \boldsymbol{\Psi}^{(k)}) = \frac{1}{1+\lambda_j^{(k)2}} + \frac{e_{ij}^{(k)}\delta(\lambda_j^{(k)})}{\sigma_j^{(k)}}r_{1i}^{(k)},$$
$$\gamma_{ij}^{(k)} = \frac{\lambda_j^{(k)}(y_i - \mu_{ij}^{(k)})}{\sigma_j^{(k)}} = \frac{\lambda_j^{(k)}e_{ij}^{(k)}}{\sigma_j^{(k)}},$$
$$\mu_{ij}^{(k)} = \boldsymbol{x}_i^{\top}\boldsymbol{\beta}_j^{(k)} - \frac{\sigma_j^{(k)}}{3}\left[m_0(\lambda_j^{(k)}) + 2\sqrt{\frac{2}{\pi}}\delta(\lambda_j^{(k)})\right].$$

• M-step. The M-step calculates parameter vector $\boldsymbol{\Psi}^{(k+1)}$ via maximizing $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})$ with respect to $\boldsymbol{\Psi}$. Thus, on the $(k+1)$-th iteration of the EM algorithm, the mixing proportions are updated by

$$\nu_j^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n}\omega_{ij}^{(k)}, \quad j = 1, 2, \ldots, m. \tag{17}$$

It is worth noting that the mixing proportions modelling should be considered in mixture of experts regression models, which can be found in Yin, Wu, Lu, et al. (2020). To improve the efficiency for selecting the number of components in real data analysis for this article, we firstly applied a clustering method to determine the optimal number of components in Subsection 5.3. By maximizing $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})$ with respect to $\boldsymbol{\Psi}$ without ν_j, namely maximizing $Q_2(\boldsymbol{\theta}; \boldsymbol{\Psi}^{(k)})$, we can compute $\boldsymbol{\theta}_j^{(k+1)}$. To obtain parameter estimation of FMMeR model without penalty, start from an initial value $\boldsymbol{\theta}^{(0)}$ and given k as the current iteration. We use the following method to update

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + [-H(\boldsymbol{\theta}^{(k)})]^{-1}S(\boldsymbol{\theta}^{(k)}), \tag{18}$$

where

$$S(\boldsymbol{\theta}) = \frac{\partial Q_2(\boldsymbol{\theta}; \boldsymbol{\Psi}^{(k)})}{\partial\boldsymbol{\theta}} = [S(\boldsymbol{\beta}), S(\sigma), S(\lambda)]^{\top}$$

is referred to as score function without penalty. $H(\boldsymbol{\theta}^{(k)})$ is an observation information matrix defined as

$$H(\boldsymbol{\theta}) = \frac{\partial^2 Q_2(\boldsymbol{\theta}; \boldsymbol{\Psi}^{(k)})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\top}}.$$

Detailed derivation can be seen in Appendix 2. We iterate between the E-steps and M-steps until algorithm converges, and the estimators $\boldsymbol{\beta}_j^{(0)}, \sigma_j^{(0)}, \lambda_j^{(0)}$ are obtained.

In order to find the non-significant variables and simplify the FMMeR model, we shrink the coefficients by the penalty function. $\boldsymbol{\beta}_j^{(0)}$ is taken as the initial value of iteration and given k as the current iteration, update

$$\boldsymbol{\beta}_j^{(k+1)} = \boldsymbol{\beta}_j^{(k)} + \left[\frac{\partial^2 Q_2(\boldsymbol{\theta}; \boldsymbol{\Psi}^{(k)})}{\partial\boldsymbol{\beta}_j\partial\boldsymbol{\beta}_j^{\top}} - n\boldsymbol{\Delta}_\tau(\boldsymbol{\beta}_j^{(k)})\right]^{-1}\left[n\boldsymbol{\Delta}_\tau(\boldsymbol{\beta}_j^{(k)})\boldsymbol{\beta}_j^{(k)} - \frac{\partial Q_2(\boldsymbol{\theta}; \boldsymbol{\Psi}^{(k)})}{\partial\boldsymbol{\beta}_j}\right],$$

with

$$\boldsymbol{\Delta}_\tau(\boldsymbol{\beta}_j^{(k)}) = \text{diag}\left\{\frac{p'_{\tau_j}(|\beta_{j1}^{(k)}|)}{|\beta_{j1}^{(k)}|}, \frac{p'_{\tau_j}(|\beta_{j2}^{(k)}|)}{|\beta_{j2}^{(k)}|}, \ldots, \frac{p'_{\tau_j}(|\beta_{jp}^{(k)}|)}{|\beta_{jp}^{(k)}|}\right\}.$$

5.2. Choice of the tuning parameters

The degree of penalty is controlled by tuning parameters. When using the method introduced in this article, we need to choose the tuning parameters. Various selection criteria, including cross-validation (CV), generalized cross-validation (GCV), Akaike information criterion (AIC) (Akaike, 1973) and Bayesian Information Criterion (BIC) (Schwarz, 1978), are often used for choosing tuning parameters. GCV has a non-negligible overfitting effect in the final model selection. H. Wang et al. (2007) suggested using BIC for the SCAD estimator in linear models and partially linear models and proved the consistency of the selection method, that is, the optimal parameter chosen by BIC can identify the true model with probability tending to 1. Considering the maximizer $\boldsymbol{\Psi}_n$ of the log-likelihood function (13), we use the estimator $\boldsymbol{\Psi}_n$ to calculate the mixing proportions in (17). The mixing proportions remain fixed throughout the tuning parameter selection process. For a given value of tuning parameter τ_j, let $(\hat{\boldsymbol{\beta}}_j, \hat{\sigma}_j, \hat{\lambda}_j)$ be the maximum regularized likelihood estimates of the parameters in the j-th component of the FMMeR model fixing the remaining elements of $\boldsymbol{\Psi}$ at $\boldsymbol{\Psi}_n$. Denote the likelihood-based deviance statistics, evaluated at $(\hat{\boldsymbol{\beta}}_j, \hat{\sigma}_j, \hat{\lambda}_j)$, corresponding to the j-th component of FMMeR model as

$$D_j(\hat{\boldsymbol{\beta}}_j, \hat{\sigma}_j, \hat{\lambda}_j) = \sum_{i=1}^{n} \omega_{ij}\left[\log SN(y_i \mid y_i, \hat{\mu}_{ij}, \hat{\sigma}_j^2, \hat{\lambda}_j) - \log SN(y_i \mid \boldsymbol{x}_i\hat{\boldsymbol{\beta}}_k, \hat{\mu}_{ij}, \hat{\sigma}_j^2, \hat{\lambda}_j)\right],$$

where $\hat{\mu}_{ij} = \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_j - \frac{\hat{\sigma}_j}{3}[m_0(\hat{\lambda}_j) + 2\sqrt{\frac{2}{\pi}}\delta(\hat{\lambda}_j)]$ and the weights ω_{ij} are given in (16). Then we define

$$\text{BIC}(\tau_j) = 2D_j(\hat{\boldsymbol{\beta}}_j, \hat{\sigma}_j, \hat{\lambda}_j) + N(\tau_j)\log(n_j), j = 1, \ldots, m,$$

where $N(\tau_j)$ is the number of nonzero elements of the vector $\hat{\boldsymbol{\beta}}_j$ and $n_j = \sum_{i=1}^{n} \omega_{ij}$. It is expected that the choice of τ_{jt} should be such that the tuning parameter for a zero coefficient is larger than that for a nonzero coefficient. Thus, we can simultaneously unbiasedly estimate the larger coefficient, and shrink the small coefficient towards zero. Hence, similar to Wu et al. (2013), we suggest

$$\tau_{jt} = \frac{\hat{\tau}_j}{|\beta_{jt}^{(0)}|}, \quad j = 1, 2, \ldots, m,\ t = 1, 2, \ldots, p,$$

where $\beta_{jt}^{(0)}$ is the MLE of β_{jt}. β_{jt}, τ_{jt} are the t-th component of $\boldsymbol{\beta}_j$ and τ_j, respectively. Tuning parameters are obtained via calculating

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

5.3. Determining the number of components

Determining the number of components of an FMR model is a challenge. In the above discussion, we assume that m is known and processing methods are either based on prior information or pre-analysis of data. A feasible method

implements reversible jump Markov chain Monte Carlo (RJMCMC) Richardson and Green (1997), since adding skewness even complicates matters, we did not pursue RJMCMC. Moreover, the component posterior probabilities evaluated in mixture modelling for Bayesian inference can be readily used as a soft clustering scheme. Alternatively, the log-likelihood maximum and two information-based criteria, AIC and BIC, can be used to select the number of components. Although some success has been shown using the model choice criteria, choosing the right number of components for a mixture model is known to be difficult.

To improve the efficiency for selecting the number of components in this article, a productive nonparametric clustering method via mode identification is applied, see J. Li et al. (2007). It deserves mentioning that this approach is robust in high dimensions and when clusters deviate substantially from Gaussian distributions. Specifically, a cluster is formed by those sample points that ascend to the same local maximum of the density function, and a pairwise separability measure for clusters is defined using the Ridgeline between the density bumps of two clusters. In this process, the Modal EM (MEM) algorithm and Ridgeline EM (REM) algorithm are used. Numerical results in Section 7 illustrated that this clustering procedure works well for determining the number of components in the FMMeR model.

6. Numerical experiments

In this section, we carry out simulation studies to investigate the finite sample performance of the proposed methodology. To be more specific, in Subsection 6.1, we conduct simulations to study the impact of the sample size on the estimation quality, and in Subsection 6.2, we investigate the quality of the performance for variable selection over different values of the skewness, and we compare the performance of the proposed FMMeR model and NMR model used in Khalili and Chen (2007) in Subsection 6.3.

6.1. Experiments 1

The experiment works to observe the impact of the sample size on the estimation quality. In addition, we compare the performance of different variable selection methods from a number of angles. We generated independently samples of size n from the following FMMeR model with two components,

$$\begin{cases} f(y_i \mid \boldsymbol{\Psi}) = \nu_1 SN(\mu_{i1}, \sigma_1^2, \lambda_1) + (1-\nu_1) SN(\mu_{i2}, \sigma_2^2, \lambda_2), \\ \text{Median}(y_{ij}) = \boldsymbol{x}_i^\top \boldsymbol{\beta}_j, \quad i = 1, 2, \ldots, n,\ j = 1, 2, \end{cases} \tag{19}$$

where μ_{i1} and μ_{i2} are defined by (9), and $\boldsymbol{\Psi} = (\nu_1, \boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \sigma_1, \sigma_2, \lambda_1, \lambda_2)^\top$. The components of the covariate $\boldsymbol{x}$ in the simulation are generated from a uniform distribution $U(-1, 1)$. The true values of parameters are set to be $\boldsymbol{\beta}_{1(0)} = (1, 0, 0, -1.5, 0)^\top$, $\boldsymbol{\beta}_{2(0)} = (-1, 0, 1, 0, 1.2)^\top$, $\sigma_{1(0)} = \sigma_{2(0)} = 2$. To test the sensitivity of the FMMeR model for positively or negatively skewed data, we set $\lambda_{1(0)} = 3$ and $\lambda_{2(0)} = -3$. A choice of mixing proportion $\nu_1 = 0.5$ and 0.35 is considered, and y is generated according to model (19). According to Karlis and Xekalaki (2003), a faster convergence rate can be achieved by setting the true value of the parameter to the initial value of the iteration. The performance of the estimators $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}$, $\hat{\lambda}$ and $\hat{\nu}$ will be assessed using the Mean Squared Error (MSE), defined as

$$\begin{aligned} \text{MSE}(\hat{\boldsymbol{\beta}}_j) &= \text{E}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_{j(0)})^\top (\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_{j(0)}), \\ \text{MSE}(\hat{\sigma}_j) &= \text{E}(\hat{\sigma}_j - \sigma_{j(0)})^2, \\ \text{MSE}(\hat{\lambda}_j) &= \text{E}(\hat{\lambda}_j - \lambda_{j(0)})^2, \\ \text{MSE}(\hat{\nu}_j) &= \text{E}(\hat{\nu}_j - \nu_{j(0)})^2. \end{aligned}$$

The average numbers of correctly (C) and incorrectly (IC) estimated zero coefficients and their standard deviation (SD) based on 500 repetitions are presented in Table 1. The results are presented in terms of mixture components 1 and 2. In addition, we report the MSEs and SD of scale, skewness and mixing proportion for $\nu_1 = 0.5$ across the repetitions in Table 2. Note that when the sample size n increases, as expected, the methods improve for a given penalty. The MSEs of estimators $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}$, $\hat{\lambda}$ and $\hat{\nu}$ tend to decrease by increasing the sample size, which illustrates the

Table 1. Three penalty functions are used for variable selection procedure.

			$\nu_1 = 0.35$			$\nu_1 = 0.5$		
Penalty	Com.	n	$C_{(SD)}$	$IC_{(SD)}$	$MSE(\hat{\boldsymbol{\beta}})_{(SD)}$	$C_{(SD)}$	$IC_{(SD)}$	$MSE(\hat{\boldsymbol{\beta}})_{(SD)}$
LASSO	Com. 1	200	$2.6500_{(0.0750)}$	$0.0650_{(0.1890)}$	$0.1677_{(1.2661)}$	$2.8250_{(0.0750)}$	$0.0550_{(0.1313)}$	$0.1556_{(1.0248)}$
		400	$2.8750_{(0.0500)}$	$0.0250_{(0.0991)}$	$0.0975_{(0.4506)}$	$2.9500_{(0.0500)}$	$0.0125_{(0.0770)}$	$0.0849_{(0.3854)}$
		800	$2.9500_{(0.0000)}$	$0.0025_{(0.0000)}$	$0.0550_{(0.1240)}$	$3.0000_{(0.0000)}$	$0.0000_{(0.0000)}$	$0.0406_{(0.0428)}$
	Com. 2	200	$1.7750_{(0.0500)}$	$0.0500_{(0.1206)}$	$0.1485_{(0.9960)}$	$1.7760_{(0.0500)}$	$0.0500_{(0.1204)}$	$0.1484_{(0.9980)}$
		400	$1.9775_{(0.0000)}$	$0.0150_{(0.0556)}$	$0.0622_{(0.2210)}$	$1.9975_{(0.0000)}$	$0.0025_{(0.0701)}$	$0.0833_{(0.3552)}$
		800	$2.0000_{(0.0000)}$	$0.0000_{(0.0000)}$	$0.0300_{(0.0728)}$	$2.0000_{(0.0000)}$	$0.0000_{(0.0000)}$	$0.0443_{(0.0462)}$
HARD	Com. 1	200	$2.7500_{(0.0500)}$	$0.0150_{(0.1210)}$	$0.1335_{(0.3055)}$	$2.8075_{(0.0500)}$	$0.0010_{(0.0500)}$	$0.1036_{(0.2840)}$
		400	$2.9975_{(0.0000)}$	$0.0000_{(0.0000)}$	$0.0712_{(0.0850)}$	$3.0000_{(0.0000)}$	$0.0000_{(0.0000)}$	$0.0511_{(0.0701)}$
		800	$3.0000_{(0.0000)}$	$0.0000_{(0.0000)}$	$0.0441_{(0.0495)}$	$3.0000_{(0.0000)}$	$0.0000_{(0.0000)}$	$0.0301_{(0.0320)}$
	Com. 2	200	$1.8750_{(0.0500)}$	$0.0200_{(0.0500)}$	$0.0623_{(0.2751)}$	$1.8750_{(0.0500)}$	$0.0100_{(0.0497)}$	$0.0954_{(0.2700)}$
		400	$2.0000_{(0.0000)}$	$0.0000_{(0.0000)}$	$0.0371_{(0.0621)}$	$2.0000_{(0.0000)}$	$0.0050_{(0.0000)}$	$0.0569_{(0.0602)}$
		800	$2.0000_{(0.0000)}$	$0.0000_{(0.0000)}$	$0.0202_{(0.0242)}$	$2.0000_{(0.0000)}$	$0.0000_{(0.0000)}$	$0.0296_{(0.0313)}$
SCAD	Com. 1	200	$2.7650_{(0.0500)}$	$0.0550_{(0.1313)}$	$0.1389_{(0.3026)}$	$2.7750_{(0.0500)}$	$0.0175_{(0.0689)}$	$0.0880_{(0.0901)}$
		400	$2.9240_{(0.0000)}$	$0.0150_{(0.0000)}$	$0.0712_{(0.0954)}$	$3.0000_{(0.0000)}$	$0.0025_{(0.0016)}$	$0.0496_{(0.0550)}$
		800	$3.0000_{(0.0000)}$	$0.0000_{(0.0000)}$	$0.0366_{(0.0424)}$	$3.0000_{(0.0000)}$	$0.0000_{(0.0000)}$	$0.0277_{(0.0297)}$
	Com. 2	200	$1.9550_{(0.0000)}$	$0.0025_{(0.0500)}$	$0.0717_{(0.2505)}$	$1.9550_{(0.0000)}$	$0.0100_{(0.0524)}$	$0.0927_{(0.1023)}$
		400	$2.0000_{(0.0000)}$	$0.0000_{(0.0000)}$	$0.0454_{(0.0524)}$	$2.0000_{(0.0000)}$	$0.0000_{(0.0000)}$	$0.0537_{(0.0689)}$
		800	$2.0000_{(0.0000)}$	$0.0000_{(0.0000)}$	$0.0197_{(0.0202)}$	$2.0000_{(0.0000)}$	$0.0000_{(0.0000)}$	$0.0315_{(0.0350)}$

Note: The first column indicates the penalty function used for variable selection method and the second column indicates the component.

Table 2. Simulation results of the parameters of scale, skewness and mixing proportion for $\nu_1 = 0.5$.

Parameters	Com.	n	$MSE(\hat{\sigma})_{(SD)}$	$MSE(\hat{\lambda})_{(SD)}$	$MSE(\hat{\nu})_{(SD)}$
LASSO	Com. 1	200	$0.0209_{(0.1656)}$	$0.0350_{(0.1202)}$	$0.0014_{(0.0045)}$
		400	$0.0098_{(0.0650)}$	$0.0149_{(0.0455)}$	$0.0006_{(0.0031)}$
		800	$0.0054_{(0.0075)}$	$0.0040_{(0.0069)}$	$0.0002_{(0.0014)}$
	Com. 2	200	$0.0221_{(0.1705)}$	$0.0342_{(0.1351)}$	$0.0014_{(0.0046)}$
		400	$0.0104_{(0.0655)}$	$0.0140_{(0.0459)}$	$0.0006_{(0.0031)}$
		800	$0.0060_{(0.0081)}$	$0.0042_{(0.0077)}$	$0.0003_{(0.0012)}$
HARD	Com. 1	200	$0.0081_{(0.0765)}$	$0.0120_{(0.0153)}$	$0.0012_{(0.0035)}$
		400	$0.0030_{(0.0089)}$	$0.0041_{(0.0058)}$	$0.0006_{(0.0016)}$
		800	$0.0015_{(0.0017)}$	$0.0027_{(0.0035)}$	$0.0003_{(0.0006)}$
	Com. 2	200	$0.0090_{(0.0790)}$	$0.0130_{(0.0147)}$	$0.0012_{(0.0031)}$
		400	$0.0035_{(0.0087)}$	$0.0049_{(0.0055)}$	$0.0006_{(0.0017)}$
		800	$0.0018_{(0.0021)}$	$0.0028_{(0.0031)}$	$0.0003_{(0.0008)}$
SCAD	Com. 1	200	$0.0087_{(0.0924)}$	$0.0113_{(0.0255)}$	$0.0013_{(0.0040)}$
		400	$0.0038_{(0.0209)}$	$0.0042_{(0.0097)}$	$0.0006_{(0.0029)}$
		800	$0.0017_{(0.0025)}$	$0.0030_{(0.0046)}$	$0.0003_{(0.0010)}$
	Com. 2	200	$0.0091_{(0.0889)}$	$0.0130_{(0.0270)}$	$0.0013_{(0.0037)}$
		400	$0.0034_{(0.0224)}$	$0.0052_{(0.0104)}$	$0.0007_{(0.0024)}$
		800	$0.0018_{(0.0020)}$	$0.0031_{(0.0049)}$	$0.0003_{(0.0009)}$

convergence property of the maximum penalized likelihood estimator of FMMeR model. For a given n, the performances of SCAD and HARD methods are similar for model complexity and better than LASSO method. When mixing proportion ν_1 reduces, and the sample size for component 1 decreases, all procedures for the component 1 of the FMMeR model become less satisfactory. Furthermore, the performances of component 1 and component 2 are similar for $\nu_1 = 0.5$, which indicates that FMMeR model is insensitive to positively or negatively skewed data.

6.2. Experiments 2

To investigate how the estimation quality is changed over different skewness, in this section, we set mixing proposition $\nu_1 = 0.5$ and the number of observations $n = 400$. Observations are generated in the same way as in Experiment 1. Table 3 shows C, IC, $MSE(\hat{\boldsymbol{\beta}})$ and their SD for different penalty function with $\lambda = -3, -1.5, -0.5, 0.5, 1.5, 3$ for 500 repetitions. Notice that the variable selection procedures perform similarly in all three cases for a given penalty function, but a larger SD is obtained by LASSO. When combined with the relevant conclusions of Experiment 1, the result indicates that the performance of the variable selection method is not affected by the choice of skewness of data.

Table 3. Varying skewness with $n = 400$ and $\nu_1 = 0.5$.

Penalty	Com.	λ	C	IC	MSE($\hat{\boldsymbol{\beta}}$)
LASSO	Com. 1	0.5	$2.9975_{(0.0000)}$	$0.0025_{(0.0557)}$	$0.0851_{(0.3558)}$
		1.5	$2.9550_{(0.0500)}$	$0.0175_{(0.0721)}$	$0.0902_{(0.3506)}$
		3	$2.9500_{(0.0000)}$	$0.0125_{(0.0770)}$	$0.0849_{(0.3854)}$
	Com. 2	−0.5	$1.9750_{(0.0500)}$	$0.0225_{(0.0725)}$	$0.0803_{(0.3724)}$
		−1.5	$1.9975_{(0.0000)}$	$0.0125_{(0.0750)}$	$0.0770_{(0.3550)}$
		−3	$1.9975_{(0.0000)}$	$0.0025_{(0.0701)}$	$0.0833_{(0.3552)}$
HARD	Com. 1	0.5	$3.0000_{(0.0000)}$	$0.0000_{(0.0015)}$	$0.0547_{(0.0684)}$
		1.5	$3.0000_{(0.0000)}$	$0.0000_{(0.0000)}$	$0.0676_{(0.0778)}$
		3	$3.0000_{(0.0000)}$	$0.0000_{(0.0000)}$	$0.0511_{(0.0701)}$
	Com. 2	−0.5	$2.0000_{(0.0000)}$	$0.0000_{(0.0050)}$	$0.0604_{(0.0599)}$
		−1.5	$2.0000_{(0.0000)}$	$0.0000_{(0.0000)}$	$0.0685_{(0.0619)}$
		−3	$2.0000_{(0.0000)}$	$0.0050_{(0.0000)}$	$0.0569_{(0.0602)}$
SCAD	Com. 1	0.5	$3.0000_{(0.0000)}$	$0.0000_{(0.0025)}$	$0.0536_{(0.0511)}$
		1.5	$3.0000_{(0.0000)}$	$0.0000_{(0.0000)}$	$0.0512_{(0.0504)}$
		3	$3.0000_{(0.0000)}$	$0.0025_{(0.0016)}$	$0.0496_{(0.0550)}$
	Com. 2	−0.5	$2.0000_{(0.0000)}$	$0.0000_{(0.0000)}$	$0.0509_{(0.0661)}$
		−1.5	$2.0000_{(0.0000)}$	$0.0000_{(0.0019)}$	$0.0632_{(0.0604)}$
		−3.	$2.0000_{(0.0000)}$	$0.0000_{(0.0000)}$	$0.0537_{(0.0689)}$

Table 4. Varying sample size n with $\lambda_1 = 3, \lambda_2 = -3$ and $\nu = 0.5$.

			FMMeR model		NMR model	
Penalty	Com.	n	$C_{(SD)}$	$IC_{(SD)}$	$C_{(SD)}$	$IC_{(SD)}$
LASSO	Com. 1	200	$2.8250_{(0.0750)}$	$0.0550_{(0.1313)}$	$2.8540_{(0.3550)}$	$0.0500_{(0.0250)}$
		400	$2.9500_{(0.0500)}$	$0.0125_{(0.0770)}$	$2.9200_{(0.0224)}$	$0.0250_{(0.0000)}$
	Com. 2	200	$1.7760_{(0.0500)}$	$0.0500_{(0.1204)}$	$1.7750_{(0.3395)}$	$0.0650_{(0.0300)}$
		400	$1.9975_{(0.0000)}$	$0.0025_{(0.0701)}$	$1.9240_{(0.0206)}$	$0.0100_{(0.0000)}$
HARD	Com. 1	200	$2.8075_{(0.0500)}$	$0.0010_{(0.0500)}$	$2.8200_{(0.1192)}$	$0.0025_{(0.0575)}$
		400	$3.0000_{(0.0000)}$	$0.0000_{(0.0000)}$	$3.0000_{(0.0000)}$	$0.0000_{(0.0000)}$
	Com. 2	200	$1.8750_{(0.0500)}$	$0.0100_{(0.0497)}$	$1.7550_{(0.0097)}$	$0.0150_{(0.0596)}$
		400	$2.0000_{(0.0000)}$	$0.0050_{(0.0000)}$	$1.9975_{(0.0000)}$	$0.0000_{(0.0000)}$
SCAD	Com. 1	200	$2.7750_{(0.0500)}$	$0.0175_{(0.0689)}$	$2.8065_{(0.1428)}$	$0.0155_{(0.0775)}$
		400	$3.0000_{(0.0000)}$	$0.0025_{(0.0016)}$	$2.9750_{(0.0042)}$	$0.0025_{(0.0000)}$
	Com. 2	200	$1.9550_{(0.0000)}$	$0.0100_{(0.0524)}$	$1.7500_{(0.1300)}$	$0.0200_{(0.0790)}$
		400	$2.0000_{(0.0000)}$	$0.0000_{(0.0000)}$	$1.9950_{(0.0000)}$	$0.0000_{(0.0000)}$

6.3. Experiments 3

To demonstrate the ability of the proposed variable selection method at selecting important variables, we compare the performance of the proposed FMMeR model and NMR model used in Khalili and Chen (2007) for a varying sample size $n = 200$, 400 and $\nu_1 = 0.5$. The data are generated exactly in the same way as in Experiment 1, and each of the two models is considered for the inference. The simulated results are reported in Table 4 based on 500 repetitions. From Table 4, it is clear that the performance of the variable selection procedure based on the FMMeR model is better than that based on the NMR model in some settings. This confirms that the FMMeR model clearly outperforms the NMR model at identifying important variables when there is skewness in the data. As expected, the MSEs indicate the convergence property of the maximum penalized likelihood estimator of FMMeR and NMR models.

7. A real-data example

FMR models have been used in the fields of biomedicine. To further demonstrate the ability of the proposed FMMeR model and variable selection method at identifying significant variables, we use a real-data example to illustrate the practical application of the proposed method of the FMMeR model in this section. The data set, analysed by Cook and Weisberg (1994), focused on the body mass index (BMI) for 102 male and 100 female athletes collected at Australian Institute of Sport. We are interested in the relationship between BMI and the 10 performance measures given as red cell count ($\boldsymbol{x}_1$), white cell count ($\boldsymbol{x}_2$), haematocrit ($\boldsymbol{x}_3$), haemoglobin ($\boldsymbol{x}_4$), plasma ferritin concentration ($\boldsymbol{x}_5$), sum of skin folds ($\boldsymbol{x}_6$), body fat percentage ($\boldsymbol{x}_7$), lean body mass ($\boldsymbol{x}_8$), height ($\boldsymbol{x}_9$) and weight ($\boldsymbol{x}_{10}$).

It can be seen from the histogram of the BMI in Figure 1 that the response is right-skewed, indicating the preference of the model with the skew-normal random errors. We determine the number of components via the method in Subsection 5.3. The clustering results are shown in Figure 2. At the level 3, 4 clusters are formed, as shown by different symbols in Figure 2(a). The 4 modes identified at level 3 are merged into 2 modes at level 4, as shown in

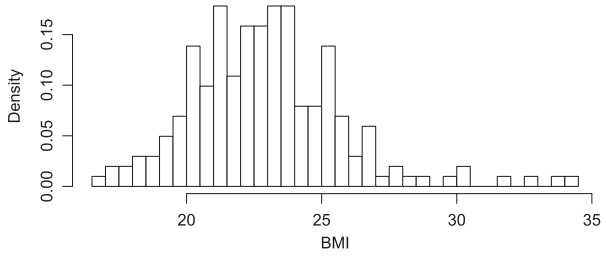

Figure 1. Histogram of the BMI.

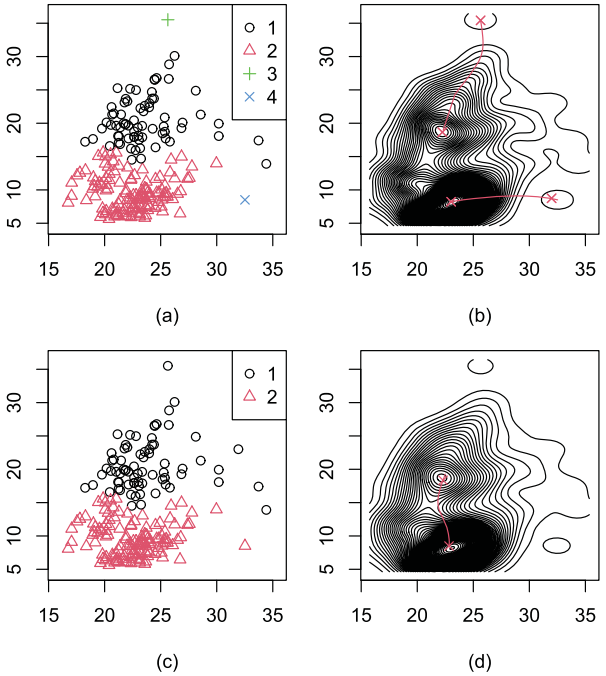

Figure 2. Clustering results for the BMI data obtained. (a) The 4 clusters at level 3. (b) The ascending paths from the modes at level 3 to those at level 4 and the contours of the density estimate at level 4. (c) The 2 clusters at level 4. (d) The ascending paths from the modes at level 4 to the next level and the contours of the density estimate at the next level.

Figure 2(b,d). Compared with level 4, two influential observations were excluded in cluster 1 and cluster 2 of level 3. Thus, it seems reasonable to use the following FMMeR model with two components to fit the BMI data,

$$\begin{cases} f(y_i \mid \boldsymbol{\Psi}) = \nu_1 SN(\mu_{i1}, \sigma_1^2, \lambda_1) + (1-\nu_1) SN(\mu_{i2}, \sigma_2^2, \lambda_2), \\ \text{Median}(y_{ij}) = \boldsymbol{x}_i^\top \boldsymbol{\beta}_j, \quad i = 1, 2, \ldots, 202, \ j = 1, 2, \end{cases} \tag{20}$$

where μ_{i1} and μ_{i2} are defined by (9), and $\boldsymbol{\Psi} = (\nu_1, \boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \sigma_1, \sigma_2, \lambda_1, \lambda_2)^\top$. $\boldsymbol{x}_i$ is a 10×1 vector consisting of all 10 potential variables. Three penalty functions are used to select significant variables.

We compare the variable selection results of the three models, including the proposed FMMeR model in this article, finite mixture of modal liner regression model and NMR model, where modal liner regression (MODLR) model was proposed by Yao and Li (2014). The results of variable selection for three models are given in Tables 5–7. In this data example, three variable selection procedures for a given model perform very similarly in terms of selecting significant variables. For FMMeR model and finite mixture of MODLR model, the same variables are removed for a given penalty function. NMR model, however, reserves more variables, resulting in a failure to select significant variables. Thus, the true structure of the model is not identified. When there is a situation of skewed data, the performances of HARD and SCAD are better than LASSO for identifying the authentic structure of the model. In FMMeR model, seven significant variables, including $\boldsymbol{x}_1$, $\boldsymbol{x}_4$, $\boldsymbol{x}_5$, $\boldsymbol{x}_7$, $\boldsymbol{x}_8$, $\boldsymbol{x}_9$, $\boldsymbol{x}_{10}$, are identified in component 1. Also seven $\boldsymbol{x}_4$, $\boldsymbol{x}_5$, $\boldsymbol{x}_6$, $\boldsymbol{x}_7$, $\boldsymbol{x}_8$, $\boldsymbol{x}_9$, $\boldsymbol{x}_{10}$ are contained in component 2. This indicates that these variables have a significant effect for BMI of athletes. We also find that there are some variables having different effects on parts one and two. For instance, red cell count ($\boldsymbol{x}_1$) and sum of skin folds ($\boldsymbol{x}_6$) are another factors affecting athletes' BMI in component

Table 5. Variable selection for BMI data set via FMMeR model.

	LASSO		HARD		SCAD	
Covariates	Com.1	Com.2	Com.1	Com.2	Com.1	Com.2
x_1			−0.7386		−0.7224	
x_2		0.0674				
x_3						
x_4	0.8752	0.7071	1.0808	0.7490	1.0726	0.7358
x_5	0.0032	0.0069	0.0027	0.0027	0.0034	0.0016
x_6		0.0501		0.0468		0.0424
x_7	0.2773	0.5750	0.7060	0.6165	0.6848	0.5997
x_8	0.2087	1.1808	0.6887	1.1974	0.6666	1.1324
x_9	−0.0446	−0.1054	−0.0668	−0.1067	−0.0656	−0.1003
x_{10}		−0.7357	−0.4254	−0.7508	−0.4060	−0.7007

Table 6. Variable selection for BMI data set via finite mixture of MODLR model.

	LASSO		HARD		SCAD	
Covariates	Com.1	Com.2	Com.1	Com.2	Com.1	Com.2
x_1			−0.7244		−0.7226	
x_2		0.0709				
x_3						
x_4	0.8790	0.6873	1.0554	0.7283	1.0496	0.7153
x_5	0.0038	0.0070	0.0033	0.0026	0.0032	0.0019
x_6		0.0486		0.0447		0.0410
x_7	0.2690	0.6033	0.7884	0.6480	0.7628	0.6284
x_8	0.2062	1.2133	0.7931	1.2285	0.7621	1.1663
x_9	−0.0454	−0.1079	−0.0714	−0.1089	−0.0687	−0.1027
x_{10}		−0.7595	−0.5188	−0.7734	−0.4929	−0.7256

Table 7. Variable selection for BMI data set via NMR model.

	LASSO		HARD		SCAD	
Covariates	Com.1	Com.2	Com.1	Com.2	Com.1	Com.2
x_1	−0.9520		−0.9582			
x_2	−0.0457	0.0824	−0.0451	0.0824		
x_3	0.0754		0.0759			
x_4	0.9064	0.6228	0.9067	0.6227	0.0794	0.0693
x_5	0.0049	0.0081	0.0049	0.0081	0.0003	0.0003
x_6	−0.019	0.0438	−0.0189	0.0437		0.0038
x_7	0.8690	0.6757	0.8703	0.6761	0.0244	0.0634
x_8	0.7615	1.2918	0.7639	1.2921	0.0195	0.1164
x_9	−0.0729	−0.1134	−0.0731	−0.1134	0.0068	0.0008
x_{10}	−0.4855	−0.8169	−0.4876	−0.8172	−0.0137	−0.0884

1 and component 2, respectively. Furthermore, $\boldsymbol{x}_4$, $\boldsymbol{x}_5$, $\boldsymbol{x}_7$ and $\boldsymbol{x}_8$ are helpful for achieving a high BMI in two components. In addition, the performance of the variable selection procedure via the FMMeR model is different from that of the variable selection procedure via the NMR model.

8. Conclusions remarks and future works

In this paper, by utilizing the skew-normal distribution as a component density to overcome the potential inappropriateness of normal mixtures in some context, we have developed a novel finite mixture of the median regression (FMMeR) model to explore asymmetrical data that arise from several subpopulations. Thanks to the stochastic representation for the skew-normal distribution, we have constructed a hierarchical representation of the finite skew-normal mixtures to address computational barriers of the parameter estimation and variable selection when fitting the FMMeR model. In addition, in order to determine the number of components, we applied a clustering method via mode identification proposed by J. Li et al. (2007) and a good performance is shown. Numerical results from simulation studies and a real-data example illustrated that the proposed FMMeR model methodology performs well in general, even when the data exhibit symmetrical behaviour.

It is worth noting that we only considered the procedures of parameter estimation and variable selection for the FMMeR model based on a mixture of the skew-normal distributions. Meanwhile, the scenario of $p > n$ has not been considered in this paper. A natural extension of the proposed methodology is to consider other skewed distributions, such as the skew-t and skew-Laplace distributions, and high-dimensional settings. In addition, another research direction is to model the mixing proportions $\boldsymbol{\nu}$, which extends the proposed model to the framework of

mixture of experts models. Finally, it will also be of interest to consider Bayesian variable selection, semi-parametric and nonparametric methods for the FMMeR model, which are currently under investigation and will be reported elsewhere.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work is partially supported by the National Natural Science Foundation of China [grant number 11861041], the Natural Science Research Foundation of Kunming University of Science and Technology [grant number KKSY201907003].

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *International Symposium on Information Theory*, *1*, 610–624. https://doi.org/10.1007/978-1-4612-1694-0_15

Atienza, N., Garcia-Heras, J., & Muñoz-Pichardo, J. (2006). A new condition for identifiability of finite mixture distributions. *Metrika*, *63*(2), 215–221. https://doi.org/10.1007/s00184-005-0013-z

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, *12*(2), 171–178. http://www.jstor.org/stable/4615982

Azzalini, A., & Capitanio, A. (2013). *The skew-normal and related families*. Cambridge University Press.

Chen, J. (2017). Consistency of the MLE under mixture models. *Statistical Science*, *32*(1), 47–63. https://doi.org/10.1214/16-sts578

Chen, J., Li, P., & Liu, G. (2020). Homogeneity testing under finite location-scale mixtures. *Canadian Journal of Statistics*, *48*(4), 670–684. https://doi.org/10.1002/cjs.11557

Chen, J., & Tan, X. (2009). Inference for multivariate normal mixtures. *Journal of Multivariate Analysis*, *100*(7), 1367–1383. https://doi.org/10.1016/j.jmva.2008.12.005

Cook, R.-D., & Weisberg, S. (1994). *An introduction to regression graphics*. John Wiley and Sons.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360. https://doi.org/10.1198/016214501753382273

Goldfeld, S., & Quandt, R. (1973). A Markov model for switching regressions. *Journal of Econometrics*, *1*(1), 3–15. https://doi.org/10.1016/0304-4076(73)90002-X

He, M., & Chen, J. (2022a). Consistency of the MLE under a two-parameter gamma mixture model with a structural shape parameter. *Metrika*. https://doi.org/10.1007/s00184-021-00856-9

He, M., & Chen, J. (2022b). Strong consistency of the MLE under two-parameter gamma mixture models with a structural scale parameter. *Advances in Data Analysis and Classification*, *16*(1), 125–154. https://doi.org/10.1007/s11634-021-00472-5

Hu, D., Gu, Y., & Zhao, W. (2019). Bayesian variable selection for median regression. *Chinese Journal of Applied Probability and Statistics*, *35*(6), 594–610.

Karlis, D., & Xekalaki, E. (2003). Choosing initial values for the EM algorithm for finite mixtures. *Computational Statistics & Data Analysis*, *41*(3–4), 577–590. https://doi.org/10.1016/S0167-9473(02)00177-9

Khalili, A., & Chen, J. (2007). Variable selection in finite mixture of regression models. *Journal of the American Statistical Association*, *102*(479), 1025–1038. https://doi.org/10.1198/016214507000000590

Kottas, A., & Gelfand, A. (2001). Bayesian semiparametric median regression modeling. *Journal of the American Statistical Association*, *96*(456), 1458–1468. https://doi.org/10.1198/016214501753382363

Li, H., Wu, L., & Ma, T. (2017). Variable selection in joint location, scale and skewness models of the skew-normal distribution. *Journal of Systems Science and Complexity*, *30*(3), 694–709. https://doi.org/10.1007/S11424-016-5193-2

Li, H., Wu, L., & Yi, J. (2016). A skew-normal mixture of joint location, scale and skewness models. *Applied Mathematics-A Journal of Chinese Universities*, *31*(3), 283–295. https://doi.org/10.1007/S11766-016-3367-2

Li, J., Ray, S., & Lindsay, B.-G. (2007). A nonparametric statistical approach to clustering via mode identification. *Journal of Machine Learning Research*, *8*(8), 1687–1723.

Lin, T.-I., Lee, J., & Yen, S. (2007). Finite mixture modelling using the skew normal distribution. *Statistica Sinica*, *17*(3), 909–927. http://www.jstor.org/stable/24307705

Liu, M., & Lin, T.-I. (2014). A skew-normal mixture regression model. *Educational and Psychological Measurement*, *74*(1), 139–162. https://doi.org/10.1177/0013164413498603

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

Otiniano, C. E. G., Rathie, P. N., & Ozelim, L. C. S. M. (2015). On the identifiability of finite mixture of skew-normal and skew-t distributions. *Statistics & Probability Letters*, *106*, 103–108. https://doi.org/10.1016/j.spl.2015.07.015

Richardson, S., & Green, P. (1997). On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *59*(4), 731–792. https://doi.org/10.1111/1467-9868.00095

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/AOS/1176344136

Tang, A., & Tang, N. (2015). Semiparametric Bayesian inference on skew-normal joint modeling of multivariate longitudinal and survival data. *Statistics in Medicine*, *34*(5), 824–843. https://doi.org/10.1002/SIM.6373

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *58*(1), 267–288. https://doi.org/10.1111/J.2517-6161.1996.TB02080.X
Titterington, D., Smith, A., & Makov, U. (1985). *Statistical analysis of finite mixture distributions*. John Wiley and Sons
Wang, H., Li, R., & Tsai, C. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, *94*(3), 553–568. https://doi.org/10.1093/BIOMET/ASM053
Wang, P., Puterman, M., Cockburn, I., & Le, N. (1996). Mixed Poisson regression models with covariate dependent rates. *Biometrics*, *52*(2), 381–400. https://doi.org/10.2307/2532881
Wu, L. (2014). Variable selection in joint location and scale models of the skew-t-normal distribution. *Communications in Statistics. Simulation and Computation*, *43*(3), 615–630. https://doi.org/10.1080/03610918.2012.712182
Wu, L., Li, S., & Tao, Y. (2020). Estimation and variable selection for mixture of joint mean and variance models. *Communications in Statistics-Theory and Methods*, *50*(24), 6081–6098. https://doi.org/10.1080/03610926.2020.1738493
Wu, L., Zhang, Z., & Xu, D. (2013). Variable selection in joint location and scale models of the skew-normal distribution. *Journal of Statistical Computation and Simulation*, *83*(7), 1266–1278. https://doi.org/10.1080/00949655.2012.657198
Yao, W., & Li, L. (2014). A new regression model: Modal linear regression. *Scandinavian Journal of Statistics*, *41*(3), 656–671. https://doi.org/10.1111/SJOS.12054
Yin, J., Wu, L., & Dai, L. (2020). Variable selection in finite mixture of regression models using the skew-normal distribution. *Journal of Applied Statistics*, *47*(16), 2941–2960. https://doi.org/10.1080/02664763.2019.1709051
Yin, J., Wu, L., Lu, H., & Dai, L. (2020). New estimation in mixture of experts models using the Pearson type VII distribution. *Communications in Statistics. Simulation and Computation*, *49*(2), 472–483. https://doi.org/10.1080/03610918.2018.1485943
Zhou, X., & Liu, G. (2016). LAD-Lasso variable selection for doubly censored median regression models. *Communications in Statistics. Theory and Methods*, *45*(12), 3658–3667. https://doi.org/10.1080/03610926.2014.904357

Appendices

Appendix 1. Regularity conditions and proofs

Regularity conditions R_1–R_4 on the joint distribution of $h = (\boldsymbol{x}, Y)$ are needed for proving the asymptotic properties of the proposed method. Let $f(h \mid \boldsymbol{\Psi})$ be the joint density function of h with the parameter space $\boldsymbol{\Psi} \in \boldsymbol{\Omega}$. We write $\boldsymbol{\Psi} = (\psi_1, \psi_2, \ldots, \psi_s)$ and s is the total number of parameters in the FMMeR model. The regularity conditions are as follows.

R_1: The density $f(h \mid \boldsymbol{\Psi})$ has common support in h for all $\boldsymbol{\Psi} \in \boldsymbol{\Omega}$, and $f(h \mid \boldsymbol{\Psi})$ is identifiable in $\boldsymbol{\Psi}$ up to a permutation of the components of the mixture.

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

R_3: For each $\boldsymbol{\Psi}_0 \in \boldsymbol{\Psi}$ and $t, l, g = 1, 2, \ldots, s$, there exist functions $B_1(h)$ and $B_2(h)$ (possibly depending on $\boldsymbol{\Psi}_0$) such that for $\boldsymbol{\Psi}$ in a neighbourhood of $N(\boldsymbol{\Psi}_0)$,

$$\left|\frac{\partial f(h \mid \boldsymbol{\Psi})}{\partial \psi_t}\right| \le B_1(h), \quad \left|\frac{\partial^2 f(h \mid \boldsymbol{\Psi})}{\partial \psi_t \partial \psi_l}\right| \le B_1(h), \quad \left|\frac{\partial^3 \log f(h \mid \boldsymbol{\Psi})}{\partial \psi_t \partial \psi_l \partial \psi_g}\right| \le B_2(h),$$

where $\int B_1(h)\, \mathrm{d}h < \infty$ and $\int B_2(h) f(h \mid \boldsymbol{\Psi})\, \mathrm{d}h < \infty$.

R_4: The Fisher information matrix,

$$I(\boldsymbol{\Psi}) = E\left\{\left[\frac{\partial}{\partial \boldsymbol{\Psi}} \log f(h \mid \boldsymbol{\Psi})\right]\left[\frac{\partial}{\partial \boldsymbol{\Psi}} \log f(h \mid \boldsymbol{\Psi})\right]^{\top}\right\},$$

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

Proof of Theorem 4.1: Let $\xi_n = n^{-1/2}(1 + a_n)$. We just have to specify that for any given $\varepsilon > 0$, there exists a large constant C such that

$$\lim_{n\to\infty} P\left\{\sup_{\|\boldsymbol{u}\|=C} L(\boldsymbol{\Psi}_0 + \xi_n \boldsymbol{u}) < L(\boldsymbol{\Psi}_0)\right\} \ge 1 - \varepsilon. \tag{A1}$$

This indicates that for sufficiently large n, with large probability namely $1 - \varepsilon$, there is a localmaximum in the ball $\{\boldsymbol{\Psi}_0 + \xi_n \boldsymbol{u} : \|\boldsymbol{u}\| \le C\}$. This localmaximizer, say $\hat{\boldsymbol{\Psi}}$, satisfies $\|\hat{\boldsymbol{\Psi}} - \boldsymbol{\Psi}_0\| = O_p(\xi_n)$.

Let $\zeta_n(\boldsymbol{u}) = L(\boldsymbol{\Psi}_0 + \xi_n \boldsymbol{u}) - L(\boldsymbol{\Psi}_0)$. Using $p_{\tau_j}(0) = 0$ and the definition of $L(\cdot)$, we have

$$\begin{aligned}
\zeta_n(\boldsymbol{u}) &= [\ell(\boldsymbol{\Psi}_0 + \xi_n \boldsymbol{u}) - \ell(\boldsymbol{\Psi}_0)] - [p(\boldsymbol{\Psi}_0 + \xi_n \boldsymbol{u}) - p(\boldsymbol{\Psi}_0)] \\
&\le [\ell(\boldsymbol{\Psi}_0 + \xi_n \boldsymbol{u}) - \ell(\boldsymbol{\Psi}_0)] - [p(\boldsymbol{\Psi}_{01} + \xi_n \boldsymbol{u}_{\boldsymbol{I}}) - p(\boldsymbol{\Psi}_{01})] \\
&\le \ell(\boldsymbol{\Psi}_0 + \xi_n \boldsymbol{u}) - \ell(\boldsymbol{\Psi}_0) - n\sum_{j=1}^{m} \nu_j \sum_{t=1}^{d_j} [p_{\tau_j}(\beta_{jt}^0 + \xi_n \boldsymbol{u}_{\boldsymbol{I}}) - p_{\tau_j}(\beta_{jt}^0)],
\end{aligned}$$

where d_j is the number of nonzero elements of the vector β_j^0. $\boldsymbol{\Psi}_{01}$ is the parameter vector with zero regression coefficients removed and $\boldsymbol{u}_{\boldsymbol{I}}$ is a subvector of $\boldsymbol{u}$ with corresponding components. By Taylor's expansion and the triangular inequality

$$\zeta_n(\boldsymbol{u}) \le \xi_n \{\ell(\boldsymbol{\Psi}_0')\}^{\top} \boldsymbol{u} - \frac{1}{2} \boldsymbol{u}^{\top} I(\boldsymbol{\Psi}_0) \boldsymbol{u} n \xi_n^2 \{1 + o_p(1)\}$$

$$- \sum_{j=1}^{m} \nu_j \sum_{t=1}^{d_j} [n\xi_n p'_{\tau_j}(\beta_{jt}^0)\boldsymbol{u_I} + n\xi_n^2 p''_{\tau_j}\boldsymbol{u_I}^2\{1 + o(1)\}]$$

$$= q_1 + q_2 + q_3. \tag{A2}$$

Regularity conditions imply that $n^{-1/2}\ell'(\boldsymbol{\Psi}_0) = O_p(1)$ and Fisher informationmatrix $I(\boldsymbol{\Psi}_0)$ is positive definite. Thus, q_1 is of the order $O_p(n^{1/2}\xi_n) = O_p(n\xi_n^2)$. By choosing a sufficiently large C, q_1 is controlled uniformly by q_2 in $\|\boldsymbol{u}\| = C$. Note that the q_3 is bounded by

$$\sum_{j=1}^{m} \nu_j\{\sqrt{d}n\xi_n a_n\|\boldsymbol{u}\| + n\xi_n^2 b_n\|\boldsymbol{u}\|^2\} = \sqrt{d}n\xi_n a_n\|\boldsymbol{u}\| + n\xi_n^2 b_n\|\boldsymbol{u}\|^2,$$

where $d = \max_j d_j$. By condition C_1 for the penalty functions, $b_n = o(1)$, this is also dominated by the q_2. Hence, by choosing a sufficiently large C, (A1) holds. This completes the proof. ■

Proof of Theorem 4.2: To prove part (a), consider the partition $\boldsymbol{\Psi} = (\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2)$ for any $\boldsymbol{\Psi}$ in the neighbourhood $\|\boldsymbol{\Psi} - \boldsymbol{\Psi}_0\| = O(n^{-1/2})$. By the definition of $L(\cdot)$, we obtain

$$L(\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2) - L(\boldsymbol{\Psi}_1, \mathbf{0}) = [\ell(\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2) - \ell(\boldsymbol{\Psi}_1, \mathbf{0})] - [p(\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2) - p(\boldsymbol{\Psi}_1, \mathbf{0})].$$

By the mean value theorem,

$$\ell(\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2) - \ell(\boldsymbol{\Psi}_1, \mathbf{0}) = \left[\frac{\partial \ell(\boldsymbol{\Psi}_1, \boldsymbol{\eta})}{\partial \boldsymbol{\Psi}_2}\right]^{\top} \boldsymbol{\Psi}_2 \tag{A3}$$

with $\|\boldsymbol{\eta}\| \le \|\boldsymbol{\Psi}_2\| = O(n^{-1/2})$. Furthermore, by using regularity condition R_3 and the mean value theorem, we have

$$\begin{aligned}
\left\| \frac{\partial \ell(\boldsymbol{\Psi}_1, \boldsymbol{\eta})}{\partial \boldsymbol{\Psi}_2} - \frac{\partial \ell(\boldsymbol{\Psi}_{01}, \mathbf{0})}{\partial \boldsymbol{\Psi}_2} \right\| &\le \left\| \frac{\partial \ell(\boldsymbol{\Psi}_1, \boldsymbol{\eta})}{\partial \boldsymbol{\Psi}_2} - \frac{\partial \ell(\boldsymbol{\Psi}_1, \mathbf{0})}{\partial \boldsymbol{\Psi}_2} \right\| + \left\| \frac{\partial \ell(\boldsymbol{\Psi}_1, \mathbf{0})}{\partial \boldsymbol{\Psi}_2} - \frac{\partial \ell(\boldsymbol{\Psi}_{01}, \mathbf{0})}{\partial \boldsymbol{\Psi}_2} \right\| \\
&\le \left[\sum_{i=1}^{n} B_1(h_i)\right] \|\boldsymbol{\eta}\| + \left[\sum_{i=1}^{n} B_1(h_i)\right] \|\boldsymbol{\Psi}_1 - \boldsymbol{\Psi}_{01}\| \\
&= \{\|\boldsymbol{\eta}\| + \|\boldsymbol{\Psi}_1 - \boldsymbol{\Psi}_{01}\|\} O_p(n) = O_p(n^{1/2}).
\end{aligned}$$

By the regularity conditions R_1–R_4, $\partial \ell(\boldsymbol{\Psi}_{01}, \mathbf{0})/\partial \boldsymbol{\Psi}_2 = O_p(n^{1/2})$. Thus, $\partial \ell(\boldsymbol{\Psi}_1, \boldsymbol{\eta})/\partial \boldsymbol{\Psi}_2 = O_p(n^{1/2})$. Applying these order assessments to (A3), we obtain

$$\ell(\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2) - \ell(\boldsymbol{\Psi}_1, \mathbf{0}) = O_p(n^{1/2}) \sum_{j=1}^{m} \sum_{t=d_j+1}^{p} |\beta_{jt}|,$$

for large n. On the other hand,

$$p(\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2) - p(\boldsymbol{\Psi}_1, \mathbf{0}) = n \sum_{j=1}^{m} \sum_{t=d_j+1}^{p} \nu_j p_{\tau_j}(\beta_{jt}).$$

Thus,

$$L(\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2) - L(\boldsymbol{\Psi}_1, \mathbf{0}) = \sum_{j=1}^{m} \sum_{t=d_j+1}^{p} \{|\beta_{jt}| O_p(n^{1/2}) - n\nu_j p_{\tau_j}(\beta_{jt})\}.$$

In a shrinking neighbourhood of 0, $|\beta_{jt}| O_p(n^{1/2}) < n\nu_j p_{\tau_j}(\beta_{jt})$ in probability by condition C_2. This completes the proof of part (a).

To prove sparsity in part (b(1)), we consider the partition $\boldsymbol{\Psi} = L(\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2)$. Let $(\widehat{\boldsymbol{\Psi}}_1, \mathbf{0})$ be the maximizer of the penalized log-likelihood function $L(\boldsymbol{\Psi}, \mathbf{0})$, which is considered as a function of $\boldsymbol{\Psi}_1$. It suffices to show that in the neighbourhood $\|\boldsymbol{\Psi} - \boldsymbol{\Psi}_0\| = O_p\{n^{-1/2}\}$, $L(\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2) - L(\widehat{\boldsymbol{\Psi}}_1, \mathbf{0}) < 0$ with probability tending to 1 as $n \to \infty$. By the result in part (a), we obtain that

$$\begin{aligned}
L(\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2) - L(\widehat{\boldsymbol{\Psi}}_1, \mathbf{0}) &= [L(\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2) - L(\boldsymbol{\Psi}_1, \mathbf{0})] + [L(\boldsymbol{\Psi}_1, \mathbf{0}) - L(\widehat{\boldsymbol{\Psi}}_1, \mathbf{0})] \\
&\le [L(\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2) - L(\boldsymbol{\Psi}_1, \mathbf{0})] < 0.
\end{aligned}$$

To prove asymptotic normality in part (b(2)), we consider $L(\boldsymbol{\Psi}, \mathbf{0})$ as a function of $\boldsymbol{\Psi}_1$. Using the same argument as in Theorem 4.1, there exists a $\sqrt{n}$-consistent local maximizer of this function, denoted by $\widehat{\boldsymbol{\Psi}}_1$, that satisfies

$$\frac{\partial L(\widehat{\boldsymbol{\Psi}})}{\partial \boldsymbol{\Psi}_1} = \left\{ \frac{\partial \ell(\boldsymbol{\Psi})}{\partial \boldsymbol{\Psi}_1} - \frac{\partial p(\boldsymbol{\Psi})}{\partial \boldsymbol{\Psi}_1} \right\}_{\widehat{\boldsymbol{\Psi}} = (\widehat{\boldsymbol{\Psi}}_1, \mathbf{0})} = \mathbf{0}.$$

By substituting the first-order Taylor's expansions of $\partial \ell(\boldsymbol{\Psi})/\partial \boldsymbol{\Psi}_1$ and $\partial p(\boldsymbol{\Psi})/\partial \boldsymbol{\Psi}_1$ into the above expression, we have

$$\left\{ \frac{\partial^2 \ell(\boldsymbol{\Psi}_{01})}{\partial \boldsymbol{\Psi}_1 \partial \boldsymbol{\Psi}_1^{\top}} - p''(\boldsymbol{\Psi}_{01}) + o_p(n) \right\} (\widehat{\boldsymbol{\Psi}}_1 - \boldsymbol{\Psi}_{01}) = \frac{\partial \ell(\boldsymbol{\Psi}_{01})}{\partial \boldsymbol{\Psi}_1} - p'(\boldsymbol{\Psi}_{01}).$$

On the other hand, under the regularity conditions, we obtain

$$\frac{\partial^2 \ell(\boldsymbol{\Psi}_{01})}{\partial \boldsymbol{\Psi}_1 \partial \boldsymbol{\Psi}_1^\top} = I_1(\boldsymbol{\Psi}_{01}) + o_p(1),$$

and

$$\frac{1}{\sqrt{n}} \frac{\partial \ell(\boldsymbol{\Psi}_{01})}{\partial \boldsymbol{\Psi}_1} \xrightarrow{d} N(\mathbf{0}, I_1(\boldsymbol{\Psi}_{01})).$$

Using the foregoing facts and Slutsky's theorem, we have

$$\sqrt{n}\left\{\left[I_1(\boldsymbol{\Psi}_{01}) - \frac{p''(\boldsymbol{\Psi}_{01})}{n}\right](\widehat{\boldsymbol{\Psi}}_1 - \boldsymbol{\Psi}_{01}) + \frac{p'(\boldsymbol{\Psi}_{01})}{n}\right\} \xrightarrow{d} N(\mathbf{0}, I_1(\boldsymbol{\Psi}_{01})),$$

which is the result in part (b(2)). ■

Appendix 2. Some technical derivations

In (5.9), the score function of j-th component is expressed as

$$\begin{aligned}
S(\boldsymbol{\beta}_j) &= -\sum_{i=1}^n \omega_{ij}^{(k)} \frac{(1+\lambda_j^2)}{\sigma_j^2}\left(e_{ij}\boldsymbol{E}_1 - \sigma_j r_{1i}^{(k)} \delta(\lambda_j)\boldsymbol{E}_1\right),\\
S(\sigma_j) &= -\frac{1}{\sigma_j}\sum_{i=1}^n \omega_{ij}^{(k)} + \frac{(1+\lambda_j^2)}{\sigma_j^2}\sum_{i=1}^n \omega_{ij}^{(k)}\left[\frac{e_{ij}^2}{\sigma_j} - e_{ij} r_{1i}^{(k)}\delta(\lambda_j)\right.\\
&\quad \left. - e_{ij}E_2 + \sigma_j r_{1i}^{(k)}\delta(\lambda_j)E_2\right],\\
S(\lambda_j) &= \sum_{i=1}^n \omega_{ij}^{(k)} \frac{\delta^2(\lambda_j)}{\lambda_j} - \sum_{i=1}^n \omega_{ij}^{(k)}\left[\frac{e_{ij}^2\lambda_j}{\sigma_j^2} + \frac{e_{ij}(1+\lambda_j^2)}{\sigma_j^2}E_3\right.\\
&\quad \left. - \frac{r_{1i}^{(k)}}{\sigma_j}\left(\frac{e_{ij}\delta(\lambda_j)}{\lambda_j} + 2e_{ij}\lambda_j\delta(\lambda_j) + \frac{\lambda_j^2}{\delta(\lambda_j)}E_3\right) + \lambda_j r_{2i}^{(k)}\right].
\end{aligned}$$

$H(\boldsymbol{\theta})$ is defined as

$$H(\boldsymbol{\theta}) = \frac{\partial^2 Q_2(\boldsymbol{\theta};\boldsymbol{\Psi}^{(k)})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} = \begin{bmatrix} \frac{\partial^2 Q_2(\boldsymbol{\theta};\boldsymbol{\Psi}^{(k)})}{\partial\boldsymbol{\beta}_j\partial\boldsymbol{\beta}_j^\top} & \frac{\partial^2 Q_2(\boldsymbol{\theta};\boldsymbol{\Psi}^{(k)})}{\partial\boldsymbol{\beta}_j\partial\sigma_j} & \frac{\partial^2 Q_2(\boldsymbol{\theta};\boldsymbol{\Psi}^{(k)})}{\partial\boldsymbol{\beta}_j\partial\lambda_j} \\ * & \frac{\partial^2 Q_2(\boldsymbol{\theta};\boldsymbol{\Psi}^{(k)})}{\partial\sigma_j\partial\sigma_j} & \frac{\partial^2 Q_2(\boldsymbol{\theta};\boldsymbol{\Psi}^{(k)})}{\partial\sigma_j\partial\lambda_j} \\ * & * & \frac{\partial^2 Q_2(\boldsymbol{\theta};\boldsymbol{\Psi}^{(k)})}{\partial\lambda_j\partial\lambda_j} \end{bmatrix},$$

where

$$\begin{aligned}
\frac{\partial^2 Q_2(\boldsymbol{\theta};\boldsymbol{\Psi}^{(k)})}{\partial\boldsymbol{\beta}_j\partial\boldsymbol{\beta}_j^\top} &= \sum_{i=1}^n \omega_{ij}^{(k)} \frac{(1+\lambda_j^2)}{\sigma_j^2}\boldsymbol{E}_1\boldsymbol{E}_1^\top,\\
\frac{\partial^2 Q_2(\boldsymbol{\theta};\boldsymbol{\Psi}^{(k)})}{\partial\boldsymbol{\beta}_j\partial\sigma_j} &= \sum_{i=1}^n \omega_{ij}^{(k)} \frac{(1+\lambda_j^2)}{\sigma_j^2}\left[(\delta(\lambda_j) + 2e_{ij} - 2\sigma_j r_{1i}^{(k)}\delta(\lambda_j)\right]\boldsymbol{E}_1,\\
\frac{\partial^2 Q_2(\boldsymbol{\theta};\boldsymbol{\Psi}^{(k)})}{\partial\boldsymbol{\beta}_j\partial\lambda_j} &= -\sum_{i=1}^n \omega_{ij}^{(k)}\left[\frac{2\lambda_j e_{ij} + (1+\lambda_j^2)E_3}{\sigma_j^2} + \frac{r_{1i}^{(k)}}{\sigma_j}\left[2\lambda_j\delta(\lambda_j) + \frac{\delta(\lambda_j)}{\lambda_j}\right]\boldsymbol{E}_1\right],\\
\frac{\partial^2 Q_2(\boldsymbol{\theta};\boldsymbol{\Psi}^{(k)})}{\partial\sigma_j\partial\sigma_j} &= \frac{1}{\sigma_j^2}\sum_{i=1}^n \omega_{ij}^{(k)} + \sum_{i=1}^n \omega_{ij}^{(k)} \frac{(1+\lambda_j^2)}{\sigma_j^2}\left[\frac{2e_{ij}(2E_2 + r_{1i}^{(k)}\delta(\lambda_j))}{\sigma_j} - \frac{3e_{ij}^2}{\sigma_j^2}\right.\\
&\quad \left. - 2r_{1i}^{(k)}\delta(\lambda_j)E_2 - e_{ij}E_{22} + \sigma_j\delta(\lambda_j)r_{1i}^{(k)}E_{22} - E_2^2\right],\\
\frac{\partial^2 Q_2(\boldsymbol{\theta};\boldsymbol{\Psi}^{(k)})}{\partial\sigma_j\partial\lambda_j} &= \sum_{i=1}^n \omega_{ij}^{(k)} \frac{2\lambda_j e_{ij}}{\sigma_j^2}\left[\frac{e_{ij}}{\sigma_j} - r_{1i}^{(k)}\delta(\lambda_j) + \frac{\sigma_j r_{1i}^{(k)}\delta(\lambda_j)E_2}{e_{ij}} - \frac{r_{1i}^{(k)}\delta(\lambda_j)}{2\lambda_j^2} - E_2\right]
\end{aligned}$$

$$-\sum_{i=1}^{n}\omega_{ij}^{(k)}\frac{(1+\lambda_j^2)}{\sigma_j^2}\left[r_{1i}^{(k)}\delta(\lambda_j)+\frac{e_{ij}E_{23}}{E_3}-\frac{2e_{ij}E_3}{\sigma_j}+E_2\right],$$

$$\frac{\partial^2 Q_2(\boldsymbol{\theta};\boldsymbol{\Psi}^{(k)})}{\partial\lambda_j\partial\lambda_j}=\sum_{i=1}^{n}\omega_{ij}^{(k)}\frac{1-\lambda_j^2}{(1+\lambda_j^2)^2}-\sum_{i=1}^{n}\omega_{ij}^{(k)}\left[\frac{e_{ij}}{\sigma_j^2}(e_{ij}+4\lambda_jE_3)+\frac{(1+\lambda_j^2)}{\sigma_j^2}(E_3^2+e_{ij}E_{33})\right.$$

$$\left.-\frac{r_{1i}^{(k)}}{\sigma_j}\left(\frac{e_{ij}\delta(\lambda_j)}{1+\lambda_j^2}+2\delta(\lambda_j)(e_{ij}+\lambda_jE_3)+\sqrt{1+\lambda_j^2}(\lambda_jE_{33}+2E_3)\right)+r_{2i}^{(k)}\right],$$

with $e_{ij}=y_i-\boldsymbol{x}_i^\top\boldsymbol{\beta}_j-\frac{\sigma_j}{3}[m_0(\lambda_j)+2\sqrt{\frac{2}{\pi}}\delta(\lambda_j)]$. Thus, we have

$$\begin{cases}
\boldsymbol{E}_1=\dfrac{\partial e_{ij}}{\partial\boldsymbol{\beta}_j}=-\boldsymbol{x}_i,\\
E_2=\dfrac{\partial e_{ij}}{\partial\sigma_j}=\dfrac{1}{3}\left[m_0(\lambda_j)+2\sqrt{\dfrac{2}{\pi}}\delta(\lambda_j)\right],\\
E_3=\dfrac{\partial e_{ij}}{\partial\lambda_j}=\dfrac{\sigma_j}{3}\left[M_1+\sqrt{\dfrac{2}{\pi}}\dfrac{2}{(1+\lambda_j^2)^{3/2}}\right],\\
\boldsymbol{E}_{11}=\dfrac{\partial^2 e_{ij}}{\partial\boldsymbol{\beta}_j\partial\boldsymbol{\beta}_j^\top}=\boldsymbol{0},\\
E_{12}=E_{21}=\dfrac{\partial^2 e_{ij}}{\partial\boldsymbol{\beta}_j\partial\sigma_j}=\dfrac{\partial^2 e_{ij}}{\partial\sigma_j\partial\boldsymbol{\beta}_j^\top}=0,\\
E_{13}=E_{31}=\dfrac{\partial^2 e_{ij}}{\partial\boldsymbol{\beta}_j\partial\lambda_j}=\dfrac{\partial^2 e_{ij}}{\partial\lambda_j\partial\boldsymbol{\beta}_j^\top}=0,\\
E_{22}=\dfrac{\partial^2 e_{ij}}{\partial\sigma_j\partial\sigma_j}=0,\\
E_{23}=E_{32}=\dfrac{\partial^2 e_{ij}}{\partial\sigma_j\partial\sigma_j}=\dfrac{1}{3}\left[M_1+\sqrt{\dfrac{2}{\pi}}\dfrac{2}{(1+\lambda_j^2)^{3/2}}\right],\\
E_{33}=\dfrac{\partial^2 e_{ij}}{\partial\lambda_j\partial\lambda_j}=\dfrac{\sigma_j}{3}\left[M_2-\sqrt{\dfrac{2}{\pi}}\dfrac{6\lambda_j}{(1+\lambda_j^2)^{5/2}}\right],
\end{cases}$$

and

$$M_1=\frac{\partial m_0(\lambda_j)}{\partial\lambda_j}=\sqrt{\frac{2}{\pi}}\frac{1}{(1+\lambda_j^2)^{3/2}}-\frac{T_1\sigma_0(\lambda_j)+S_1t_0(\lambda_j)}{2}-\frac{\pi\,\mathrm{sign}^2(\lambda_j)}{\lambda_j^2}\exp\left(-\frac{2\pi}{|\lambda_j|}\right),$$

$$M_2=\frac{\partial^2 m_0(\lambda_j)}{\partial\lambda_j\partial\lambda_j}=-\sqrt{\frac{2}{\pi}}-\frac{3\lambda_j}{(1+\lambda_j^2)^{5/2}}-\frac{T_2\sigma_0(\lambda_j)+2T_1S_1+S_2t_0(\lambda_j)}{2}$$

$$-\frac{2\pi[\pi\,\mathrm{sign}(\lambda_j)-\lambda_j\,\mathrm{sign}^2(\lambda_j)]}{\lambda_j^4}\exp\left(-\frac{2\pi}{|\lambda_j|}\right),$$

$$S_1=\frac{\partial\sigma_0(\lambda_j)}{\partial\lambda_j}=-\sqrt{\frac{2}{\pi}}\frac{\mu_0(\lambda_j)}{\sigma_0(\lambda_j)(1+\lambda_j^2)^{3/2}},$$

$$S_2=\frac{\partial^2\sigma_0(\lambda_j)}{\partial\lambda_j\partial\lambda_j}=\frac{1}{\sigma_0(\lambda_j)(1+\lambda_j^2)^3}\left[3\lambda_j\mu_0(\lambda_j)\sqrt{1+\lambda_j^2}-\frac{2}{\pi\sigma_0(\lambda_j)}\right],$$

$$T_1=\frac{\partial t_0(\lambda_j)}{\partial\lambda_j}=\frac{3(4-\pi)}{2\sigma_0^4(\lambda_j)}\left[\sqrt{\frac{2}{\pi}}\frac{\mu_0^2(\lambda_j)\sigma_0(\lambda_j)}{(1+\lambda_j^2)^{3/2}}-\mu_0^3(\lambda_j)S_1\right],$$

$$T_2=\frac{\partial^2 t_0(\lambda_j)}{\partial\lambda_j\partial\lambda_j}=\frac{3(4-\pi)}{2\sigma_0^5(\lambda_j)}\left[\frac{4\mu_0(\lambda_j)\sigma_0^2(\lambda_j)}{\pi(1+\lambda_j^2)^{3/2}}-\sqrt{\frac{2}{\pi}}\frac{3\lambda_j\mu_0^2(\lambda_j)\sigma_0^2(\lambda_j)}{(1+\lambda_j^2)^{5/2}}\right.$$

$$\left.-\sqrt{\frac{2}{\pi}}\frac{6\mu_0^2(\lambda_j)\sigma_0(\lambda_j)S_1}{(1+\lambda_j^2)^{3/2}}-S_2\mu_0^3(\lambda_j)\sigma_0(\lambda_j)+4\mu_0^3(\lambda_j)S_1^2\right].$$