International Journal of Mathematics Trends and Technology
ISSN: 2231-5373/ https://doi.org/10.14445/22315373/IJMTT-V70I1P104

Volume 70 Issue 1, 27-39, January 2024

Original Article

Modal Regression for Varying Coefficient Model in High Dimensions

Zhaoliang Wang[1], Suting Zhang[2]

[1,2]*School of Mathematics and Information Science, Henan Polytechnic University, Henan China.*

[2]*Corresponding Author : 2370430659@qq.com*

Received: 01 December 2023 Revised: 04 January 2024 Accepted: 17 January 2024 Published: 31 January 2024

Abstract *- This paper consider modal regression in varying coefficient model with high dimensionality under a sparsity assumption. We apply the B-spline basis to approximate the varying coefficient functions. First, we demonstrate the convergence rates of the oracle estimator when the nonzero components are known in advance, but their numbers is diverging with the sample size. Then, we propose a nonconvex group SCAD penalized estimator and derive its oracle property under some regularity conditions. That is, under mild conditions, we prove that the oracle estimator is a local solution of the group SCAD penalized estimator of modal regression in varying coefficient model with high dimensionality. Furthermore, we address issues of numerical implementation and of data adaptive choice of the tuning parameters. Some Monte Carlo simulations are provided to corroborate our theoretical findings in finite samples.*

Keywords *- High dimensionality, Modal regression, Oracle property, Varying coefficient model, Variable selection.*

1. Introduction

With the rapid development of date acquisition and storage, high dimensional data sets are especially commonplace in many scientific fields. Examples abound from collaborative filtering [1] to signal processing [2], genome studies [3] and so on. A key feature is that the number of unknow parameters is comparable or even exceeds the sample size. Under the sparsity assumption of the high dimensional parameter vector, a widely used approach is to optimize a suitably penalized loss function (or negative log-likelihood). Fan and Li [4] proposed the Smoothly Clipped Absolute Deviation (SCAD) penalty, which can identify and estimate the nonzero predictors consistently. Tibshirani [5] proposed Least Absolute Shrinkage Operator (Lasso) penalty. Zhang [6] proposed minimax concave penalty (MCP) and considered variable selection in high dimensional linear regression models. Such methods have been proved to possess high computational efficiency as well as desirable statistical properties in a variety of settings. Readers are referred to the review article in Fan and Lv [7] and the monograph in Bühlmann and Van de Geer [8] for a general survey.

To relax the linearity assumption in the classical linear model, many semiparametric models, which retain the flexibility of nonparametric models while avoiding the "curse of dimensionality", have been proposed and studied, c.f.[9]. A leading example of semiparametric models is the varying coefficient model (VCM):

$$Y_i = \sum_{l=1}^{p} \beta_l(U_i)X_{il} + \varepsilon_i = X_i^{\top}\beta(U_i) + \varepsilon_i,\ i = 1, \dots, n, \tag{1}$$

Where $X_i = \left(X_{i1}, \dots, X_{ip}\right)^{\top}$ is a p -dimensional vector of predictor, $\beta(\cdot) = \left(\beta_1(\cdot), \dots, \beta_p(\cdot)\right)^{\top}$ is unknow smooth functions, index variable $U_i \in [0,1]$ for simplicity, and ε_i is a random noise. If $X_{i1} \equiv 1$, the model (1) allow varying intercept term. Throughout the paper, we assume that $\{(Y_i, X_i, U_i), 1 \le i \le n\}$is an independent identically distributed random sample.

Model (1) includes many commonly used parametric, semiparametric and nonparametric models as its special cases. If $\beta_l(u) \equiv \beta_l$ (the constant function) for $l = 1, \dots, q$, and $\beta_l(u)$is unspecified function of $u$for $q+1, \dots, p$, Equation 1 corresponds to semivarying coefficient models or partial linear varying coefficient models. When $q = p$, Equation 1 reduces to the classical linear model. When $q = p - 1$, Equation 1 reduces to the partially linear model. If X_i is a vector of ones, this model becomes the well-know additive model. If $p = 1$ and $X_{ip} = 1$, Equation 1 becomes nonparametric model. Model (1) has received considerable attention during the past decades. They arise in many real applications, see [10] and [11] for various applications of the models. For mean regression of varying coefficient models, estimation can be performed based on local polynomial regression, B-spline expansion, or smoothing splines [12][13][14][15],[17]. For variable selection problems, Wang, Li and Huang [18] considered the varying coefficient model in a longitudinal data setting built on the SCAD approach, Wang and Xia [19] proposed the use of local polynomial regression with an adaptive Lasso penalty, Wei, Huang and Li [20] proposed an adaptive group Lasso approach using B-spline basis approximation.

Studies of the model (1) have yielded promising results. No longer listed here. The aforementioned existing researches were mainly built on either the least-square or empirical likelihood method, which are expected to be very sensitive to the outliers and its efficiency may be significantly decreased for many commonly used non-normal or skewed errors.

Skewed or heavy-tailed data (e.g. wages, prices, scores on a difficult exam, movie ticket sales, and expenditures) appear in a broad variety of practical applications, including economic, statistical, social and educational research studies, among others . In such instances, the mean estimate may not adequately disclose the dataąrs characteristics, and the mode estimate (one of the center measures) should be considered as a supplemental measure to capture the 'most likely' element of the data. In light of the robustness of mode, an increasing amount of literature pays attention to the conditional modal regression. Yao [21] studied the local modal regression for nonparametric models, which is robust when the data sets have heavy-tail or non-normal distributional error, and asymptotically efficient even if there are no outliers or the error follows normal distribution. Yao and Li [22] further proposed a robust modal linear regression for parametric models. A distinguishing characteristic of the mode regression is that it introduces an additional tuning parameter (i.e., bandwidth h) that is automatically selected using the observed data in order to achieve both robustness and efficiency of the resulting estimation. Zhang, Liu and Yang[23][24][25] investigated this robust estimation method for partial linear varying coefficient models, single-index models and single index varying coefficient models, respectively.

However, the new modal regression approach was only considered for fixed dimension. These facts motivate us to extend the modal regression method to the VCM in high dimensions. Our primary interest is to investigate the variable selection and estimation for model (1) in high dimensional setting. With high dimensionality, we allow$p \to \infty$as $n \to \infty$and denote it by p_n. In particular, we allow $\ln(p_n) = O(n^a)$ for some constant $a > 0$, but the number of predictors related to the response grows slowly with the sample size n. Thus, this work fulfills an important gap in the existing literature on semiparametric models by developing variable selection methodology that allows high dimensional parameter vector.

Local polynomial regression is a most popular approach, but it requires solving many similar optimization problems on a fine grid on the support of the index variable. See [26] for more details. Thus here we choose the B-spline expansion approach, which is more convenient to implement. That is we approximate the nonparametric coefficient functions using B-spline basis. We first demonstrate the convergence rates of the nonparametric coefficient functions for the oracle estimator, that is, the one obtained when the nonzero components are known in advance. Of course, it is infeasible in practice for unknown true active set. It is worth pointing out that our asymptotic framework allows the number of the nonzero components grows with the sample size. This resonates with the perspective that a more complex statistical model can be fit when more data are collected. Then, we propose a nonconvex group penalized estimator for simultaneous variable selection and estimation when p_n is of an exponential order of the sample size n and the model has a sparse structure. With a proper choice of the regularization parameters and the penalty function, such as the popular SCAD, we derive the oracle property of the proposed estimator under relaxed conditions. Specifically, we prove that the oracle estimator is a local solution of the group SCAD penalized modal regression problem. This indicates that the penalized estimators work as well as if the subset of true nonzero coefficients was already known. Moreover, a modified version of a modal expectation-maximization (MEM) algorithm is proposed to obtain the solutions for the object function. Lastly, we address issues of practical implementation of the proposed method.

The rest of the paper is organized as follows. In Section 2, we first present the oracle estimator, then develop the methodology for group SCAD penalized modal varying coefficient regression model, and give some the theoretical properties. In Section 3, we discuss the computation approach by combining MEM algorithm and local quadratic algorithm, and selection methods for the tuning parameters. In Section 4, simulation studies are provided to illustrate good numerical results of the proposed methodology. Section 5 provides all technical proofs. Section 6 concludes the paper with a brief discussion.

3.

4. 2.Methodology and asymptotic properties

For high dimensional statistical inference, it is often assumed that the true coefficient $\beta_0(u)$ in model (1) is sparse vector, where $\beta_0(u)$ is the true value of $\beta(u)$. Let $\mathcal{S} = \{l: \|\beta_{0l}(u)\|_{L^2} \neq 0, l \in \{1, \dots, p_n\}\}$ be the index set of the nonzero varying coefficients, then its cardinality $|\mathcal{S}| = q_n < n$. The asymptotic framework also allows $q_n \to \infty$ as $n \to \infty$, which is of independent interests. The set $\mathcal{S}$ is unknown. The main goal is to identify the true model $\mathcal{S}$ and derive the optimal rate of convergence for $\beta_l(u)$ with $l \in \mathcal{S}$.

2.1. Oracle Estimator

Nonparametric functions $\beta_l(u)$ with $l \in \{1, \dots, p_n\}$ in model (1) can be approximated using a linear combination of B-spline basis functions. First, one definition is provided to define the class of functions that can be estimated with B-

splines. Define $\mathcal{H}_r$ as the collection of functions $h(\cdot)$ on [0,1] whose $\lfloor r \rfloor$-th derivative $h^{(\lfloor r \rfloor)}(\cdot)$ satisfies the Hölder condition of order $r - \lfloor r \rfloor$, where $\lfloor r \rfloor$ denotes the largest integer strictly smaller than r. That is, for each $h(\cdot) \in \mathcal{H}_r$, there exists some positive constant C such that $\left|h^{(\lfloor r \rfloor)}(u_1) - h^{(\lfloor r \rfloor)}(u_2)\right| \leq C|u_1 - u_2|^{r-\lfloor r \rfloor}$, for any $0 \leq u_1,\ u_2 \leq 1$.

Let $\pi(u) = \left(B_1(u), \dots, B_{k_n+\hbar}(u)\right)^\top$ be a vector of normalized B-spline basis functions of order $\hbar$ with k_n quasi-uniform internal knots on [0,1]. Under condition (C1) below, $\beta_{0l}(u),\ l = 1, \dots, p_n$, can be approximated using a linear combination of $\pi(u)$. The readers are referred to De Boor[27] for details of the B-spline construction, and the result that there exists $\gamma_{0l} \in \mathbb{R}^{K_n}$, where$K_n = k_n + \hbar$, such that $\sup_u|R_l(u)| = O(K_n^{-r})$ with $R_l(u) = \pi(u)^\top\gamma_{0l} - \beta_{0l}(u)$. For ease of notation and simplicity of proofs, we use the same number of basis functions for different varying coefficient in model (1). In practice, such restrictions are not necessary.

Using B-spline expansion, each varying coefficient function $\beta_l(u)$ with $l \in \{1, \dots, p_n\}$ in model (1) can be approximated by

$$\beta_l(u) \approx \sum_{j=1}^{K_n} B_j(u)\gamma_{lj} = \pi(u)^\top\gamma_l, \tag{2}$$

Where $\gamma_l = \left(\gamma_{l1}, \dots, \gamma_{lK_n}\right)^\top$. Then, the varying coefficient modal regression can be approximated by

$$Y_i = \sum_{l=1}^{p_n} \beta_l(U_i)X_{il} + \varepsilon_i \approx \sum_{l=1}^{p_n} \pi(U_i)^\top\gamma_l X_{il} + \varepsilon_i = \Pi_i^\top\gamma + \varepsilon_i, \tag{3}$$

Where $\Pi_i = \left(X_{i1}\pi(U_i)^\top, \dots, X_{ip_n}\pi(U_i)^\top\right)^\top$ and $\gamma = \left(\gamma_1^\top, \dots, \gamma_{p_n}^\top\right)^\top$.

Now we consider the oracle estimator with the oracle information that the index set $\mathcal{S}$ is known in advance. Without loss of generality, the first q_n coefficient functions among $\beta_{01}(u), \dots, \beta_{0p_n}(u)$ are nonzero and the remaining $p_n - q_n$ coefficient functions are zero. In other words, $\mathcal{S} = \{1, \dots, q_n\}$. The oracle estimator is obtained assuming that one only uses the predictor $X_{il},\ l \in \mathcal{S}$ in fitting model (1).

Let $\gamma = \left(\gamma_{\mathcal{S}}^\top, \gamma_{\mathcal{S}^c}^\top\right)^\top$ and $\Pi_i = \left(\Pi_{i,\mathcal{S}}^\top, \Pi_{i,\mathcal{S}^c}^\top\right)^\top$, where $\gamma_{\mathcal{S}}$, $\Pi_{i,\mathcal{S}} \in \mathbb{R}^{q_nK_n}$ and $\gamma_{\mathcal{S}^c}$, $\Pi_{i,\mathcal{S}^c} \in \mathbb{R}^{(p_n-q_n)K_n}$. The oracle estimator of the mode regression parameter γ is $\hat{\gamma}^o = \left(\hat{\gamma}_{\mathcal{S}}^{o\top}, 0_{(p_n-q_n)K_n}^\top\right)^\top$, where $\hat{\gamma}_{\mathcal{S}}^o$ can be obtained by maximizing the kernel based objective function

$$Q_h(\gamma_{\mathcal{S}}) = \sum_{i=1}^{n} \phi_h\left(Y_i - \Pi_{i,\mathcal{S}}^\top\gamma_{\mathcal{S}}\right), \tag{4}$$

With respect to $\gamma_{\mathcal{S}}$. The oracle estimator for the coefficient function $\beta_{0l}(u)$ is $\hat{\beta}_l^o(u) = \pi(u)^\top\hat{\gamma}_l^o$for $l = 1, \dots, p_n$. In objective function Equation 4, $\phi_h(\cdot) = h^{-1}\phi(\cdot/h)$, $\phi(\cdot)$ is a kernel density function symmetric about 0 and $h > 0$ is a bandwidth. For the remainder of the paper, we will assume that $\phi(\cdot)$ is the standard normal density (for the simplicity of computation). It should be noted that all the asymptotic results presented in this article still hold if other kernels are used, see[22].

The asymptotic properties of the oracle estimators as q_n diverges are presented. Let

$$F(x, u, h) = E\left(\phi_h''(\varepsilon) \mid X = x, U = u\right),$$

and

$$G(x, u, h) = E\left(\phi_h'(\varepsilon)^2 \mid X = x, U = u\right).$$

The following technical conditions are imposed for theoretical analysis.

(C1) For $l = 1, \dots, p_n, \beta_{0l}(u) \in \mathcal{H}_r$ for some $r > 1.5$.

(C2) There exist positive constants c_2 and c_3 such that $c_2I_{q_n} \leq E(X_{\mathcal{S}}X_{\mathcal{S}}^\top \mid U) \leq c_3I_{q_n}$, where I_{q_n} is a $q_n \times q_n$ identity matrix and $X_{\mathcal{S}} = \left(X_1, \dots, X_{q_n}\right)^\top$. In addition, we assume $\max_i\|X_i\|/\sqrt{n} = o_P(1)$.

(C3) The index variable U has a compact support on [0,1] and its density is absolutely continuous and bounded away from 0 and infinity.

(C4) $q_n = O(n^\kappa)$ for some $\kappa \in (0,1/4)$.

(C5) The number of interior knots $k_n \asymp n^{1/(2r+1)}$; where r is defined by the Condition (C1). Throughout, we use $a_n \asymp b_n$ to mean that a_n and b_n have the same order as $n \to \infty$.

(C6) For each $i \in \{1, \ldots, n\}$, let $\varepsilon_i = Y_i - \sum_{l=1}^{q_n} \beta_{0l}(U_i) X_{il}$. The conditional distribution of ε_i given (U_i, X_i) have a density function $f_i(\cdot \mid U_i, X_i)$ with $0 < c < f_i(0|U_i, X_i) < C < \infty$ for some constants c, C. The density function $f_i(\cdot \mid U, X)$ has a bounded first derivative in neighborhood of zero, uniformly over i.

(C7) There exists $c_3 \in (1/(2r+1), 1)$and a positive constant $c_4$such that $\min_{l \in \mathcal{S}} \|\beta_{0l}(u)\|_{L^2} \geq c_4\left(n^{-(1-c_3)/2} k_n^{1/2} + k_n^{-r}\right)$.

(C8) $F(x, u, h)$and $G(x, u, h)$are continuous with respect to (x, u). And $F(x, u, h) < 0$ for any $h > 0$.

(C9) $E(\phi_h{}'(\varepsilon_i) \mid X, U) = 0$, for some $\zeta > 0$, $\max_{1 \leq i \leq n} E\left(|\phi_h{}'(\varepsilon_i)|^{2+\zeta} \mid X, U\right) < \infty$ $\max_{1 \leq i \leq n} E\left(|\phi_h{}'(\varepsilon_i)|^{2+\zeta} \mid X, U\right) < \infty$, and $E(\phi_h{}''(\varepsilon)^2 \mid X = x, U = u)$, $E(\phi_h{}'(\varepsilon)^3 \mid X = x, U = u)$ and $E(\phi_h{}'''(\varepsilon) \mid X = x, U = u)$ are continuous with respect to x and u.

The theorem below summarizes the convergence rates of the oracle estimators.

Theorem 1. Assume that regularity Conditions (C1)-(C9) hold, as $n \to \infty$, we have

$$\frac{1}{n} \sum_{l \in \mathcal{S}} \sum_{i=1}^{n} \left\{\hat{\beta}_l^o(U_i) - \beta_{0l}(U_i)\right\}^2 = O_p\{q_n(K_n n^{-1} + K_n^{-2r})\}.$$

An interesting observation is that since we allow q_n to diverge with n, it affects the convergence rates for estimating $\beta_l(\cdot)$. If q_n is fixed, the convergence rate reduces to $n^{-2r/2r+1}$ for estimating $\beta_l(\cdot)$, which is the optimal rate of convergence.

2.2. Variable Selection

In real data analysis, it does not know which of the p_n covariates in X_i are important. To encourage sparse estimation, we define the following group penalized estimation for model (1) based on modal regression as

$$L_n(\gamma) = \sum_{i=1}^{n} \quad \phi_h(Y_i - \Pi_i^{\top} \gamma) - n \sum_{l=1}^{p_n} \quad p_\lambda(\gamma_l), \tag{5}$$

Where $\|\cdot\|$ denotes the Euclidean metric and $p_\lambda(\cdot)$ is a penalty function with tuning parameter $\lambda > 0$ which controls the complexity of the selected model and goes to zero as $n \to \infty$. Although it is not necessarily that the tuning parameter λ is the same for all γ_l in practice, we make the above choices for simplicity.

Here, we focus on the popular nonconvex SCAD penalty [4] given by

$$p_\lambda'(|t|) = \lambda \left\{ I(|t| \leq \lambda) + \frac{(a\lambda - |t|)_+}{(a-1)\lambda} I(|t| > \lambda) \right\},$$

for some $a > 2$, where $x_+ = \max(x, 0)$, $I(\cdot)$is the indicator function. Note that the SCAD penalty is continuously differentiable on $(-\infty, 0) \cup (0, \infty)$ but singular at 0 and its derivative vanishes outside $[-a\lambda, a\lambda]$. As illustrated in[28], these features of SCAD penalty result in a solution with three desirable properties: asymptotic unbiasedness, sparsity and continuity. Other choices of penalty, such as MCP, are expected to produce similar results in both theory and practice. In comparison, Lasso is known to over-penalize large coefficients, tends to be biased and requires strong conditions on the design matrix to achieve selection consistency. This is usually not a concern for prediction, but can be undesirable if the goal is to identify the underlying model.

The theorem below shows that the oracle estimator is a local maximizer of Equation 5 using SCAD penalty with probability tending to one, provided the following additional condition (C10), which is needed to identify the underlying model. The condition (C10) (i) is how quickly a nonzero signal can decay which is not a concern when the dimension is fixed, and the condition (C10) (ii) is concerning the divergence rate of p_n.

(C10) (i) $\min_{1 \leq l \leq q_n} \gamma_{0l} \gg \lambda \gg \sqrt{q_n(K_n n^{-1} + K_n^{-2r})}$;

(ii) $\sqrt{n K_n^{-1} \ln(p_n K_n)} + n q_n K_n^{-(2r-1)} \log(p_n \vee n) \ll n\lambda$.

Theorem 2. Consider the group SCAD penalty with tuning parameter λ, let $\varepsilon(\lambda)$ be the set of local maximizers of Equation 5 for a given tuning parameter λ, under regularity Conditions (C1)-(C10), as $n \to \infty$ we have

$$Pr\{\hat{\gamma}^o \in \varepsilon(\lambda)\} \to 1.$$

Theorem 2 shows that the oracle estimator is one, possibly among multiple, local maximizers of the group SCAD penalized objective. This of course leaves many open questions, including whether the oracle estimator is a global maximizer and whether the computationally feasible estimator returned by the algorithm has desired oracle property. For some other models, such questions have been considered in[28, 29]. We leave the detailed investigations along this direction to the future.

3. Materials and Methods

This section introduces a computational algorithm for obtaining the maximizers of Equation 5 and selection methods for the tuning parameters.

3.1. Algorithm

For given the tuning parameters, finding the solution that maximizes Equation 5 poses a number of interesting challenges because there is no closed-form expression of the maximizer of Equation 5 and the SCAD penalty function is nondifferentiable at the origin and nonconvex. We combine the MEM algorithm for modal regression, proposed by Li et al, [30] and Yao et al, [21], and an approximation based on local quadratic approximation, proposed by Fan and Li [4], to solve Equation 5.

Firstly, the penalty function $p_\lambda(\|\gamma_l\|)$is approximated by local quadratic approximation. More specifically, given the initial value $\hat{\gamma}_l^{(0)}$, $l = 1, \dots, p_n$, and a specified small positive number ξ, when $\|\hat{\gamma}_l^{(0)}\| < \xi$, let $\hat{\gamma}_l = 0$; when $\|\hat{\gamma}_l^{(0)}\| \geq \xi$, a local quadratic approximation penalty function can be used

$$p_\lambda(\|\gamma_l\|) \approx p_\lambda\left(\|\hat{\gamma}_l^{(0)}\|\right) + \frac{1}{2}\frac{p_\lambda'\left(\|\hat{\gamma}_l^{(0)}\|\right)}{\|\hat{\gamma}_l^{(0)}\|}\left(\|\gamma_l\|^2 - \|\hat{\gamma}_l^{(0)}\|^2\right).$$

With the aid of the local quadratic approximation and by extending the MEM algorithm, optimization problem Equation 5 is further implemented iteratively using the following algorithm.

Step 0 (Initializing). The initial estimator $\hat{\gamma}^{(0)}$ is obtained by using the group Lasso penalty

$$\hat{\gamma}^{(0)} = \underset{\gamma}{argmax} \sum_{i=1}^{n} \phi_h\left(Y_i - \Pi_i^\top \gamma\right) - n\sum_{l=1}^{p_n}\|\gamma_l\|,$$

and set $k = 0$.

Step 1 (E-step). Update weights $\pi\left(i \mid \hat{\gamma}^{(k)}\right)$, for $i = 1, \dots, n$ as

$$\pi\left(i \mid \hat{\gamma}^{(k)}\right) = \frac{\phi_h\left(Y_i - \Pi_i^\top \hat{\gamma}^{(k)}\right)}{\sum_{j=1}^{n}\phi_h\left(Y_j - \Pi_j^\top \hat{\gamma}^{(k)}\right)} \propto \phi_h\left(Y_i - \Pi_i^\top \hat{\gamma}^{(k)}\right). \tag{6}$$

Step 2 (M-step). Update $\hat{\gamma}^{(k+1)}$

$$\begin{aligned}\hat{\gamma}^{(k+1)} &= \underset{\gamma}{argmax} \sum_{i=1}^{n}\left\{\pi\left(i \mid \hat{\gamma}^{(k)}\right) log\, \phi_h\left(Y_i - \Pi_i^\top \gamma\right)\right\} - n\sum_{l=1}^{p_n}\frac{1}{2}\frac{p_\lambda'\left(\|_l\hat{\gamma}_l^{(k)}\|\right)}{\|_l\hat{\gamma}_l^{(k)}\|}\|\gamma_l\|^2 \\ &= \left(\Pi^\top W^{(k)}\Pi + n\Sigma_\lambda^{(k)}\right)^{-1}\Pi^\top W^{(k)}Y,\end{aligned} \tag{7}$$

where $Y = (Y_1, \dots, Y_n)^\top$, $\Pi = (\Pi_1, \dots, \Pi_n)^\top$, $W^{(k)}$is an $n \times n$ diagonal matrix with diagonal elements $\pi\left(j \mid \hat{\gamma}^{(k)}\right)$, and $\Sigma_\lambda^{(k)} = diag\left\{\frac{p_\lambda'\left(\|\hat{\gamma}_1^{(k)}\|\right)}{\|\hat{\gamma}_1^{(k)}\|}\mathbb{I}_{K_n}, \dots, \frac{p_\lambda'\left(\|\hat{\gamma}_{pn}^{(k)}\|\right)}{\|\hat{\gamma}_{p_n}^{(k)}\|}\mathbb{I}_{K_n}\right\}$, $\mathbb{I}_{K_n}$ is a $K_n \times K_n$ identity matrix

Step 3 Repeat Step 1 and Step 2 until the algorithm convergence (In this iteration, take $\xi = 10^{-3}$), denote the estimator of γ by $\hat{\gamma}$.

Similar to the EM algorithm, the above MEM algorithm for the varying coefficient model within each step also consists of two steps: E-step and M-step. The ascending property of the proposed MEM algorithm can be established along the lines of the study of [30]. Note that the converged value may depend on the starting point as the usual EM algorithms, and there is no guarantee that the MEM algorithm will converge to the global optimal solution. Therefore, it is prudent to run the algorithm from several starting-points and choose the best local optima found.

3.2. Tuning Parameters Selection

To implement the above estimation procedures and achieve good numerical performance, the spline order $\hbar$ and the number of basis K_n, bandwidth h as well as the regularization parameter λ should be chosen appropriately. Due to the computation complexity, it is often impractical to automatically select all four components based on the observable data. As a commonly adopted strategy, it is often to fix $\hbar = 4$ (cubic splines). Note that K_n should not be too large since the larger the $K_n$is, the larger the estimation variance is, and the more difficult it is to distinguish important variables from unimportant ones. On the other hand, K_n should not be too small to create probing biases. For computation convenience, let $K_n = \lfloor n^{1/5} \rfloor + \hbar$. In the following simulations, it also conducts a sensitivity analysis by setting K_n to be different

values. There are similar numerical results if K_n varies in a reasonable range. A same arguments for bandwidth h can also be found in subsection 3.1 of [22].

Fixed $\hbar$, $K_n$and h, it is critical for the performance of the estimators to employ a data-driven method to chooseλ. Cross validation is a common approach, but is known to often result in overfitting. In high dimensional context, it is suitable for us to use the extended Bayesian information criterion (EBIC) in Chen and Chen [31] to select λ. More specifically, the EBIC can be defined as

$$EBIC(\lambda) = log\Big(n^{-1}\sum_{i=1}^{n}\phi_h(Y_i - \Pi_i^{\top}\hat{\gamma}_\lambda)\Big) + \hat{q}_{n\lambda}\frac{log\, n}{n} + 2C_n\hat{q}_{n\lambda}\frac{log\, p_n}{n}, \tag{8}$$

Where $\hat{\gamma}_\lambda$ is the solutions based on Equation 5 for given λ, $\hat{q}_{n\lambda}$ is the number of nonzero values in $\hat{\gamma}_\lambda$ and C_n is a tuning parameter which is taken as $1 - \log(n)/\big(3\log(p)\big)$ suggested by Chen and Chen [31]. Note that when $C_n = 0$, the EBIC is the BIC. From the following numerical studies, it can find that the above data-driven procedure works well.

4. Results and Discussion

4.1. Simulations

In this section, two simulations are conducted to demonstrate the finite sample performance of the proposed method. To examine the efficiency of the proposed modal regression with group SCAD estimator (mSCAD), we compare it the following alternative estimators: (1) lsLasso: the least squares with group Lasso estimator; (2) lsSCAD: the least squares with group SCAD estimator and (3) Oracle: the least squares when the nonzero subset of $\beta(\cdot)$ is known. However, in reality, we seldom know this nonzero subset. The Oracle estimator serves as a benchmark, omniscient estimator to check how well the estimators. The lsLasso and lsSCAD are computed by the R package grpreg with tuning parameter λ being selected by cross-validation.

In this simulation study, we generate the data with different sample sizes of $n \in \{100{,}300\}$, and consider both p =10, 200 and 500 to examine the performance of model selection and estimation when p is smaller than, close to, or exceeds the sample size. A total of 100 simulation replications are conducted for each model setup.

The performance of the nonparametric estimate $\hat{\beta}(\cdot)$will be assessed by using the square root of average square errors (RASE), defined by

$$RASE(\hat{\beta}(\cdot)) = \left[\frac{1}{n_{\text{grid}}}\sum_{j=1}^{n_{\text{grid}}}\left\|\hat{\beta}(u_j) - \beta(u_j)\right\|^2\right]^{1/2} \tag{9}$$

where $\{u_j, j = 1, \ldots, n_{\text{grid}}\}$ is a set of grid points uniformly placed on [0,1] at which the functions $\hat{\beta}(\cdot)$are evaluated. We considered $n_{\text{grid}} = 200$. The sample mean of the RASEs over 100 simulations are presented. In addition, we calculate the average number of the true zero coefficients that were correctly set to zero (CZ) and the average number of the truly nonzero coefficients that were incorrectly set to zero (IZ). We also report the proportion of selecting the true model (Correct), the proportion of including at least one irrelevant predictor but does not miss any relevant one (Over) and the proportion of excluding at least one relevant predictor (Under), respectively.

Consider the following sparse model

$$Y_i = \beta_0(U_i)X_{i0} + \beta_1(U_i)X_{i1} + \beta_2(U_i)X_{i2} + \varepsilon_i, \tag{10}$$

Where $X_{i0} \equiv 1$ represents the intercept, the index variable U_i 's are sampled uniformly on [0,1]. The other covariates $(X_{i1}, \ldots, X_{ip})^{\top}$ are independently drawn from multivariate normal distribution $N_p(0, \Sigma)$, where Σ is the Toeplitz covariance matrix with $\Sigma_{j_1 j_2} = \rho^{|j_1 - j_2|}$ for $1 \le j_1$, $j_2 \le p$. Here we consider $\rho = 0$ and 0.5 to test the effect of correlation structure. The error ε_i is independent of covariates and follows three different distributions: (i) $N(0,1)$, (ii) t - distribution with 3 degrees of freedom and (iii) mixture of normals: $0.5N(-1, 2.5^2) + 0.5N(1, 0.5^2)$. All the generated random errors are standardized to have mean 0 and variance 1 . For case (i) and (ii), the conditional mean regression function is the same as the conditional modal regression function. That is, $E(Y \mid X, U) = \text{Mode}(Y \mid X, U) = \beta_0(U) + \beta_1(U)X_1 + \beta_2(U)X_2$. For case (iii), the conditional mean regression function is $E(Y \mid X, U) = \beta_0(U) + \beta_1(U)X_1 + \beta_2(U)X_2$, however, the conditional modal regression function is $\text{Mode}(Y \mid X, U) = \beta_0(U) + \sqrt{4/17} + \beta_1(U)X_1 + \beta_2(U)X_2$.

In model (10), let$\beta_0(u) = \exp(2u - 1)$, $\beta_1(u) = 8u(1 - u)$ and $\beta_2(u) = 2\sin(2\pi u)$. All other coefficients $\beta_j(u)$, $j = 3, \ldots, p$, are set to be zero. Then, the first three covariates (including intercept) are relevant for predicting the response variable from the point of view both mean regression and modal regression even though their regression parameters are different.

4.2. Result Analysis

Simulation results are summarized in Tables 1-3. Several observations can been seen from these tables.

(1) It can be seen that for the normal error, the RASE of MSCAD are slightly larger than those for lsSCAD due to the emphasis of cross-validation on prediction.

(2) For non-normal distribution error, the mSCAD estimate has great gain of efficiency and robustness over the lsLasso and lsSCAD methods.

(3) For a given error distribution, the performance of mSCAD becomes better and better and its advantages are more prominent over others when the sample size increases. MSCAD can identify the true model with high probability. There is little difference for the mSCAD estimate with Oracle in terms of all criteria especially when the sample size increases.

Table 1. Simulation results over 100 replications when $\varepsilon \sim N(0,1)$.

			$\rho = 0$						$\rho = 0.5$					
n	p	Method	RASE	CZ	IZ	Under	Correct	Over	RASE	CZ	IZ	Under	Correct	Over
100	10	lsLasso	0.831	4.23	0.00	0.00	0.06	0.94	0.712	3.86	0.00	0.00	0.00	1.00
		lsSCAD	0.584	6.15	0.00	0.00	0.28	0.72	0.637	6.30	0.00	0.00	0.29	0.74
		mSCAD	0.601	7.97	0.00	0.00	0.97	0.03	0.700	7.80	0.00	0.00	0.89	0.11
		Oracle	0.520	8.00	0.00	0.00	1.00	0.00	0.573	8.00	0.00	0.00	1.00	0.00
	200	lsLasso	0.831	182.78	0.00	0.00	0.02	0.98	0.952	183.58	0.00	0.00	0.01	0.99
		lsSCAD	0.658	187.18	0.00	0.00	0.07	0.93	0.692	188.74	0.00	0.00	0.05	0.95
		mSCAD	0.567	197.96	0.01	0.01	0.95	0.04	0.623	197.97	0.05	0.05	0.92	0.03
		Oracle	0.539	198.00	0.00	0.00	1.00	0.00	0.552	198.00	0.00	0.00	1.00	0.00
	500	lsLasso	0.893	479.41	0.00	0.00	0.00	1.00	1.029	480.40	0.00	0.00	0.01	0.99
		lsSCAD	0.675	485.97	0.00	0.00	0.05	0.95	0.780	483.77	0.00	0.00	0.03	0.97
		mSCAD	0.550	497.99	0.01	0.01	0.98	0.01	0.700	497.96	0.07	0.07	0.89	0.04
		Oracle	0.534	498.00	0.00	0.00	1.00	0.00	0.607	498.00	0.00	0.00	1.00	0.00
300	10	lsLasso	0.393	3.22	0.00	0.00	0.01	0.99	0.406	3.10	0.00	0.00	0.01	0.99
		lsSCAD	0.307	6.31	0.00	0.00	0.35	0.65	0.317	6.91	0.00	0.00	0.49	0.51
		mSCAD	0.336	8.00	0.00	0.00	1.00	0.00	0.361	7.99	0.00	0.00	0.99	0.01
		Oracle	0.292	8.00	0.00	0.00	1.00	0.00	0.311	8.00	0.00	0.00	1.00	0.00
	200	lsLasso	0.468	177.08	0.00	0.00	0.01	0.99	0.564	177.78	0.00	0.00	0.00	1.00
		lsSCAD	0.320	189.28	0.00	0.00	0.16	0.84	0.337	190.80	0.00	0.00	0.12	0.88
		mSCAD	0.305	198.00	0.00	0.00	1.00	0.00	0.311	197.99	0.00	0.00	0.99	0.01
		Oracle	0.290	198.00	0.00	0.00	1.00	0.00	0.306	198.00	0.00	0.00	1.00	0.00
	500	lsLasso	0.492	472.82	0.00	0.00	0.00	1.00	0.573	476.33	0.00	0.00	0.01	0.99
		lsSCAD	0.329	485.60	0.00	0.00	0.08	0.92	0.342	488.65	0.00	0.00	0.06	0.94
		mSCAD	0.298	498.00	0.00	0.00	1.00	0.00	0.317	498.00	0.00	0.00	1.00	0.00
		Oracle	0.291	498.00	0.00	0.00	1.00	0.00	0.307	498.00	0.00	0.00	1.00	0.00

(4) For fixed n, the performance of the mSCAD does not deteriorate rapidly when p increases, while for fixed p the performance improves substantially as the sample size increases. In turn, these results show that the sample size n is more important than the dimension of the covariates for high dimensional statistical inference.

(5) It is easy to see that the mSCAD is not sensitive to the different correlations between variables.

(6) It is very interesting to see that the values of RASE for the mSCAD estimate become smaller than other methods when the error follows a mixture normal. The main reason for this is that when there are some very large outliers in the data, the modal regression will put more weight on the "most likely" data around the true value, which leads to a robust and efficient estimator. These findings agree with our asymptotic properties.

Table 2. Simulation results over 100 replications when $\varepsilon \sim t(3)$.

			$\rho = 0$						$\rho = 0.5$					
n	p	Method	RASE	CZ	IZ	Under	Correct	Over	RASE	CZ	IZ	Under	Correct	Over
100	10	lsLasso	0.614	4.69	0.00	0.00	0.01	0.99	0.666	4.30	0.00	0.00	0.03	0.97
		lsSCAD	0.554	6.50	0.00	0.00	0.34	0.66	0.594	6.39	0.00	0.00	0.34	0.66
		mSCAD	0.443	7.88	0.00	0.00	0.94	0.06	0.488	7.97	0.00	0.00	0.97	0.03
		Oracle	0.478	8.00	0.00	0.00	1.00	0.00	0.538	8.00	0.00	0.00	1.00	0.00
	200	lsLasso	0.807	182.81	0.00	0.00	0.01	0.99	0.934	181.94	0.00	0.00	0.00	1.00
		lsSCAD	0.646	189.23	0.00	0.00	0.09	0.91	0.706	188.64	0.00	0.00	0.03	0.97
		mSCAD	0.452	197.95	0.02	0.02	0.95	0.03	0.552	197.91	0.04	0.04	0.89	0.07
		Oracle	0.491	198.00	0.00	0.00	1.00	0.00	0.551	198.00	0.00	0.00	1.00	0.00

	500	lsLasso	0.875	476.00	0.00	0.00	0.00	1.00	0.973	477.23	0.00	0.00	0.00	1.00
		lsSCAD	0.674	485.18	0.00	0.00	0.04	0.96	0.683	485.93	0.00	0.00	0.06	0.94
		mSCAD	0.501	497.89	0.02	0.02	0.94	0.04	0.547	497.93	0.06	0.05	0.90	0.05
		Oracle	0.508	498.00	0.00	0.00	1.00	0.00	0.531	498.00	0.00	0.00	1.00	0.00
300	10	lsLasso	0.356	3.34	0.00	0.00	0.01	0.99	0.409	3.29	0.00	0.00	0.00	1.00
		lsSCAD	0.276	6.92	0.00	0.00	0.50	0.50	0.307	6.92	0.00	0.00	0.49	0.51
		mSCAD	0.230	8.00	0.00	0.00	1.00	0.00	0.239	8.00	0.00	0.00	1.00	0.00
		Oracle	0.265	8.00	0.00	0.00	1.00	0.00	0.297	8.00	0.00	0.00	1.00	0.00
	200	lsLasso	0.459	178.38	0.00	0.00	0.01	0.99	0.559	181.49	0.00	0.00	0.00	1.00
		lsSCAD	0.314	191.93	0.00	0.00	0.17	0.83	0.350	192.25	0.00	0.00	0.16	0.84
		mSCAD	0.233	198.00	0.00	0.00	1.00	0.00	0.260	198.00	0.00	0.00	1.00	0.00
		Oracle	0.276	198.00	0.00	0.00	1.00	0.00	0.310	198.00	0.00	0.00	1.00	0.00
	500	lsLasso	0.480	472.28	0.00	0.00	0.00	1.00	0.575	474.75	0.00	0.00	0.00	1.00
		lsSCAD	0.314	488.04	0.00	0.00	0.08	0.92	0.358	489.20	0.00	0.00	0.11	0.89
		mSCAD	0.238	498.00	0.00	0.00	1.00	0.00	0.263	498.00	0.00	0.00	1.00	0.00
		Oracle	0.278	498.00	0.00	0.00	1.00	0.00	0.314	498.00	0.00	0.00	1.00	0.00

Table 3. Simulation results over 100 replications when $\boldsymbol{\varepsilon \sim 0.5N(-1, 2.5^2) + 0.5N(1, 0.5^2)}$.

			$\rho = 0$						$\rho = 0.5$					
n	p	Method	RASE	CZ	IZ	Under	Correct	Over	RASE	CZ	IZ	Under	Correct	Over
100	10	lsLasso	0.806	3.78	0.00	0.00	0.01	0.99	0.840	4.29	0.00	0.00	0.05	0.95
		lsSCAD	0.765	6.17	0.00	0.00	0.27	0.73	0.807	6.17	0.00	0.00	0.29	0.71
		mSCAD	0.594	7.88	0.00	0.00	0.89	0.11	0.648	7.89	0.00	0.00	0.92	0.08
		Oracle	0.713	8.00	0.00	0.00	1.00	0.00	0.748	8.00	0.00	0.00	1.00	0.00
	200	lsLasso	0.960	182.86	0.00	0.00	0.02	0.98	1.112	181.15	0.00	0.00	0.01	0.99
		lsSCAD	0.798	187.55	0.00	0.00	0.06	0.94	0.886	187.23	0.00	0.00	0.03	0.97
		mSCAD	0.639	197.99	0.00	0.00	0.99	0.01	0.789	197.96	0.06	0.06	0.90	0.04
		Oracle	0.688	198.00	0.00	0.00	1.00	0.00	0.763	198.00	0.00	0.00	1.00	0.00
	500	lsLasso	1.003	480.23	0.00	0.00	0.00	1.00	1.111	478.13	0.00	0.00	0.00	1.00
		lsSCAD	0.835	485.67	0.00	0.00	0.04	0.96	0.888	484.64	0.00	0.00	0.01	0.99
		mSCAD	0.727	497.94	0.05	0.05	0.90	0.05	0.815	497.88	0.06	0.06	0.84	0.10
		Oracle	0.693	498.00	0.00	0.00	1.00	0.00	0.744	498.00	0.00	0.00	1.00	0.00
300	10	lsLasso	0.614	3.52	0.00	0.00	0.02	0.98	0.633	3.32	0.00	0.00	0.01	0.99
		lsSCAD	0.569	6.96	0.00	0.00	0.50	0.50	0.581	6.97	0.00	0.00	0.58	0.42
		mSCAD	0.295	8.00	0.00	0.00	1.00	0.00	0.335	7.99	0.00	0.00	0.99	0.01
		Oracle	0.562	8.00	0.00	0.00	1.00	0.00	0.576	8.00	0.00	0.00	1.00	0.00
	200	lsLasso	0.674	179.58	0.00	0.00	0.00	1.00	0.731	180.01	0.00	0.00	0.00	1.00
		lsSCAD	0.579	190.60	0.00	0.00	0.09	0.91	0.588	191.16	0.00	0.00	0.13	0.87
		mSCAD	0.463	198.00	0.00	0.00	1.00	0.00	0.472	198.00	0.00	0.00	1.00	0.00
		Oracle	0.563	198.00	0.00	0.00	1.00	0.00	0.570	198.00	0.00	0.00	1.00	0.00
	500	lsLasso	0.694	473.86	0.00	0.00	0.00	1.00	0.762	471.45	0.00	0.00	0.01	0.99
		lsSCAD	0.579	488.12	0.00	0.00	0.07	0.93	0.608	485.89	0.00	0.00	0.06	0.94
		mSCAD	0.474	498.00	0.00	0.00	1.00	0.00	0.529	498.00	0.00	0.00	1.00	0.00
		Oracle	0.557	498.00	0.00	0.00	1.00	0.00	0.591	498.00	0.00	0.00	1.00	0.00

5. Proof

In this section, we outline the key idea of the proofs. Note the $c, c_1, c_2, \ldots$ denote generic positive constants. Their values may vary from expression to expression.

Lemma 1. Let $Y_1, \ldots, Y_n$ be independent random variables with zero mean such that $E|Y_i|^m \leq m!\, M^{m-2} v_i/2$, for every $m \geq 2$ (and all i), some constants M and $v_i = EY_i^2$. Let $v = v_1 + \cdots + v_n$, for $x > 0$,

$$\Pr(|\textstyle\sum_{i=1}^{n} \quad Y_i| > x) \leq 2\exp\left\{-\frac{x^2}{2(v+Mx)}\right\}.$$

Lemma 2. If there exists $\gamma \in \mathbb{R}^{p_n K_n}$ such that (i) $\sum_{i=1}^{n} \phi_h{}' (Y_i - \Pi_i^{\top}\gamma)\Pi_{il} = 0$ and $\|\gamma_l\| \geq a\lambda$ for $l = 1, \ldots, q_n$ and (ii) $\|\sum_{i=1}^{n} \phi_h{}' (Y_i - \Pi_i^{\top}\gamma)\Pi_{il}\| \leq n\lambda$ and $\|\gamma_l\| < \lambda$ for $l = q_n + 1, \ldots, p_n$, where $a = 3.7$, $\Pi_{il} = X_{il}\pi(U_i)$, then γ is a local maximizer of Equation 5.

This lemma is a direct extension of Theorem 1 in[28]. Thus, we omit the proof.

5.1. Proof of Theorem 1

Let γ_{0l} is the best approximating spline coefficient for $\beta_{0l}(u)$ and $R_l(u) = \pi(u)^\top \gamma_{0l} - \beta_{0l}(u),\ l = 1, \dots, q_n$. By the conditions (C1), (C3), (C5) and Corollary 6.21 in[32]., we have

$$\sup_{u\in[0,1]} |R_l(u)| = O(K_n^{-r}). \tag{11}$$

Note that

$$\begin{aligned}\tfrac{1}{n}\textstyle\sum_{l\in\mathcal{S}}\sum_{i=1}^{n}\{\hat{\beta}_l^o(U_i) - \beta_{0l}(U_i)\}^2 &= \tfrac{1}{n}\textstyle\sum_{l\in\mathcal{S}}\sum_{i=1}^{n}\{\pi(U_i)^\top(\hat{\gamma}_l^o - \gamma_{0l}) + R_l(U_i)\}^2\\ &\le \tfrac{2}{n}\textstyle\sum_{l\in\mathcal{S}}\sum_{i=1}^{n}\left\{\left(\pi(U_i)^\top(\hat{\gamma}_l^o - \gamma_{0l})\right)^2 + R_l^2(U_i)\right\}\\ &\le c_1 K_n^{-1}\|\hat{\gamma}_{\mathcal{S}}^o - \gamma_{0\mathcal{S}}\|^2 + c_2 q_n K_n^{-2r}.\end{aligned}$$

where in the last step above we used the well-know relation $\|\pi(t)^\top a\|^2 \sim K_n^{-1}\|a\|^2$ for any $a \in \mathbb{R}^{K_n}$. In order to get the rate of convergence, it sufficient to show that $\|\hat{\gamma}_{\mathcal{S}}^o - \gamma_{0\mathcal{S}}\| = O_p\left(q_n^{1/2} K_n n^{-1/2}\right)$.

Let $\delta_n = q_n^{1/2} K_n n^{-1/2}$ and μ be a vector, $\gamma_{\mathcal{S}} = \gamma_{0\mathcal{S}} + \delta_n \mu$. We want to show that for any given $\eta > 0$, there exists a large constant $c > 0$ such that

$$Pr\left\{\sup_{\|\mu\|=c} Q_h(\gamma_{0\mathcal{S}} + \delta_n\mu) < Q_h(\gamma_{0\mathcal{S}})\right\} \ge 1 - \eta. \tag{12}$$

This implies that, with probability at least $1 - \eta$, there is a maximum in the ball $\{\gamma_{0\mathcal{S}} + \delta_n\mu : \|\mu\| \le c\}$. Hence, their exists a local maximizer $\hat{\gamma}_{\mathcal{S}}^o$ such that $\|\hat{\gamma}_{\mathcal{S}}^o - \gamma_{0\mathcal{S}}\| = O_p\left(q_n^{1/2} K_n n^{-1/2}\right)$.

To prove (12), we consider the optimization problem in Equation 5 rewritten as

$$Q_h(\gamma_{\mathcal{S}}) = \textstyle\sum_{i=1}^{n} \quad \phi_h\left(\varepsilon_i - R_{i,\mathcal{S}}^* - \Pi_{i,\mathcal{S}}^\top(\gamma_{\mathcal{S}} - \gamma_{0\mathcal{S}})\right),$$

where $R_{i,\mathcal{S}}^* = \sum_{l\in\mathcal{S}} R_l(U_i) X_{il}$. By condition (C2) and the approximation of splines,

$$R_{i,\mathcal{S}}^* = O_P\left(q_n^{1/2} K_n^{-r}\right) \text{ and } \|R_{\mathcal{S}}^*\| = O\left(n^{1/2} q_n^{1/2} K_n^{-r}\right) \tag{13}$$

where $R_{\mathcal{S}}^* = \left(R_{1,\mathcal{S}}^*, \dots, R_{n,\mathcal{S}}^*\right)^\top$. Using the Taylor expansion, we have that

$$\begin{aligned} & Q_h(\gamma_{0\mathcal{S}} + \delta_n\mu) - Q_h(\gamma_{0\mathcal{S}})\\ = \quad & \delta_n\mu^\top \textstyle\sum_{i=1}^{n} \quad \phi_h'\left(\varepsilon_i - R_{i,\mathcal{S}}^*\right)\Pi_{i,\mathcal{S}} + \frac{1}{2}\delta_n^2\mu^\top \sum_{i=1}^{n} \quad \phi_h''(\zeta_i)\Pi_{i,\mathcal{S}}\Pi_{i,\mathcal{S}}^\top\mu\\ = \quad & I_1 + I_2\end{aligned}$$

where ζ_i is between $\varepsilon_i - R_{i,\mathcal{S}}^*$ and $\varepsilon_i - R_{i,\mathcal{S}}^* - \delta_n\Pi_{i,\mathcal{S}}^\top\mu$.

For I_1, using the Cauchy-Schwartz inequality, we have $|I_1| \le \delta_n\|\mu\|\left\|\sum_{i=1}^{n}\phi_h'\left(\varepsilon_i - R_{i,\mathcal{S}}^*\right)\Pi_{i,\mathcal{S}}\right\|$. Note that

$$\textstyle\sum_{i=1}^{n} \quad \phi_h'\left(\varepsilon_i - R_{i,\mathcal{S}}^*\right)\Pi_{i,\mathcal{S}} = \sum_{i=1}^{n} \quad \left\{\phi_h'(\varepsilon_i) - \phi_h''(\varepsilon_i)R_{i,\mathcal{S}}^* + \frac{1}{2}\phi_h'''(\varepsilon_i^*)R_{i,\mathcal{S}}^{*2}\right\}\Pi_{i,\mathcal{S}},$$

where ε_i^* is between ε_i and $\varepsilon_i - R_{i,\mathcal{S}}^*$. Invoking condition (C9), $\sum_{i=1}^{n}\phi_h'(\varepsilon_i) = O_p\left(n^{-1/2}\right)$. Then, from Equation 13, we have

$$\left\|\textstyle\sum_{i=1}^{n} \quad \phi_h'\left(\varepsilon_i - R_{i,\mathcal{S}}^*\right)\Pi_{i,\mathcal{S}}\right\| = O_p\left(n^{1/2} q_n^{1/2}\right)$$

Hence, we have $I_1 = O_p(q_n K_n \parallel \mu \parallel)$. For I_2, with the same argument, it is not difficult to prove that

$$I_2 = E\left(F(X, U, h)\right)O_p(q_n K_n \parallel \mu \parallel^2).$$

Therefore, by choosing a sufficiently large c, I_2 dominates I_1 uniformly $\parallel \mu \parallel = c$ for sufficiently large n. Note that I_2 is negative for sufficiently large c. Then the proof is completed.

5.2. Proof of Theorem 2

Let $\hat{\gamma} = \hat{\gamma}^o = \left(\hat{\gamma}_{\mathcal{S}}^{o\top}, 0_{(p_n - q_n)K_n}^\top\right)^\top$, we will show that $\hat{\gamma}$ satisfies equation (i)-(ii) of Lemma 2. This will immediately imply this theorem. Since $\hat{\gamma}_{\mathcal{S}}^o$ is the solution of the optimization problem Equation 4, we have

$$\textstyle\sum_{i=1}^{n} \phi_h'\left(Y_i - \Pi_{i,\mathcal{S}}^\top\hat{\gamma}_{\mathcal{S}}^o\right)\Pi_{il} = 0,\ l = 1, \dots, q_n \tag{14}$$

where $\Pi_{il} = X_{il}\pi(U_i)^\top$.

First, for $l = 1, \dots, q_n$, note that $\|\hat{\gamma}_l\| = \|\hat{\gamma}_l - \gamma_{0l} + \gamma_{0l}\| \ge min \ \ _{1\le l\le q_n}\|\gamma_{0l}\| - \|\hat{\gamma}_l - \gamma_{0l}\|$, then $\|\hat{\gamma}_l\| \ge a\lambda$ is implied by

$$\min_{1\le l\le q_n}\|\gamma_{0l}\| \gg \lambda,$$

and

$$\|\hat{\gamma}_l - \gamma_{0l}\| \ll \lambda,$$

and both equations above are implied by condition (C10) as well as Theorem 1. By Equation 14, it follows that (i) trivially hold since $\Pi_i^\top\hat{\gamma} = \Pi_{i,\mathcal{S}}^\top\hat{\gamma}_{\mathcal{S}}^o$.

Now it remains to show (ii). For $l = q_n + 1, \dots, p_n$, $\|\hat{\gamma}_l\| < \lambda$ is trivial since $\hat{\gamma}_l = 0$. Furthermore, by condition (C6), we have $\varepsilon_i = Y_i - \Pi_{i,\mathcal{S}}^{\top}\gamma_{0\mathcal{S}} + R_{i,\mathcal{S}}^{*}$. Then, by Taylor expansion, we have

and

$$\begin{aligned}\sum_{i=1}^{n}\phi_h'(Y_i - \Pi_i^{\top}\hat{\gamma})\Pi_{il} &= \sum_{i=1}^{n}\phi_h'\left(Y_i - \Pi_{i,\mathcal{S}}^{\top}\hat{\gamma}_{\mathcal{S}}^{o}\right)\Pi_{il}\\ &= \sum_{i=1}^{n}\phi_h'\left(\varepsilon_i - R_{i,\mathcal{S}}^{*} - \Pi_{i,\mathcal{S}}^{\top}(\hat{\gamma}_{\mathcal{S}}^{o} - \gamma_{0\mathcal{S}})\right)\Pi_{il}\\ &= \sum_{i=1}^{n}\phi_h'(\varepsilon_i)\Pi_{il} - \sum_{i=1}^{n}\phi_h''(\varepsilon_i)\Pi_{il}\left(\Pi_{i,\mathcal{S}}^{\top}(\hat{\gamma}_{\mathcal{S}}^{o} - \gamma_{0\mathcal{S}}) + R_{i,\mathcal{S}}^{*}\right)\{1 + o_p(1)\}\\ &= I_3 - I_4\{1 + o_p(1)\}.\end{aligned}$$

and

$$\left\|\sum_{i=1}^{n}\phi_h'(Y_i - \Pi_i^{\top}\hat{\gamma})\Pi_{il}\right\| \le \|I_3\| + \|I_4\|.$$

Next we will consider I_3 and I_4 respectively. For I_3, note that

$$\begin{aligned}\max_{1\le l\le p_n}\left\|\sum_{i=1}^{n}\phi_h'(\varepsilon_i)\Pi_{il}\right\|^2 &= \max_{1\le l\le p_n}\left\|\sum_{i=1}^{n}\phi_h'(\varepsilon_i)X_{il}\pi(U_i)^{\top}\right\|^2\\ &\max_{1\le l\le p_n}\sum_{k=1}^{K_n}\left|\sum_{i=1}^{n}\phi_h'(\varepsilon_i)X_{il}B_k(U_i)\right|^2\\ &\le K_n\max_{1\le l\le p_n,1\le k\le K_n}\left|\sum_{i=1}^{n}\phi_h'(\varepsilon_i)X_{il}B_k(U_i)\right|^2\end{aligned}$$

Let $T_{lk} = \sum_{i=1}^{n}\phi_h'(\varepsilon_i)X_{il}B_k(U_i)$ and $s_n^2 = \max_{1\le l\le p_n,1\le k\le K_n}\sum_{i=1}^{n}X_{il}^2B_k^2(U_i)$. By condition (C9), we can prove

$$E\left(\max_{1\le l\le p_n,1\le k\le K_n}\| T_{lk} \| X_{il}, U_i, 1\le l\le p_n, 1\le i\le n\right) \le c_1 s_n\sqrt{\log(p_nK_n)}$$

Therefore

$$E\left(\max_{1\le l\le p_n,1\le k\le K_n}|T_{lk}|\right) \le c_1\sqrt{\log(p_nK_n)}E(s_n) \tag{15}$$

By condition (C3) and the properties of B-splines, we have

$$\sum_{k=1}^{K_n} B_k(U_i) = 1 \text{ and } c_1K_n^{-1} \le E\left(B_k^2(U_i)\right) \le c_2K_n^{-1} \tag{16}$$

Thus, by condition (C2), we can get

$$\sum_{i=1}^{n} E\left[X_{il}^2B_k^2(U_i) - E\left(X_{il}^2B_k^2(U_i)\right)\right]^2 \le c_3nK_n^{-2} \tag{17}$$

and

$$\max_{1\le l\le p_n,1\le k\le K_n}\sum_{i=1}^{n} E\left(X_{il}^2B_k^2(U_i)\right) \le c_2nK_n^{-1} \tag{18}$$

By Lemma A. 1 of van de Geer (2008), (16) and (17) imply

$$E\left(\max_{1\le l\le p_n,1\le k\le K_n}\left|\sum_{i=1}^{n}\left[X_{il}^2B_k^2(U_i) - E\left(X_{il}^2B_k^2(U_i)\right)\right]\right|\right) = O\left(\sqrt{nK_n^{-2}\ln(p_nK_n)} + \ln(p_nK_n)\right) \tag{19}$$

Therefore, from (18) and the triangle inequality

$$E(s_n^2) = O\left(\sqrt{nK_n^{-2}\ln(p_nK_n)} + \ln(p_nK_n) + nK_n^{-2}\right) = O(\ln(p_nK_n) + nK_n^{-2})$$

Now since $E(s_n) \le \sqrt{E(s_n^2)}$, we have

$$E(s_n) = O\left(\sqrt{\ln(p_nK_n) + nK_n^{-2}}\right) \tag{20}$$

From (15) and (20), we have

$$E\left(\max_{1\le l\le p_n,1\le k\le K_n}|T_{lk}|\right) \le O\left(\log(p_nK_n) + \sqrt{nK_n^{-2}\ln(p_nK_n)}\right)$$

Integrate the above discussion, for I_3, we have

$$\max_{1\le l\le p_n}\left\|\sum_{i=1}^{n}\phi_h'(\varepsilon_i)\Pi_{il}\right\| = O\left(\sqrt{nK_n^{-1}\ln(p_nK_n)}\right) \tag{21}$$

For I_4, firstly it is easy to see (using Theorem 1) that

$$\left|\Pi_{i,\mathcal{S}}^{\top}(\hat{\gamma}_{\mathcal{S}}^{o} - \gamma_{0\mathcal{S}})\right| \asymp K_n^{-1}\|\hat{\gamma}_{\mathcal{S}}^{o} - \gamma_{0\mathcal{S}}\| = O_p\left(q_n^{1/2}n^{-1/2}\right)$$

and

$$\left\|R_{i,\mathcal{S}}^{*}\right\| = O_p\left(q_n^{1/2}K_n^{-r}\right) \tag{22}$$

Let $a = \left(\phi_h''(\varepsilon_1), \dots, \phi_h''(\varepsilon_n)\right)^{\top}$ and $b_{lk} = \left(b_{lk,1}, \dots, b_{lk,n}\right)^{\top}$ with $b_{lk,i} = X_{il}B_k\left(U_i\right)\left(\Pi_{i,\mathcal{S}}^{\top}(\hat{\gamma}_{\mathcal{S}}^{o} - \gamma_{0\mathcal{S}}) + R_{i,\mathcal{S}}^{*}\right)$, and

we have $\max_i |b_{lk,i}| \le cq_n^{1/2}\left(n^{-1/2}+K_n^{-r}\right) \le cq_n^{1/2}K_n^{-r}$. Then

$$\begin{aligned} I_4 &= \sum_{i=1}^n \phi_h''(\varepsilon_i)\Pi_{il}\left(\Pi_{i,\mathcal{S}}^\top(\hat{\gamma}_{\mathcal{S}}^o-\gamma_{0\mathcal{S}})+R_{i,\mathcal{S}}^*\right) \\ &= \sum_{i=1}^n \phi_h''(\varepsilon_i)X_{il}\pi(U_i)^\top\left(\Pi_{i,\mathcal{S}}^\top(\hat{\gamma}_{\mathcal{S}}^o-\gamma_{0\mathcal{S}})+R_{i,\mathcal{S}}^*\right) \\ &= \left(a^\top b_{l1}, a^\top b_{l2}, \dots, a^\top b_{lK_n}\right)^\top. \end{aligned}$$

Note that

$$\max_{1\le l\le p_n}\left\|\sum_{i=1}^n \phi_h''(\varepsilon_i)\Pi_{il}\left(\Pi_{i,\mathcal{S}}^\top(\hat{\gamma}_{\mathcal{S}}^o-\gamma_{0\mathcal{S}})+R_{i,\mathcal{S}}^*\right)\right\|^2 \le K_n \max_{1\le l\le p_n,1\le k\le K_n}|a^\top b_{lk}|^2$$

By condition (C9), we have $E|\phi_h''(\varepsilon_i)|^m \le \frac{m!}{2}S^2T^{m-2}$,$m = 2,3,\dots$, for some constants S and T. Then we have

$$E|a_i b_{lk,i}|^m \le \frac{m!}{2}\left(b_{lk,i}S\right)^2\left(b_{lk,i}T\right)^{m-2} \le \frac{m!}{2}\left(b_{lk,i}S\right)^2\left(cq_n^{1/2}K_n^{-r}T\right)^{m-2}$$

and

$$\sum_{i=1}^n E|a_{lk,i}b_i|^2 \le S^2\sum_{i=1}^n b_{lk,i}^2 \le c^2S^2nq_nK_n^{-2r}$$

By Lemma 1 and a simple union bound

$$\begin{aligned} P\left(\max_{1\le l\le p_n,1\le k\le K_n}|a^\top b_{lk}|>\epsilon\right) &= P\left(\max_{1\le l\le p_n,1\le k\le K_n}\left|\sum_{i=1}^n a_i b_{lk,i}\right|>\epsilon\right) \\ &\le 2p_n\exp\left\{-\frac{\epsilon^2}{2nc^2S^2q_nK_n^{-2r}+2cq_n^{1/2}K_n^{-r}T\epsilon}\right\} \end{aligned}$$

Taking $\epsilon = c_1\sqrt{nq_n}K_n^{-r}\log(p_n\vee n)$ for some c_1 large enough, the above probability tends to zero, thus we have

$$\|I_4\| \le O_p\left(\frac{nq_n}{K_n^{2r-1}}\log(p_n\vee n)\right)$$

From condition (C10), we prove (ii) in Lemma 2. This completes the proof.

6. Conclusion

The varying coefficient model is flexible and powerful for modeling the dynamic changes of regression coefficients. It is important to identify significant covariates associated with response variables, especially for high dimensional settings. It has received considerable attention during the past decades. However, many papers were built on either least square or likelihood based methods, which are expected to be very sensitive to outliers and their efficiency may be significantly reduced for many commonly used non-normal errors. Due to the well-known advantages of modal regression, these facts motivate us to extend the modal regression method to the varying coefficient model in high dimensions.

This paper has investigated the robust estimator of the varying coefficient models based on modal regression when the number of nonparametric components $\beta(u)$ diverges with sample size increasing. The varying coefficients are approximated by B-splines and the relevant variables are selected automatically by the SCAD penalty. Theoretically, the oracle theory was derived under mild conditions. In addition, a computation algorithm is developed based on local quadratic approximation and EM algorithm. Some Monte Carlo simulations are provided to corroborate our theoretical findings in finite samples.

There are some several possible extensions that further study. First, our approach described in this paper can be easily extended to other models, such as the partially linear single index model and the partially linear additive model. Second, a challenging problem, particularly for high dimensional data, is how to identify which covariates are parametric or nonparametric terms. Another problem of practical interest is to construct prediction intervals based on the observed date. Last, it would be interesting to take into account complex data in high dimensional semiparametric models, such as missing data, measurement error data, censored data.

Funding Statement

This work was supported by the Humanities and Social Sciences Research Projects of the Ministry of Education of China (20YJC910010), and the Doctoral Foundation of Henan Polytechnic University (B2020-37).

References

[1] Yehuda Koren, Robert Bell, and Chris Volinsky, "Matrix Factorization Techniques for Recommender Systems," *Computer*, vol. 42, no. 8, pp. 30-37, 2009. [Google Scholar] [Publisher Link]

[2] Yongdai Kim, Hosik Choi, and Hee-Seok Oh, "Smoothly Clipped Absolute Deviation on High Dimensions," *Journal of the American Statistical Association*, vol. 103, no. 484, pp. 1665-1673, 2008. [CrossRef] [Google Scholar] [Publisher Link]

[3] Laura J. van't Veer et al., "Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer," *Nature*, vol. 415, pp. 530-536, 2002. [CrossRef] [Google Scholar] [Publisher Link]

[4] Jianqing Fan, and Runze Li, "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348-1360, 2001. [CrossRef] [Google Scholar] [Publisher Link]

[5] Robert Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 58, no. 1, pp. 267-288, 1996. [CrossRef] [Google Scholar] [Publisher Link]

[6] Cun-Hui Zhang, "Nearly Unbiased Variable Selection under Minimax Concave Penalty," *The Annals of Statistics*, vol. 38, no. 2, pp. 894-942, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[7] Jianqing Fan, and Jinchi Lv, "A Selective Overview of Variable Selection in High Dimensional Feature Space," *Statistica Sinica*, vol. 20, no. 1, pp. 101-148, 2010. [Google Scholar] [Publisher Link]

[8] Peter Bühlmann, and Sara van de Geer, *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer Berlin Heidelberg, 2011. [Google Scholar] [Publisher Link]

[9] Peter J. Bickel et al., *Efficient and Adaptive Estimation for Semiparametric Models*, New York: Springer, 1998. [Google Scholar] [Publisher Link]

[10] Trevor Hastie, and Robert Tibshirani, "Varying Coefficient Models," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 55, no. 4, pp. 757-796, 1993. [CrossRef] [Google Scholar] [Publisher Link]

[11] Jianqing Fan, and Wenyang Zhang, "Statistical Methods with Varying Coefficient Models," *Statistics and Its Interface*, vol. 1, no. 1, pp. 179-195, 2008. [CrossRef] [Google Scholar] [Publisher Link]

[12] Jianqing Fan, and Wenyang Zhang, "Statistical Estimation in Varying Coefficient Models," *The Annals of Statistics*, vol. 27, no. 5, pp. 1491-1518, 1999. [CrossRef] [Google Scholar] [Publisher Link]

[13] Jianqing Fan, and Jin-Ting Zhang, "Two-Step Estimation of Functional Linear Models with Applications to Longitudinal Data," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 62, no. 2, pp. 303-322, 2000. [CrossRef] [Google Scholar] [Publisher Link]

[14] Chin-Tsang Chiang, John A. Rice, and Colin O. Wu, "Smoothing Spline Estimation for Varying Coefficient Models with Repeatedly Measured Dependent Variables," *Journal of the American Statistical Association*, vol. 96, no. 454, pp. 605-619, 2001. [CrossRef] [Google Scholar] [Publisher Link]

[15] Jianhua Z. Huang, Colin O. Wu, and Lan Zhou, "Varying-coefficient Models and Basis Function Approximations for the Analysis of Repeated Measurements," *Biometrika*, vol. 89, no. 1, pp. 111-128, 2002. [CrossRef] [Google Scholar] [Publisher Link]

[16] Jianhua Z. Huang, Colin O. Wu, and Lan Zhou "Polynomial Spline Estimation and Inference for Varying Coefficient Models with Longitudinal Data," *Statistica Sinica*, vol. 14, no. 3, pp. 763-788, 2004. [Google Scholar] [Publisher Link]

[17] R.L. Eubank et al., "Smoothing Spline Estimation in Varying Coefficient Models," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 66, no. 3, pp. 653-667, 2004. [CrossRef] [Google Scholar] [Publisher Link]

[18] Lifeng Wang, Hongzhe Li, and Jianhua Z. Huang, "Variable Selection in Nonparametric Varying Coefficient Models for Analysis of Repeated Measurements," *Journal of the American Statistical Association*, vol. 103, no. 484, pp. 1556-1569, 2008. [CrossRef] [Google Scholar] [Publisher Link]

[19] Hansheng Wang, and Yingcun Xia, "Shrinkage Estimation of the Varying Coefficient Model," *Journal of the American Statistical Association*, vol. 104, no. 486, pp. 747-757, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[20] Fengrong Wei, and Jian Huang and Hongzhe Li, "Variable Selection and Estimation in High-Dimensional Varying-Coefficient Models," *Statistica Sinica*, vol. 21, no. 4, pp. 1515-1540, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[21] Weixin Yao, Bruce G. Lindsay, and Runze Li, "Local Modal Regression," *Journal of Nonparametric Statistics*, vol. 24, no. 3, pp. 647-663, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[22] Weixin Yao, and Longhai Li, "A New Regression Model: Modal Linear Regression," *Scandinavian Journal of Statistics*, vol. 41, no. 3, pp. 656-671, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[23] Riquan Zhang, Weihua Zhao, and Jicai Liu, "Robust Estimation and Variable Selection for Semiparametric Partially Linear Varying Coefficient Model Based on Modal Regression," *Journal of Nonparametric Statistics*, vol. 25, no. 2, pp. 523-544, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[24] Jicai Liu et al., "A Robust and Efficient Estimation Method for Single Index Models," *Journal of Multivariate Analysis*, vol. 122, pp. 226-238, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[25] Hu Yang, Chaohui Guo, and Jing Lv, "A Robust and Efficient Estimation Method for Single-index Varyig Coefficient Models," *Statistics and Probability Letters*, vol. 94, pp. 119-127, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[26] Jianqing Fan, *Local Polynomial Modeling and Its Applications, Monographs on Statistics and Applied Probability 66*, 1st ed., Routledge, 1996. [CrossRef] [Google Scholar] [Publisher Link]

[27] Carl de Boor, *A Practical Guide to Splines*, Springer-Verlag New York, 2001. [Google Scholar] [Publisher Link]

[28] Jianqing Fan, and Jinchi Lv, "Nonconcave Penalized Likelihood with NP-dimensionality," *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 5467-5484, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[29] Cun-Hui Zhang, and Tong Zhang, "A General Theory of Concave Regularization for High Dimensional sparse estimation problems," *Statistical Science*, vol. 27, no. 4, pp. 576-593, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[30] Jia Li, Surajit Ray, and Bruce G. Lindsay, "A Nonparametric Statistical Approach to Clustering via Mode Identification," *Journal of Machine Learning Research*, vol. 8, no. 4, pp. 1687-1723, 2007. [Google Scholar] [Publisher Link]

[31] Jiahua Chen, and Zehua Chen, "Extended Bayesian Information Criteria for Model Selection with Large Model Spaces," *Biometrika,* vol. 95, no. 3, pp. 759-771, 2008. [CrossRef] [Google Scholar] [Publisher Link]

[32] Larry L. Schumaker, *Spline Functions: Basic Theory*, Wiley, 1981. [Google Scholar] [Publisher Link]