| Statistica Sinica Preprint No: SS-2020-0170 | |
|---|---|
| **Title** | High-dimensional Varying Index Coefficient Quantile Regression Model |
| **Manuscript ID** | SS-2020-0170 |
| **URL** | http://www.stat.sinica.edu.tw/statistica/ |
| **DOI** | 10.5705/ss.202020.0170 |
| **Complete List of Authors** | Jing Lv and Jialiang Li |
| **Corresponding Author** | Jialiang Li |
| **E-mail** | stalj@nus.edu.sg |

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170

Statistica Sinica

# High-dimensional Varying Index Coefficient Quantile Regression Model

Jing Lv[1] and Jialiang Li[2*]

[1]*School of Mathematics and Statistics, Southwest University, Chongqing, China*

[2]*Department of Statistics and Applied Probability, National University of Singapore, Singapore*

*Abstract:* Statistical learning is evolving quickly, with increasingly sophisticated models seeking to incorporate the complicated data structures from modern scientific and business problems. Varying-index coefficient models extend varying-coefficient models and single-index models for semiparametric regressions. This new class of model offers greater flexibility in terms of characterizing complicated nonlinear interaction effects in a regression analysis. To safeguard against outliers and extreme observations, we consider a robust quantile regression approach to estimate the model parameters. High-dimensional loading parameters are allowed in our development, under reasonable theoretical conditions. Thus, we propose a regularized estimation procedure to select the significant nonzero loading parameters, identify linear functions in varying-index coefficient models, and consistently estimate the parametric and nonparametric components. Under some technical assumptions, we show that the proposed procedure is consistent

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170



in terms of variable selection and linear function identification, and that the proposed parameter estimation enjoys the oracle property. Extensive simulation studies are carried out to assess the finite-sample performance of the proposed method. We illustrate our methods using an example based on New Zealand workforce data.

*Key words and phrases:* High-dimensional data, Penalty, Quantile regression, Semiparametric regression, Varying index coefficient model.

## 1. Introduction

Semiparametric regression models are powerful statistical learning approaches that are popular in scientific and business research studies because they enjoy the merits of both parametric and nonparametric models. We consider the varying-index coefficient model (VICM) proposed by Ma and Song (2015). This new class of model extends varying-coefficient models (Fan and Zhang (1999)), single-index models (Xia et al. (2002)), single-index coefficient models (Xue and Wang (2012)), and almost all other familiar semiparametric models. To safeguard against outliers and extreme observations, we consider a robust quantile regression (QR) approach to fit the VICM. Specifically, for a given quantile level $\tau \in (0, 1)$, varying-index coefficient QR models are given by

$$Q_\tau\left(Y|\boldsymbol{X}, \boldsymbol{Z}\right) = \sum_{l=1}^{d} m_{\tau,l}(\boldsymbol{Z}^T\boldsymbol{\beta}_{\tau,l})X_l, \tag{1.1}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170



where $\boldsymbol{X} = (X_1, ..., X_d)^T$, $X_1 \equiv 1$, $\boldsymbol{Z} = (Z_1, ..., Z_p)^T$ are covariates for the response variable $Y \in \mathbb{R}$, $\boldsymbol{\beta}_{\tau,l} = (\beta_{\tau,l1}, ..., \beta_{\tau,lp})^T$ are unknown loading parameters for the $l$th covariate $X_l$, and $m_{\tau,l}(\cdot)$ are unknown nonparametric functions, for $l = 1, ..., d$. Let $\varepsilon_\tau = Y - Q_\tau(Y|\boldsymbol{X}, \boldsymbol{Z})$ be the model error with an unspecified conditional density function $f_{\varepsilon_\tau}(\cdot|\boldsymbol{X}, \boldsymbol{Z})$ and a conditional cumulative distribution function $F_{\varepsilon_\tau}(\cdot|\boldsymbol{X}, \boldsymbol{Z})$ of $\varepsilon_\tau$ given $(\boldsymbol{X}, \boldsymbol{Z})$. In the remainder of the paper, we drop the subscript $\tau$ from $\boldsymbol{\beta}_{\tau,l}$, $m_{\tau,l}(\cdot)$, $\varepsilon_\tau$, $f_{\epsilon_\tau}(\cdot|\boldsymbol{X}, \boldsymbol{Z})$, and $F_{\varepsilon_\tau}(\cdot|\boldsymbol{X}, \boldsymbol{Z})$ to simplify the notation. However, it is helpful to bear in mind that these quantities are $\tau$-specific. Note that $\varepsilon$'s conditional $\tau$th quantile is equal to zero; that is, $P(\varepsilon \leq 0|\boldsymbol{X}, \boldsymbol{Z}) = F_\varepsilon(0|\boldsymbol{X}, \boldsymbol{Z}) = \tau$. For the sake of identifiability, we assume that $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^T, ..., \boldsymbol{\beta}_d^T\right)^T$ belongs to the following parameter space:

$$\Theta = \left\{\boldsymbol{\beta} = \left(\boldsymbol{\beta}_l^T : 1 \leq l \leq d\right)^T : \|\boldsymbol{\beta}_l\|_2 = 1, \beta_{l1} > 0, \boldsymbol{\beta}_l \in \mathbb{R}^p\right\},$$

where $\|.\|_2$ denotes the $L_2$ norm such that $\|\boldsymbol{\xi}\|_2 = (\xi_1^2 + \cdots + \xi_s^2)^{1/2}$, for any vector $\boldsymbol{\xi} = (\xi_1, ..., \xi_s)^T \in \mathbb{R}^s$. Model (1.1) is quite general, and includes many other existing models as special cases: (i) when $m_l(\cdot)$ is assumed to be constant or a linear function, it reduces to the linear regression model with interactions; (ii) when $d = 1$ and $X_l = 1$, it is the single-index model; (iii) when $m_l(\cdot)$ is constant for $l \geq 2$ and $X_1 = 1$, it is the partial linear single-index model; (iv) when the common coefficient vector $\boldsymbol{\beta}_l$ is used, it

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170



is the single-index coefficient model; and (v) when we consider $\boldsymbol{X} = \boldsymbol{Z}$ and the common coefficient vector $\boldsymbol{\beta}_l$, it reduces to the well-known adaptive varying-coefficient model.

The VICM offers a flexible way to model and assess the nonlinear interaction effects between the covariates $\boldsymbol{X}$ and $\boldsymbol{Z}$. Note that the choice of these specialized model forms may depend on the application. For example, in econometric studies, it is often of interest to summarize the effects of multiple input variables within a single *variable*, and then to perform a regression analysis on the combined *variable* and other ordinary variables. The well-known capital asset pricing model (CAPM) and the Fama-French three-factor model both introduce derived *variables* in their model representations. Such *variables* may invoke linear or nonlinear interactions in the regression function with other variables, even those used to create the index *variables*. Experienced data analysts may suggest that predictors be properly used in different components of $\boldsymbol{X}$ and $\boldsymbol{Z}$. How to design $\boldsymbol{X}$ and $\boldsymbol{Z}$ objectively from data remains an interesting question. However, fully addressing this question is very challenging, and beyond the scope of this study. Thus, similarly to traditional studies on index models, we assume that $\boldsymbol{X}$ and $\boldsymbol{Z}$ are given in the data set, and that there is no overlapped term between $\boldsymbol{X}$ and $\boldsymbol{Z}$. Our main interest is to perform statistical infer-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170



ences on both the loading coefficients $\boldsymbol{\beta}_l$ and the nonparametric functions $m_l(\cdot)$, for $l = 1, ..., d$.

Ma and Song (2015) proposed a profile least squares estimation procedure for the VICM and established its theoretical properties. Their work focused on a mean regression, which is suitable for nicely distributed data, such as Gaussian data, but may perform badly in the presence of outliers and heavy-tailed errors. Our model (1.1) imposes different assumptions on the error structure and, thus, produces a novel and robust framework applicable to a wider variety of applications. The estimation methods and the associated asymptotic theories are thus different to those of Ma and Song (2015).

Since the seminal work of Koenker and Bassett (1978), QRs have emerged as an important alternative to the mean regression. It is well known that an inference based on a QR is more robust against distribution contamination (Koenker (2005)). A full range of quantile analyses can provide a more complete description of the conditional distribution. It is now widely acknowledged that an analysis based on a QR may lead to more appropriate findings. For example, climatologists often pay close attention to how the high quantiles of tropical cyclone intensity change over time (Elsner et al. (2008)), because these generate strong winds and waves, often resulting in

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170



heavy rain and storm surges. In the health sciences, medical scientists often study the effects of maternal behaviors on the low quantiles of birth-weight distributions (Abrevaya (2001)). Furthermore, in business and economics, petroleum is a primary source of nonrenewable energy, and has important effects on industrial production, electric power generation, and transportation (Marimoutou et al. (2009)). Thus, most analysts focus on the high quantiles of oil prices, because oil price fluctuations have considerable effects on economic activity. The QR framework considered in this study may affect all of these fields, where a direct application of a mean regression is inappropriate.

Another important contribution of this study is that we consider high-dimensional learning issues for the quantile VICM. In fact, recent advances in technologies for cheaper and faster data acquisition and storage have led to explosive growth in data complexity in a variety of scientific areas, such as medicine, economics, and environmental science. We have to consider a realistic solution to the "large $n$, diverging $p$" data setting. Specifically, we allow the dimension of the covariates $\boldsymbol{Z}$ to increase to infinity as the sample size increases. Many penalty-based estimation methods have been proposed to address the high-dimensional issue (Fan and Li (2006); Giraud (2015); Hastie et al. (2015)). This framework can effectively reduce the model bias

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170



and improve the prediction performance of the fitted model. Fan and Peng (2004) first studied the nonconcave penalized likelihood estimation when the number of covariates increases with the sample size. Later, Wang et al. (2012) extended their method to generalized linear models for longitudinal measurements. The high-dimensional issue has also been investigated for semiparametric models. Wang and Wang (2015) applied the smoothly clipped absolute deviations (SCAD) penalty to perform variable selection for single-index prediction models with a diverging number of index parameters. Fan et al. (2017) presented a penalized empirical likelihood approach for high-dimensional semiparametric models.

Variable selection for model (1.1) is challenging, because the high-dimensional loading parameter is structured within the unknown nonparametric function coefficients. We adopt a spline basis approximation for the estimation of $m_l(\cdot)$, and estimate the unknown vector of the loading parameters $\boldsymbol{\beta}_l$ under the sparsity assumption. In addition, we correctly identify the linear interaction effects between the covariates. That is, we want to decide whether it is necessary to model $m_l(\cdot)$ nonparametrically for all $d$ varying index functions. Ma and Song (2015) constructed a generalized likelihood ratio statistic to test whether there exists a linear interaction effect between covariates. Although this test approach works very well for

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170



low-dimensional problems, it is computationally infeasible when the number of covariates is large. To this end, we develop a group penalization method that can quickly and effectively differentiate linear functions from nonparametric functions. The theoretical justification is also nontrivial for this complicated setting.

## 2. QR Estimation of Functions and Loadings

### 2.1 Estimation procedures

Suppose that $\{(\boldsymbol{X}_i, \boldsymbol{Z}_i, Y_i), 1 \leq i \leq n\}$ is an independent and identically distributed (i.i.d.) sample from model (1.1). Similarly to Wang and Wang (2015), we assume that each $Z_{ik}$, for $i = 1, ..., n, k = 1, ..., p$ takes a value in $[a, b]$, where $a$ and $b$ are some finite numbers. B-spline basis functions are commonly used to approximate the unknown smooth functions, owing to their desirable numerical stability in practice (de Boor (2001)). We thus adopt a nonparametric approach to estimate the index functions. More specifically, let $\boldsymbol{B}(u) = (B_s(u) : 1 \leq s \leq J_n)^T$ be a set of normalized B-spline basis functions of order $q$ ($q \geq 2$) with $N_n$ internal knots and $J_n = q + N_n$. We then approximate $m_l(\cdot)$ using a linear combination of B-spline basis functions $m_l(\cdot) \approx \boldsymbol{B}(\cdot)^T \boldsymbol{\lambda}_l$, where $\boldsymbol{\lambda} = \left(\boldsymbol{\lambda}_1^T, ..., \boldsymbol{\lambda}_d^T\right)^T$ is the spline coefficient vector with $\boldsymbol{\lambda}_l = (\lambda_{ls} : 1 \leq s \leq J_n)^T$, for $l = 1, ..., d$.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170



Let $\rho_\tau(u) = u\{\tau - I(u \leq 0)\}$ be the quantile loss function, where $I(\cdot)$ is an indicator function. We obtain the estimators of the spline coefficients $\boldsymbol{\lambda}$ and the loading parameters $\boldsymbol{\beta}$ by minimizing

$$\mathcal{L}_{\tau n}(\boldsymbol{\lambda}, \boldsymbol{\beta}) = \sum_{i=1}^{n} \rho_\tau \left\{ Y_i - \sum_{l=1}^{d} \boldsymbol{B}\left(\boldsymbol{Z}_i^T \boldsymbol{\beta}_l\right)^T \boldsymbol{\lambda}_l X_{il} \right\}, \tag{2.1}$$

subject to the constraints $\|\boldsymbol{\beta}_l\|_2 = 1$ and $\beta_{l1} > 0$. Minimizing (2.1) with respect to all unknown quantities requires nonstandard nonlinear programming, and the solution is usually difficult to obtain directly. To address this computing difficulty, we adopt the profile iterative procedure to estimate $\boldsymbol{\beta}_l$ and $m_l(\cdot)$. The detailed steps are given below.

*Step 0.* Initialization step: Obtain an initial value $\hat{\boldsymbol{\beta}}^{(0)}$, with $\left\|\hat{\boldsymbol{\beta}}^{(0)}\right\|_2 = 1$. Further details on how to generate the initial values can be found in Appendix A of the Supplemental Material.

*Step 1.* For a given $\boldsymbol{\beta}$, $\hat{\boldsymbol{\lambda}}(\boldsymbol{\beta})$ can be attained using $\hat{\boldsymbol{\lambda}}(\boldsymbol{\beta}) = \underset{\boldsymbol{\lambda} \in \mathbb{R}^{dJ_n}}{\arg\min}\, \mathcal{L}_{\tau n}(\boldsymbol{\lambda}, \boldsymbol{\beta})$. This leads to $\hat{m}_l(\cdot, \boldsymbol{\beta}) = \boldsymbol{B}(\cdot)^T \hat{\boldsymbol{\lambda}}_l(\boldsymbol{\beta})$, for $l = 1, ..., d$. According to (de Boor (2001), page 116), the first-order derivative $\dot{m}_l(\cdot)$ can be approximated by the spline functions of one order lower than that of $m_l(\cdot)$. Then, we have $\hat{\dot{m}}_l(\cdot, \boldsymbol{\beta}) = \dot{\boldsymbol{B}}(\cdot)^T \hat{\boldsymbol{\lambda}}_l(\boldsymbol{\beta})$, where $\dot{\boldsymbol{B}}$ is the first-order derivative of $\boldsymbol{B}$.

The parameter space $\Theta$ specifies that $\boldsymbol{\beta}$ lies on the boundary of a unit ball. Therefore, for a given $\boldsymbol{\lambda}$, the function $\mathcal{L}_{\tau n}(\boldsymbol{\lambda}, \boldsymbol{\beta})$ is not differentiable at point $\boldsymbol{\beta}$. To handle this constraint, we employ the "remove-one-component"

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170



method to change the restricted QR to an unrestricted QR. Specifically, for $\boldsymbol{\beta}_l = (\beta_{l1}, \beta_{l2}, ..., \beta_{lp})^T$, let $\boldsymbol{\beta}_{l,-1} = (\beta_{l2}, ..., \beta_{lp})^T$ be a $(p-1)$-dimensional parameter vector after removing $\beta_{l1}$ in $\boldsymbol{\beta}_l$. Then, $\boldsymbol{\beta}_l$, for $l = 1, ..., d$, can be rewritten as

$$\boldsymbol{\beta}_l = \boldsymbol{\beta}_l(\boldsymbol{\beta}_{l,-1}) = (\sqrt{1 - \|\boldsymbol{\beta}_{l,-1}\|_2^2}, \boldsymbol{\beta}_{l,-1}^T)^T, \quad \|\boldsymbol{\beta}_{l,-1}\|_2^2 < 1. \tag{2.2}$$

It is obvious that $\boldsymbol{\beta}_l$ is infinitely differentiable with respect to $\boldsymbol{\beta}_{l,-1}$, and the Jacobian matrix is given by

$$\boldsymbol{J}_l\left(\boldsymbol{\beta}_{l,-1}\right) = \frac{\partial \boldsymbol{\beta}_l}{\partial \boldsymbol{\beta}_{l,-1}^T} = \begin{pmatrix} -\boldsymbol{\beta}_{l,-1}^T \Big/ \sqrt{1 - \|\boldsymbol{\beta}_{l,-1}\|_2^2} \\ \boldsymbol{I}_{p-1} \end{pmatrix},$$

where $\boldsymbol{I}_p$ is a $p \times p$ identity matrix. Denote $\boldsymbol{\beta}_{-1} = \left(\boldsymbol{\beta}_{1,-1}^T, ..., \boldsymbol{\beta}_{d,-1}^T\right)^T$. Then, $\boldsymbol{\beta}_{-1}$ belongs to

$$\Theta_{-1} = \left\{\boldsymbol{\beta}_{-1} = \left(\boldsymbol{\beta}_{l,-1}^T : 1 \le l \le d\right)^T : \|\boldsymbol{\beta}_{l,-1}\|_2^2 < 1, \boldsymbol{\beta}_{l,-1} \in \mathbb{R}^{p-1}\right\}.$$

*Step 2.* Let $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{\beta}_{-1})$, with the aforementioned definition $\boldsymbol{\beta}_l = \boldsymbol{\beta}_l(\boldsymbol{\beta}_{l,-1})$, for $1 \le l \le d$. Based on the estimators $\hat{\boldsymbol{\lambda}}_l$, $\hat{m}_l$, and $\hat{\dot{m}}_l$ from Step 1, we can construct the QR estimating equations for $\boldsymbol{\beta}_{-1}$ by setting $\partial \mathcal{L}_{\tau n}\left(\hat{\boldsymbol{\lambda}}(\boldsymbol{\beta}), \boldsymbol{\beta}\right) \Big/ \partial \boldsymbol{\beta}_{-1} = \mathbf{0}$. However, the equations involve a non-smooth function $\psi_\tau(u) = \dot{\rho}_\tau(u) = \tau - I(u \le 0)$. This adds difficulty to the computation, despite there being a linear programming solver (e.g.,

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170



Jin et al. (2003)). We circumvent this problem by smoothing the function $\partial \mathcal{L}_{\tau n}\left(\hat{\boldsymbol{\lambda}}(\boldsymbol{\beta}), \boldsymbol{\beta}\right) \Big/ \partial \boldsymbol{\beta}_{-1}$, that is, by replacing $\psi_\tau(\cdot)$ with a smooth function $\psi_{\tau h}(\cdot)$ (Whang (2006)). For this purpose, we introduce $G_h(x) = G(x/h)$, where $G(x) = \int_{u<x} K(u)du$, $K(\cdot)$ is a kernel function and $h$ is a bandwidth. Then, we construct the approximation function $\psi_{\tau h}(\cdot) = \tau - 1 + G_h(\cdot)$, and the smoothed estimating equations for $\boldsymbol{\beta}_{-1}$ are given as

$$\begin{aligned}\mathcal{R}_{\tau nh}(\boldsymbol{\beta}_{-1}) = &-\sum_{i=1}^{n} \psi_{\tau h}\left\{Y_i - \sum_{l=1}^{d} \boldsymbol{B}\left(\boldsymbol{Z}_i^T \boldsymbol{\beta}_l\right)^T \hat{\boldsymbol{\lambda}}_l(\boldsymbol{\beta}) X_{il}\right\} \\ &\times \begin{bmatrix} \hat{\dot{m}}_1\left(\boldsymbol{Z}_i^T \boldsymbol{\beta}_1, \boldsymbol{\beta}\right) X_{i1} \boldsymbol{J}_1^T \boldsymbol{Z}_i + \left(\partial \hat{\boldsymbol{\lambda}}(\boldsymbol{\beta})^T \Big/ \partial \boldsymbol{\beta}_{1,-1}\right) \boldsymbol{D}_i(\boldsymbol{\beta}) \\ \vdots \\ \hat{\dot{m}}_d\left(\boldsymbol{Z}_i^T \boldsymbol{\beta}_d, \boldsymbol{\beta}\right) X_{id} \boldsymbol{J}_d^T \boldsymbol{Z}_i + \left(\partial \hat{\boldsymbol{\lambda}}(\boldsymbol{\beta})^T \Big/ \partial \boldsymbol{\beta}_{d,-1}\right) \boldsymbol{D}_i(\boldsymbol{\beta}) \end{bmatrix} = \mathbf{0}, \end{aligned} \tag{2.3}$$

where $\boldsymbol{D}_i(\boldsymbol{\beta}) = (D_{i,sl}(\boldsymbol{\beta}_l), 1 \le s \le J_n, 1 \le l \le d)^T$, with $D_{i,sl}(\boldsymbol{\beta}_l) = B_s(\boldsymbol{Z}_i^T \boldsymbol{\beta}_l) X_{il}$. Then, we employ the Fisher scoring algorithm to obtain the estimates,

$$\boldsymbol{\beta}_{-1}^{(k+1)} = \boldsymbol{\beta}_{-1}^{(k)} - \left[\partial \mathcal{R}_{\tau nh}(\boldsymbol{\beta}_{-1}) / \partial \boldsymbol{\beta}_{-1}^T\right]^{-1} \mathcal{R}_{\tau nh}(\boldsymbol{\beta}_{-1}) \big|_{\boldsymbol{\beta}_{-1} = \boldsymbol{\beta}_{-1}^{(k)}}. \tag{2.4}$$

*Step 3.* Repeat Steps 1 and 2 until convergence, and denote the final estimators as $\hat{\boldsymbol{\beta}}_{-1}$ and $\hat{\boldsymbol{\lambda}}$. Then, we apply formula (2.2) to obtain $\hat{\boldsymbol{\beta}}$, and construct the estimators of $m_l(\cdot)$ as $\hat{m}_l(\cdot, \hat{\boldsymbol{\beta}}) = \boldsymbol{B}(\cdot)^T \hat{\boldsymbol{\lambda}}_l(\hat{\boldsymbol{\beta}})$, for $l = 1, ..., d$.

**Remark 1.** Another merit of the kernel smoothing method is that we can quickly obtain the covariance matrix estimation of $\hat{\boldsymbol{\beta}}$ by using the sand-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170



wich formula, which effectively avoids estimating the density function of the random error.

## 2.2 Theoretical properties

Let $\boldsymbol{\beta}^0 = \left\{(\boldsymbol{\beta}_1^0)^T, ..., (\boldsymbol{\beta}_d^0)^T\right\}^T$ be the true parameters in model (1.1), where $\boldsymbol{\beta}_l^0 = \left\{\beta_{l1}^0, \left(\boldsymbol{\beta}_{l,-1}^0\right)^T\right\}^T$ and $\boldsymbol{\beta}_{l,-1}^0 = (\beta_{l2}^0, ..., \beta_{lp_n}^0)^T$, for $1 \leq l \leq d$. Here, the subscript $n$ in $p_n$ is used to make it explicit that the dimension of the loading parameters $p_n$ may depend on $n$. Let $\|g\|_2 = \left\{\int g^2(x)dx\right\}^{1/2}$ be the $L_2$ norm of a function $g$. Now, we define the space $\mathcal{M}$ as a collection of functions with finite $L_2$ norm on $[a,b]^d \times \mathbb{R}^d$ by $\mathcal{M} = \left\{g(\boldsymbol{u}, \boldsymbol{x}) = \sum_{l=1}^{d} g_l(u_l)x_l, Eg_l^2(\boldsymbol{Z}^T\boldsymbol{\beta}_l) < \infty\right\}$, where $\boldsymbol{u} = (u_1, ..., u_d)^T$ and $\boldsymbol{x} = (x_1, ..., x_d)^T$. For $1 \leq k \leq p_n$, we assume that $g_k^0$ is a minimizer in $\mathcal{M}$ for the following optimization problem:

$$\begin{aligned}
\mathbb{P}(Z_k) &= g_k^0\left(\boldsymbol{U}(\boldsymbol{\beta}^0), \boldsymbol{X}\right) \\
&= \sum_{l=1}^{d} g_{l,k}^0(\boldsymbol{Z}^T\boldsymbol{\beta}_l^0)X_l \\
&= \arg\min_{g\in\mathcal{M}} E[f_\varepsilon\left(0|\boldsymbol{X}, \boldsymbol{Z}\right)\left\{Z_k - g\left(\boldsymbol{U}\left(\boldsymbol{\beta}^0\right), \boldsymbol{X}\right)\right\}^2],
\end{aligned}$$

where $\boldsymbol{U}(\boldsymbol{\beta}^0) = (\boldsymbol{Z}^T\boldsymbol{\beta}_1^0, ..., \boldsymbol{Z}^T\boldsymbol{\beta}_d^0)^T$. Next, let $\mathbb{P}(\boldsymbol{Z}) = \{\mathbb{P}(Z_1), \cdots, \mathbb{P}(Z_{p_n})\}^T$, $\tilde{\boldsymbol{Z}} = \boldsymbol{Z} - \mathbb{P}(\boldsymbol{Z})$, $\boldsymbol{H}(\boldsymbol{\beta}_{-1}^0) = E\left\{f_\varepsilon\left(0|\boldsymbol{X}, \boldsymbol{Z}\right)\left[\left(\dot{m}_l\left(\boldsymbol{Z}^T\boldsymbol{\beta}_l^0, \boldsymbol{\beta}^0\right)X_l\boldsymbol{J}_l^{0T}\tilde{\boldsymbol{Z}}\right)_{l=1}^{d}\right]^{\otimes 2}\right\}$, and $\boldsymbol{M}(\boldsymbol{\beta}_{-1}^0) = E\left[\left(\dot{m}_l\left(\boldsymbol{Z}^T\boldsymbol{\beta}_l^0, \boldsymbol{\beta}^0\right)X_l\boldsymbol{J}_l^{0T}\tilde{\boldsymbol{Z}}\right)_{l=1}^{d}\right]^{\otimes 2}$, with $\boldsymbol{J}_l^0 = \boldsymbol{J}_l\left(\boldsymbol{\beta}_{l,-1}^0\right)$, for $1 \leq l \leq d$, $\boldsymbol{A}^{\otimes 2} = \boldsymbol{A}\boldsymbol{A}^T$ for any matrix $\boldsymbol{A}$, and $(\boldsymbol{a}_l)_{l=1}^d = \left(\boldsymbol{a}_1^T, ..., \boldsymbol{a}_d^T\right)^T$ for any vector $\boldsymbol{a}_l$. For any positive numbers $a_n$ and $b_n$, we denote $a_n \ll b_n$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170



if $a_n/b_n = o(1)$. Denote the space of the $r$th-order smooth function as $C^{(r)}[a,b] = \left\{\varphi \left| \varphi^{(r)} \in C[a,b] \right.\right\}$, where $f^{(i)}(v) = d^i f(v)/dv^i$. Let $f_Y(y \mid \boldsymbol{X}, \boldsymbol{Z})$ and $F_Y(y \mid \boldsymbol{X}, \boldsymbol{Z})$ be the condition density and the conditional cumulative distribution function of $Y$ given $(\boldsymbol{X}, \boldsymbol{Z})$, respectively, and let $\nu \geq 2$ be an integer. To prove the theoretical results of the proposed estimators of the nonparametric functions and loading coefficients, we need the following technical conditions.

(C1) For $\boldsymbol{\beta}_l$ in the neighborhood of $\boldsymbol{\beta}_l^0$, the density function $f_{U_l(\boldsymbol{\beta}_l)}(\cdot)$ of the random variable $U_l(\boldsymbol{\beta}_l) = \boldsymbol{Z}^T\boldsymbol{\beta}_l$ is bounded away from zero on $[a,b]$, for $1 \leq l \leq d$, and satisfies the Lipschitz condition of order 1 on $[a,b]$.

(C2) For every $1 \leq l \leq d$ and $1 \leq k \leq p_n$, $g_{l,k}^0 \in C^{(1)}[a,b]$ and $m_l \in C^{(r)}[a,b]$, for some integer $r \geq 2$. At the same time, the spline order $q$ satisfies $q \geq r+2$.

(C3) $\boldsymbol{X}$ has bounded support, and $E\left(\boldsymbol{X}\boldsymbol{X}^T | \boldsymbol{Z}^T\boldsymbol{\beta}_l^0 = u_l\right)$ is positive definite, for all $u_l \in [a,b]$.

(C4) $E\left[\left(\dot{m}_l\left(\boldsymbol{Z}^T\boldsymbol{\beta}_l^0, \boldsymbol{\beta}^0\right) X_l \boldsymbol{J}_l^{0T}\tilde{\boldsymbol{Z}}\right)_{l=1}^{d}\right]^{\otimes 2}$ has eigenvalues that are bounded and bounded away from zero.

(C5) For all $u$ in a neighborhood of 0, $f_\varepsilon(u \mid \boldsymbol{X}, \boldsymbol{Z})$ is bounded away from zero and is $\nu$ times continuously differentiable with respect to $u$.

(C6) The kernel function $K(u)$ is nonnegative, bounded, symmetrical,

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170



continuous, and compactly supported on $[-1, 1]$. Furthermore, for some constant $C_K \neq 0$, $K(\cdot)$ is a $\nu$th-order kernel function. For example, $\int u^j K(u)\, du$ is equal to one if $j = 0$, zero if $1 \leq j \leq \nu - 1$, and $C_K$ if $j = \nu$.

(C7) The positive bandwidth $h$ satisfies $nh^{2\nu} \to 0$.

**Remark 2.** Conditions (C1)–(C2) are standard conditions for a VICM, and are similar to conditions (C1), (C2), and (C5) in Ma and Song (2015). Condition (C3) is similar to condition (C3) in Ma and Xu (2015) and Assumption 3 in Whang (2006). Condition (C4) is similar to condition (C7) in Xue and Wang (2012), and ensures that the asymptotic variance for the estimator of $\boldsymbol{\beta}^0$ exists. Condition (C5) is similar to Assumption 4 in Whang (2006). From condition (C5) and the fact that $\varepsilon = Y - \sum_{l=1}^{d} m_l(\boldsymbol{Z}^T\boldsymbol{\beta}_l^0)X_l$, the conditional density $f_Y(y\,|\boldsymbol{X}, \boldsymbol{Z})$ satisfies the Lipschitz condition of order one and $f_Y\left(\sum_{l=1}^{d} m_l\left(\boldsymbol{Z}^T\boldsymbol{\beta}_l\right) X_l\,|\boldsymbol{X}, \boldsymbol{Z}\right)$ is bounded away from zero for $\boldsymbol{\beta}$ in a neighborhood of $\boldsymbol{\beta}^0$. Conditions (C6)–(C7) are necessary conditions on the kernel function and the bandwidth $h$, which are also required in Whang (2006). Condition (C7) ensures that the smoothing has an asymptotically negligible bias on the estimator of $\boldsymbol{\beta}^0$.

**Theorem 1.** *Assume conditions (C1)–(C7) and* $n^{1/(2r+2)} \ll J_n \ll n^{1/4}$ *hold. If* $n^{-1}p_n^3 = o(1)$, *then* $\forall \boldsymbol{e}_n \in \mathbb{R}^{d(p_n-1)}$, *such that* $\boldsymbol{e}_n^T\boldsymbol{e}_n = 1$, *and we*

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170



*have*

*(i)* $\left\|\hat{\boldsymbol{\beta}}_{-1}-\boldsymbol{\beta}_{-1}^{0}\right\|_{2}=O_{p}\left(\sqrt{p_{n}/n}\right)$,

*(ii)* $n^{1/2}\boldsymbol{e}_{n}^{T}\boldsymbol{M}^{-1/2}\left(\boldsymbol{\beta}_{-1}^{0}\right)\boldsymbol{H}\left(\boldsymbol{\beta}_{-1}^{0}\right)\left(\hat{\boldsymbol{\beta}}_{-1}-\boldsymbol{\beta}_{-1}^{0}\right)\xrightarrow{d}N\left(0,\tau(1-\tau)\right)$, *where* $\xrightarrow{d}$ *denotes convergence in distribution.*

**Theorem 2.** *Under the same conditions of Theorem 1, for* $1\leq l\leq d$, *we have* $|\hat{m}_{l}(u_{l},\hat{\boldsymbol{\beta}})-m_{l}(u_{l})|=O_{p}\left(\sqrt{J_{n}/n}+J_{n}^{-r}\right)$ *uniformly, for any* $u_{l}\in[a,b]$.

In practice, we approximate $\mathbb{P}(Z_{ik})$ using its spline estimator $\mathbb{P}_{n}(Z_{ik})$, with its explicit form given in (A.29) of the Supplementary Material. Let $\hat{\mathbb{P}}_{n}(Z_{ik})=\boldsymbol{D}_{i}\left(\hat{\boldsymbol{\beta}}\right)^{T}\left\{\sum_{i=1}^{n}\hat{w}_{i}\boldsymbol{D}_{i}(\hat{\boldsymbol{\beta}})\boldsymbol{D}_{i}(\hat{\boldsymbol{\beta}})^{T}\right\}^{-1}\sum_{i=1}^{n}\hat{w}_{i}\boldsymbol{D}_{i}(\hat{\boldsymbol{\beta}})Z_{ik}$, $\hat{w}_{i}=h^{-1}K\left(\hat{\varepsilon}_{i}/h\right)$, $\hat{\varepsilon}_{i}=Y_{i}-\sum_{l=1}^{d}\hat{m}_{l}(\boldsymbol{Z}_{i}^{T}\hat{\boldsymbol{\beta}}_{l},\hat{\boldsymbol{\beta}})X_{il}$, $\hat{\mathbb{P}}_{n}(\boldsymbol{Z}_{i})=\{\hat{\mathbb{P}}_{n}(Z_{i1}),...,\hat{\mathbb{P}}_{n}(Z_{ip_{n}})\}^{T}$, $\hat{\boldsymbol{Z}}_{i}=\boldsymbol{Z}_{i}-\hat{\mathbb{P}}_{n}(\boldsymbol{Z}_{i})$, $\hat{\boldsymbol{J}}_{l}=\boldsymbol{J}_{l}\left(\hat{\boldsymbol{\beta}}_{l,-1}\right)$, $\mathbb{H}_{n}\left(\hat{\boldsymbol{\beta}}_{-1}\right)=\sum_{i=1}^{n}\hat{w}_{i}\left[\left(\hat{\dot{m}}_{l}\left(\boldsymbol{Z}_{i}^{T}\hat{\boldsymbol{\beta}}_{l},\hat{\boldsymbol{\beta}}\right)X_{il}\hat{\boldsymbol{J}}_{l}^{T}\hat{\boldsymbol{Z}}_{i}\right)_{l=1}^{d}\right]^{\otimes2}$, and $\mathbb{M}_{n}\left(\hat{\boldsymbol{\beta}}_{-1}\right)=\sum_{i=1}^{n}\left[\psi_{\tau}\left\{\hat{\varepsilon}_{i}\right\}\left(\hat{\dot{m}}_{l}\left(\boldsymbol{Z}_{i}^{T}\hat{\boldsymbol{\beta}}_{l},\hat{\boldsymbol{\beta}}\right)X_{il}\hat{\boldsymbol{J}}_{l}^{T}\hat{\boldsymbol{Z}}_{i}\right)_{l=1}^{d}\right]^{\otimes2}$.

**Remark 3.** Based on the above results, we apply the following sandwich formula to consistently estimate the asymptotic covariance of $\hat{\boldsymbol{\beta}}_{-1}$:

$$\widehat{Cov}\left(\hat{\boldsymbol{\beta}}_{-1}\right)=\mathbb{H}_{n}^{-1}\left(\hat{\boldsymbol{\beta}}_{-1}\right)\mathbb{M}_{n}\left(\hat{\boldsymbol{\beta}}_{-1}\right)\mathbb{H}_{n}^{-1}\left(\hat{\boldsymbol{\beta}}_{-1}\right). \tag{2.5}$$

Furthermore, we define $\hat{\mathbb{J}}=\oplus_{l=1}^{d}\hat{\boldsymbol{J}}_{l}=\mathrm{diag}(\hat{\boldsymbol{J}}_{1},...,\hat{\boldsymbol{J}}_{d})$ as the direct sum of the Jacobian matrices $\hat{\boldsymbol{J}}_{1},...,\hat{\boldsymbol{J}}_{d}$ with dimension $dp_{n}\times d(p_{n}-1)$. Then,

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170



we can obtain the estimated asymptotic covariance of $\hat{\boldsymbol{\beta}}$ as $\widehat{Cov}\left(\hat{\boldsymbol{\beta}}\right) = \hat{\mathbb{J}}\widehat{Cov}\left(\hat{\boldsymbol{\beta}}_{-1}\right)\hat{\mathbb{J}}^T$.

## 3. Penalized Estimation for High-dimensional Loading Parameters

Thus far, all covariates $\boldsymbol{Z}$ in model (1.1) have been assumed to be important for predicting the response variable. However, the true model is often unknown. On the one hand, fitted models may be seriously biased and non-informative if important predictors are omitted. On the other hand, including spurious covariates may unnecessarily increase the complexity and further reduce the estimation efficiency. Thus, a fundamental issue is selecting variables for the VICM with a diverging number of loading parameters. As usual, we assume the model is sparse, in the sense that most of the components of $\boldsymbol{\beta}$ are essentially zero. Recall from the preceding section, after profiling, we obtain a single objective function as a function of $\boldsymbol{\beta}$. We can then introduce a common penalty toward sparsity to regularize the coefficient. More specifically, we modify (2.3) to be

$$\mathcal{R}_{\tau nh}\left(\boldsymbol{\beta}_{-1}\right) + n\boldsymbol{b}_{\alpha_1}\left(\boldsymbol{\beta}_{-1}\right) = \mathbf{0}, \tag{3.1}$$

where $\boldsymbol{b}_{\alpha_1}\left(\boldsymbol{\beta}_{-1}\right) = [\dot{p}_{\alpha_1}\left(|\beta_{12}|\right)\operatorname{sgn}\left(\beta_{12}\right), ..., \dot{p}_{\alpha_1}\left(|\beta_{1p_n}|\right)\operatorname{sgn}\left(\beta_{1p_n}\right), ..., \dot{p}_{\alpha_1}\left(|\beta_{dp_n}|\right) \times\operatorname{sgn}\left(\beta_{dp_n}\right)]$ is a $d(p_n - 1)$ vector with $\operatorname{sgn}\left(t\right) = I\left(t > 0\right) - I\left(t < 0\right)$, and

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170



$\dot{p}_{\alpha_1}(\cdot)$ is the first-order derivative of the SCAD penalty function, defined by

$$\dot{p}_{\alpha_1}(x) = \alpha_1 \left\{ I\left(x \leq \alpha_1\right) + \frac{\left(a\alpha_1 - x\right)_+}{\left(a-1\right)\alpha_1} I\left(x > \alpha_1\right) \right\},$$

where $a > 2$, $p_{\alpha_1}(0) = 0$, and $\alpha_1$ is a nonnegative penalty parameter that regulates the complexity of the model. In our simulation studies and real-data analysis, we set $a = 3.7$. The iterative majorize-minorize (MM) algorithm proposed by Hunter and Li (2005) can be incorporated to estimate $\boldsymbol{\beta}_{-1}$ in (3.1). Specifically, for a fixed $\alpha_1$, we obtain the estimator $\bar{\boldsymbol{\beta}}_{\alpha_1,-1}$ of $\boldsymbol{\beta}_{-1}$ using the following iterative procedure:

$$\begin{aligned} \boldsymbol{\beta}_{\alpha_1,-1}^{(k+1)} = \boldsymbol{\beta}_{\alpha_1,-1}^{(k)} - \Big\{ & \left[\partial \mathcal{R}_{\tau nh}\left(\boldsymbol{\beta}_{-1}\right) / \partial \boldsymbol{\beta}_{-1}^T + n\boldsymbol{\Delta}_{\alpha_1}\left(\boldsymbol{\beta}_{-1}\right)\right]^{-1} \\ & \times \left[\mathcal{R}_{\tau nh}\left(\boldsymbol{\beta}_{-1}\right) + n\boldsymbol{b}_{\alpha_1}\left(\boldsymbol{\beta}_{-1}\right)\right] \Big\} \mid_{\boldsymbol{\beta}_{-1} = \boldsymbol{\beta}_{\alpha_1,-1}^{(k)}}, \end{aligned} \tag{3.2}$$

where $\boldsymbol{\Delta}_{\alpha_1}\left(\boldsymbol{\beta}_{-1}\right) = diag\left( \frac{\dot{p}_{\alpha_1}(|\beta_{12}|)}{\kappa + |\beta_{12}|}, ..., \frac{\dot{p}_{\alpha_1}(|\beta_{1p_n}|)}{\kappa + |\beta_{1p_n}|}, ..., \frac{\dot{p}_{\alpha_1}\left(\left|\beta_{dp_n}\right|\right)}{\kappa + \left|\beta_{dp_n}\right|} \right)$, and $\kappa$ is a small number, such as $10^{-6}$. The above iterative formula is similar to the MM algorithm of Hunter and Li (2005), and its convergence can be similarly justified using their Proposition 3.3 under the stationary and continuity assumptions.

We next study the asymptotic properties for the proposed penalized estimator, including the well-known sparsity and oracle properties. In general, we define the true coefficients as $\boldsymbol{\beta}_{l,-1}^0 = \left( \left(\boldsymbol{\beta}_{l,-1}^{0(1)}\right)^T, \left(\boldsymbol{\beta}_{l,-1}^{0(2)}\right)^T \right)^T$, with $\boldsymbol{\beta}_{l,-1}^{0(1)} = \left(\beta_{l2}^0, ..., \beta_{ls_l}^0\right)^T$ and $\boldsymbol{\beta}_{l,-1}^{0(2)} = \left(\beta_{l(s_l+1)}^0, ..., \beta_{lp_n}^0\right)^T$, where $\beta_{lj}^0 \neq 0$ for $j =$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170



$2, ..., s_l$, and $\beta_{lj}^0 = 0$ for $j = s_l+1, ..., p_n$; $\boldsymbol{\beta}_{-1}^{0(1)} = \left( \left( \boldsymbol{\beta}_{1,-1}^{0(1)} \right)^T, ..., \left( \boldsymbol{\beta}_{d,-1}^{0(1)} \right)^T \right)^T$; and $\boldsymbol{\beta}_{-1}^{0(2)} = \left( \left( \boldsymbol{\beta}_{1,-1}^{0(2)} \right)^T, ..., \left( \boldsymbol{\beta}_{d,-1}^{0(2)} \right)^T \right)^T$. Correspondingly, we also divide $\bar{\boldsymbol{\beta}}_{\alpha_1 l,-1}$ into two parts, namely, $\bar{\boldsymbol{\beta}}_{\alpha_1 l,-1} = \left( (\bar{\boldsymbol{\beta}}_{\alpha_1 l,-1}^{(1)})^T, (\bar{\boldsymbol{\beta}}_{\alpha_1 l,-1}^{(2)})^T \right)^T$, with $\bar{\boldsymbol{\beta}}_{\alpha_1 l,-1}^{(1)} = \left( \bar{\beta}_{\alpha_1 l 2}, ..., \bar{\beta}_{\alpha_1 l s_l} \right)^T$ and $\bar{\boldsymbol{\beta}}_{\alpha_1 l,-1}^{(2)} = \left( \bar{\beta}_{\alpha_1 l (s_l+1)}, ..., \bar{\beta}_{\alpha_1 l p_n} \right)^T$. Here, we assume the number of nonzero components in $\boldsymbol{\beta}_l$ is fixed, for $l = 1, ..., d$; that is, $s_l$ does not vary with $n$. Define $\bar{\boldsymbol{\beta}}_{\alpha_1,-1}^{(1)} = \left( \left( \bar{\boldsymbol{\beta}}_{\alpha_1 1,-1}^{(1)} \right)^T, ..., \left( \bar{\boldsymbol{\beta}}_{\alpha_1 d,-1}^{(1)} \right)^T \right)^T$ and $\bar{\boldsymbol{\beta}}_{\alpha_1,-1}^{(2)} = \left( \left( \bar{\boldsymbol{\beta}}_{\alpha_1 1,-1}^{(2)} \right)^T, ..., \left( \bar{\boldsymbol{\beta}}_{\alpha_1 d,-1}^{(2)} \right)^T \right)^T$. We need to introduce some additional conditions to derive the asymptotic theory.

(C8) $\liminf_{n\to\infty} \liminf_{x\to 0_+} \dot{p}_{\alpha_1}(x)/\alpha_1 > 0$.

(C9) $a_n = \max_{2\le j\le p_n, 1\le l\le d} \{ \dot{p}_{\alpha_1}(|\beta_{lj}^0|), \beta_{lj}^0 \neq 0 \} = O\left(n^{-1/2}\right)$.

(C10) $b_n = \max_{2\le j\le p_n, 1\le l\le d} \{ |\ddot{p}_{\alpha_1}(|\beta_{lj}^0|)|, \beta_{lj}^0 \neq 0 \} \to 0$ as $n \to \infty$.

(C11) There are constants $C_1$ and $C_2$ such that $|\ddot{p}_{\alpha_1}(x_1) - \ddot{p}_{\alpha_1}(x_2)| \le C_2 |x_1 - x_2|$ when $x_1, x_2 > C_1\alpha_1$.

(C12) Assume $\left\{ \beta_{l2}^0, ..., \beta_{ls_l}^0 \right\}_{l=1}^d$ satisfy $\min_{1\le l\le d, 2\le j\le s_l} |\beta_{lj}^0|/\alpha_1 \to \infty$ as $n \to \infty$.

**Remark 4.** Conditions (C8)–(C11) are the regularity conditions on the penalty given in Fan and Peng (2004), and condition (C12) is similar to condition (H) of Fan and Peng (2004), which is used to obtain the oracle property.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170



**Theorem 3.** *Under conditions (C1)–(C11) and* $n^{1/(2r+2)} \ll J_n \ll n^{1/4}$*, if* $n^{-1}p_n^3 = o(1)$ *as* $n \to \infty$*, we have* $\left\|\bar{\boldsymbol{\beta}}_{\alpha_1,-1} - \boldsymbol{\beta}^0_{-1}\right\|_2 = O_p\left(\sqrt{p_n}\left(n^{-1/2} + a_n\right)\right)$*, where* $a_n$ *is given in condition (C9).*

Let $\mathbb{M}^{(1)}$ and $\mathbb{H}^{(1)}$ be the $\sum_{l=1}^{d}(s_l-1) \times \sum_{l=1}^{d}(s_l-1)$ sub-matrices of $\boldsymbol{M}\left(\boldsymbol{\beta}^0_{-1}\right)$ and $\boldsymbol{H}\left(\boldsymbol{\beta}^0_{-1}\right)$, respectively, corresponding to $\boldsymbol{\beta}^{0(1)}_{-1}$.

**Theorem 4.** *Under conditions (C1)–(C12) and* $n^{1/(2r+2)} \ll J_n \ll n^{1/4}$*, if* $\alpha_1 \to 0$*,* $\sqrt{n/p_n}\alpha_1 \to \infty$*, and* $n^{-1}p_n^3 = o(1)$ *as* $n \to \infty$*, with probability tending to one, the consistent estimator* $\bar{\boldsymbol{\beta}}_{\alpha_1,-1}$ *in Theorem 3 satisfies*

*(i)* $\bar{\boldsymbol{\beta}}^{(2)}_{\alpha_1 l,-1} = \mathbf{0}$ *for* $1 \le l \le d$*;*

*(ii)* $\sqrt{n}\left(\bar{\boldsymbol{\beta}}^{(1)}_{\alpha_1,-1} - \boldsymbol{\beta}^{0(1)}_{-1}\right) \xrightarrow{d} N\left(\mathbf{0}, \tau(1-\tau)(\mathbb{H}^{(1)})^{-1}\mathbb{M}^{(1)}(\mathbb{H}^{(1)})^{-1}\right)$.

Now, we define $\mathbb{J}^0 = \oplus_{l=1}^{d}\boldsymbol{J}^0_l = \mathrm{diag}(\boldsymbol{J}^0_1, ..., \boldsymbol{J}^0_d)$ as the direct sum of the Jacobian matrices $\boldsymbol{J}^0_1, ..., \boldsymbol{J}^0_d$ with dimension $dp_n \times d(p_n-1)$. For $1 \le l \le d$, $\boldsymbol{\beta}_l$ can be estimated as $\bar{\boldsymbol{\beta}}_{\alpha_1,l} = (\bar{\beta}_{\alpha_1,l1}, ..., \bar{\beta}_{\alpha_1,lp_n})^T$, with $\bar{\beta}_{\alpha_1,l1} = \left(1 - \sum_{k=2}^{p_n}\bar{\beta}^2_{\alpha_1,lk}\right)^{1/2}$. From Theorem 4 (ii), we can use the multivariate delta method to obtain the asymptotic normality of $\bar{\boldsymbol{\beta}}^{(1)}_{\alpha_1} = (\bar{\boldsymbol{\beta}}^{(1)T}_{\alpha_1,1}, ..., \bar{\boldsymbol{\beta}}^{(1)T}_{\alpha_1,d})^T$, with $\bar{\boldsymbol{\beta}}^{(1)}_{\alpha_1,l} = \left(\bar{\beta}_{\alpha_1 l1}, \bar{\beta}_{\alpha_1 l2}, ..., \bar{\beta}_{\alpha_1 l s_l}\right)^T$, for $1 \le l \le d$. That is, $\sqrt{n}(\bar{\boldsymbol{\beta}}^{(1)}_{\alpha_1} - \boldsymbol{\beta}^{0(1)}) \xrightarrow{d} N\left(\mathbf{0}, \tau(1-\tau)\mathbb{J}^{0(1)}(\mathbb{H}^{(1)})^{-1}\mathbb{M}^{(1)}(\mathbb{H}^{(1)})^{-1}\mathbb{J}^{0(1)T}\right)$, where $\mathbb{J}^{0(1)}$ is a sub-matrix of $\mathbb{J}^0$ corresponding to $\boldsymbol{\beta}^{0(1)}$, and $\boldsymbol{\beta}^{0(1)} = \left(\beta^0_{11}, ..., \beta^0_{1s_1}, ..., \beta^0_{d1}, ..., \beta^0_{ds_d}\right)^T$.

**Remark 5.** Theorem 3 shows that $\bar{\boldsymbol{\beta}}_{\alpha_1,-1}$ is a $\sqrt{n/p_n}$-consistent estimator

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170



if $a_n = O(n^{-1/2})$. Theorem 4 indicates that $\bar{\boldsymbol{\beta}}_{\alpha_1,-1}$ is consistent in terms of variable selection and has the oracle property when the number of loading parameters diverges. These results provide a theoretical guarantee for the application of our proposed estimation for the high-dimensional QR VICM. Based on the iterative procedure (3.2), we estimate the asymptotic covariance matrix of $\bar{\boldsymbol{\beta}}_{\alpha_1,-1}$ using the following sandwich formula:

$$\widehat{Cov}\left(\bar{\boldsymbol{\beta}}_{\alpha_1,-1}\right) = \bar{\mathbb{H}}_n^{-1}\left(\bar{\boldsymbol{\beta}}_{\alpha_1,-1}\right)\mathbb{M}_n\left(\bar{\boldsymbol{\beta}}_{\alpha_1,-1}\right)\bar{\mathbb{H}}_n^{-1}\left(\bar{\boldsymbol{\beta}}_{\alpha_1,-1}\right), \tag{3.3}$$

where $\bar{\mathbb{H}}_n\left(\bar{\boldsymbol{\beta}}_{\alpha_1,-1}\right) = \mathbb{H}_n\left(\bar{\boldsymbol{\beta}}_{\alpha_1,-1}\right) + n\boldsymbol{\Delta}_{\alpha_1}\left(\bar{\boldsymbol{\beta}}_{\alpha_1,-1}\right)$, and $\mathbb{M}_n$ and $\mathbb{H}_n$ are defined as in subsection 2.2.

**Remark 6.** The main reason that we do not consider variable selection for $X_k, 1 \leq k \leq d$ is as follows. Based on our model $Q_\tau\left(Y|\boldsymbol{X},\boldsymbol{Z}\right) = \sum_{l=1}^{d} m_{\tau,l}(\boldsymbol{Z}^T\boldsymbol{\beta}_{\tau,l})X_l$, it is easy to see that $m_{\tau,l}\left(\cdot\right) = 0$ implies that $\boldsymbol{\beta}_{\tau,l}$ can take any value. In fact, $\boldsymbol{\beta}_{\tau,l}$ has no impact on $Q_\tau\left(Y|\boldsymbol{X},\boldsymbol{Z}\right)$ after fixing $m_{\tau,l}\left(\cdot\right) = 0$. In this case, our considered model is unidentifiable. Consequently, for the sake of model identification, we assume that all components $m_{\tau,l}\left(\cdot\right)$, for $l = 1, ..., d$ are nonzero. Thus, it is not practical for us to implement variable selection for $X_k$, for $1 \leq k \leq d$ (because this is usually equivalent to finding $m_{\tau,l}(\cdot) = 0$). In addition, in practice, many variables are usually used to construct the index function (thus, a high-dimensional $\boldsymbol{Z}$), but relatively

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170



fewer variables used for $\boldsymbol{X}$. Hence, we consider variable selection to be a more relevant issue for $\boldsymbol{Z}$, and examine it in detail.

**Remark 7.** Ultrahigh-dimensional variable selection ($p \gg n$) has recently become very popular, especially for genetic studies. There are practical challenges to allowing $p > n$ in our methods. First, note that we assume $\|\boldsymbol{\beta}_l\|_2 = 1$, for $l = 1, ..., d$, for the sake of model identifiability, indicating that $\mid \beta_{lk} \mid < 1$, for $l = 1, ..., d$ and $k = 1, ..., p$. Thus, it is practically difficult to separate all nonzero coefficients in $\boldsymbol{\beta}$ from ultrahigh-dimensional background noise, because the true signal is rather weak ($< 1$). Thus far, even recently, existing research findings on nonparametric index models have been based mostly on fixed $p$ or on diverging dimensionality with $p < n$; see Wang and Wang (2015), Huang et al. (2014), Lian and Liang (2016), Zhang et al. (2016), Ma and He (2016), Zhao et al. (2017), Zhao and Lian (2017), and Zhang et al. (2017), among others. Furthermore, implementing the proposed estimation procedures for $p > n$ is computationally prohibitive. We recommend that alternative dimension-reduction statistical methodologies be developed to deal with the ultrahigh-dimensional case. This an intriguing extension is left to future research.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170



## 4. Identification of Linear Components in a QR VICM

In varying-index coefficient models, identifying the linear interaction components is also an important issue. Ma and Song (2015) proposed a generalized likelihood ratio test to distinguish linear functions from nonparametric functions. However, the classical significance tests may not be useful in high-dimensional settings, owing to computational and theoretical concerns. Therefore, we develop a penalized procedure to investigate whether there is a linear interaction effect between $\boldsymbol{Z}^T\boldsymbol{\beta}_l$ and $X_l$.

Let $\ddot{m}_l$ be the second derivative of $m_l$. Clearly, $\|\ddot{m}_l\|_2 = 0$ if $m_l$ is a linear function, for $1 \leq l \leq d$. Thus, by shrinking $\|\ddot{m}_l\|_2$ toward zero, we can automatically identify the linear and nonlinear components in model (1.1). Note that $\|\ddot{m}_l\|_2 = \left\{\int \ddot{m}_l^2(x)dx\right\}^{1/2}$ can be equivalently written as $\sqrt{\boldsymbol{\lambda}_l^T\boldsymbol{D}\boldsymbol{\lambda}_l} \equiv \|\boldsymbol{\lambda}_l\|_{\boldsymbol{D}}$, owing to the well-known algebraic property of the B-spline approximation, where $\boldsymbol{D}$ is a $J_n \times J_n$ matrix with the $(k, k')$ entry being $\int_a^b \ddot{B}_k(x)\,\ddot{B}_{k'}(x)\,dx$. Specifically, we minimize

$$\bar{\boldsymbol{\lambda}} = \underset{\boldsymbol{\lambda}\in\mathbb{R}^{dJ_n}}{\arg\min}\, \mathcal{L}^*_{\tau n}\left(\boldsymbol{\lambda}, \bar{\boldsymbol{\beta}}_{\alpha_1}\right) \equiv \underset{\boldsymbol{\lambda}\in\mathbb{R}^{dJ_n}}{\arg\min} \left\{\mathcal{L}_{\tau n}\left(\boldsymbol{\lambda}, \bar{\boldsymbol{\beta}}_{\alpha_1}\right) + n\sum_{l=1}^{d} p_{\alpha_2}\left(\|\boldsymbol{\lambda}_l\|_{\boldsymbol{D}}\right)\right\}, \tag{4.1}$$

where $p_{\alpha_2}(\cdot)$ is the SCAD penalty with a penalty parameter $\alpha_2$, and $\bar{\boldsymbol{\beta}}_{\alpha_1}$ is given in Section 3. This is still a complicated nonlinear programming

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170



problem, and we use the "ucminf" function in R to find the minimum of (4.1) using numerical computing methods. This R function was developed by Hans Bruun Nielsen and Stig Bousgaard Mortensen for general-purpose unconstrained nonlinear optimization. It is a quasi-Newton-type algorithm, with Broyder-Fletcher-Goldfarb-Shanno updating of the inverse Hessian, and a soft line search with a trust region monitoring method.

**Remark 8.** One may combine two types of penalties in (2.1) to perform variable selection for the loading parameters, and to detect linear/nonlinear functions simultaneously; that is, $\mathcal{Q}_{\tau n}(\boldsymbol{\lambda}, \boldsymbol{\beta}) = \mathcal{L}_{\tau n}(\boldsymbol{\lambda}, \boldsymbol{\beta}) + n\sum_{l=1}^{d}\sum_{j=2}^{p_n} p_{\alpha_1}(|\beta_{lj}|)$ $+n\sum_{l=1}^{d} p_{\alpha_2}(\|\boldsymbol{\lambda}_l\|_{\boldsymbol{D}})$. However, $\boldsymbol{\lambda}$ depends on $\boldsymbol{\beta}$, which indicates that we cannot simultaneously obtain the estimators of $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$ by minimizing $\mathcal{Q}_{\tau n}(\boldsymbol{\lambda}, \boldsymbol{\beta})$. To address this difficulty, an iterative procedure is proposed to select the loading parameter and detect the linear/nonlinear components. That is, for a given $\lambda$, we use the penalized robust estimating equations in (3.1) to estimate and select the loading parameters. Then, we minimize (4.1) to detect the linear/nonlinear components for a given $\boldsymbol{\beta}$.

Let $\bar{\boldsymbol{\lambda}} = \left(\bar{\boldsymbol{\lambda}}_1^T, ..., \bar{\boldsymbol{\lambda}}_d^T\right)^T$ be the minimizer of $\mathcal{L}_{\tau n}^*\left(\boldsymbol{\lambda}, \bar{\boldsymbol{\beta}}_{\alpha_1}\right)$. Consequently, the final estimator of $m_l(\cdot)$ is $\bar{m}_l(\cdot) = \boldsymbol{B}(\cdot)^T\bar{\boldsymbol{\lambda}}_l$, for $1 \le l \le d$. Without loss of generality, we suppose that $m_l$ is truly nonlinear for $1 \le l \le d_1$, and is linear for $d_1 + 1 \le l \le d$. We have the following theoretical results.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170



**Theorem 5.** *Suppose that conditions (C1)–(C12) are satisfied. Then, together with* $n^{1/(2r+2)} \ll J_n \ll n^{1/4}$ *and* $\alpha_2 \to 0$, *we have for each* $1 \le l \le d$, $|\bar{m}_l(u_l, \bar{\boldsymbol{\beta}}_{\alpha_1}) - m_l(u_l)| = O_p\left(\sqrt{J_n/n} + J_n^{-r}\right)$ *uniformly, for any* $u_l \in [a, b]$.

**Theorem 6.** *In addition to the conditions in Theorem 5, we further assume that* $(\sqrt{J_n/n} + J_n^{-r})^{-1}\alpha_2 \to \infty$. *Then, with probability approaching one,* $\|\bar{\boldsymbol{\lambda}}_l\|_{\boldsymbol{D}} = 0$, *and* $\bar{m}_l$ *is a linear function, for* $1 + d_1 \le l \le d$.

## 5. Numerical Illustration

### 5.1 Selection of tuning parameters

in all our numerical studies, we use the cubic spline ($q = 4$) to approximate the nonparametric functions $m_l(\cdot)$ in our simulations. We choose the number of interior knots as $N_n = \left[n^{1/(2q+1)}\right]$ to satisfy the theoretical requirement, where $[a]$ stands for the largest integer not greater than $a$. The kernel function $K(\cdot)$ is set as the second-order Bartlett kernel ($\nu = 2$); that is, $K(u) = \frac{3}{4\sqrt{5}}(1 - u^2/5)\, I\left(|u| \le \sqrt{5}\right)$. The smoothed estimating equations (2.3) depend on the bandwidth $h$. We conduct a sensitivity analysis for the selection of $h$ in the finite samples. Let $\{T^v, v = 1, \cdots, 5\}$ be a random partitioning with size $n/5$ of the full data set $T = (T - T^v) \bigcup T^v$, where $T - T^v$ and $T^v$ are the cross-validated training and test sets, respectively, for $v = 1, ..., 5$. The prediction error (PE) from the fivefold cross-validation

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170



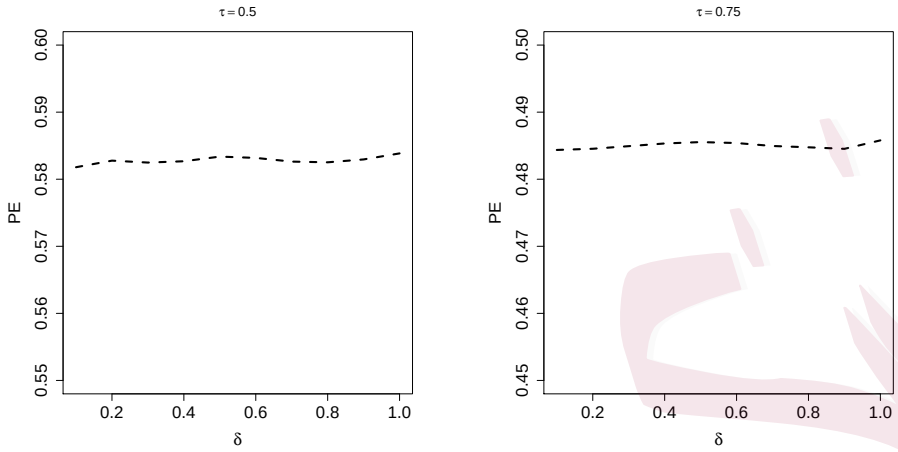


Figure 1: Prediction error from 5-fold cross-validation with different bandwidth $h = n^{-\delta}$ with $\delta = 0.1, 0.2, ..., 1$.

is given by $\text{PE} = n^{-1} \sum_{v=1}^{5} \sum_{(Y_i, \boldsymbol{X}_i, \boldsymbol{Z}_i) \in T^v} \rho_\tau \left( Y_i - \sum_{l=1}^{d} \hat{m}_l^{(v)}(\boldsymbol{Z}_i^T \hat{\boldsymbol{\beta}}_l^{(v)}) X_{il} \right)$, where $\hat{m}_l^{(v)}$ and $\hat{\boldsymbol{\beta}}_l^{(v)}$ are estimators of $m_l$ and $\boldsymbol{\beta}_l$, respectively, using the training set $T - T^v$, for $l = 1, ..., d$. For the quantile levels $\tau = 0.5, 0.75$, we conduct 200 replicates for example 1, given below, with a normal error distribution. Figure 1 depicts the prediction error for the fivefold cross-validation with different bandwidths $h = n^{-\delta}$ for $\delta = 0.1, 0.2, ..., 1$. It is easy to see that the PE does not vary significantly with $h$, indicating that the proposed method is not sensitive to the bandwidth $h$. Thus, we fix $h = n^{-0.3}$ in the simulation studies to reduce the computational burden. This choice also satisfies the theoretical requirement $nh^{2\nu} \to 0$, with $\nu = 2$.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170



The tuning parameter $\alpha_1$ is used to control the sparsity of $\boldsymbol{\beta}$, and the tuning parameter $\alpha_2$ is used to identify the linear functions. Under fixed dimensions, Lian (2012) demonstrated that the Schwartz information criterion (SIC) is consistent in terms of variable selection in the SCAD penalized QR. However, the traditional SIC may not work very well for a diverging number of parameters. Therefore, we adopt the following modified SIC (MSIC) to select $\alpha_1$:

$$\mathrm{MSIC}\left(\alpha_1\right)=\log\left(\mathcal{L}_{\tau n}\left(\hat{\boldsymbol{\lambda}},\bar{\boldsymbol{\beta}}_{\alpha_1}\right)\right)+df_1 C_n\log\left(n\right)/(2n),$$

where $\bar{\boldsymbol{\beta}}_{\alpha_1}$ is the estimated parameter for a given $\alpha_1$, $\hat{\boldsymbol{\lambda}}$ is the unpenalized estimator given in section 2, $df_1$ is the number of nonzero coefficients in $\bar{\boldsymbol{\beta}}_{\alpha_1}$, and $C_n$ is required to be diverging. In our simulations and applications, we choose $C_n$ as $C_n=\max\left\{1,\log\left(\log(dp_n)\right)\right\}$ (Chen and Chen (2008)). The optimal tuning parameter $\hat{\alpha}_1^{opt}$ is defined as $\hat{\alpha}_1^{opt}=\min\limits_{\alpha_1}\mathrm{MSIC}\left(\alpha_1\right)$. Similarly, for $\alpha_2$,

$$\mathrm{MSIC}\left(\alpha_2\right)=\log\left(\mathcal{L}_{\tau n}\left(\bar{\boldsymbol{\lambda}}_{\alpha_2},\bar{\boldsymbol{\beta}}_{\hat{\alpha}_1^{opt}}\right)\right)+df_2 J_n\log\left(n\right)/(2n),$$

where $\bar{\boldsymbol{\lambda}}_{\alpha_2}$ is the estimated parameter for a given $\alpha_2$, and $df_2$ is the number of nonlinear components. Then, we have $\hat{\alpha}_2^{opt}=\min\limits_{\alpha_2}\mathrm{MSIC}\left(\alpha_2\right)$. Note that every $m_l(\cdot)$ is characterized by a spline coefficient vector $\boldsymbol{\lambda}_l$ with dimension $J_n$. Thus, $df_2 J_n$ is regarded as the dimension of the nonlinear function coef-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170



ficients. Our simulation results confirm that the two proposed MSIC criteria work well for variable selection and the identification of linear components.

## 5.2 Simulation studies

**Example 1.** In this example, our goal is to compare the proposed QR estimator with the least squares estimator (LS; Ma and Song (2015)). We generate random samples from the following model:

$$Y_i = \sum\nolimits_{l=1}^{d} m_l\left(\boldsymbol{Z}_i^T\boldsymbol{\beta}_l\right) X_{il} + \sigma\epsilon_i, \tag{5.1}$$

where $\sigma = 0.5$, $d = p = 3$, $X_{i1} = 1$, $(X_{i2}, X_{i3})^T$, and $\boldsymbol{Z}_i = (Z_{i1}, Z_{i2}, Z_{i3})^T$ follow multivariate normal distributions with mean zero, variance one, and constant correlation coefficient 0.5. Here, we set the true loading parameters as $\boldsymbol{\beta}_1 = \frac{1}{\sqrt{14}}(2,1,3)^T$, $\boldsymbol{\beta}_2 = \frac{1}{\sqrt{14}}(3,2,1)^T$, and $\boldsymbol{\beta}_3 = \frac{1}{\sqrt{14}}(2,3,1)^T$, and set the true coefficient functions as $m_1(u_1) = \exp(u_1)/5, m_2(u_2) = \sin(0.5\pi u_2)$, and $m_3(u_3) = u_3^2$. In order to investigate the effects of relatively heavy-tailed error distributions or outliers, we consider the following four error distributions of $\epsilon_i$: the standard normal distribution (SN), $t$-distribution with three degrees of freedom ($t_3$), Laplace distribution (LA) with location parameter zero and shape parameter one and mixed normal distribution (MN$(\rho, \sigma_1, \sigma_2)$), which is a mixture of N$(0, \sigma_1^2)$ and N$(0, \sigma_2^2)$, with weights $1-\rho$ and $\rho$, respectively. In this example, we consider $\rho = 0.1$, $\sigma_1 = 1$ and

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170



$\sigma_2 = 5$. For the purpose of comparison, we consider $\tau = 0.5$ and the sample size $n = 500$ and 1500, with 200 simulation replications. For a fixed $\tau = 0.5$, we have $Q_{0.5}(Y|\boldsymbol{X}, \boldsymbol{Z}) = E(Y|\boldsymbol{X}, \boldsymbol{Z}) = \sum_{l=1}^{d} m_l\left(\boldsymbol{Z}_i^T\boldsymbol{\beta}_l\right) X_{il}$, because the median and the mean of $\epsilon_i$ are both zero under the four error distributions. Therefore, it is fair to compare the proposed QR estimator with the least squares estimator under this setting.

For the parametric part, we report the bias (Bias); empirical standard deviation (ESD), calculated as the sample standard deviation of 200 estimates; estimated asymptotic standard deviation (ASD), based on the sandwich formula (2.5); and mean absolute deviation (MAD), calculated as the mean absolute deviation of 200 estimates. We compute the root average squared error (RASE) to measure the accuracy of the nonparametric estimators $\hat{m}_l$ $\text{RASE}(\hat{m}_l) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{m}_l(u_{il}) - m_l(u_{il})\right)^2}, u_{il} = \boldsymbol{Z}_{\boldsymbol{i}}^T\hat{\boldsymbol{\beta}}_l$, for $l = 1, 2, 3$. To conserve space, we report the corresponding results for the proposed QR estimator with $\tau = 0.5$ and the least squares estimator in Tables S2–S6 in Appendix B of the Supplementary Material. Both the mean regression and the median regression in this example are consistent to the true parameters and functions, owing to their small bias, MAD, and RASE. The aforementioned tables show that the performance of the proposed QR is much more stable than that of LS, especially in cases with non-normal

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170



errors, demonstrating the robust feature of our proposed approach. Finally, the estimated ASD is very close to the ESD, especially for $n = 1500$. This demonstrates that the sandwich covariance formula (2.5) works reasonably well.

**Example 2.** In this example, we specify the conditional quantile function $Q_\tau(Y_i|\boldsymbol{X}_i, \boldsymbol{Z}_i)$ as

$$Q_\tau(Y_i|\boldsymbol{X}_i, \boldsymbol{Z}_i) = m_{\tau,1}(\boldsymbol{Z}_i^T\boldsymbol{\beta}_{\tau,1})X_{i1} + m_{\tau,2}(\boldsymbol{Z}_i^T\boldsymbol{\beta}_{\tau,2})X_{i2} + m_{\tau,3}(\boldsymbol{Z}_i^T\boldsymbol{\beta}_{\tau,3})X_{i3},$$

where $\beta_{\tau,1} = \frac{\left(\tau^{1/2},\tau,2\tau\right)^T}{\sqrt{5\tau^2+\tau}}$, $\beta_{\tau,2} = \frac{\left(\tau,\tau^{1/2},2\tau\right)^T}{\sqrt{5\tau^2+\tau}}$, $\beta_{\tau,3} = \frac{\left(2\tau,\tau,\tau^{1/2}\right)^T}{\sqrt{5\tau^2+\tau}}$, $m_{\tau,1}(u_1) = \tau^{1/2}u_1$, $m_{\tau,2}(u_2) = \tau\sin(0.5\pi u_2)$, and $m_{\tau,3}(u_3) = -0.5\log(1-\tau)u_3^2$. The covariate $X_{i1} = 1$ and $(X_{i2}, X_{i3})^T$ are generated from an independent standard normal distribution. The covariates $\boldsymbol{Z}_i = (Z_{i1}, Z_{i2}, Z_{i3})^T$ are independently generated from the uniform distribution U(0,1). Similarly to Ma and He (2016) and Frumento and Bottai (2016), we generate $Y_i$ as

$$Y_i = m_{U_i,1}(\boldsymbol{Z}_i^T\boldsymbol{\beta}_{U_i,1})X_{i1} + m_{U_i,2}(\boldsymbol{Z}_i^T\boldsymbol{\beta}_{U_i,2})X_{i2} + m_{U_i,3}(\boldsymbol{Z}_i^T\boldsymbol{\beta}_{U_i,3})X_{i3},$$

where $U_i$ follows the uniform distribution U(0,1). In this example, it is easy to see that the loading coefficients $\boldsymbol{\beta}_{\tau,l}$ and nonparametric functions $m_{\tau,l}$, for $l = 1, 2, 3$, are functions of $\tau$, suggesting that the covariate effects vary with the quantile level. Thus, the VICM model structure is more sophisticated than that of example 1, and the mean regression method is

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170



no longer appropriate.

In this example, we consider an estimation at the quartiles $\tau = 0.5$ and $\tau = 0.75$, and simulate 200 data sets with $n = 500$ and $n = 1500$. Tables S7 and S8 in Appendix B of the Supplementary Material give the bias, ESD, ASD, and MAD of $\boldsymbol{\beta}_{\tau,l}$, and the RASE for $m_{\tau,l}$ for the proposed method, for $l = 1, 2, 3$. Note that the true loading coefficients $\boldsymbol{\beta}_{\tau,l}$ and the nonparametric functions $m_{\tau,l}$ are different at $\tau = 0.5$ and 0.75. The proposed estimation is also consistent, with small biases and RASE, and the ESD, ASD, MAD, and RASE become smaller as the sample size increases.

**Example 3.** The main goal of this example is to investigate the finite-sample performance of the proposed penalized estimation approach for variable selection and identifying linear components. We generate random samples from model (5.1), with $\sigma = 0.2$, $d = 4$, $m_1(u_1) = 0.2u_1^3$, $m_2(u_2) = cos(0.5\pi u_2)$, $m_3(u_3) = 0.5u_3$, and $m_4(u_4) = -0.5u_4$. In this case, we allow the last two nonparametric components to be linear functions. The true loading parameters are $\boldsymbol{\beta}_l = \boldsymbol{\varsigma}_l / \|\boldsymbol{\varsigma}_l\|_2$ with $\boldsymbol{\varsigma}_l = (\varsigma_{l1}, ..., \varsigma_{ld_n}, \mathbf{0}_{p_n - d_n})^T$, for $(1 \leq l \leq 4)$, where $\varsigma_{lk}$ is generated from a uniform distribution $U(0.5, 1)$, for $k = 1, ..., d_n$, and $\mathbf{0}_m$ denotes an $m$-vector of zeros. The dimension of $\boldsymbol{\beta}_l$ is set as $p_n = [n^{1/3}]$, and the number of nonzero coefficients in $\boldsymbol{\beta}_l$ is taken as $d_n = [n^{1/4}]$. In this example, we focus on the quantile levels at

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170



$\tau = 0.1, 0.5, 0.75$, and 0.9. To ensure $Q_\tau\left(Y|\boldsymbol{X}, \boldsymbol{Z}\right) = \sum_{l=1}^{d} m_l\left(\boldsymbol{Z}_i^T\boldsymbol{\beta}_l\right) X_{il}$ at $\tau = 0.1, 0.5, 0.75$, and 0.9, we consider $\epsilon_i = \varsigma_i - c_\tau$, and $c_\tau$ is the $\tau$th quantile of the random error $\varsigma_i$, resulting in $Q_\tau\left(\epsilon_i|\boldsymbol{X}_i, \boldsymbol{Z}_i\right) = 0$. Here, $\{\varsigma_i\}$ is an i.i.d. random sample from SN, $t_3$, LA, or MN. The other settings are the same as those of example 1.

To evaluate the performance of the variable selection and the identification of the linear components for our proposed method, we consider the following five criteria: (1) the average number of zero coefficients that are correctly estimated to be zero (C); (2) the average number of nonzero coefficients that are incorrectly estimated to be zero (IC); (3) the average correctly fit (CF) percentage, which measures the accuracy of the variable selection procedure, where " correctly fit" means that the procedure correctly selects significant components from all $\boldsymbol{\beta}_l$, for $l = 1, 2, 3, 4$; (4) the proportion of $m_l$ identified as the linear component for $l = 1, 2, 3, 4$ (ILC$_l$); and (5) the proportion of correctly identified linear components (CIL) among the four nonparametric functions. For the loading parameters, we compute the mean square error of the oracle estimators (O.MSE), penalized estimators (P.MSE), and unpenalized estimators (U.MSE). We also consider the RASE of the penalized estimators (P.RASE) and the unpenalized estimators (U.RASE), which measure the accuracy of the non-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170



parametric estimations. In each case, 200 data sets are generated. The simulation results are summarized in Tables S9–S11 in Appendix B of the Supplementary Material.

Tables S9–S11 show the following observations. First, the values in the column labeled C are very close to the true number of zero-loading parameters. The CF values increase steadily with the sample size $n$, and approach one quickly, indicating that the proposed procedure is consistent in terms of variable selection. Second, the proposed penalized estimator performs similarly to the oracle estimator in terms of the MSE, and significantly improves the estimation accuracy of the unpenalized estimator. Third, only the last two functions $m_3$ and $m_4$ are linear, in this example. Thus, note that $\text{ILC}_l$ is close to zero, for $l = 1, 2$, and $\text{ILC}_l$ approaches one, for $l = 3, 4$, as the sample size increases. These results show that our penalized method can correctly distinguish linear components from nonparametric functions, with a high probability. Fourth, for the nonlinear functions ($m_1$ and $m_2$), there is a small difference between the RASE of the penalized and unpenalized estimators. However, our proposed penalized estimator is obviously more efficient for the linear components $m_3$ and $m_4$, because it reduces about 40%-60% of the RASE relative to that of the unpenalized estimator. This is because $m_3$ and $m_4$ are truly identified as linear functions by the

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170



regularized method. In summary, the proposed methods are satisfactory at different quantile levels in terms of variable selection and the identification of linear components.

## 5.3 Real-data analysis

In this subsection, we illustrate the proposed approaches by analyzing a cross-sectional data set of a workforce company, plus another health survey, in New Zealand during the early 1990s (McCulloch (1995)). This data set consists of physical, lifestyle, and psychological variables, and can be freely downloaded from the R package *VGAMdata*. Three binary variables (*sex*, *diabetes*, *nervous*) and seven continuous variables (*age*, *cholest*, *dmd*, *feethour*, *sleep*, *sbp*, *dbp*) are considered here as predictors. These factors may affect the body mass index (*BMI*) of the subject (Yee (2015)). In this study, our goal is to explore the functional dependency of the body mass index on the risk factors. Thus, we take the *BMI* as the response ($Y$), and set the three binary variables as $\boldsymbol{X}$ ($d = 4$, including an intercept) and the seven continuous variables as the covariate $\boldsymbol{Z}$ ($p = 7$). Detailed definitions of the variables are reported in Table S12. Before implementing the estimation procedure, we normalize all continuous predictor variables to have mean zero and variance one, and take a logarithm transformation

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170



of the response variable. We consider both an unpenalized estimator ($\hat{\boldsymbol{\beta}}_l$ and $\hat{m}_l$) and a penalized estimator ($\bar{\boldsymbol{\beta}}_l$ and $\bar{m}_l$) at the quantile levels $\tau = 0.1, 0.25, 0.5, 0.75$, and 0.9.

We plot a histogram and Q-Q plot of *BMI* in Figure S1 of Appendix B. The figure suggests that the response does not follow a normal distribution. Moreover, we find that the $p$-value is less than $10^{-3}$ using the Shapiro–Wilk test (Shapiro and Wilk (1965)), and therefore reject the null hypothesis of a normal distribution. Thus, a QR analysis may be more suitable here.

The estimated values $\hat{\boldsymbol{\beta}}$ and $\bar{\boldsymbol{\beta}}$ of the loading parameters are presented in Table S13 of Appendix B. Based on the penalized approach, the loadings are automatically estimated as zero and produce sparse solutions. It seems that the estimated loading parameters for $\tau = 0.25, 0.75$, and 0.9 are sparser. For example, at $\tau = 0.25$, the loading parameters for $Z_1$ and $Z_7$ are nonzero for sex ($X_2$), suggesting *age* and *dbp* have interaction effects with gender on the response *BMI* at the first quartile. For the diabetes status $X_3$, *age*, *feethour* and *sleep* may include interaction effects. The other parameter estimates can be interpreted similarly.

After using the penalized estimate $\bar{\boldsymbol{\lambda}}_l$ discussed in Section 4, Table S14 displays the estimated functional norms $\| \bar{\boldsymbol{\lambda}}_l \|_{\boldsymbol{D}}$, for $l = 1, 2, 3, 4$, clearly indicating that $m_3$ is identified as nonlinear for $\tau = 0.1, 0.25, 0.5, 0.75$, and

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170



that $m_4$ is regarded as nonlinear only for $\tau = 0.1$. In all other cases, the functions can be treated as linear. Figure 2 reports the estimated curves and their 95% confidence bands at different quantiles. The graphs agree with the numerical results. In particular, the estimated function $\bar{m}_1(\cdot)$ appears to be an increasing function of index $\boldsymbol{Z}^T\bar{\boldsymbol{\beta}}_{\alpha_1,1}$, which indicates that the combination of seven continuous factors has a positive effect on *BMI*. Other functions can be interpreted similarly for their effects on the response. These nonlinear interaction effects between the covariates $\boldsymbol{Z}^T\boldsymbol{\beta}_l$ and $X_l$ cannot be detected easily without using the proposed QR VICM.

Additional numerical results for this data analysis can be found in Appendix B.

## Supplementary Material

The online Supplementary Material contains the procedure for generating the initial values, additional numerical results, and technical proofs of the theoretical results.

## Acknowledgments

We thank the Editor, an Associate editor, and two referees for their constructive comments and suggestions that have greatly improved the paper.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170



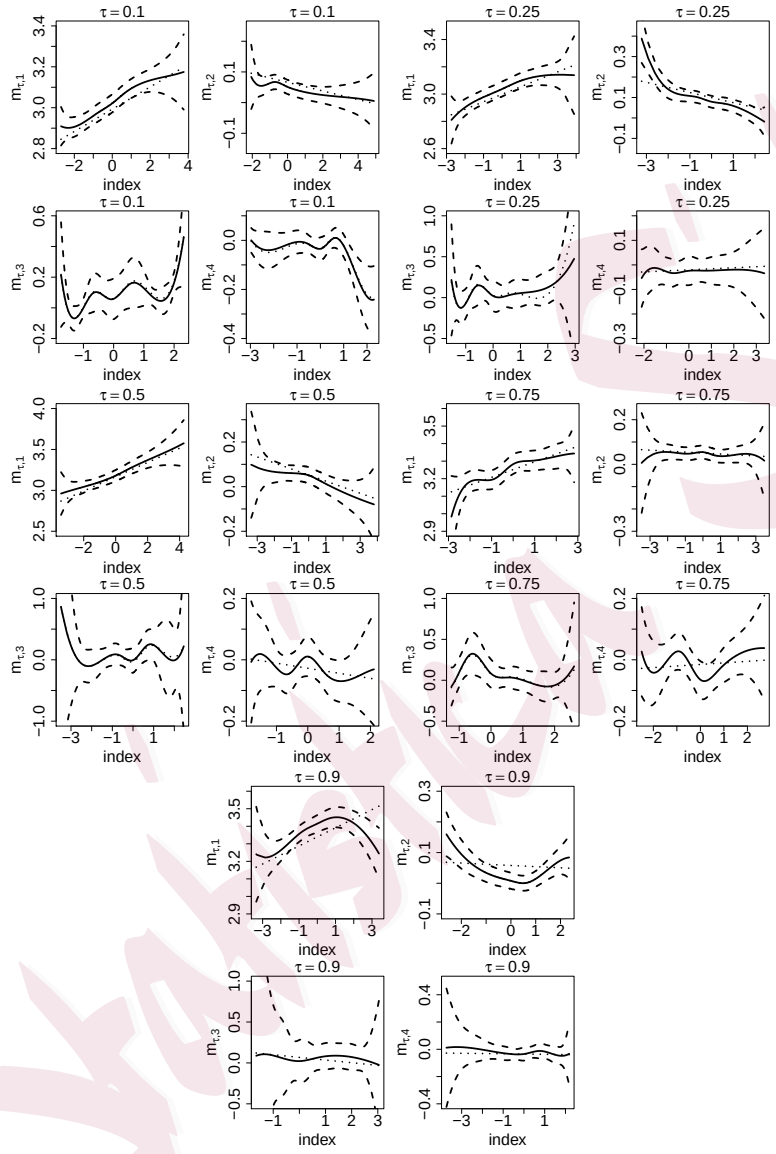


Figure 2: Plots of the unpenalized estimator $\hat{m}_l(\cdot)$ (solid line) and its 95% pointwise confidence intervals (dashed line), and the penalized estimator $\bar{m}_l(\cdot)$ (dotted line).

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170



Jing Lv was partially supported by the National Natural Science Foundation of China Grant 11801466, Fundamental Research Funds for the Central Universities Grant XDJK2019C105, and Basic and Frontier Research Program of Chongqing Grant cstc2017jcyjAX0182. Jialiang Li was partially supported by Academic Research Funds R-155-000-205-114 and R-155-000-195-114, and Tier 2 Ministry of Education funds in Singapore MOE2017-T2-2-082: R-155-000-197-112 (Direct cost) and R-155-000-197-113 (IRC).

## References

Abrevaya, J. (2001). The effect of demographics and maternal behavior on the distribution of birth outcomes. *Empir. Econ.* **26**, 247–259.

Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95**, 759–771.

de Boor, C. (2001). A practical guide to splines. Springer, New York.

Elsner, J. B. Kossin, J. P. and Jagger, T. H. (2008). The increasing intensity of the strongest tropical cyclones. *Nature* **455**, 92–95.

Fan, J. and Li, R. (2006). Statistical challenges with high dimensionality: feature selection in knowledge discovery. Proceedings of the Madrid International Congress of Mathematicians, III: 595–622.

Fan, J. Liu, W. and Lu, X. (2017). Penalized empirical likelihood for semiparametric models

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170



with a diverging number of parameters. *J. Stat. Plan. Infer.* **186**, 42–57.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32**, 928–961.

Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. *Ann. Statist.* **27**, 1491–1518.

Frumento P. and Bottai, M. (2016). Parametric modeling of quantile regression coefficient functions. *Biometrics* **72**, 74–84.

Giraud, C. (2015). Introduction to high-dimensional statistics. Chapman & Hall/CRC Monographs on Statistics & Applied Probability.

Hastie, T. Tibshirani, R. and Wainwright, M. J. (2015). Statistical Learning with Sparsity: the Lasso and Generalizations. Chapman & Hall/CRC Press, Series in Statistics and Applied Probability.

Huang, Z., Pan, Z., Lin, B. and Shao, Q. (2014). Model structure selection in single-index-coefficient regression models. *J. Multivariate Anal.* **125**, 159–175.

Hunter, D. and Li, R. (2005). Variable selection using MM algorithms. *Ann. Statist.* **33**, 1617–1642.

Jin, Z. Lin, D. Y. Wei, L. J. and Ying, Z. (2003). Rank-Based Inference for the Accelerated Failure Time Model. *Biometrika* **90**, 341–353.

Koenker, R. (2005). Quantile regression. Combridge University Press, New York.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170

## REFERENCES

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica* **46**, 33–50.

Lian, H. (2012). A note on the consistency of Schwarz's criterion in linear quantile regression with the SCAD penalty. *Stat. Probabil. Lett.* **82**, 1224–1228.

Lian, H. and Liang, H. (2016). Separation of linear and index covariates in partially linear single-index models. *J. Multivariate Anal.* **143**, 56–70.

Ma, S. and He, X. (2016). Inference for single-index quantile regression models with profile optimization. *Ann. Statist.* **44**, 1234–1268.

Ma, S. and Song, P. X.-K. (2015). Varying index coefficient models.*J. Amer. Statist. Assoc.* **110**, 341–356.

Ma, S. and Xu, S. (2015). Semiparametric nonlinear regression for detecting gene and environment interactions. *J. Stat. Plan. Infer.* **156**, 31–47.

Marimoutou, V. Raggad, B. and Trabelsi, A. (2009). Extreme value theory and value at risk: application to oil market. *Energ. Econ.* **31**, 519–530.

McCulloch, A. (1995). Fletcher Challenge-University of Auckland Heart and Health Study: design and baseline findings. *New Zeal. Med. J.* **108**, 499–502.

Shapiro, S. S. and Wilk, M.B. (1965). An analysis of variance test for normality (complete samples). *Biometrika* **52**, 591–611.

Wang, G. and Wang, L. (2015). Spline estimation and variable selection for single-index prediction models with diverging number of index parameters. *J. Stat. Plan. Infer.* **162**, 1–19.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170

## REFERENCES

Wang, L. Zhou, J. and Qu, A. (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics* **68**, 353–360.

Whang, Y. J. (2006). Smoothed empirical likelihood methods for quantile regression models. *Economet. Theor.* **22**, 173–205.

Xia, Y. Tong, H. Li, W. K. and Zhu, L. (2002). An adaptive estimation of dimension reduction space. *J. R. Stat. Soc. B* **64**, 363–410.

Xue, L. and Wang, Q. (2012). Empirical likelihood for single-index varying-coefficient models. *Bernoulli* **18**, 836–856.

Yee, T. W. (2015). Vector Generalized Linear and Additive Models. Springer. New York.

Zhang, R., Lv, Y., Zhao, W., Liu, J. (2016). Composite quantile regression and variable selection in single-index coefficient model. *J. Stat. Plan. Infer.* **176**, 1–21.

Zhang, Y., Lian, H., Yu, Y. (2017). Estimation and variable selection for quantile partially linear single-index models. *J. Multivariate Anal.* **162**, 215–234.

Zhao, W., Lian, H. (2017). Quantile index coefficient model with variable selection. *J. Multivariate Anal.* **154**, 40–58.

Zhao, W., Zhang, R., Lv, Y., Liu, J. (2017). Quantile regression and variable selection of single-index coefficient model. *Ann. I. Stat. Math.* **69**, 761–789.

School of Mathematics and Statistics, Southwest University, Chongqing 400715, China

E-mail: (lvjing@swu.edu.cn)

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0170

## REFERENCES

Department of Statistics and Applied Probability, National University of Singapore, 119077,

Singapore

E-mail: (stalj@nus.edu.sg)