INSTITUT DE STATISTIQUE
BIOSTATISTIQUE ET
SCIENCES ACTUARIELLES
(ISBA)

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

DISCUSSION
PAPER

2012/20

VARIABLE SELECTION OF VARYING COEFFICIENT MODELS IN QUANTILE REGRESSION

NOH, H. and K. CHUNG and I. VAN KEILEGOM

Variable Selection of Varying Coefficient Models in Quantile Regression

Hohsuk Noh
Université catholique de Louvain

Kwanghun Chung
Hongik University

Ingrid Van Keilegom
Université catholique de Louvain

May 30, 2012

Abstract

Varying coefficient (VC) models are commonly used to study dynamic patterns in many scientific areas. In particular, VC models in quantile regression are known to provide a more complete description of the response distribution than in mean regression. In this paper, we develop a variable selection method for VC models in quantile regression using a shrinkage idea. The proposed method is based on the basis expansion of each varying coefficient and the regularization penalty on the Euclidean norm of the corresponding coefficient vector. We show that our estimator is obtained as an optimal solution to the second order cone programming (SOCP) problem and that the proposed procedure has consistency in variable selection under suitable conditions. Further, we show that the estimated relevant coefficients converge to the true functions at the univariate optimal rate. Finally, the method is illustrated with numerical simulations including the analysis of forced expiratory volume (FEV) data.

1 Introduction

Varying coefficient (VC) models have been widely used as a useful generalization of the linear regression model to depict dynamic behaviors in scientific research. Because of their flexibility and interpretability, much work has been done on their parameter estimation and hypothesis testing, but mostly in the mean regression setting until some authors such as Honda [6], Cai and Xu [2] and Kim [9] paid attention to the quantile regression setting in the early 2000s.

Quantiles themselves can be defined without moment conditions. Inspecting several conditional quantiles can provide us with a more complete description of the data than inspecting only the conditional mean. These advantages have propelled many researchers to consider the quantile regression framework not only in parametric models but also in semi-parametric or nonparametric models. Regarding VC models, Honda [6] and Cai and Xu [2] considered a VC model for the conditional quantiles using local polynomials. Kim [9] proposed a polynomial-spline-based methodology for the same model. For longitudinal data analysis, Wang et al. [22] developed the partially linear VC model in quantile regression. In spite of remarkable progress in estimation and hypotheses testing, it is not well understood how to conduct variable selection efficiently for the VC model in the quantile regression framework.

Variable selection is important for any regression problem in that ignoring important predictors brings out seriously biased results, whereas including spurious predictors leads to substantial loss in estimation efficiency. Due to their computational efficiency, various shrinkage methods such as the nonnegative garrotte, the Least Absolute Shrinkage and Selection Operator (LASSO) and the Smoothly Clipped Absolute Deviation (SCAD) have been used in parametric models and recognized as promising methods to allow us to do estimation and variable selection simultaneously. Furthermore, the past decade has observed their extensions to semi-parametric and nonparametric models using basis approximation techniques. Regarding variable selection in VC models, Wang et al. [23] and Wang and Xia [21] proposed penalization methods for selecting nonzero coefficients using the SCAD penalty [4] and the adaptive LASSO penalty [26], respectively. Noh and Park [14] improved the performance of the estimator in Wang et al. [23] by extending the result of Zou and Li [27] to VC models. However, most existing research about the variable selection for VC models is concentrated on mean regression. Since there are few such works in the quantile regression context, we are stimulated to develop a shrinkage method for the VC model in quantile regression.

While Kai et al. [8] recently developed a variable selection method for the parametric part of a

partially linear VC model in quantile regression using SCAD, they assumed that there is no sparsity of the variables in the nonparametric part. Different from their work, we propose a variable selection method for estimating coefficient functions using a nonparametric approach. We approximate each varying coefficient function with a B-spline basis [3] and consider a penalized check loss function based on the Euclidean norm of the corresponding coefficient vector. Our variable selection method uses a penalization of the norm of each coefficient vector. Therefore, our work shares the same motivation as in Wang et al. [23] and Xue [24]. However, because our interest lies in the estimation of the conditional quantile of the response, we need to use the check loss instead of the squared loss. Therefore, we take a different approach both theoretically and computationally. In particular, the non-differentiability of the check loss function requires us to develop a different computational algorithm. Furthermore, to show its asymptotic properties, we have to adopt a different approach from the one used in mean regression to handle the issues caused by the non-differentiability of the check loss. We will discuss these issues in the proof of our main results, which are given in detail in the Appendix.

The rest of the article is organized as follows. Section 2 introduces the estimator that we propose. We present a computational algorithm for obtaining the estimator and selection methods for the tuning parameters in Section 3. Its asymptotic properties are fully described in Section 4. Finally, we report numerical simulation results as well as an application of our methodology to the forced expiratory volume (FEV) data in Section 5. All the technical proofs are provided in the Appendix.

2 Methodology

2.1 Varying coefficient model

Suppose that $\{(Y_i, U_i, \mathbf{X}_i)\}_{i=1}^n$ is an independent and identically distributed (*i.i.d.*) random sample, where $Y_i \in \mathbb{R}$ is the response of interest, $\mathbf{X}_i = (X_i^{(0)}, X_i^{(1)}, \ldots, X_i^{(p)}) \in \mathbb{R}^{p+1}$ is the $(p+1)$-dimensional covariate vector with $X_i^0 \equiv 1$ and $U_i \in [0,1]$ is the univariate index variable. We consider the following VC model for the conditional quantile of Y_i given $(\mathbf{X}_i, U_i)$:

$$Y_i = Q_\tau(\mathbf{X}_i, U_i) + e_{i,\tau} = \mathbf{X}_i^\top \boldsymbol{\alpha}_\tau(U_i) + e_{i,\tau} = \sum_{k=0}^{p} X_i^{(k)} \alpha_{k,\tau}(U_i) + e_{i,\tau},\ i = 1, \cdots, n, \tag{2.1}$$

where $\boldsymbol{\alpha}_\tau(u) = (\alpha_{0,\tau}(u), \alpha_{1,\tau}(u), \ldots, \alpha_{p,\tau}(u))^\top$ is a coefficient vector and the errors $e_{i,\tau}$ are independent random variables with the τth quantile 0 and $e_{i,\tau}$ is independent of $(U_i, \mathbf{X}_i)$. We assume that only s covariates among the $X_i^{(k)}$'s are relevant in model (2.1). It is unknown which s covariates are relevant,

nor what the value of s is. Without loss of generality, we let $\alpha_k(\cdot),\ k = 1, \ldots, s$ be the nonzero coefficient functions, and $\alpha_k(\cdot),\ k = s+1, \ldots, p$, be identically zero. Additionally we suppress the subscript "τ" whenever no confusion is caused hereafter.

2.2 Regularized estimation using one-step group SCAD

We assume that each coefficient function $\alpha_k(u),\ k = 0, \ldots, p$, can be approximated by a set of basis functions, that is,

$$\alpha_k(u) \approx \sum_{l=1}^{q_k} \gamma_{kl} B_{kl}(u),\ k = 0, \ldots, p, \tag{2.2}$$

where $\{B_{kl}(\cdot),\ l = 1, \ldots, \infty\}$ for all $k = 0, \ldots, p$ span a function space $\mathcal{F}_k$ which is assumed to contain $\alpha_k(u)$, and q_k is the number of basis functions that are needed to approximate $\alpha_k(u)$. Following the approximation (2.2), model (2.1) can be rewritten as

$$Y_i \approx \sum_{k=0}^{p} \sum_{l=1}^{q_k} \gamma_{kl} X_i^{(k)} B_{kl}(U_i) + e_i. \tag{2.3}$$

The parameters γ_{kl} in the basis expansion can be estimated by minimizing

$$l_0(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \rho \left(Y_i - \sum_{k=0}^{p} \sum_{l=1}^{q_k} \gamma_{kl} X_i^{(k)} B_{kl}(U_i) \right), \tag{2.4}$$

where $\rho(t) = 2(\tau - I(t < 0))t$ is the check loss function at a given quantile level $0 < \tau < 1$. We denote the minimizer of (2.4) as $\tilde{\boldsymbol{\gamma}} = (\tilde{\boldsymbol{\gamma}}_0^\top, \tilde{\boldsymbol{\gamma}}_1^\top, \ldots, \tilde{\boldsymbol{\gamma}}_p^\top)^\top$, where $\tilde{\boldsymbol{\gamma}}_k = (\tilde{\gamma}_{k1}, \ldots, \tilde{\gamma}_{kq_k})^\top$. The statistical properties of the estimator of $\alpha_k(\cdot)$ based on $\tilde{\boldsymbol{\gamma}}$ are fully addressed in Kim [9].

Now suppose that some variables are irrelevant in (2.1) so that the corresponding coefficients are zero functions. Since, using the approximation (2.2), each function $\alpha_k(u)$ in (2.1) is characterized by a set of parameters $\boldsymbol{\gamma}_k = (\gamma_{k1}, \ldots, \gamma_{kq_k})^\top$, we should not select nonzero individual components γ_{kl}, but choose the whole nonzero vector $\boldsymbol{\gamma}_k$. For that purpose, we use the regularized estimation by adding penalties not to an individual element γ_{kl} but to the Euclidean norm $\|\boldsymbol{\gamma}_k\|_2$ of a coefficient vector $\boldsymbol{\gamma}_k$. To gain some insights into the proposed method, consider the approximation of $\alpha_k(\cdot)$ as $g_k(\cdot) = \sum_{l=1}^{q_k} \gamma_{kl} B_{kl}(\cdot)$. We note that its squared L_2-norm can be rewritten as $\|g_k\|_{L_2}^2 = \boldsymbol{\gamma}_k^\top H_k \boldsymbol{\gamma}_k$, where H_k is a $q_k \times q_k$ matrix with entries $h_{ll'} = \int_0^1 B_{kl}(u) B_{kl'}(u)\, du$. Since the relevance of a coefficient function $\alpha_k(\cdot)$ is equivalent to $\|\alpha_k\|_{L_2} > 0$, we add to (2.4) thresholding penalties based on $\|\boldsymbol{\gamma}_k\|_w$ where $\|\boldsymbol{\gamma}_k\|_w \equiv (\boldsymbol{\gamma}_k^\top H_k \boldsymbol{\gamma}_k)^{1/2}$ is the weighted Euclidean norm of a vector $\boldsymbol{\gamma}_k \in \mathbb{R}^{q_k}$. Examples of penalties include the LASSO or the SCAD penalty. Such penalties are called group LASSO [25] penalty and group SCAD [23] penalty, respectively.

In this paper, we use a one-step group SCAD penalty, which is a local linear approximation of the group SCAD. It is known that the one-step group SCAD outperforms the group SCAD both in theoretical and computational aspects. For details, we refer to Noh and Park [14]. Let $p_\lambda(\cdot)$ be the SCAD penalty function. The function p_λ is defined on $\mathbb{R}^+$ by its derivative as

$$p'_\lambda(x) = \lambda I(x \le \lambda) + \frac{(a\lambda - x)_+}{a-1} I(x > \lambda)$$

for some constant $a > 2$ and $I(\cdot)$ is the indicator function. In this paper, we define the one-step group SCAD regularized estimator of $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_0^\top, \boldsymbol{\gamma}_1^\top, \ldots, \boldsymbol{\gamma}_p^\top)^\top$ as the minimizer $\hat{\boldsymbol{\gamma}} = (\hat{\boldsymbol{\gamma}}_0^\top, \ldots, \hat{\boldsymbol{\gamma}}_p^\top)^\top$ of

$$l(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \rho \left(Y_i - \sum_{k=0}^{p} \sum_{l=1}^{q_k} \gamma_{kl} X_i^{(k)} B_{kl}(U_i) \right) + n \sum_{k=1}^{p} \nu_k \|\boldsymbol{\gamma}_k\|_w \tag{2.5}$$

where $\nu_k = p'_\lambda(\|\tilde{\boldsymbol{\gamma}}_k\|_w)$. Note that for a given basis $\{B_{k1}(\cdot), \ldots, B_{kq_k}(\cdot)\}$, if there exist constants $\alpha \ge 0$ and $0 < N_1, N_2 < \infty$, not depending on q_k, such that

$$N_1 q_k^{-\alpha} \sum_{l=1}^{q_k} \gamma_{kl}^2 \le \int_0^1 \left[\sum_{l=1}^{q_k} \gamma_{kl} B_{kl}(t) \right]^2 dt \le N_2 q_k^{-\alpha} \sum_{l=1}^{q_k} \gamma_{kl}^2, \ \forall k = 0, \ldots, p, \tag{2.6}$$

the Euclidean norm of $\boldsymbol{\gamma}_k$, $\|\boldsymbol{\gamma}_k\|_2$, can be easily used in the penalty of (2.5) instead of the weighted Euclidean norm, $\|\boldsymbol{\gamma}_k\|_w$. This is because the condition (2.6) enables the direct translation between the Euclidean norm of the estimated coefficient vector $\hat{\boldsymbol{\gamma}}_k$ and the L_2-norm of the estimated function $\hat{\alpha}_k$. B-splines ($\alpha = 1$) and Riesz bases ($\alpha = 0$) are examples of such bases. For these reasons, we will use $\|\boldsymbol{\gamma}_k\|_2$ in the penalty throughout this paper.

3 Implementation of the Proposed Estimator

Since quantile regression typically requires a non-differentiable and asymmetric check loss function, the computation of the estimator defined in (2.5) is quite demanding when penalizing the Euclidean norm of $\boldsymbol{\gamma}_k$. In particular, the iterative algorithm using local quadratic approximation of $\sum_{k=1}^{p} \nu_k \|\boldsymbol{\gamma}_k\|_2$ is not so useful for our estimator as for the one of Wang et al. [23]. On account of this fact, for variable selection of model (2.1), Tang et al. [20] considered the penalty based on the ℓ_1-norm of $\boldsymbol{\gamma}_k$ instead. However, their proposal needs to be justified theoretically in that such penalization cannot generally guarantee groupwise sparsity in the estimator $\boldsymbol{\gamma}$, which is necessary for variable selection in nonparametric models. In our work, while still using the Euclidean norm of $\boldsymbol{\gamma}_k$, we show that the minimization of $l(\boldsymbol{\gamma})$ in (2.5) is equivalent to a second-order cone programming problem. Since it is

a well-known convex optimization problem, the estimator $\hat{\boldsymbol{\gamma}}$ can be calculated using computationally efficient methods from the optimization literature.

3.1 Computational algorithms

Let $\boldsymbol{\pi}(\cdot) = (B_1(\cdot), \ldots, B_{q_k}(\cdot))^\top$ be a set of basis functions for the estimation of $\alpha_k(\cdot)$. We define $\boldsymbol{\Pi}(U, \mathbf{X}) = (X^{(0)}\boldsymbol{\pi}(U)^\top, \ldots, X^{(p)}\boldsymbol{\pi}(U)^\top)^\top$, $\boldsymbol{\pi}_i = \boldsymbol{\pi}(U_i)$ and $\boldsymbol{\Pi}_i = \boldsymbol{\Pi}(U_i, \mathbf{X}_i)$, $i = 1, \ldots, n$. Using these notations, the optimization problem (2.5) is reformulated as:

$$\begin{aligned} \min_{\boldsymbol{\gamma}, \mathbf{v}, \boldsymbol{\eta}^+, \boldsymbol{\eta}^-} & \left(\tau \sum_{i=1}^{n} \eta_i^+ + (1-\tau) \sum_{i=1}^{n} \eta_i^- + n \sum_{k=1}^{p} \nu_k v_k \right) \\ \text{such that } & \eta_i^+ - \eta_i^- = Y_i - \boldsymbol{\Pi}_i^\top \boldsymbol{\gamma}, \ i = 1, \ldots, n \\ & \|\boldsymbol{\gamma}_k\|_2 \leq v_k, \ k = 1, \ldots, p \\ & \eta_i^+ \geq 0, \ \eta_i^- \geq 0, \ i = 1, \ldots, n \end{aligned} \tag{3.1}$$

The reformulation (3.1) shows that the problem to minimize (2.5) is expressed as one of the second order cone programming (SOCP) problems in which a linear objective function is minimized over the intersection of an affine set and second-order (quadratic) cones. For more details, we refer to Lobo et al. [13] and Alizadeh and Goldfarb [1]. It is clear that (3.1) always has a feasible solution because the original problem (2.5) is an unconstrained optimization problem. Therefore, an optimal solution to (3.1) can be obtained by using the convex optimization algorithms such as primal-dual interior point methods. In our simulations, we use `CVX` [5] to solve the SOCP problem (3.1).

3.2 Selection of tuning parameters

Variable selection in nonparametric models needs to determine two regularization parameters. One is a smoothing parameter for controlling the smoothness of the estimated coefficient functions. The other is a shrinkage parameter for managing the complexity that is the number of covariates in the model. In our model, they are the number of basis functions (q_k) to approximate each coefficient $\alpha_k(\cdot)$ and the penalty parameter (λ).

For an initial estimator $\tilde{\boldsymbol{\gamma}}$, we need to choose the number q_k of basis for each coefficient function. To choose the q_k, we use the Schwarz-type Information Criterion (SIC) [17] as follows:

$$SIC(\mathbf{q}^{ini}) = \log \sum_{i=1}^{n} \rho(Y_i - \boldsymbol{\Pi}_i^\top \tilde{\boldsymbol{\gamma}}_{\mathbf{q}^{ini}}) + \frac{\log n}{2n} \sum_{k=0}^{p} q_k^{ini}. \tag{3.2}$$

where $\mathbf{q}^{ini} = (q_0^{ini}, q_1^{ini}, \ldots, q_p^{ini})^\top$. For the proposed estimator, we need to choose q_k's again for the minimization of (2.5). Further, we also have to determine the penalty parameter λ. For the simplicity of implementation, we use the same q_k's for an initial estimator. This is justified by the fact that the optimal order of q_k for the proposed estimator is the same as that of the initial estimator. For the selection of λ, we use SIC as follows:

$$SIC(\lambda) = \log \sum_{i=1}^{n} \rho(Y_i - \mathbf{\Pi}_i^\top \hat{\boldsymbol{\gamma}}_\lambda) + \frac{\log n}{2n} edf, \tag{3.3}$$

where *edf* is the number of zero residuals. The number of zero residuals is widely used as a measure of the effective dimension of the fitted models in quantile regression. Li and Zhu [12] provided its justification in L_1-norm penalized quantile regression and Koenker et al. [10] heuristically argued that in the case of univariate quantile smoothing splines the number of zero residuals is a plausible measure for the effective dimension. Additionally, a recent work by Lee and Noh [11] showed that the proposed SIC in (3.3) with such *edf* gives consistency in model selection for model (2.1).

4 Asymptotic Properties

As a desirable property of nonparametric variable selection, Storlie et al. [19] defined a selection procedure to be nonparametric oracle (np-oracle) if it identifies all relevant variables and estimates the nonzero coefficient functions at the optimal nonparametric rate simultaneously. Wang et al. [23] showed that the estimator for (2.1) in mean regression is np-oracle. In this section, we extend their results to the case of quantile regression.

We focus our asymptotic analysis on the case of B-spline basis functions. Note that if we use normalized B-splines of the order $d+1$ with b_k uniform internal knots to approximate $\alpha_k(\cdot)$, the number of basis functions q_k is equal to $b_k + d + 1$. The extension to general basis expansions satisfying (2.6) or to the case of the penalty based on $\|\boldsymbol{\gamma}_k\|_w$ can be obtained using similar technical arguments given in this paper.

To investigate asymptotic properties of our estimator, we consider the case where q_k tends to infinity as n goes to infinity. Hence, we should use the notation $q_{k,n}$ because q_k depends on n. However, since the assumption that $q_{k,n} = q_n$ for all k does not incur any loss of generality in the asymptotic analysis as long as $\limsup_{n\to\infty}(\max_{0\le k\le p} q_{k,n}/\min_{0\le k\le p} q_{k,n}) < \infty$ is guaranteed, we suppress the dependence of $q_{k,n}$ on k for simplicity. Since we focus on B-splines in our asymptotic analysis, we use the notation b_n instead of $b_{k,n}$ in this section and in the Appendix. Similarly, we use the notation λ_n

since λ needs to vary as n increases. Finally, we use $a_n \approx b_n$ to indicate that there exist constants $0 < A < B < \infty$ such that $A \leq a_n/b_n \leq B$ for sufficiently large n. To derive asymptotic properties of the proposed estimator, we make the following assumptions.

(A1) $\alpha_k(\cdot) \in \mathcal{H}_r,\ k = 0, \ldots, p$ for a constant $r > 3/2$, where $\mathcal{H}_r$ is the collection of all functions on $[0, 1]$ for which the mth order derivative satisfies the Hölder condition of the order γ with $r \equiv m + \gamma$ and $0 < \gamma \leq 1$.

(A2) The conditional distribution of U, given $\mathbf{X} = \mathbf{x}$, has a bounded density $f_{U|\mathbf{X}} : \ 0 < c_1 \leq f_{U|\mathbf{X}}(u|\mathbf{x}) \leq c_2 < \infty$ uniformly in $\mathbf{x}$ and u for some positive constants c_1 and c_2.

(A3) $\mathrm{E}(X^{(k)}|U) = 0$ and $P(|X^{(k)}| < M) = 1$ for some $M < \infty,\ k = 1, \ldots, p$. There exist two positive definite matrices Σ_1 and Σ_2 such that $\Sigma_1 \leq \mathrm{Var}(\mathbf{X}|U) \leq \Sigma_2$ uniformly in U, where $\mathrm{Var}(\mathbf{X}|U)$ denotes the conditional covariance matrix of $\mathbf{X}$ given U and $A \leq B$ means that $B - A$ is a positive semi-definite matrix.

(A4) The random variables $e_1, \ldots, e_n$ are *i.i.d.* and have a density function $f_e(\cdot)$ that is continuous at 0 with $0 < f_e(0) < N < \infty$ for some positive constant N.

This set of assumptions is the same as those used by Kim [9] to develop the convergence of the unpenalized estimator for model (2.1). Now, we describe the main result of our paper.

Theorem 4.1 *Suppose that Assumptions (A1)-(A4) hold and that* $b_n \approx n^{1/(2r+1)}$. *Further, we assume that* $\lambda_n \to 0$ *and* $\lambda_n/(n^{-1/2}b_n) \to \infty$ *as* n *goes to infinity. Then, we have*

(a) with probability approaching 1, $\hat{\alpha}_k(\cdot)$ *are nonzero varying coefficients for* $k = 1, \ldots, s$ *and* $\hat{\alpha}_k(\cdot) = 0$ *for* $k = s + 1, \ldots, p$.

(b) $\|\hat{\alpha}_k - \alpha_k\|_{L_2} = O_p(n^{1/(2r+1)}),\ k = 0, 1, \ldots, s$, *where* $\|f\|_{L_2}$ *is the* L_2*-norm of the function* f.

The detailed proof of Theorem 4.1 is given in the Appendix. We briefly provide the main idea of the proof. First, to handle the issue of non-differentiability of the check function, we uniformly approximate by a quadratic function of $\boldsymbol{\gamma}$ the difference between $l_0(\boldsymbol{\gamma})$ and $l_0(\boldsymbol{\gamma}^0)$ (which is non-differentiable) when $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}^0$ differ by a term of order $n^{-1/2}b_n$. Here, $\boldsymbol{\gamma}^0$ is a coefficient vector that makes the corresponding coefficient function vector $\boldsymbol{\alpha}^0(u)$ best approximate the true one $\boldsymbol{\alpha}(u)$ within the function space under consideration. For the precise definition of $\boldsymbol{\gamma}^0$, we refer to Lemma 6.1 in

Appendix. Using this approximation, we then show asymptotic properties of our estimator with the idea of the proof in Wang et al. [23].

Part (a) of Theorem 4.1 shows that the proposed estimator consistently identifies the relevant covariates with probability tending to 1. Part (b) provides the convergence rate of the nonzero coefficient functions. If $\alpha_k(\cdot)$ has a bounded second derivative, and if piecewise linear splines with $b_n \sim n^{1/5}$ are used, then Theorem 4.1 (b) implies that $\|\hat{\alpha}_k - \alpha_k\|_{L_2} = O_p(n^{-2/5})$, which is the same as the univariate optimal rate for nonparametric regression with i.i.d. data [18]. As a result, Theorem 4.1 shows that our estimator is np-oracle.

5 Numerical Studies

In this section, we use B-splines as the basis functions in our implementation of the proposed method. Following the recommendation of Ruppert [16], in our simulation we use a common value of q_k, which we will call q as before. Thus the number of basis functions q is equal to $b + d + 1$, where b is the number of interior knots for approximating a coefficient function and d is the degree of the spline. We use equally spaced knots for all simulations in this paper. As for the degree of the spline, Kim [9] recommended the use of lower degree splines because higher degree splines would induce unnecessary interactions between the spline basis and the collinearity among variables in varying coefficient models. In our simulations, we use $d = 3$ corresponding to cubic splines.

5.1 Simulation examples

When the underlying error distribution has a fat tail, is not normal or has infinite variance, the median regression (MR) estimator that minimizes the sum of absolute errors is often considered as a robust alternative to the least squares regression (LSR) estimator in order to understand the conditional central tendency in a dataset. In this section, we compare the performance of the median regression estimator for the model (2.1) with that of the least squares regression estimator while the tails of the error becomes gradually fat using the contaminated normal distribution. The LSR estimator can be obtained by substituting the check loss with the squared loss in (2.4) and (2.5). Since the LSR estimator in the model (2.1) is regarded as a special case of the estimator in Wang et al. [23] where the number of repeated measurements is one, it is easy to see that the estimator has the same asymptotic properties as the MR estimator. For the LSR estimator, we use the Bayesian Information Criterion

(BIC) to select the tuning parameters as follows:

$$\begin{aligned} BIC^{ini}(b) &= \log \sum_{i=1}^{n} (Y_i - \mathbf{\Pi}_i^\top \tilde{\boldsymbol{\gamma}}^{LS})^2 + \frac{\log n}{n}(p+1)(b+d+1), \\ BIC^{fin}(\lambda) &= \log \sum_{i=1}^{n} (Y_i - \mathbf{\Pi}_i^\top \hat{\boldsymbol{\gamma}}_\lambda^{LS})^2 + \frac{\log n}{n} edf, \end{aligned}$$

where $\hat{\boldsymbol{\gamma}}_\lambda^{LS}$ is the estimator in mean regression and *edf* is the number of nonzero elements in $\hat{\boldsymbol{\gamma}}_\lambda^{LS}$. Theoretical background of this type of BIC criterion can be found partly in Huang and Yang [7] and the same type of BIC was also considered in Tang et al. [20] to select the tuning parameters.

In our simulations, we consider the following two models:

$$\begin{aligned} \text{Model (I)} &: Y_i = 2\sin(2\pi U_i) + 8U_i(1-U_i)X_i^{(1)} + e_i \\ \text{Model (II)} &: Y_i = 4U_i + 2\sin(2\pi U_i)X_i^{(1)} + 2X_i^{(2)} - 1.5X_i^{(3)} + e_i, \end{aligned}$$

where $(X_i^{(1)}, \ldots, X_i^{(6)})^\top$ is generated from a truncated multivariate normal distribution with mean vector consisting of all zeros and covariance matrix given by $\text{cov}(X_i^{(k_1)}, X_i^{(k_2)}) = 0.5^{|k_1 - k_2|}$ for any $1 \le k_1, k_2 \le 6$. The support of each marginal distribution is restricted to $[-5, 5]$ and e is a symmetric random error independent of the covariates. Because the median and mean of the error coincide when the error is symmetric, both the LSR and the MR estimator aim for the same quantity and hence are directly comparable. The index variable U is randomly generated from Uniform(0,1). It is clear that only the covariate $X_i^{(1)}$ is relevant in model (I) and that the covariates $X_i^{(1)}$, $X_i^{(2)}$, and $X_i^{(3)}$ are relevant in model (II). To investigate the effect of relatively heavy tail error distributions, we consider the contaminated normal distribution $CN(\rho; \sigma_1, \sigma_2)$ that is the mixture of $N(0, \sigma_1^2)$ and $N(0, \sigma_2^2)$ with weights $1 - \rho$ and ρ, respectively. To be more specific for our simulations, we set $\sigma_1 = 1$ for model (I) and $\sigma_1 = 1.5$ for model (II). For both models, we set $\sigma_2 = 5$ and take the values of ρ equal to 0, 0.1 and 0.2. As ρ increases, the error distribution gradually changes from a normal distribution to a relatively heavy tail distribution. For each model, we generate two random samples of $n = 200$ and $n = 400$ and then repeat the simulations 500 times.

To compare the MR and LSR estimates for $\boldsymbol{\alpha}(u)$, we first calculate the absolute deviation error (ADE) defined as

$$\text{ADE}(\hat{\boldsymbol{\alpha}}) = \frac{1}{n_{grid}} \sum_{j=1}^{p} \sum_{r=1}^{n_{grid}} |\hat{\alpha}_j(u_r) - \alpha_j(u_r)|,$$

where the u_r's are the equally spaced grid points on the support of U with $n_{grid} = 201$. We define the

relative absolute deviation error (RADE) by

$$\mathrm{RADE}(\hat{\boldsymbol{\alpha}}) = \frac{\mathrm{ADE}\,(\hat{\boldsymbol{\alpha}}_{MR})}{\mathrm{ADE}\,(\hat{\boldsymbol{\alpha}})}$$

for an estimator $\hat{\boldsymbol{\alpha}}$ of the coefficient function vector, where $\hat{\boldsymbol{\alpha}}_{MR}(\cdot)$ is the MR estimator. Table 1 shows the sample mean and standard deviation of the RADE with respect to the LSR and unpenalized MR (uMR) estimators over 500 simulations. From the RADE with respect to the LSR estimator, we see that the LSR estimator slightly outperforms the MR estimator in terms of the estimation accuracy when the error is normal ($\rho = 0$). However, when the error distribution becomes relatively heavier which is the case when $\rho = 0.1$ or 0.2, the performance of the LSR estimator deteriorates more rapidly than that of the MR estimator. This shows that the quantile regression estimator in the VC model is more useful than the least squares estimator when the error deviates from a normal distribution. The RADE with respect to the uMR estimator implies that the estimation efficiency has substantially improved by the variable selection. Additionally, when comparing models (I) and (II) in terms of the RADE with respect to the uMR, we observe that the sparser model has more benefits from the variable selection.

The results of variable selection for each estimator is given in Table 2. As a performance measure of the variable selection, we report in the column "Correct" the average number of irrelevant coefficient functions correctly estimated to be zero functions and in the column "Incorrect" the average number of relevant coefficient functions incorrectly estimated to be zero functions. In another aspect, we summarize the variable selection results by discriminating three different situations. In the column "Correct fit", we present the percentage of trials for which the estimated model coincides with the true model. When the estimated model misses at least one relevant covariate (even if some irrelevant covariates are selected), we count it as an underfitted model and report its percentage in the column "Underfit". Finally, we show in the column "Overfit" the percentage of overfitted cases in which the estimated model includes all relevant covariates and some irrelevant covariates. Overall, similar conclusions are obtained in terms of variable selection as in the result of estimation accuracy. Additionally, Table 2 suggests that both the MR and the LSR estimator have a tendency to overfit more than to underfit. A possible explanation for this phenomenon is that both estimators determine the amount of penalization for variable selection using the estimator of the L_2 norm of each coefficient. When the sample size is relatively small, the instability of estimation could bring out a situation where the L_2 norm estimates of some irrelevant coefficients are larger than those of other irrelevant ones, which results in overfit. However, even in such a situation, the L_2 norm estimates of relevant coefficients are

not likely to be smaller than those of irrelevant ones so underfit is less frequent than overfit. Similar results can be found in Wang and Xia [21], who also uses the L_2 norm estimate of each coefficient for variable selection.

Throughout the simulation, we observed that the performance of our estimator is fairly dependent on that of the initial estimator because our variable selection method uses the initial estimator to determine the amount of shrinkage for each coefficient function. However, this dependency can be eliminated in a certain degree by using the final estimate as a new initial estimator if necessary. An example of this kind of iterations is found in Tang et al. [20].

5.2 Application to forced expiratory volume data

The forced expiratory volume (FEV) is the amount of air which is forcibly exhaled from the lungs in the first second. The FEV data from Rosner [15] contain measurements of FEV in liters, age (U) in years, height (H) in inches, sex ($S = 0$ for girls/$S = 1$ for boys), and smoking status ($SM = 1$ for a regular smoker/$SM = 0$ otherwise) which were collected from 654 children aged 3-19. To assess the effect of smoking status on FEV and the lower conditional quantiles of FEV as a measure for poor pulmonary functioning, Kim [9] considered the VC model (2.1) in quantile regression as follows:

$$q_\tau = \beta_0(u) + \beta_1(u) \times S + \beta_2(u) \times SM + \beta_3(u) \times H + \beta_4(u) \times (H \times S). \tag{5.1}$$

In this subsection, we illustrate our proposed method by showing how it selects relevant covariates in model (5.1) with respect to different quantile levels.

We consider the conditional first vigintile ($q_{0.2}$) and median ($q_{0.5}$) for the analysis. In order to detect smoking effects accurately, we only consider children over the age 10 (345 subjects) because there are no smokers among the children below the age 10. We use piecewise linear ($d = 1$) and quadratic ($d = 2$) splines with equispaced knots. Since the number of covariates in model (5.1) is relatively small, choosing different number of knots for each coefficient function is computationally allowed. Hence, we determine the degree of splines and the number of equispaced knots for each coefficient function simultaneously using the SIC values in (3.2).

Through the proposed procedure, all the covariates except the smoking status are selected as relevant for both quantiles. FEV is known to increase in accordance with the body growth, which is measured by the height to a certain extent. Typically boys tend to have a larger FEV on average than girls as they grow. Based on this information, our selection results of covariates coincide with

the previous results commonly known in the medical science. Figure 1 shows the coefficient function estimates for the selected covariates. Since only a small number of data are available beyond the age of 15, we restrict our attention to the estimates below the age of 15 in Figure 1. At both quantiles, the estimated baseline function $\beta_0(u)$ increases linearly, which implies that FEV increases with age. This makes sense because all the children in the data were in the middle of growth. Regarding the effect of sex on FEV, the difference between boys and girls becomes bigger in both quantiles if we ignore the age range beyond 15.

One of the interesting results in our analysis is that the relevance of smoking status on FEV is found to be different according to the quantile level. The smoking status is not selected for the median level ($q_{0.5}$), while it is selected for the relatively low quantile ($q_{0.2}$). In Figure 2, we observe that the coefficient function of smoking status in the first vigintile is negative during all the period except at age 10 and for the age range 17-18. The coefficients in those ranges are not reliable because the smoking period of a 10-year old boy is supposed to be very short and few observations exist in the age range 17-18. Consequently, our result shows that the negative effect of smoking on FEV is not clear at the median level. However, smoking smoking should at least be considered as a relevant factor which has some negative effects on the group who has weak lungs, especially at the period of growth. This kind of heterogeneity would not have been revealed by variable selection procedures for mean regression. However, to investigate the relationship between the smoking status and FEV more rigorously, we need to know not only the smoking status of each subject but also the years they have been smoking for. This relationship should be checked again with such data.

6 Conclusion and Future Work

In this paper, we proposed an efficient variable selection method for varying coefficient models in quantile regression when all coefficient functions are assumed to be varying. Theoretically, we showed that our method has a nonparametric oracle property, which is desirable as a variable selection procedure. From a practical point of view, we illustrated through a simulation study that when the error is asymmetric or has a fat tail, our estimator is more robust than the one in mean regression in terms of the consistency of the variable selection method and the efficiency of the estimation procedure.

For more efficient estimation through variable selection, it is also important to separate varying and constant coefficients among the nonzero coefficients. For that purpose, it is possible to use the hypothesis testing given in Section 4 of Wang et al. [22] after identifying the nonzero coefficients based

on our method.

Finally we addressed variable selection issues of varying coefficient models under the assumption that there are no repeated measurements for each subject. To extend our work to a longitudinal data setting seems a promising and useful project for practitioners. We leave it as a future work.

Acknowledgements

The authors would like to thank the Associate Editor, one referee and Dr. Eun Ryung Lee for their valuable suggestions, which much improved the paper. H. Noh and I. Van Keilegom acknowledge financial support from IAP research network P6/03 of the Belgian Government (Belgian Science Policy). K. Chung acknowledges financial support by the Hongik University new faculty research support fund. I. Van Keilegom additionally acknowledges financial support from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement No. 203650, and from the contract 'Projet d'Actions de Recherche Concertées' (ARC) 11/16-039 of the 'Communauté française de Belgique', granted by the 'Académie universitaire Louvain'.

References

[1] F. Alizadeh and D. Goldfarb. Second-order cone programming. *Mathematical Programming*, B95: 3–51, 2003.

[2] Z. Cai and X. Xu. Nonparametric quantile estimations for dynamic smooth coefficient models. *Journal of the American Statistical Association*, 103:1595–1608, 2008.

[3] C. de Boor. *A Practical Guide to Splines*. Springer-Verlag, 1978.

[4] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 99:710–723, 2001.

[5] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 1.21. `http://cvxr.com/cvx`, Apr. 2011.

[6] T. Honda. Quantile regression in varying coefficient models. *Journal of Statistical Planning and Inference*, 121:113–125, 2004.

[7] J. Z. Huang and L. Yang. Identification of non-linear additive autoregressive models. *Journal of the Royal Statistical Society*, B66:463–477, 2004.

[8] B. Kai, R. Li, and H. Zou. New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models. *Annals of Statistics*, 39:305–332, 2011.

[9] M. O. Kim. Quantile regression with varying coefficients. *Annals of Statistics*, 35:92–108, 2007.

[10] R. Koenker, P. NG, and S. Portnoy. Quantile smoothing splines. *Biometrika*, 81:673–680, 1994.

[11] E. R. Lee and H. Noh. Model selection via Bayesian information critetion for quantile regression models. Technical report, Universit catholique de Louvain, 2012.

[12] Y. Li and J. Zhu. L1-norm quantile regression. *Journal of Computational and Graphical Statistics*, 17:163–185, 2008.

[13] M. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret. Applications of second-order cone programming. *Linear Algebra and its Applications*, 284:193–228, 1998.

[14] H. Noh and B. Park. Sparse varying coefficient models for longitudinal data. *Statistica Sinica*, 20:1183–1202, 2010.

[15] B. Rosner. *Fundamentals of Biostatistics*. Duxbury, 2000.

[16] D. Ruppert. Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, 11:735–757, 2002.

[17] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[18] C. J. Stone. Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, 10:1348–1360, 1982.

[19] C. B. Storlie, H. D. Bondell, B. J. Reich, and H. H. Zhang. Surface estimation, variable selection, and the nonparametric oracle property. *Statistica Sinica*, 21:679–705, 2011.

[20] Y. Tang, H. J. Wang, Z. Zhu, and X. Song. A unified variable selection approach for varying coefficient models. *Statistica Sinica*, 22:601–628, 2012.

[21] H. Wang and Y. Xia. Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association*, 104:747–757, 2009.

[22] H. J. Wang, Z. Zhu, and J. Zhou. Quantile regression in partially linear varying coefficient models. *Annals of Statistics*, 37:3841–3866, 2009.

[23] L. Wang, H. Li, and J. Z. Huang. Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association*, 103: 1556–1569, 2008.

[24] L. Xue. Variable selection in additive models. *Statistica Sinica*, 19:1281–1296, 2009.

[25] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B*, 68:49–67, 2006.

[26] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006.

[27] H. Zou and R. Li. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36:1509–1533, 2008.

Appendix

For the proof of Theorem 4.1, we recall three lemmas from Kim [9].

Lemma 6.1 *[9, Lemma A.1] Assume that (A1)-(A4) hold. Then, there exists a spline coefficient vector* $\boldsymbol{\gamma}^0 = ({\boldsymbol{\gamma}_0^0}^\top, \ldots, {\boldsymbol{\gamma}_p^0}^\top)^\top$ *and some positive constants* W_1 *and* W_1^* *that depend only on* m, W_0, p *and* M *(but not necessarily all of them), such that:*

(a) $\sup_{u\in[0,1]} |\alpha_k(u) - \boldsymbol{\pi}(u)^\top \boldsymbol{\gamma}_k^0| \leq W_1 b_n^{-r}$

(b) $\sup_{(u,\mathbf{X})\in[0,1]\times\mathbb{R}^{p+1}} |\boldsymbol{\Pi}(u,\mathbf{X})^\top \boldsymbol{\gamma}_0 - \mathbf{x}^\top \alpha(u)| \leq W_1^* b_n^{-r}$

Lemma 6.2 *[9, Lemma A.4] Define* $H_n = \sum_{i=1}^n \boldsymbol{\Pi}_i \boldsymbol{\Pi}_i^\top$. *Assume that (A1)-(A4) hold and that* $\lim_{n\to\infty} b_n n^{\delta-1} = 0$ *for some* $0 < \delta < 1$. *Then, the eigenvalues of* $n^{-1} b_n H_n$ *are uniformly bounded away from zero and infinity in probability.*

Lemma 6.3 *[9, Lemma A.7 (i)(ii)]*

(i) For any sequence $\{L_n\}$ *satisfying* $1 \le L_n \le b_n^{\eta_0/10}$ *for some* $0 < \eta_0 < (r-1/2)/(2r+1)$*, we have:*

$$\sup_{(\boldsymbol{\gamma}-\boldsymbol{\gamma}^0)^\top H_n(\boldsymbol{\gamma}-\boldsymbol{\gamma}^0)\le L_n^2 b_n} b_n^{-1}\left|\sum_{i=1}^n \Big\{\rho(e_i-\boldsymbol{\Pi}_i^\top(\boldsymbol{\gamma}-\boldsymbol{\gamma}^0)-R_{ni})-\rho(e_i-R_{ni})+\boldsymbol{\Pi}_i^\top(\boldsymbol{\gamma}-\boldsymbol{\gamma}^0)(2\tau-2I(e_i<0)) \right.$$
$$\left. -\mathrm{E}_e\Big(\rho(e_i-\boldsymbol{\Pi}_i^\top(\boldsymbol{\gamma}-\boldsymbol{\gamma}^0)-R_{ni})-\rho(e_i-R_{ni})\Big)\Big\}\right| = o_p(1), \tag{6.1}$$

where E_e *stands for the conditional expectation given* $(\mathbf{X}_i, U_i),\ i=1,\ldots,n$ *and* $R_{ni} = \boldsymbol{\Pi}_i^\top\boldsymbol{\gamma}^0 - \mathbf{X}_i\alpha(U_i)$*.*

(ii) For any $\epsilon > 0$*, there exists* $L \equiv L_\epsilon$ *(sufficiently large) such that as* $n \to \infty$*,*

$$P\left\{b_n^{-1}\left(\inf_{(\boldsymbol{\gamma}-\boldsymbol{\gamma}^0)^\top H_n(\boldsymbol{\gamma}-\boldsymbol{\gamma}^0)\le L b_n}\sum_{i=1}^n\Big[\mathrm{E}_e(\rho(e_i-\boldsymbol{\Pi}_i^\top(\boldsymbol{\gamma}-\boldsymbol{\gamma}^0)-R_{ni})-\rho(e_i-R_{ni}))\Big]\right.\right.$$
$$\left.\left. -\left|\sum_{i=1}^n\boldsymbol{\Pi}_i(2\tau-2I(e_i<0))\right|\right)>1\right\} > 1-\epsilon \tag{6.2}$$

Lemma 6.4 *Under the same assumptions as in Theorem 4.1, we have that* $\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^0\|_2 = O_p(n^{-1/2}b_n)$*.*

Proof of Lemma 6.4.

From Theorem 1 in Kim [9], it follows that $\|\tilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^0\|_2 = O_p(n^{-1/2}b_n)$. Since $\boldsymbol{\gamma}_k^0 = \mathbf{0},\ k=s+1,\ldots,p$, from (2.6) we have:

$$\|\tilde{\boldsymbol{\gamma}}_k\|_2 = O_p(b_n^{1/2}),\ 1\le k\le s \tag{6.3}$$
$$\|\tilde{\boldsymbol{\gamma}}_k\|_2 = O_p(b_n^{1/2}\max\{n^{-1/2}b_n^{1/2}, b_n^{-r}\}),\ s+1\le k\le p. \tag{6.4}$$

Assume that $\|\boldsymbol{\gamma}-\boldsymbol{\gamma}^0\|_2 = C_1 n^{-1/2}b_n$ and that C_1 is large enough. From (6.3), it follows that $\|\tilde{\boldsymbol{\gamma}}_k\|_2 > a\lambda_n$ for all $1\le k\le s$ with probability tending to one, where a appears in the definition of p_{λ_n}. This means that, with probability tending to one, $p'_{\lambda_n}(\|\tilde{\boldsymbol{\gamma}}_k\|_2)=0$ for all $1\le k\le s$. Since $\|\boldsymbol{\gamma}_k\|_2 - \|\boldsymbol{\gamma}_k^0\|_2 \le \|\boldsymbol{\gamma}-\boldsymbol{\gamma}^0\|_2 = O_p(n^{-1/2}b_n) = o_p(1)$, it follows from the definition of p_{λ_n} that

$$n\sum_{k=1}^s p'_{\lambda_n}(\|\tilde{\boldsymbol{\gamma}}_k\|_2)(\|\boldsymbol{\gamma}_k\|_2-\|\boldsymbol{\gamma}_k^0\|_2) = o_p(b_n) \ \text{ and } \ n\sum_{k=s+1}^p p'_{\lambda_n}(\|\tilde{\boldsymbol{\gamma}}_k\|_2)(\|\boldsymbol{\gamma}_k\|_2-\|\boldsymbol{\gamma}_k^0\|_2)\ge 0 \tag{6.5}$$

because $\|\boldsymbol{\gamma}_k^0\|_2 = 0$ for $k=s+1,\ldots,p$. Define $Q_n(\boldsymbol{\gamma}) = \sum_{i=1}^n[\rho(Y_i-\boldsymbol{\Pi}_i^\top\boldsymbol{\gamma})-\rho(Y_i-\boldsymbol{\Pi}_i^\top\boldsymbol{\gamma}^0)] = \sum_{i=1}^n[\rho(e_i-\boldsymbol{\Pi}_i^\top(\boldsymbol{\gamma}-\boldsymbol{\gamma}^0)-R_{ni})-\rho(e_i-R_{ni})]$. By Lemma 6.3, we have

$$P\left(\inf_{\|\boldsymbol{\gamma}-\boldsymbol{\gamma}^0\|_2 = C_1 n^{-1/2}b_n} Q_n(\boldsymbol{\gamma}) > b_n\right)\to 1,$$

and hence it follows from (6.5) that

$$P\left(\inf_{\|\boldsymbol{\gamma}-\boldsymbol{\gamma}^0\|_2 = C_1 n^{-1/2}b_n} (l(\boldsymbol{\gamma})-l(\boldsymbol{\gamma}^0)) > 0\right)\to 1,$$

as $n \to \infty$. By the convexity of $l(\boldsymbol{\gamma})$ and the fact that $l(\hat{\boldsymbol{\gamma}}) - l(\boldsymbol{\gamma}) \leq 0$, there exists some C_ξ, for any $\xi > 0$, such that as $n \to \infty$,

$$P(\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^0\|_2 \leq C_\xi n^{-1/2} b_n) > 1 - \xi.$$

□

Proof of Theorem 4.1 (a) Different from the mean regression, the gradient function of $l_0(\boldsymbol{\gamma}) = \sum_{i=1}^n \rho(Y_i - \mathbf{\Pi}_i^\top \boldsymbol{\gamma})$ is not useful to establish asymptotic properties of the estimator $\hat{\boldsymbol{\gamma}}$. The reason is because the check function is not differentiable at zero and some of the estimated residuals are exactly zero in quantile regression. Because of these two reasons, the gradient function of $l_0(\boldsymbol{\gamma})$ evaluated at $\hat{\boldsymbol{\gamma}}$ is not directly applicable since it involves the first derivative of the check function evaluated at each residual $Y_i - \mathbf{\Pi}_i^\top \hat{\boldsymbol{\gamma}}$. As an alternative, we may consider the subgradient function of $l_0(\boldsymbol{\gamma})$. However, the subgradient function of $l_0(\boldsymbol{\gamma})$ is so complicated that it is very difficult to derive easy-to-check optimality conditions. Even though Tang et al. [20] seemed to derive the concise optimality conditions as in the case of mean regression when penalizing by the Euclidean norm of $\boldsymbol{\gamma}_k$, their optimality conditions are not correct. So in our work we derive the consistency in variable selection using a certain lower bound of the difference of two check loss functions. Since the selection consistency regarding the relevant coefficients is clear from Theorem 4.1 (b), which we will show, we focus on the case of the irrelevant coefficients.

Suppose that there exists a $(s+1) \leq k_0 \leq p$ such that the probability of $\hat{\alpha}_{k_0}(\cdot)$ being a zero function does not converge to one. Then, there exists $\epsilon > 0$ such that, for infinitely many n,

$$P(\hat{\boldsymbol{\gamma}}_{k_0} \neq \mathbf{0}) = P(\hat{\alpha}_{k_0} \neq 0) \geq \epsilon.$$

Let $\boldsymbol{\gamma}^*$ be the vector obtained from $\hat{\boldsymbol{\gamma}}$ with $\hat{\boldsymbol{\gamma}}_{k_0}$ being replaced by $\mathbf{0}$. It will be shown that there exists a $\eta > 0$ such that $l(\hat{\boldsymbol{\gamma}}) - l(\boldsymbol{\gamma}^*) > 0$ with probability at least η for infinitely many n, which contradicts with the fact that $l(\hat{\boldsymbol{\gamma}}) - l(\boldsymbol{\gamma}^*) \leq 0$.

From Lemma 6.4 and because $n^{-1/2} b_n \ll \lambda$, we may assume that $\nu_k = \lambda$ for $s+1 \leq k \leq p$. Since $\rho(u) - \rho(v) \geq 2(\tau - I(v \leq 0))(u - v)$ for any $u, v \in \mathbb{R}$, we have

$$\begin{aligned}
& l(\hat{\boldsymbol{\gamma}}) - l(\boldsymbol{\gamma}^*) \\
& \geq -\sum_{i=1}^n (2\tau - 2I(e_i \leq 0)) \mathbf{\Pi}_i^\top (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*) - 2\sum_{i=1}^n (I(e_i \leq 0) - I(Y_i \leq \mathbf{\Pi}_i^\top \boldsymbol{\gamma}^*)) \mathbf{\Pi}_i^\top (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*) + n\lambda \|\hat{\boldsymbol{\gamma}}_{k_0}\|_2 \\
& \geq \left(-2 \left\| \sum_{i=1}^n (\tau - I(e_i \leq 0)) \mathbf{\Pi}_i \right\|_2 - 2 \left\| \sum_{i=1}^n (I(e_i \leq 0) - I(e_i \leq r_{ni})) \mathbf{\Pi}_i \right\|_2 + n\lambda \right) \|\hat{\boldsymbol{\gamma}}_{k_0}\|_2, \qquad (6.6)
\end{aligned}$$

where $r_{ni} = R_{ni} + \mathbf{\Pi}_i^\top(\hat{\boldsymbol{\gamma}}^* - \boldsymbol{\gamma}^0)$.

From assumptions (A3) and (A4), we obtain that for any $L > 0$,

$$\begin{aligned}
& E\sum_{k=0}^{p}\sum_{l=1}^{q_n}\left\{\sum_{i=1}^{n}(I(e_i \le Ln^{-1/2}b_n) - I(e_i \le -Ln^{-1/2}b_n))|X_i^{(k)}B_{kl}(U_i)|\right\}^2 \\
\le\ & \sum_{k=1}^{p}\sum_{l=1}^{q_n} nM^2 E\left\{(I(e \le Ln^{-1/2}b_n) - I(e \le -Ln^{-1/2}b_n))|B_{kl}(U)|\right\}^2 \\
& \quad + \sum_{k=1}^{p}\sum_{l=1}^{q_n} n(n-1)M^2\left\{E(I(e \le Ln^{-1/2}b_n) - I(e \le -Ln^{-1/2}b_n))|B_{kl}(U)|\right\}^2 \\
\le\ & M^2\left\{n(2Ln^{-1/2}b_nN) + n^2(2Ln^{-1/2}b_nN)^2\right\} = O(nb_n^2).
\end{aligned}$$

This implies that

$$\left\|\sum_{i=1}^{n}(I(e_i \le 0) - I(e_i \le r_{ni}))\mathbf{\Pi}_i\right\|_2 = O(n^{1/2}b_n) \tag{6.7}$$

because $\max_{1\le i\le n}|r_{ni}| \le O(b_n^{-r}) + \|\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}^0\| = O_p(n^{-1/2}b_n)$. By simple calculations, one has that $\|\sum_{i=1}^{n}(\tau - I(e_i < 0))\mathbf{\Pi}_i\|_2 = O_p(n^{1/2})$. From this fact combined with (6.7) and $\lambda/(n^{-1/2}b_n) \to \infty$, we can conclude that $n\lambda\|\hat{\boldsymbol{\gamma}}_{k_0}\|_2$ dominates the other terms in (6.6), which contradicts to $l(\hat{\boldsymbol{\gamma}}) - l(\boldsymbol{\gamma}^*)$ being negative. □

Proof of Theorem 4.1 (b) It is easy to see that, for all $k = 0, 1, \ldots, s$,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}[\hat{\alpha}_k(U_i) - \alpha_k(U_i)]^2 &\le \frac{2}{n}\sum_{i=1}^{n}[\boldsymbol{\pi}_i^\top(\hat{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma}_k^0)]^2 + \frac{2}{n}\sum_{i=1}^{n}R_{ni}^2 \\
&\le \frac{2}{n}(\hat{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma}_k^0)^\top\sum_{i=1}^{n}\boldsymbol{\pi}_i\boldsymbol{\pi}_i^\top(\hat{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma}_k^0) + 2Cb_n^{-2r}.
\end{aligned}$$

By Lemma 6.2 and 6.4, the desired result is obtained. □

Table 1: Sample mean and standard deviation of the RADE with respect to LSR and uMR estimators

			RADE	
Model	n	Error	wrt LSR	wrt uMR
I	$n = 200$	$\rho = 0$	1.706(1.351)	0.168(0.106)
		$\rho = 0.1$	0.913(0.947)	0.190(0.128)
		$\rho = 0.2$	0.698(0.802)	0.221(0.164)
	$n = 400$	$\rho = 0$	1.369(0.754)	0.137(0.052)
		$\rho = 0.1$	0.811(0.495)	0.139(0.061)
		$\rho = 0.2$	0.610(0.456)	0.145(0.079)
II	$n = 200$	$\rho = 0$	1.363(0.421)	0.553(0.151)
		$\rho = 0.1$	1.044(0.418)	0.570(0.168)
		$\rho = 0.2$	0.933(0.503)	0.641(0.248)
	$n = 400$	$\rho = 0$	1.262(0.281)	0.501(0.101)
		$\rho = 0.1$	0.973(0.263)	0.499(0.110)
		$\rho = 0.2$	0.880(0.279)	0.517(0.121)

Table 2: Variable selection results of LSR and MR estimators

			No. of estimated zeros				Proportion of models					
			Correct		Incorrect		Underfit		Correct fit		Overfit	
Model	n	Error	MR	LSR	MR	LSR	MR	LSR	MR	LSR	MR	LSR
I	$n = 200$	$\rho = 0$	4.750	4.912	0.000	0.000	0.00	0.00	81.00	92.80	19.00	7.20
		$\rho = 0.1$	4.568	4.356	0.000	0.026	0.00	2.60	70.20	55.80	29.80	41.60
		$\rho = 0.2$	4.412	4.268	0.012	0.140	1.20	14.00	59.00	40.00	39.80	46.00
	$n = 400$	$\rho = 0$	4.980	5.000	0.000	0.000	0.00	0.00	98.00	100.00	2.00	0.00
		$\rho = 0.1$	4.950	4.808	0.000	0.000	0.00	0.00	95.40	84.40	4.60	15.60
		$\rho = 0.2$	4.888	4.574	0.000	0.012	0.00	1.20	90.60	67.20	9.40	31.60
II	$n = 200$	$\rho = 0$	2.570	2.766	0.002	0.002	0.20	0.20	67.40	79.80	32.40	20.00
		$\rho = 0.1$	2.446	2.506	0.008	0.062	0.80	4.60	58.20	58.00	41.00	37.40
		$\rho = 0.2$	2.384	2.386	0.100	0.236	8.00	16.60	49.40	42.80	42.60	40.60
	$n = 400$	$\rho = 0$	2.908	2.964	0.000	0.000	0.00	0.00	91.40	96.60	8.60	3.40
		$\rho = 0.1$	2.870	2.848	0.000	0.000	0.00	0.00	87.80	86.40	12.20	13.60
		$\rho = 0.2$	2.784	2.752	0.000	0.004	0.00	0.40	81.20	78.20	18.80	21.40

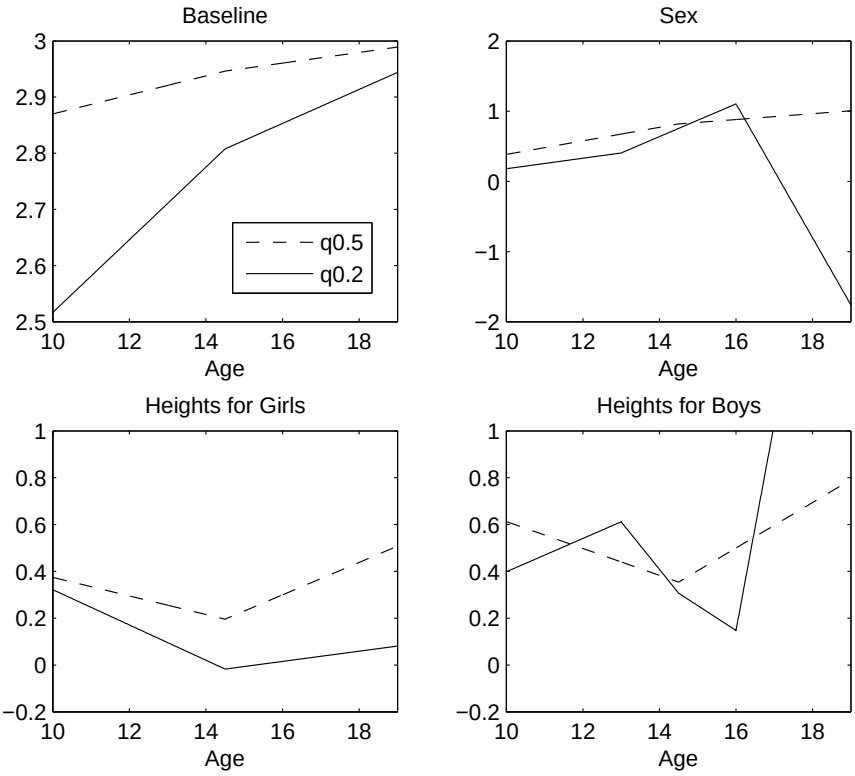

Figure 1: Coefficient estimates by piecewise linear splines

Figure 2: Coefficient estimates of smoking status when $q = 0.5$ and 0.2