INSTITUT DE STATISTIQUE
BIOSTATISTIQUE ET
SCIENCES ACTUARIELLES
(ISBA)

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

DISCUSSION
PAPER

2012/21

COMPONENT SELECTION IN ADDITIVE QUANTILE REGRESSION MODELS

NOH, H. and E.R. LEE

Component Selection in Additive Quantile Regression Models

Hohsuk Noh *
Université catholique de Louvain

Eun Ryung Lee †
University of Mannheim

July 30, 2012

Abstract

Nonparametric additive models are powerful techniques for multivariate data analysis. Although many procedures have been developed for estimating additive components both in mean regression and quantile regression, the problem of selecting relevant components has not been addressed much especially in quantile regression. In this article, we present a doubly-penalized estimation procedure for component selection in additive quantile regression models that combines basis function approximation with a variant of the smoothly clipped absolute deviation penalty and a ridge-type penalty. We show that the proposed estimator identifies relevant and irrelevant components consistently and achieves the nonparametric optimal rate of convergence for the relevant components. We also provide some numerical evidence of the estimator, and illustrate its usefulness through a real data example to identify important body measurements to predict percentage of body fat of an individual.

Key words: component selection, additive quantile regression, penalized estimation, nonparametric regression.

1 Introduction

Suppose that Y is a response of interest which depends on a vector of random covariates $\mathbf{X} = (X^1, \ldots, X^p) \in \mathbb{R}^p,\ p \geq 2$. We are interested in the conditional quantile of Y given $\mathbf{X}$. The re-

*H. Noh acknowledges financial support from ERC Grant agreement No. 203650.

†E. R. Lee acknowledges financial support from the Research Center (SFB) 884 "Political Economy of Reforms" (Project A3), funded by the German Research Foundation (DFG).

lationship between Y and $\mathbf{X}$ can be modeled as

$$Y = Q_\tau(\mathbf{X}) + U_\tau,$$

where $Q_\tau(\cdot)$ is an unknown real-valued function and U_τ is an unobserved random variable whose τth quantile conditional on $\mathbf{X} = \mathbf{x}$ is 0 for almost every $\mathbf{x}$. Many non-parametric methods are available to estimate $Q_\tau(\cdot)$ yet they are all severely affected by the so-called curse of dimensionality when p is large. To tackle this issue, several authors have proposed dimensional reduction methods, a popular one of which is to consider the following model based on the additive structure of $Q_\tau(\mathbf{X})$:

$$Y_i = Q_\tau(\mathbf{X}_i) + U_{i,\tau} = \mu_\tau + \sum_{k=1}^{p} m_{k,\tau}(X_i^k) + U_{i,\tau} \tag{1.1}$$

where $(Y_i, \mathbf{X}_i, U_{i,\tau}) \in \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}$ for $1 \leq i \leq n$ are independent and identically distributed and $P(U_{i,\tau} \leq 0 | \mathbf{X}_i = \mathbf{x}_i) = \tau$ for almost every $\mathbf{x}_i$. To fit this additive quantile regression model (1.1), there have been many proposals including De Gooijer and Zerom (2003), Yu and Lu (2004), Horowitz and Lee (2005), Lee et al. (2010) and Cheng et al. (2011). Most of them have more or less similar implication that it is possible to estimate the additive component functions in multi-dimension settings at the same accuracy as in univariate cases. However, all these works are based on the assumption that the given covariates are all relevant to the conditional quantile of interest, which is not always the case in practice. Motivated by this concern, we consider the problem of selecting the relevant covariates in the model (1.1).

Although there have been a few works in the literature which deal with problem of covariate selection in additive mean regression framework such as Huang and Yang (2004), Xue (2009) and Fan et al. (2011), to the best of our knowledge such works do not exist in the context of quantile regression. The only exception that we have found is the empirical analysis of Fenske et al. (2011) about child malnutrition data in India. They adapted an idea of Bühlmann and Yu (2003) (in framework of boosting with squared error loss) for covariate selection in additive mean regression models to their semi-parametric quantile regression models with additive structures in order to identify the relevant factors (covariates) in the data, but did not discuss its statistical properties. Contrary to them, in this paper we not only propose a method to identify the relevant covariates in the model (1.1) and estimate the corresponding additive component functions, but also derive its theoretical properties. Because the dependence pattern between the response and covariates can be different across the quantile levels in quantile regression, our work has quite different implications from the works for covariate selection

in additive mean regression models. From a point of view of component selection, in a regression model with a heteroscedastic error, the set of relevant component functions for the conditional quantile of interest can vary across the quantile levels. This is the characteristic feature of component selection in quantile regression, which differentiate itself from that in mean regression. It is the main motivation for this work.

The main idea of component selection in our work is that the magnitude of the L_2-norm of m_k is closely related to relevance of the covariate X^k. Motivated by this observation, we propose two-stage procedure for estimating $m_{k,\tau}(\cdot)$ simultaneously with the identification of irrelevant covariates. The initial estimator is computed for estimating $\|m_k\|_{L_2}$. Then we obtain the final estimator which automatically discards presumably irrelevant component functions by referencing to the estimates of the L_2 norm of each component, $\|\tilde{m}_k\|_{L_2},\ k = 1, \ldots, p$, in a lasso-type penalized framework with basis approximation. This work can be seen as an extension of the idea of Zou and Li (2008) and Zou (2006) for parametric linear models to nonparametric additive quantile regression models. Since we focus on the problem of component selection, we use our initial estimator only in the first stage for estimating $\|m_k\|_{L_2}$. However, we found that the initial estimator shows more stable numerical performance than the estimator discussed by Horowitz and Lee (2005) for the same model with basis approximation. We will discuss further the comparison between our initial estimator and the one of Horowitz and Lee (2005) in Section 4 and 5.

The rest of the paper is organized as follows. Section 2 describes our initial and final estimators. Section 3 shows how the proposed estimators can be implemented. In Section 4, the asymptotic properties of the estimators are established including the consistency of the final estimator in component selection. We provide the finite sample properties of the estimators via a simulation study and illustrate the use of the estimator while analyzing real data for body fat percentage in Section 5. Technical details are given in the Appendix. Regarding the notation, we use subscripts to represent observations of random variables and superscripts to index components of vectors. Additionally we suppress the subscript "τ" whenever no confusion is caused.

2 Description of the Estimators

In this section, first we propose an estimator for $\|m_k\|_{L_2}$ based on the idea of P-spline in Eilers and Marx (1996). Then we will explain how to use it in component selection when constructing the final estimator. We assume the support $\mathcal{X}_k$ of X^k is compact for $k = 1, \ldots, p$ and simply let $\mathcal{X}_k = [0, 1]$.

Additionally, we assume that $m_k(\cdot)$ is centered in the sense that

$$\int_{\mathcal{X}_k} m_k(x)\,dx = 0,\ k = 1, \ldots, p. \tag{2.1}$$

This assumption guarantees identifiability of the model (1.1) (in L_2 sense) because there exists an universal constant $c > 0$ such that for any $\mu \in \mathbb{R}$ and any p functions $m_k : [0,1] \to \mathbb{R},\ k = 1, \ldots, p$ satisfying (2.1), $\|\mu + \sum_{k=1}^p m_k\|_{L_2} \geq c(|\mu| + \sum_{k=1}^p \|m_k\|_{L_2})$. Hereafter, we let $\|f\|_{L_2} = (\int f^2)^{1/2}$ be the L_2 norm of a function f. Since we are focusing on situations where the problem of component selection should be considered, we assume that only s covariates among the X_i^k's are relevant in the model (1.1). It is unknown which s covariates are relevant, nor what the value of s is. Without loss of generality, we let $m_k(\cdot),\ k = 1, \ldots, s$ be the nonzero component functions and $m_k(\cdot),\ k = s+1, \ldots, p$ be identically zero.

2.1 The initial estimator $\tilde{m}_k$ without component selection

We suppose that each additive component $m_k(x),\ k = 1, \ldots, p$, can be approximated by a set of basis functions $\boldsymbol{\pi}^k = (B_{k1}(\cdot), \ldots, B_{kq_k}(\cdot))^\top$ which are centered in the sense of (2.1) and orthonormal, that is,

$$m_k(x) \approx \sum_{l=1}^{q_k} \beta_{kl} B_{kl}(x),\ k = 1, \ldots, p. \tag{2.2}$$

Obviously, then the approximation $\sum_{l=1}^{q_k} \beta_{kl} B_{kl}(\cdot)$ of $m_k(\cdot)$ for $k = 1, \ldots, p$ is also centered like $m_k(\cdot)$. The linear space spanned by $\boldsymbol{\pi}^k = (B_{k1}(\cdot), \ldots, B_{kq_k}(\cdot))^\top$ approximates a function space which is assumed to contain m_k. Conditions that the basis functions should satisfy are given in Section 4. Following the approximation (2.2), the model (1.1) can be rewritten as

$$Y_i \approx \beta_0 + \sum_{k=1}^{p} \sum_{l=1}^{q_k} \beta_{kl} B_{kl}(X_i^k) + U_i. \tag{2.3}$$

The parameters β_0 and β_{kl} in the basis approximation can be estimated by minimizing

$$l_0(\boldsymbol{\beta}) = \sum_{i=1}^{n} \rho\left(Y_i - \beta_0 - \sum_{k=1}^{p} \sum_{l=1}^{q_k} \beta_{kl} B_{kl}(X_i^k)\right) + \kappa \sum_{k=1}^{p} \sum_{l=1}^{q_k} \beta_{kl}^2, \tag{2.4}$$

where $\rho(t) = (\tau - I(t \leq 0))t$ is the check loss function at a given quantile level $0 < \tau < 1$ and $\kappa \geq 0$. We denote the minimizer of (2.4) as $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}_1^\top, \ldots, \tilde{\boldsymbol{\beta}}_p^\top)^\top$, where $\tilde{\boldsymbol{\beta}}_k = (\tilde{\beta}_{k1}, \ldots, \tilde{\beta}_{kq_k})^\top$. Then, we estimate m_k by $\tilde{m}_k(\cdot) = \sum_{l=1}^{q_k} \tilde{\beta}_{kl} B_{kl}(\cdot)$ and $\|m_k\|_{L_2}$ by $\|\tilde{m}_k\|_{L_2}$. The estimator $\tilde{m}_k(\cdot)$ when $\kappa = 0$ was already considered by Horowitz and Lee (2005) but its statistical properties were not appropriately

derived there. Here, we define a more general estimator than theirs by adopting the concept of ridge penalty so that our estimator $\tilde{m}_k$ performs more stably and effectively. Further, we provide asymptotic properties of our estimator which includes the one of Horowitz and Lee (2005) as a special case in a rigorous way.

Actually, the penalty term $\kappa \sum_{k=1}^{p} \sum_{l=1}^{q_k} \beta_{kl}^2$ that appears in (2.4) is a special case of the penalties that Eilers and Marx (1996) proposed to prevent overfitting in a situation where a relatively large number of spline basis functions are used for nonparametric estimation. One attractive feature of the resulting 'P-spline' estimator is that it has no boundary effects. For a similar purpose, roughness penalties have been considered in a large literature on nonparametric function estimation with spline basis functions. For example, see Eubank (1988) and Wahba (1990). Intuitively speaking, our penalized estimator manages to get stable like ridge regression estimators in linear models by preventing the square of L_2 norm of each component function from growing too big because the term $\sum_{l=1}^{q_k} \beta_{kl}^2$ corresponds to the squared L_2 norm of the approximate function $\sum_{l=1}^{q_k} \beta_{kl} B_{kl}(\cdot)$ for $m_k(\cdot)$. In our simulation study, we observed that the estimator with the appropriately chosen $\kappa > 0$ gives more stable results than the one with $\kappa = 0$ especially when the covariates are correlated.

2.2 The final estimator $\hat{m}_k(\cdot)$ with component selection

When some covariates are irrelevant in (1.1), it means that the corresponding additive components are zero functions hence their L_2 norms are exactly zero. Considering that from the approximation (2.2) each function $m_k(x)$ in (1.1) is fully characterized by a set of parameters $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kq_k})^\top$, we should make the vector $\boldsymbol{\beta}_k$ zero when $\|\tilde{m}_k\|_{L_2}$ is considerably close to zero in order to select relevant covariates. For that purpose, we consider adding to (2.4) an extra penalty acting on $\|\boldsymbol{\beta}_k\|_2$ such as the LASSO (Tibshirani (1996)) or SCAD penalty (Fan and Li (2001)). Such penalties on $\|\boldsymbol{\beta}_k\|_2$ are called group LASSO (Yuan and Lin, 2006) penalty and group SCAD (Wang et al., 2008) penalty, respectively.

In this paper, we use a local linear approximation of the group SCAD penalty, which is called one-step group SCAD penalty. It is known that the one-step group SCAD method is better than the group SCAD method both in theoretical and computational aspects. For details, we refer to Noh and Park (2010). Let $p_\lambda(\cdot)$ be the SCAD penalty function. The function p_λ is defined on $\mathbb{R}^+$ by its derivative as

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

for some constant $a > 2$, where $x_+ = \max\{x, 0\}$. Following the recommendation of Fan and Li (2001), we use a=3.7 in our simulation. In this paper, we define the one-step group SCAD regularized estimator of $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_p^\top)^\top$ as the minimizer $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\boldsymbol{\beta}}_1^\top, \ldots, \hat{\boldsymbol{\beta}}_p^\top)^\top$ of

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \rho \left(Y_i - \beta_0 - \sum_{k=1}^{p} \sum_{l=1}^{q_k} \beta_{kl} B_{kl}(X_i^k) \right) + n \sum_{k=1}^{p} \nu_k \sqrt{q_k} \, \|\boldsymbol{\beta}_k\|_2 + \kappa \sum_{k=1}^{p} \sum_{l=1}^{q_k} \beta_{kl}^2, \tag{2.5}$$

where $\nu_k = p'_\lambda(\sqrt{q_k} \, \|\tilde{\boldsymbol{\beta}}_k\|_2)$ and $\kappa \geq 0$. Note that since the function p'_λ is decreasing and becomes exactly zero when the argument exceeds $a\lambda$, it enables us to adaptively penalize $\|\boldsymbol{\beta}_k\|_2$ depending on the magnitude of $\|\tilde{\boldsymbol{\beta}}_k\|_2 = \|\tilde{m}_k\|_{L_2}$. The multiplication by the constant $\sqrt{q_k}$ (q_k is the dimension of $\boldsymbol{\beta}_k$) in the penalty is for balancing the size of the group. In penalized estimation problems to select a group of variables simultaneously, several works including Yuan and Lin (2006) and Antoniadis and Gijbels (2012) use the same multiplier for a similar purpose. Finally, the ridge-type penalty used for stabilization in the first step (2.4) is used again for the same reason. Once we solve the minimization problem (2.5), we get

$$\hat{m}_k(\cdot) = \sum_{l=1}^{q_k} \hat{\beta}_{kl} B_{kl}(\cdot) \tag{2.6}$$

as an estimator of $m_k(\cdot)$ for $k = 1, \ldots, p$.

3 Numerical Implementation

3.1 Construction of basis functions

In numerical applications, B-spline basis is one of commonly used bases because of its good numerical properties. However, the use of the (ordinary) B-spline basis itself for approximating $m_k(\cdot)$ is not appropriate for our settings. It is because each B-spline basis $\{B_{kl}^0\}_{l=1}^{q_k^0}$ satisfies $\sum_{l=1}^{q_k^0} B_{kl}^0(x) = 1$ for every $x \in [0, 1]$ so the resulting design matrix $[\boldsymbol{\Pi}_1 \cdots \boldsymbol{\Pi}_n]^\top$ from B-spline approximation is singular (refer to the next subsection for the definition of $\boldsymbol{\Pi}_i$), and moreover B_{kl}^0 is not centered. Due to this reasons, we construct a new basis which is orthonormal and centered from the B-spline basis so that the problems explained above do not occur. To avoid the singularity coming from the relation $\sum_{l=1}^{q_k^0} B_{kl}^0(x) = 1$, of the B-spline basis we discard one B-spline function (for simplicity, $B_{kq_k^0}^0$). Then we make the remaining B-spline functions centered and orthonormal via the Gram-Schmidt process

to get the new basis. Precisely speaking, the new basis for m_k is inductively defined as

$$
\begin{aligned}
B_{k1}(\cdot) &= \frac{1}{\|B^0_{k1} - \int B^0_{k1}\|_{L_2}} \left(B^0_{k1}(\cdot) - \int B^0_{k1} \right), \\
B_{k2}(\cdot) &= \frac{1}{\|B^0_{k2} - \int B^0_{k2} - (\int B^0_{k2} B_{k1}) B_{k1}\|_{L_2}} \left(B^0_{k2}(\cdot) - \int B^0_{k2} - \left(\int B^0_{k2} B_{k1} \right) B_{k1}(\cdot) \right) \\
&\vdots \\
B_{kq_k}(\cdot) &= \frac{1}{\|B^0_{kq_k} - \int B^0_{kq_k} - \sum_{l'=1}^{q_k-1} (\int B^0_{kq_k} B_{kl'}) B_{kl'}\|_{L_2}} \left(B^0_{kq_k}(\cdot) - \int B^0_{kq_k} - \sum_{l'=1}^{q_k-1} \left(\int B^0_{kq_k} B_{kl'} \right) B_{kl'}(\cdot) \right),
\end{aligned}
$$

where $q_k = q^0_k - 1$. The obtained basis $\{B_{kl} : 1 \le l \le q_k\}$, $k = 1, \ldots, p$ satisfies the conditions on basis functions, which will be described by the assumption (A3) in Section 4. We use the basis $\{B_{kl} : 1 \le l \le q_k\}$ for $1 \le k \le p$ in our numerical study.

3.2 Computational algorithms

In order to calculate the final estimator $\hat{m}_k(\cdot)$, we should solve two optimization problems (2.4) and (2.5). Since (2.5) reduces to (2.4) when $\nu_1 = \cdots = \nu_p = 0$, we focus on the case (2.5). Let $\boldsymbol{\pi}^k(\cdot) = (B_{k1}(\cdot), \ldots, B_{kq_k}(\cdot))^\top$ be a set of basis functions for the estimation of $m_k(\cdot)$. Define $\boldsymbol{\Pi}(\mathbf{X}) = (1, \boldsymbol{\pi}^1(X^1)^\top, \ldots, \boldsymbol{\pi}^p(X^p)^\top)^\top$, $\boldsymbol{\pi}^k_i = \boldsymbol{\pi}^k(X^k_i)$ and $\boldsymbol{\Pi}_i = \boldsymbol{\Pi}(\mathbf{X}_i)$, $i = 1, \ldots, n$. With these notations, the optimization problem (2.5) is equivalently reformulated as:

$$
\begin{aligned}
\min_{\boldsymbol{\beta}, \mathbf{v}, \boldsymbol{\eta}^+, \boldsymbol{\eta}^-} \left(\tau \sum_{i=1}^n \eta_i^+ + (1-\tau) \sum_{i=1}^n \eta_i^- + n \sum_{k=1}^p \sqrt{q_k} \nu_k v_k \right) & \qquad (3.1) \\
\text{such that } \ \eta_i^+ - \eta_i^- = Y_i - \boldsymbol{\Pi}_i^\top \boldsymbol{\beta}, \ i = 1, \ldots, n & \\
\|\boldsymbol{\beta}_k\|_2 \le v_k, \ k = 1, \ldots, p & \\
\sum_{k=1}^p \|\boldsymbol{\beta}_k\|_2^2 \le \alpha(\kappa) & \\
\eta_i^+ \ge 0, \ \eta_i^- \ge 0, \ i = 1, \ldots, n &
\end{aligned}
$$

where $\alpha(\kappa)$ is a parameter corresponding one-to-one to κ. The reformulation (3.1) shows that the problem to minimize (2.5) is a second order cone programming (SOCP) problem in which a linear function is minimized over the intersection of an affine linear manifold with the Cartesian product of second-order cones. 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

determined using the convex optimization algorithms such as primal-dual interior point methods. Especially, when when $\nu_1 = \cdots = \nu_p = 0$, note that (3.1) is just a quadratic programming problem, which is easier to optimize. In our simulations, we use `CVX` to solve the SOCP problem (3.1).

3.3 Tuning parameters

For the initial estimator $\tilde{m}_k$, we should select the number q_k of basis for each component function and the stabilization parameter κ. To select the q_k and κ, we consider the K-fold cross-validation criterion. The original sample $\mathcal{S} = \{(\mathbf{X}_i, Y_i) : 1 \leq i \leq n\}$ is randomly partitioned into K groups of subsamples. Then, for each j, $\mathcal{S}$ with the jth partition removed is used for estimation while the jth partition is used for validation. The cross-validation criterion is given as

$$CV(\mathbf{q}, \kappa) = \sum_{j=1}^{K} \sum_{i \in I_j} \rho(Y_i - \mathbf{\Pi}_i^\top \tilde{\boldsymbol{\beta}}_{-j}(\mathbf{q}, \kappa)) \tag{3.2}$$

where for $1 \leq j \leq K$, I_j is the set of the indices of the jth partition of $\mathcal{S}$ and $\tilde{\boldsymbol{\beta}}_{-j}(\mathbf{q}, \kappa)$ is the estimate computed via (2.4) with $\mathbf{q} = (q_1, \ldots, q_p)^\top$ and κ from $\mathcal{S}$ with the jth partition deleted. In our simulations, to reduce the computation burden, we select $\mathbf{q}$ to minimize the criterion (3.2) given $\kappa = 0$ (with respect to $\mathbf{q}$) under the constraints $q_k = q$ for $k = 1, \ldots, p$ and then κ to minimize (3.2) given the chosen value of $\mathbf{q}$. This choice of $\mathbf{q}$ and κ gives reasonable numerical results for $\tilde{m}_k$ in the simulations.

For the final estimator $\hat{m}_k$, we need to choose q_k's and κ at (2.5). Additionally, we also should choose the penalty parameter λ. For conveinience of computation, we use the same q_k's and κ of the estimator $\tilde{m}_k$ and choose only λ for the final estimator. This is justified by the fact that the theoretical order of q_k is the same for $\hat{m}_k$ and $\tilde{m}_k$ and the κ's for $\hat{m}_k$ and $\tilde{m}_k$ have the same theoretical upper bound in asymptotic order. Once the q_k's and κ's are decided, we use a Schwarz-type Information Criterion (SIC) to select λ. If we denote the minimizer of (2.5) by $\hat{\boldsymbol{\beta}}(\lambda)$, the SIC for λ is given by

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

where $df(\lambda) = 1 + \sum_{k \in \mathcal{S}} q_k$ and $\mathcal{S} = \{1 \leq k \leq p : \|\hat{\boldsymbol{\beta}}_k(\lambda)\|_2 \neq 0\}$ denotes the index set of the selected component functions.

4 Asymptotic Properties

Let $m(\mathbf{x}) = \mu + \sum_{k=1}^{p} m_k(x^k)$ be the true τth conditional quantile function. For approximating m_k, the number of basis functions, q_k, tends to go to infinity as $n \to \infty$ so q_k depends on n although we suppress the dependence in the notation. We also suppress the dependence of κ and λ on n for simplicity of the notations. Suppose that $\limsup_{n\to\infty}(\max_k q_k / \min_k q_k) < \infty$. Then, without loss of generality we can assume $q_k = q$ for all k. We make the following assumptions to facilitate our asymptotic analysis.

(A1) The distribution of $\mathbf{X}$ is absolute continuous with density $f_{\mathbf{X}}$. Furthermore, the density $f_{\mathbf{X}}(\cdot)$ of $\mathbf{X}$ is bounded away from 0.

(A2) The conditional distribution $F_{U|\mathbf{X}}(\cdot|\mathbf{x})$ of U given $\mathbf{X} = \mathbf{x}$, has a bounded density $f_{U|\mathbf{X}} : 0 < b \leq f_{U|\mathbf{X}}(u|\mathbf{x}) \leq B < \infty$ uniformly in $\mathbf{x}$ and u for some positive constants b and B. The derivative $f'(u|\mathbf{x}) := \partial f(u|\mathbf{x})/\partial u$ is bounded, that is, $|f'(u|\mathbf{x})| \leq C$ for some $C > 0$.

(A3) The basis functions $B_{k1}(\cdot), \ldots, B_{kq_k}(\cdot),\ k = 1, \ldots, p$ satisfy the following conditions:

(a) Each $B_{kl}(x), k = 1, \ldots, p, l = 1, \ldots, q_k$ is continuous

(b) $\int_{\mathcal{X}_k} B_{kl}(x)dx = 0$

(c) $\int_{\mathcal{X}_k} B_{kl}(x)B_{kl'}(x)dx = \begin{cases} 1 & \text{if } l = l' \\ 0 & \text{otherwise} \end{cases}$

(d) $\sup_{\mathbf{x}\in[0,1]^p} \|\mathbf{\Pi}(\mathbf{x})\|_2 = O(q^{1/2})$.

(e) There is a vector $\boldsymbol{\beta}^*$ such that $\sup_{\mathbf{x}\in[0,1]^d} |m(\mathbf{x}) - \mathbf{\Pi}(\mathbf{x})^\top \boldsymbol{\beta}^*| = O(q^{-r})$ for some $r > 1/2$.

(A4) $q \approx n^{1/(2r+1)}$.

In (A4), the relation $a_n \approx b_n$ means that the ratio a_n/b_n is bounded away from zero and infinity. The condition (e) of (A3) requires that the true component function $m_k(\cdot)$ can be uniformly well-approximated by the basis $\{B_{kl}\}_{l=1}^{q_k}$. The validity of this condition depends on the smoothness of $m_k(\cdot)$'s as well as the basis functions. For example, when the dth order derivative of $m_k(\cdot)$ satisfies the Hölder condition of order γ, it is well-known that B-splines of order $d+1$ give the best approximation for m_k at accuracy $O(b_n^{-r})$, where b_n is the number of internal knots for the B-splines and $r = d + \gamma$ (Schumaker, 1981, Corollary 6.21). Thus, the condition (e) holds with $r = d+\gamma$ for the basis functions

that we construct from B-splines as described in Section 3, provided that the dth order derivatives of m_k for $1 \leq k \leq p$ are all Hölder continuous of order γ. Under a typical assumption that m_k for $1 \leq k \leq p$ have bounded second derivatives, then $r = 2$. Moreover, we assume the specific order of magnitude of q which balances estimation and approximation errors of the estimators in (A4). It is a typical assumption in nonparametric quantile regression (see He and Shi (1994), for example).

In the following theorem, we obtain the convergence rate of $\tilde{\boldsymbol{\beta}}$ which is the minimizer of $l_0(\boldsymbol{\beta})$ at (2.4) and also that of $\tilde{m}_k(\cdot)$ that we use as an initial estimator in this work.

Theorem 4.1 *Under the assumptions (A1)-(A4) and* $\kappa^2/(nq) \to 0$*, we have* $\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O_p(n^{-1/2}q^{1/2})$*. Thus, it follows that* $\|\tilde{m}_k - m_k\|_{L_2} = O_p(n^{-r/(2r+1)})$ *and* $\sup_x |\tilde{m}_k(x) - m_k(x)| = O_p(n^{-(2r-1)/(2(2r+1))})$ *for* $k = 1, \ldots, p$*.*

Theorem 4.1 states that the initial estimator $\tilde{m}_k$ has the rate of convergence $n^{-r/(2r+1)}$ in L_2-norm, which is the same as the optimal convergence rate in univariate nonparametric regression (Stone, 1982). This attractive rate of convergence is possible because of the additive structure and the similar results in kernel smoothing context for the same model can be found in the literature, for example, Lee et al. (2010) and Cheng et al. (2011). Actually, Horowitz and Lee (2005) studied the same estimator as ours except for the ridge penalty for the stabilization and seems to have derived the convergence of rate of their estimator in Theorem 1 there. However, their proof for it does not seem to be correct although the result itself is correct. To be more specific, they took an inappropriate function as an estimating equation for the additive quantile regression model and tried to derive the asymptotic results from the estimating equation. As a consequence of that, they were not successful to prove the main results correctly. For example, in their proof, the convexity of $M_{n\kappa}(\boldsymbol{\theta})$ (in their notation) and the subsequent claim from it play an important role to establish the results but it is clear from its definition that $M_{n\kappa}(\boldsymbol{\theta})$ is not convex because it is a discontinuous function involving indicator functions in it.

In general it is rather difficult to find easy-to-treat estimating equations from the check loss function in quantile regression because of its non-differentiability at zero. One of commonly used techniques to get around it in quantile regression is to obtain an uniform approximation of the check loss, which is easier to deal with (He and Shi, 1994, Lemma 3.2) and then to derive asymptotic properties of the estimator from the uniform approximation. Adapting the technique to our settings we provide a rigorous proof for the asymptotic behavior of the estimator $\tilde{m}_k$ which includes the one of Horowitz and Lee (2005) as a special case.

Now we present the main theorem which states asymptotic properties of the final estimator $\hat{m}_k$ at (2.6).

Theorem 4.2 *Suppose that the assumptions (A1)-(A4) hold and* $\kappa^2/(nq) \to 0$. *Further, we assume that* $\lambda \to 0$ *and* $n^{-1/2}q\,\lambda^{-1} \to 0$. *Then, we have*

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

(ii) $\|\hat{m}_k - m_k\|_{L_2} = O_p(n^{-r/(2r+1)})$ *and* $\sup_x |\hat{m}_k(x) - m_k(x)| = O_p(n^{-(2r-1)/(2(2r+1))})$ *for* $k = 1, \ldots, s$.

The first part of Theorem 4.2 implies the consistency of our doubly-penalized procedure in selecting the relevant component functions. The second part provides the rate of convergence for the relevant components, which is the optimal nonparametric rate as mentioned before. Therefore the theorem implies that the proposed doubly-penalized method is nonparametric oracle using the terminology of Storlie et al. (2011). For coherence of the paper, we develop the initial estimator $\tilde{m}_k$ based on basis approximation in the model (1.1) and used $\|\tilde{\boldsymbol{\beta}}_k\|_2 = \|\tilde{m}_k\|_{L_2}$ as an estimator of $\|m_k\|_{L_2}$ in the definition of ν_k at (2.5) for the final estimators $\hat{m}_k$, $1 \le k \le p$. However, the specific initial estimators $\|\tilde{\boldsymbol{\beta}}_k\|_2$ are not required for $\hat{m}_k$'s. In the definition of ν_k, $\|\tilde{\boldsymbol{\beta}}_k\|_2$ can be replaced by other consistent estimator for $\|m_k\|_{L_2}$ in the model (1.1). Denote a consistent estimator of $\|m_k\|_{L_2}$ by $\|\tilde{m}_k^0\|_{L_2}$ and consider the case where $\nu_k = p'_\lambda(\sqrt{q_k}\ \|\tilde{m}_k^0\|_{L_2})$. Suppose that there exists a sequence of $\{a_n\}$ with $a_n \to 0$ such that $\|\tilde{m}_k^0 - m_k\|_{L_2} = O_p(a_n)$ for $k = 1, \ldots, p$. Then, Theorem 4.2 still holds if the condition $n^{-1/2}q\,\lambda^{-1} \to 0$ is replace by $q^{1/2}a_n\,\lambda^{-1} \to 0$.

5 Examples

In this section we first illustrate how the introduction of ridge-type penalty can contribute to the stabilization of the estimators via a simulation study. Then, we show the finite sample performance of our final estimator and illustrate its use with an analysis of data about body fat percentage.

5.1 Simulated example

In the simulation, we consider the following additive model:

$$Y_i = \sum_{k=1}^{8} f_k(X_i^k) + \left\{ \sum_{k=1}^{8} \sigma_k(X_i^k) \right\} U_i,\ i = 1, \ldots, n$$

where U_i are i.i.d. $N(0,1)$ independent of $\mathbf{X}_i = (X_i^1, \ldots, X_i^8)^\top$, $f_1(x) = 5x$, $f_2(x) = 4\sin(2\pi x)/(2 - \sin(2\pi x))$, $f_3(x) = 6(0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin^2(2\pi x) + 0.4\cos^3(2\pi x) + 0.5\sin^3(2\pi x))$, $f_4(x) = \cdots = f_8(x) \equiv 0$, $\sigma_4(x) = 4.5(2x-1)^2$ and $\sigma_k(x) \equiv 0$ for all $k \neq 4$. Under our simulated model, the centered version of the kth additive component in the τth conditional quantile function is given as

$$m_{k,\tau}(x^k) = a_k + f_k(x^k) + \sigma_k(x^k)\Phi^{-1}(\tau),$$

where $\Phi^{-1}(\tau)$ is the τth quantile of the standard normal distribution and a_k is the constant to make the identifiability constraint $\int_{\mathcal{X}_k} m_{k,\tau}(x)dx = 0$ hold. Then, the true index sets of the relevant covariates for a given quantile level $\tau = 0.5$ and $\tau \neq 0.5$ are $\{1,2,3\}$ and $\{1,2,3,4\}$, respectively. The covariates $\mathbf{X}$ are generated to have a compound symmetry covariance structure as in Xue (2009): $X^k = (W^k + tU)/(1+t)$, $k = 1, \ldots, 8$, where $(W^1, \ldots, W^8)$ and U are i.i.d. from $U[0,1]$. The case when $t = 0$ corresponds to the case where the covariates $\mathbf{X}$ are independent, whereas the case when $t \neq 0$ assumes that they are dependent with $\mathrm{corr}(X^k, X^{k'}) = t^2/(1+t^2)$ for $k \neq k'$.

As a measure of performance, we calculate Monte Carlo estimates, based on 100 random samples, of the mean integrated squared errors

$$\begin{aligned} MISE &= E\int_{[0,1]^p} \Big\{\sum_{k=1}^p \bar{m}_k(x^k) - \sum_{k=1}^p m_k(x^k)\Big\}^2 f_{\mathbf{X}}(\mathbf{x})d\mathbf{x} \\ &\simeq \frac{1}{100}\sum_{r=1}^{100} E_{\mathbf{X}}\left\{\sum_{k=1}^p \bar{m}_k^{(r)}(X^k) - \sum_{k=1}^p m_k(X^k)\right\}^2 \end{aligned}$$

where $\bar{m}_k(\cdot)$, $1 \leq k \leq p$ denotes the generic estimate, $\bar{m}_k^{(r)}$ is the estimate obtained from the rth random sample and $E_{\mathbf{X}}$ is the expectation taken only for the test data $\mathbf{X}$. In the simulation, we use a test sample of size 300 to compute the value of the integrated squared error $E_{\mathbf{X}}\big\{\sum_{k=1}^p \bar{m}_k^{(r)}(X^k) - \sum_{k=1}^p m_k(X^k)\big\}^2$ for the rth sample. Let $\mathcal{S}$ and $\mathcal{S}_0$ denote the estimated and true index set of relevant covariates, respectively. Following the terms of Huang and Yang (2004), we say $\mathcal{S}$ is correct if $\mathcal{S} = \mathcal{S}_0$; $\mathcal{S}$ overfits if $\mathcal{S} \supset \mathcal{S}_0$ and $\mathcal{S} \neq \mathcal{S}_0$; and $\mathcal{S}$ underfits if $\mathcal{S} \not\supset \mathcal{S}_0$.

First, we focus on investigating the effect of the ridge penalty for estimation. We tried $t = 0, 0.5, 1$ to see the effect when the dependence of the covariates varies. For both the initial estimator $\tilde{m}_k(\cdot)$ and the oracle estimator, we compute the MISE's when κ is set to zero and the selected value by the cross-validation (CV) criterion as described in Section 3. Note that the initial estimator with $\kappa = 0$ corresponds to the estimator considered in Horowitz and Lee (2005). The oracle estimator is obtained from fitting (2.4) with only the relevant covariates, X_i^k for $k \in S_0$, instead of the whole covariates $\mathbf{X}_i$.

We choose the optimal q from the 30-fold CV criterion. The reason why a large number of folds are used is that it was observed that the CV with a rather small number of folds was unreliable due to its many local minima. As for the selection of κ, we use the 5-fold CV.

(Insert Table 1 and Figure 1 about here)

For Table 1, we see that both estimators clearly enjoy the advantage of shrinkage from the ridge penalty especially when the correlations between the covariates become large. Figure 1 depicts two versions of estimates of all the component function from one random sample when $\tau = 0.5$, $t = 1$ and $n = 300$. One is the initial estimator with $\kappa = 0$ and the other is the initial estimator using the ridge penalty. For each version of the estimator, the sample is chosen for which the value of the integrated square error $E_{\mathbf{X}}\{\sum_{k=1}^{p} \bar{m}_k(X^k) - \sum_{k=1}^{p} m_k(X^k)\}^2$ is the median of these values obtained from the 100 random samples. From Figure 1, we see that the ridge-penalty with the κ chosen by the CV prevents the estimator from being unstable without distorting the main feature of the curve. The use of the ridge penalty is highly recommended especially when the covariates are considered to be fairly correlated.

To assess the performance of the final estimator $\hat{m}_k(\cdot)$, $1 \le k \le p$, in terms of estimation accuracy and consistency in component selection, we compute the MISE's of the initial estimators, the penalized estimator and the oracle estimator and their results of component selection. We took $t = 0, 1$. Note that the case with $t = 0$ deals with independent covariates whereas the case with $t = 1$ has dependent covariates whose pairwise correlations are 0.5. Since it is already observed that an adaptive well-chosen κ is advantageous in estimation for both the initial and oracle estimators, both are computed with the κ chosen by the CV. For the final estimator, the initial estimator with such choice of κ is used to estimate $\|m_k\|_{L_2}$. To see whether the ridge penalty at (2.5) is also necessary for the final estimator $\hat{m}_k(\cdot)$, we compute two final estimators by solving the problems (2.5) when $\kappa = 0$ and κ is chosen by the CV.

(Insert Table 2 and 3 about here)

Table 2 and 3 summarize the results. The numbers under "C", "U" and "O", respectively, are how many times the selected index sets $\mathcal{S}$ are correct, overfit and underfit over 100 Monte Carlo replications. The results in Table 2 and 3 implies that our final estimator $\hat{m}_k(\cdot)$ for $1 \le k \le p$ enjoys oracle properties which mean that it asymptotically behaves as if all the relevance information about the covariates is given in advance. Additionally, the comparison of the MISE's of two versions of the final estimator $\hat{m}_k$, $1 \le k \le p$ suggests that the ridge penalty is also needed for stabilization of $\hat{m}_k$'s,

because the other penalty that is the second term of the right-hand side at (2.5) serves mainly for component selection.

5.2 Application to body fat data

To demonstrate the effectiveness of the proposed covariate selection method in additive quantile regression models, we present a result from the analysis of body fat data. The data consists of percentage of body fat, age, weight, height and various body circumference measurements from 252 men. The body fat percentage is regarded as a simple and effective measure of an individual's fitness level so many health websites have a so-called body fat calculator which yields an estimate of it from body circumference measurements that visitors type in. In our work, based on four selected covariates from the data, which are frequented used to predict body fat percentage in many calculators, we try to see whether we might observe different pattern of relevance of the covariates across the quantile level.

We estimate the conditional quantile functions when $\tau = 0.5$ and $\tau = 0.8$ with four covariates: age, height, abdomen circumference and hip circumference. Before the analysis, the covariates are standardized to have mean 0 and variance 1. As we develop our theory based on the assumption that the support of each covariate is compact, we excluded 15 rather extreme observations which has at least one covariate, the absolute value of which exceeds 3 after the standardization. For more accurate analysis, unlike the simulated examples, we allowed each component function to have different spline order and number of internal knots from others.

(Insert Figure 2 about here)

Figure 2 summarized the estimation and covariate selection results when $\tau = 0.5$ and $\tau = 0.8$. Each panel of the figure shows two versions of the estimated component function. The dotted curve is from the initial estimator and the solid one from the final estimator. When $\tau = 0.5$, we see that our covariate selection method enhances interpretability of the estimation by thresholding the component functions of age and hip circumference which exhibit strange behavior against common knowledge in health sciences hence are noncontributory. Additionally, the strictly increasing trend of the component function of abdomen circumference, which are selected by our method, is in good accord with the well-known fact that men are prone to accumulate fat around their abdomens.

However, when we consider estimation of a higher conditional quantile, the result becomes a bit different from in conditional median quantile estimation. As we see in Figure 2, our procedure additionally select hip circumference as well as the covariates selected when $\tau = 0.5$. This suggests

that as for men with high body fat percentage, fat may tend to be accumulated around their hips as well as around their abdomens. This finding is new and interesting because it is a well-known knowledge in health care that hip circumference is only related to body fat percentage of women but it is not in the case of men. Further, our finding is partially supported by a recent study about body fat by Bergman et al. (2011) who considered hip circumference as an important factor related to body fat percentage both for men and women.

6 Conclusion

This article has developed a doubly-penalized estimation for nonparametric additive quantile regression models that can simultaneously perform component selection and estimation of smooth component functions. We have established the theoretical properties of our procedure, including consistency in component selection and the nonparametric optimal rate of convergence in estimation. Different from the works concerned with component selection in the literature, we introduced an extra penalty that controls the L_2-norms of component functions for the stability of the estimation using the idea of P-spline, independently from the penalty for component selection. Finally, we have shown that in the numerical study our doubly-penalized estimator performs more stably when the covariates are fairly correlated than a singly-penalized one.

Appendix

For the vector $\boldsymbol{\beta}^*$ in the condition (e) of (A3), we let $\boldsymbol{\beta}^* = (\beta_0^*, \boldsymbol{\beta}_1^{*\top}, \ldots, \boldsymbol{\beta}_p^{*\top})^\top$. Define $H = E[\boldsymbol{\Pi}(\mathbf{X})\boldsymbol{\Pi}(\mathbf{X})^\top]$ and $H_n = \sum_{i=1}^n \boldsymbol{\Pi}_i \boldsymbol{\Pi}_i^\top$. First, we present two lemmas that are necessary to prove the theorems.

Lemma 6.1 *Under (A1) and $q^2/n \to 0$, the eigenvalues of $n^{-1}H_n$ are uniformly bounded away from zero and infinity in probability.*

Proof. Let $\ell_{\min}(S)$ and $\ell_{\max}(S)$ denote the smallest and largest, respectively, eigenvalues of a real-valued square matrix S. We denote the Frobenius norm of a matrix by $\|\cdot\|_F$. From (A1) and orthonormality of basis functions, we can take $c > 0$ such that $0 < c^{-1} \le \ell_{\min}(H) \le \ell_{\max}(H) \le c < \infty$. By Corollary 8.1.6 of Golub and Van Loan (1996) and the fact that $\|n^{-1}H_n - H\|_F^2 = O_p(q^2/n) = o_p(1)$,

we obtain that

$$0 < (2c)^{-1} \leq \ell_{\min}\big(n^{-1}H_n\big) \leq \ell_{\max}(n^{-1}H_n) \leq 2c < \infty$$

on a set with probability tending to one. □

Lemma 6.2 *Suppose that (A1)-(A4) hold. Then, for any sequence* $\{L_n\}$ *satisfying* $1 \leq L_n \leq q^{\delta_0/10}$ *for some* $0 < \delta_0 < (r-1/2)/(2r+1)$

$$\sup_{\boldsymbol{\theta}^\top H_n \boldsymbol{\theta} \leq L_n^2 q} \Big| \sum_{i=1}^n \rho(U_i - \boldsymbol{\Pi}_i^\top \boldsymbol{\theta} - R_{ni}) - \rho(U_i - R_{ni}) + \boldsymbol{\Pi}_i^\top \boldsymbol{\theta}(\tau - I(U_i < 0)) \\ - E_{U_i|\mathbf{X}_i}(\rho(U_i - \boldsymbol{\Pi}_i^\top \boldsymbol{\theta} - R_{ni}) - \rho(U_i - R_{ni})) \Big| = o_p(q),$$

where $R_{ni} = \boldsymbol{\Pi}_i^\top \boldsymbol{\beta}^* - m(\mathbf{X}_i)$.

Using the similar arguments as described to prove Lemma 3.2 of He and Shi (1994), Lemma 6.2 can be proven. The detail of the proof is available on request.

Proof of Theorem 4.1

Let $a_n = n^{-1/2}q^{1/2}$. By convexity of l_0 at (2.4), it is enough to show that for any given $\epsilon > 0$, there exists a large constant $L > 0$ such that

$$P\left(\inf_{\|\boldsymbol{\theta}\|_2 = L a_n} l_0(\boldsymbol{\beta}^* + \boldsymbol{\theta}) > l_0(\boldsymbol{\beta}^*) \right) > 1 - \epsilon. \tag{6.1}$$

Let $\delta = La_n$. The condition (e) of (A3) and orthonormality of basis functions give $\|\boldsymbol{\beta}^*\|_2 = O(1)$, so it leads to

$$\sup_{\|\boldsymbol{\theta}\|_2 = \delta} |\|\boldsymbol{\beta}^* + \boldsymbol{\theta}\|^2 - \|\boldsymbol{\beta}^*\|^2| = o(a_n). \tag{6.2}$$

Then, we have

$$\begin{aligned} D_n(\boldsymbol{\theta}) &\equiv l_0(\boldsymbol{\beta}^* + \boldsymbol{\theta}) - l_0(\boldsymbol{\beta}^*) \\ &\geq \sum_{i=1}^n \Big[-\boldsymbol{\Pi}_i^\top \boldsymbol{\theta}(\tau - I(U_i < 0)) + E_{U_i|\mathbf{X}_i}(\rho(U_i - \boldsymbol{\Pi}_i^\top \boldsymbol{\theta} - R_{ni}) - \rho(U_i - R_{ni})) \Big] + o_p(q) \\ &\geq A_n(\boldsymbol{\theta}) + B_n(\boldsymbol{\theta}) + C_n(\boldsymbol{\theta}) + o_p(q), \end{aligned} \tag{6.3}$$

where $A_n(\boldsymbol{\theta}) = -\sum_{i=1}^n \boldsymbol{\Pi}_i^\top \boldsymbol{\theta}(\tau - I(U_i < 0))$, $B_n(\boldsymbol{\theta}) = -\sum_{i=1}^n E_{U_i|\mathbf{X}_i}[\boldsymbol{\Pi}_i^\top \boldsymbol{\theta}(\tau - I(U_i - R_{ni} \leq 0))]$ and $C_n(\boldsymbol{\theta}) = \sum_{i=1}^n E_{U_i|\mathbf{X}_i}[\int_0^{\boldsymbol{\Pi}_i^\top \boldsymbol{\theta}} (I(U_i - R_{ni} \leq s) - I(U_i - R_{ni} \leq 0))\, ds]$. The first inequality follows from Lemma 6.2, (6.2) and $\kappa^2/(nq) \to 0$. For the last inequality, we use the Knight's identity:

$$\rho(u - v) - \rho(u) = -v(\tau - I(u \leq 0)) + \int_0^v (I(u \leq s) - I(u \leq 0)) ds.$$

By simple calculations, one has that

$$|A_n(\boldsymbol{\theta})| \le \delta \, \| \sum_{i=1}^{n} \boldsymbol{\Pi}_i(\tau - I(U_i < 0))\|_2 \text{ and } \| \sum_{i=1}^{n} \boldsymbol{\Pi}_i(\tau - I(U_i < 0))\|_2 = O_p((nq)^{1/2}). \tag{6.4}$$

Note that $|F_{U|\mathbf{X}}(0|\mathbf{x}) - F_{U|\mathbf{X}}(R_{ni}|\mathbf{x})| \le B|R_{ni}|$ for all $\mathbf{x}$, where B is the constant in the assumption (A2). Consequently, we can take a constant $M_1 > 0$ such that $\sup_{\|\boldsymbol{\theta}\|_2=\delta} |B_n(\boldsymbol{\theta})| \le M_1 q^{-r} n\delta$ by the condition (e) of (A3) and Lemma 6.1. From Taylor's theorem, we have

$$\begin{aligned} C_n(\boldsymbol{\theta}) &\ge [\frac{b}{2}\sum_{i=1}^{n}(\boldsymbol{\Pi}_i^\top \boldsymbol{\theta})^2 - \frac{C}{6}\sum_{i=1}^{n}(\boldsymbol{\Pi}_i^\top \boldsymbol{\theta})^3] \\ &\ge \sum_{i=1}^{n}(\boldsymbol{\Pi}_i^\top \boldsymbol{\theta})^2 (b/2 + o_p(1)) \\ &\ge (n\delta^2) M_2 \end{aligned}$$

for some constant $M_2 > 0$ (that is independent of $\boldsymbol{\theta}$) and sufficiently large n, where b and C are the constants in the assumption (A2). The last inequality follows from Lemma 6.1. Thus, for sufficiently large L, $C_n(\boldsymbol{\theta})$ dominates all the other terms in (6.3) uniformly in $\|\boldsymbol{\theta}\|_2 = \delta$, which implies (6.1). The remaining parts of the theorem are immediate since $\|\tilde{m}_k - \boldsymbol{\beta}_k^{*\top} \boldsymbol{\pi}^k\|_{L_2} = \|\tilde{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^*\|_2$ and $\sup_{x\in[0,1]} \|\boldsymbol{\pi}^k(x)\|_2 = O(q^{1/2})$. □

Proof of Theorem 4.2

First, we prove (ii). Since $\sqrt{q}\, \|\tilde{\beta}_k\| > a\lambda$ for $1 \le k \le s$ with probability tending to one, one has that

$$\sup_{\boldsymbol{\beta}} \left| \sum_{k=1}^{s} \nu_k \sqrt{q} (\|\boldsymbol{\beta}_k\|_2 - \|\boldsymbol{\beta}_k^*\|_2) \right| = o_p(k_n) \tag{6.5}$$

for any sequence $\{k_n\}$ such that $k_n \to 0$ as $n \to \infty$. Note that $\sum_{k=s+1}^{p} \nu_k \sqrt{q}\, (\|\boldsymbol{\beta}_k\|_2 - \|\boldsymbol{\beta}_k^*\|_2) \ge 0$ for any $\boldsymbol{\beta}$ because $\boldsymbol{\beta}_k^* = \mathbf{0}$ for $k = s+1, \ldots, p$. From Lemma 6.2, (6.2), (6.5) and $\kappa^2/(nq) \to 0$, we have that for any $\boldsymbol{\theta}$ with $\|\boldsymbol{\theta}\|_2 = Ln^{-1/2}q^{1/2}$,

$$\begin{aligned} D_n'(\boldsymbol{\theta}) &\equiv l(\boldsymbol{\beta}^* + \boldsymbol{\theta}) - l(\boldsymbol{\beta}^*) \\ &\ge \sum_{i=1}^{n} \Big[-\boldsymbol{\Pi}_i^\top \boldsymbol{\theta}(\tau - I(U_i < 0)) + E_{U_i|\mathbf{X}_i}(\rho(U_i - \boldsymbol{\Pi}_i^\top \boldsymbol{\theta} - R_{ni}) - \rho(U_i - R_{ni})) \Big] + o_p(q). \end{aligned}$$

Following the lines of the proof of Theorem 4.1 with $D_n'(\boldsymbol{\theta})$, we get $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O_p(n^{-1/2}q^{1/2})$, which implies (ii).

Next, we move on to prove (i) which implies the consistency in component selection. Different from existing proofs for consistency in variable selection in framework of mean regression, the gradient

function (with respect to $\boldsymbol{\beta}$) of $\sum_{i=1}^n \rho(Y_i - \boldsymbol{\Pi}_i^\top \boldsymbol{\beta})$ is not applicable here. This is because the check loss function ρ is not differentiable at zero. Instead, we derive (i) directly from a certain lower bound of the difference of two check loss functions. Since the selection consistency regarding the relevant components is clear from Theorem 4.2 (ii), we focus on the case of the irrelevant components.

Without loss of generality, we may assume that $\nu_k = \lambda$ for $s+1 \le k \le p$ because $\sqrt{q}\|\tilde{\boldsymbol{\beta}}_k\|_2 = O_p(n^{-1/2}q)$ by Theorem 4.1 and $n^{-1/2}q\lambda^{-1} \to 0$. Suppose that there exists an index $k_0 \in \{s+1, \ldots, p\}$ such that $\hat{m}_{k_0} \not\equiv 0$, that is, $\hat{\boldsymbol{\beta}}_{k_0} \not\equiv 0$. Let $\hat{\boldsymbol{\beta}}^*$ be the vector obtained from $\hat{\boldsymbol{\beta}}$ with $\hat{\boldsymbol{\beta}}_{k_0}$ being replaced by 0. Since $\rho(u) - \rho(v) \ge (\tau - I(v \le 0))(u - v)$ for any $u, v \in \mathbb{R}$, we have

$$
\begin{aligned}
& l(\hat{\boldsymbol{\beta}}) - l(\hat{\boldsymbol{\beta}}^*) \\
&\ge -\sum_{i=1}^n (\tau - I(Y_i \le \boldsymbol{\Pi}_i^\top \hat{\boldsymbol{\beta}}^*))\boldsymbol{\Pi}_i^\top (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^*) + n\lambda\sqrt{q}\|\hat{\boldsymbol{\beta}}_{k_0}\|_2 \\
&= -\sum_{i=1}^n (\tau - I(U_i \le 0))\boldsymbol{\Pi}_i^\top (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^*) - \sum_{i=1}^n (I(U_i \le 0) - I(U_i \le r_{ni}))\boldsymbol{\Pi}_i^\top (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^*) + n\lambda\sqrt{q}\|\hat{\boldsymbol{\beta}}_{k_0}\|_2 \\
&\ge \left(- \left\| \sum_{i=1}^n (\tau - I(U_i \le 0))\boldsymbol{\Pi}_i \right\|_2 - \left\| \sum_{i=1}^n (I(U_i \le 0) - I(U_i \le r_{ni}))\boldsymbol{\Pi}_i \right\|_2 + n\lambda\sqrt{q} \right) \|\hat{\boldsymbol{\beta}}_{k_0}\|_2, \qquad (6.6)
\end{aligned}
$$

where $r_{ni} = R_{ni} + \boldsymbol{\Pi}_i^\top (\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}^*)$. From the assumptions (A2) and (A3), we obtain that for any $L > 0$,

$$
\begin{aligned}
& E \sum_{k=1}^p \sum_{l=1}^q \left\{ \sum_{i=1}^n (I(U_i \le Ln^{-1/2}q) - I(U_i \le -Ln^{-1/2}q))|B_{kl}(X_i^k)| \right\}^2 \\
= \; & \sum_{k=1}^p \sum_{l=1}^q nE \left\{ (I(U \le Ln^{-1/2}q) - I(U \le -Ln^{-1/2}q))|B_{kl}(X^k)| \right\}^2 \\
& + \sum_{k=1}^p \sum_{l=1}^q n(n-1) \left\{ E(I(U \le Ln^{-1/2}q) - I(U \le -Ln^{-1/2}q))|B_{kl}(X^k)| \right\}^2 \\
\le \; & \left\{ n(2Ln^{-1/2}qB) + n^2(2Ln^{-1/2}qB)^2 \right\} \sup_{\mathbf{x} \in [0,1]^p} \|\boldsymbol{\Pi}(\mathbf{x})\|_2 = O(nq^3).
\end{aligned}
$$

This implies that

$$
\| \sum_{i=1}^n (I(U_i \le 0) - I(U_i \le r_{ni}))\boldsymbol{\Pi}_i \|_2 = O_p(n^{1/2}q^{3/2}) \qquad (6.7)
$$

because $\max_{1 \le i \le n} |r_{ni}| \le O(q^{-r}) + q^{1/2}\|\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}^*\| = O_p(n^{-1/2}q)$. By the second part of (6.4), (6.6),(6.7) and $n^{-1/2}q\lambda^{-1} \to 0$, we have

$$
l(\hat{\boldsymbol{\beta}}) - l(\hat{\boldsymbol{\beta}}^*) \ge \frac{n\lambda\sqrt{q}}{2}\|\hat{\boldsymbol{\beta}}_{k_0}\|_2 > 0
$$

with probability tending to one, which contradicts to the fact that $\hat{\boldsymbol{\beta}}$ is the minimizer of (2.5). This completes the proof of the first part of the theorem. □

References

A. Antoniadis and I. Gijbels. Penalized estimation in additive varying coefficient models using grouped regularization. Technical report, 2012.

R. N. Bergman, D. Stefanovski, T. A. Buchanan, A. E. Sumner, J. C. Reynolds, N. G. Sebring, A. H. Xiang, and R. M. Watanabe. A better index of body adiposity. *Obesity*, 19:1083–1089, 2011.

P. Bühlmann and B. Yu. Boosting with the l_2 loss:regression and classification. *Journal of the American Statistical Association*, 98:324–339, 2003.

Y. Cheng, J. G. De Gooijer, and D. Zerom. Efficient Estimation of an Additive Quantile Regression Model. *Scandinavian Journal of Statistics*, 38:46–62, 2011.

J. G. De Gooijer and D. Zerom. On Additive Conditional Quantiles With High-Dimensional Covariates. *Journal of the American Statistical Association*, 98:135–146, 2003.

P. H. C. Eilers and B. D. Marx. Flexible smoothing with b -splines and penalties. *Statistical Science*, 11:89–121, 1996.

R. L. Eubank. *Spline Smoothing and Nonparametric Regression.* New York: Marcel Dekker, Inc., 1988.

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.

J. Fan, Y. Feng, and R. Song. Nonparametric Independence Screening in Sparse Ultra-High-Dimensional Additive Models. *Journal of the American Statistical Association*, 106:544–557, 2011.

N. Fenske, T. Kneib, and T. Hothorn. Identifying Risk Factors for Severe Childhood Malnutrition by Boosting Additive Quantile Regression. *Journal of the American Statistical Association*, 106: 494–510, 2011.

G. H. Golub and C. F. Van Loan. *Matrix Computations.* Johns Hopkins University Press, 1996.

X. He and P. Shi. Convergence rate of B-spline estimators of nonparametric conditional quantile functions. *Journal of Nonparametric Statistics*, 3:299–208, 1994.

J. L. Horowitz and S. Lee. Nonparametric estimation of an additive quantile regression model. *Journal of the American Statistical Association*, 100:1238–1249, 2005.

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

Y. K. Lee, E. Mammen, and B. U. Park. Backfitting and smooth backfitting for additive quantile models. *Annals of Statistics*, 38:2857–2883, 2010.

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

L. L. Schumaker. *Spline Functions: Basis Theory*. Wiley, New York., 1981.

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

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.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, B58:267–288, 1996.

G. Wahba. *Spline Models for Observational Data*. SIAM, Philadelphia, 1990.

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.

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

K. Yu and Z. Lu. Local Linear Additive Quantile Regression. *Scandinavian Journal of Statistics*, 31: 333–346, 2004.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society*, B68:49–67, 2006.

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

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

Table 1: MISE(×100)'s for the respective choice of $\kappa = 0$ and κ_{CV} selected by CV

n	Method	$t = 0$		$t = 0.5$		$t = 1$	
		$\tau = 0.5$	$\tau = 0.75$	$\tau = 0.5$	$\tau = 0.75$	$\tau = 0.5$	$\tau = 0.75$
300	initial estimator with κ_{CV}	69.400	80.853	47.833	60.461	50.384	65.759
	initial estimator with $\kappa = 0$	72.257	82.158	86.906	85.946	108.438	122.329
	oracle estimator with κ_{CV}	34.202	59.054	23.737	47.361	26.729	48.091
	oracle estimator with $\kappa = 0$	33.997	59.753	32.315	78.912	57.179	120.053
500	initial estimator with κ_{CV}	45.958	54.590	32.675	42.963	34.931	47.481
	initial estimator with $\kappa = 0$	47.077	54.570	54.433	56.802	67.181	74.470
	oracle estimator with κ_{CV}	19.498	40.484	16.177	36.446	17.949	42.525
	oracle estimator with $\kappa = 0$	19.494	40.733	18.926	52.611	26.985	63.954

Table 2: MISE(×100)'s and component selection results when $t = 0$.

n	Method	$\tau = 0.5$				$\tau = 0.75$			
		C	U	O	MISE	C	U	O	MISE
300	initial estimator with κ_{CV}	0	0	100	69.400	0	0	100	80.853
	final estimator with κ_{CV}	100	0	0	33.436	87	13	0	63.140
	final estimator with $\kappa = 0$	100	0	0	32.566	88	12	0	63.112
	oracle estimator with κ_{CV}	100	0	0	34.202	100	0	0	59.053
500	initial estimator with κ_{CV}	0	0	100	45.958	0	0	100	54.590
	final estimator with κ_{CV}	99	0	1	18.066	100	0	0	40.052
	final estimator with $\kappa = 0$	99	0	1	17.833	100	0	0	40.404
	oracle estimator with κ_{CV}	100	0	0	19.498	100	0	0	40.484

Table 3: MISE(×100)'s and component selection results when $t = 1$.

n	Method	$\tau = 0.5$				$\tau = 0.75$			
		C	U	O	MISE	C	U	O	MISE
300	initial estimator with κ_{CV}	0	0	100	50.364	0	0	100	65.765
	final estimator with κ_{CV}	87	0	13	36.588	74	20	6	58.287
	final estimator with $\kappa = 0$	80	0	20	59.198	77	14	9	88.802
	oracle estimator with κ_{CV}	100	0	0	26.800	100	0	0	48.268
500	initial estimator with κ_{CV}	0	0	100	34.930	0	0	100	47.528
	final estimator with κ_{CV}	89	0	11	21.836	95	1	4	40.304
	final estimator with $\kappa = 0$	81	0	19	38.424	95	1	4	62.733
	oracle estimator with κ_{CV}	100	0	0	17.948	100	0	0	42.623

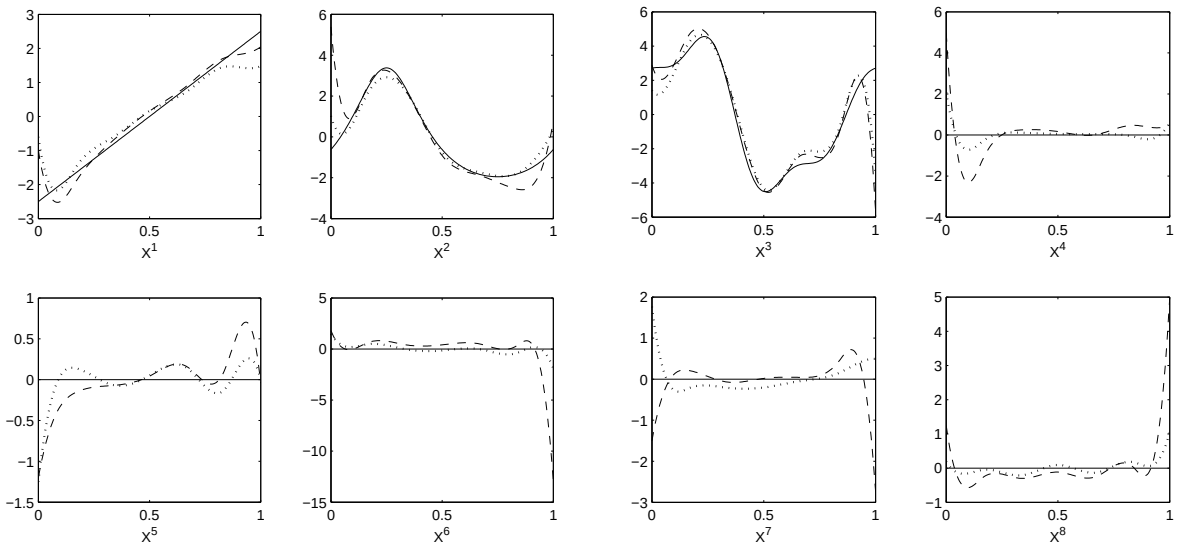

Figure 1: The estimated component functions when $\tau = 0.5$, $t = 1$ and $n = 300$. In each panel, the dashed line represents the estimate with $\kappa = 0$ and the dotted line represents the one with κ chosen by CV.

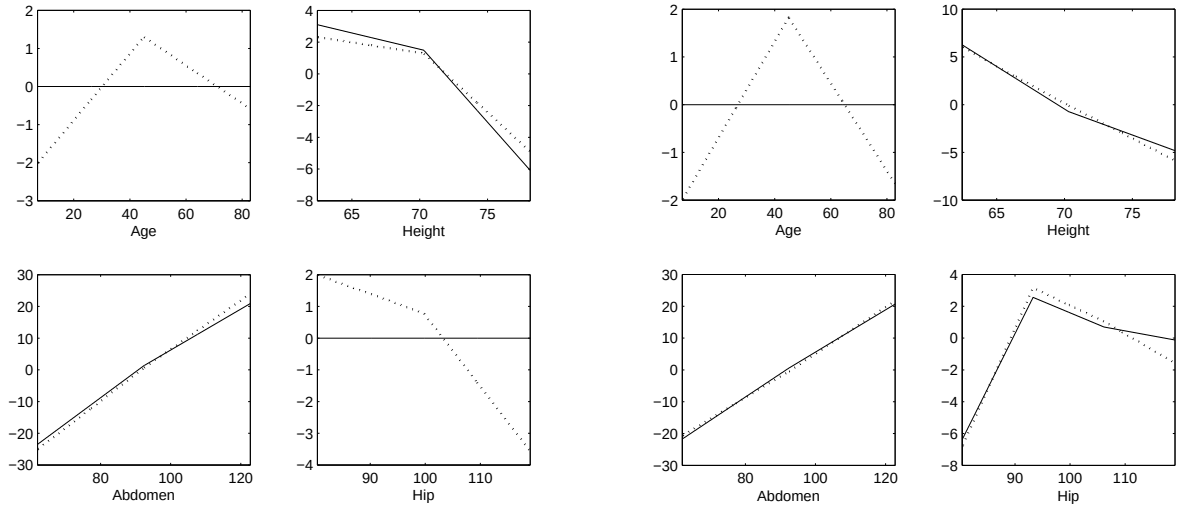

Figure 2: The estimated component functions of age, height, abdomen circumference and hip circumference when $\tau = 0.5$ (left) and $\tau = 0.8$ (right)