Statistica Sinica **22** (2012), 601-628
doi:http://dx.doi.org/10.5705/ss.2010.121

# A UNIFIED VARIABLE SELECTION APPROACH FOR VARYING COEFFICIENT MODELS

Yanlin Tang[1], Huixia Judy Wang[2], Zhongyi Zhu[1] and Xinyuan Song[3]

[1]*Fudan University,* [2]*North Carolina State University and* [3]*Chinese University of Hong Kong*

*Abstract:* In varying coefficient models, three types of variable selection problems are of practical interests: separation of varying and constant effects, selection of variables with nonzero varying effects, and selection of variables with nonzero constant effects. Existing variable selection methods in the literature often focus on only one of the three types. In this paper, we develop a unified variable selection approach for both least squares regression and quantile regression models with possibly varying coefficients. The developed method is carried out by using a two-step iterative procedure based on basis expansion and a double adaptive-LASSO-type penalty. Under some regularity conditions, we show that the proposed procedure is consistent in both variable selection and the separation of varying and constant coefficients. In addition, the estimated varying coefficients possess the optimal convergence rate under the same smoothness assumption, and the estimated constant coefficients have the same asymptotic distribution as their counterparts obtained when the true model is known. Finally, we investigate the finite sample performance of the proposed method through a simulation study and the analysis of the Childhood Malnutrition Data in India.

*Key words and phrases:* Adaptive LASSO, B-spline, least squares regression, quantile regression, separation of varying and constant effects.

## 1. Introduction

Suppose $(Y_i, \boldsymbol{X}_i, U_i)$, $i = 1, \ldots, n$, is an independent and identically distributed (*i.i.d.*) sample. We consider the following varying coefficient (VC) model

$$Y_i = \boldsymbol{X}_i^T \boldsymbol{\alpha}(U_i) + e_i, \tag{1.1}$$

where $\boldsymbol{\alpha}(u) = \{\alpha_0(u), \ldots, \alpha_p(u)\}^T$ is a $(p+1)$-vector of unknown smooth functions, $\boldsymbol{X}_i$ is the $(p+1)$-dimensional design vector with the first element $X_i^{(0)} \equiv 1$, $U_i \in \mathbb{R}^1$ is the univariate index variable such as the measurement time, and $e_i$ is the random error.

The varying coefficient model (1.1), first proposed by Hastie and Tibshirani (1993), provides more flexibility than the parametric linear models by allowing

the regression coefficients to depend on some covariate $U$. In the last decade, many estimation and hypothesis testing methods have been developed for VC models; see for example Wu and Chiang (2000), Huang, Wu, and Zhou (2002) and Kim (2007). Similar to parametric models, variable selection in VC models is equally important and even more complex, since ignoring important predictors can lead to biased results, while including irrelevant predictors or modeling constant coefficients as nonparametric can over-fit the data and lead to efficiency loss. In this paper, we aim to develop a unified variable selection method for VC models.

To summarize, three types of variable selection problems are of practical interests in VC models: (i) separation of varying and constant coefficients (Huang, Wu, and Zhou (2002)); Wang, Zhu, and Zhou (2009)); (ii) selection of variables with nonzero varying coefficients (Cai, Fan, and Li (2000), Qu and Li (2006)); (iii) selection of variables with nonzero constant coefficients (Fan and Huang (2005), Wang, Zhu, and Zhou (2009)). The existing variable selection methods for VC models often focus on only one of the above problems, and are confined to least squares regression. For selecting between constant and varying coefficients, Xia, Zhang, and Tong (2004) proposed a stepwise cross-validation-based procedure; Leng (2009) developed a penalization method in the framework of smoothing spline ANOVA models; Hu and Xia (2010) proposed a penalized procedure via the LASSO (least absolute shrinkage and selection operator, Tibshirani (1996)) penalty, where the varying coefficients were approximated by kernel smoothing and the penalty was applied to the $L_2$ norm of $\{\alpha_k(U_2) - \alpha_k(U_1), \ldots, \alpha_k(U_n) - \alpha_k(U_{n-1})\}$, $k = 1, \ldots, p$. For selecting variables with nonzero varying coefficients, Wang, Li, and Huang (2008) and Wang and Xia (2009) developed penalization methods via the SCAD (smoothly clipped absolute deviation, Fan and Li (2001)) penalty and the LASSO penalty, respectively. Assuming a partially linear varying coefficient model where the varying and constant coefficients are separated *a priori*, Zhao and Xue (2009) proposed methods for selecting variables in the parametric components and in the nonparametric components separately.

In this paper, we propose a unified approach that solves all the three types of variable selection problems for varying coefficient models, in both least squares and quantile regressions. To our best knowledge, this is a first attempt to do so. We estimate the varying coefficients in the VC models using the method of basis expansion because of its computational simplicity and stability; see illustrations in Huang, Wu, and Zhou (2002) and Wang, Zhu, and Zhou (2009). To conduct variable selection, we adopt the adaptive LASSO penalty (Zou (2006)) for both least squares and quantile regressions. The variable selection method is carried

out by using a two-step iterative procedure. We show that, under some regularity conditions, the penalized estimators are consistent in both variable selection and the separation of varying and constant coefficients. In addition, the resulting varying coefficient estimates possess the optimal convergence rate under the same smoothness assumption, and the constant coefficient estimates have the same asymptotic distribution as their counterparts obtained when the true model is known.

The rest of the paper is organized as follows. In Section 2, we describe the proposed variable selection method, and give the computational algorithms for both least squares regression and quantile regression. In Section 3, we present the theoretical results, including consistency in variable selection, the convergence rate of the nonzero varying coefficient estimates, and the asymptotic normality of the nonzero constant coefficient estimates. We assess the finite sample performance of the proposed method through an extensive simulation study in Section 4, and in the analysis of the Childhood Malnutrition Data in India in Section 5. All proofs are deferred to the Appendix.

## 2. The Proposed Variable Selection Method

### 2.1. The penalized estimation via adaptive LASSO

Throughout the paper we use superscript $T$ to denote matrix transpose. Without loss of generality, we assume that the index variable $U \in [0,1]$.

Let $k_n$ be the number of uniform internal knots, and $\hbar$ be the degree of the polynomial, $\hbar = 1$ corresponds to linear splines, $\hbar = 2$ corresponds to quadratic splines and so on. Let $\tilde{k}_n = k_n + 1$, $\hbar' = \hbar - 1$, and $I_{nj} = [(j-1)/\tilde{k}_n, j/\tilde{k}_n)$ for $1 \le j < \tilde{k}_n$, $I_{n\tilde{k}_n} = [(\tilde{k}_n - 1)/\tilde{k}_n, 1]$. Let $\mathcal{F}_n$ denote the collection of functions $f$ on $[0,1]$ such that (i) the restriction of $f$ to $I_{nj}$ is a polynomial of degree $\hbar$ (or less) for $1 \le j \le \tilde{k}_n$; (ii) $f$ is $\hbar'$-times continuous differentiable on $[0,1]$. Let $\tilde{\boldsymbol{\pi}}(\cdot) = (B_1(\cdot), \ldots, B_{k_n+\hbar+1}(\cdot))^T$ be a set of normalized B-spline basis for $\mathcal{F}_n$; see Schumaker (1981, Chap. 4) for details on the construction of B-spline basis functions. From now on, we write $q_n = k_n + \hbar + 1$.

Recall that our interest lies in selecting variables with nonzero varying and constant effects. By Schumaker (1981, Chap. 4), there exists a transformation matrix $G$ such that $G\tilde{\boldsymbol{\pi}}(u) = \big(1, \bar{\boldsymbol{\pi}}(u)^T\big)^T \doteq \boldsymbol{\pi}(u)$, where each component of $\bar{\boldsymbol{\pi}}(u)$ depends on $u$. Therefore, we can approximate each $\alpha_k(u)$ by $\alpha_k(u) \approx \boldsymbol{\pi}(u)^T\boldsymbol{\gamma}_k \doteq \gamma_{k,1} + \bar{\boldsymbol{\pi}}(u)^T\boldsymbol{\gamma}_{k*}$. Here $\boldsymbol{\gamma}_k \doteq (\gamma_{k,1}, \boldsymbol{\gamma}_{k*}^T)^T$ is the $k$th spline coefficient vector, where $\gamma_{k,1}$ corresponds to the constant part of the coefficient functional, and $\boldsymbol{\gamma}_{k*} = (\gamma_{k,2}, \ldots, \gamma_{k,q_n})^T$ corresponds to the varying part. To fix notation, we take $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_0^T, \ldots, \boldsymbol{\gamma}_p^T)^T$, $\boldsymbol{\Pi}(u, \boldsymbol{X}) = \big(X^{(0)}\boldsymbol{\pi}(u)^T, \ldots, X^{(p)}\boldsymbol{\pi}(u)^T\big)^T$, $\boldsymbol{\pi}_i = \boldsymbol{\pi}(U_i)$,

and $\mathbf{\Pi}_i = \mathbf{\Pi}(U_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$. With the B-spline approximation, (1.1) can be rewritten as

$$Y_i \approx \mathbf{\Pi}_i^T \boldsymbol{\gamma} + e_i. \tag{2.1}$$

If $\|\boldsymbol{\gamma}_{k*}\|_{L_1} = \sum_{l=2}^{q_n} |\gamma_{k,l}| = 0$, it means that the $k$th covariate has only a constant effect and, if in addition $\gamma_{k,1} = 0$, then the $k$th covariate has no effect at all. Therefore, $\boldsymbol{\gamma}_{k*}$ can be treated as a group.

Let the loss function be $g(t) = t^2$ for least squares regression, and $g(t) = \{\tau - I(t < 0)\}t$ for quantile regression at a given quantile level $0 < \tau < 1$. While mean regression is confined to estimating the mean function of the response, quantile regression offers a systematic strategy for examining how covariates influence the location, scale, and shape of the entire response distribution. Fitting data at a set of quantiles provides a more complete description of the response distribution than does the mean.

Since quantile regression involves a non-differentiable and asymmetric $L_1$ loss function, computation is challenging when penalizing the $L_2$ norm of $\boldsymbol{\gamma}_{k*}$, though it is more commonly used in the literature for group selection (Yuan and Lin (2006), Wang, Li, and Huang (2008)). This motivated us to consider the penalization based on group $L_1$ norm, for which standard linear programming can be employed to solve the optimization problem. To conduct unified variable selection, we minimize the penalized objective function

$$\begin{aligned} l(\boldsymbol{\gamma}) = &\sum_{i=1}^{n} g(Y_i - \mathbf{\Pi}_i^T \boldsymbol{\gamma}) + n\lambda_{1,n} \sum_{k=1}^{p} \tilde{\omega}_{k*} \|\boldsymbol{\gamma}_{k*}\|_{L_1} \\ &+ n\lambda_{2,n} \sum_{k=1}^{p} \tilde{\omega}_{k,1} |\gamma_{k,1}| I\left(\|\boldsymbol{\gamma}_{k*}\|_{L_1} = 0\right), \end{aligned} \tag{2.2}$$

where $\lambda_{1,n}$ and $\lambda_{2,n}$ are the penalization parameters, $\tilde{\omega}_{k*}$ and $\tilde{\omega}_{k,1}$ are the adaptive weights, $\|\boldsymbol{\gamma}_{k*}\|_{L_1} = \sum_{j=2}^{q_n} |\gamma_{k,j}|$ is the $L_1$-norm of $\boldsymbol{\gamma}_{k*}$. The second term of $l(\boldsymbol{\gamma})$ aims to separate varying and constant effects by penalizing the $L_1$ norm of the varying parts of each varying coefficient. To select variables with nonzero constant coefficients, we include the indicator function in the third term of $l(\boldsymbol{\gamma})$ to penalize only those variables that tend to have constant effects. Therefore, if the variable $X^{(k)}$ has a constant effect, then all components of $\boldsymbol{\gamma}_{k*}$ are shrunk exactly to zero. On the other hand, if the variable $X^{(k)}$ has no effect at all, both $\boldsymbol{\gamma}_{k*}$ and $\gamma_{k,1}$ are shrunk to zero. In practice, with the indicator function involved, it is very difficult to minimize the penalized objective function $l(\boldsymbol{\gamma})$. Motivated by (2.2), we propose an alternative two-step iterative procedure that is computationally convenient. Following arguments used for proving Theorems

1 and 2 in the Appendix, it can be shown that the minimizer of (2.2) and the estimator from our iterative two-step procedure are asymptotically equivalent. However, in finite samples, the iterative procedure does not necessarily converge to the minimizer of (2.2).

We use notations $V$, $C$ and $Z$ to denote the subsets of covariates with varying effects, nonzero constant effects and zero effects, respectively. The full model assigns all the covariates to $V$. In the following, for ease of representation, we do not distinguish the estimators from quantile regression and least squares regression wherever this is clear from the context. The proposed procedure is as follows.

*Step* 1. Obtain the penalized estimator $\hat{\boldsymbol{\gamma}}^{VC}$ by minimizing

$$l_1(\boldsymbol{\gamma}) = \sum_{i=1}^{n} g(Y_i - \boldsymbol{\Pi}_i^T \boldsymbol{\gamma}) + n\lambda_{1,n} \sum_{k=1}^{p} \tilde{\omega}_{k*} \|\boldsymbol{\gamma}_{k*}\|_{L_1} \tag{2.3}$$

with respect to $\boldsymbol{\gamma}$, where $\tilde{\omega}_{k*}$ are the adaptive weights. In (2.3), we penalize each varying coefficient in a group manner by assigning the same penalty to each component of $\boldsymbol{\gamma}_{k*}$ for $k = 1, \ldots, p$. Note that our proposed method differs from the group LASSO of Yuan and Lin (2006), in which the penalty is applied to the $L_2$ norm of each coefficient group. We use $\tilde{\omega}_{k*} = \|\tilde{\boldsymbol{\gamma}}_{k*}\|_{L_1}^{-1}$, where $\tilde{\boldsymbol{\gamma}}_k = (\tilde{\gamma}_{k,1}, \tilde{\boldsymbol{\gamma}}_{k*}^T)^T = (\tilde{\gamma}_{k,1}, \ldots, \tilde{\gamma}_{k,q_n})^T$ as the unpenalized estimator obtained by minimizing $l_1(\boldsymbol{\gamma})$ with $\lambda_{1,n} = 0$. With this penalization, the $\boldsymbol{\gamma}_{k*}$ are shrunk toward zero if the $k$th covariate has a constant effect, leading to an automatic separation of the varying and constant effects. We move the $k$th covariate from $V$ to $C$ if $\|\hat{\boldsymbol{\gamma}}_{k*}^{VC}\|_{L_1} = 0$, otherwise retain it in $V$. After Step 1, the full model is reduced to a partially linear varying coefficient model.

*Step* 2. For each $k = 1, \ldots, p$, we take $\boldsymbol{\gamma}_k^1 = (\gamma_{k,1}, 0_{q_n-1}^T)^T$ if $X^{(k)} \in C$, and $\boldsymbol{\gamma}_k^1 = \boldsymbol{\gamma}_k$ if $X^{(k)} \in V$, where $C$ and $V$ are the covariate subsets obtained after the previous step, and $\boldsymbol{\gamma}^1 = (\boldsymbol{\gamma}_0^{1T}, \boldsymbol{\gamma}_1^{1T}, \ldots, \boldsymbol{\gamma}_p^{1T})^T$. We obtain the estimator $\hat{\boldsymbol{\gamma}}^{CZ}$ by minimizing

$$l_2(\boldsymbol{\gamma}^1) = \sum_{i=1}^{n} g(Y_i - \boldsymbol{\Pi}_i^T \boldsymbol{\gamma}^1) + n\lambda_{2,n} \sum_{k=1}^{p} \hat{\omega}_{k,1} |\gamma_{k,1}^1| I\left(\|\hat{\boldsymbol{\gamma}}_{k*}^{VC}\|_{L_1} = 0\right) \tag{2.4}$$

with respective to the unknown parameters in $\boldsymbol{\gamma}^1$, where the adaptive weights are set as $\hat{\omega}_{k,1} = |\hat{\gamma}_{k,1}^{VC}|^{-1}$. Since $\hat{\boldsymbol{\gamma}}_{k*}^{VC}$ is obtained in the previous step, the second term in (2.4) corresponds to an adaptive $L_1$ penalty. In this step, we aim to exclude the irrelevant variables by assigning the adaptive LASSO penalty only to the terms that have been determined to be constant. We move any $X^{(k)}$ that is assigned to $C$ in the previous step to $Z$ if the corresponding $|\hat{\gamma}_{k,1}^{CZ}| = 0$.

*Step* 3. Iterate Steps 1 and 2 to convergence. We take the estimator at convergence as $\hat{\boldsymbol{\gamma}}^{final}$. Throughout the iteration, the covariates assigned to $C$ have only constant effects, and those assigned to $Z$ are excluded from the model. In the end, for each $k = 0, 1, \ldots, p$, $\alpha_k(u)$ can be estimated by $\hat{\alpha}_k(u) = \pi(u)^T \hat{\boldsymbol{\gamma}}_k^{final}$ if it is chosen as a varying coefficient, and by $\hat{\gamma}_{k,1}^{final}$ if it is chosen as a nonzero constant.

**Remark 1.** With the current iteration procedure, if the varying parts $\gamma_{k*}$ are not shrunk to zero in Step 1, no penalty is applied to the constant parts $\gamma_{k,1}$ in Step 2. Consequently, after Step 2, a redundant variable may be selected to have a constant effect, and a constant coefficient may be selected as varying. However, such over-fitting affects only the efficiency of the subsequent steps, and a non-varying coefficient still may be shrunk to a constant or zero as the iteration continues. On the other hand, if one varying coefficient is incorrectly shrunk as a constant in Step 1, due to over-shrinkage, the coefficient is selected as either a constant or zero at convergence. As a result, when the procedure under-fits the model at convergence, it is more likely to select a varying coefficient to be constant. As suggested by one referee, we investigated the estimator obtained by flipping the orders of Steps 1 and 2. The resulting estimator has the same asymptotic properties as the proposed estimator. However, numerical results for finite samples suggested that flipping the orders of Steps 1 and 2 introduces unnecessary bias to the varying coefficient estimates.

### 2.2. Computational algorithms

In this subsection, we describe the variable selection algorithms for least squares and quantile regression VC models, respectively.

For least squares regression, the minimization of (2.3) involves quadratic loss and a LASSO-type penalty. We propose an iterative algorithm based on the local quadratic approximation of Fan and Li (2001), which performs well and is computationally simple.

Suppose we have an initial value $\boldsymbol{\gamma}^{(0)}$ that is close to the minimizer of (2.3). For example, we can choose $\boldsymbol{\gamma}^{(0)}$ as the unpenalized estimator

$$\boldsymbol{\gamma}^{(0)} = \Big( \sum_{i=1}^{n} \boldsymbol{\Pi}_i \boldsymbol{\Pi}_i^T \Big)^{-1} \sum_{i=1}^{n} \boldsymbol{\Pi}_i Y_i.$$

By the local quadratic approximation, for $k = 1, \ldots, p, l = 2, \ldots, q_n$,

$$|\gamma_{k,l}| \approx |\gamma_{k,l}^{(0)}| + \operatorname{sign}\Big(\gamma_{k,l}^{(0)}\Big)\Big(\gamma_{k,l} - \gamma_{k,l}^{(0)}\Big) \approx |\gamma_{k,l}^{(0)}| + \frac{1}{2|\gamma_{k,l}^{(0)}|}\Big(\gamma_{k,l}^2 - {\gamma_{k,l}^{(0)}}^2\Big),$$

where sign$(\cdot)$ is the sign function. Except for a constant term, (2.3) can be approximated by

$$\sum_{i=1}^{n}(Y_i - \mathbf{\Pi}_i^T\boldsymbol{\gamma})^2 + n\lambda_{1,n}\sum_{k=1}^{p}\tilde{\omega}_{k*}\sum_{l=2}^{q_n}\frac{\gamma_{k,l}^2}{2|\gamma_{k,l}^{(0)}|}.$$

Then, we can obtain $\hat{\boldsymbol{\gamma}}^{VC}$ by iteratively conducting the following ridge regression to convergence

$$\boldsymbol{\gamma}^{(j)} = \Big(\sum_{i=1}^{n}\mathbf{\Pi}_i\mathbf{\Pi}_i^T + n\lambda_{1,n}\Omega^{(j-1)}\Big)^{-1}\sum_{i=1}^{n}\mathbf{\Pi}_i Y_i,$$

where

$$\Omega^{(j-1)} = \operatorname{diag}\left\{\mathbf{0}_{q_n}^T, \Omega_{1,(j-1)}^T, \ldots, \Omega_{p,(j-1)}^T\right\},$$

with $\Omega_{k,(j-1)} = \tilde{\omega}_{k*}\Big\{0, \Big(2|\gamma_{k,2}^{(j-1)}|\Big)^{-1}, \ldots, \Big(2|\gamma_{k,q_n}^{(j-1)}|\Big)^{-1}\Big\}^T$ for $k = 1, \ldots, p$, and $\gamma_{k,l}^{(j-1)}$ is the estimate of $\gamma_{k,l}$ at the ($j$-1)th iteration with $j \geq 1$. During the iteration, once $\|\boldsymbol{\gamma}_{k*}^{(j)}\|_{L_1} < \epsilon$, we set $\hat{\boldsymbol{\gamma}}_{k*}^{VC} = \mathbf{0}$, where $\epsilon > 0$ is a small positive value. In our implementation, we use $\epsilon = 10^{-4}$.

The algorithm for minimizing (2.4) is similar and thus is omitted.

For quantile regression, with the aid of slack variables, the minimization of (2.3) and (2.4) can be easily cast as a linear programming problem, then solved by using existing linear programming packages such as the R package *quantreg*.

### 2.3. Selection of tuning parameters

To implement the proposed method, we have to choose the tuning parameters, the degree of B-splines $\hbar$, the number of interior knots $k_n$, and the penalization parameters $\lambda_{j,n}$, $j = 1, 2$.

For varying coefficient models, since the effect of splines is multiplicative, higher degree splines induce complicated interactions of the form $xu$, $xu^2$, $xu^3$, $xu^4, \cdots$, and collinearity between variables in the model. Therefore, we suggest using lower degree splines such as linear, quadratic, or cubic splines. In our numerical studies, we use $\hbar = 3$ corresponding to cubic splines, but lower orders can also be used if we believe the functional coefficients are less smooth.

At each iteration of the two-step procedure, we choose $k_n$ and $\lambda_{1,n}$, $\lambda_{2,n}$ by minimizing the Schwarz-type Information Criterion (SIC, Schwarz (1978)) as described below. The locations of the interior knots are taken equally spaced on $[0, 1]$.

Suppose that after the $m$th iteration, we have reduced the full varying coefficient model to a partially linear varying coefficient model. We use $V_m$, $C_m$, and $Z_m$ to denote the subsets of covariates with varying effects, nonzero constant effects, and zero effects, and use $v_m$, $c_m$, and $z_m$ to denote the number of covariates in these subsets, respectively. Let $\boldsymbol{\gamma}_k^m = \boldsymbol{\gamma}_k$ if $X^{(k)} \in V_m$, $\boldsymbol{\gamma}_k^m = (\gamma_{k,1}, \mathbf{0}_{q_n-1}^T)^T$ if $X^{(k)} \in C_m$, and $\boldsymbol{\gamma}_k^m = \mathbf{0}_{q_n}$ if $X^{(k)} \in Z_m$.

In Step 1, we first choose $k_n$ as the minimizer of

$$SIC_0(k) = \log \sum_{i=1}^{n} g\Big(Y_i - \boldsymbol{\Pi}_i^T \hat{\boldsymbol{\gamma}}_{m,k}\Big) + \frac{\log n}{\varrho\ n}\{v_m(k + \hbar + 1) + c_m\},$$

where $\varrho = 1$ and 2 for least squares regression and quantile regression, respectively, and $\hat{\boldsymbol{\gamma}}_{m,k}$ is the minimizer of

$$l_1(\boldsymbol{\gamma}^m) = \sum_{i=1}^{n} g\Big(Y_i - \boldsymbol{\Pi}_i^T \boldsymbol{\gamma}^m\Big) + n\lambda_{1,n} \sum_{k=1}^{p} \tilde{\omega}_{m,k*} \|\boldsymbol{\gamma}_{k*}^m\|_{L_1} \tag{2.5}$$

with respect to the unknown components of $\boldsymbol{\gamma}^m$ with $k_n = k$ and $\lambda_{1,n} = 0$; see Wang, Zhu, and Zhou (2009). for a similar criterion for knots selection. Conditional on the selected $k_n$, we take $\lambda_{1,n}$ as the minimizer of

$$SIC_1(\lambda_1) = \log \sum_{i=1}^{n} g\Big(Y_i - \boldsymbol{\Pi}_i^T \hat{\boldsymbol{\gamma}}_{m,\lambda_1}\Big) + \frac{\log n}{\varrho\ n} edf_1,$$

where $\hat{\boldsymbol{\gamma}}_{m,\lambda_1}$ is the minimizer of (2.5) with respect to the unknown components of $\boldsymbol{\gamma}^m$ with $\lambda_{1,n} = \lambda_1$, $\tilde{\omega}_{m,k*} = \|\tilde{\boldsymbol{\gamma}}_{k*}^m\|_{L_1}^{-1}$, with $\tilde{\boldsymbol{\gamma}}^m$ being the unpenalized estimator obtained by minimizing $l_1(\boldsymbol{\gamma}^m)$ with $\lambda_{1,n} = 0$. For least squares regression, $edf_1$ is defined as the total number of varying and nonzero constant coefficients (Wang and Xia (2009)). For quantile regression, $edf_1$ is the number of interpolated $Y_i$'s, i.e., the number of zero residuals (Koenker, Ng, and Portnoy (1994)), which provides a measure of the effective dimensionality of the fitted model.

In Step 2, we first replace $\boldsymbol{\gamma}_{k*}^m$ with $\mathbf{0}_{q_n-1}$ in $\boldsymbol{\gamma}^m$ if $\hat{\boldsymbol{\gamma}}_{m,\lambda_{1,n};k*} = \mathbf{0}$, and take the new coefficient vector to be $\boldsymbol{\gamma}^{m,2}$. Conditional on $\hat{\boldsymbol{\gamma}}_{m,\lambda_{1,n}}$, a function of $\lambda_{1,n}$, obtained in Step 1, we select $\lambda_{2,n}$ as the minimizer of

$$SIC_2(\lambda_2) = \log \sum_{i=1}^{n} g\Big(Y_i - \boldsymbol{\Pi}_i^T \hat{\boldsymbol{\gamma}}_{m,\lambda_2}\Big) + \frac{\log n}{\varrho\ n} edf_2,$$

where $\hat{\boldsymbol{\gamma}}_{m,\lambda_2}$ is the minimizer of

$$l_2(\boldsymbol{\gamma}^{m,2}) = \sum_{i=1}^{n} g\Big(Y_i - \boldsymbol{\Pi}_i^T \boldsymbol{\gamma}^{m,2}\Big) + n\lambda_{2,n} \sum_{k=1}^{p} \hat{\omega}_{m,k,1} |\gamma_{k,1}^{m,2}| I\left(\|\hat{\boldsymbol{\gamma}}_{m,\lambda_{1,n};k*}\|_{L_1} = 0\right)$$

with respect to the unknown components of $\boldsymbol{\gamma}^{m,2}$, where $\hat{\omega}_{m,k,1} = |\hat{\gamma}_{m,\lambda_{1,n};k,1}|^{-1}$, and $edf_2$ is defined similarly as $edf_1$ in $SIC_1$.

Even though the implementation seems complicated, the computational cost is not great as the algorithm often converges within two iterations. For Example 3 with $n = 500$ and $p = 50$ variables, we generated 100 replicates and searched $k_n, \lambda_{1,n}, \lambda_{2,n}$ on a 3-point, 40-point, 40-point grid. The simulation was carried out using R on a computer with 2.40GHz CPU and 4.00GB RAM. The average computing time (and standard error) are 61.85 (4.69) and 129.00 (0.33) seconds for median regression and least squares regression, respectively.

## 3. Asymptotic Properties

Throughout this paper, we use $a_n \sim b_n$ to mean that $a_n$ and $b_n$ have the same order as $n \to \infty$. Suppose that there are $s$ true relevant covariates in model (1.1), in which $\nu$ of them have varying effects on the response, and $s-\nu$ covariates have constant effects. Without loss of generality, assume $\{\alpha_k(u), k = 1, \ldots, \nu\}$ are the varying coefficients, $\{\alpha_k(u) = \alpha_k, k = \nu+1, \ldots, s\}$ are the nonzero constant coefficients, and $\alpha_k(u) \equiv 0, k = s+1, \ldots, p$. Let $\boldsymbol{\alpha}_{(c)} = (\alpha_{\nu+1}, \ldots, \alpha_s)^T$ be the true constant coefficient vector. To establish the asymptotic results in this paper, we make the following assumptions.

A1. $\alpha_k(u) \in \mathcal{H}_r, k = 0, 1, \ldots, \nu$, for some $r > 3/2$, where $\mathcal{H}_r$ is the collection of all functions on $[0, 1]$ whose $d$th order derivative is *Hölder* of order $\upsilon$, $r \equiv d + \upsilon$.

A2. The random design vectors $\{\boldsymbol{X}_i, i = 1, \ldots, n\}$ are uniformly bounded in probability. The eigenvalues of the matrix $n^{-1}\boldsymbol{X}^T\boldsymbol{X}$ are bounded away from zero and infinity in probability, where $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^T$.

A3. The density function of $U_i$, $f_U(u)$, is continuous and bounded away from zero and infinity on [0,1].

A4. The penalization parameters satisfy $n^{1/2}k_n^{1/2}\max\{\lambda_{1,n}, \lambda_{2,n}\} \to 0$ and $nk_n^{-3/2}\min\{\lambda_{1,n}, \lambda_{2,n}\} \to \infty$.

Theorems 1 and 2 state the asymptotic properties of the penalized least squares estimator $\hat{\boldsymbol{\alpha}}(u)^{LS}$ and the penalized quantile regression estimator $\hat{\boldsymbol{\alpha}}(u)^{QR}$, respectively, both obtained by the two-step iterative procedure.

**Theorem 1.** *Suppose* A1−A4 *hold and* $k_n \sim n^{1/(2r+1)}$. *Assume that* $e_i$ *are independent with zero conditional mean given* $(\boldsymbol{X}_i, U_i)$, *and there exists positive constants* $M$ *and* $\delta$ *such that* $\max_{i=1}^n E(e_i^{2+\delta}) \le M$. *Then we have*

(a) *with probability approaching 1,* $\hat{\alpha}_k(u)^{LS}$ *are nonzero constants for* $k = \nu + 1, \ldots, s$, *and* $\hat{\alpha}_k(u)^{LS} = 0$ *for* $k = s+1, \ldots, p$;

(b) $n^{-1}\sum_{i=1}^{n}\left\{\hat{\alpha}_k(U_i)^{LS}-\alpha_k(U_i)\right\}^2=O_p\left(n^{-2r/(2r+1)}\right), k=0,1,\ldots,\nu$;

(c) $\tilde{\Lambda}_{n*}^{-1/2}\tilde{K}_{n*}\left\{\hat{\boldsymbol{\alpha}}_{(c)}^{LS}-\boldsymbol{\alpha}_{(c)}\right\}\to N(0,I_{s-\nu})$, *where* $\hat{\boldsymbol{\alpha}}_{(c)}^{LS}$ *is the penalized least squares estimator of* $\boldsymbol{\alpha}_{(c)}$, *and* $\tilde{K}_{n*}$ *and* $\tilde{\Lambda}_{n*}$ *are defined as in* (A.15) *in the Appendix.*

**Theorem 2.** *Suppose* A1−A4 *hold and* $k_n\sim n^{1/(2r+1)}$. *For a given quantile level* $0<\tau<1$, *assume that* $e_i$ *are independently distributed with zero conditional* $\tau$*-th quantile given* $(\boldsymbol{X}_i,U_i)$, *and the conditional density function of* $e_i$ *given* $(\boldsymbol{X}_i,U_i)$, $f_i(\cdot)$, *is uniformly bounded from infinity, and continuous and bounded away from zero in a neighborhood of zero. Then we have*

(a) *with probability approaching 1,* $\hat{\alpha}_k(u)^{QR}$ *are nonzero constants for* $k=\nu+1,\ldots,s$, *and* $\hat{\alpha}_k(u)^{QR}=0$ *for* $k=s+1,\ldots,p$;

(b) $n^{-1}\sum_{i=1}^{n}\left\{\hat{\alpha}_k(U_i)^{QR}-\alpha_k(U_i)\right\}^2=O_p\left(n^{-2r/(2r+1)}\right), k=0,1,\ldots,\nu$;

(c) $\Lambda_{n*}^{-1/2}K_{n*}\left\{\hat{\boldsymbol{\alpha}}_{(c)}^{QR}-\boldsymbol{\alpha}_{(c)}\right\}\to N(0,I_{s-\nu})$, *where* $\hat{\boldsymbol{\alpha}}_{(c)}^{QR}$ *is the penalized quantile regression estimator of* $\boldsymbol{\alpha}_{(c)}$, *and* $\Lambda_{n*}$ *and* $K_{n*}$ *are defined as in* (A.11) *in the Appendix.*

The theorems establish the asymptotic properties of the proposed estimators. Part (a) suggests that the proposed penalized procedure provides consistent variable selection and automatic separation of different types of effects. Part (b) states that the estimated varying coefficients achieve the optimal nonparametric convergence rate under the smoothness assumption A1 (Stone (1982)). Part (c) shows that the estimated constant coefficients have the same asymptotic distribution as their counterparts obtained when the true model is known.

**Remark 2.** Asymptotically, even without iteration, the estimator $\hat{\boldsymbol{\gamma}}^{CZ}$ at the second step already possesses the same asymptotic properties as $\hat{\boldsymbol{\gamma}}^{final}$. However, our experience suggests that iteration provides finite-sample improvement when some covariates either have no effects or have constant effects. One explanation is that, for sparse models, the first two steps lead to more efficient penalized estimators than the $\tilde{\boldsymbol{\gamma}}$ obtained under the full model. Therefore, the adaptive weights based on the penalized estimators shrink the effects of irrelevant covariates more effectively toward zero. In our work, convergence is often achieved within two iterations.

## 4. Simulation Study

In this section, we assess the finite sample performance of the proposed method through three simulation examples.

**Example 1.** We compare the performance of the proposed method for least squares regression (referred to as *LSR*) to two existing methods, the modified

cross-validation-based method (*mCV*) of Xia, Zhang, and Tong (2004) and the component selection and smoothing operator (*COSSO*) of Leng (2009). The *COSSO* was first studied by Lin and Zhang (2006) for additive models, and it was extended by Leng (2009) to varying coefficient models. Both *mCV* and *COSSO* were designed to determine which coefficients are varying, but neither is able to select variables with nonzero constant effects. Therefore, for fair comparison, we focus on the separation of variables with varying and constant effects.

As in Xia, Zhang, and Tong (2004) and Leng (2009), 100 replicates were randomly generated from

$$Y_i = \alpha_0(U_i) + \alpha_1(U_i)X_i^{(1)} + a\alpha_2(U_i)X_i^{(2)} + \sum_{k=3}^{6} \alpha_k(U_i)X_i^{(k)} + \sigma_0 e_i,\ i = 1, \ldots, n,$$

where $U_i \sim Uniform(0,1)$, and $X_i^{(1)}, \ldots, X_i^{(6)}$ and $e_i$ are independent $N(0,1)$ variables. The coefficient functions $\alpha_k(u)$ were $\alpha_0(u) = \exp\{-32(u-0.5)^2\}$, $\alpha_1(u) = \sin(2\pi u)$, $\alpha_2(u) = \cos(2\pi u)$, $\alpha_3 = 1, \alpha_4 = -1, \alpha_5 = 1$, and $\alpha_6 = 0$. The parameter $a$ determines the extent to which $\alpha_2(u)$ varies with $u$. The parameter $\sigma_0$ controls the signal-to-noise level. For fair comparison, we make a slight modification to our algorithm, so that the intercept term $\alpha_0(u)$ is also penalized, as in the other two competing methods.

Table 1 summarizes the automatic separation results. The results of *mCV* and *COSSO* are from Xia, Zhang, and Tong (2004) and Leng (2009). When the variability of the varying coefficient $\alpha_2(u)$ is small ($a = 0.3$) and the noise is large ($\sigma_0 = 0.5$), *COSSO* seems to perform the best, while our method is comparable to it. However, for data sets with larger variability of $\alpha_2(u)$ or smaller noise, our method performs similarly to *mCV*, and both are clearly better than *COSSO* in terms of separating the constant and varying coefficients. Note that, in this example, $X_i^{(6)}$ has no effect on $Y_i$ and thus is redundant. Our method is able to select the exact true model with high proportion; see the last column of Table 1. However, neither *mCV* nor *COSSO* can eliminate $X_i^{(6)}$ in the final model.

**Example 2.** In this example, we investigate the performance of the proposed method for least squares regression (*LSR*) and median regression (*MR*). We generated 1,000 replicates, each consisting of $n = 500$ observations obtained as

$$Y_i = \alpha_0(U_i) + \sum_{k=1}^{10} \alpha_k(U_i)X_i^{(k)} + e_i,$$

where $U_i \sim Uniform(0,1)$, and $X_i^{(k)}, k = 1, \ldots, 10$, are independent $N(0,1)$ variables. Here $X_i^{(k)}, k = 1, 2$, have varying effects, $X_i^{(k)}, k = 3, 4$, have nonzero

Table 1. Frequencies that $\alpha_k(u), k = 0, 1, \ldots, 6$ are selected to be constant in 100 replicates.

| $a$ | $\sigma_0$ | $n$ | Method | $\alpha_0$ | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $\alpha_4$ | $\alpha_5$ | $\alpha_6$ | Md. | Oracle |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.3 | 0.5 | 100 | *LSR* | 1 | 0 | 25 | 88 | 93 | 90 | 90 | 52 | 52 |
| | | | *mCV* | 0 | 0 | 66 | 99 | 99 | 100 | 100 | 34 | / |
| | | | *COSSO* | 2 | 0 | 18 | 91 | 93 | 87 | 91 | 57 | / |
| | | 200 | *LSR* | 0 | 0 | 2 | 95 | 96 | 97 | 95 | 84 | 82 |
| | | | *mCV* | 0 | 0 | 33 | 100 | 100 | 100 | 100 | 67 | / |
| | | | *COSSO* | 0 | 0 | 0 | 96 | 96 | 93 | 98 | 87 | / |
| | 0.2 | 100 | *LSR* | 0 | 0 | 7 | 97 | 98 | 98 | 98 | 86 | 86 |
| | | | *mCV* | 0 | 0 | 2 | 99 | 99 | 100 | 100 | 97 | / |
| | | | *COSSO* | 0 | 0 | 0 | 98 | 100 | 95 | 97 | 90 | / |
| | | 200 | *LSR* | 0 | 0 | 1 | 100 | 100 | 100 | 100 | 99 | 99 |
| | | | *mCV* | 0 | 0 | 0 | 100 | 100 | 100 | 100 | 100 | / |
| | | | *COSSO* | 0 | 0 | 0 | 96 | 94 | 95 | 98 | 86 | / |
| $5n^{-1/2}$ | 0.5 | 100 | *LSR* | 2 | 0 | 3 | 95 | 97 | 94 | 94 | 81 | 78 |
| | | | *mCV* | 0 | 0 | 14 | 99 | 100 | 100 | 100 | 85 | / |
| | | | *COSSO* | 2 | 0 | 0 | 95 | 94 | 88 | 90 | 68 | / |
| | | 200 | *LSR* | 0 | 0 | 2 | 98 | 100 | 100 | 100 | 96 | 94 |
| | | | *mCV* | 0 | 0 | 10 | 100 | 100 | 100 | 100 | 90 | / |
| | | | *COSSO* | 0 | 0 | 0 | 95 | 95 | 93 | 98 | 86 | / |
| | 0.2 | 100 | *LSR* | 0 | 0 | 0 | 100 | 100 | 99 | 100 | 99 | 99 |
| | | | *mCV* | 0 | 0 | 0 | 99 | 99 | 100 | 100 | 99 | / |
| | | | *COSSO* | 0 | 0 | 0 | 95 | 96 | 90 | 93 | 74 | / |
| | | 200 | *LSR* | 0 | 0 | 0 | 100 | 100 | 100 | 100 | 100 | 100 |
| | | | *mCV* | 0 | 0 | 3 | 100 | 100 | 100 | 100 | 97 | / |
| | | | *COSSO* | 0 | 0 | 0 | 94 | 93 | 94 | 99 | 83 | / |

Md.: frequency that the varying and constant coefficients are separated correctly; Oracle: frequency that the true model is selected.

constant effects, and the others are redundant variables. We took two distributions for generating the random error $e_i$: standard normal and $t(3)$, the t-distribution with 3 degrees of freedom. We set $\alpha_0(u) = 15+20\sin(\pi u/2), \alpha_1(u) = 2 - 3\cos\{(6u-5)\pi/3\}, \alpha_2(u) = 6 - 6u$, $\alpha_3(u) \equiv 1.5$, $\alpha_4(u) \equiv 2$, and $\alpha_k(u) \equiv 0$ for $k = 5, \ldots, 10$.

Table 2 summarizes the results for the normal error and the $t(3)$ error. For comparison, we also include the results of median regression and least squares regression obtained under the true model, referred to as Oracle.M and Oracle.LS, respectively. Parts (I) and (I') show the variable selection results, including the percentage of correctly selecting the true model (Oracle Perc.), the average number of effects (excluding the intercept) that are selected as varying (Aver.v), and the average number of redundant variables that are incorrectly selected (Aver.r). The ideal values of Aver.v and Aver.r are 2 and 0, respectively. Parts (II) and

(II') summarize the estimation results, including the mean squared error (MSE) of $\hat{\alpha}_k$ conditional on the models where $\alpha_k$ is selected as constant, $k = 3, 4$, and the integrated mean squared error (IMSE) of $\hat{\alpha}_k(u)$, $k = 0, 1, 2$, defined as

$$\text{IMSE}\{\hat{\alpha}_k(u)\} = \frac{1}{1,000}\sum_{i=1}^{1,000}\frac{1}{100}\sum_{a=1}^{100}\{\hat{\alpha}_{k,i}(u_a) - \alpha_k(u_a)\}^2,$$

where $\{u_a, a = 1, \ldots, 100\}$ is a grid equally spaced on $[0.02, 0.98]$, $\hat{\alpha}_{k,i}(u_a) = \hat{\boldsymbol{\gamma}}_{k,i}^T\boldsymbol{\pi}(u_a)$, and $\hat{\boldsymbol{\gamma}}_{k,i}$ are the estimates of $\alpha_k(u_a)$ and $\boldsymbol{\gamma}_k$ in the $i$th replicate, respectively. The values in parentheses are the Monte Carlo standard errors of the MSE and IMSE estimates.

Simulation results have the estimates of *MR* and *LSR* close to those of their oracle counterparts. For normal errors, *MR* and *LSR* perform similarly in terms of variable selection, but *LSR* is slightly more efficient for estimating the nonzero constant and varying coefficients. However, when the error distribution has heavy tails, *MR* produces not only more accurate variable selection, but also much more efficient estimation than *LSR*.

**Example 3.** In this example, we considered the performance of the proposed method with high dimensional covariates. We generated 1,000 replicates, each consisting of $n = 500$ observations obtained as

$$Y_i = \alpha_0(U_i) + \sum_{k=1}^{50}\alpha_k(U_i)X_i^{(k)} + e_i,$$

where the index variable $U_i \sim Uniform(0, 1)$, and $X_i^{(k)}, k = 1, \ldots, 50$, are independent $N(0, 1)$ variables. Here $X_i^{(k)}, k = 1, 2$, have varying effects on the response, $X_i^{(k)}, k = 3, \ldots, 6$, have nonzero constant effects, and the others are redundant variables. As in Example 2, we considered two distributions for generating the random error $e_i$: standard normal and $t(3)$. We set $\alpha_k(u), k = 0, 1, \ldots, 4$, as in Example 2, $\alpha_5(u) \equiv -1$, $\alpha_6(u) \equiv 1$, and $\alpha_k(u) \equiv 0$ for $k = 7, \ldots, 50$. We used cubic splines, therefore the dimension of spline coefficients to estimate was $50(k_n + 4)$ with $k_n$ selected as 2 or 3, comparable to the sample size, $n = 500$.

Table 3 summarizes the results for the normal error and the $t(3)$ error. Overall, we can see that the proposed *MR* and *LSR* methods perform quite well even when the dimension of spline coefficients is comparable to the sample size, and the results are similar to Example 2. It is surprising that the estimated varying coefficients of *LSR* have smaller IMSE's than the oracle estimators obtained under the true model. Further investigation shows that the smaller IMSE of *LSR* is mainly due to the smaller bias, which is possibly caused by the under-shrinkage

Table 2. Variable selection and estimation results in Example 2.

| Methods | *MR* | Oracle.M | *LSR* | Oracle.LS |
|---|---|---|---|---|
| | **(I): variable selection results for $N(0,1)$ error** | | | |
| Oracle Perc. | 86.8 | 100 | 83.1 | 100 |
| Aver.v | 2.07 | 2 | 2.18 | 2 |
| Aver.r | 0.14 | 0 | 0.14 | 0 |
| | **(II): estimation results for $N(0,1)$ error** | | | |
| $10^2\times$ IMSE | | | | |
| $\alpha_0(u)$ | 20.29 (0.38) | 20.31 (0.38) | 19.64 (0.35) | 19.86 (0.35) |
| $\alpha_1(u)$ | 7.71 (0.12) | 7.70 (0.12) | 7.06 (0.10) | 7.08 (0.10) |
| $\alpha_2(u)$ | 3.11 (0.06) | 3.12 (0.06) | 2.33 (0.04) | 2.35 (0.04) |
| $10^3\times$ MSE | | | | |
| $\alpha_3$ | 3.43 (0.16) | 3.39 (0.15) | 2.24 (0.10) | 2.11 (0.10) |
| $\alpha_4$ | 3.26 (0.15) | 3.25 (0.15) | 2.40 (0.11) | 2.31 (0.11) |
| Methods | *MR* | Oracle.M | *LSR* | Oracle.LS |
| | **(I'): variable selection results for $t(3)$ error** | | | |
| Oracle Perc. | 85.3 | 100 | 59.9 | 100 |
| Aver.v | 2.11 | 2 | 2.60 | 2 |
| Aver.r | 0.13 | 0 | 0.47 | 0 |
| | **(II'): estimation results for $t(3)$ error** | | | |
| $10^2\times$ IMSE | | | | |
| $\alpha_0(u)$ | 21.00 (0.40) | 21.00 (0.41) | 21.95 (0.44) | 22.05 (0.44) |
| $\alpha_1(u)$ | 8.47 (0.14) | 8.45 (0.14) | 9.50 (0.17) | 9.50 (0.18) |
| $\alpha_2(u)$ | 3.68 (0.07) | 3.67 (0.07) | 5.01 (0.14) | 4.91 (0.13) |
| $10^3\times$ MSE | | | | |
| $\alpha_3$ | 4.29 (0.19) | 4.14 (0.19) | 5.98 (0.28) | 5.67 (0.29) |
| $\alpha_4$ | 4.04 (0.18) | 3.90 (0.17) | 6.09 (0.30) | 5.98 (0.28) |

Oracle.M: median estimation obtained under the true model; Oracle.LS: least squares estimation obtained under the true model; Oracle Perc.: the percentage of replications that the exact true model is selected; Aver.v: average number of varying effects selected (excluding the intercept); Aver.r: average number of redundant variables that are incorrectly selected; IMSE: integrated mean squared error; MSE: mean squared error. The values in parentheses are the Monte Carlo standard error of the IMSE and MSE estimates.

of the varying coefficients but over-shrinkage of the constant coefficients. Consequently, the MSE's of *LSR* for constant coefficients are much larger than those of *Oracle*.

## 5. Application to the Childhood Malnutrition Data in India

We apply the proposed method to a subset of the Childhood Malnutrition Data in India, to study risk factors for early childhood malnutrition. The data is based on the Demographic and Health Surveys data from 2005–2006,

Table 3. Variable selection and estimation results in Example 3.

| Methods | *MR* | Oracle.M | *LSR* | Oracle.LS |
|---|---|---|---|---|
| | **(I): variable selection results for $N(0,1)$ error** | | | |
| Oracle Perc. | 78.9 | 100 | 77.3 | 100 |
| Aver.v | 2.10 | 2 | 2.30 | 2 |
| Aver.r | 0.28 | 0 | 0.28 | 0 |
| | **(II): estimation results for $N(0,1)$ error** | | | |
| $10^2\times$ IMSE | | | | |
| $\alpha_0(u)$ | 20.34 (0.38) | 20.33 (0.38) | 17.82 (0.35) | 19.63 (0.36) |
| $\alpha_1(u)$ | 7.97 (0.13) | 8.00 (0.13) | 6.75 (0.10) | 6.96 (0.11) |
| $\alpha_2(u)$ | 3.26 (0.06) | 3.24 (0.06) | 2.23 (0.04) | 2.40 (0.04) |
| $10^3\times$ MSE | | | | |
| $\alpha_3$ | 3.14 (0.16) | 2.83 (0.13) | 2.50 (0.12) | 2.09 (0.10) |
| $\alpha_4$ | 3.35 (0.15) | 3.26 (0.15) | 2.13 (0.10) | 1.87 (0.08) |
| $\alpha_5$ | 3.78 (0.16) | 3.32 (0.14) | 3.04 (0.14) | 2.15 (0.10) |
| $\alpha_6$ | 3.65 (0.17) | 3.18 (0.14) | 3.12 (0.14) | 2.18 (0.11) |
| | **(I'): variable selection results for $t(3)$ error** | | | |
| Oracle Perc. | 79.1 | 100 | 47.4 | 100 |
| Aver.v | 2.17 | 2 | 3.23 | 2 |
| Aver.r | 0.26 | 0 | 1.25 | 0 |
| | **(II'): estimation results for $t(3)$ error** | | | |
| $10^2\times$ IMSE | | | | |
| $\alpha_0(u)$ | 21.02 (0.38) | 21.02 (0.38) | 20.70 (0.46) | 22.22 (0.43) |
| $\alpha_1(u)$ | 8.62 (0.14) | 8.63 (0.14) | 9.61 (0.20) | 9.74 (0.21) |
| $\alpha_2(u)$ | 3.76 (0.07) | 3.74 (0.07) | 5.12 (0.17) | 5.15 (0.15) |
| $10^3\times$ MSE | | | | |
| $\alpha_3$ | 4.19 (0.19) | 3.94 (0.18) | 8.56 (0.44) | 6.19 (0.33) |
| $\alpha_4$ | 4.36 (0.21) | 3.99 (0.18) | 7.05 (0.37) | 6.23 (0.41) |
| $\alpha_5$ | 4.85 (0.22) | 3.93 (0.16) | 11.29 (0.50) | 6.52 (0.32) |
| $\alpha_6$ | 4.88 (0.23) | 4.20 (0.16) | 10.75 (0.47) | 5.73 (0.27) |

Oracle.M: median estimation obtained under the true model; Oracle.LS: least squares estimation obtained under the true model; Oracle Perc.: the percentage of replications that the exact true model is selected; Aver.v: average number of varying effects selected (excluding the intercept); Aver.r: average number of redundant variables that are incorrectly selected; IMSE: integrated mean squared error; MSE: mean squared error. The values in parentheses are the Monte Carlo standard error of the corresponding estimates.

available free of charge for research purposes at `http://www.measuredhs.com/countries/start.cfm`. Fenske, Kneib, and Hothorn (2009), and Koenker (2010) analyzed the whole data set by fitting nonparametric additive models. The two papers focused on nonparametric estimation instead of variable selection. Here we analyze a subset of the data including 606 children from New Delhi City between the ages of 0 and 5. Similar to Koenker (2010), we used child height as

the response variable, child age as the index variable, with 16 covariates. A brief description of the variables is as follows; more detailed information can be found in Koenker (2010).

| | |
|---|---|
| $Y$: Child's height (cm); | $U$: Child's age (months); |
| $X_1$: Breastfeeding (months); | $X_2$: Mother's Body Mass Index(BMI); |
| $X_3$: Mother's age (years); | $X_4$: Mother's education (years); |
| $X_5$: Father's education (years); | $X_6$: Child's sex, 1=Female, 0=Male; |
| $X_7$: 1=First child in the family, 0=not; | $X_8$: 1=Mothe is employed, 0=not; |
| $X_9$: 1=Religion not Hindu, 0=Hindu; | $X_{10}$: 1=Rich family, 0=not; |
| $X_{11}$: 1=Has radio, 0=not; | $X_{12}$: 1=Has television, 0=not; |
| $X_{13}$: 1=Has refrigerator, 0=not; | $X_{14}$: 1=Has bicycle, 0=not; |
| $X_{15}$: 1=Has motorcycle, 0=not; | $X_{16}$: 1=Has car, 0=not. |

We applied least squares regression and quantile regression to the data at two quantile levels $\tau = 0.1$ and 0.5, since we are not only interested in the typical nutrition in terms of the mean and median, but also in severe malnutrition. In median regression, our approach selects the model

$$Y = \alpha_0(u) + \sum_{k\in\{1,15\}} \alpha_k(u)X_k + \sum_{k\in\{3,4,5,6,7,14\}} \alpha_k X_k + \epsilon.$$

The quantile regression approach at $\tau = 0.1$ selects the model

$$Y = \alpha_0(u) + \sum_{k\in\{1,11,15\}} \alpha_k(u)X_k + \sum_{k\in\{3,4,8,14\}} \alpha_k X_k + \epsilon.$$

The least squares regression selects the model

$$Y = \alpha_0(u) + \sum_{k\in\{1,2,4,15\}} \alpha_k(u)X_k + \sum_{k\in\{3,5,7,10,14\}} \alpha_k X_k + \epsilon.$$

Of those selected, we see something in common: $X_1$ and $X_{15}$ have varying effects, and $X_3$ and $X_{14}$ have nonzero constant effects.

Table 4 summarizes the estimates of the nonzero constant coefficients, where the standard error in parenthese is obtained by using the bootstrap method with 200 bootstrap estimates obtained under the selected model. Figure 1 shows the estimated varying coefficients in the median regression, the 0.1th quantile regression, and the least squares regression. The shaded areas are the 90% pointwise confidence bands for the least squares estimates, obtained by using the bootstrap method based on 200 bootstrap samples with the bias being ignored.

From Table 4, we see that mother's age and education have positive effects on the child's height at both the lower quantile and the center. In contrast, father's

Table 4. Nonzero constant estimates and their standard errors (values in the parenthese) for the Childhood Malnutrition Data in India.

| Variable | $\alpha_{LSR}$ | $\alpha_{MR}$ | $\alpha_{0.1}$ |
|---|---|---|---|
| $X_1$ | $V$ | $V$ | $V$ |
| $X_2$ | $V$ | 0 (–) | 0 (–) |
| $X_3$ | 0.158 (0.060) | 0.231 (0.064) | 0.227 (0.091) |
| $X_4$ | $V$ | 0.139 (0.051) | 0.349 (0.074) |
| $X_5$ | 0.104 (0.055) | 0.142 (0.061) | 0 (–) |
| $X_6$ | 0 (–) | –0.856 (0.509) | 0 (–) |
| $X_7$ | –0.943 (0.471) | –1.302(0.553) | 0 (–) |
| $X_8$ | 0 (–) | 0 (–) | –2.306 (1.027) |
| $X_9$ | 0 (–) | 0 (–) | 0 (–) |
| $X_{10}$ | 0.876 (0.582) | 0 (–) | 0 (–) |
| $X_{11}$ | 0 (–) | 0 (–) | $V$ |
| $X_{12}$ | 0 (–) | 0 (–) | 0 (–) |
| $X_{13}$ | 0 (–) | 0 (–) | 0 (–) |
| $X_{14}$ | 1.326(0.463) | 1.377(0.531) | 1.630(0.853) |
| $X_{15}$ | $V$ | $V$ | $V$ |
| $X_{16}$ | 0 (–) | 0 (–) | 0 (–) |

Notation: $\alpha_{LSR}$: estimates from least squares regression; $\alpha_{MR}$: estimates from median regression; $\alpha_{0.1}$: estimates from the 0.1th quantile regression; $V$: varying effect.

education has positive effect only on the mean and the median of child's height. Girls are shorter than boys at the median and, on average, the first born child is taller than younger siblings. For the child with severe malnutrition, working mother has a negative effect on the response. The mean height is larger in families with higher household economic status. Bicycle ownership is associated with better health.

From Figure 1, we see that breastfeeding has a positive effect for new born children, and the effect decreases quickly as the children grow, reaching zero at around 10 months. However, this positive effect lasts longer for children with severe malnutrition; breastfeeding is the most important, maybe the only, nutrition for these children. The mother's BMI has positive effect as the child grows. The effect of radio is positive on the children with severe malnutrition only when they are young enough, but is negative later on. If we ignore the boundary region, the effect of the motorcycle is negative for children aged 2 to 20 months, but positive for children aged 25 to 45 months.

## Acknowledgement

Dr. Wang's Research is supported by NSF grant DMS-0706963, Dr. Zhu's Research is supported by NFC grants 10931002 and 1091112038, and Dr. Song's

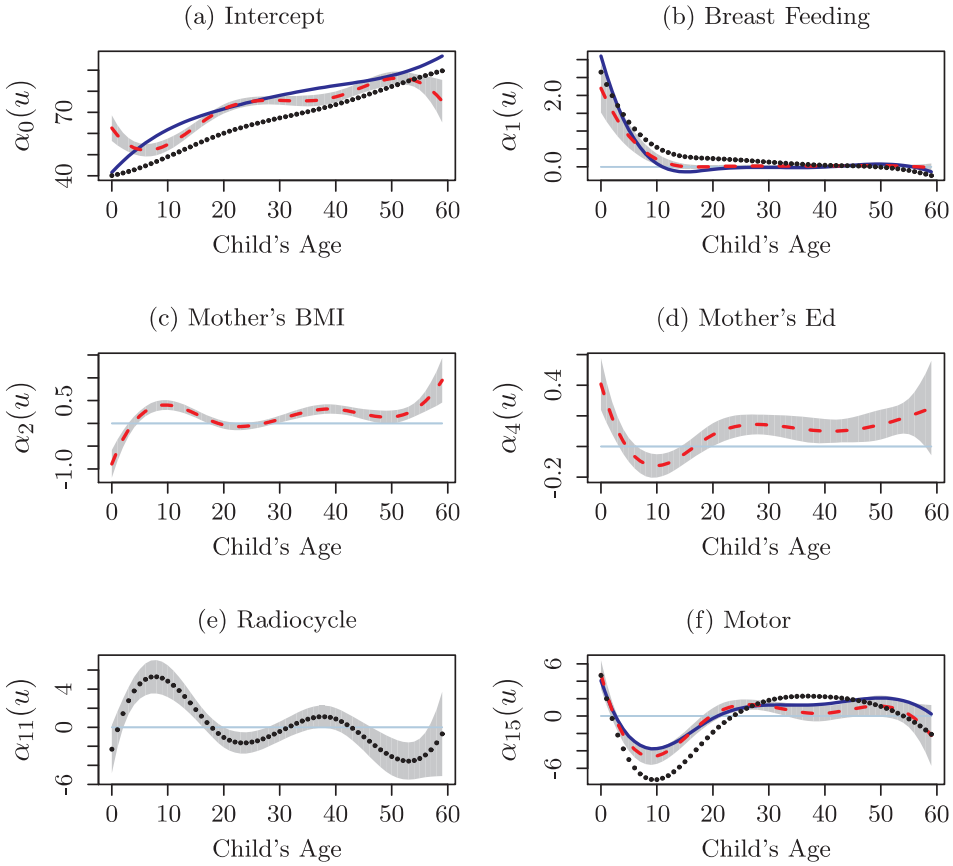


Figure 1. The estimated varying coefficients for two quantiles, $\tau$=0.1($\cdots$), $\tau$=0.5(—) and the least squares regression (- -). The shaded area indicates the 90% pointwise bootstrap confidence band from the least squares regression regression, except that panel (e) is for $\tau = 0.1$.

Research is supported by grant GRF 403109 from Hong Kong Special Administration Region. We would like to thank the Editor, an associate editor and three anonymous reviewers for their constructive comments that led to a major improvement of this article.

## Appendix

### A.1. Some useful lemmas

We first provide some useful lemmas that facilitate the proofs of Theorems 1 and 2. The method used in the proof of Lemma A.1 is similar to that used in Theorem 4 and Proposition 4 of Chen (1991). Lemma A.2 follows directly from Corollary 6.21 of Schumaker (1981, Chap. 6). Lemma A.3 is a special case

of Lemma A.3 in Wang, Li, and Huang (2008). Lemma A.4 is a special case of Lemma A.7 in Huang, Wu, and Zhou (2004). Lemma A.5 follows from Theorem 2 of He and Shi (1994) and Theorem 2 of Huang, Wu, and Zhou (2002). We omit the proofs of these lemmas.

**Lemma A.1.** Suppose A1-A3 hold and $k_n \sim n^{1/(2r+1)}$, the eigenvalues of $n^{-1}k_n\boldsymbol{V}_n$ are uniformly bounded away from zero and infinity in probability, where $\boldsymbol{V}_n = (\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_n)(\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_n)^T$.

**Lemma A.2.** Suppose A1−A3 hold, then there exists a spline coefficient vector $\boldsymbol{\gamma}^0 = (\boldsymbol{\gamma}_0^{0T}, \boldsymbol{\gamma}_1^{0T}, \ldots, \boldsymbol{\gamma}_p^{0T})^T$, and some positive constants $a_1$, $a_2$, and $\epsilon_1$, such that

$$\begin{aligned}
&\text{(i)}\ \ \|\boldsymbol{\gamma}_{k*}^0\|_{L_1} > \epsilon_1 \text{ if } k \in \{1, \ldots, \nu\},\ \gamma_{k,1}^0 = \alpha_k, \boldsymbol{\gamma}_{k*}^0 = \mathbf{0}_{k_n+\hbar} \text{ if } k \in \{\nu+1, \ldots, s\},\\
&\qquad \text{and}\ \ \boldsymbol{\gamma}_k^0 = \mathbf{0}_{q_n} \ \text{ if } \ k \in \{s+1, \ldots, p\}; \qquad \text{(A.1)}\\
&\text{(ii)}\ \sup_{(u,\boldsymbol{X})\in[0,1]\times\mathbb{R}^{p+1}} |\boldsymbol{\Pi}(u, \boldsymbol{X})^T\boldsymbol{\gamma}^0 - \boldsymbol{x}^T\alpha(u)| \le a_1 k_n^{-r};\\
&\text{(iii)}\ \sup_{u\in[0,1]} \left|\alpha_k(u) - \pi(u)^T\boldsymbol{\gamma}_k^0\right| \le a_2 k_n^{-r}, k = 0, 1, \ldots, \nu,
\end{aligned}$$

where $\boldsymbol{\gamma}_k^0 = (\gamma_{k,1}^0, \boldsymbol{\gamma}_{k*}^{0T})^T$, $k = 0, 1, \ldots, p$.

For simplicity, we adopt the notation

$$\begin{aligned}
Y &= (Y_1, \ldots, Y_n)^T,\ \tilde{\boldsymbol{\Pi}} = (\boldsymbol{\Pi}_1, \ldots, \boldsymbol{\Pi}_n)^T,\ \boldsymbol{e} = (e_1, \ldots, e_n)^T,\\
\boldsymbol{\Pi}_{(1)i} &= \left(X_i^{(0)}\boldsymbol{\pi}_i^T, \ldots, X_i^{(\nu)}\boldsymbol{\pi}_i^T\right)^T,\ \boldsymbol{\Pi}_{(2)i} = \left(X_i^{(\nu+1)}, \ldots, X_i^{(p)}\right)^T,\\
\boldsymbol{\Pi}_{(c)i} &= \left(X_i^{(\nu+1)}, \ldots, X_i^{(s)}\right)^T,\\
\boldsymbol{\gamma}_{(1)} &= (\boldsymbol{\gamma}_0^T, \ldots, \boldsymbol{\gamma}_\nu^T)^T,\ \boldsymbol{\gamma}_{(2)} = (\gamma_{\nu+1,1}, \ldots, \gamma_{p,1})^T,\ \boldsymbol{\gamma}_{(c)} = (\gamma_{\nu+1,1}, \ldots, \gamma_{s,1})^T,\\
\boldsymbol{\gamma}_{(1)}^0 &= (\boldsymbol{\gamma}_0^{0T}, \ldots, \boldsymbol{\gamma}_\nu^{0T})^T,\ \boldsymbol{\gamma}_{(2)}^0 = (\gamma_{\nu+1,1}^0, \ldots, \gamma_{p,1}^0)^T = (\boldsymbol{\alpha}_c^T, 0_{s-\nu}^T)^T,\\
\boldsymbol{\gamma}_{(c)}^0 &= \boldsymbol{\alpha}_c = (\alpha_{\nu+1}, \ldots, \alpha_s)^T,\\
D_{ni} &= \boldsymbol{\Pi}_i^T\boldsymbol{\gamma}^0 - \boldsymbol{X}_i^T\boldsymbol{\alpha}(U_i),\quad D_n = (D_{n1}, \ldots, D_{nn})^T,\quad d_{nik} = \pi_i^T\boldsymbol{\gamma}_k^0 - \alpha_k(U_i).
\end{aligned}$$

By $(ii)$ and $(iii)$ of Lemma A.2, it is easy to see that $\max_i |D_{ni}| \le a_1 k_n^{-r}$, $\|D_n\|_{L_2} = O_p(n^{1/2}k_n^{-r})$, $\max_{i,k} |d_{nik}| \le a_2 k_n^{-r}$, and $d_{nik} = 0$ for $k > \nu$. By Lemma A.5 of Kim (2007), $\max_{i=1}^n \left\|\boldsymbol{\Pi}_{(1)i}\right\|_{L_2} = O_p(1)$.

**Lemma A.3.** Suppose A1−A3 hold and $k_n \sim n^{1/(2r+1)}$, then we have

$$n^{-1/2} \sup_{\boldsymbol{v}\in R^{(p+1)q_n}} \frac{\left|\boldsymbol{v}^T\tilde{\boldsymbol{\Pi}}^T\boldsymbol{e}\right|}{\|\boldsymbol{v}\|_{L_2}} = O_p(1).$$

**Lemma A.4.** Suppose A1−A3 hold and $k_n \sim n^{1/(2r+1)}$. Let $\bar{Y}_i = E(Y_i|\boldsymbol{X}_i, U_i)$, $\bar{Y} = (\bar{Y}_1, \ldots, \bar{Y}_n)^T$, and $\bar{\boldsymbol{\gamma}} = (\tilde{\boldsymbol{\Pi}}^T\tilde{\boldsymbol{\Pi}})^{-1}\tilde{\boldsymbol{\Pi}}^T\bar{Y}$. Then $\|\bar{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^0\|_{L_2} = O_p(n^{-1/2}k_n)$.

**Lemma A.5.** Let $\tilde{\boldsymbol{\gamma}}$ be the minimizer of (2.3) with $\lambda_{1,n} = 0$. Then if A1−A3 hold and $k_n \sim n^{1/(2r+1)}$, we have, for both quantile and least squares regressions, $\|\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^0\|_{L_2} = O_p(n^{-1/2}k_n)$.

Let $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_0^T, \ldots, \boldsymbol{\gamma}_p^T)^T$ be any given $(p+1)q_n$-vector, with $\boldsymbol{\gamma}_k = (\gamma_{k,1}, \boldsymbol{\gamma}_{k*}^T)^T$ and $\boldsymbol{\gamma}_{k*} = (\gamma_{k,2}, \ldots, \gamma_{k,q_n})^T$.

**Lemma A.6.** Under the conditions of Theorems 1 and 2 we have, for both quantile and least squares regressions, $\|\hat{\boldsymbol{\gamma}}^{VC} - \boldsymbol{\gamma}^0\|_{L_2} = O_p(n^{-1/2}k_n)$, where $\hat{\boldsymbol{\gamma}}^{VC}$ is the minimizer of (2.3).

**Proof.** Combining (A.1) and Lemma A.5, there exist some positive constants $b_k$ such that $\|\tilde{\boldsymbol{\gamma}}_{k*}\|_{L_1} > b_k$, and therefore $\tilde{\omega}_{k*} = \|\tilde{\boldsymbol{\gamma}}_{k*}\|_{L_1}^{-1} < b_k^{-1}$, hold in probability for $k = 1, \ldots, \nu$. By routine calculation, $\|\boldsymbol{\gamma}_{k*} - \boldsymbol{\gamma}_{k*}^0\|_{L_1} \le q_n^{1/2}\|\boldsymbol{\gamma} - \boldsymbol{\gamma}^0\|_{L_2}$. Assume that $\|\boldsymbol{\gamma} - \boldsymbol{\gamma}^0\|_{L_2} = C_1 n^{-1/2}k_n$ and $C_1$ is large enough. Then combining (A.1) and A4,

$$\begin{aligned} & n\lambda_{1,n}\sum_{k=1}^{p}\tilde{\omega}_{k*}\left(\|\boldsymbol{\gamma}_{k*}\|_{L_1} - \|\boldsymbol{\gamma}_{k*}^0\|_{L_1}\right) \\ & \ge -n\lambda_{1,n}\sum_{k=1}^{\nu}\tilde{\omega}_{k*}\left(\|\boldsymbol{\gamma}_{k*} - \boldsymbol{\gamma}_{k*}^0\|_{L_1}\right) \\ & = O_p\Big(n\lambda_{1,n}\sum_{k=1}^{\nu}b_k^{-1}q_n^{1/2}C_1 n^{-1/2}k_n\Big) = o_p(k_n). \end{aligned} \tag{A.2}$$

Let $B_n(\boldsymbol{\gamma}) \doteq \sum_{i=1}^n \left[g(Y_i - \boldsymbol{\Pi}_i^T\boldsymbol{\gamma}) - g(Y_i - \boldsymbol{\Pi}_i^T\boldsymbol{\gamma}^0)\right] = \sum_{i=1}^n \Big[g(e_i - \boldsymbol{\Pi}_i^T(\boldsymbol{\gamma} - \boldsymbol{\gamma}^0) - D_{ni}) - g(e_i - D_{ni})\Big]$. For quantile regression, by Lemma 3.2 and the arguments used for proving Lemma 3.3 in He and Shi (1994),

$$P\Big(\inf_{\|\boldsymbol{\gamma}-\boldsymbol{\gamma}^0\|_{L_2}=C_1n^{-1/2}k_n} B_n(\boldsymbol{\gamma}) > k_n\Big) \longrightarrow 1. \tag{A.3}$$

For least squares regression, by decomposing $\boldsymbol{\gamma}$ as $(\boldsymbol{\gamma} - \boldsymbol{\gamma}^0) + (\boldsymbol{\gamma}^0 - \bar{\boldsymbol{\gamma}}) + \bar{\boldsymbol{\gamma}}$, and facilitated by Lemmas A.1, A.3, and A.4, we can prove that there exists a positive constant $C_2$ such that

$$P\Big(\inf_{\|\boldsymbol{\gamma}-\boldsymbol{\gamma}^0\|_{L_2}=C_1n^{-1/2}k_n} B_n(\boldsymbol{\gamma}) > C_2C_1^2k_n\Big) \longrightarrow 1. \tag{A.4}$$

By the definition of $B_n(\cdot)$,

$$l_1(\boldsymbol{\gamma}) - l_1(\boldsymbol{\gamma}^0) = B_n(\boldsymbol{\gamma}) + n\lambda_{1,n} \sum_{k=1}^{p} \tilde{\omega}_{k*} \left( \|\boldsymbol{\gamma}_{k*}\|_{L_1} - \|\boldsymbol{\gamma}_{k*}^0\|_{L_1} \right). \tag{A.5}$$

Equations (A.2), (A.3), (A.4), and (A.5) together suggest that $l_1(\boldsymbol{\gamma}) - l_1(\boldsymbol{\gamma}^0) > 0$ in probability for both quantile and least squares regressions. By the convexity of $l_1(\boldsymbol{\gamma})$ and the fact that $l_1(\hat{\boldsymbol{\gamma}}^{VC}) - l_1(\boldsymbol{\gamma}^0) \leq 0$, there exists some $C_{\varsigma_1}$, for any $\varsigma_1 > 0$, such that as $n \to \infty$,

$$P\left( \|\hat{\boldsymbol{\gamma}}^{VC} - \boldsymbol{\gamma}^0\|_{L_2} \leq C_{\varsigma_1} n^{-1/2} k_n \right) > 1 - \varsigma_1.$$

Therefore, Lemma A.6 is proved.

In the following, we prove Theorem 1 and Theorem 2 with $\hat{\boldsymbol{\alpha}}(u)$ estimated from $\hat{\boldsymbol{\gamma}}^{CZ}$, the estimator obtained after Step 2 of the first iteration, as $\hat{\boldsymbol{\gamma}}^{CZ}$ and the estimator at convergence $\hat{\boldsymbol{\gamma}}^{final}$ have the same asymptotic properties. For ease of representation, we omit the statement "with probability approaching one" whenever it is clear. For example, the statement $\hat{\gamma}_{k,l} = 0$ means that $\hat{\gamma}_{k,l} = 0$ with probability approaching one. We give the proof for Theorem 2. The proof of Theorem 1 is similar and thus is not given in detail.

### A.2. The Proof of Theorem 2 in four steps

**(I) Proof of the first part of Theorem 2(a)**: as $n \to \infty$, with probability approaching 1, $\hat{\boldsymbol{\gamma}}_{k*}^{VC} = 0, k = \nu + 1, \ldots, p$, therefore $\hat{\alpha}_k(u)$ is constant. We prove this using the Karush-Kuhn-Tucker conditions (Yuan and Lin (2006), Huang, Horowitz, and Wei (2010)).

If $M(\boldsymbol{\gamma})$ is the derivative of $l_1(\boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$,

$$M(\boldsymbol{\gamma}) = \frac{\partial l_1(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} = \sum_{i=1}^{n} -\psi_\tau(Y_i - \boldsymbol{\Pi}_i^T \boldsymbol{\gamma}) \boldsymbol{\Pi}_i + n\lambda_{1,n}(K_{\boldsymbol{\gamma}_0}^T, K_{\boldsymbol{\gamma}_1}^T, \ldots, K_{\boldsymbol{\gamma}_p}^T)^T,$$

where $\psi_\tau(t) = \tau - I(t < 0)$, $K_{\boldsymbol{\gamma}_0} = \mathbf{0}_{q_n}$, and

$$K_{\boldsymbol{\gamma}_k} = \tilde{\omega}_{k*} \frac{\partial \|\boldsymbol{\gamma}_{k*}\|_{L_1}}{\partial \boldsymbol{\gamma}_k} = \tilde{\omega}_{k*} \left(0, \operatorname{sign}(\gamma_{k,2}), \ldots, \operatorname{sign}(\gamma_{k,q_n})\right)^T, \quad k = 1, \ldots, p.$$

According to the Karush-Kuhn-Tucker conditions, a necessary and sufficient condition for $(\hat{\boldsymbol{\gamma}}_0^{VC^T}, \hat{\boldsymbol{\gamma}}_1^{VC^T}, \ldots, \hat{\boldsymbol{\gamma}}_p^{VC^T})^T$ to be the minimizer of $l_1(\boldsymbol{\gamma})$ is that

$$-\sum_{i=1}^{n} \psi_\tau(Y_i - \boldsymbol{\Pi}_i^T \hat{\boldsymbol{\gamma}}^{VC}) X_i^{(k)} \pi_i + n\lambda_{1,n} K_{\hat{\boldsymbol{\gamma}}_k^{VC}} = 0 \quad \text{for any } \|\hat{\boldsymbol{\gamma}}_{k*}^{VC}\|_{L_2} \neq 0, \quad k \geq 0,$$

$$\left\| \sum_{i=1}^{n} \psi_\tau(Y_i - \boldsymbol{\Pi}_i^T \hat{\boldsymbol{\gamma}}^{VC}) X_i^{(k)} \pi_i \right\|_{L_2} \leq n\lambda_{1,n} \tilde{\omega}_{k*} \quad \text{for any } \|\hat{\boldsymbol{\gamma}}_{k*}^{VC}\|_{L_2} = 0, \quad k \geq 1.$$

To prove the first part of Theorem 2(a), we only need prove that for $k = \nu+1,\ldots,p$, $\left\|\sum_{i=1}^n \psi_\tau(Y_i - \mathbf{\Pi}_i^T\hat{\boldsymbol{\gamma}}^{VC})X_i^{(k)}\pi_i\right\|_{L_2} \le n\lambda_{1,n}\tilde{\omega}_{k*}$.

Let $S_n^0 = \sum_{i=1}^n \psi_\tau(e_i)\mathbf{\Pi}_i$ and $S_n(\boldsymbol{\gamma}) = \sum_{i=1}^n \psi_\tau(Y_i - \mathbf{\Pi}_i^T\boldsymbol{\gamma})\mathbf{\Pi}_i$. By Lemma 1, we can easily establish that $\|S_n^0\|_{L_2} = O_p(n^{1/2}k_n^{-1/2})$. Since $\hat{\boldsymbol{\gamma}}^{VC} - \boldsymbol{\gamma}^0 = O_p(n^{-1/2}k_n)$, similar to the lines used in the proof of Lemmas 8.4, 8.5, and Theorem 4.1 in Wei and He (2006), with Lemma A.1 we can show that $\|S_n(\hat{\boldsymbol{\gamma}}^{VC}) - S_n(\boldsymbol{\gamma}^0)\|_{L_2} = o_p(n^{1/2})$.

Note that $S_n(\boldsymbol{\gamma}^0) = \sum_{i=1}^n \psi_\tau(e_i - D_{ni})\mathbf{\Pi}_i$ and $\max_i |D_{ni}| \le a_1 k_n^{-r}$. We want to point out that

$$\begin{aligned}&\left\|S_n(\boldsymbol{\gamma}^0) - S_n^0\right\|_{L_2}\\ &\quad= O_p\left(\left\|E\left[S_n(\boldsymbol{\gamma}^0) - S_n^0\right]\right\|_{L_2} + \left\{E\left[S_n(\boldsymbol{\gamma}^0) - S_n^0\right]^T\left[S_n(\boldsymbol{\gamma}^0) - S_n^0\right]\right\}^{1/2}\right).\end{aligned}$$

Let $H_n = \sum_{i=1}^n \mathbf{\Pi}_i\mathbf{\Pi}_i^T$, then by Lemma A.1, the maximum eigenvalue of $H_n$, $\lambda_{\max}(H_n) = O_p(n/k_n)$. By the Cauchy-Schwarz inequality,

$$\begin{aligned}\left\|E\left[S_n(\boldsymbol{\gamma}^0) - S_n^0\right]\right\|_{L_2} &= \left\|\sum_{i=1}^n \mathbf{\Pi}_i E\left[\psi_\tau(e_i - D_{ni}) - \psi_\tau(e_i)\right]\right\|_{L_2}\\ &= O_p\Big(\Big\|\sum_{i=1}^n \mathbf{\Pi}_i f_i(0) D_{ni}\Big\|_{L_2}\Big)\\ &\le O_p\left(\lambda_{\max}^{1/2}(H_n)(nk_n^{-2r})^{1/2}\right) = O_p(n^{1/2}). \qquad \text{(A.6)}\end{aligned}$$

By the independence of $e_i$ and Lemma A.1, there exists a large enough positive constant $C_3$ such that

$$\begin{aligned}&E\left[S_n(\boldsymbol{\gamma}^0) - S_n^0\right]^T\left[S_n(\boldsymbol{\gamma}^0) - S_n^0\right]\\ &\le \sum_{i=1}^n \mathbf{\Pi}_i^T\mathbf{\Pi}_i E\left[\psi_\tau(e_i) - \psi_\tau(e_i - D_{ni})\right]^2\\ &\quad+ \sum_{i_1=1}^n\sum_{i_2\ne i_1} \mathbf{\Pi}_{i_1}^T\mathbf{\Pi}_{i_2} E\left[\psi_\tau(e_{i_1}) - \psi_\tau(e_{i_1} - D_{ni_1})\right] E\left[\psi_\tau(e_{i_2}) - \psi_\tau(e_{i_2} - D_{ni_2})\right]\\ &\le C_2\Big[(n/k_n)a_1k_n^{-r} + n\lambda_{\max}(H_n)k_n^{-2r}\Big] = O_p(n). \qquad \text{(A.7)}\end{aligned}$$

Combining (A.6) and (A.7), we have $\left\|S_n(\boldsymbol{\gamma}^0) - S_n^0\right\|_{L_2} = O_p(n^{1/2})$. Therefore,

$$\|S_n(\hat{\boldsymbol{\gamma}}^{VC})\|_{L_2} \le \|S_n(\hat{\boldsymbol{\gamma}}^{VC}) - S_n(\boldsymbol{\gamma}^0)\|_{L_2} + \|S_n(\boldsymbol{\gamma}^0) - S_n^0\|_{L_2} + \|S_n^0\|_{L_2} = O_p(n^{1/2}).$$

For $k = v+1, \ldots, p$, it follows by Lemma A.5 that $\|\tilde{\boldsymbol{\gamma}}_k\|_{L_1} = O_p(k_n^{1/2} n^{-1/2} k_n)$, which indicates that $\tilde{\omega}_{k*} = \|\tilde{\boldsymbol{\gamma}}_k\|_{L_1}^{-1} \geq C_4 n^{1/2} k_n^{-3/2}$ for some positive constant $C_4$. By assumption A4,

$$\begin{aligned} n\lambda_{1,n}\tilde{\omega}_{k*} &\geq (C_4 n\lambda_{1,n} k_n^{-3/2}) n^{1/2} \geq \|S_n(\hat{\boldsymbol{\gamma}}^{VC})\|_{L_2} \\ &\geq \Big\| \sum_{i=1}^{n} \psi_\tau(Y_i - \boldsymbol{\Pi}_i^T \hat{\boldsymbol{\gamma}}^{VC}) X_i^{(k)} \pi_i \Big\|_{L_2}. \end{aligned} \tag{A.8}$$

**(II) Proof of Theorem 2(b)**: the optimal convergence rate of the varying coefficient estimates.

In Lemma A.6 and (I), we have proved that, when $n \to \infty$, with probability approaching 1, $\|\hat{\boldsymbol{\gamma}}_{k*}^{VC}\|_{L_1} > 0, k \leq \nu$ and $\|\hat{\boldsymbol{\gamma}}_{k*}^{VC}\|_{L_1} = 0, k > \nu$. Recall the definitions of $\boldsymbol{\gamma}_{(1)}$ and $\boldsymbol{\gamma}_{(2)}$, and $\boldsymbol{\gamma}^1$ in the iteration estimation procedure. For the quantile regression, with probability approaching 1, we have

$$l_2(\boldsymbol{\gamma}^1) = \sum_{i=1}^{n} \rho_\tau \left( Y_i - \boldsymbol{\Pi}_{(1)i}^T \boldsymbol{\gamma}_{(1)} - \boldsymbol{\Pi}_{(2)i}^T \boldsymbol{\gamma}_{(2)} \right) + n\lambda_{2,n} \sum_{k=\nu+1}^{p} \hat{\omega}_{k,1} |\gamma_{k,1}| \doteq l_2\left(\boldsymbol{\gamma}_{(1)}, \boldsymbol{\gamma}_{(2)}\right).$$

For $k = \nu+1, \ldots, s$, by (A.1) and Lemma A.6, there exists some positive constant $\tilde{b}$ such that $|\hat{\gamma}_{k,1}^{VC}| > \tilde{b}$; therefore $\hat{\omega}_{k,1} = |\hat{\gamma}_{k,1}^{VC}|^{-1} < \tilde{b}^{-1}$ holds in probability. Similar to the proof of Lemma A.6, there exists some $C_{\varsigma_2}$, for any $\varsigma_2 > 0$, such that as $n \to \infty$,

$$P\left( \left\| \hat{\boldsymbol{\gamma}}_{(1)} - \boldsymbol{\gamma}_{(1)}^0 \right\|_{L_2} \leq C_{\varsigma_2} n^{-1/2} k_n, \left\| \hat{\boldsymbol{\gamma}}_{(2)} - \boldsymbol{\gamma}_{(2)}^0 \right\|_{L_2} \leq C_{\varsigma_2} n^{-1/2} k_n^{1/2} \right) > 1 - \varsigma_2, \tag{A.9}$$

where $\left(\hat{\boldsymbol{\gamma}}_{(1)}^T, \hat{\boldsymbol{\gamma}}_{(2)}^T\right)^T$ is the minimizer of $l_2\left(\boldsymbol{\gamma}_{(1)}, \boldsymbol{\gamma}_{(2)}\right)$. By the definition of $\boldsymbol{\gamma}^1$, (A.9) indicates that

$$P\left( \left\| \hat{\boldsymbol{\gamma}}_{(1)}^{CZ} - \boldsymbol{\gamma}_{(1)}^0 \right\|_{L_2} \leq C_{\varsigma_2} n^{-1/2} k_n \right) > 1 - \varsigma_2, \tag{A.10}$$

where $\hat{\boldsymbol{\gamma}}_{(1)}^{CZ}$ is the corresponding sub-vector of $\hat{\boldsymbol{\gamma}}^{CZ}$. It is easy to see that, for $k = 0, 1, \ldots, \nu$,

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} [\hat{\alpha}_k(U_i) - \alpha_k(U_i)]^2 &\leq \frac{2}{n} \sum_{i=1}^{n} \left[ \boldsymbol{\pi}_i^T (\hat{\boldsymbol{\gamma}}_k^{CZ} - \boldsymbol{\gamma}_k^0) \right]^2 + \frac{2}{n} \sum_{i=1}^{n} d_{nik}^2 \\ &\leq \frac{2}{n} (\hat{\boldsymbol{\gamma}}_k^{CZ} - \boldsymbol{\gamma}_k^0)^T \boldsymbol{V}_n (\hat{\boldsymbol{\gamma}}_k^{CZ} - \boldsymbol{\gamma}_k^0) + 2a_2^2 k_n^{-2r}. \end{aligned}$$

As $k_n \sim n^{1/(2r+1)}$, by Lemma A.1 and (A.10) we have proved Theorem 2(b).

**(III) Proof of the second part of Theorem 2(a)**: the sparsity of the penalized estimator.

In (I), we proved the constancy of $\hat{\alpha}_k(u)$ for $k = \nu+1, \ldots, p$. To prove $\hat{\alpha}_k(u) \equiv 0, k = s+1, \ldots, p$, we only need to prove that $\hat{\gamma}_{k,1}^{CZ} = 0$. Similar to (I), by the Karush-Kuhn-Tucker conditions, it suffices to prove that $n\lambda_{2,n}\hat{\omega}_{k,1} \geq \left\| \sum_{i=1}^n X_i^{(k)} \psi_\tau \left( Y_i - \mathbf{\Pi}_{(1)i}^T \hat{\boldsymbol{\gamma}}_{(1)} - \mathbf{\Pi}_{(2)i}^T \hat{\boldsymbol{\gamma}}_{(2)} \right) \right\|_{L_2}$.

Following similar lines as in (I), we obtain that

$$\left\| \sum_{i=1}^n X_i^{(k)} \psi_\tau \left( Y_i - \mathbf{\Pi}_{(1)i}^T \hat{\boldsymbol{\gamma}}_{(1)} - \mathbf{\Pi}_{(2)i}^T \hat{\boldsymbol{\gamma}}_{(2)} \right) \right\|_{L_2} = O_p(n^{1/2} k_n^{1/2}).$$

Combining (A.1) and Lemma A.6, $\hat{\gamma}_{k,1}^{VC} = O_p(n^{-1/2} k_n^{1/2})$, and then there exists some positive constant $C_5$ such that $\hat{\omega}_{k,1} \geq C_5 n^{1/2} k_n^{-1/2}$. The result follows from assumption A4.

**(IV) Proof of Theorem 2(c)**: the asymptotic normality of the constant coefficient estimates. For simplicity, we omit $CZ$ from $\hat{\boldsymbol{\gamma}}^{CZ}$ and its sub-vector.

Let

$$\begin{aligned}
\breve{\mathbf{\Pi}}_{(1)} &= \left( \mathbf{\Pi}_{(1)1}, \ldots, \mathbf{\Pi}_{(1)n} \right)^T, \ \breve{\mathbf{\Pi}}_{(c)} = \left( \mathbf{\Pi}_{(c)1}, \ldots, \mathbf{\Pi}_{(c)n} \right)^T, \\
\breve{B} &= diag\Big( f_1(0), \ldots, f_n(0) \Big), \\
\breve{P} &= \breve{\mathbf{\Pi}}_{(1)} \left( \breve{\mathbf{\Pi}}_{(1)}^T \breve{B} \breve{\mathbf{\Pi}}_{(1)} \right)^{-1} \breve{\mathbf{\Pi}}_{(1)}^T \breve{B}, \ \breve{\mathbf{\Pi}}_{(c)*} = (I - \breve{P}) \breve{\mathbf{\Pi}}_{(c)}, \\
K_{n*} &= \breve{\mathbf{\Pi}}_{(c)*}^T \breve{B} \breve{\mathbf{\Pi}}_{(c)*}, \quad \Lambda_{n*} = \tau(1-\tau) \breve{\mathbf{\Pi}}_{(c)*}^T \breve{\mathbf{\Pi}}_{(c)*}.
\end{aligned} \tag{A.11}$$

By (I) and (III), as $n \to \infty$, with probability approaching 1, $\hat{\boldsymbol{\gamma}}_{k*} = 0, k = \nu+1, \ldots, s$ and $\hat{\boldsymbol{\gamma}}_k = 0, k = s+1, \ldots, p$. Therefore, with probability approaching 1, we have

$$\begin{aligned}
l_2 \left( \boldsymbol{\gamma}_{(1)}, \boldsymbol{\gamma}_{(2)} \right) = & \sum_{i=1}^n \rho_\tau \left( Y_i - \mathbf{\Pi}_{(1)i}^T \boldsymbol{\gamma}_{(1)} - \mathbf{\Pi}_{(c)i}^T \boldsymbol{\gamma}_{(c)} \right) \\
& + n\lambda_{2,n} \sum_{k=\nu+1}^s \hat{\omega}_{k,1} |\gamma_{k,1}| \doteq l_2 \left( \boldsymbol{\gamma}_{(1)}, \boldsymbol{\gamma}_{(c)} \right).
\end{aligned}$$

Let

$$\begin{aligned}
\boldsymbol{\zeta} \left( \boldsymbol{\gamma}_{(1)}, \boldsymbol{\gamma}_{(c)} \right) &= \begin{pmatrix} \boldsymbol{\zeta}_1 \\ \boldsymbol{\zeta}_2 \end{pmatrix} \\
&= \begin{pmatrix} \Lambda_{n*}^{-1/2} K_{n*} \left( \boldsymbol{\gamma}_{(c)} - \boldsymbol{\gamma}_{(c)}^0 \right) \\ k_n^{-1/2} \breve{H}_n \left( \boldsymbol{\gamma}_{(1)} - \boldsymbol{\gamma}_{(1)}^0 \right) + k_n^{1/2} \breve{H}_n^{-1} \breve{\mathbf{\Pi}}_{(1)}^T \breve{B} \breve{\mathbf{\Pi}}_{(c)} \left( \boldsymbol{\gamma}_{(c)} - \boldsymbol{\gamma}_{(c)}^0 \right) \end{pmatrix},
\end{aligned}$$

$\hat{\boldsymbol{\zeta}} = \boldsymbol{\zeta}\left(\hat{\boldsymbol{\gamma}}_{(1)}, \hat{\boldsymbol{\gamma}}_{(c)}\right) = (\hat{\boldsymbol{\zeta}}_1^T, \hat{\boldsymbol{\zeta}}_2^T)^T$, where $\breve{H}_n^2 = k_n \breve{\boldsymbol{\Pi}}_{(1)}^T \breve{B} \breve{\boldsymbol{\Pi}}_{(1)}$, and $\left(\hat{\boldsymbol{\gamma}}_{(1)}^T, \hat{\boldsymbol{\gamma}}_{(c)}^T\right)^T$ is the minimizer of $l_2\left(\boldsymbol{\gamma}_{(1)}, \boldsymbol{\gamma}_{(c)}\right)$. From (A.9), $\|\hat{\boldsymbol{\gamma}}_{(1)} - \boldsymbol{\gamma}_{(1)}^0\|_{L_2} = O_p(n^{-1/2}k_n)$, $\|\hat{\boldsymbol{\gamma}}_{(c)} - \boldsymbol{\gamma}_{(c)}^0\|_{L_2} = O_p(n^{-1/2}k_n^{1/2})$; therefore $\|\hat{\boldsymbol{\zeta}}\|_{L_2} = O_p(k_n^{1/2})$.

Let $\hat{\boldsymbol{\zeta}}_1^* = \Lambda_{n*}^{-1/2} \breve{\boldsymbol{\Pi}}_{(c)*}^T \psi_\tau(\boldsymbol{e})$, where $\psi_\tau(e_i) = \tau - I(e_i < 0)$ and $\psi_\tau(\boldsymbol{e}) = \{\psi_\tau(e_1), \ldots, \psi_\tau(e_n)\}^T$. By Lemmas 1 and 2 in Wang, Zhu, and Zhou (2009), $\hat{\boldsymbol{\zeta}}_1^*$ is asymptotically normal distributed with variance-covariance matrix $I_{s-\nu}$. Therefore, to prove part (c) of Theorem 2, we only need to prove $\|\hat{\boldsymbol{\zeta}}_1 - \hat{\boldsymbol{\zeta}}_1^*\|_{L_2} = o_p(1)$.

By the definition of $\hat{\boldsymbol{\zeta}}_1$ and $\hat{\boldsymbol{\zeta}}_1^*$, there exist two positive constants $C_6$ and $C_7$ such that $P\left(\|\hat{\boldsymbol{\zeta}}_1^*\|_{L_2} < C_6\right) \to 1$ and $P\left(\|\hat{\boldsymbol{\zeta}}_1\|_{L_2} < C_7 k_n^{1/2}\right) \to 1$. Let

$$\begin{aligned} U_i(\boldsymbol{\zeta}_1, \hat{\boldsymbol{\zeta}}_1^*) = {} & \rho_\tau\left(e_i - \boldsymbol{\zeta}_1^T \tilde{\boldsymbol{\Pi}}_{(c)i} - \hat{\boldsymbol{\zeta}}_2^T \tilde{\boldsymbol{\Pi}}_{(1)i} - D_{ni}\right) \\ & - \rho_\tau\left(e_i - \hat{\boldsymbol{\zeta}}_1^{*T} \tilde{\boldsymbol{\Pi}}_{(c)i} - \hat{\boldsymbol{\zeta}}_2^T \tilde{\boldsymbol{\Pi}}_{(1)i} - D_{ni}\right), \end{aligned}$$

where $\tilde{\boldsymbol{\Pi}}_{(1)i} = k_n^{1/2} H_n^{-1} \boldsymbol{\Pi}_{(1)i}$, $\tilde{\boldsymbol{\Pi}}_{(c)i} = \Lambda_{n*}^{1/2} K_{n*}^{-1} \breve{\boldsymbol{\Pi}}_{(c)*i}$. By Lemmas 8.1 and 8.3 of Wei and He (2006) and the orthogonality of $\breve{\boldsymbol{\Pi}}_{(1)}$ and $\breve{\boldsymbol{\Pi}}_{(c)*}$, for any given $\eta > 0$,

$$\sup_{\|\boldsymbol{\zeta}_1 - \hat{\boldsymbol{\zeta}}_1^*\|_{L_2} < \eta} \left| \sum_{i=1}^n \left[ U_i(\boldsymbol{\zeta}_1, \hat{\boldsymbol{\zeta}}_1^*) + (\boldsymbol{\zeta}_1 - \hat{\boldsymbol{\zeta}}_1^*)^T \tilde{\boldsymbol{\Pi}}_{(c)i} \psi_\tau(e_i) - E U_i(\boldsymbol{\zeta}_1, \hat{\boldsymbol{\zeta}}_1^*) \right] \right| = o_p(1),$$

$$\sup_{\|\boldsymbol{\zeta}_1 - \hat{\boldsymbol{\zeta}}_1^*\|_{L_2} < \eta} \left| \sum_{i=1}^n E U_i(\boldsymbol{\zeta}_1, \hat{\boldsymbol{\zeta}}_1^*) - \frac{1}{2}\left( \boldsymbol{\zeta}_1^T \Lambda_{n*}^{1/2} K_{n*}^{-1} \Lambda_{n*}^{1/2} \boldsymbol{\zeta}_1 - \hat{\boldsymbol{\zeta}}_1^{*T} \Lambda_{n*}^{1/2} K_{n*}^{-1} \Lambda_{n*}^{1/2} \hat{\boldsymbol{\zeta}}_1^* \right) \right| = o_p(1).$$

By the definition of $\hat{\boldsymbol{\zeta}}_1^*$ and $\tilde{\boldsymbol{\Pi}}_{(c)i}$,

$$(\boldsymbol{\zeta}_1 - \hat{\boldsymbol{\zeta}}_1^*)^T \tilde{\boldsymbol{\Pi}}_{(c)}^T \psi_\tau(\boldsymbol{e}) = (\boldsymbol{\zeta}_1 - \hat{\boldsymbol{\zeta}}_1^*)^T \Lambda_{n*}^{1/2} K_{n*}^{-1} \Lambda_{n*}^{1/2} \hat{\boldsymbol{\zeta}}_1^*,$$

where $\tilde{\boldsymbol{\Pi}}_{(c)} = \left(\tilde{\boldsymbol{\Pi}}_{(c)1}, \ldots, \tilde{\boldsymbol{\Pi}}_{(c)n}\right)^T$. Therefore,

$$\begin{aligned} & \sup_{\|\boldsymbol{\zeta}_1 - \hat{\boldsymbol{\zeta}}_1^*\|_{L_2} < \eta} \left| \sum_{i=1}^n U_i(\boldsymbol{\zeta}_1, \hat{\boldsymbol{\zeta}}_1^*) + (\boldsymbol{\zeta}_1 - \hat{\boldsymbol{\zeta}}_1^*)^T \tilde{\boldsymbol{\Pi}}_{(c)} \psi_\tau(\boldsymbol{e}) \right. \\ & \qquad \left. - \frac{1}{2}\left( \boldsymbol{\zeta}_1^T \Lambda_{n*}^{1/2} K_{n*}^{-1} \Lambda_{n*}^{1/2} \boldsymbol{\zeta}_1 - \hat{\boldsymbol{\zeta}}_1^{*T} \Lambda_{n*}^{1/2} K_{n*}^{-1} \Lambda_{n*}^{1/2} \hat{\boldsymbol{\zeta}}_1^* \right) \right| \\ & = \sup_{\|\boldsymbol{\zeta}_1 - \hat{\boldsymbol{\zeta}}_1^*\|_{L_2} < \eta} \left| \sum_{i=1}^n U_i(\boldsymbol{\zeta}_1, \hat{\boldsymbol{\zeta}}_1^*) - 1/2 (\boldsymbol{\zeta}_1 - \hat{\boldsymbol{\zeta}}_1^*)^T \Lambda_{n*}^{1/2} K_{n*}^{-1} \Lambda_{n*}^{1/2} (\boldsymbol{\zeta}_1 - \hat{\boldsymbol{\zeta}}_1^*) \right| \\ & = o_p(1). \end{aligned} \tag{A.12}$$

Let $\hat{\boldsymbol{\gamma}}^*_{(c)} \doteq (\hat{\gamma}^*_{\nu+1,1}, \ldots, \hat{\gamma}^*_{s,1})^T = K^{-1}_{n*}\Lambda^{1/2}_{n*}\hat{\boldsymbol{\zeta}}^*_1 + \boldsymbol{\gamma}^0_{(c)}$. Then by A4, the definition of $\hat{\omega}_{k,1}$, and the fact that $\max\left\{\|\hat{\boldsymbol{\zeta}}^*_1\|_{L_2}, \|\hat{\boldsymbol{\zeta}}_1\|_{L_2}\right\} = O_p(k_n^{1/2})$,

$$
\begin{aligned}
&n\lambda_{2,n}\sum_{k=\nu+1}^{s}\hat{\omega}_{k,1}(|\hat{\gamma}_{k,1}| - |\hat{\gamma}^*_{k,1}|) \geq -n\lambda_{2,n}\sum_{k=\nu+1}^{s}\hat{\omega}_{k,1}(|\hat{\gamma}_{k,1} - \hat{\gamma}^*_{k,1}|)\\
&\quad = O_p\left(n\lambda_{2,n}\left\|K^{-1}_{n*}\Lambda^{1/2}_{n*}(\hat{\boldsymbol{\zeta}}_1 - \hat{\boldsymbol{\zeta}}^*_1)\right\|_{L_1}\right) = O_p\left(n^{1/2}k_n^{1/2}\lambda_{2,n}\right) = o_p(1). \quad \text{(A.13)}
\end{aligned}
$$

When $\|\boldsymbol{\zeta}_1 - \hat{\boldsymbol{\zeta}}^*_1\|_{L_2} > \eta$, $(\boldsymbol{\zeta}_1 - \hat{\boldsymbol{\zeta}}^*_1)^T(\boldsymbol{\zeta}_1 - \hat{\boldsymbol{\zeta}}^*_1) > 0$. Combining (A.12) and (A.13),

$$
\begin{aligned}
&\lim_{n\to\infty} P\Bigg(\inf_{\|\boldsymbol{\zeta}_1 - \hat{\boldsymbol{\zeta}}^*_1\|_{L_2}\geq\eta}\sum_{i=1}^{n}\rho_\tau\Big(e_i - \boldsymbol{\zeta}_1^T\tilde{\boldsymbol{\Pi}}_{(c)i} - \hat{\boldsymbol{\zeta}}_2^T\tilde{\boldsymbol{\Pi}}_{(1)i} - D_{ni}\Big) + n\lambda_{2,n}\sum_{k=\nu+1}^{s}\hat{\omega}_{k,1}|\hat{\boldsymbol{\gamma}}_{(c)}|\\
&\quad > \sum_{i=1}^{n}\rho_\tau\Big(e_i - \hat{\boldsymbol{\zeta}}_1^{*T}\tilde{\boldsymbol{\Pi}}_{(c)i} - \hat{\boldsymbol{\zeta}}_2^T\tilde{\boldsymbol{\Pi}}_{(1)i} - D_{ni}\Big) + n\lambda_{2,n}\sum_{k=\nu+1}^{s}\hat{\omega}_{k,1}|\hat{\boldsymbol{\gamma}}^*_{(c)}|\Bigg) = 1. \quad \text{(A.14)}
\end{aligned}
$$

By the convexity of $l_2(\cdot)$, the definition of $\hat{\boldsymbol{\zeta}}^*_1$ and the fact that $l_2\left(\boldsymbol{\gamma}_{(1)}, \boldsymbol{\gamma}_{(c)}\right)$ is minimized at $(\hat{\boldsymbol{\gamma}}^T_{(1)}, \hat{\boldsymbol{\gamma}}^T_{(c)})^T$, (A.14) implies that for any $\eta > 0$, $P(\|\hat{\boldsymbol{\zeta}}_1 - \hat{\boldsymbol{\zeta}}^*_1\|_{L_2} > \eta) \to 0$, that is, $\|\hat{\boldsymbol{\zeta}}_1 - \hat{\boldsymbol{\zeta}}^*_1\|_{L_2} = o_p(1)$. This completes the proof of part (c) of Theorem 2.

### A.3. The Proof of Theorem 1

Let

$$
\begin{aligned}
\Phi_n &= n^{-1}\sum_{i=1}^{n}\boldsymbol{\Pi}_{(1)i}\boldsymbol{\Pi}^T_{(1)i}, \quad \Psi_n = n^{-1}\sum_{i=1}^{n}\boldsymbol{\Pi}_{(1)i}\boldsymbol{\Pi}^T_{(c)i},\\
\boldsymbol{\Pi}_{(c)*i} &= \boldsymbol{\Pi}_{(c)i} - \Psi_n^T\Phi_n^{-1}\boldsymbol{\Pi}_{(1)i},\\
\tilde{K}_{n*} &= \sum_{i=1}^{n}\boldsymbol{\Pi}_{(c)*i}\boldsymbol{\Pi}^T_{(c)*i}, \quad \tilde{\Lambda}_{n*} = \sum_{i=1}^{n}\sigma_i^2\boldsymbol{\Pi}_{(c)*i}\boldsymbol{\Pi}^T_{(c)*i}, \qquad \text{(A.15)}
\end{aligned}
$$

where $\sigma_i^2$ is the conditional variance of $e_i$ given $(\boldsymbol{X}_i, U_i)$.

The proof of Theorem 1 can be obtained in four similar steps as in the proof of Theorem 2. Lemmas A.4, A.5, and A.6 are needed in the first step, Lemmas A.1 and A.6 are needed in the second and the third steps, and the fourth step follows similar arguments as used in the proof of Theorem 3 in Zhao and Xue (2009). We omit the details.

## References

Cai, Z., Fan, J. and Li, R. (2000). Efficient estimation and inferences for varying-coefficient models. *J. Amer. Statist. Assoc.* **95**, 888-902.

Chen, H. (1991). Polynomial splines and nonparametric regression. *J. Nonparametric Statist.* **1**, 143-156.

Fan, J. and Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli* **11**, 1031-1057.

Fan, J. and Li, R. (2001). Variable selection via non-concave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

Fenske, N., Kneib, T. and Hothorn, T. (2009). Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression. Technical report, number 052, Department of Statistics, University of Munich. Available at `http://www.stat.uni-muenchen.de`.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *J. Roy. Statist. Soc. Ser. B* **55**, 757-796.

He, X. and Shi, P. (1994). Convergence rate of B-spline estimators of nonparametric conditional quantile functions. *J. Nonparametric Statist.* **3**, 299-308.

Hu, T. and Xia, Y. (2010). Adaptive semi-varying coefficient model selection. Manuscript.

Huang, J., Horowitz, J. and Wei, F. (2010). Variable selection in nonparametric additive models. *Ann. Statist.* **38**, 2282-2313.

Huang, J. Z., Wu, CO. and Zhou, L. (2002). Varying-coefficient models and basis function approximation for the analysis of repeated measurements. *Biometrika* **89**, 111-128.

Huang, J. Z., Wu, CO. and Zhou, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statist. Sinica* **14**, 763-788.

Kim, M. (2007). Quantile regression with varying coefficients. *Ann. Statist.* **35**, 92-108.

Koenker, R. (2010). Additive models for quantile regression: model selection and confidence bandaids. Working paper, number CWP33/10, Institute for Fiscal Studies and University of Illinois. Available at `http://www.cemmap.ac.uk/wps/cwp3310.pdf`

Koenker, R., Ng, P. and Portnoy, S. (1994). Quantile smoothing splines. *Biometrika* **81**, 673-680.

Leng, C. (2009). A simple approach for varying-coefficient model selection. *J. Statist. Plann. Inference* **139**, 2138-2146.

Lin, Y. and Zhang, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist.* **34**, 2272-2297.

Qu, A. and Li, R. (2006). Quadratic inference functions for varying coefficient models with longitudinal data. *Biometrics* **62**, 379-391.

Schumaker, L. L. (1981). *Spline Functions: Basic Theory.* Wiley, New York.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10**, 1040-1053.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Wang, L., Li, H., and Huang, J. Z. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *J. Amer. Statist. Assoc.* **103**, 1556-1569.

Wang, H. and Xia, Y. (2009). Shrinkage estimation of the varying coefficient model. *J. Amer. Statist. Assoc.* **104**, 747-757.

Wang, H., Zhu, Z. and Zhou, J. (2009). Quantile regression in partially linear varying coefficient models. *Ann. Statist.* **37**, 3841-3866.

Wei, Y. and He, X. (2006). Conditional growth charts(with discussion). *Ann. Statist.* **19**, 801-817.

Wu, C. O. and Chiang, C. T. (2000). Kernel smoothing on varying coefficient models with longitudinal dependent variable. *Statist. Sci.* **10**, 433-456.

Xia, Y., Zhang, W. and Tong, H. (2004). Efficient estimation for semivarying-coefficient models. *Biometrika* **91**, 661-681.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. Roy. Statist. Soc. Ser. B* **68**, 49-67.

Zhao, P. and Xue, L. (2009). Variable selection for semiparametric varying coefficient partially linear models. *Statist. Probab. Lett.* **79**, 2148-2157.

Zou, H. (2006). The Adaptive Lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418-1429.

Department of Statistics, Fudan University, Shanghai, China.

E-mail: tangyl@fudan.edu.cn

Department of Statistics, North Carolina State University, 2311 Stinson Drive, 4270 SAS Hall Raleigh, NC 27695-8203, USA.

E-mail: wang@stat.ncsu.edu

Department of Statistics, Fudan University, Shanghai, China.

E-mail: zhuzy@fudan.edu.cn

Department of Statistics, Chinese University of Hong Kong, Shatin, N.T., Hong Kong.

E-mail: xysong@sta.cuhk.edu.hk

(Received June 2010; accepted March 2011)