Statistics and Its Interface Volume 12 (2019) 397–413

Robust variable selection of varying coefficient partially nonlinear model based on quantile regression

Jing Yang*, Fang Lu, Guoliang Tian, Xuewen Lu, and Hu Yang

Quantile regression has been a popular topic for robust inference in semi-parametric models. However, there does not exist related literature for the varying coefficient partially nonlinear model (VCPNLM), which is the focus of this paper. Let alone on the quantile variable selection of VCPNLM. Specifically, via iteratively minimizing an average check loss estimation procedure based on quantile loss function, we propose a profile-type nonlinear quantile regression method for the VCPNLM, and further establish the asymptotic properties of the resulting estimators under some mild regularity conditions. In addition, to achieve sparsity when there exist irrelevant variables, we develop a variable selection procedure for high-dimensional VCPNLM by using the idea of shrinkage, and then demonstrate its oracle property. Two most important parameters including the smoothing parameter and the tuning parameter are also discussed, respectively. Finally, extensive numerical simulations with various errors are conducted to evaluate the finite sample performance of estimation and variable selection, and a real data analysis is further presented to illustrate the application of the proposed methods.

AMS 2000 subject classifications: Primary 62F35; secondary 62G20, 62G08, 62P12.
Keywords and phrases: Varying coefficient partially nonlinear model, Quantile regression, Variable selection, Oracle property, Robustness.

1. INTRODUCTION

Varying coefficient partially linear model (VCPLM), an important extension of partially linear model and varying-coefficient model, plays a crucial role among the semiparametric modeling. Due to its merits of interpretability of parametric model and flexibility of nonparametric model, extensive statistical inference on the estimation and variable selection methods about this model have been researched. Related literature include but not be restricted to [4, 10, 26, 3, 2, 17]. Recently, to capture some more complicated potential relationship between the response variable and the covariates, Li and Mei [16] proposed the so-called varying coefficient partially nonlinear model (VCPNLM) by replacing the linear component of VCPLM with a known nonlinear function of the covariates. Specifically, the VCPNLM has the form of

$$Y = g(X, \beta_0) + Z^T \alpha(U) + \varepsilon, \quad (1)$$

where "T" denotes the transpose of a vector or matrix throughout this paper, Y is the response variable, $X \in R^r$, $Z \in R^q$ and $U \in R$ are the associated covariates, $\beta_0 = (\beta_{01}, \ldots, \beta_{0p})^T$ is a vector of unknown coefficients that do not necessarily have the same dimension with X, $g(\cdot, \cdot)$ is a pre-specified nonlinear function, $\alpha(\cdot) = (\alpha_1(\cdot), \ldots, \alpha_q(\cdot))^T$ is a q-dimensional vector of unknown coefficient functions, ε is the random error. We assume the first component of Z is 1 and thus no separate intercept is explicitly written.

Obviously, model (1) is flexible enough to contain some classical models as its special cases, for instance, linear model, partially linear model, nonlinear model, partially nonlinear model, varying coefficient model and VCPLM. Then, it is necessary and meaningful to do some research on VCPNLM, but the relevant literature is relatively less so far. Li and Mei [16] proposed a profile nonlinear least squares estimation approach for the parametric vector β_0 and coefficient function vector $\alpha(\cdot)$, and established the asymptotic properties of the resulting estimators. Yang and Yang [34] presented a new estimation procedure for β_0 based on an orthogonality-projection method, and further studied the variable selection of $\alpha(\cdot)$ via smooth-threshold estimating equations. Qian and Huang [21] and Xiao and Chen [32] developed a corrected profile least-squared estimation procedure for the VCPNLM with measurement errors, and then did some statistical tests by using of the likelihood ratio test approach. Jiang et al. [9] proposed a robust estimation procedure based on exponential squared loss function. Zhou et al. [37] investigated the empirical likelihood inferences of model (1).

Note that, nearly all the existing estimation approaches for VCPNLM immerse in mean regression, based on either least squares or likelihood method. Although the mean estimate performs the best in the case of normally distributed error which is a standard assumption in mean regression, its performance can be highly influenced by the data set that

*Corresponding author.

contains some outliers or a thick tail. Worse still, the mean estimate will break down under the error of Cauchy distribution and is not a consistent estimate any more. Therefore, considering on the uncertain distributions of the data sets encountered in many practical problems, it is of interest and desirable to develop some robust inference on VCPNLM. As one of the most popular alternative ways to overcome the drawback of mean regression method, the quantile regression proposed by Koenker and Bassett [14] has been widely used as a robust statistical tool to explore the underlying relationship between the covariates and the response. Later on, this method has been quickly extended to the semiparametric models for the sake of robust inference, the recent literatures include but are not restricted to [36, 30] for the varying coefficient model, [28, 10] for the VCPLM, [8, 19] for the single-index model, [33, 18] for the partially linear single-index model, and a comprehensive book by Koenker [13] for an in-depth discussion about quantile method. However, even though the quantile regression has been well developed in various semiparametric models, there exists no research on the VCPNLM as far as we know.

Motivated by these observations, we extend the quantile methodology to model (1) and propose a variable selection procedure for the parametric vector in this paper. To the best of our knowledge, this is the first attempt at providing robust estimate and the first considering variable selection for the VCPNLM via quantile regression. More specifically, we first propose a profile nonlinear quantile regression approach for the parametric vector and the coefficient function vector based on an average check loss estimation procedure, and then consider a variable selection procedure by combining the proposed estimation approach with adaptive LASSO penalty. Theoretical properties including the asymptotic normalities of estimation and the oracle property of variable selection are established. It is worth noting that, here we employ the profile quantile technique to avoid undersmoothing, which should be a necessary condition in the frequently used backfitting approach for semiparametric models. Thus, our proposed method provides an easy way to get the optimal bandwidth for quantile regression, while brings some difficulty in proving the theoretical properties since one can not write the expressions of the solutions. Accordingly, the results derived in Theorem 2.1 plays a key role throughout our theoretical proofs. Some simulations with various distributed errors are conducted to evaluate the finite sample performance of estimation and variable selection, and a real data analysis is presented to further illustrate the application of our proposed methods.

The rest of this paper is organized as follows. In Section 2, we first present the detailed profile nonlinear quantile regression methodology and its calculation procedure, then the asymptotic properties of the resulting estimators are established. In Section 3, we employ the adaptive LASSO penalty to the parametric components for a sparse estimate, and the oracle property of the penalized estimate is also derived. In Section 4, we conduct some numerical simulations and apply the proposed methods to analyze the Boston housing price data. A short conclusion is summarized in Section 5. Regularity conditions and all technical proofs are collected in the Appendix.

2. A PROFILE-TYPE QUANTILE REGRESSION

2.1 Estimation methodology

Let $\rho_\tau(t) = t[\tau - I(t < 0)]$ be the check loss function for $\tau \in (0, 1)$. Quantile regression (QR) is often used to estimate the conditional quantile of the response variable Y, which has the following definition

$$Q_\tau(x, z, u) = \arg\min_a E\{\rho_\tau(Y - a) \mid (X, Z, U) = (x, z, u)\}.$$

The VCPNLM assumes that the τ-th conditional quantile function of Y can be expressed as $Q_\tau(x, z, u) = g(x, \beta_\tau) + z^T\alpha_\tau(u)$. Thus, for any given τ, quantile regression can be applied to estimate the parametric vector β_τ and the coefficient function vector $\alpha_\tau(\cdot)$.

Suppose that $\{X_i, Y_i, Z_i, U_i\}_{i=1}^n$ are independent identically distributed (i.i.d.) random samples generated from

$$Y = g(X, \beta_\tau) + Z^T\alpha_\tau(U) + \varepsilon_\tau, \tag{2}$$

where ε_τ is the random error with its τ-th quantile equaling to zero when given (X, Z, U). For the convenience of notation, we omit τ from β_τ, $\alpha_\tau(\cdot)$ and ε_τ in model (2) wherever clear from the context, but we should keep in mind that those quantities are τ-specific.

From the idea of [14], the τ-th QR estimate of β and $\alpha(\cdot)$ can be obtained by minimizing the following nonlinear quantile loss function

$$\sum_{i=1}^n \rho_\tau\{Y_i - g(X_i, \beta) - Z_i^T\alpha(U_i)\}. \tag{3}$$

Obviously, the objective function (3) involves both the nonparametric and parametric components, which should be estimated with different rates of convergence, so we propose the following two-step procedure to estimate $\alpha(\cdot)$ and β.

In the first step, for given β, model (2) can be rewrite as

$$Y - g(X, \beta) = Z^T\alpha(U) + \varepsilon.$$

When U is closed to u, the coefficient functions $\alpha_k(U)$ can be locally linearly approximated as

$$\alpha_k(U) \approx \alpha_k(u) + \alpha_k'(u)(U - u), \quad k = 1, \ldots, q.$$

So the coefficient function vector $\alpha(u)$ and its derivative $\alpha'(u)$ can be estimated by minimizing

$$\sum_{i=1}^n \rho_\tau\{Y_i - g(X_i, \beta) - Z_i^T[\alpha(u) + \alpha'(u)U_{i0}]\}K_h(U_{i0}), \tag{4}$$

where $U_{i0} = U_i - u$, $K_h(\cdot) = K(./h)/h$ with $K(\cdot)$ being a kernel weight function and h being the bandwidth. Taking u_0 in (4) be $U_1, \ldots, U_n$, the estimates of $\alpha(U_j)$ and $\alpha'(U_j)$ for $j = 1, \ldots, n$ are obtained, which are denoted by $\tilde{\alpha}(U_j)$ and $\tilde{\alpha}'(U_j)$, respectively.

It is worthy pointing out here that $\tilde{\alpha}(U_j)$ and $\tilde{\alpha}'(U_j)$ depend on β, so write them as $\tilde{\alpha}(U_j, \beta)$ and $\tilde{\alpha}'(U_j, \beta)$ should be more accurate. However, considering this is just a notation issue and the latter expression is somewhat tiring, we abbreviate $\tilde{\alpha}(U_j, \beta)$ and $\tilde{\alpha}'(U_j, \beta)$ as $\tilde{\alpha}(U_j)$ and $\tilde{\alpha}'(U_j)$ whenever no confusion is caused hereafter, but we should keep in mind that the dependence is truly existed.

In the second step, we update the estimate of β based on $\{\tilde{\alpha}(U_j), \tilde{\alpha}'(U_j)\}_{j=1}^n$. In details, the parameter β can be estimated via minimizing

$$\sum_{j=1}^n \sum_{i=1}^n \rho_\tau\{Y_i - g(X_i, \beta) - Z_i^T[\tilde{\alpha}(U_j) + \tilde{\alpha}'(U_j)U_{ij}]\}w_{ij},$$

where $U_{ij} = U_i - U_j$, $w_{ij} = K_h(U_{ij})/\sum_{l=1}^n K_h(U_{lj})$ satisfy $\sum_{i=1}^n w_{ij} = 1$ for $\forall j = 1, \ldots, n$.

Let $a_j = \alpha(U_j) = (\alpha_1(U_j), \ldots, \alpha_q(U_j))^T$, $b_j = \alpha'(U_j) = (\alpha_1'(U_j), \ldots, \alpha_q'(U_j))^T$, $\tilde{a}_j = \tilde{\alpha}(U_j)$ and $\tilde{b}_j = \tilde{\alpha}'(U_j)$ for $j = 1, \ldots, n$. Consequently, a profile estimation algorithm for estimating β_0 can be summarized as follows.

Step 1. Given the current estimator of β_0 by $\tilde{\beta}^{(t)}$, then a_j and b_j can be estimated by

$$\begin{aligned}(\tilde{a}_j, \tilde{b}_j) \quad = \quad & \arg\min_{a_j, b_j} \sum_{i=1}^n \rho_\tau\{Y_i - g(X_i, \tilde{\beta}^{(t)}) \\ & -Z_i^T[a_j + b_j U_{ij}]\}w_{ij}\end{aligned}$$

for $j = 1, \ldots, n$.

Step 2. Based on $\{\tilde{a}_j, \tilde{b}_j\}_{j=1}^n$, we update the estimator $\tilde{\beta}^{(t)}$ by

$$\begin{aligned}\tilde{\beta}^{(t+1)} \quad = \quad & \arg\min_{\beta} \sum_{j=1}^n \sum_{i=1}^n \rho_\tau\{Y_i - g(X_i, \beta) \\ & -Z_i^T[\tilde{a}_j + \tilde{b}_j U_{ij}]\}w_{ij}.\end{aligned}$$

Step 3. Iterate Step 1-Step 2 until convergence. Denote the final estimate of β_0 by $\hat{\beta}$.

Note that, the idea for estimate β_0 is somewhat like the profile least squares (PLS) of Fan and Huang [4] and Li and Mei [16] for the VCPLM and VCPNLM except for the following small difference: The PLS approach has a explicit expression of the solutions due to its quadratic loss function but quantile do not have, thus here we use an iterative procedure. Meanwhile, this procedure enables us to take use of the estimates of $\alpha'(\cdot)$ in addition to the estimates of $\alpha(\cdot)$, which could avoid the requirement of under-smoothing in the estimation of semi-parametric models (see the discussions on Page 1166 of Xia and Härdle [31]). Therefore, we call our proposed methodology the profile-type nonlinear quantile regression (PQR) method.

With the estimate $\hat{\beta}$ of β_0, for any inner point u in the support of U, the estimators of $\alpha(u)$ and $\alpha'(u)$ will be $\hat{a} = \hat{\alpha}(u, \hat{\beta})$ and $\hat{b} = \hat{\alpha}'(u, \hat{\beta})$, where $\hat{a}$ and $\hat{b}$ are obtained from

$$\begin{aligned}(5) \qquad (\hat{a}, \hat{b}) \quad = \quad & \arg\min_{a,b} \sum_{i=1}^n \rho_\tau\{Y_i - g(X_i, \hat{\beta}) \\ & -Z_i^T[a + bU_{i0}]\}K_h(U_{i0}).\end{aligned}$$

2.2 Theoretical properties

This subsection aims at to establish the asymptotic properties of the resulting estimators. We first introduce some notations and definitions. Denote by $f_Y(\cdot \mid (X, Z, U))$ and $F_Y(\cdot \mid (X, Z, U))$ the conditional density function and cumulative distribution function of Y on (X, Z, U), respectively. Let $g'(x, \beta) = \partial g(x, \beta)/\partial\beta \in R^{p\times 1}$ and $g''(x, \beta) = \partial^2 g(x, \beta)/\partial\beta\partial\beta^T \in R^{p\times p}$ be the first and second order derivatives of $g(x, \beta)$ with respect to β. Let $f_U(\cdot)$ be the marginal density function of U and

$$\mu_j = \int u^j K(u)du, \qquad \nu_j = \int u^j K^2(u)du, \ j = 0, 1, 2, \ldots$$

Theorem 2.1. *Suppose that the regularity conditions* (C1)-(C6) *given in the Appendix hold. If $h = O(n^{-\delta})$ with $\delta \in (1/6, 1/4)$, then we have*

$$\begin{aligned}\hat{\alpha}(u, \beta) = \alpha(u) + \frac{1}{2}\mu_2 h^2 \alpha''(u) - E(ZZ^T \mid U = u)^{-1} \cdot \\ E(Zg'(X, \beta)^T \mid U = u)d_\beta + Q_{n1}(u) + O(h^4 + \delta_\beta^2), \\ \hat{\alpha}'(u, \beta) = \alpha'(u) + \frac{1}{h}Q_{n2}(u) + O(h^2 + \delta_\beta),\end{aligned}$$

where

$$\begin{aligned}Q_{n1}(u) = \big\{nf_U(u)E\big[f_Y(Q_\tau(X, Z, U))ZZ^T \mid U = u\big]\big\}^{-1} \cdot \\ \sum_{i=1}^n K_{i,h} Z_i \psi_\tau(\varepsilon_i), \\ Q_{n2}(u) = \big\{nh\mu_2 f_U(u)E\big[f_Y(Q_\tau(X, Z, U))ZZ^T \mid U = u\big]\big\}^{-1} \cdot \\ \sum_{i=1}^n K_{i,h} Z_i U_{i0} \psi_\tau(\varepsilon_i),\end{aligned}$$

$d_\beta = \beta - \beta_0$, $\delta_\beta = |\beta - \beta_0|$, $K_{i,h} = K_h(U_{i0})$, $U_{i0} = U_i - u$ *and* $\psi_\tau(t) = \tau - I(t < 0)$.

It is worth pointing out that the results derived in Theorem 2.1 always act as a key role throughout our theoretical proofs. From now on, we assume in the following context that the initial value β lies in a small neighborhood of β_0: $\Theta_n = \{\|\beta - \beta_0\| \le C_0 n^{-1/2+c_0}\}$, where C_0 and $c_0 < 1/20$ are some positive constants. These assumptions are feasible because such an initial estimator can be obtained directly from the PLS approach of Li and Mei [16] or the orthogonality-projection method of Yang and Yang [34].

Theorem 2.2. *Suppose that the regularity conditions* (C1)-(C6) *given in the Appendix hold. If* $h = O(n^{-\delta})$ *with* $\delta \in (1/6, 1/4)$*, then we have*

$$\|\hat{\beta} - \beta_0\| = O_p(1/\sqrt{n}). \tag{6}$$

Theorem 2.3. *Under the same conditions assumed in Theorem 2.2, we have*

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N\big(0,\ \tau(1-\tau)\Sigma^{-1}\Sigma_0\Sigma^{-1}\big), \tag{7}$$

where $\Sigma_0 = E\{\Xi\}$, $\Sigma = E\{f_Y(Q_\tau(X, Z, U))\Xi\}$ *and*

$$\begin{aligned}\Xi &= g'(X, \beta_0)g'(X, \beta_0)^T - E(g'(X, \beta_0)Z^T \mid U) \cdot \\ &\quad E(ZZ^T \mid U)^{-1}E(Zg'(X, \beta_0)^T \mid U).\end{aligned}$$

The results of Theorems 2.2 and 2.3 indicate that without undersmoothing, our proposed profile nonlinear quantile regression procedure would eventually provide a $\sqrt{n}$-consistent estimate of β_0 with its asymptotic distribution being (7). The following theorem presents the asymptotic normalities of the estimates of $\alpha(\cdot)$ as well as its first derivative $\alpha'(\cdot)$.

Theorem 2.4. *Suppose that* $\hat{\beta} - \beta_0 = O_p(n^{-1/2})$*,* u *is an interior point on its support* $\mathcal{U}$*. If the same conditions given in Theorem 2.3 hold, then we have*

$$\sqrt{nh}\left\{\begin{pmatrix} \hat{\alpha}(u, \hat{\beta}) - \alpha(u) \\ h\big(\hat{\alpha}'(u, \hat{\beta}) - \alpha'(u)\big) \end{pmatrix} - \frac{1}{2}\mu_2 h^2 \Omega^{-1}(u)\Gamma(u)\alpha''(u)\right\}$$
$$\xrightarrow{d} N\Big(0,\ \tau(1-\tau)\Omega^{-1}(u)\Omega_0(u)\Omega^{-1}(u)\Big),$$

where

$$\begin{aligned}\Omega(u) &= f_U(u)E\left\{f_Y(Q_\tau(X, Z, U))\varpi \mid U = u\right\}, \\ \Omega_0(u) &= f_U(u)E\left\{f_Y(Q_\tau(X, Z, U))\varpi_0 \mid U = u\right\}, \\ \Gamma(u) &= E\left\{f_Y(Q_\tau(X, Z, U))\begin{pmatrix} ZZ^T \\ 0 \end{pmatrix} \mid U = u\right\},\end{aligned}$$

$$\varpi = \begin{pmatrix} ZZ^T & 0 \\ 0 & \mu_2 ZZ^T \end{pmatrix} \text{ and } \varpi_0 = \begin{pmatrix} \nu_0 ZZ^T & 0 \\ 0 & \nu_2 ZZ^T \end{pmatrix}$$

2.3 Bandwidth selection

It is well known that the bandwidth always plays a crucial role in local polynomial smoothing wherever in mean regression and quantile regression, because this parameter controls the curvature of the fitted function. One advantage of the proposed method lies in that undersmoothing is not necessary, which can be observed from the condition $h = O(n^{-\delta})$ with $\delta \in (1/6, 1/4)$. As a result, some existing bandwidth selection criteria such as the rule of thumb [24], k-fold cross-validation (k-CV, [1]) and the generalized cross-validation (GCV, [25]) can be used. Here following the similar arguments on quantile regression by [35], we take the bandwidth as

$$h_\tau = h_m\{\tau(1-\tau)/\phi(\Phi^{-1}(\tau))^2\}^{1/5}, \tag{8}$$

where h_m is the optimal bandwidth used in mean regression, ϕ and Φ are the probability density function and the cumulative distribution function of the standard normal distribution, respectively. Based on the facts that many existing algorithms can be employed to the selection of h_m (see the discussions of [22]) and our proposed procedure is not sensitive to the choice of bandwidth, thus in this paper we select h_m through the plug-in bandwidth selector developed by Ruppert et al. [22], which is easily implemented via the function "*dpill*" in R software, for the purpose of reducing computational burden.

Remark 1. Compared with the k-CV and GCV approaches, the bandwidth given in (8) not only provides an easy approach to get the optimal bandwidth for quantile regression, but also can effectively reduce the burden of calculations. For detailed discussion about this method, we recommend to turn to Yu and Jones [35].

3. VARIABLE SELECTION PROCEDURE

In practice, the true model is often unknown, which permits the possibility of selecting an overfitted or an underfitted model, leading to inefficient predictions or biased estimators. With high-dimensional covariates, sparse modeling is often considered superior owing to enhanced model predictability and interpretability. This motivates us to develop a variable selection procedure for the purpose of selecting significant parametric components in model (1).

With this goal in mind, we employ the adaptive LASSO penalty proposed by Zou [38] to simultaneously select significant variables and estimate their effects. To this end, we construct the following adaptive penalized quantile loss function

$$\begin{aligned}\Phi_n(\lambda, \beta) &= \sum_{j=1}^n \sum_{i=1}^n \rho_\tau\Big\{Y_i - g(X_i, \beta) - Z_i^T[\hat{\alpha}(U_j) \\ &\quad + \hat{\alpha}'(U_j)U_{ij}]\Big\}w_{ij} + \sum_{k=1}^p \lambda_k|\beta_k|,\end{aligned}$$

where $\lambda = (\lambda_1, \ldots, \lambda_p)^T$ is the tuning parameter. Thus given λ, the penalized estimate of β_0, denoted as $\hat{\beta}^\lambda$, can be obtained from

$$\hat{\beta}^\lambda = \arg\min_\beta \Phi_n(\lambda, \beta).$$

As will be shown in Theorem 3.1 that the penalized estimate $\hat{\beta}^\lambda$ enjoys the oracle property if the adaptive tuning parameters $\lambda = (\lambda_1, \ldots, \lambda_p)^T$ is properly selected. Thus, the values of $\lambda_k, k = 1, \ldots, p$ should be determined before implementation in practice, and many existing selection criteria based on data-driven method such as GCV and BIC can be used. In view of the expensive computation to simultaneously select a p-dimensional parameter, we simply set $\lambda_k = \lambda_n/|\hat{\beta}_k|^2$, $k = 1, \ldots, p$ according to [38], where λ_n is a common penalty parameter with its dimension being one, and $\hat{\beta}$ denotes the unpenalized estimator generated from Section 2. That is to say, the considered adaptive penalized objective function is

$$\Phi_n(\lambda_n, \beta) = \sum_{j=1}^{n}\sum_{i=1}^{n} \rho_\tau\Big\{Y_i - g(X_i, \beta) - Z_i^T[\hat{\alpha}(U_j) + \hat{\alpha}'(U_j)U_{ij}]\Big\}w_{ij} + \lambda_n \sum_{k=1}^{p} \frac{|\beta_k|}{|\hat{\beta}_k|^2}.$$

Finally, we adopt a BIC-type criterion to select the one dimension tuning parameter λ_n following Wang and Leng [27]. In detail, we select λ_n as

$$\hat{\lambda}_n^{BIC} = \arg\min_{\lambda} \log\Big\{\sum_{j=1}^{n}\sum_{i=1}^{n} \rho_\tau\{Y_i - g(X_i, \hat{\beta}^\lambda) - Z_i^T[\hat{\alpha}(U_j) + \hat{\alpha}'(U_j)U_{ij}]\}w_{ij}\Big\} + \frac{\log n}{n} df_\lambda, \tag{9}$$

where df_λ is the effective degrees of freedom measured by the number of nonzero coefficients in $\hat{\beta}^\lambda$ for any candidate penalty parameter λ. In this paper, a residual is regarded as zero if its absolute value is smaller than 10^{-6}, and the simulation studies later show that the selected $\hat{\lambda}_n^{BIC}$ performs very well.

Therefore, the detailed penalized estimation procedure of β_0 can be summarized as follows:

Step 1*. Same to Step 1 presented in Section 2.1. Specifically, given the current estimator of β_0 by $\tilde{\beta}^{(t)}$, a_j and b_j are estimated by

$$(\tilde{a}_j, \tilde{b}_j) = \arg\min_{a_j, b_j} \sum_{i=1}^{n} \rho_\tau\{Y_i - g(X_i, \tilde{\beta}^{(t)}) - Z_i^T[a_j + b_jU_{ij}]\}w_{ij}$$

for $j = 1, \ldots, n$.

Step 2*. For any one alternative λ_n, we can calculate an estimator $\tilde{\beta}^{\lambda_n}$ from

$$\tilde{\beta}^{\lambda_n} = \arg\min_{\beta} \sum_{j=1}^{n}\sum_{i=1}^{n} \rho_\tau\Big\{Y_i - g(X_i, \beta) - Z_i^T[\tilde{a}_j + \tilde{b}_jU_{ij}]\Big\}w_{ij} + \lambda_n \sum_{k=1}^{p} \frac{|\beta_k|}{|\hat{\beta}_k|^2},$$

where the estimators $\{\tilde{a}_j, \tilde{b}_j\}_{j=1}^n$ are obtained from Step 1*. Then, we use the BIC-type criterion (9) to search an optimal $\hat{\lambda}_n^{BIC}$, and the related estimator $\tilde{\beta}^{\hat{\lambda}_n^{BIC}}$ is selected as the updated estimator $\tilde{\beta}^{(t+1)}$ of $\tilde{\beta}^{(t)}$.

Step 3*. Iterate Steps 1-2 until convergence. Denote the finally penalized estimator of β_0 by $\hat{\beta}^\lambda$.

In the next, we will establish the oracle property of the adaptive penalized estimator $\hat{\beta}^\lambda$. Without loss of generality, we assume that $\beta_{0I} = (\beta_{01}, \ldots, \beta_{0l})^T$ consists all of the nonzero components of β_0, and $\beta_{0II} = (\beta_{0,l+1}, \ldots, \beta_{0,p})^T$ corresponding to all the zero coefficients. Divide $\hat{\beta}^\lambda$ into $\hat{\beta}^\lambda = \big(\hat{\beta}_I^{\lambda^T}, \hat{\beta}_{II}^{\lambda^T}\big)^T$, then we have the following theorem holds.

Theorem 3.1. *Suppose that the same conditions given in Theorem 2.3 hold. If $\lambda_n \to \infty$ and $\lambda_n/\sqrt{n} \to 0$ as $n \to \infty$, then we have*
(i) *Selection consistency: $\hat{\beta}_{II}^\lambda = 0$ with probability tending to 1,*
(ii) *Asymptotic normality:*

$$\sqrt{n}(\hat{\beta}_I^\lambda - \beta_{0I}) \xrightarrow{d} N\big(0,\ \tau(1-\tau)\Sigma_I^{-1}\Sigma_{0I}\Sigma_I^{-1}\big), \tag{10}$$

where Σ_I and Σ_{0I} are the top-left l-by-l submatrices of Σ and Σ_0 in Theorem 2.3, respectively.

Remark 2. It is worth noting that here we choose adaptive LASSO penalty solely for the convenience of computation. Other penalty such as smoothly clipped absolute deviation (SCAD, [5]) can also be considered, and the oracle property can be similarly established.

4. NUMERICAL RESULTS

In this section, we will conduct some Monte Carlo simulations with various distributed errors to evaluate the finite sample performance of the proposed estimation and variable selection procedures, and then apply the new methods to Boston housing price data for further application. In all examples, we fix the kernel function $K(u)$ to be the Epanechnikov kernel, that is $K(u) = 0.75(1-u^2)I(|u| \leq 1)$, where $I(\cdot)$ is the indicative function.

4.1 Monte Carlo simulations

Example 1. In this example, we are interested in comparing the estimation performance of our proposed profile-type quantile regression (PQR) method when $\tau = 0.25$, 0.5, 0.75 with the profile least squares (PLS) approach developed in [16]. The data are generated from the model

$$Y = g(X_1, X_2; \beta_1, \beta_2) + Z_1\alpha_1(U) + Z_2\alpha_2(U) + \varepsilon, \tag{11}$$

where the covariates $X = (X_1, X_2)^T$ comes from a bivariate normal distribution with mean zero and $cov(X_i, X_j) = 0.5^{|i-j|}$, $Z_1 \sim N(0,1)$, $Z_2 \sim U(0,3)$ and $U \sim U(0,1)$, $g(X_1, X_2; \beta_1, \beta_2) = \exp(X_1\beta_1 + X_2\beta_2)$ with the true parameter $\beta_0 = (\beta_{01}, \beta_{02})^T = (1, 1.5)^T$, $\alpha_1(U) = \sin(2\pi U)$ and $\alpha_2(U) = 3.5\{\exp(-(4U-1)^2) + \exp(-(4U-3)^2) - 1.5\}$.

To evaluate the efficiencies of the considered estimators, we introduce some measurement criteria: the mean absolute deviation (MAD) and the corresponding standard deviations (SD) for the parametric components; the average square root of average square error

Table 1. The estimation results in Example 1 with sample size $n = 200$

Error type	Method	MAD($\hat{\beta}_1$)	SD	MAD($\hat{\beta}_2$)	SD	$RASE_1$	SD	$RASE_2$	SD
N(0,1)	PLS	0.0030	0.0026	0.0027	0.0024	0.1937	0.0463	0.1636	0.0397
	PQR(0.25)	0.0033	0.0030	0.0029	0.0027	0.2418	0.0541	0.2055	0.0504
	PQR(0.5)	0.0031	0.0028	0.0027	0.0026	0.2245	0.0527	0.1909	0.0465
	PQR(0.75)	0.0034	0.0030	0.0031	0.0029	0.2488	0.0530	0.2126	0.0493
t(2)	PLS	0.0094	0.0300	0.0097	0.0486	0.7592	0.5076	0.7261	0.4494
	PQR(0.25)	0.0074	0.0076	0.0065	0.0071	0.3123	0.0845	0.2677	0.0821
	PQR(0.5)	0.0061	0.0059	0.0052	0.0052	0.2926	0.0811	0.2404	0.0747
	PQR(0.75)	0.0074	0.0079	0.0061	0.0073	0.3104	0.0843	0.2658	0.0829
CN(0.1,10)	PLS	0.0082	0.0097	0.0075	0.0092	0.6110	0.2372	0.4751	0.2162
	PQR(0.25)	0.0056	0.0056	0.0050	0.0054	0.2731	0.0711	0.2304	0.0623
	PQR(0.5)	0.0047	0.0056	0.0043	0.0052	0.2649	0.0635	0.2138	0.0558
	PQR(0.75)	0.0060	0.0059	0.0056	0.0058	0.2817	0.0736	0.2348	0.0654
Cauchy	PLS	0.0679	0.1550	0.0606	0.1381	6.9564	16.548	5.5902	16.1843
	PQR(0.25)	0.0112	0.0155	0.0113	0.0161	0.4119	0.1490	0.3405	0.1162
	PQR(0.5)	0.0082	0.0083	0.0077	0.0090	0.3877	0.1381	0.3172	0.1009
	PQR(0.75)	0.0119	0.0163	0.0114	0.0157	0.4146	0.1502	0.3447	0.1195

Table 2. The estimation results in Example 1 with sample size $n = 400$

Error type	Method	MAD($\hat{\beta}_1$)	SD	MAD($\hat{\beta}_2$)	SD	$RASE_1$	SD	$RASE_2$	SD
N(0,1)	PLS	0.0016	0.0014	0.0015	0.0014	0.1347	0.0289	0.1155	0.0269
	PQR(0.25)	0.0019	0.0016	0.0016	0.0016	0.1750	0.0380	0.1484	0.0321
	PQR(0.5)	0.0017	0.0015	0.0015	0.0014	0.1610	0.0328	0.1395	0.0290
	PQR(0.75)	0.0018	0.0015	0.0016	0.0015	0.1704	0.0383	0.1491	0.0334
t(2)	PLS	0.0057	0.0101	0.0049	0.0087	0.4736	0.1981	0.3901	0.1520
	PQR(0.25)	0.0035	0.0036	0.0030	0.0027	0.2005	0.0537	0.1840	0.0491
	PQR(0.5)	0.0030	0.0027	0.0026	0.0022	0.1853	0.0495	0.1702	0.0456
	PQR(0.75)	0.0033	0.0034	0.0028	0.0025	0.2039	0.0551	0.1884	0.0506
CN(0.1,10)	PLS	0.0050	0.0061	0.0044	0.0049	0.4287	0.1379	0.3375	0.1094
	PQR(0.25)	0.0028	0.0030	0.0024	0.0022	0.1931	0.0450	0.1524	0.0378
	PQR(0.5)	0.0024	0.0023	0.0021	0.0019	0.1797	0.0413	0.1497	0.0346
	PQR(0.75)	0.0030	0.0037	0.0028	0.0030	0.1998	0.0469	0.1560	0.0395
Cauchy	PLS	0.0780	0.3157	0.0677	0.2718	19.111	176.47	9.564	57.339
	PQR(0.25)	0.0069	0.0103	0.0065	0.0107	0.2518	0.0730	0.2157	0.0624
	PQR(0.5)	0.0043	0.0051	0.0036	0.0046	0.2432	0.0691	0.1984	0.0531
	PQR(0.75)	0.0066	0.0108	0.0058	0.0101	0.2534	0.0739	0.2113	0.0647

(RASE) as well as its SD for the coefficient functions $\alpha_1(u), \ldots, \alpha_q(u)$, in which the RASE is defined as $\text{RASE}_k = \left\{\frac{1}{n_{grid}} \sum_{i=1}^{n_{grid}} \left(\hat{\alpha}_k(u_i) - \alpha_k(u_i)\right)^2\right\}^{1/2}$ for any estimator $\hat{\alpha}_k$ of α_k, where $\{u_i, i = 1, \ldots, n_{grid}\}$ is a set of grid points. In addition, to show the robustness of our proposed method, four different error distributions for ε are considered: standard normal distribution N(0,1), t(2) distribution, contaminated normal distribution $0.9\text{N}(0,1^2)+0.1\text{N}(0,10^2)$ abbreviate as CN(0.1,10), and standard Cauchy distribution.

The corresponding results with 200 simulation runs are summarized in Tables 1 and 2. As we can clearly see that for the normal error, the PLS method performs the best, and the PQR methods lose some efficiency but are still very comparable. Whereas, for the other three non-normal errors, the performances of our proposed PQR procedures are significantly better than that of PLS approach no matter in terms of estimation accuracy or stability. Especially when the error follows Cauchy distribution, the PLS approach is

Table 3. Estimation results in Example 1 at $\tau = 0.1$ and $\tau = 0.9$ with sample size $n = 200$

Error type	Method	MAD($\hat{\beta}_1$)	SD	MAD($\hat{\beta}_2$)	SD	RASE$_1$	SD	RASE$_2$	SD
t(2)	$\tau = 0.1$	0.0086	0.0117	0.0074	0.0111	0.3851	0.1214	0.3217	0.1169
	$\tau = 0.9$	0.0081	0.0123	0.0078	0.0116	0.3775	0.1253	0.3232	0.1146
Cauchy	$\tau = 0.1$	0.0193	0.0241	0.0171	0.0230	0.5224	0.2116	0.4348	0.1945
	$\tau = 0.9$	0.0178	0.0212	0.0167	0.0238	0.5177	0.2135	0.4376	0.1912

Table 4. Results of the proposed estimator PQR(0.5) for model (11) across different bandwidth selection procedures with $n = 400$

Error type	Bandwidth	MAD($\hat{\beta}_1$)	SD	MAD($\hat{\beta}_2$)	SD	RASE$_1$	SD	RASE$_2$	SD
N(0,1)	Plug-in	0.0018	0.0016	0.0017	0.0015	0.1618	0.0312	0.1389	0.0291
	5-CV	0.0019	0.0016	0.0018	0.0016	0.1626	0.0320	0.1398	0.0294
	10-CV	0.0015	0.0015	0.0016	0.0014	0.1604	0.0313	0.1390	0.0285
	GCV	0.0017	0.0016	0.0015	0.0015	0.1601	0.0305	0.1384	0.0287
t(2)	Plug-in	0.0028	0.0027	0.0029	0.0025	0.1848	0.0490	0.1713	0.0457
	5-CV	0.0026	0.0028	0.0030	0.0021	0.1852	0.0494	0.1701	0.0466
	10-CV	0.0026	0.0023	0.0025	0.0022	0.1839	0.0479	0.1688	0.0453
	GCV	0.0026	0.0024	0.0028	0.0021	0.1838	0.0483	0.1704	0.0449

Table 5. The empirical coverage probabilities of confidence intervals for model (11) with sample size $n = 200$

Error type	Method	β_1		β_2	
		90%	95%	90%	95%
N(0,1)	PQR(0.25)	0.8960	0.9460	0.8980	0.9440
	PQR(0.5)	0.8980	0.9500	0.9020	0.9480
	PQR(0.75)	0.8940	0.9460	0.8960	0.9480
CN(0.1,10)	PQR(0.25)	0.8940	0.9440	0.9000	0.9440
	PQR(0.5)	0.9020	0.9480	0.8960	0.9480
	PQR(0.75)	0.8960	0.9460	0.8980	0.9460

corrupted due to the infinite variance of model error, and the superiorities of PQR methods become more remarkable. On the other hand, as the sample size increases from 200 to 400, there is an obvious tendency that the performances of all the considered estimates get better and better except for the PLS method of Cauchy error, this phenomenon is reasonable and coincide with the theory that the PLS-based estimators no longer enjoy root-n consistency in this case. Besides, although the PQR procedures with quantiles being 0.25, 0.5 and 0.75 have similar performances, the median method seems to be slightly better in most scenarios. Furthermore, we have added some simulations to evaluate the performance of our proposed method at the tails of the distributions in Table 3. Although the effects of $\tau = 0.1, 0.9$ are not as good as $\tau = 0.25, 0.5, 0.75$ that presented in Table 1, the overall performances are acceptable and the estimators are significantly better than the ones of PLS, which imply the validation of our proposed method.

Some extra simulations are conducted in this example to examine the performance of our developed method. Table 4 shows some results of the PQR estimator across different bandwidth selection, in which only $\tau = 0.5$ is consider since other two cases are similar from previous studies and "Plug-in" means that the bandwidth is generated by (8) with h_m obtained by the plug-in bandwidth selector. It is clear that the results of four different bandwidth selection procedures are much similar, which indicates the stability of our proposed new approach with respect to the bandwidth selection. Thus, we use the formula (8) to obtain the bandwidth for its convenience and time-saving benefit in the latter numerical studies. In addition, Table 5 presents the empirical coverage rates at the nominal level 90% and 95%

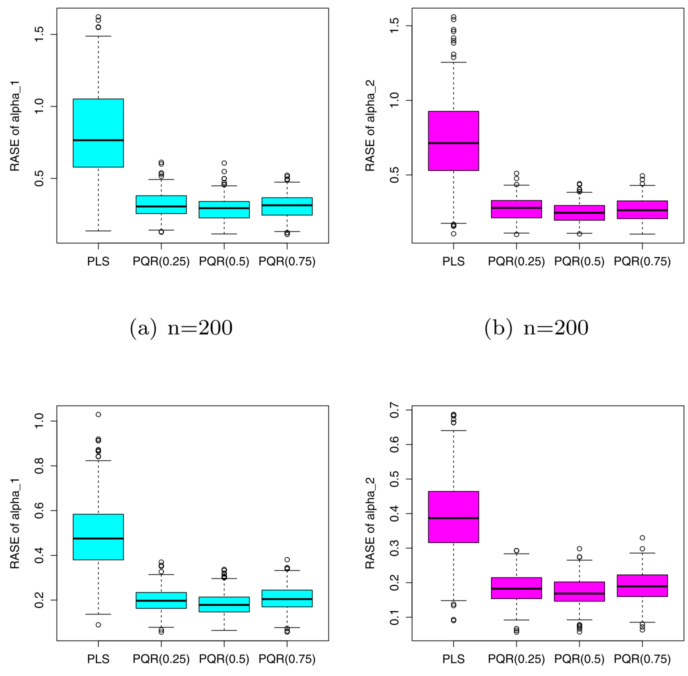
Table 6. The results of variable selection in Example 2

n	Error type	Method	NC	NIC	U.fit	O.fit	C.fit
200	N(0,1)	PQR(0.25)	4.950	0.005	0.005	0.040	0.955
		PQR(0.5)	4.965	0	0	0.035	0.965
		PQR(0.75)	4.960	0	0	0.040	0.960
		PQR(0.5)-SCAD	4.960	0.005	0.005	0.035	0.960
	t(2)	PQR(0.25)	4.935	0.015	0.010	0.055	0.935
		PQR(0.5)	4.940	0.005	0.005	0.055	0.945
		PQR(0.75)	4.930	0.015	0.010	0.060	0.930
		PQR(0.5)-SCAD	4.925	0.015	0.010	0.060	0.930
	CN(0.1,10)	PQR(0.25)	4.950	0	0	0.045	0.955
		PQR(0.5)	4.960	0	0	0.040	0.960
		PQR(0.75)	4.955	0.005	0.005	0.045	0.950
		PQR(0.5)-SCAD	4.935	0.005	0.005	0.040	0.955
	Cauchy	PQR(0.25)	4.920	0.025	0.020	0.065	0.915
		PQR(0.5)	4.925	0.005	0.005	0.065	0.930
		PQR(0.75)	4.920	0.010	0.010	0.070	0.920
		PQR(0.5)-SCAD	4.925	0.035	0.015	0.065	0.920
400	N(0,1)	PQR(0.25)	4.990	0	0	0.010	0.990
		PQR(0.5)	4.995	0	0	0.005	0.995
		PQR(0.75)	4.990	0	0	0.005	0.995
		PQR(0.5)-SCAD	4.995	0	0	0.005	0.995
	t(2)	PQR(0.25)	4.985	0	0	0.015	0.985
		PQR(0.5)	4.990	0	0	0.010	0.990
		PQR(0.75)	4.980	0	0	0.015	0.985
		PQR(0.5)-SCAD	4.980	0	0	0.010	0.990
	CN(0.1,10)	PQR(0.25)	4.995	0	0	0.005	0.995
		PQR(0.5)	4.995	0	0	0.005	0.995
		PQR(0.75)	4.990	0	0	0.010	0.990
		PQR(0.5)-SCAD	4.990	0	0	0.010	0.990
	Cauchy	PQR(0.25)	4.980	0	0	0.015	0.985
		PQR(0.5)	4.985	0	0	0.010	0.990
		PQR(0.75)	4.980	0	0	0.020	0.980
		PQR(0.5)-SCAD	4.985	0	0	0.015	0.985

of parameters over 500 simulations for model (11) with the errors of Normal and CN(0.1,10) distributions. Obviously, the empirical coverage rates are close to the true values for all cases as expected, this result confirms the validation of the proposed new method and is consistent to the asymptotic normalities of the parameter estimator established in Theorem 2.3.

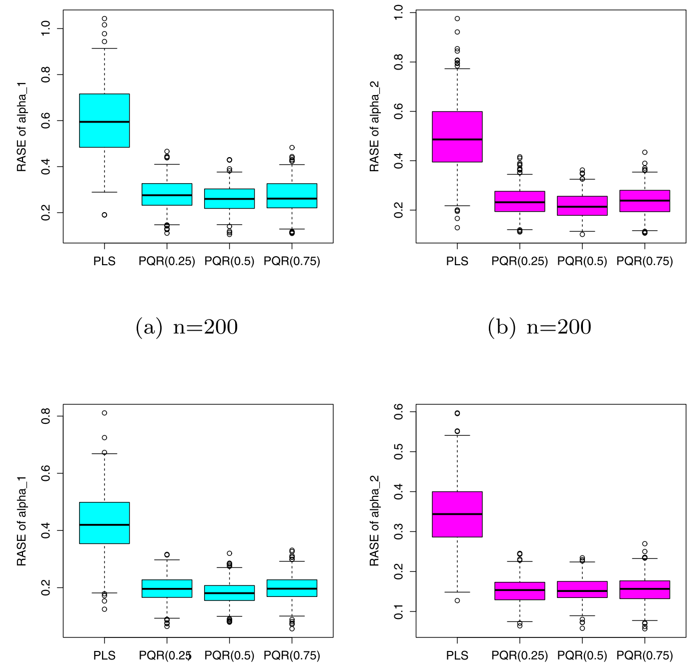
Example 2. The main purpose of this example is to evaluate the variable selection performance of our proposed adaptive penalized PQR procedure with $\tau = 0.25,\ 0.5,\ 0.75$. Similar model is considered as model (11) except that the covariates $X = (X_1, \ldots, X_7)^T$ comes from a multi-normal distribution with mean zero and $cov(X_i, X_j) = 0.5^{|i-j|}$ for $i, j = 1, \ldots, 7$, and the true parameter $\beta = (\beta_1, \ldots, \beta_7)^T = (1, 1.5, 0, 0, 0, 0, 0)^T$. As done in Example 1, four different error distributions for ε are also considered. Besides, the average number of the true zero coefficients correctly identified as zero (NC), the average number of the true nonzero coefficients erroneously identified as zero (NIC), the proportion of trials excluding at least one important variable in the selected final model (U.fit), the proportion of trials selecting all significant variables and at least including one noise variables (O.fit), and the proportion of trials selecting the exact sub-model (C.fit), are presented to show the performance of variable selection.

The corresponding results with 200 simulation runs are summarized in Table 6 and some observations can be obtained. Specifically, the penalized PQR procedure with dif-



(a) n=200 (b) n=200 (c) n=400 (d) n=400

Figure 1. The boxplots of RASEs for the varying coefficients in Example 2 when the error is t(2) distributed.



(a) n=200 (b) n=200 (c) n=400 (d) n=400

Figure 2. The boxplots of RASEs for the varying coefficients in Example 2 when the error is CN(0,1) distributed.

ferent quantiles performs quite well in correctly selecting the significant covariates as well as distinguishing the noise covariates, which indicates that the BIC selection strategy of tuning parameters presented in previous section works well. In addition, the method based on different quantiles perform similarly in each situation and all become better as the sample size increasing. Similar conclusions can also be observed from Figures 1 and 2, which present the boxplots of RASEs for the varying coefficients with the errors come from t(2) and CN(0,1) distributions, respectively. Note that, similar boxplots for the rest two errors can also be obtained, but we omit presenting them to reduce the length of this paper. Furthermore, we also present some simulation results of the proposed PQR estimator under SCAD penalty in Table 6. Since the estimation procedure is relatively time-consuming, we only consider the case of $\tau = 0.5$ and abbreviate the estimator as PQR(0.5)-SCAD. As we can see, although both the adaptive LASSO and SCAD penalty obtain some satisfactory results, the overall performance of adaptive LASSO appears to be a little better than SCAD in most cases. In a word, all these results corroborate the theoretical properties and reflect the robustness of our proposed method.

4.2 Real data analysis

As an illustration, we apply our proposed estimation and variable selection procedures to the Boston housing price data, which can be available freely from in R package. This data set contains 506 observations and has been analyzed by [4, 29, 16] via different semiparametric models. Following the previous studies, here we take the median value of owner-occupied homes in \$1000's (*medv*) as the response variable Y, and $\sqrt{lstat}$ as the index variable U, where *lstat* denotes the percentage of lower status of the population. The six covariate variables considered are per capita crime rate by town (*crim*), average number of rooms per dwelling (*rm*), full-value property tax per \$10,000 (*tax*), nitric oxides concentration per 10 million (*nox*), pupil-teacher ratio by town (*ptratio*) and proportion of owner-occupied units built prior to 1940 (*age*), which are denoted by Z_2, Z_3, Z_4, Z_5, Z_6 and Z_7, respectively. On the other hand, existing literature including Lesage and Pace [15] and Wang and Xue [29] have provided some evidence support for an exponential relationship between *medv* and the variables *ptratio* and *age*. Therefore, we analyze this data set based on the following varying coefficient partially nonlinear models

$$Y = \exp\left(X_1\beta_1 + X_2\beta_2\right) + \sum_{j=1}^{5} \alpha_j(U) Z_j + \varepsilon,$$

where $X_1 = Z_6$ and $X_2 = Z_7$, $Z_1 = 1$ corresponding to the baseline function.

In the beginning, we give a rough understanding of this data set through two simple figures, which include the boxplot and histogram of response Y presented in Figure 3

Table 7. The results of estimation and model fitting in Boston housing price data

Method	$\hat{\beta}_1$	$\hat{\beta}_2$	R^2	MAPE
PLS	-0.1634 (0.0258)	-0.0255 (0.0217)	0.7994 (0.1106)	0.2792 (0.0781)
PQR(0.25)	-0.1443 (0.0105)	-0.0509 (0.0091)	0.8322 (0.0531)	0.2049 (0.0446)
PQR(0.5)	-0.1418 (0.0093)	-0.0481 (0.0077)	0.8380 (0.0497)	0.2015 (0.0418)
PQR(0.75)	-0.1450 (0.0108)	-0.0528 (0.0089)	0.8339 (0.0514)	0.2060 (0.0453)

Enclosed in parentheses are the corresponding standard errors.

(a)-(b). As we can clearly seen that the distribution of Y is left-skewed and some outliers may be existed, thus robust statistical methods shall be preferred here for the sake of a convincing result. Next, we standardize both the response and covariate variables so that they have zero mean and unit sample standard deviation, and transform the index variable U such that its marginal distribution is U[0,1]. Particularly, U is transformed by taking a operation $\Phi(\cdot)$ on it with $\Phi(\cdot)$ being the cumulative distribution function (CDF) of a standard normal distribution. For a fair evaluation, we randomly pick up 400 samples from the data for modeling fitting, the rest 106 observations are used to assess the predictive power of the estimated model. A measurement of goodness of fit is given by R^2 with its definition as $R^2 = 1 - \mathrm{RSS}/\sum_{i\in S_1}(Y_i - \bar{Y})^2$ with $\bar{Y} = \sum_{i\in S_1} Y_i/|S_1|$, and a measurement of prediction is given by $\mathrm{MAPE} = \sum_{i\in S_2}(Y_i - \hat{Y}_i)/|S_2|$, where RSS is the residual sum of squares, $\hat{Y}_i$ is the fitted value of Y_i, $|S_1|$ and $|S_2|$ denote the modes of the training data set S_1 and the test data set S_2, respectively. Therefore, the lager of R^2 the better of model fitting, and the smaller of MAPE the better of prediction.

The corresponding results with 200 times operation are presented in Table 7, where the values in parentheses are their associated standard errors. As we can see from this table, all the considered procedures indicate that both *ptratio* and *age* have some effects and the same impact trend on Boston housing price, which is coincide with the discovery of [29]. On the other hand, our proposed PQR approach not only has a better performance in terms of model fitting, but also is much robust to the data set, because the PQR methods have larger values of R^2 and significantly smaller standard errors versus PLS. As to the performance of model prediction, similar results can also be observed from the boxplots of MAPE displayed in Figure 3 (c). These conclusions are reasonable and consistent to the theoretical findings. Moreover, the Shapiro-Wilk normal test proposed by Shapiro and Wilk [23] is further applied to study the applicability of PQR method, the obtained testing p-value is 2.2e-12, which significantly implies the non-normality of Boston housing data. Similar conclusion is also confirmed by the normal QQ-plot of residuals presented in Figure 3 (d). Consequently, taking into account of the robustness of estimation as well as performance of prediction, our proposed PQR method is preferred for analyzing this data set.

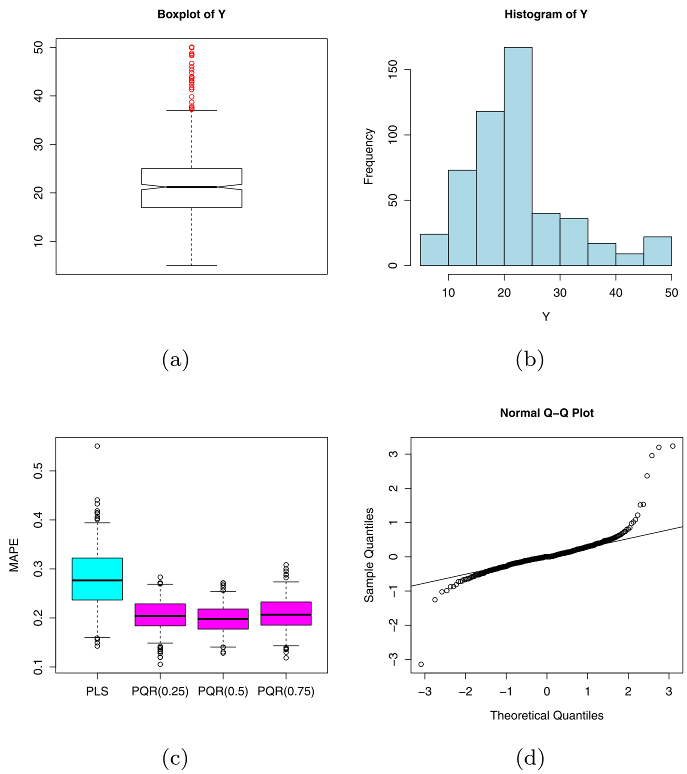


Figure 3. (a) and (b) are the histogram and boxplot of the response Y, respectively; (c) is the boxplots of MAPE for the Boston housing price data based on 200 time operations; (d) is the normal QQ-plot of residuals for the PQR(0.5) method.

5. CONCLUSION

In this paper, we focus on the robust estimation and variable selection methods for the varying coefficient partially nonlinear models. A novel profile-type nonlinear quantile regression approach is developed by minimizing the average check loss estimation procedure, and the asymptotic properties of the resulting estimators under some mild regularity conditions are derived. In addition, combining with the adaptive LASSO penalty, we consider a variable selection procedure for the parametric components and further demonstrate its oracle property. Finally, our limited numerical studies show the superiority of the proposed profile quantile regression method versus the well-known profile least squares approach in terms of robustness. For future study, one may consider the quantile VCPNLMs with complex data structures such as measurement errors, longitudinal data or missing data. The studies of above mentioned interesting questions are part of ongoing research work, but beyond the scope of this article.

APPENDIX

To establish the asymptotic properties and the oracle property of the proposed methods, the following regularity conditions are required.

(C1) The kernel function $K(\cdot)$ is a symmetric probability density function with compact support and satisfies Lipschitz condition.

(C2) The random variable U has a bounded support $\mathcal{U}$ and its marginal density function $f_U(\cdot)$ is Lipschitz continuous and bounded away from 0 on $\mathcal{U}$.

(C3) The covariates X and Z have bounded support. Define the matrices

$$\pi_1(u) = E\left\{f_Y\big(Q_\tau(X,Z,U)\big)ZZ^T \mid U=u\right\},$$
$$\pi_2(u) = E\left\{f_Y\big(Q_\tau(X,Z,U)\big)Zg'(X,\beta_0)^T \mid U=u\right\},$$

then $\pi_1(u)$ is positive defined for each $u \in \mathcal{U}$, $\pi_1(u)$, $\pi_1(u)^{-1}$ and $\pi_2(u)$ are all Lipschitz continuous.

(C4) All the true coefficient functions $\alpha_1(\cdot),\ldots,\alpha_q(\cdot)$ have continuous second order derivatives on $\mathcal{U}$.

(C5) For any x, $g(x,\beta)$ is a continuous function of β and the second derivative of $g(x,\beta)$ with respect to β exists and is continuous. Besides, the matrices $E\left\{g'(X,\beta)^{\otimes 2}\right\}$ and $E\left\{E\big(g'(X,\beta)\mid U\big)^{\otimes 2}\right\}$ are all bounded in a neighborhood of β, where $A^{\otimes 2} = AA^T$ for any matrix A.

(C6) There exists a large enough open subset $\mathcal{B} \in R^p$ that contains the true parameter β_0, such that for all x and any $\beta_1, \beta_2 \in \mathcal{B}$, the second derivative matrix $g''(x,\beta)$ satisfies

$$\|g''(x,\beta_1) - g''(x,\beta_2)\| \le R(x)\|\beta_1-\beta_2\|,$$
$$\left|\frac{\partial^2 g(x,\beta)}{\partial\beta_j\partial\beta_k}\right| \le H_{jk}(x)$$

for all $\beta \in \mathcal{B}$, with $E[R^2(x)] < \infty$, $E[H^2_{jk}(x)] < C_1 < \infty$ for all j,k.

In addition, we need the following two lemmas which will be frequently used in the sequel.

Lemma 1. *Suppose that $A_n(s)$ is convex and can be represented as $\frac{1}{2}s^TV_ns + J_n^Ts + C_n + r_n(s)$, where V_n is symmetric and positive definite, J_n is stochastically bounded, C_n is arbitrary and $r_n(s)\xrightarrow{p}0$ for each s. Then α_n, the minimizer of A_n, is only $o_p(1)$ away from $\beta_n = -V_n^{-1}J_n$, the minimizer of $\frac{1}{2}s^TV_ns + J_n^Ts + C_n$. If also $J_n \xrightarrow{d} J$, then $\alpha_n \xrightarrow{d} -V^{-1}J$.*

Lemma 2. *Let $(X_1,Y_1),\ldots,(X_n,Y_n)$ be i.i.d. random vectors, where Y_i's are scalar random variables, f denotes the joint density of (X,Y). Let K be a bounded positive function with bounded support, satisfying Lipschitz condition. Further assume that $\sup_{\mathbf{x}}\int |y|^r f(\mathbf{x},y)dy < \infty$ and $E|Y|^r < \infty$ for some $r > 0$. Then,*

$$\sup_{\mathbf{x}\in D}\left|\frac{1}{n}\sum_{i=1}^n\{K_h(X_i-\mathbf{x})Y_i - E[K_h(X_i-\mathbf{x})Y_i]\}\right| = O_p\left(\frac{\log^{1/2}(1/h)}{\sqrt{nh}}\right),$$

provided that $n^{2\varepsilon-1}h \to \infty$ for some $\varepsilon < 1-r^{-1}$ and D is the support set of $\mathbf{x}$.

The proofs of Lemma 1 and Lemma 2 can be referred to [7] and [20], respectively.

Proof of Theorem 2.1. For any given β, let $\hat{a}_\beta = \hat{\alpha}(u,\beta)$ and $\hat{b}_\beta = \hat{\alpha}'(u,\beta)$. Then we have

$$(\hat{a}_\beta^T, \hat{b}_\beta^T)^T = \arg\min_{(a,b)}\sum_{i=1}^n \rho_\tau\Big\{Y_i - g(X_i,\beta) - Z_i^T[a+bU_{i0}]\Big\}K(U_{i0}/h),$$

where $U_{i0} = U_i - u$. Denote by

$$\hat{\xi}_n = \sqrt{nh}\begin{pmatrix}\hat{a}_\beta - \alpha(u)\\ h(\hat{b}_\beta - \alpha'(u))\end{pmatrix}, \qquad A_i = \begin{pmatrix}Z_i\\ Z_iU_{i0}/h\end{pmatrix},$$

$$r_i(u) = -\big\{[g(X_i,\beta)-g(X_i,\beta_0)] + Z_i^T[\alpha(U_i)-\alpha(u)-\alpha'(u)U_{i0}]\big\}.$$

It is easy to demonstrate that $\hat{\xi}_n$ is also the minimizer of the following objective function

$$L_n(\xi_n) = \sum_{i=1}^n\Big\{\rho_\tau(\varepsilon_i - r_i(u) - \xi_n^TA_i/\sqrt{nh}) - \rho_\tau(\varepsilon_i - r_i(u))\Big\}K_i$$

with respect to ξ_n, where $K_i = K(U_{i0}/h)$.

Following the identity of Knight [11], that is

(12)
$$\rho_\tau(x-y) - \rho_\tau(x) = -y\psi_\tau(x) + \int_0^y [I(x\le s) - I(x\le 0)]ds,$$

where $\psi_\tau(x) = \tau - I(x \le 0)$. Based on some simple calculations, $L_n(\xi_n)$ can be rewritten as

(13) $$L_n(\xi_n) = -\xi_n^TW_n + B_n(\xi_n)$$

with the definitions of W_n and $B_n(\xi_n)$ are

$$W_n = \frac{1}{\sqrt{nh}} \sum_{i=1}^{n} K_i A_i \psi_\tau(\varepsilon_i)$$

$$B_n(\xi_n) = \sum_{i=1}^{n} K_i \int_{r_i(u)}^{r_i(u)+\xi_n^T A_i/\sqrt{nh}} [I(\varepsilon_i \le s) - I(\varepsilon_i \le 0)]ds.$$

Note that, the conditional expectation of $B_n(\xi_n)$ given U is

$$\begin{aligned} & E\{B_n(\xi_n) \mid U = u\} \\ = \ & \frac{1}{2} f_Y\left(Q_\tau(X, Z, U) \mid u\right) \xi_n^T \left[\frac{1}{nh}\sum_{i=1}^{n} K_i A_i A_i^T\right] \xi_n \\ & + \left\{\frac{1}{\sqrt{nh}} f_Y\left(Q_\tau(X, Z, U) \mid u\right) \sum_{i=1}^{n} K_i r_i(u) A_i\right\}^T \xi_n \\ & + o_p(1). \end{aligned}$$

In addition, as $B_n(\xi_n)$ is a summation of i.i.d. random variables of the kernel form, it follows from Lemma 2 and some simple calculations that

$$\begin{aligned} B_n(\xi_n) &= E\{B_n(\xi_n)\} + O_p\left(\log^{1/2}(1/h)/\sqrt{nh}\right) \\ &= E\{E[B_n(\xi_n) \mid u]\} + O_p\left(\log^{1/2}(1/h)/\sqrt{nh}\right) \\ &= \frac{1}{2}\xi_n^T \Omega(u) \xi_n + B_{n2}(\xi_n) + o_p(1), \end{aligned} \tag{14}$$

in which the definition of $B_{n2}(\xi_n)$ are

$$\begin{aligned} B_{n2}(\xi_n) = \ & E\Big\{\frac{1}{\sqrt{nh}} f_Y\left(Q_\tau(X, Z, U)\right) \cdot \\ & \sum_{i=1}^{n} K_i r_i(u) A_i \mid U = u\Big\}^T \xi_n. \end{aligned} \tag{15}$$

Applying the Taylor expansion to $r_i(u)$ yields $r_i(u) = g'(X_i, \beta_0)d_\beta - \frac{1}{2}U_{i0}^2 Z_i^T \alpha''(u) + O(U_{i0}^3 + \delta_\beta^2)$, then we have

$$\begin{aligned} & E\left\{\frac{1}{\sqrt{nh}} f_Y\left(Q_\tau(X, Z, U)\right) \sum_{i=1}^{n} K_i Z_i r_i(u) \mid u\right\} \\ = \ & -\frac{\sqrt{nh}}{2}\mu_2 h^2 f_U(u)\pi_1(u)\alpha''(u) + \sqrt{nh} f_U(u)\pi_2(u) d_\beta \\ & + O\left(\sqrt{nh}(h^4 + \delta_\beta^2)\right) \end{aligned} \tag{16}$$

and

$$\begin{aligned} & E\left\{\frac{1}{\sqrt{nh}} f_Y\left(Q_\tau(X, Z, U)\right) \sum_{i=1}^{n} K_i Z_i r_i(u) \frac{U_{i0}}{h} \mid u\right\} \\ = \ & O\left(\sqrt{nh}(h^3 + h\delta_\beta)\right), \end{aligned} \tag{17}$$

where $\pi_1(u)$ and $\pi_2(u)$ are defined in condition (C3).

Therefore, based on Equations (13)-(17), $L_n(\xi_n)$ can be expressed as

$$\begin{aligned} & L_n(\xi_n) = o_p(1) + \frac{1}{2}\xi_n^T \Omega(u)\xi_n - W_n^T \xi_n + \sqrt{nh} f_U(u) \times \\ & E\begin{pmatrix} \frac{-1}{2}\mu_2 h^2 \pi_1(u)\alpha''(u) + \pi_2(u) d_\beta + O(h^4 + \delta_\beta^2) \\ O(h^3 + h\delta_\beta) \end{pmatrix}^T \xi_n, \end{aligned}$$

It follows from Lemma 1 that the minimizer of $L_n(\xi_n)$ is

$$\begin{aligned} & \hat{\xi}_n = o_p(1) + \Omega(u)^{-1} W_n - \sqrt{nh} \times \\ & E\begin{pmatrix} \frac{-1}{2}\mu_2 h^2 \alpha''(u) + \pi_1(u)^{-1}\pi_2(u) d_\beta + O(h^4 + \delta_\beta^2) \\ O(h^3 + h\delta_\beta) \end{pmatrix}. \end{aligned}$$

Consequently, combing the expressions of $\hat{\xi}_n$ and W_n, we complete the proof of Theorem 2.1.

Proof of Theorem 2.2. To prove this theorem, it is sufficient to show that for any given small η, there exists a sufficiently large constant C such that

$$P\left\{\inf_{\|\mathbf{v}\|=C} \mathcal{D}_n(\beta_0 + n^{-1/2}\mathbf{v}) > \mathcal{D}_n(\beta_0)\right\} \ge 1 - \eta, \tag{18}$$

where $\mathbf{v}$ is a p-dimensional vector and

$$\begin{aligned} \mathcal{D}_n(\beta) = \ & \sum_{j=1}^{n}\sum_{i=1}^{n} \rho_\tau\Big\{Y_i - g(X_i, \beta) \\ & - Z_i^T[\hat{\alpha}(U_j) + \hat{\alpha}'(U_j)U_{ij}]\Big\} w_{ij}. \end{aligned}$$

Let $r_{ij} = Z_i^T[-\alpha(U_i) + \hat{\alpha}(U_j) + \hat{\alpha}'(U_j)U_{ij}]$, applying the Taylor expansion to $g(X_i, \beta_0 + n^{-1/2}\mathbf{v})$ yields

$$\begin{aligned} D_n(\mathbf{v}) &= \mathcal{D}_n(\beta_0 + n^{-1/2}\mathbf{v}) - \mathcal{D}_n(\beta_0) \\ &= \sum_{j=1}^{n}\sum_{i=1}^{n}\Big\{\rho_\tau(\varepsilon_i - r_{ij} - n^{-1/2}\mathbf{v}^T g'(X_i, \beta_0)) \\ & \quad - \rho_\tau(\varepsilon_i - r_{ij})\Big\} w_{ij}. \end{aligned}$$

Following the identity (12) and some similar arguments in the proof of Theorem 2.1, we can obtain that

$$D_n(\mathbf{v}) = \frac{1}{2}\mathbf{v}^T \Lambda \mathbf{v} + D_{n1}(\mathbf{v}) + D_{n2}(\mathbf{v}) + o_p(1), \tag{19}$$

where

$$\begin{aligned} & \Lambda = E\left\{f_Y\left(Q_\tau(X, Z, U)\right) g'(X, \beta_0) g'(X, \beta_0)^T \mid U\right\}, \\ & D_{n1}(\mathbf{v}) = -\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\sum_{i=1}^{n} w_{ij}\psi_\tau(\varepsilon_i) g'(X_i, \beta_0)^T \mathbf{v}, \\ & D_{n2}(\mathbf{v}) = E\Big\{\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\sum_{i=1}^{n} f_Y\left(Q_\tau(X, Z, U)\right) r_{ij} \times \\ & g'(X_i, \beta_0) w_{ij} \mid U_j\Big\}^T \mathbf{v}. \end{aligned}$$

Note that, r_{ij} can be expressed by the Taylor expansion as

$$r_{ij} = (Z_i^T, Z_i^T U_{ij}/h) \begin{pmatrix} \hat{\alpha}(U_j) - \alpha(U_j) \\ h\,(\hat{\alpha}'(U_j) - \alpha'(U_j)) \end{pmatrix} - \frac{1}{2} Z_i^T \alpha''(U_j) U_{ij}^2 + O(U_{ij}^3).$$

Substituting this expression into $D_{n2}(\mathbf{v})$ yields

$$D_{n2}(\mathbf{v}) = (D_{n21} + D_{n22})^T \mathbf{v} + O(h^3), \tag{20}$$

where

$$D_{n21} = E\Big\{ \frac{1}{\sqrt{n}} \sum_{j=1}^n \sum_{i=1}^n f_Y\left(Q_\tau(X,Z,U)\right) \left(Z_i^T, Z_i^T U_{ij}/h\right) \cdot \begin{pmatrix} \hat{\alpha}(U_j) - \alpha(U_j) \\ h\,(\hat{\alpha}'(U_j) - \alpha'(U_j)) \end{pmatrix} g'(X_i, \beta_0) w_{ij} \mid U_j \Big\},$$

$$D_{n22} = -E\Big\{ \frac{1}{\sqrt{n}} \sum_{j=1}^n \sum_{i=1}^n f_Y\left(Q_\tau(X,Z,U)\right) \frac{1}{2} Z_i^T \alpha''(U_j) U_{ij}^2 g'(X_i, \beta_0) w_{ij} \mid U_j \Big\}.$$

Based on the results of Theorem 2.1, we have

$$\begin{aligned} D_{n21} &= E\Big\{ \frac{1}{\sqrt{n}} \sum_{j=1}^n \sum_{i=1}^n f_Y\left(Q_\tau(X,Z,U) \mid U_j\right) \cdot \left(Z_i^T, Z_i^T U_{ij}/h\right) \begin{pmatrix} Q_{n1}(U_j) \\ Q_{n2}(U_j) \end{pmatrix} g'(X_i, \beta_0) w_{ij} \Big\} \\ &\quad + E\Big\{ \frac{1}{\sqrt{n}} \sum_{j=1}^n \sum_{i=1}^n f_Y\left(Q_\tau(X,Z,U) \mid U_j\right) Z_i^T \cdot \frac{1}{2} \mu_2 h^2 \alpha''(U_j) g'(X_i, \beta_0) w_{ij} \Big\} + o_p(1) \\ &\triangleq T_1 + T_2 + o_p(1). \end{aligned} \tag{21}$$

For the first term T_1, it follows from the expressions of $Q_{n1}(u)$ and $Q_{n2}(u)$ that T_1 is

$$\begin{aligned} & E\Big\{ \frac{1}{\sqrt{n}} \sum_{j=1}^n \sum_{i=1}^n \frac{w_{ij}\left(Z_i^T, Z_i^T U_{ij}/h\right)}{n f_U(U_j)} \begin{pmatrix} E(ZZ^T \mid U_j)^{-1} \\ E(ZZ^T \mid U_j)^{-1} \end{pmatrix} \cdot \sum_{k=1}^n K_h(U_{kj}) \psi_\tau(\varepsilon_k) \begin{pmatrix} Z_k \\ Z_k U_{kj}/(\mu_2 h) \end{pmatrix} g'(X_i, \beta_0) \Big\} \\ &= \frac{1}{\sqrt{n}} \sum_{k=1}^n \sum_{j=1}^n \psi_\tau(\varepsilon_k) w_{kj} E(g'(X, \beta_0) Z^T \mid U_j) \cdot E(ZZ^T \mid U_j) E(Z \mid U_j) + o_p(1), \end{aligned}$$

where the second equality holds by interchanging the summations and the symmetry of kernel function $K(\cdot)$. Thus,

$$\begin{aligned} & D_{n1}(\mathbf{v}) + T_1^T \mathbf{v} \\ &= o_p(1) - \frac{1}{\sqrt{n}} \sum_{i=1}^n \sum_{j=1}^n \psi_\tau(\varepsilon_i) w_{ij} \Big[g'(X_i, \beta_0) - E(g'(X, \beta_0) Z^T \mid U_j) E(ZZ^T \mid U_j) E(Z \mid U_j) \Big]^T \mathbf{v} \\ &\triangleq -\sqrt{n} P_{n1}^T \mathbf{v} + o_p(1), \end{aligned} \tag{22}$$

where

$$P_{n1} = -\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \psi_\tau(\varepsilon_i) w_{ij} \Big[g'(X_i, \beta_0) - E(g'(X, \beta_0) Z^T \mid U_j) E(ZZ^T \mid U_j) E(Z \mid U_j) \Big].$$

On the other hand, with some simple calculations based on the expressions of D_{n22} and T_2, we can verify that

$$D_{n22} + T_2 = o_p(1). \tag{23}$$

Consequently, combining Equations (19)-(23), $D_n(\mathbf{v})$ can be expressed as

$$D_n(\mathbf{v}) = \frac{1}{2} \mathbf{v}^T \Lambda \mathbf{v} - \sqrt{n} P_{n1}^T \mathbf{v} + (D_{n22} + T_2)^T \mathbf{v} + o_p(1). \tag{24}$$

Since it is not difficult to derive that $\sqrt{n} P_{n1}^T \mathbf{v} = O_p\big(\|\mathbf{v}\|\big)$ and $(D_{n22} + T_2)^T \mathbf{v} = o_p\big(\|\mathbf{v}\|\big)$, which means that $D_n(\mathbf{v})$ is dominated by the positive term $\frac{1}{2}\mathbf{v}^T \Lambda \mathbf{v}$ as long as C is large enough, then we have

$$P\left\{ \inf_{\|\mathbf{v}\|=C} \mathcal{D}_n(\beta_0 + n^{-1/2}\mathbf{v}) > \mathcal{D}_n(\beta_0) \right\} \geq 1 - \eta.$$

Therefore, (18) holds and we complete the proof of Theorem 2.2.

Proof of Theorem 2.3. The $\sqrt{n}$-consistency of $\hat{\beta}$ is derived in Theorem 2.2. In the next, we focus our attention on proving the asymptotic normality of $\hat{\beta}$. Using the similar notations as above and let $\hat{\beta}^* = \sqrt{n}(\hat{\beta} - \beta_0)$, it follows from the results of Theorem 2.2 that $\hat{\beta}^*$ is also the minimizer of the following objective function

$$R_n(\beta^*) = \sum_{j=1}^n \sum_{i=1}^n \Big\{ \rho_\tau(\varepsilon_i - r_{ij} - g'(X_i, \beta_0)^T \beta^* / \sqrt{n}) - \rho_\tau(\varepsilon_i - r_{ij}) \Big\} w_{ij}.$$

Based on some similar arguments as previous done, we can obtain that

$$R_n(\beta^*) = \frac{1}{2} {\beta^*}^T \Lambda \beta^* - \sqrt{n} P_{n1}^T \beta^* + (D_{n22} + T_2^*)^T \beta^* + o_p(1), \tag{25}$$

where

$$T_2^* = E\Big\{\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\sum_{i=1}^{n} f_Y\left(Q_\tau(X,Z,U)\mid U_j\right) Z_i^T \cdot \Big[\frac{1}{2}\mu_2 h^2\alpha''(U_j) - E(ZZ^T\mid U_j)^{-1}\cdot E(Zg'(X,\beta_0)^T\mid U_j)d_\beta\Big] g'(X_i,\beta_0)w_{ij}\Big\} + o_p(1). \tag{26}$$

Thus, do some basic calculations based on the expressions of D_{n22} and T_2^*, we have

$$\begin{aligned} & D_{n22}+T_2^* \\ = & -\frac{1}{\sqrt{n}}\sum_{j=1}^{n} f_Y\left(Q_\tau(X,Z,U)\mid U_j\right)\cdot E(g'(X,\beta_0)Z^T\mid U_j)E(ZZ^T\mid U_j)\cdot E(Zg'(X,\beta_0)^T\mid U_j)d_\beta + o_p(1) \\ = & -\sqrt{n}P_2(\beta-\beta_0)+o_p(1), \end{aligned} \tag{27}$$

where

$$P_2 = E\Big\{f_Y(Q_\tau(X,Z,U))E(g'(X,\beta_0)Z^T\mid U)\cdot E(ZZ^T\mid U)E(Zg'(X,\beta_0)^T\mid U)\Big\}.$$

Consequently, from Equations (25) and (27), $R_n(\beta^*)$ can be written as

$$R_n(\beta^*) = \frac{1}{2}{\beta^*}^T\Lambda\beta^* - \sqrt{n}[P_{n1}+P_2(\beta-\beta_0)]^T\beta^* + o_p(1).$$

Based on the conclusions of Lemma 1, we have

$$\hat\beta^* = \sqrt{n}\Lambda^{-1}P_{n1} + \sqrt{n}\Lambda^{-1}P_2(\beta-\beta_0)+o_p(1).$$

Combining this expression with the definition of $\hat\beta^*$ yields

$$\hat\beta-\beta_0 = \Lambda^{-1}P_{n1}+\Lambda^{-1}P_2(\beta-\beta_0)+o_p(1/\sqrt{n}). \tag{28}$$

Note that, by condition (C5) and the expressions of Λ and P_2, we can obtain that Λ, P_2 and $\Lambda - P_2$ are positive matrices. So $\tilde\Lambda = \Lambda^{-1/2}P_2\Lambda^{-1/2}$ is also a positive matrix with its eigenvalues are all less than 1.

Denote by $\tilde\beta^k$ be the estimator of the k-th iteration in our proposed procedure, then Equation (28) holds with $\hat\beta$ and β replaced by $\tilde\beta^{k+1}$ and $\tilde\beta^k$ for each k, respectively. Let $\tilde\gamma^k = \Lambda^{1/2}(\tilde\beta^k-\beta_0)$, then we have

$$\tilde\gamma^{k+1} = \Lambda^{-1/2}P_{n1}+\tilde\Lambda\tilde\gamma^k+o_p(1/\sqrt{n}).$$

By the fact that all the eigenvalues of $\tilde\Lambda$ are smaller than 1, thus the convergence of our proposed procedure can be guaranteed following a similar analysis in Xia and Härdle [31]. Specifically, for some sufficiently large k, we have

$$\Lambda^{1/2}(\hat\beta-\beta_0) = \Lambda^{-1/2}P_{n1}+\tilde\Lambda\Lambda^{1/2}(\hat\beta-\beta_0)+o_p(1/\sqrt{n})$$

holds, which is equivalent to

$$(\Lambda-\Lambda^{1/2}\tilde\Lambda\Lambda^{1/2})(\hat\beta-\beta_0) = P_{n1}+o_p(1/\sqrt{n}).$$

Finally, based on the Cramér-Wald device and the Center Limit Theorem, we can derive that the asymptotic distribution of $\hat\beta$ is (7). This competes the proof of Theorem 2.3.

Proof of Theorem 2.4. Combining the conclusion that $\hat\beta$ is a $\sqrt{n}$-consistent estimate of β_0 with the conditions (C3) and (C5), the asymptotic properties of $\hat\alpha(u,\hat\beta)$ can be established through a similar proof in Theorem 2.1, so we briefly give some major steps here. Let

$$\hat\zeta_n = \sqrt{nh}\begin{pmatrix}\hat\alpha(u,\hat\beta)-\alpha(u)\\ h(\hat\alpha'(u,\hat\beta)-\alpha'(u))\end{pmatrix},$$

$$\zeta_n = \sqrt{nh}\begin{pmatrix}\alpha(u,\hat\beta)-\alpha(u)\\ h(\alpha'(u,\hat\beta)-\alpha'(u))\end{pmatrix},$$

and $s_i(u) = -Z_i^T[\alpha(U_i)-\alpha(u)-\alpha'(u)U_{i0}]$. Then, we can verify that $\hat\zeta_n$ is also the minimizer of the following objective function

$$\Pi_n(\zeta_n) = \sum_{i=1}^{n}\Big\{\rho_\tau(\varepsilon_i - s_i(u)-\zeta_n^T A_i/\sqrt{nh}) - \rho_\tau(\varepsilon_i-s_i(u))\Big\}K_i.$$

Via some similar arguments in the proof of Theorem 2.1 leads to

$$\Pi_n(\zeta_n) = \frac{1}{2}\zeta_n^T\Omega\zeta_n - W_n^T\zeta_n + \sqrt{nh}f_U(u)\begin{pmatrix}-\frac{1}{2}\mu_2h^2\pi_1(u)\\ o_p(1)\end{pmatrix}^T\zeta_n + o_p(1).$$

Then, it follows from Lemma 1 that the minimizer of $\Pi_n(\zeta_n)$ is

$$\hat\zeta_n = \Omega^{-1}W_n - \sqrt{nh}f_U(u)\begin{pmatrix}-\frac{1}{2}\mu_2h^2\pi_1(u)\\ o_p(1)\end{pmatrix}+o_p(1).$$

In addition, based on some basic calculations, we can prove that $W_n \xrightarrow{d} N\big(0,\tau(1-\tau)\Omega_0(u)\big)$. Therefore, from the definition of $\hat\zeta_n$, we have

$$\sqrt{nh}\left\{\begin{pmatrix}\hat\alpha(u,\hat\beta)-\alpha(u)\\ h(\hat\alpha'(u,\hat\beta)-\alpha'(u))\end{pmatrix} - \frac{1}{2}\mu_2h^2\Omega^{-1}(u)\Gamma(u)\alpha''(u)\right\}$$

$$\xrightarrow{d} N\Big(0,\ \tau(1-\tau)\Omega^{-1}(u)\Omega_0(u)\Omega^{-1}(u)\Big)$$

holds, where $\Omega(u)$, $\Omega_0(u)$ and $\Gamma(u)$ are defined in Theorem 2.4. This completes the proof.

Proof of Theorem 3.1. We first prove the asymptotic normality of $\hat\beta^\lambda$. Similar to the first part proof of Theorem 2.3, we can show that there exists a local minimizer $\hat\beta^\lambda$ of $\Phi_n(\lambda,\beta)$ satisfying $\|\hat\beta^\lambda-\beta_0\|=O_p(1/\sqrt{n})$. Let $\hat\theta=\sqrt{n}(\hat\beta^\lambda-\beta_0)$, then $\hat\theta$ is also the minimizer of

$$\begin{aligned}&\Psi_n(\theta)\\ =&\sum_{j=1}^n\sum_{i=1}^n\Big\{\rho_\tau(\varepsilon_i-r_{ij}-g'(X_i,\beta_0)^T\theta/\sqrt{n})\\ &-\rho_\tau(\varepsilon_i-r_{ij})\Big\}w_{ij}+\lambda_n\sum_{k=1}^p\frac{1}{|\hat\beta_k|^2}\left(|\beta_{0k}+\theta_k/\sqrt{n}|-|\beta_{0k}|\right).\end{aligned}$$

It follows from some similar arguments in Theorem 2.3 that

$$\begin{aligned}\Psi_n(\theta)&=\frac12\theta^T\Lambda\theta-\sqrt{n}[P_{n1}+P_2(\beta-\beta_0)]^T\theta\\ &+\lambda_n\sum_{k=1}^p\frac{1}{|\hat\beta_k|^2}\left(|\beta_{0k}+\theta_k/\sqrt{n}|-|\beta_{0k}|\right)+o_p(1).\end{aligned}$$

Now, considering the penalty term of $\Psi_n(\theta)$. For any $k=1,2,\ldots,l$, we have $\beta_{0k}\neq0$, $|\hat\beta_k|^2\xrightarrow{p}|\beta_{0k}|^2$ and $\sqrt{n}\left(|\beta_{0k}+\theta_k/\sqrt{n}|-|\beta_{0k}|\right)\to\theta_k\cdot\mathrm{sgn}(\beta_{0k})$. From the condition $\lambda_n/\sqrt{n}\to0$ and slutsky's theorem, so

$$\begin{aligned}&\frac{\lambda_n}{|\hat\beta_k|^2}\left(|\beta_{0k}+\theta_k/\sqrt{n}|-|\beta_{0k}|\right)\\ =&\frac{\lambda_n}{\sqrt{n}|\hat\beta_k|^2}\sqrt{n}\left(|\beta_{0k}+\theta_k/\sqrt{n}|-|\beta_{0k}|\right)\to0.\end{aligned}$$

Besides, for $k=l+1,l+2,\ldots,p$, we have $\beta_{0k}=0$ and $\sqrt{n}\left(|\beta_{0k}+\theta_k/\sqrt{n}|-|\beta_{0k}|\right)=|\theta_k|$. It follows from the condition $\lambda_n\to\infty$ that $\frac{\lambda_n}{|\hat\beta_k|^2}\left(|\beta_{0k}+\theta_k/\sqrt{n}|-|\beta_{0k}|\right)\to\infty$ for any $\theta_k\neq0$. Therefore, we have

$$\begin{aligned}&\frac{\lambda_n}{|\hat\beta_k|^2}\left(|\beta_{0k}+\theta_k/\sqrt{n}|-|\beta_{0k}|\right)\\ \xrightarrow{d}&\begin{cases}0, & \text{if } \beta_{0k}\neq0\\ 0, & \text{if } \beta_{0k}=0 \text{ and } \theta_k=0\\ \infty, & \text{otherwise}\end{cases}\end{aligned}$$

holds, which indicates $\Psi_n(\theta)\xrightarrow{d}\Psi_n^*(\theta)$ hold and $\Psi_n^*(\theta)$ is

$$\begin{cases}\frac12\theta_I^T\Lambda_I\theta_I-\sqrt{n}\rho^T\theta_I, & \text{if } \theta_k=0 \text{ for } k=l+1,\ldots,p,\\ \infty, & \text{otherwise,}\end{cases}$$

where $\rho=P_{n1I}+P_{2I}(\beta_I-\beta_{0I})$, θ_I, Λ_I, P_{n1I}, P_{2I}, β_I and β_{0I} are the correspondingly first l components or top-left l-by-l submatrices of θ, Λ, P_{n1}, P_2, β and β_0, respectively.

Note that $\Psi_n(\theta)$ is convex and $\Psi_n^*(\theta)$ has an unique minimum, so from the epi-convergence results of Geyer [6] and Knight and Fu [12], the asymptotic normality part of Theorem 3.1 is proved by some similar arguments as done in Theorem 2.3.

In the following, we devoted to proving the consistency of model selection. To this end, we only need to show that $P(\hat\beta_{II}^\lambda=0)\to1$, which is equivalent to prove that $P(\hat\beta_k^\lambda\neq0)\to0$ if $\beta_{0k}=0$ for $k=l+1,\ldots,p$. Recall that $\hat\beta^\lambda$ minimize $\Phi_n(\lambda,\beta)$, if $\beta_{0k}=0$ but $\hat\beta_k^\lambda\neq0$, we must have

$$\begin{aligned}&\sum_{j=1}^n\sum_{i=1}^n\rho_\tau\Big\{Y_i-g_k'(X_i,\hat\beta^\lambda)\hat\beta_k^\lambda-Z_i^T[\hat\alpha(U_j)+\\ &\hat\alpha'(U_j)U_{ij}]\Big\}w_{ij}+\frac{\lambda_n}{|\hat\beta_k|^2}|\hat\beta_k^\lambda|\\ \leq&\sum_{j=1}^n\sum_{i=1}^n\rho_\tau\left\{Y_i-Z_i^T[\hat\alpha(U_j)+\hat\alpha'(U_j)U_{ij}]\right\}w_{ij},\end{aligned}$$

where $g_k'(X_i,\hat\beta^\lambda)$ represents the k-th component of $g'(X_i,\hat\beta^\lambda)$. Taking into account of the inequality that

$$\left|\frac{\rho_\tau(x_1)-\rho_\tau(x_2)}{x_1-x_2}\right|\leq\max(\tau,1-\tau)\leq1,$$

then we have

$$\frac{\lambda_n}{\sqrt{n}|\hat\beta_k|^2}\leq\frac{1}{\sqrt{n}}\sum_{j=1}^n\sum_{i=1}^n|g_k'(X_i,\hat\beta^\lambda)|w_{ij}.$$

This implies that the following inequality holds, that is

$$P(\hat\beta_k^\lambda\neq0)\leq P\left(\frac{\lambda_n}{\sqrt{n}|\hat\beta_k|^2}\leq\frac{1}{\sqrt{n}}\sum_{j=1}^n\sum_{i=1}^n|g_k'(X_i,\hat\beta^\lambda)|w_{ij}\right).$$

Obviously, $P(\hat\beta_k^\lambda\neq0)\to0$ holds. In fact, since $k=l+1,\ldots,p$ here, we have $\frac{\lambda_n}{\sqrt{n}|\hat\beta_k|^2}\to\infty$, whereas $\frac{1}{\sqrt{n}}\sum_{j=1}^n\sum_{i=1}^n|g_k'(X_i,\hat\beta^\lambda)|w_{ij}$ is bounded from conditions (C5) and (C6), hence

$$P\left(\frac{\lambda_n}{\sqrt{n}|\hat\beta_k|^2}\leq\frac{1}{\sqrt{n}}\sum_{j=1}^n\sum_{i=1}^n|g_k'(X_i,\hat\beta^\lambda)|w_{ij}\right)\to0.$$

Consequently, for $k=l+1,\ldots,p$, we can prove that $P(\hat\beta_k^\lambda\neq0)\to0$ holds, which is equivalent to $P(\hat\beta_k^\lambda=0)\to1$. This completes the proof.

ACKNOWLEDGEMENTS

The authors are grateful to the Editor, Associate Editor and two anonymous referees whose constructive comments lead to a significant improvement of the paper. Jing Yang's research was partially supported by the National Natural Science Foundation of China (Grant 11801168) and the Natural Science Foundation of Hunan Province (Grant 2018JJ3322). Fang Lu's research was partially supported by the National Natural Science Foundation of China (Grants 11801169, 11571055). Guoliang Tian's research was partially supported by the National Natural Foundation of China

(Grant 11771199). Xuewen Lu's research was supported by Discovery Grants (RG/PIN261567-2013) from National Science and Engineering Council of Canada. Hu Yang's research was partially supported by the National Natural Science Foundation of China (Grant 11671059).

Received 3 June 2018

REFERENCES

[1] Breiman, L. (1995). Better subset regression using the Nonnegative Garrote. *Technometrics* **37** 373–384. MR1365720

[2] Fan, G., Liang, H. and Shen, Y. (2016). Penalized empirical likelihood for high-dimensional partially linear varying coefficient model with measurement errors. *J. Multivariate Anal.* **147** 183–201.

[3] Fan, G., Liang, H. and Wang, J. (2013). Statistical inference for partially time-varying coefficient errors-in-variables models. *J. Statist. Plann. Inference.* **143** 505–519. MR2995111

[4] Fan, J. and Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli* **11** 1031–1057. MR2189080

[5] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

[6] Geyer, C. (1994). On the asymptotics of constrained M-estimation. *Ann. Statist.* **22** 1993–2010.

[7] Hjort, N. and Pollard, D. (1993). Asymptotics for minimisers of convex processes. *Statistical Research Report*, Yale University.

[8] Jiang, R., Zhou, Z., Qian, M. and Chen, Y. (2013). Two step composite quantile regression for single-index models. *Comput. Statist. Data. Anal.* **64** 180–191.

[9] Jiang, Y., Ji, Q. and Xie, B. (2017). Robust estimation for the varying coefficient partially nonlinear models. *J. Comput. Appl. Math.* **326** 31–43.

[10] Kai, B., Li, R. and Zou, H. (2011). New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models. *Ann. Statist.* **39** 305–332.

[11] Knight, K. (1998). Limiting distributions for L_1 regression estimators under general conditions. *Ann. Statist.* **26** 755–770. MR1626024

[12] Knight, K. and Fu, W. (2000). Asymptotics for Lasso-type estimators. *Ann. Statist.* **28** 1356–1378.

[13] Koenker, R. (2005). Quantile Regression. Cambridge University Press, Cambridge.

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

[15] Lesage, J. and Pace, R. (2009). Introduction to Spatial Econometrics, CRC Press, Boca Raton. MR2485048

[16] Li, T. and Mei, C. (2013). Estimation and inference for varying coefficient partially nonlinear models. *J. Statist. Plann. Inference.* **143** 2023–2037.

[17] Li, Y., Li, G., Lian, H. and Tong, T. (2017). Profile forward regression screening for ultra-high dimensional semiparametric varying coefficient partially linear models. *J. Multivariate Anal.* **155** 133–150.

[18] Lv, Y., Zhang, R., Zhao, W. and Liu, J. (2015). Quantile regression and variable selection of partial linear single-index model. *Ann. Inst. Statist. Math.* **67** 375–409.

[19] Ma, S. and He, X. (2016). Inference for single-index quantile regression models with profile optimization. *Ann. Statist.* **44** 1234–1268. MR3485959

[20] Mack, Y. and Silverman, B. (1982). Weak and strong uniform consistency of kernel regression estimates. *Probab. Theory Related Fields.* **61** 405–415.

[21] Qian, Y. and Huang, Z. (2016). Statistical inference for a varying-coefficient partially nonlinear model with measurement errors. *Stat. Methodol.* **32** 122–130.

[22] Ruppert, D., Sheather, S. J. and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.* **90** 1257–1270.

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

[24] Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

[25] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B* **58** 267–288.

[26] Wang, D. and Kulasekera, K. B. (2012). Parametric component detection and variable selection in varying-coefficient partially linear models. *J. Multivariate Anal.* **112** 117–129.

[27] Wang, H. and Leng, C. (2007). Unified lasso estimation via least squares approximation. *J. Amer. Statist. Assoc.* **102** 1039–1048.

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

[29] Wang, Q. and Xue, L. (2011). Statistical estimation in partially-varying-coefficient single-index models. *J. Multivariate Anal.* **102** 1–19.

[30] Wu, W. and Zhou, Z. (2017). Nonparametric Inference for Time-Varying Coefficient Quantile Regression. *J. Bus. Econom. Statist.* **35** 98–109.

[31] Xia, Y. and Härdle, W. (2006). Semi-parametric estimation of partially linear single-index models. *J. Multivariate Anal.* **97** 1162–1184.

[32] Xiao, Y. and Chen, Z. (2018). Bias-corrected estimations in varying-coefficient partially nonlinear models with measurement error in the nonparametric part. *J. Appl. Stat.* **45** 586–603.

[33] Yang, H. and Yang, J. (2014). The adaptive L1-penalized LAD regression for partially linear single-index models. *J. Statist. Plann. Inference.* **151** 73–89.

[34] Yang, J. and Yang, H. (2016). Smooth-threshold estimating equations for varying coefficient partially nonlinear models based on orthogonality-projection method. *J. Comput. Appl. Math.* **302** 24–37.

[35] Yu, K. and Jones, M. (1998). Local linear quantile regression. *J. Amer. Statist. Assoc.* **93** 228–237.

[36] Zhao, W., Zhang, R. and Liu, J. (2014). Regularization and model selection for quantile varying coefficient model with categorical effect modifiers. *Comput. Statist. Data. Anal.* **79** 44–62.

[37] Zhou, X., Zhao, P. and Wang, X. (2017). Empirical likelihood inferences for varying coefficient partially nonlinear models. *J. Appl. Stat.* **44** 474–492.

[38] Zou, H. (2006). The adaptive LASSO and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

Jing Yang
Key Laboratory of High Performance Computing
and Stochastic Information Processing
(Ministry of Education of China)
College of Mathematics and Statistics
Hunan Normal University
Changsha 410081, Hunan Province
China
E-mail address: yj@hunnu.edu.cn

Fang Lu
Key Laboratory of High Performance Computing
and Stochastic Information Processing
(Ministry of Education of China)
College of Mathematics and Statistics
Hunan Normal University
Changsha 410081, Hunan Province
China
E-mail address: lufang8771030@163.com

Guoliang Tian
Department of Mathematics
Southern University of Science and Technology
Shenzhen 518055, Guangdong Province
China
E-mail address: tiangl@sustc.edu.cn

Xuewen Lu
Department of Mathematics and Statistics
University of Calgary
2500 University Drive NW
Calgary, AB T2N 1N4
Canada
E-mail address: xlu@ucalgary.ca

Hu Yang
College of Mathematics and Statistics
Chongqing University
Chongqing 401331
China
E-mail address: yh@cqu.edu.cn