Published in final edited form as:

Commun Stat Theory Methods. 2017 ; 46(3): 1031–1049. doi:10.1080/03610926.2015.1010009.

Consistent model identification of varying coefficient quantile regression with BIC tuning parameter selection

Qi Zheng and Limin Peng

Department of Biostatistics and Bioinformatics, Emory University Atlanta GA, 30322, USA

Abstract

Quantile regression provides a flexible platform for evaluating covariate effects on different segments of the conditional distribution of response. As the effects of covariates may change with quantile level, contemporaneously examining a spectrum of quantiles is expected to have a better capacity to identify variables with either partial or full effects on the response distribution, as compared to focusing on a single quantile. Under this motivation, we study a general adaptively weighted LASSO penalization strategy in the quantile regression setting, where a continuum of quantile index is considered and coefficients are allowed to vary with quantile index. We establish the oracle properties of the resulting estimator of coefficient function. Furthermore, we formally investigate a BIC-type uniform tuning parameter selector and show that it can ensure consistent model selection. Our numerical studies confirm the theoretical findings and illustrate an application of the new variable selection procedure.

Keywords

Bayesian information criterion; Shrinkage estimation; Varying covariate effects; Quantile regression

1 Introduction

Quantile regression (Koenker and Bassett, 1978) offers a useful tool to investigate the varying covariate effect model, which has emerged to deal with applications with considerable inhomogeneous associations arising from various scientific fields. By allowing the covariates effects to change smoothly over quantile indices, quantile regression does not involve estimation on kernel smoothing, which is required by many existing procedures (e.g., Wang and Xia, 2009; Kai et al., 2011; Wang and Kulasekera, 2012), and hence can preserve the desired $\sqrt{n}$ -consistency.

In contemporary statistical analysis, the challenge of analyzing the varying covariate effect model is often entangled with variable selection problems. Lots of literature (e.g., Zou and Yuan, 2008; Wu and Liu, 2009; Zheng et al., 2013; Wang et al., 2007a, 2012) incorporated penalization into quantile regression to study variable selection and heterogeneous associations simultaneously. However, these methods select variables based on their effects on a single or multiple prespecified quantiles. It is easy to perceive that they can miss important variables whose impact range does not include the target quantile indices.

Retaining these variables would contribute a better understanding of how covariates impact response distribution.

Motivated by this consideration, Peng et al. (2014) proposed a uniform adaptive-LASSO (UAL) varying coefficient quantile regression over a spectrum of quantiles, which was designed to remove only covariates with no effect on any quantile index of interest. Their key strategy, following the idea of adaptive-LASSO (Zou, 2006), is to penalize the estimation by the maximum departure of the coefficient function from zero. They show that the UAL estimator is able to (i) correctly identify all zero and nonzero coefficient functions with probability approaching 1; (ii) converge weakly to a Gaussian process in the same manner as if all irrelevant covariates were precluded. That is, the UAL estimator enjoys the oracle properties (Fan and Li, 2001). However, there still remains two important questions: (1) Is there a general type of adaptive weights to achieve oracle properties? (2) The numerical studies in Peng et al. (2014) suggest that a BIC-type tuning parameter selector generally yields satisfactory rates of correct model identification. Is there any formal justification for the BIC-type selector?

In this work, we attempt to address those questions. First, we generalize the UAL estimator by allowing for a more general class of adaptive weights, which now includes the classical adaptive weight. The specific constraints on the proposed adaptive weights are elaborated in Section 2. The general adaptively weighted LASSO estimator, like the UAL estimator, enjoys the oracle properties uniformly across all quantile indices of interest. Second and more importantly, we formally investigate a BIC-type tuning parameter selector for the new estimator. We prove that the proposed BIC selector ensures the true model identification uniformly across all quantile levels under consideration. Our numerical studies confirm our theoretical findings and support the utility of our proposal.

It is remarkable that our proposed BIC selector is not a trivial extension of BIC proposed by Wang et al. (2007b). Wang and Leng (2007) proposed the unified lasso via least squares approximation (LSA) to study different types of lasso objective functions including single quantile regression. They showed that LSA is able to transfer those functions into their asymptotically equivalent least-squares problems, and the BIC selector for LSA can consistently identify the true model. However, the varying coefficient quantile can not be expressed as a least square approximation. Moreover, it is not clear what properties the BIC owns across a compact quantile index set. Our BIC selector is also quite distinct from the one recently proposed by Lee et al. (2014), which only considers selecting tuning parameter at a single quantile. As reported in our simulation studies, the varying coefficient quantile regression with BIC conducted pointwisely (i.e. separately at each quantile index) can yield a severely overfitted model. To our best knowledge, little work in literature has concerned the properties of BIC in varying coefficient quantile regression over a spectrum of quantiles.

In this work, we focus on the situation where the number of covariates is fixed. Nevertheless, our proposal can be extended to the high dimensional scenario where the number of covariates can be much larger than the sample size. Our efforts in both directions suggest that the consistency of the BIC selector can be achieved under weaker assumptions

in the fixed p case as compared to the high dimensional scenario, along with a better estimation consistency rate. Therefore, it is worth a separate report on the fixed p case.

The rest of the paper is organized as follows. In Section 2, we propose a general adaptively weighted LASSO estimator and discuss its asymptotic properties. A BIC-type tuning parameter selector and its properties are presented in Section 3. Numerical studies are reported in Section 4. We provide a short discussion and concluding remarks in Section 5, and relegate all technical details to the Appendix.

2 The general adaptively weighted LASSO estimator for varying coefficient quantile regression

We consider the linear varying coefficient quantile model,

$$Q_Y(\tau|\mathbf{X}) = \mathbf{X}'\boldsymbol{\beta}_0(\tau), \quad \tau \in \Delta := (\tau_L, \tau_U), \quad (1)$$

where Y is a response variable, $\mathbf{X} = (1, \mathbf{Z}')'$ is a $p \times 1$ vector of covariates,

$\boldsymbol{\beta}_0(\tau) = (\alpha_0(\tau), \boldsymbol{\beta}'_0(\tau))'$ is a $p \times 1$ vector unknown coefficients vector, $Q_Y(\tau|\mathbf{X}) := \inf\{y: P(Y \leq y|\mathbf{X}) = \tau\}$ denotes the τ th quantile of Y given $\mathbf{X}$, and (τ_L, τ_U) indicates the range of targeted quantile indices with $0 < \tau_L < \tau_U < 1$. Here $\alpha_0(\tau)$ and $\boldsymbol{\beta}_0(\tau)$ denote the intercept and covariate effects on the τ th quantile of Y respectively.

Contemporaneously examining covariate effects on a spectrum of quantiles as in model (1) necessitates a different perspective for variable selection. In model (1), some covariates may have strong effects throughout all $\tau \in \Delta$ (i.e. full effects), while some may only influence part of (but not all) quantiles considered (i.e. partial effects). It is important to retain both types of variables to achieve a complete picture about the covariate-response relationship. Thus, the true model S_T over Δ is defined as follows:

$$S_T := \{j: \exists \tau \in \Delta, \beta_0^{(j)}(\tau) \neq 0, j \geq 2\}. \quad (2)$$

Here and hereafter, $u^{(j)}$ and $u^{(1:j)}$ denote the j th component and the first j component of the vector $\mathbf{u}$ respectively. Our aim is to consistently identify S_T , all variables that impact the conditional distribution over Δ .

The choice of Δ reflects the segment of conditional distribution of interest. For example, if the interest is to identify variables that impact the upper quantiles, one can choose $\Delta = (0.7, 0.9)$; if one intends to have an overall understanding of the influence of covariates on the response distribution, Δ may be chosen to be $(0.1, 0.9)$. In our numerical studies, we set $\Delta = (0.1, 0.9)$.

Let $(Y_i, \mathbf{X}_i)$, $i = 1, \dots, n$ be a random sample from (1). We propose a general adaptively weighted LASSO estimator for model (1), defined as

$$\hat{\beta}_{\lambda_n}(\tau) = \arg \min_{\beta} \sum_{i=1}^n \rho_{\tau}(Y_i - \mathbf{X}_i' \beta) + \lambda_n \sum_{j=2}^p \omega_j(\tau) |\beta^{(j)}|, \quad (3)$$

where $\rho_{\tau}(t) := t(\tau - 1\{t < 0\})$ is the loss function, λ_n is the tuning parameter which does not depend on τ , and $\omega_j(\tau)$ is a nonnegative penalty assigned to $\beta^{(j)}$. The adaptive weight $\omega_j(\tau)$ is required to satisfy the following conditions: **(A1)** $\sup_{\tau \in \Delta} \omega_j(\tau) = O_p(1)$ for $j \in S_T$; **(A2)** $1/\inf_{\tau \in \Delta} \omega_j(\tau) = O_p(n^{-1/2} \log \log n)$ for $j \notin S_T$. The resulting model is

$$S_{\lambda_n} = \{2 \leq j \leq p : \sup_{\tau \in \Delta} |\hat{\beta}_{\lambda_n}^{(j)}(\tau)| \neq 0\}.$$

The general specification of the adaptive weight function adds flexibility to the proposed shrinkage estimator. Let $\tilde{\beta}(\tau)$ be the unpenalized quantile regression estimator at level τ . When $\omega_j(\tau) = 1/(\sup_{\tau \in \Delta} |\tilde{\beta}^{(j)}(\tau)|)$, the proposed estimator reduces to the UAL estimator (Peng et al., 2014). It can be shown that the classical adaptive weight, $1/|\tilde{\beta}^{(j)}(\tau)|$, is also a qualified adaptive weight function. Another useful special case is $\omega_j(\tau) = 1/(\int_{\tau \in \Delta} |\tilde{\beta}^{(j)}(\tau)| d\tau)$, which indicates penalization based on the average signal strength over $\tau \in \Delta$ (Volgushev et al., 2012).

It is worth emphasizing that the proposed shrinkage estimator of $\beta_0(\cdot)$ adopts a common tuning parameter λ_n , not depending on τ . This constitutes a major distinction between the proposed penalization approach and the naive pointwise penalization approach, which is to fit penalized quantile regression separately at each $\tau \in \Delta$ (including tuning parameter selection). The common tuning parameter selection plays a key role in achieving a consistent identification of all relevant variables in model (1). This is evidenced by our simulation studies, where we compare the naive pointwise adaptive-LASSO estimator with the proposed estimator with the classic adaptive weight. The former approach demonstrates a strong tendency to select an overfitted model (1), while the latter one demonstrates a high likelihood of correctly identifying the true model.

2.1 Technical conditions

The following technical conditions are required to study the theoretical properties of the proposed penalized varying coefficient quantile regression procedure and the uniform BIC-type tuning parameter selector.

C1. (i) $E(\mathbf{X}\mathbf{X}') > 0$; (ii) $\beta_0(\tau)$ is Lipschitz in τ for $\tau \in \Delta$; (iii) $E[|\boldsymbol{\delta}'\mathbf{X}|^3]/\|\boldsymbol{\delta}\|^3$ is bounded by some positive constant q for any $\boldsymbol{\delta} \neq 0$, where $\|\cdot\|$ denotes the Euclidean norm.

C2. (i) For some $A_0 > 0$ and $\delta_1 > 0$, $\sup_{\tau \in \Delta} \int_{\mathcal{X}} |f_{\tau}(tx) - f_{\tau}(0|x)| A_0 |t| dx < \delta_1$ for $|t| < \delta_1$. Here $f_{\tau}(tx)$ denotes the density $Y - \beta_0'(\tau)\mathbf{X}$ given $\mathbf{X} = \mathbf{x}$; (ii) $\sup_{\tau \in \Delta} \int_{\mathcal{X}} f_{\tau}(tx) dx < \bar{f}$ and $\inf_{\tau \in \Delta} \int_{\mathcal{X}} f_{\tau}(0|x) dx > \underline{f} > 0$.

C1(i) is crucial for the identifiability of $\beta_0(\tau)$, and is commonly adopted in literature (for example, Wang et al., 2007b; Peng et al., 2014). C1(ii) simplifies our arguments based on empirical process, and more importantly it ensures that we can approximate the entire

coefficient curve $\beta_0(\tau)$ by a sufficiently fine grid over $\tau \in \Delta$. C2(i) imposes a mild smoothness condition on the conditional density, which is crucial for establishing the uniform consistency of $\hat{\beta}(\tau)$ across $\tau \in \Delta$. It is worthy pointing out that C2(i) does not require any homoscedasticity assumption. The assumption that the conditional density is bounded below by a positive constant in C2(ii) is often met in practice.

2.2 Asymptotic Properties

Without loss of generality, we assume $\sup_{\tau \in \Delta} |\beta_0^{(j)}| \neq 0$ for $2 \leq j \leq s$ and $\sup_{\tau \in \Delta} |\beta_0^{(j)}| = 0$ for $j > s$. Then $S_T = \{2, \dots, s\}$ with some unknown s .

Theorem 1—Assume that $(n^{1/2} \log \log n)^{-1} \lambda_n = O(1)$. If $\omega_j(\tau)$'s satisfy $\sup_{\tau \in \Delta} \omega_j(\tau) = O_p(1)$ for $2 \leq j \leq s$ and $1/\inf_{\tau \in \Delta} \omega_j(\tau) = O_p(n^{-1/2} \log \log n)$ for $s+1 \leq j \leq p$, then we have $\sup_{\tau \in \Delta} \|\hat{\beta}_{\lambda_n}(\tau) - \beta_0(\tau)\| = O_p(n^{-1/2} \log \log n)$, under the assumptions C1 and C2.

Theorem 1 states that the supreme of the deviation of $\hat{\beta}_{\lambda_n}(\tau)$ from the true $\beta_0(\tau)$ is bounded by $n^{-1/2} \log \log n$ in probability, which implies the uniform consistency of $\hat{\beta}_{\lambda_n}(\tau)$. The uniform upper bound further facilitates us to establish the following oracle properties.

Theorem 2—Assume that $\lim_{n \rightarrow \infty} (\log \log n)^{-2} \lambda_n = \infty$ and $\lim_{n \rightarrow \infty} n^{-1/2} \lambda_n = 0$. Under the same assumptions in Theorem 1, we have

i. (selection consistency)

$$Pr(\{2 \leq j \leq p: \sup_{\tau \in \Delta} |\hat{\beta}_{\lambda_n}^{(j)}(\tau)| = 0\} = \{s+1, \dots, p\}) \rightarrow 1;$$

ii.

(weak convergence) $n^{1/2}(\hat{\beta}_{\lambda_n}^{(1:s)}(\tau) - \beta_0^{(1:s)}(\tau))$ converges weakly to a mean zero Gaussian process with a covariance matrix $\Sigma(\tau, \tau')$, the same limit covariance matrix of $n^{1/2}(\hat{\beta}_{\text{oracle}}(\tau) - \beta_0(\tau))$, where $\hat{\beta}_{\text{oracle}}(\tau)$ is the unpenalized estimator with all irrelevant variables precluded in advance.

The oracle properties established in Theorem 2 assert the uniform behavior of the estimated coefficient function across all $\tau \in \Delta$. They are stronger statements than the “local” oracle properties targeted at variable selection at a single quantile index.

3 The proposed BIC-type tuning parameter selector

It is known that the performance of a model selection procedure greatly rely on the choice of the tuning parameter. Although many existing criteria including AIC and GCV could be potentially employed to select the tuning parameter, Wang et al. (2007b) showed that the tuning parameter selected by AIC or GCV may fail to consistently identify the true model. Wang et al. (2007b) novelly imposed BIC for the model selection of smoothly clipped absolute deviation (SCAD) method. They select the optimal λ_n by minimizing

$$BIC(\lambda)^{LS} = \log \left(\sum_{i=1}^n (Y_i - \mathbf{X}_i' \hat{\beta}_{\lambda})^2 / n \right) + |S_{\lambda}| \log(n)/n,$$

where $\hat{\beta}_\lambda$ is the estimated coefficient vector by SCAD method and S_λ is the resulting model using the tuning parameter λ . They demonstrated that the BIC tuning parameter selector for SCAD method can consistently identify the true model. Wang and Leng (2007) further confirmed that the BIC-type selector for adaptive lasso with least square enjoys the same property. For varying coefficient quantile regression, Peng et al. (2014) provided numerical evidence to support the use of BIC-type selector. However, no formal justifications have been provided. One notable distinction from previous work is that the tuning parameter now controls the uniform shrinkage of coefficient function.

Following Peng et al. (2014), we consider the uniform BIC tuning parameter selector over λ_n , defined as the minimizer of

$$BIC(\lambda) = \int_{\Delta} \log \hat{\sigma}_\lambda(\tau) d\tau + \log(n) |S_\lambda| / n, \quad (4)$$

Here, $\hat{\sigma}_\lambda(\tau) = n^{-1} \rho_\tau(Y_i - \mathbf{X}_i' \hat{\beta}_\lambda(\tau))$ and $S_\lambda = \{j=2, \dots, p, \sup_{\tau \in \Delta} |\hat{\beta}_\lambda^{(j)}| \neq 0\}$. To facilitate the investigation of $BIC(\lambda)$, we examine an auxiliary BIC of model S , defined as

$$BIC_S = \int_{\Delta} \log \hat{\sigma}(S, \tau) d\tau + \log(n) |S| / n, \quad (5)$$

where $\hat{\sigma}(S, \tau) = n^{-1} \rho_\tau(Y_i - \mathbf{X}_i' \hat{\beta}(S, \tau))$ and $\hat{\beta}(S, \tau)$ are the estimated coefficients of model S at τ th quantile. We can see BIC_S is a model selection criterion for the unpenalized varying coefficient quantile regression.

We first demonstrate that with probability approaching 1, BIC_S is minimized when S is the true model. Since we consider smooth covariates effects, the true model S_T is equivalent to $\{j: \Delta_j := \{\tau \in \Delta: \beta_0^{(j)}(\tau) \neq 0\}, \text{ and } L(\Delta_j) > 0\}$, where L denotes the Lebesgue measure. We define the collection of overfitted models as $OF := \{S: S \supsetneq S_T\}$, and the collection of underfitted models as $UF := \{S: S \not\supset S_T\}$.

Theorem 3

Under the same assumptions in Theorem 1, we have

$$Pr\left\{\min_{S \in OF \cup UF} BIC_S > BIC_{S_T}\right\} \rightarrow 1.$$

Theorem 3 indicates that the minimal BIC value of overfitted or underfitted models is larger than that of the true model almost surely. Therefore, any overfitted or underfitted model can not be selected by the proposed BIC-type selector. In other words, BIC_S is a consistent model selection criterion for the unpenalized varying coefficient quantile regression.

By Theorems 1 and 2, there exists a sequence of tuning parameters, $\tilde{\lambda}_n$, such that $\sup_{\tau \in \Delta} \|\hat{\beta}_{\tilde{\lambda}_n}(\tau) - \beta_0(\tau)\| = O_p(n^{-1/2} \log \log n)$ and with probability tending to 1, $S_{\tilde{\lambda}_n} = S_T$. Recall that S_λ is the identified model from the proposed method with the tuning

parameter λ . By the definitions of $\hat{\beta}(S_\lambda, \tau)$ and $\log \hat{\sigma}(S_\lambda, \tau)$, we have $\log \hat{\sigma}(S_\lambda, \tau) = \log \hat{\sigma}_\lambda(\tau)$. Thus, $BIC(\lambda) \geq BIC_{S_\lambda}$, for all $\lambda > 0$.

Let Ω denote the collection of λ that can not lead to a correct identification of the true model. We know that with probability tending to 1,

$$\min_{\lambda \in \Omega} BIC(\lambda) \geq \min_{\lambda \in \Omega} BIC_{S_\lambda} > BIC_{S_T} = BIC_{S_{\tilde{\lambda}_n}}. \quad (6)$$

If we can show $BIC(\tilde{\lambda}_n) = BIC_{S_{\tilde{\lambda}_n}} + o_p(\log n)/n$, then from (6), we establish the consistency of the proposed BIC-type tuning parameter selector:

Corollary 1

Under the same assumptions in Theorem 1,

$$Pr \left(\min_{\lambda \in \Omega} BIC(\lambda) > BIC(\tilde{\lambda}_n) \geq BIC(\hat{\lambda}_n) \right) \rightarrow 1,$$

where $\hat{\lambda}_n$ is the minimizer of (4).

It is worth pointing out that a naive pointwise BIC tuning parameter selector, $\hat{\lambda}(\tau)$ is flawed for identifying S_T . Here, $\hat{\lambda}(\tau)$ is obtained by minimizing

$$BIC(\lambda; \tau) = \log \hat{\sigma}_\lambda(\tau) + \log(n) |S_\lambda(\tau)| / n, \quad (7)$$

for each $\tau \in \mathcal{T}$, where $S_\lambda(\tau) := \{j: \hat{\beta}_\lambda^{(j)}(\tau) \neq 0, 2 \leq j \leq p\}$, indicates the identified model at the τ th quantile. Although such a pointwise BIC can provide a tuning parameter leading to the consistent identification of a local model $S_{(\tau)} := \{j: \beta_0^{(j)}(\tau) \neq 0, 2 \leq j \leq p\}$, it can yield severely overfitted models for S_T , as demonstrated in the next section.

4 Numerical studies

4.1 Computation algorithm

To evaluate the general adaptively weighted LASSO estimator, we can take a pre-defined grid $G = \{\tau_1, \dots, \tau_m\}$ of $\mathcal{T}$. Our theoretical studies suggest that using a grid of size $O(n^{-1/2})$ would only bring asymptotically negligible deviations from the exact $\hat{\beta}_{\lambda_n}(\tau)$, which can be computationally expensive in many practical settings. Below we provide a detailed description of evaluating the general adaptively weighted LASSO estimator with a pre-defined grid.

Step 1: Evaluate the unpenalized quantile regression $\tilde{\beta}(\cdot)$ at $\tau_1, \dots, \tau_m$.

Step 2: Compute the adaptive weights $\omega_j(\tau)$ ($j = 2, \dots, p$) based on $\tilde{\beta}(\cdot)$.

Step 3: Calculate $BIC(\lambda)$ on a sequence of tuning parameter candidates. More specifically,

Step 3.1: Evaluate $\hat{\beta}(\cdot)$ at $\tau_1, \dots, \tau_m$ using existing software, such as `rq.fit.lasso()` in **R**.

Step 3.2: Calculate $\hat{\sigma}_\lambda(\tau)$ and $S_\lambda(\tau)$ at $\tau_1, \dots, \tau_m$.

Step 3.3: Calculate the grid approximated $BIC(\lambda)$ as follows

$$\sum_{k=1}^m \log \hat{\sigma}_\lambda(\tau_k) L(\Delta) / m + \log(n) \left| \bigcup_{k=1}^m S_\lambda(\tau_k) \right| / n$$

Step 4: Find $\hat{\lambda}_n$ that minimizes $BIC(\lambda)$.

Step 5: Obtain $\hat{\beta}_{\lambda_n}(\cdot)$ corresponding to $\lambda = \hat{\lambda}_n$ at $\tau_1, \dots, \tau_m$.

In our simulation studies, we chose the grid of size $2n/5$. The estimator based on grid approximation performs well with realistic sample sizes.

4.2 Simulations

We conduct extensive Monte-Carlo simulations to examine the finite sample performance of the proposed general adaptive lasso estimator equipped with the proposed BIC-type tuning parameter selector. We choose the UAL, the LASSO estimator adaptively weighted by $\omega_j(\tau) = 1/|\beta^{(j)}(\tau)|$ and $\omega_j(\tau) = 1/\int_{\tau \in \cdot} |\beta^{(j)}(\tau)| d\tau$, as representatives of the proposed method. We implement these estimators with the proposed BIC-type tuning parameter selector, denoted by UAL, UPS, and UAP respectively. We compare them with several other estimators: (i) the pointwise penalization approach with pointwise BIC, denoted by PPS; (ii) the adaptive-LASSO quantile regression estimator with the BIC from (7) at single predetermined quantile level, $\tau = 0.25, 0.50$ or 0.75 , denoted by SS(0.25), SS(0.50), and SS(0.75). We set $\tau_L = 0.1$, $\tau_U = 0.6$, and hence $\tau = [0.1, 0.9]$. The tuning parameter λ_n is selected as the minimizer of the corresponding BIC criterion among the $n/4$ equally-spaced grid points between 0 and $n/10$.

We consider the three simulation setups considered in Peng et al. (2014), denoted by (I), (II), and (III). More specifically, for each setup, the response variable Y is generated based on (1) with eight covariates, $Z^{(1)}, \dots, Z^{(8)}$. The effects of $Z^{(1)}, Z^{(2)}$, and $Z^{(5)}$ smoothly vary across τ , while the other covariates have zero effects. In setup (I), the intercept coefficient function is $\sqrt{2}$ times quantile function of standard normal distribution. The covariates $\mathbf{Z}_i$ are generated from multivariate normal distribution $N_p(\mathbf{0}, \Sigma)$ with $\Sigma = (\sigma_{jk})_{p \times p}$ and $\sigma_{jk} = 0.5^{|j-k|}$. The effects of $Z^{(1)}, Z^{(2)}$ and $Z^{(5)}$ are set to be 2, 1.5, and 3 constantly over τ . It is easy to see that setup (I) is a standard linear regression model with normal random errors. In setup (II), (III), the intercept coefficient function is set to be 0. We plot the coefficient functions for $Z^{(1)}, Z^{(2)}$ and $Z^{(5)}$ in Figure 1. In setups (II) and (III), we first generate $\tilde{\mathbf{Z}}_i$ from $N_p(\mathbf{0}, \Sigma)$ and then set $Z_i^{(j)} = \Phi(\tilde{Z}_i^{(j)})$ for $j = 1, \dots, p$, where $\Phi(\cdot)$ is the standard normal distribution function. Setups (II) and (III) are designed to assess the performance of the proposed estimator in varying covariate effects models.

For each set-up, we carried out 400 simulation replications with sample size $n = 200, 400$. We assess the model selection by calculating NC: mean number of correctly identified zero effects; NIC: mean number of incorrectly identified zero effects; PUF: percentage of under-fitted models, PCF: percentage of correctly fitted models; and POF: percentage of over-fitted models. We evaluate the estimation accuracy by computing their relative estimation errors with respect to the oracle estimator (REEO), for which we only use the true model. The REEO of $\beta(\tau)$ is defined as

$$REEO(\beta) = 100 \times \frac{\int_{\Delta} \sum_{i=1}^n \sum_{j=1}^p |\beta^{(j)}(\tau) - \beta_0^{(j)}(\tau)| d\tau}{\int_{\Delta} \sum_{i=1}^n \sum_{j=1}^p |\hat{\beta}_{oracle}^{(j)}(\tau) - \beta_0^{(j)}(\tau)| d\tau},$$

To calculate the REEO of $SS(\tau)$, we take the estimate from $SS(\tau)$ as the constant value over τ . The results are summarized in the Table 1.

As we can see from Table 1, in set-up (I), where the true model is a homoscedastic linear model, all methods, except for PPS, have similar performance in terms of variable selection. PPS demonstrates a strong tendency to select irrelevant variables with coefficient functions being zero over τ . Note that, PPS and UPS use the same adaptive weight function and differ only by the selection method of tuning parameter. The better model selection behavior of UPS (compared to that of PPS) indicates that the overfitting issue associated with PPS is primarily due to the pointwise BIC strategy for tuning parameter selection, which determines tuning parameter separately at each $\tau \in \Delta$.

In set up (II) and (III), the penalized quantile regression at single quantile level tends to miss relevant variables when the target quantile level is outside of the quantile index range, where these covariates have effects. For example, in set-up (II), we observed $NIC=0.99$ and $PCF=1.3\%$ for $SS(0.50)$ when $n = 200$. This indicates that one important variable was frequently missed by the penalized median regression. This is because the penalized median regression only extracts information at $\tau = 0.5$, where $Z^{(1)}$ has zero effect, despite that $Z^{(1)}$ is influential at other quantile levels. A similar phenomena is also observed in set-up (III), where $Z^{(1)}$ and $Z^{(2)}$ have partial effects on two disjoint quantile ranges. This suggests that the “local” quantile regression targeting certain specified quantile levels is subject to a higher risk of missing important variables in the presence of heterogeneous covariate-response associations, and hence a more thorough assessment of covariate effects on various quantile levels is needed for a more robust and sensible selection of variables that impact the response distribution. It is seen from Table 1 that the proposed estimators, UAL, UAP and UPS, all demonstrate satisfactory performance; PCF is always above 95% when $n = 400$ and the median of REEO decreases to approach 1 as n increases. This indicates that the proposed general class of adaptive weight functions is able to identify the true model consistently, and is a viable extension of the adaptive weight considered by Peng et al. (2014).

Another interesting finding is that the pointwise BIC fails to identify the true model consistently. The high POF in three set-ups exhibits its strong tendency to overfit. One intuitive explanation is that at a single quantile, the estimation may be affected by some perturbations of the data so that some irrelevant variables may be selected. If we collect

those selected variables across all quantiles, then overfitting is almost inevitable. On the other hand, the proposed BIC-type selector assesses the importance of a variable by its overall effect over τ . For a variable that is truly irrelevant but exhibits some artifactual effect at certain local quantiles, its overall effect across $\tau \in [0, 1]$ is expected to be very slight, and thus would be excluded by the proposed BIC-type selector. Meanwhile, those important variables having true strong impacts over τ are retained. Thus, the proposed BIC-type selector is able to identify the true model consistently. Our simulation results clearly confirm the consistency of the proposed BIC-type selector.

We also consider another simulation set-up (IV). In (IV), covariates are generated in the same manner as in (II) and (III). There are only two relevant covariates over quantile index interval $[0.1, 0.9]$. The true coefficient functions for $Z^{(1)}$ and $Z^{(2)}$ are plotted in Figure 1. $Z^{(2)}$ has a constant effect over all quantiles while $Z^{(1)}$ only impacts lower quantiles with $\tau < 0.25$. Based on the results in Table 1, it is quite clear that penalized quantile regression that targets $\tau = 0.25$ would have a high chance to fail to identify $Z^{(1)}$ as a variable that impacts the response distribution. Thus, for set-up (IV), we consider $SS(0.15)$ and $SS(0.20)$ instead of $SS(0.25)$, $S(0.50)$, $S(0.75)$. The simulation results are reported in Table 2.

We garner some additional interesting findings from Table 2. It seems that even though $Z^{(1)}$ and $Z^{(2)}$ are active variables at quantile levels $\tau = 0.15$ and $\tau = 0.20$, $SS(0.15)$ and $SS(0.20)$ demonstrate substantially lower power to correctly identify all relevant variables, as compared to UAL and UPS, particularly with sample size $n = 200$. This is somewhat different from what we observe from Table 1, which shows that $SS(\tau)$ generally produces good PCFs when all relevant variables have effects at τ th quantile. Given that signal strengths are quite comparable among all simulation set-ups, we think that the reduced power of $SS(\tau)$ in set-up (IV) is mainly because smaller τ 's are considered for $SS(\tau)$ in set-up (IV). Estimation of quantile regression generally has larger variability at lower or upper quantiles (compared to middle quantiles), and naturally $SS(\tau)$ has poorer performance with τ closer to 0 or 1. Table 2 suggests that the proposed variable selection method is less prone to such an issue, and therefore is expected to outperform the classical penalized quantile regression to capture variables that primarily influence the upper or lower end of the response distribution.

We also examine the proposed penalized quantile regression procedure with AIC and GCV tuning parameter selector. We use UAL_A , UAP_A , and UPS_A to denote the three general adaptive LASSO varying coefficient quantile regression estimators with AIC selector, and use UAL_G , UAP_G , and UPS_G to denote the three estimators with GCV selector. The simulation results are summarized in Table 3. From Table 3, we can see that the AIC and GCV based procedures have competitive performance in estimation accuracy but have a substantially larger chance to overfit model (1) compared to our BIC based procedure. This phenomena is consistent with the theoretical results in Wang et al. (2007b).

4.3 Real data example

We use the Air Pollution to Mortality Data to illustrate the utility of our proposal. 60 measures of total mortality on different US metropolitan areas in 1959-1961 were recorded in the dataset, along with 15 variables including precipitation (x_1), January temperature (x_2),

July temperature(x_3), the percentage of age 60 or older(x_4), population per household(x_5), education(x_6), the percentage of soundhouse (x_7), population per square mile (x_8), the percentage of non-white (x_9), the percentage of white collar (x_{10}), the percentage of family income under 3000 dollars (x_{11}), HC potential (x_{12}), NOx potential (x_{13}), SO2 potential (x_{14}), and relative humidity (x_{15}). This dataset is available in **R** package *SMPracticals*. Of scientific interest is to identify which variables have significant effect on y and to predict y given those variables. For our analysis, we standardize all variables to have unit variance, and choose $\lambda = [0.1, 0.9]$.

We applied the UAL, UAP, UPS with the proposed BIC-type selector, PPS with pointwise BIC selector, SS(0.25), SS(0.50), and SS(0.75) to the data set. The selected variables by those methods are reported in Table 4. To evaluate different methods, we compute the prediction errors (PE) as follow. First, we randomly split the 60 observations into a training set and a test set each consisting of 30 observations. We apply the method under consideration to the training data set and obtain the corresponding estimator, denoted by $\beta^{\text{train}}(\tau)$, and then calculate its prediction error over the test set as

$$PE = \frac{\sum_{i=1}^n 1\{i \text{ in test set}\} \int_{\Delta} \rho_{\tau}(y_i - x_i' \beta^{\text{train}}(\tau)) d\tau}{\sum_{i=1}^n 1\{i \text{ in test set}\}}.$$

The third column of Table 4 presents the average and standard deviation of PE based on 4000 replications of random splitting of training and test sets.

From Table 4, it is apparent that PPS with pointwise BIC selects too many variables; 12 variables are selected. This suggests that collecting variable selection results from locally identified models without any adjustment would yield an overfitted model. We also notice that there are quite discrepancies in variable selection results from SS(τ) with different τ 's. Among SS(0.25), SS(0.50), SS(0.75), only x_9 and x_{14} are common variables being selected. This may imply certain varying covariates effects on response. Meanwhile, we also need to be cautious about potential artifactual effects identified by SS(τ), as suggested by the overfitting phenomenon of PPS observed in simulation studies. The proposed method, UAL, UAP and UPS, select fewer variables compared to PPS. Moreover, UAL and UAP select the same set of variables, which cover major variables overlapped among SS(0.25), SS(0.50), SS(0.75). The comparisons of prediction errors also suggest that the model selected by UAL, UAP or UPS has better prediction performance compared to the other methods.

5 Conclusion remarks

Varying coefficient quantile regression offers an excellent platform to accommodate and investigate the potential heterogeneous covariates-response associations. In this work, we extend Peng et al. (2014) to study a class of adaptively weighted LASSO penalties. In addition, we provide formal justifications for a BIC-type tuning parameter selector, tailored to varying coefficient quantile regression. Our theoretical results and empirical studies both demonstrate that the proposed general adaptively weighted LASSO estimator coupled with the BIC-type selector can consistently identify the true model. In this work, we only concern the situation where p is fixed. Our study of variable selection based on high dimensional

varying coefficient quantile regression, where p grows as n increases, will be reported separately.

Acknowledgments

This work was supported by National Institutes of Health grant R01HL-113548.

Qi Zheng was supported by NIH grant R01HL113548. Limin Peng was supported by NSF Award DMS-1007660 and NIH grant R01HL113548.

7 Appendix

Proof of Theorems 1 and 2

The proofs of Theorem 2.1 and 2.2 follows the same arguments in Peng et al. (2014). Hence, we omit the details here.

Proof of Theorem 3

We consider **OF** and **UF** separately.

Part I (OF)

Given $S \in OF$, we would like to show that

$$Pr \{ \inf_{S \in OF} (BIC_S - BIC_{S_T}) > 0 \} \rightarrow 1. \quad (8)$$

We separate the proof into three steps:

Step 1: we first show

$$\sup_{\tau \in \Delta, S \in OF} \| \hat{\beta}(S, \tau) - \beta_0(\tau) \| = O_p(n^{-1/2} \log \log n). \quad (9)$$

Define $B_{n,\tau}(C) = \{ \beta : \beta_0(\tau) + (n^{-1/2} \log \log n) \mathbf{u}, \|\mathbf{u}\| \leq C \}$, and let $B_{n,\tau}$ denote the boundary set of $B_{n,\tau}(C)$. Define $D_n(\tau, \beta) = \sum_{i=1}^n \rho_\tau(Y_i - \beta \mathbf{X}_i) - \sum_{i=1}^n \rho_\tau(Y_i - \beta'_0(\tau) \mathbf{X}_i)$, where $\beta \in B_{n,\tau}(C)$.

Let $D(U) = U - E(U)$. We have $D_n(\tau, \beta) = I + \sum_{k=1}^6 II_k$, where

$$\begin{aligned}
I &= nE \left(1\{0 \leq Y_i - \beta'_0(\tau)\mathbf{X}_i < (\beta - \beta_0(\tau))'\mathbf{X}_i\} [-(Y_i - \beta'_0(\tau)\mathbf{X}_i) + (\beta - \beta_0(\tau))'\mathbf{X}_i] \right) \\
&\quad + nE \left(1\{(\beta - \beta_0(\tau))'\mathbf{X}_i \leq Y_i - \beta'_0(\tau)\mathbf{X}_i < 0\} [(Y_i - \beta'_0(\tau)\mathbf{X}_i) - (\beta - \beta_0(\tau))'\mathbf{X}_i] \right), \\
II_1 &= \sum_{i=1}^n D \left(1\{0 \leq Y_i - \beta'_0(\tau)\mathbf{X}_i < (\beta - \beta_0(\tau))'\mathbf{X}_i\} [-(Y_i - \beta'_0(\tau)\mathbf{X}_i) + (\beta - \beta_0(\tau))'\mathbf{X}_i] \right), \\
II_2 &= \sum_{i=1}^n D \left(1\{(\beta - \beta_0(\tau))'\mathbf{X}_i \leq Y_i - \beta'_0(\tau)\mathbf{X}_i < 0\} [(Y_i - \beta'_0(\tau)\mathbf{X}_i) - (\beta - \beta_0(\tau))'\mathbf{X}_i] \right), \\
II_3 &= \sum_{i=1}^n D \left(1\{Y_i - \beta'_0(\tau)\mathbf{X}_i \geq 0, Y_i - \beta'_0(\tau)\mathbf{X}_i \geq 0\} (-\tau)(\beta - \beta_0(\tau))'\mathbf{X}_i \right), \\
II_4 &= \sum_{i=1}^n D \left(1\{Y_i - \beta'_0(\tau)\mathbf{X}_i < 0, Y_i - \beta'_0(\tau)\mathbf{X}_i < 0\} (1 - \tau)(\beta - \beta_0(\tau))'\mathbf{X}_i \right), \\
II_5 &= \sum_{i=1}^n D \left(1\{(\beta - \beta_0(\tau))'\mathbf{X}_i \leq y_i - \beta'_0(\tau)\mathbf{X}_i < 0\} (1 - \tau)(\beta - \beta_0(\tau))'\mathbf{X}_i \right), \\
II_6 &= \sum_{i=1}^n D \left(1\{0 \leq Y_i - \beta'_0(\tau)\mathbf{X}_i < (\beta - \beta_0(\tau))'\mathbf{X}_i\} (-\tau)(\beta - \beta_0(\tau))'\mathbf{X}_i \right).
\end{aligned}$$

For the term II_1 , we first show the function classes

$$F = \{f_{\tau, \beta}(y, \mathbf{x}) := 1\{0 \leq y - \beta'_0(\tau)\mathbf{x} < (\beta - \beta_0(\tau))'\mathbf{x}\} \mathbf{x}', \tau \in \Delta, \beta \in A(C_0), y \in \mathbb{R}, \mathbf{x} \in \mathbb{R}^p\}$$

and

$$G = \{g_{\tau, \beta}(y, \mathbf{x}) := 1\{0 \leq y - \beta'_0(\tau)\mathbf{x} < (\beta - \beta_0(\tau))'\mathbf{x}\} (y - \beta'_0(\tau)\mathbf{x}), \tau \in \Delta, \beta \in A(C_0), y \in \mathbb{R}, \mathbf{x} \in \mathbb{R}^p\}$$

are Donsker classes, where $A(C_0) = \{\mathbf{b} \in \mathbb{R}^p : \inf_{\tau \in \Delta} \|\mathbf{b} - \beta_0(\tau)\| \geq C_0\}$.

Consider $W_1 := \{y - \beta'_0(\tau)\mathbf{x}, \tau \in \Delta\}$, $W_2 := \{1\{0 \leq y - \beta'_0(\tau)\mathbf{x}\}, \tau \in \Delta\}$ and $V := \{1\{y - \beta'_0(\tau)\mathbf{x} < 0\}, \beta \in A(C_0)\}$. Since $\tau \mapsto \beta'_0(\tau)\mathbf{x}$ is monotone according to the definition of quantiles. Then the VC index of W_1 and W_2 is bounded by 2. The VC index of $\{y - \beta'_0(\tau)\mathbf{x}, \beta \in A(C_0)\}$ is $p + 3$ by **Lemma 2.6.15** in Van Der Vaart and Wellner (2000). Thus, the VC index of V is bounded by $p + 3$ according to **Lemma 2.6.18** in Van Der Vaart and Wellner (2000). The VC index of G is equal to the VC index of the class of sets $\{(Z, t) : g_{\tau, \beta}(Z, t) \leq t\}$ where $Z = (y, \mathbf{x}')$, $g_{\tau, \beta} \in G$, $t \in \mathbb{R}$.

We have

$$\begin{aligned}
\{(Z, t) : g_{\tau, \beta}(Z) \leq t\} &= \{(Z, t) : (y - \beta'_0(\tau)\mathbf{x}) 1\{0 \leq y - \beta'_0(\tau)\mathbf{x} < (\beta - \beta_0(\tau))'\mathbf{x}\} \leq t\} \\
&= \{(Z, t) : y - \beta'_0(\tau)\mathbf{x} \leq t, y - \beta'_0(\tau)\mathbf{x} < 0, y - \beta'_0(\tau)\mathbf{x} \geq 0\} \cup \\
&\quad \{(Z, t) : y - \beta'_0(\tau)\mathbf{x} \geq 0, t > 0\} \cup \{(Z, t) : y - \beta'_0(\tau)\mathbf{x} < 0, t > 0\}
\end{aligned}$$

Thus, each set $\{(Z, t) : g_{\tau, \beta}(Z) \leq t\}$ is created by taking finite unions, intersections, and complements of the basic sets

$\{(Z, t) : y - \beta'_0(\tau)\mathbf{x} \leq t\}$, $\{Z : y \leq \beta'_0(\tau)\mathbf{x}\}$, $\{Z : y \leq \beta'_0(\tau)\mathbf{x}\}$, $\{t > 0\}$. These basic sets from VC classes, each having VC index no greater than $p + 3$. Therefore, by **Lemma 2.6.17** in Van Der Vaart and Wellner (2000), we can see that the VC index of G is bounded by $4p + 12$. Let $G(Z) = (C_0 + 1)\|Z\|$. Clearly, $G(\cdot)$ is an envelop function for G with a finite second moment. Since G is a parametric class of functions, where (τ, β) belongs to a compact subset of $\mathbb{R}^{p+1}$, G is a separable class. By **Theorem 2.6.14** in Van Der Vaart and Wellner (2000), G is Donsker. Similarly, we can prove F is a Donsker.

We have

$$\sup_{\beta \in B_{n,t}(C_0), \tau \in \Delta} |II_1| \leq 2 \sup_{\beta \in B_{n,t}(C_0), \tau \in \Delta} \|\beta - \beta_0(\tau)\| \times \sup_{\beta \in B_{n,t}(C_0), \tau \in \Delta} \left\| \sum_{i=1}^n D(1 \{0 \leq Y_i - \beta'_0(\tau) \mathbf{X}_i < (\beta - \beta_0(\tau))' \mathbf{X}_i\}) \mathbf{X}_i' \right\| \quad (10)$$

Since the functional classes F and G are Donsker and Suslin, we can apply the functional law of the iterated logarithm in Kuelbs and Dudley (1980) to (10), and get

$\sup_{\beta \in B_{n,\tau}(C_0), \tau \in \Delta} |II_1| \leq O_p(\log \log n)(C_0 \log \log n) = C_0 O_p(l_2^2 n)$. By similar reasoning, we have $\sup_{\beta \in B_{n,\tau}(C_0), \tau \in \Delta} |II_k| \leq C_0 O_p(l_2^2 n)$, for $k = 2$. Therefore, we obtain

$$\sup_{\beta \in B_{n,\tau}(C_0), \tau \in \Delta} \left| \sum_{k=1}^6 II_k \right| \leq C_0 O_p(l_2^2 n) \quad (11)$$

Next we consider the term I . According to Condition C2 (i) and (ii), there exist $\delta_2 > 0$, such that $\inf_{\tau \in \Delta, |t| \leq \delta_2, x} f_\tau(t|x) > \underline{f}/2$

$$I = n E \left[\int_0^{(\beta - \beta_0(\tau))' \mathbf{X}_i} (-t + (\beta - \beta_0(\tau))' \mathbf{X}_i) f_\tau(t | \mathbf{X}_i) dt \right] \geq \frac{1}{2} n \underline{f} E[(\beta - \beta_0(\tau))' \mathbf{X}_i]^2 - \frac{4}{3} A_0 E[|(\beta - \beta_0(\tau))' \mathbf{X}_i|^3].$$

If $\beta \in B_{n,\tau}(C_0)$, then we get

$$\inf_{\beta \in \partial B_{n,\tau}(C_0), \tau \in \Delta} I \geq \lambda_{\min} C_0^2 l_2^2 n / 2 - 4 A_0 q C_0^3 l_2^3 n / (3 \sqrt{n}), \quad (12)$$

where $\lambda_{\min}$ denotes the smallest eigenvalue of $E[\mathbf{X}_i \mathbf{X}_i']$. From equations (11) and (12), it follows that $\inf_{\beta \in \partial B_{n,\tau}(C_0), \tau \in \Delta} D_n(\tau, \beta) > 0$ for a large C_0 and sufficiently large n . Since $\rho_\tau(y - \beta'x)$ is convex in β , $\inf_{\beta \in B_{n,\tau}(C_0), \tau \in \Delta} D_n(\tau, \beta) > 0$, which implies (9).

Step 2: Next we show that

$$r_n := \sup_{\beta \in B_{n,\tau}(C_0), \tau \in \Delta} \left| \frac{D_n(\tau, \beta) - D_n(\tau, \hat{\beta}(S_T, \tau))}{D_n(\tau, \hat{\beta}(S_T, \tau)) + \sum_{i=1}^n \rho_\tau(Y_i - \beta'_0(\tau) \mathbf{X}_i)} \right| = o_p(1) \\ r_n \leq C_1 |O_p(l_2^2 n)| \quad \text{for some constant } C_1.$$

From Step 1, we can also observe that $\sup_{\beta \in B_{n,\tau}(C_0), \tau \in \Delta} |I| \leq 3 \bar{f} \lambda_{\max} C_0^2 l_2^2 n / 2$ where $\lambda_{\max}$ denotes the largest eigenvalue of $E[\mathbf{X}_i \mathbf{X}_i']$. This and (11) together imply that

$$\sup_{\beta \in B_{n,\tau}(C_0), \tau \in \Delta} |D_n(\tau, \beta)| \leq C_0 |O_p(l_2^2 n)| + 3\bar{f} \lambda_{max} C_0^2 l_2^2 n. \quad (13)$$

Also, we can show that $\{\rho_\tau(y - \beta'_0(\tau)x), \tau \in \Delta\}$ is Donsker, and hence Glivenko-Cantelli.

Since Δ is a compact set, we have

$0 < \inf_{\tau \in \Delta} E[\rho_\tau(Y_i - \beta'_0(\tau)\mathbf{X}_i)] \leq \sup_{\tau \in \Delta} E[\rho_\tau(Y_i - \beta'_0(\tau)\mathbf{X}_i)]$. We have

$$\begin{aligned} r_n &= \sup_{\beta \in B_{n,\tau}(C_0), \tau \in \Delta} \left| \frac{D_n(\tau, \beta) - D_n(\tau, \hat{\beta}(S_T, \tau))}{D_n(\tau, \hat{\beta}(S_T, \tau)) + \sum_{i=1}^n \rho_\tau(Y_i - \beta'_0(\tau)\mathbf{X}_i)} \right| \\ &\leq \left| \frac{\sup_{\beta \in B_{n,\tau}(C_0), \tau \in \Delta} |D_n(\tau, \beta)| + \sup_{\tau \in \Delta} |D_n(\tau, \hat{\beta}(S_T, \tau))|}{-\inf_{\tau \in \Delta} |D_n(\tau, \hat{\beta}(S_T, \tau))| + \inf_{\tau \in \Delta} \sum_{i=1}^n \rho_\tau(Y_i - \beta'_0(\tau)\mathbf{X}_i)} \right| \\ &\leq \left| \frac{(2C_0 |O_p(l_2^2 n)| + 3\bar{f} \lambda_{max} C_0^2 l_2^2 n)/n}{-(C_0 |O_p(l_2^2 n)| + 3/2 \bar{f} \lambda_{max} C_0^2 l_2^2 n)/n + \inf_{\tau \in \Delta} E[\rho_\tau(Y_i - \beta'_0(\tau)\mathbf{X}_i)]} \right| = |o_p(1)|, \end{aligned}$$

and $r_n(\tau)n \leq C_1 |O_p(l_2^2 n)|$, where $C_1 := 3\bar{f} \lambda_{max} C_0^2 / (\inf_{\tau \in \Delta} E[\rho_\tau(Y_i - \beta'_0(\tau)\mathbf{X}_i)])$.

Step 3: Now, we show

$$Pr \left\{ \inf_{S \in OF} n(BIC_S - BIC_{S_T}) > 0 \right\} \rightarrow 1. \quad (14)$$

We observe that

$$\inf_{S \in OF} n(BIC_S - BIC_{S_T}) \geq \int_{\Delta} \log \left(1 + \inf_{S \in OF, \tau \in \Delta} \eta_n(\tau) \right)^n d\tau + \log(n),$$

where

$$\eta_n(\tau) = \frac{D_n(\tau, \hat{\beta}(S, \tau)) - D_n(\tau, \hat{\beta}(S_T, \tau))}{D_n(\tau, \hat{\beta}(S_T, \tau)) + \sum_{i=1}^n \rho_\tau(Y_i - \beta'_0(\tau)\mathbf{X}_i)}$$

In Step 1, we showed $\sup_{\tau \in \Delta, S \in OF} \|\hat{\beta}(S, \tau) - \beta_0(\tau)\| = O_p(n^{-1/2} \log \log n)$. Thus,

$$\begin{aligned} \inf_{S \in OF, \tau \in \Delta} \eta_n(\tau) &\geq - \sup_{\beta \in B_{n,\tau}(C_0), \tau \in \Delta} \left| \frac{D_n(\tau, \beta) - D_n(\tau, \hat{\beta}(S_T, \tau))}{D_n(\tau, \hat{\beta}(S_T, \tau)) + \sum_{i=1}^n \rho_\tau(Y_i - \beta'_0(\tau)\mathbf{X}_i)} \right| \\ &= -r_n = o_p(1). \end{aligned}$$

It follows

$$\inf_{S \in OF} n(BIC_S - BIC_{S_T}) \geq r_n n \int_{\Delta} \log(1 - r_n)^{\frac{1}{r_n}} d\tau + \log(n) \geq -C_1 O_p(l_2^2 n) + \log(n), \quad (15)$$

which directly implies (14). Consequently, we obtain (8).

Part II (UF)

Given $S \in UF$, we would like to show that

$$\Pr \left\{ \inf_{S \in UF} (BIC_S - BIC_{S_T}) > 0 \right\} \rightarrow 1. \quad (16)$$

Since $S \in UF$, there exists some j_0 , s.t. $j_0 \in S_T$ but $j_0 \notin S$. Since $j_0 \in S_T$, then $L(\Delta_{j_0}) > 0$, where $\Delta_{j_0} := \{\tau \in \Delta : \beta_0^{(j_0)}(\tau) \neq 0\}$. Let $\Delta_{j_0}^c = \Delta \setminus \Delta_{j_0}$. Given $\Delta_{j_0} = \bigcup_{c \in \mathbb{Q}} \{\tau : |\beta_0^{(j_0)}(\tau)| \geq c\}$ and $L(\Delta_{j_0}) > 0$, there exists $C_{3,j_0} > 0$, s.t. $\Delta_{j_0}^0 := \{\tau : |\beta_0^{(j_0)}(\tau)| \geq C_{3,j_0}\}$ has a positive measure, i.e. $L(\Delta_{j_0}^0) > 0$. Let $\Delta_{j_0}^1 = \Delta_{j_0} \setminus \Delta_{j_0}^0$.

Step 1: We first attempt to show that for some $C_{2,j_0} > 0$, in probability we have

$$\int_{\Delta_{j_0}} (\log \hat{\sigma}(S, \tau) - \log \hat{\sigma}(S_T, \tau)) d\tau > C_{2,j_0} \quad (17)$$

Let $B_\tau(C) = \{\beta : \beta_0(\tau) + \mathbf{u}, \|\mathbf{u}\| \leq C\}$. Applying similar arguments in **Part I** yields

$$\sup_{\beta \in B_\tau(C_{3,j_0}), \tau \in \Delta_{j_0}} \left| \sum_{k=1}^6 II_k \right| \leq O_p(\sqrt{n} \log \log n) C_{3,j_0} \text{ and}$$

$$\inf_{\beta \in \partial B_\tau(C_{3,j_0}), \tau \in \Delta_{j_0}} I \geq n \lambda_{\min} C_{3,j_0}^2 l_2^2 n / 2 - 4n A_0 q C_{3,j_0}^3 l_2^3 n / 3. \text{ Consequently,}$$

$$\inf_{\beta \in \partial B_\tau(C_{3,j_0}), \tau \in \Delta_{j_0}} D_n(\tau, \beta) \geq -|O_p(\sqrt{n} \log \log n)| C_{3,j_0}, \quad (18)$$

with sufficiently small C_{3,j_0} . By convexity, we obtain

$$\inf_{\beta \notin B_\tau(C_{3,j_0}), \tau \in \Delta_{j_0}} D_n(\tau, \beta) > \inf_{\beta \in \partial B_\tau(C_{3,j_0}), \tau \in \Delta_{j_0}} D_n(\tau, \beta). \quad (19)$$

We can write

$$\int_{\Delta_{j_0}} (\log \hat{\sigma}(S_T, \tau) - \log \hat{\sigma}(S, \tau)) d\tau \geq \int_{\Delta_{j_0}^0} \inf_{\tau \in \Delta_{j_0}^0} \log(1 + \eta_n(\tau)) d\tau + \int_{\Delta_{j_0}^1} \inf_{\tau \in \Delta_{j_0}^1} \log(1 + \eta_n(\tau)) d\tau.$$

For $\tau \in \Delta_{j_0}^0$, $\|\hat{\beta}(S, \tau) - \beta_0(\tau)\| \geq |0 - \beta_0^{(j_0)}(\tau)| \geq C_{3,j_0}$. Then by equations (13), (18) and (19),

$$\begin{aligned}
& \inf_{\tau \in \Delta_{j_0}^0} \log(1 + \eta_n(\tau)) \geq \log(1 + \inf_{\tau \in \Delta_{j_0}^0} \eta_n(\tau)) \\
& \geq \log \left(1 + \frac{\inf_{\tau \in \Delta_{j_0}^0} D_n(\tau, \hat{\beta}(S, \tau)) - \sup_{\tau \in \Delta} D_n(\tau, \hat{\beta}(S_T, \tau))}{\sup_{\tau \in \Delta} D_n(\tau, \hat{\beta}(S_T, \tau)) + \sup_{\tau \in \Delta} \sum_{i=1}^n \rho_\tau(Y_i - \beta'_0(\tau) \mathbf{X}_i)} \right) \\
& \geq \log \left(1 + \frac{(-|O_p(\sqrt{n} \log \log n) C_{3,j_0}| + \frac{n}{2} \lambda_{\min} C_{3,j_0}^2 l_2^2 n - \frac{4n}{3} A_0 q C_{3,j_0}^3 l_2^3 n - |O_p(l_2^2 n)|)/n}{(|O_p(l_2^2 n)| + \sup_{\tau \in \Delta} \sum_{i=1}^n \rho_\tau(Y_i - \beta'_0(\tau) \mathbf{X}_i))/n} \right)
\end{aligned}$$

Taking the limit yields

$$\lim_{n \rightarrow \infty} \inf_{\tau \in \Delta_{j_0}^0} \log(1 + \eta_n(\tau)) = \log \left(1 + \frac{\frac{1}{2} \lambda_{\min} C_{3,j_0}^2 l_2^2 n - \frac{4}{3} A_0 q C_{3,j_0}^3 l_2^3 n}{\sup_{\tau \in \Delta} E[\rho_\tau(Y_i - \beta'_0(\tau) \mathbf{X}_i)]} \right) := C_{2,j_0}/L(\Delta_{j_0}^0) > 0.$$

For $\tau \in \Delta_{j_0}^1$ by equations (13), (18) and (19),

$$\inf_{\Delta_{j_0}^1} \log(1 + \eta_n(\tau)) \geq \log(1 + \inf_{\tau \in \Delta_{j_0}^1} \eta_n(\tau)) = o_p(1).$$

Therefore, (17) can be obtained by combining the above two results.

Step 2: Next, we show that almost surely,

$$BIC_S - BIC_{S_T} > C_{2,j_0}. \quad (20)$$

Let $d = |S_T|$. Following (17), we have

$$\begin{aligned}
& BIC_S - BIC_{S_T} \\
& \geq \int_{\Delta_{j_0}^c} (\log \hat{\sigma}(S, \tau) - \log \hat{\sigma}(S_T, \tau)) d\tau + \int_{\Delta_{j_0}^c} (\log \hat{\sigma}(S, \tau) - \log \hat{\sigma}(S_T, \tau)) d\tau - d \log(n)/n \\
& \geq C_{2,j_0} + \int_{\Delta_{j_0}^c} (\log \hat{\sigma}(S, \tau) - \log \hat{\sigma}(S_T, \tau)) d\tau - d \log(n)/n
\end{aligned}$$

For $\tau \in \Delta_{j_0}^c$, S could be underfitted, correctly fitted or overfitted. If S is either correctly fitted or overfitted $\tau \in \Delta_{j_0}^c$, based on similar arguments in part I, we have that

$$\int_{\Delta_{j_0}^c} (\log \hat{\sigma}(S, \tau) - \log \hat{\sigma}(S_T, \tau)) d\tau = o_p(1). \text{ Thus, (20) holds almost surely. If } S \text{ is}$$

underfitted in $\Delta_{j_0}^c$, we can find another missing active variable j_1 , and apply the same process as presented above to split $\Delta_{j_0}^c$ into two parts: Δ_{j_1} and $\Delta_{j_1}^c := \Delta_{j_0}^c \setminus \Delta_{j_1}$. We keep doing this process until S is either correctly fitted or overfitted on some measurable set. Each integral considered is either positive or $o_p(1)$. Therefore, (20) holds almost surely.

Step 3: Now we show (16). Without loss of generality, we assume that the first d variables are active. For each active variable j , we can find $C_{2,j}$, $j = 1, \dots, d$. For any $S \in UF$, some active variables is missing. Therefore, uniformly for

$S \in UF$, $BIC_S - BIC_{S_T} \geq \min_{j \leq d} C_{2,j}$ with probability approaching 1. Take minimum of the left hand side and we complete the proof of (16).

We complete the proof of Theorem 3 by combining (8) and (16).

Proof of Corollary 1

By (15) and (20) in the proof of Theorem 3, we can conclude that in probability

$\inf_{S \in OF \cup UF} (BIC_S - BIC_{S_T}) \geq \log n / (2n)$. By Theorems 1 and 2, we observe that with probability approaching 1, $BIC(\tilde{\lambda}_n) - BIC_{S_{\tilde{\lambda}_n}} = BIC(\tilde{\lambda}_n) - BIC_{S_T} = O_p(l_2^2 n / n)$. Then by the minimality of $BIC(\hat{\lambda}_n)$ in probability tending to 1, we have

$$\begin{aligned} \inf_{\lambda \in \Omega} BIC(\lambda) &\geq \inf_{\lambda \in \Omega} BIC_{S_\lambda} = \inf_{S \in OF \cup UF} BIC_S \geq BIC_{S_T} + \log n / (2n) \\ &= BIC_{S_{\tilde{\lambda}_n}} + \log n / (2n) \geq BIC(\tilde{\lambda}_n) \geq BIC(\hat{\lambda}_n). \end{aligned}$$

This completes the proof of Corollary 1.

References

- Fan J, Li R. Variable selection via nonconcave penalised likelihood and its oracle properties. *Journal of the American Statistical Association*. 2001; 96:1348–1360.
- Kai B, Li R, Zou H. New efficient estimation and variable selection methods for semi-parametric varying-coefficient partially linear models. *Annals of Statistics*. 2011; 39:305–332. [PubMed: 21666869]
- Koenker R, Bassett GW. Regression quantiles. *Econometrica*. 1978; 46:33–50.
- Kuelbs J, Dudley R. Log log laws for empirical measures. *Annals of Probability*. 1980; 8:405–418.
- Lee E, Noh H, Park B. Model selection via bayesian information criterion for quantile regression models. *Journal of the American Statistical Association*. 2014; 109:216–229.
- Peng L, Xu J, Kutner N. Shrinkage estimation of varying covariate effects based on quantile regression. *Statistics and Computing*. 2014
- Van Der Vaart, A.; Wellner, J. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer; New York: 2000.
- Volgushev S, Wagener J, Dette H. Censored quantile regression processes under dependence and penalization. 2012 arxiv.org/abs/1208.5467.
- Wang D, Kulasekera K. Parametric component detection and variable selection in varying-coefficient partially linear models. *Journal of Multivariate Analysis*. 2012; 112:117–129.
- Wang H, Leng C. Unified lasso estimation by least squares approximation. *Journal of the American Statistical Association*. 2007; 102:1039–1048.
- Wang H, Li G, Jiang G. Robust regression shrinkage and consistent variable selection via the lad-lasso. *Journal of Business and Economic Statistics*. 2007a; 25:347–355.
- Wang H, Li R, Tsai CL. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*. 2007b; 94:553–568. [PubMed: 19343105]
- Wang H, Xia Y. Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association*. 2009; 104:747–757.
- Wang L, Wu Y, Li R. Quantile regression for analyzing heterogeneity in ultrahigh dimension. *Journal of the American Statistical Association*. 2012; 101:1418–1429.
- Wu Y, Liu Y. Variable selection in quantile regression. *Statistica Sinica*. 2009; 19:801–817.
- Zheng Q, Gallagher C, Kulasekera KB. Adaptive penalized quantile regression for high dimensional data. *Journal of Statistical Planning and Inference*. 2013; 143:1029–1038.
- Zou H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*. 2006; 101:1418–1429.

Zou H, Yuan M. Composite quantile regression and the oracle model selection theory. *Annals of Statistics*. 2008; 36:1108–1126.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1.
The coefficient functions in simulation set-ups (II) & (III)

Figure 2.
The coefficient functions in simulation set-up (IV)

Table 1

Simulation results for three set-ups

set-up	Method	n	NC	NIC	PUF	PCF	POF	Median of REEO(%)
(I)	UAL	200	4.98	.00	.0	98.8	1.2	108.1
		400	4.99	.00	.0	99.3	0.7	104.7
	UAP	200	4.99	.00	.0	99.0	1.0	102.7
		400	5.00	.00	.0	99.8	0.2	101.6
	UPS	200	4.99	.00	.0	99.0	1.0	108.7
		400	5.00	.00	.0	100.0	.0	104.5
	PPS	200	4.23	.00	.0	48.0	52.0	105.7
		400	4.39	.00	.0	56.3	43.7	103.3
	SS(0.25)	200	4.93	.00	.0	94.0	6.0	107.6
		400	4.97	.00	.0	97.3	2.7	102.3
	SS(0.50)	200	4.98	.00	.0	97.5	2.5	92.7
		400	4.99	.00	.0	98.8	1.2	90.7
(II)	UAL	200	4.94	.00	.0	93.8	6.2	105.2
		400	4.97	.00	.0	96.5	3.5	103.4
	UAP	200	4.93	.00	.0	93.3	6.7	155.2
		400	4.99	.00	.0	99.0	1.0	123.1
	UPS	200	4.95	.00	.5	94.7	4.8	136.3
		400	4.99	.00	.0	99.0	1.0	114.3
	PPS	200	4.91	.00	.3	91.5	8.2	177.0
		400	4.98	.00	.0	98.0	2.0	140.8
	SS(0.25)	200	4.36	.00	.0	52.5	47.5	149.7
		400	4.78	.00	.0	80.9	19.1	128.0
	SS(0.50)	200	4.95	.25	16.2	79.3	4.5	368.4
		400	4.98	.00	.0	98.5	1.5	485.1
SS(0.75)	SS(0.50)	200	4.99	.99	98.7	1.3	.2	262.1
		400	5.00	1.00	100.0	.0	.0	360.6
	SS(0.75)	200	4.96	0.05	5.3	91.0	3.7	340.4
		400	4.99	.00	.2	99.0	0.8	411.1

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

set-up	Method	n	NC	NIC	PUF	PCF	POF	Median of REEO(%)
(III)	UAL	200	4.93	0.02	1.0	92.5	6.5	107.9
		400	5.00	0.00	.0	99.5	.5	96.2
	UAP	200	4.91	0.12	9.5	82.0	8.5	113.7
	UPS	400	4.98	0.00	.0	98.3	1.7	99.4
		200	4.89	0.06	5.5	92.0	2.5	132.6
	PPS	400	4.99	0.00	.0	98.5	1.5	88.7
		200	4.42	.00	.0	57.0	43.0	92.6
		400	4.81	.00	.0	85.5	14.5	82.3
	SS(0.25)	200	4.95	1.00	98.0	1.8	.2	293.4
		400	4.99	1.00	99.8	0.2	.0	421.5
	SS(0.50)	200	4.98	1.93	99.8	.2	.0	237.1
		400	5.00	1.95	100.0	.0	.0	340.8
		200	4.95	1.01	99.3	.5	.2	292.2
	SS(0.75)	400	4.99	1.00	99.8	.2	.0	418.5

Table 2

Simulation results for set-up (IV)

set-up	Method	n	NC	NIC	PUF	PCF	POF	Median of REEO(%)
(IV)	UAL	200	6.00	.11	11.3	88.5	.2	157.08
		400	6.00	.01	1.3	98.8	1.2	141.6
	UAP	200	5.87	.09	9.3	73.0	17.7	149.4
		400	6.00	.02	2.0	97.8	.2	123.0
	UPS	200	6.00	.22	21.5	78.3	.2	190.2
		400	5.99	.01	1.3	98.2	.5	143.7
	PPS	200	4.97	.01	.5	30.4	69.6	157.9
		400	5.34	.00	.0	57.9	42.1	150.3
	SS(0.10)	200	5.73	.20	19.8	62.5	17.7	277.2
		400	5.86	.02	1.5	86.3	12.2	302.3
	SS(0.15)	200	5.84	.55	55.3	36.3	8.4	318.0
		400	5.91	0.32	32.3	62.5	5.2	374.7
	SS(0.20)	200	5.88	.65	64.8	29.3	5.9	328.8
		400	5.93	0.44	44.3	52.0	3.7	403.7

Table 3

Simulation results with AIC and Cross-Validation

set-up	Method	n	NC	NIC	PUF	PCF	POF	Median of REEO(%)
(I)	UAL _A	200	4.72	.00	.0	83.8	16.2	110.0
		400	4.79	.00	.0	89.3	10.7	106.3
	UAP _A	200	4.78	.00	.0	85.8	14.2	104.1
		400	4.84	.00	.0	90.3	9.7	102.3
	UPS _A	200	4.77	.00	.0	86.5	13.5	109.7
		400	4.86	.00	.0	93.3	6.7	105.1
	UAL _C	200	4.77	.00	.0	85.3	14.7	110.0
		400	4.83	.00	.0	90.3	9.7	106.3
	UAP _C	200	4.81	.00	.0	86.0	14.0	104.0
		400	4.83	.00	.0	90.0	10.0	102.3
	UPS _C	200	4.80	.00	.0	87.8	12.2	109.6
		400	4.88	.00	.0	94.0	6.0	105.1
(II)	UAL _A	200	4.68	.00	.0	78.3	21.7	149.8
		400	4.91	.00	.0	92.0	8.0	127.1
	UAP _A	200	4.78	.00	.0	84.3	15.7	135.0
		400	4.93	.00	.0	93.7	6.3	115.6
	UPS _A	200	4.58	.00	.0	70.8	29.2	166.3
		400	4.85	.00	.0	87.2	12.8	138.8
	UAL _C	200	4.72	.00	.0	79.3	20.7	150.1
		400	4.93	.00	.0	93.5	6.5	127.2
	UAP _C	200	4.80	.00	.0	85.0	15.0	135.2
		400	4.94	.00	.0	95.0	5.0	115.6
	UPS _C	200	4.64	.00	.0	74.0	26.0	167.4
		400	4.86	.00	.0	88.2	11.8	140.2
(III)	UAL _A	200	4.72	.00	.0	75.8	24.2	103.5
		400	4.91	.00	.0	92.3	7.7	96.5
	UAP _A	200	4.65	.02	2.1	73.0	24.9	107.2
		400	4.78	.00	.0	85.4	14.5	100.5

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

set-up	Method	n	NC	NIC	PUF	PCF	POF	Median of REEO(%)
	UPS _A	200	4.59	.00	.0	70.9	29.1	99.6
		400	4.81	.00	.0	85.3	14.7	88.1
	UAL _C	200	4.75	.00	.0	78.7	21.3	104.0
		400	4.93	.00	.0	93.5	6.5	96.5
	UAP _C	200	4.68	.05	5.0	73.0	22.0	107.2
		400	4.82	.00	.0	87.0	13.0	100.3
	UPS _C	200	4.61	.00	.0	71.6	28.4	99.6
		400	4.81	.00	.0	84.8	15.2	88.2

Table 4

Analysis of the air pollution to mortality data

Methods	Variables selected	Prediction error mean (sd)
UAL	X_6, X_9, X_{12}, X_{14}	.236 (.0016)
UAP	X_6, X_9, X_{12}, X_{14}	.249 (.0018)
UPS	X_6, X_9, X_{12}	.249 (.0018)
PPS	$X_1, X_2, X_3, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{12}, X_{13}, X_{14}$	.310 (.0030)
SS90.25)	$X_1, X_3, X_9, X_{10}, X_{13}, X_{14}$	.271 (.0024)
SS(0.50)	$X_1, X_8, X_9, X_{10}, X_{12}, X_{14}$	.264 (.0022)
SS(0.75)	$X_6, X_8, X_9, X_{12}, X_{14}$	.292 (.0029)