Annals of the Institute of Statistical Mathematics (2023) 75:649–681
https://doi.org/10.1007/s10463-022-00857-z

# Model averaging for semiparametric varying coefficient quantile regression models

**Zishu Zhan[1] · Yang Li[2] · Yuhong Yang[3] · Cunjie Lin[2]**

Received: 30 November 2021 / Revised: 14 July 2022 / Accepted: 8 November 2022 /
Published online: 22 December 2022


## Abstract

In this study, we propose a model averaging approach to estimating the conditional quantiles based on a set of semiparametric varying coefficient models. Different from existing literature on the subject, we consider a particular form for all candidates, where there is only one varying coefficient in each sub-model, and all the candidates under investigation may be misspecified. We propose a weight choice criterion based on a leave-more-out cross-validation objective function. Moreover, the resulting averaging estimator is more robust against model misspecification due to the weighted coefficients that adjust the relative importance of the varying and constant coefficients for the same predictors. We prove out statistical properties for each sub-model and asymptotic optimality of the weight selection method. Simulation studies show that the proposed procedure has satisfactory prediction accuracy. An analysis of a skin cutaneous melanoma data further supports the merits of the proposed approach.

**Keywords** Model averaging · Quantile regression · Local linear fitting · Semiparametric model · Varying coefficient

✉ Cunjie Lin
lincunjie@ruc.edu.cn

Zishu Zhan
zishu927@hotmail.com

Yang Li
yang.li@ruc.edu.cn

Yuhong Yang
yangx374@umn.edu

[1] School of Statistics, Renmin University of China, No. 59 Zhongguancun Street, Haidian District, Beijing 100872, People's Republic of China

[2] Center for Applied Statistics and School of Statistics, Renmin University of China, No. 59 Zhongguancun Street, Haidian District, Beijing 100872, People's Republic of China

[3] School of Statistics, University of Minnesota, 313 Ford Hall, 224 Church Street SE, Minneapolis, MN 55455, USA

## 1 Introduction

Quantile regression (Koenker and Bassett 1978) has attracted considerable interest in the literature not only because it provides a more complete view of the conditional distribution of a response variable than the conditional mean, but also because it is more robust to outliers compared to the method of least squares. It has been extensively applied in economics, biology, medicine, finance and many other disciplines (Fitzenberger et al. 2002; Wheelock and Wilson 2008; Li et al. 2010).

In many practical applications, however, the linear parametric quantile regression may not be flexible enough to capture the relationship between response and its covariates. For example, in genetic analysis, many studies have shown that the genetic influences on disease risk are modified by environmental factors and this phenomenon is coined as gene-environment ($G \times E$) interaction (Liu et al. 2013; Sharafeldin et al. 2015; Wu et al. 2017). However, most of existing interaction analysis methods assume a linear relationship between a disease response and the genetic, and environmental factors as well as their interactions, which cannot assess the varying (or dynamic) patterns of genetic effects responsive to environmental changes (Ma et al. 2011). For example, a preliminary analysis of the skin cutaneous melanoma (SKCM) data (For more details of this dataset, see Section 4) from The Cancer Genome Atlas (TCGA) provides strong evidence of nonlinear $G \times E$ interactions, in the sense that the effects of genes CCNB1 and POU5F1B on the quantiles of Breslow's thickness fluctuate according to patients' age. It means that the assumption of linear $G \times E$ interaction does not hold for this dataset.

In recent years, there has been a rapidly growing literature on the extension of linear quantile regression to overcome the limitation of parametric assumptions, including both nonparametric and semiparametric quantile regression. Among them, the semiparametric varying coefficient model is the most frequently used approach, due to its appeal that it inherits the simple structure of the linear model and retains the flexibility of the varying coefficient model. Some important works along the line of semiparametric quantile regression models include Kai et al. (2011), Cai and Xiao (2012), Lian (2015), Shen and Liang (2017), Cai et al. (2018). In applications, however, it is difficult to decide which coefficients are really varying and which are not, and incorrect identification of nonparametric and parametric coefficients may result in substantial biases and poor predictions.

One possible approach to dealing with this uncertainty of models is to use model averaging, which fits a number of candidate models and combines them according to some criterion (Yang 2001; Hjort and Claeskens 2003; Hansen 2007). If we consider all candidate models with various combinations of parametric and nonparametric components, it will be quite computationally demanding even for only a few covariates. In this study, we consider model averaging for quantile regression with semiparametric candidate models. This is motivated by appealing properties of quantile regression and desirable flexibility of semiparametric varying coefficient models. We note that this research is more than a direct application of model averaging. In particular, we take a set of semiparametric varying coefficients quantile regression models as the candidates, where each sub-model involves only one varying

coefficient and treats the others as constant coefficients. These candidate models are more flexible to describe the nonlinear interactions than the linear parametric candidate models (Lu and Su 2015). And the semiparametric one-varying-coefficient models show preferable prediction performance to the univariate sub-models of Li et al. (2015a). Although averaging similar candidate models has been studied in the framework of mean regression (Li et al. 2018a), weight choice of quantile regression model averaging is significantly more complicated than that of the least squares estimation, because we do not have the usual bias-variance decomposition for the MSE-based evaluation criterion, and it is difficult to define a Mallows-type criterion as in Hansen (2007). Accordingly, new and more challenging technical developments are needed for selecting weights optimally. We propose a leave-more-out cross-validation weight choice criterion by minimizing the quantile prediction error (QPE), which is quite different from Li et al. (2018a). Indeed, the distinct natures of QPE-based criterion from MSE-based criterion present a non-trivial work to calculate the optimal weight and derive asymptotic properties for our quantile regression model averaging. Numerical studies show that our work provides a practically useful new model averaging approach for quantile regression with improved prediction accuracy and computational efficiency.

The remainder of the paper is organized as follows. Section 2 introduces the basic setup of model averaging for semiparametric varying coefficient quantile regression models and proposes weight choice methods. Theoretical properties are also investigated. Sections 3 and 4 present the numerical results in simulation and real data examples, respectively. Section 5 contains concluding remarks. The proofs of the theorems are in Appendix. Additional simulation results can be found in supplementary materials.

## 2 Methodology

In this work, we assume that variable selection has been done and focus on the uncertainty of coefficients. Let $Y$ be the response variable of interest, $\mathbf{X} = (X_1, ..., X_p)^\top$ be a $p$-vector of covariates and $U$ is univariate. Here, we assume that $X_1 = 1$. For a given $\tau\,(0 < \tau < 1)$, we are interested in estimating the conditional quantile function of $Y$ given covariates $\mathbf{X} = \mathbf{x}, U = u$, $Q_\tau(\mathbf{x}, u) = \arg\min_a E\{\rho_\tau(Y - a)|\mathbf{X} = \mathbf{x}, U = u\}$, where $\rho_\tau(\varepsilon) = \varepsilon(\tau - I(\varepsilon \le 0))$ is the check loss function at $\tau \in (0, 1)$. For the varying coefficient quantile regression model, the expression is as follows:

$$Q_\tau(\mathbf{x}, u) = \sum_{j=1}^{p} \alpha_j(u) x_j,$$

where $\alpha_j(\cdot)$ is the varying coefficient for the $j$th covariate, and we may write the parameter without $\tau$ for simplicity.

It is possible that some covariates have constant coefficients, and in this case, it is unwise to estimate the constant coefficient by treating it as a function. If we could

identify the constant coefficients, say the last $p-q$ covariates, we would consider the semiparametric varying coefficient model:

$$Q_\tau(\mathbf{x},u)=\sum_{j=1}^{q}\alpha_j(u)x_j+\sum_{k=q+1}^{p}\beta_k x_k.$$

However, it is not easy to identify which coefficients are really varying and which are not. One possible approach to dealing with this problem is to use model selection tools. Nevertheless, it can be highly unstable in the sense that a small change to the data may sometimes lead to drastically different outcomes (Nan and Yang 2014). This motivates us to use the idea of model averaging to reduce the loss of misspecification of models and avoid the additional uncertainty induced by model selection.

### 2.1 Candidate models and estimation

Without any constraints, there are a vast number of candidate models with various combinations of parametric and nonparametric components, and it is computationally infeasible to implement model averaging with all combinations of candidates. Considering the superiority of semiparametric varying coefficient model, we propose to approximate the conditional quantile function with the following sub-models:

$$Q_\tau^{(j)}(\mathbf{x},u)=\alpha_j(u)x_j+\sum_{k\neq j}^{p}\beta_{jk}x_k,\ \ j=1,...,p.$$

Note that each sub-model involves one varying coefficient and $p-1$ constant coefficients, which is preferable to the parametric candidates (Lu and Su 2015) for addressing the nonlinear interactions. This study shares the same motivation of improving flexibility and prediction accuracy via incorporating one-dimensional marginal regression functions as some existing studies in other contexts (Stock and Watson 2004; Li et al. 2015a). Different from these studies, we use the semiparametric one-varying-coefficient models, not the univariate sub-models, to better account for the potentially confounding effects among predictors.

Suppose that we have a random sample $\mathcal{D}_n=\{Y_i,\mathbf{X}_i,U_i\}_{i=1}^n$, where $\mathbf{X}_i=(X_{i1},...,X_{ip})^\top$. Then we can estimate $\alpha_j(\cdot)$ and $\beta_{jk}$ by minimizing the quantile loss function

$$\frac{1}{n}\sum_{i=1}^{n}\rho_\tau\left(Y_i-\alpha_j(U_i)X_{ij}-\sum_{k\neq j}^{p}\beta_{jk}X_{ik}\right).$$

Assume that $\alpha_j(\cdot)$ is twice continuously differentiable so that it can be approximated locally by $\alpha_j(u)\approx a_j+b_j(u-u_0)$, with $u$ in the neighborhood of a given point $u_0$. Write $\boldsymbol{\beta}_{(j)}=(\beta_{j1},...,\beta_{j(j-1)},\beta_{j(j+1)},...,\beta_{jp})^\top$, then the local quantile regression estimator for $\boldsymbol{\theta}_{(j)}=(a_j,b_j,\boldsymbol{\beta}_{(j)}^\top)^\top$ is obtained by minimizing

$$\ell_j(\boldsymbol{\theta}_{(j)}) = \frac{1}{n}\sum_{i=1}^{n}\rho_\tau\big(Y_i - (a_j + b_j(U_i - u))X_{ij} - \sum_{k\neq j}^{p}\beta_{jk}X_{ik}\big)K_{h_j}\big(U_i - u\big), \qquad (1)$$

where $u$ is a given point, $K_{h_j}(\cdot) = K(\cdot/h_j)/h_j$. Here $K(\cdot)$ is a kernel function and $h_j$ is a bandwidth.

It should be noted that we may use different bandwidths $h_j$ $(j = 1, ..., p)$ for different sub-models. But it can be quite time-consuming if we select optimal bandwidth for each sub-model, especially when $p$ is large. In our numerical studies, we use a simple bandwidth selection method for ease of computation, and all models adopt a common bandwidth. Nonetheless, one can use cross-validation to select the optimal bandwidth $h_j^{opt}$ by minimizing the nonparametric version of the Akaike information criterion in the context of quantile regression (Cai and Xu 2008).

Let $\widehat{\boldsymbol{\theta}}_{(j)} = (\widehat{a}_j, \widehat{b}_j, \widehat{\boldsymbol{\beta}}_{(j)}^{\top})^{\top}$ be the estimator of $\boldsymbol{\theta}_{(j)}$ obtained by minimizing (1) and $\widehat{\alpha}_j(u) = \widehat{a}_j$, we can estimate the $\tau$th conditional quantile of $Y$ given $\mathbf{X} = \mathbf{x}$ and $U = u$ by $\sum_{j=1}^{p} w_j \widehat{Q}_\tau^{(j)}(\mathbf{x}, u)$, where $\widehat{Q}_\tau^{(j)}(\mathbf{x}, u) = \widehat{\alpha}_j(u)x_j + \sum_{k\neq j}^{p}\widehat{\beta}_{jk}x_k$, and $\mathbf{w} = (w_1, ..., w_p)^{\top}$ is the weight vector, belonging to the set $\mathcal{H} = \left\{\mathbf{w} \in [0,1]^p : \sum_{j=1}^{p} w_j = 1\right\}$. Then, for a given $\mathbf{w}$, the conditional quantile of $Y$ can be approximated by

$$\widehat{Q}_\tau(\mathbf{x}, u) = \sum_{j=1}^{p} w_j \widehat{Q}_\tau^{(j)}(\mathbf{x}, u) = \sum_{j=1}^{p}\left(w_j\widehat{\alpha}_j(u) + \sum_{k\neq j}^{p} w_k\widehat{\beta}_{kj}\right)x_j.$$

***Remark 1*** If we denote $\widehat{\varphi}_j(u) = w_j\widehat{\alpha}_j(u) + \sum_{k\neq j}^{p} w_k\widehat{\beta}_{kj}$, a weighted version of coefficient by assigning different weights to a varying coefficient and $p-1$ constant coefficients, the estimated conditional quantile shares a similar form with varying coefficient model, i.e., $\sum_{j=1}^{p}\widehat{\varphi}_j(u)x_j$. But the unstructured varying coefficient model may overfit, especially when some varying coefficients should have been constant. In contrast, the model averaging estimator is more robust against model misspecification due to the weighted coefficients that adjust the relative importance of the varying and constant coefficients for the same predictor. Moreover, our method does not require a rigid designation of an oracle model but instead integrates a number of possible sub-models. The plausibility of each sub-model is then reflected by its weight in the averaging step. The weight, to a certain extent, may help us to identify the model structure. Hence, this flexible approach avoids making a fixed parametric or nonparametric model assumption and its flexibility with relative simplicity yields good prediction results, as witnessed in our numerical studies.

Asymptotic theory for semiparametric varying coefficient quantile regression models has been well studied in Kai et al. (2011), but it cannot be applied here due to misspecification of sub-models. In this study, we adopt the idea of Angrist et al. (2006) to derive the asymptotic properties for $\widehat{\boldsymbol{\theta}}_{(j)}$ when the corresponding sub-model is potentially misspecified.

Suppose that the relationship between $Y$ and $\{\mathbf{X}, U\}$ follows $Y = Q_\tau(\mathbf{X}, U) + \varepsilon_\tau$, where $\varepsilon_\tau$ is random error with conditional $\tau$th quantile being zero. Let $f(\cdot|\mathbf{x}, u)$ and

$F(\cdot|\mathbf{x},u)$ be the conditional density function and cumulative distribution function of $\varepsilon_\tau$, respectively. And the conditional density function of $Y$ given $\{\mathbf{x},u\}$ is denoted by $f_y(\cdot|\mathbf{x},u)$. Besides, we denote by $f_U(\cdot)$ the marginal density function of the covariate $U$. The kernel $K(\cdot)$ is chosen as a symmetric density function and we let

$$\mu_k = \int u^k K(u)du, \quad \nu_k = \int u^k K^2(u)du, \quad k = 0, 1, 2, \ldots$$

To proceed, for a given $u$, we define the pseudo-true parameter $\boldsymbol{\theta}^*_{(j)} = (a^*_j, b^*_j, \boldsymbol{\beta}^{*\top}_{(j)})^\top$

$$\boldsymbol{\theta}^*_{(j)} = \arg\min_{\boldsymbol{\theta}_{(j)}} E\Big[\rho_\tau(Y - \mathbf{X}^\top_{(j)}\boldsymbol{\theta}_{(j)})K_{h_j}(U-u)\Big],$$

where $\mathbf{X}_{(j)} = (X_j, (U-u)X_j, X_1, ..., X_{(j-1)}, X_{(j+1)}, ..., X_p)^\top$ and $\boldsymbol{\beta}^*_{(j)} = (\beta^*_{j1}, ..., \beta^*_{j(j-1)}, \beta^*_{j(j-1)}, \beta^*_{j(j+1)} ...s, \beta^*_{jp})^\top$. Define

$$\mathbf{A}_{(j)}(u) = E\Big[f(-r_{(j)}|\mathbf{X},U)\mathbf{X}_{(j)}\mathbf{X}^\top_{(j)}|U=u\Big], \ \mathbf{B}_{(j)}(u) = E\Big[\psi^2_\tau(\varepsilon_\tau + r_{(j)})\mathbf{X}_{(j)}\mathbf{X}^\top_{(j)}|U=u\Big],$$
$$\mathbf{C}_{(j)}(u) = E\Big[f(-r_{(j)}|\mathbf{X},U)\widetilde{\mathbf{X}}_{(j)}\widetilde{\mathbf{X}}^\top_{(j)}|U=u\Big], \ \mathbf{D}_{(j)}(u) = E\Big[\psi^2_\tau(\varepsilon_\tau + r_{(j)})\widetilde{\mathbf{X}}_{(j)}\widetilde{\mathbf{X}}^\top_{(j)}|U=u\Big],$$

where $r_{(j)} = Q_\tau(\mathbf{X},U) - \mathbf{X}^\top_{(j)}\boldsymbol{\theta}^*_{(j)}$ indicates the approximation bias for the $j$th sub-model, $\widetilde{\mathbf{X}}_{(j)} = (X_j, X_1, ..., X_{(j-1)}, X_{(j+1)}, ..., X_p)^\top$ and $\psi_\tau(u) = \tau - I(u<0)$. Let $\boldsymbol{\theta}^*_{(-j)} = (a^*_j, \boldsymbol{\beta}^{*\top}_{(j)})^\top$, $\widehat{\boldsymbol{\theta}}_{(-j)} = (\widehat{a}_j, \widehat{\boldsymbol{\beta}}^\top_{(j)})^\top$, and make the following assumptions:

(A1) $(Y_i, \mathbf{X}_i, U_i), i = 1, ..., n$ are i.i.d.

(A2) The conditional density $f_y(\cdot|\mathbf{x},u)$ exists a.s. and is bounded above by a finite constant $C_f$ and is continuous over its support a.s.;

(A3) $E\|\mathbf{X}_{(j)}\| < \infty$, and $\boldsymbol{\theta}^*_{(j)} \in \mathbb{R}^{p+1}$ is the unique solution to

$$E\Big\{\Big[\tau - I(Y \le \mathbf{X}^\top_{(j)}\boldsymbol{\theta}_{(j)})\Big]\mathbf{X}_{(j)}K_{h_j}(U-u)\Big\} = 0, \tag{2}$$

where $\|\cdot\|$ is the $L_2$ norm;

(A4) The random variable $U$ has a bounded support $\mathcal{U}$ and its density function $f_U(\cdot)$ has a continuous second derivative in its support (The derivative is left/right continuous at the right/left endpoint of $\mathcal{U}$);

(A5) The varying coefficients $\alpha(\cdot)$ have continuous second derivatives in $u \in \mathcal{U}$;

(A6) $K(\cdot)$ is a symmetric density with bounded support and satisfies a Lipschitz condition;

(A7) $f(\cdot|\mathbf{x},u)$ is bounded away from zero and has a continuous and uniformly bounded derivative;

(A8) For any $u \in \mathcal{U}$, there exist constants $\underline{C}_{A_{(j)}}$, $\overline{C}_{A_{(j)}}$, $\underline{C}_{C_{(j)}}$ and $\overline{C}_{C_{(j)}}$ such that

$$0 < \underline{C}_{A_{(j)}} \le \lambda_{\min}(\mathbf{A}_{(j)}(u)) \le \lambda_{\max}(\mathbf{A}_{(j)}(u)) \le C_f\lambda_{\max}(E[\mathbf{X}_{(j)}\mathbf{X}^\top_{(j)}]) \le \overline{C}_{A_{(j)}} < \infty,$$

and

$$0 < \underline{C}_{C_{(j)}} \le \lambda_{\min}(\mathbf{C}_{(j)}(u)) \le \lambda_{\max}(\mathbf{C}_{(j)}(u)) \le C_f \lambda_{\max}(E[\widetilde{\mathbf{X}}_{(j)}\widetilde{\mathbf{X}}_{(j)}^{\top}]) \le \overline{C}_{C_{(j)}} < \infty;$$

(A9) For any $u \in \mathcal{U}$, there exist constants $\underline{C}_{B_{(j)}}$, $\overline{C}_{B_{(j)}}$, $\underline{C}_{D_{(j)}}$ and $\overline{C}_{D_{(j)}}$ such that

$$0 < \underline{C}_{B_{(j)}} \le \lambda_{\min}(\mathbf{B}_{(j)}(u)) \le \lambda_{\max}(\mathbf{B}_{(j)}(u)) \le \overline{C}_{B_{(j)}} < \infty,$$

and

$$0 < \underline{C}_{D_{(j)}} \le \lambda_{\min}(\mathbf{D}_{(j)}(u)) \le \lambda_{\max}(\mathbf{D}_{(j)}(u)) \le \overline{C}_{D_{(j)}} < \infty;$$

(A10) $E(Q_{\tau}^4(\mathbf{X}, U)) < \infty$ and $\max_{1\le j\le p} E(X_j^8) \le C_{\mathbf{X}}$ for some $C_{\mathbf{X}} < \infty$.

**Theorem 1** *Suppose Assumptions* (A1)-(A10) *hold, for a given u and* $\tau$, *we have* $\|\widehat{\boldsymbol{\theta}}_{(j)} - \boldsymbol{\theta}^*_{(j)}\| = o_p(1)$.

Theorem 1 shows that the estimation error for each candidate model is at a negligible order. The proof can be found in Appendix. In Theorem 2, we give the asymptotic distribution of the estimators.

**Theorem 2** *Suppose* $\mathbf{C}_j(u)$ *is nonsingular for all* $1 \le j \le p$ *and* $u \in \mathcal{U}$, *and Assumptions* (A1)-(A10) *hold, for a given u and* $\tau$, *if* $h_j \to 0$ *and* $nh_j^2 \to \infty$ *as* $n \to \infty$, *then*

$$\sqrt{nh_j}\left(\widehat{\boldsymbol{\theta}}_{(-j)} - \boldsymbol{\theta}^*_{(-j)}\right) \xrightarrow{d} N\left(\mathbf{0}, \frac{v_0}{f_U(u)}\mathbf{C}_j^{-1}(u)\mathbf{D}_j(u)\mathbf{C}_j^{-1}(u)\right).$$

***Remark 2*** As discussed in Li et al. (2018a), another theoretical issue of interest is approximation error, which is closely related to the approximation bias $r_{(j)}$. Assumptions (A8) and (A9) imply that there is no candidate model for which the approximation error is zero. That is, all the candidate models are misspecified. But the approximation errors of all candidate models can be bounded according to Assumptions (A8) and (A9), which indicate that the difference between each candidate model and the true model cannot tend to infinity.

### 2.2 Weight choice criterion

An important issue with model averaging is how to choose weights. Here, we propose a weight choice method based on leave-more-out cross-validation. The procedure is as follows:

Fix $\tau \in (0, 1)$, let $1 \le n_0 \le n - 1$ be an integer (typically $n_0$ is of the same order as or slightly larger than $n - n_0$).

Step 1: Randomly partition the data into two parts: $\mathcal{D}^{(1)} = (Y_i, \mathbf{X}_i^{\top}, U_i)_{i=1}^{n_0}$ for training and $\mathcal{D}^{(2)} = (Y_i, \mathbf{X}_i^{\top}, U_i)_{i=n_0+1}^{n}$ for testing.

Step 2: Based on $\mathcal{D}^{(1)}$, obtain the estimate $\widehat{\boldsymbol{\theta}}_{(j),n_0} = (\widehat{a}_{j,n_0}, \widehat{b}_{j,n_0}, \widehat{\boldsymbol{\beta}}_{(j),n_0}^{\top})^{\top}$ by minimizing (1) for $j = 1, ..., p$.

Step 3: For $\mathcal{D}^{(2)}$, compute the predicted quantiles $\widehat{Q}_{\tau,n_0}^{(j)}(\mathbf{X}_i, U_i) = \widehat{a}_{j,n_0} X_{ij} + \sum_{k\neq j}^{p} \widehat{\beta}_{jk,n_0} X_{ik}$ for $i = n_0 + 1, ..., n$.

Step 4: Compute the weights by minimizing the loss function

$$CV_{n_0}(\mathbf{w}) = \frac{1}{n - n_0} \sum_{i=n_0+1}^{n} \rho_{\tau}\left( Y_i - \sum_{j=1}^{p} w_j \widehat{Q}_{\tau,n_0}^{(j)}(\mathbf{X}_i, U_i) \right), \tag{3}$$

over the set $\mathcal{H}$. And denote $\widehat{\mathbf{w}}^{*} = \arg\min_{\mathbf{w}\in\mathcal{H}} CV_{n_0}(\mathbf{w})$.

Step 5: Repeat Steps (1)-(4) $K$-1 more times and average the weights over $K$ random permutations. Denote the averaging weights as $\widehat{\mathbf{w}}$, and the final model averaging estimator of the $\tau$th conditional quantile of $Y$ given $\mathbf{X} = \mathbf{x}$ and $U = u$ is given by $\widehat{Q}_{\tau}(\mathbf{x}, u) = \sum_{j=1}^{p} \widehat{w}_j \widehat{Q}_{\tau,n_0}^{(j)}(\mathbf{x}, u)$.

***Remark 3*** Different from the adaptive quantile regression by mixing (AQRM) in Shan and Yang (2009), the proposed weights have no explicit expression. But AQRM is more time-consuming due to the selection of tuning parameter. Here, we can convert the constrained minimization problem (3) into a linear programming problem:

$$\min_{\mathbf{w},\mathbf{u},\mathbf{v}} \left\{ \tau \mathbf{1}_{n-n_0}^{\top} \mathbf{u} + (1-\tau)\mathbf{1}_{n-n_0}^{\top} \mathbf{v} \,\middle|\, \sum_{j=1}^{p} w_j \widehat{Q}_{\tau,n_0}^{(j)}(\mathbf{X}_i, U_i) + u_i - v_i = y_i, \right.$$
$$\left. i = n_0 + 1, ..., n \right\}$$
$$\text{subject to} \ : \ 0 \le u_i, \ 0 \le v_i, \ \text{ for } \ i = n_0 + 1, ..., n,$$
$$0 \le w_j \le 1, \ \sum_{j=1}^{p} w_j = 1, \ \text{ for } \ j = 1, ..., p.$$

This linear programming can be implemented in standard software, for example, the *linprog* package in R and the algorithm *linprog* in MATLAB.

***Remark 4*** It is natural as done in Li et al. (2018a) to consider the MSE-based weight choice criterion in the least squares regression, since MSE balances the asymptotic bias and variance in a good way. But it may not be a suitable criterion choice for quantile regression as pointed out by Lu and Su (2015). In the framework of parametric quantile regression, Lu and Su (2015) proposed a jackknife model averaging estimator which selects the weights by minimizing a leave-one-out cross-validation criterion. As mentioned above, however, the linear parametric candidate models may not be rich enough to describe the nonlinear interactions, which motivates us to propose the cross-validation model averaging for semiparametric varying coefficient quantile regression model. When $n_0 = n - 1$, the proposed weight criterion reduces

to the jackknife weight, but numerical studies show that leave-more-out will help to boost the performance of prediction (See Zhan and Yang (2022) for some recent results on the choice of $n_0$ in cross-validation).

Based on Theorem 1, by $n_0/n = O_p(1)$, we have that $\|\widehat{\boldsymbol{\theta}}_{(j),n_0} - \boldsymbol{\theta}^*_{(j)}\| = o_p(1)$. Now we can establish the asymptotic behavior of the estimated weight $\widehat{\mathbf{w}}$. Let $\{Y, \mathbf{X}, U\}$ be an independent copy of $\{Y_i, \mathbf{X}_i, U_i\}$, and define the quantile prediction error (QPE) as

$$QPE_{\tau,n}(\mathbf{w}) = E\left[\rho_\tau\left(Y - \sum_{j=1}^{p} w_j \widehat{Q}^{(j)}_\tau(\mathbf{X}, U)\right)\middle|\mathcal{D}_n\right].$$

We suppress the subscript $\tau$ in $QPE_{\tau,n}(\mathbf{w})$ when $\tau$ is fixed. And we impose the following assumption:

(A11) For any $u \in \mathcal{U}$, there exist constants $\underline{C}_J$ and $\overline{C}_J$ such that

$$0 < \underline{C}_J \le \lambda_{\min}(\mathbf{J}(u)) \le \lambda_{\max}(\mathbf{J}(u)) \le C_f \lambda_{\max}(E[\widetilde{\mathbf{X}}_{(j)}\widetilde{\mathbf{X}}^\top_{(j)}]) \le \overline{C}_J < \infty,$$

where $\mathbf{J}(u) = E\left[f(-r(\mathbf{w})|\mathbf{X}, U)\widetilde{\mathbf{X}}_{(j)}\widetilde{\mathbf{X}}^\top_{(j)}|U = u\right]$ and $r(\mathbf{w}) = Q_\tau(\mathbf{X}, U) - \sum_{j=1}^{p} w_j \mathbf{X}^\top_{(j)}\boldsymbol{\theta}^*_{(j)}$.

**Theorem 3** *Suppose Assumptions* (A1)-(A11) *hold, as* $n \to \infty$, *for a given* $\tau$, *then* $\widehat{\mathbf{w}}$ *is asymptotically optimal in the sense that*

$$\frac{QPE_n(\widehat{\mathbf{w}})}{\inf_{\mathbf{w}\in\mathcal{H}} QPE_n(\mathbf{w})} = 1 + o_p(1).$$

**Theorem 4** *Let* $\mathcal{T} = [t, 1-t]$ *be a compact subset of* $(0, 1)$ *with* $0 < t \le 1/2$. *Suppose Assumptions* (A1)-(A11) *hold, as* $n \to \infty$, *then* $\widehat{\mathbf{w}}$ *is asymptotically optimal uniformly for* $\tau \in \mathcal{T}$ *in the sense that*

$$\sup_{\tau\in\mathcal{T}} \frac{QPE_{\tau,n}(\widehat{\mathbf{w}})}{\inf_{\mathbf{w}\in\mathcal{H}} QPE_{\tau,n}(\mathbf{w})} = 1 + o_p(1).$$

The proofs are presented in Appendix. The optimality property in Theorem 3 indicates that the optimal weight $\widehat{\mathbf{w}}$ is asymptotically equivalent to the best weight to minimize $QPE_n(\mathbf{w})$ over $\mathcal{H}$ for a single given $\tau$. Similar to Wang et al. (2021), Theorem 4 shows that the asymptotic optimality of $\widehat{\mathbf{w}}$ is uniform in the set of quantile indices.

## 3 Finite sample analysis by simulations

In this section, we conduct simulation studies to evaluate the performances of the proposed method. We simulate data from the following semiparametric varying coefficient model

$$Y = \alpha_1(U)X_1 + \cdots + \alpha_4(U)X_4 + \beta_5 X_5 + \cdots + \beta_{10}X_{10} + \varepsilon,$$

where $\boldsymbol{\beta} = (\beta_5, ..., \beta_{10})^\top = (0.5, 1, -1.5, 2, 3, -2.5)^\top$, $\alpha_1(u) = \cos(2\pi u)$, $\alpha_2(u) = \frac{2+u^2}{1+u^2}$, $\alpha_3(u) = \frac{2\exp(-0.5u^2)}{\exp(-0.5u^2)+1}$ and $\alpha_4(u) = \sin(6\pi u)$. The covariate $U$ is generated from the uniform distribution on [0, 1], $X_1 = 1$ and the covariate vector $(X_2, ..., X_{10})^\top$ follows $N(0, \Sigma)$ with $\Sigma = (0.5^{|i-j|})_{i,j=1}^{9}$. We consider two settings for error distributions, $\varepsilon \sim N(0, 1)$ in Setting I (homoscedastic case); $\varepsilon = \sum_{j=1}^{10} j^{-1}X_j\epsilon$ in Setting II (heteroscedastic case), where $\epsilon$ follows $N(0, 1)$ and is independent of $\mathbf{X}$.

In all sets of simulation studies, we generate 500 datasets, each with a training sample size of $n = 100, 200, 300$. For each replication, we generate $n_{\text{test}} = 500$ observations as the testing sample. And we use the local linear fitting method to estimate the varying coefficients for the competing methods. The Gaussian kernel is adopted here and the bandwidth we select is $h = c\widehat{\sigma}_u n^{-1/5}$, where $c$ is a constant and $\widehat{\sigma}_u$ is the standard deviation of the sample of observations $U$. In numerical studies, for simplicity, we use $c = 1.06$, which is the optimal bandwidth choice based on the Silverman's rule-of-thumb method (Silverman 1986). One may use cross-validation to select the optimal $c$ for practical use, but it is time-consuming. We consider the quantiles $\tau = 0.1, 0.2, 0.5, 0.8, 0.9$.

When evaluating the prediction performances of different methods, we consider several measures based on quantile prediction error (QPE) and absolute prediction error (APE), defined as

$$\text{QPE} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \rho_\tau\Big(Y_i - \widehat{Q}_\tau(\mathbf{X}_i, U_i)\Big),$$

$$\text{APE} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \Big|\widehat{Q}_\tau(\mathbf{X}_i, U_i) - Q_\tau(\mathbf{X}_i, U_i)\Big|,$$

respectively. Analogous to Zhu et al. (2019), the measures used include **normalized QPE**, which is calculated by dividing the QPE of one estimator by the smallest QPE among the methods considered; **Paired *t*-test**, which is used to test the difference between two methods in QPE; **optimality rate** of QPE (or APE), which is defined as the proportion of times of each methods producing the smallest QPE (or APE) across 500 replications. Here, we present the results for $n = 200$ due to the space limitation, while the results with $n = 100$ and 300 are available in the supplementary material. In general, they exhibit patterns that are similar to those shown here and it is not surprising that all methods tend to perform better when sample size increases.

**Case 1 (Different weight choice criteria)**. In this case, we investigate the performances of the following methods:

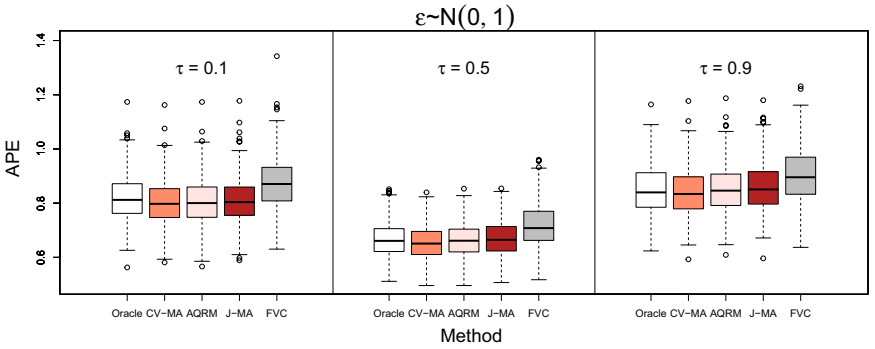
(a) CV-MA: the proposed method with $n_0 = n/2$.
(b) AQRM: proposed in Shan and Yang (2009).
(c) J-MA: Jackknife model averaging which selects the weights by minimizing a leave-one-out cross-validation criterion, that is $n_0 = n - 1$ in the proposed method.
(d) FVC: estimation based on the fully varying coefficient model without model selection.
(e) Oracle: estimation based on the true model.

Note that CV-MA, AQRM and J-MA are methods of model averaging with the same candidate models but different weight choice criteria, while FVC and Oracle are one model-based methods. The results of normalized QPE and $t$-test for the differences in QPE between CV-MA and alternatives are summarized in Table 1, where normalized QPE and $t$-test of QPE are computed over 500 replications. A positive $t$-statistic indicates that the estimator in the numerator produces a larger QPE than the estimator in the denominator. Besides, we plot the results of APE and optimality rate in Figure 1. The main conclusions are as follows:

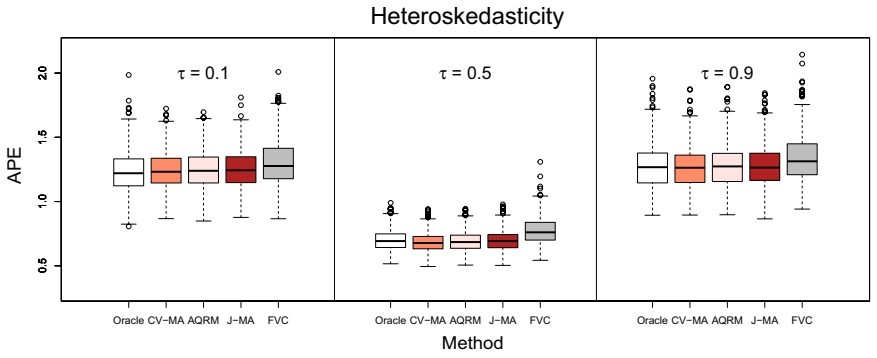
(1) From the results of normalized QPE, we find that FVC method produces the largest QPE because of overfitting and the oracle method also performs worse than model averaging methods in QPE. Among the three model averaging methods, CV-MA and AQRM perform better than J-MA. The results of $t$ statistics and $p$-values indicate that there are statistically significant differences in QPE between CV-MA and the alternatives for all quantiles. For almost all quantiles, CV-MA yields smaller QPE than AQRM, and the differences are significant, showing that CV-MA performs better than AQRM significantly.

(2) The box-plots of APE show that CV-MA enjoys advantages in prediction accuracy over other methods, since it produces the smallest APE in almost all situations. AQRM also has satisfactory performance in terms of APE, while slightly worse than CV-MA. J-MA still performs worst among three model averaging methods in APE. Remarkably, the oracle method is not always the victor in APE, and the smaller range of APE in the box-plots for model averaging methods indicates that the model averaging methods are more stable than the methods based on a single model, even the true one, which reconfirms the superiority of the approach of model averaging. Regarding the model averaging methods beating the oracle method, similar results can be found in Yang (2007) and Li et al. (2018a). Among all methods, FVC performs poorly and has the largest APE.

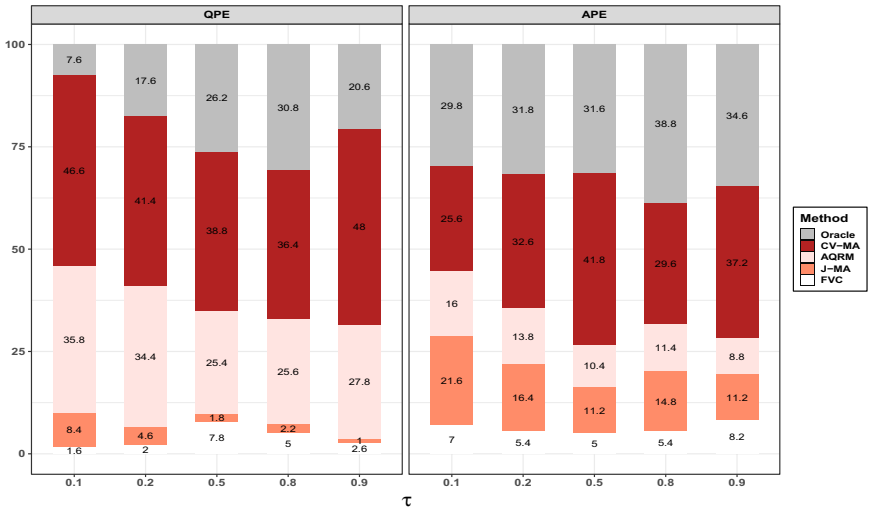
(3) The results of optimality rate of QPE show that CV-MA attains a proportion over 30% for almost all situations, meaning that in more than one-third of the replications, CV-MA yields the smallest QPE among all competitors in Case 1. The Oracle method performs best in terms of the optimality rate of APE and there is little chance for FVC to produce the smallest prediction error when $n = 200$. Besides, the results with $n = 100$ provided in supplementary material show that the Oracle




(a) Box-plot of APE for Setting I

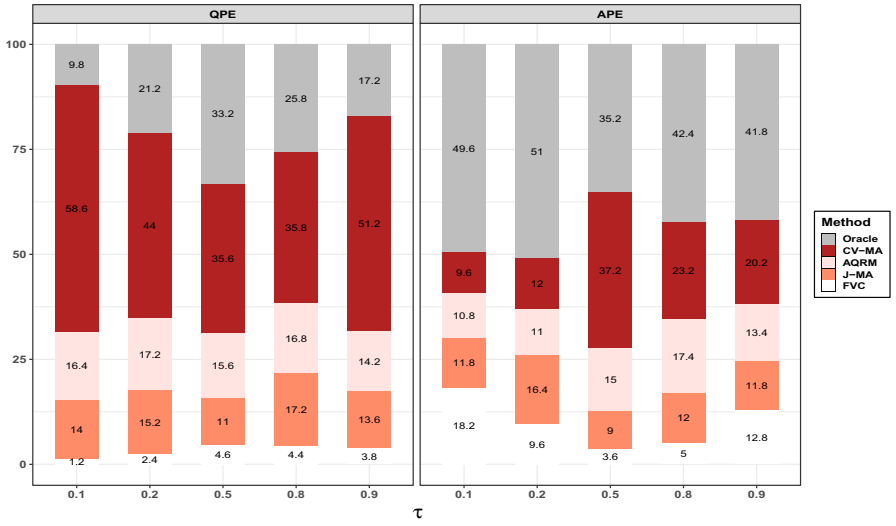



(b) Box-plot of APE for Setting II




(c) Stacked bar plot of optimality rates for Setting I

**Fig. 1** Simulation results for Case 1 with $n = 200$




(d) Stacked bar plot of optimality rates for Setting II

**Fig. 1** (continued)

method is not the shoo-in in terms of the optimality rate of APE. This is due to the fact that when the sample size is small, it is difficult to capture the nonparametric coefficient function in the true model. As the sample size increases, we are then gradually able to precisely estimate the four varying coefficients.

**Case 2 (Model misspecification)**. In this case, we assess the performances of methods (a)-(e) when the models are all misspecified. Specifically, we wrongly use $X_2$ as the index variable for all candidate models, while $U$ is mistaken for $X$. We denote the method (e) regarding misspecification with SVC. The results of normalized QPE and $t$-test over 500 replications are presented in Table 2. A positive $t$-statistic indicates that the estimator in the numerator produces a larger QPE than the estimator in the denominator. Besides, Figure 2 shows the results of APE and optimality rate.

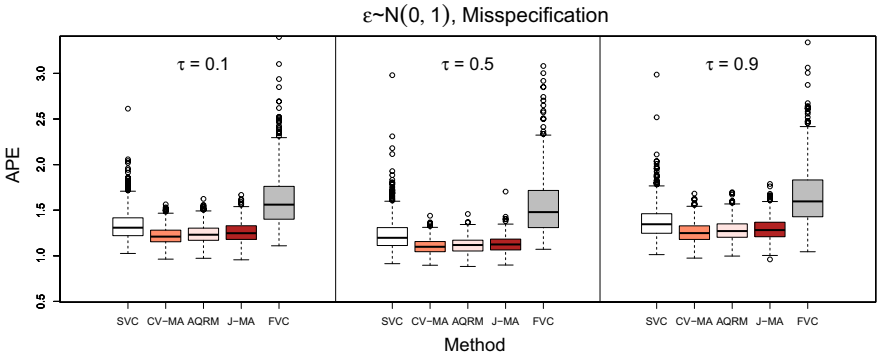
Across all results for Case 2, CV-MA is observed to have superior or similar performance compared to the alternatives. Specifically, CV-MA yields the smallest normalized QPE, but the differences in QPE between CV-MA and AQRM are not always significant. And they substantially outperform J-MA. FVC, not surprisingly, has the poorest prediction performances. SVC also loses to model averaging methods because of the model misspecification, which indicates that averaging method is more robust against model misspecification. The results of optimality rate show that CV-MA has a clear advantage over other methods because it has the largest optimality rate of APE for almost all scenarios.

**Case 3 (Different candidate models)**. In this case, we examine the performances of the following candidate models:

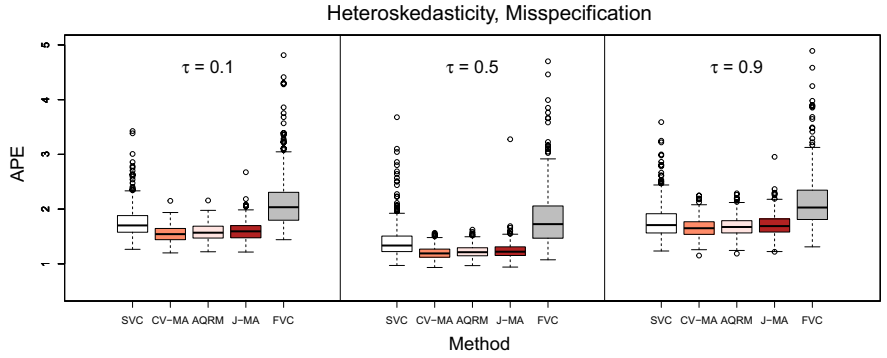
**Table 1** Simulation results for Case 1 with $n = 200$

| Setting | $\tau$ | Oracle | CV-MA | AQRM | J-MA | FVC |
|---|---|---|---|---|---|---|
| Normalized QPE | | | | | | |
| Setting I | 0.1 | 1.131 | **1.014** | 1.017 | 1.032 | 1.253 |
| | 0.2 | 1.056 | **1.010** | 1.013 | 1.024 | 1.127 |
| | 0.3 | 1.024 | **1.011** | 1.015 | 1.023 | 1.063 |
| | 0.4 | 1.034 | **1.015** | 1.018 | 1.029 | 1.095 |
| | 0.5 | 1.067 | **1.016** | 1.020 | 1.040 | 1.176 |
| Setting II | 0.1 | 1.109 | **1.011** | 1.017 | 1.032 | 1.224 |
| | 0.2 | 1.046 | **1.010** | 1.013 | 1.026 | 1.120 |
| | 0.3 | 1.021 | **1.011** | 1.015 | 1.021 | 1.068 |
| | 0.4 | 1.034 | **1.013** | 1.017 | 1.024 | 1.102 |
| | 0.5 | 1.070 | **1.013** | 1.020 | 1.036 | 1.168 |
| Setting | $\tau$ | | $\frac{\text{FVC}}{\text{CV-MA}}$ | $\frac{\text{Oracle}}{\text{CV-MA}}$ | $\frac{\text{J-MA}}{\text{CV-MA}}$ | $\frac{\text{AQRM}}{\text{CV-MA}}$ |
| Paired t-test | | | | | | |
| Setting I | 0.1 | $t$ | 35.508 | 25.564 | 14.098 | 5.219 |
| | | $p$-value | 0.000 | 0.000 | 0.000 | 0.000 |
| | 0.2 | $t$ | 31.181 | 18.058 | 15.574 | 6.656 |
| | | $p$-value | 0.000 | 0.000 | 0.000 | 0.000 |
| | 0.3 | $t$ | 23.88 | 8.869 | 17.496 | 9.519 |
| | | $p$-value | 0.000 | 0.000 | 0.000 | 0.000 |
| | 0.4 | $t$ | 24.786 | 9.117 | 15.687 | 7.273 |
| | | $p$-value | 0.000 | 0.000 | 0.000 | 0.000 |
| | 0.5 | $t$ | 28.204 | 15.532 | 16.538 | 7.891 |
| | | $p$-value | 0.000 | 0.000 | 0.000 | 0.000 |
| Setting II | 0.1 | $t$ | 32.063 | 21.912 | 17.221 | 6.600 |
| | | $p$-value | 0.000 | 0.000 | 0.000 | 0.000 |
| | 0.2 | $t$ | 27.582 | 15.229 | 18.339 | 8.193 |
| | | $p$-value | 0.000 | 0.000 | 0.000 | 0.000 |
| | 0.3 | $t$ | 25.638 | 7.483 | 15.113 | 8.431 |
| | | $p$-value | 0.000 | 0.000 | 0.000 | 0.000 |
| | 0.4 | $t$ | 26.871 | 11.181 | 12.533 | 5.228 |
| | | $p$-value | 0.000 | 0.000 | 0.000 | 0.000 |
| | 0.5 | $t$ | 27.473 | 17.445 | 18.089 | 8.423 |
| | | $p$-value | 0.000 | 0.000 | 0.000 | 0.000 |

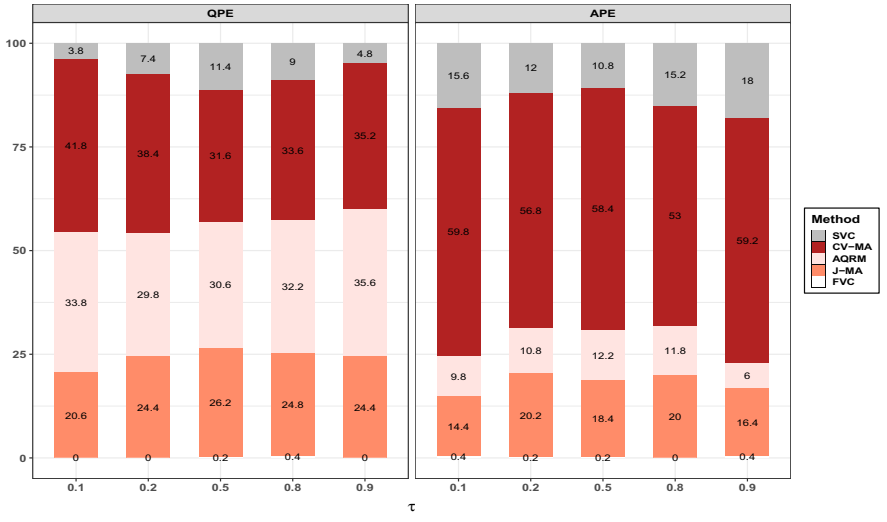
(a) Candidate A: the proposed semiparametric one-varying-coefficient candidate models.
(b) Candidate B: the nested-varying-coefficient candidate models containing true model (Hansen 2007), i.e., the $j$-th candidate model is given by $Q_\tau^{(j)}(\mathbf{x}, u) = \sum_{l=1}^{j} \alpha_{jl}(u)x_l + \sum_{k=j+1}^{p} \beta_{jk}x_k,\ j = 1, ..., p.$
(c) Candidate C: the marginal candidate models (Li et al. 2015a), $Q_\tau^{(j)}(\mathbf{x}, u) = \alpha_j(u)x_j, j = 1, ..., p.$




(a) Box-plot of APE for Setting I

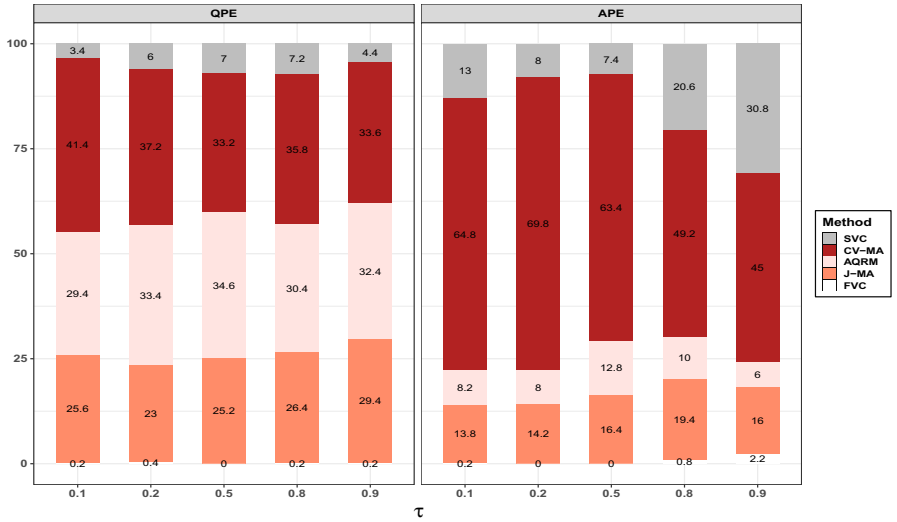



(b) Box-plot of APE for Setting II




(c) Stacked bar plot of optimality rates for Setting I

**Fig. 2** Simulation results for Case 2 with $n = 200$




(d) Stacked bar plot of optimality rates for Setting II

**Fig. 2** (continued)

Here, we adopt the same weight choice criterion with CV-MA for Candidate A, B and C. The results of normalized QPE and $t$-test for heteroscedastic case over 500 replications are presented in Table 3. A positive $t$-statistic indicates that the estimator in the numerator produces a larger QPE than the estimator in the denominator. Also, the results of APE and optimality rate are shown in Figure 3. We can draw the following conclusions from the results:

CV-MA with Candidate A always yields the smallest normalized QPE among all candidates, and Candidate B performs slightly better than Candidate A in terms of APE across 500 replications, and yet, somewhat worse considering the optimality rate. The optimality rates of QPE demonstrate that Candidate A is the favorite of all candidates, because it most likely yields the smallest prediction error among the three methods. Candidate C has the poorest performance due to ignoring the potentially confounding effects among predictors. In summary, the performance of one-varying-coefficient candidate models (all misspecified) indicates better-performing deals over other model sets even for the true-model-involved nested-varying-coefficient models in terms of QPE. Although Candidate A performs slightly worse than Candidate B in terms of APE, in practice, it is difficult to determine the order of the variables when applying the nested candidate models. Thus, Candidate A is more desirable than other candidates because it is relatively easy to implement and has a satisfactory prediction accuracy.

**Case 4 (Time cost)**. In this case, we compare the computing time of CV-MA and AQRM. For AQRM, the tuning parameter is taken from $\lambda = \min\{\tau, 1-\tau\} \times \{0, 0.5, 1, 3, 6, 10\}$. We select the optimal $\lambda$ as the one that produces the smallest Weighted Integrated Coverage Error (WICE) defined in Shan and Yang (2009) by cross-validation. Furthermore, to approximate the integrals in

**Table 2** Simulation results for Case 2 with $n = 200$

| Setting | $\tau$ | SVC | CV-MA | AQRM | J-MA | FVC |
|---|---|---|---|---|---|---|
| Normalized QPE | | | | | | |
| Setting I | 0.1 | 1.309 | **1.014** | 1.015 | 1.046 | 1.782 |
| | 0.2 | 1.176 | **1.011** | 1.012 | 1.031 | 1.505 |
| | 0.3 | 1.109 | **1.008** | 1.009 | 1.021 | 1.348 |
| | 0.4 | 1.181 | **1.012** | 1.013 | 1.034 | 1.543 |
| | 0.5 | 1.329 | **1.019** | 1.021 | 1.050 | 1.870 |
| Setting II | 0.1 | 1.372 | **1.016** | 1.018 | 1.054 | 1.867 |
| | 0.2 | 1.230 | **1.013** | 1.014 | 1.041 | 1.592 |
| | 0.3 | 1.138 | **1.008** | 1.009 | 1.027 | 1.407 |
| | 0.4 | 1.220 | **1.013** | 1.014 | 1.034 | 1.597 |
| | 0.5 | 1.374 | **1.021** | 1.023 | 1.055 | 1.917 |
| Setting | $\tau$ | | $\frac{\text{FVC}}{\text{CV-MA}}$ | $\frac{\text{SVC}}{\text{CV-MA}}$ | $\frac{\text{J-MA}}{\text{CV-MA}}$ | $\frac{\text{AQRM}}{\text{CV-MA}}$ |
| Paired t-test | | | | | | |
| Setting I | 0.1 | $t$ | 28.138 | 16.587 | 7.800 | 2.246 |
| | | $p$-value | 0.000 | 0.000 | 0.000 | 0.025 |
| | 0.2 | $t$ | 26.684 | 15.689 | 7.076 | 3.283 |
| | | $p$-value | 0.000 | 0.000 | 0.000 | 0.001 |
| | 0.3 | $t$ | 27.946 | 16.351 | 8.541 | 2.270 |
| | | $p$-value | 0.000 | 0.000 | 0.000 | 0.024 |
| | 0.4 | $t$ | 26.351 | 15.895 | 9.342 | 1.786 |
| | | $p$-value | 0.000 | 0.000 | 0.000 | 0.075 |
| | 0.5 | $t$ | 24.915 | 18.547 | 8.753 | 0.949 |
| | | $p$-value | 0.000 | 0.000 | 0.000 | 0.343 |
| Setting II | 0.1 | $t$ | 27.471 | 15.851 | 6.293 | 4.785 |
| | | $p$-value | 0.000 | 0.000 | 0.000 | 0.000 |
| | 0.2 | $t$ | 25.529 | 15.329 | 5.010 | 3.134 |
| | | $p$-value | 0.000 | 0.000 | 0.000 | 0.002 |
| | 0.3 | $t$ | 23.105 | 16.021 | 6.050 | 1.959 |
| | | $p$-value | 0.000 | 0.000 | 0.000 | 0.051 |
| | 0.4 | $t$ | 21.511 | 15.122 | 8.173 | 3.878 |
| | | $p$-value | 0.000 | 0.000 | 0.000 | 0.000 |
| | 0.5 | $t$ | 20.605 | 17.229 | 6.302 | 2.805 |
| | | $p$-value | 0.000 | 0.000 | 0.000 | 0.005 |

WICE, we need to calculate coverage error under a number of discrete $\tau$ values. In this study, we choose $\tau \in \{0.1, 0.2, ..., 0.9\}$. All these reasons force AQRM to spend more time in estimation than CV-MA. For the sake of fairness, we record the time that the two methods use in Case 1 in a laptop with 2.4 GHz Intel Core i5 and 8GB of RAM. The results are shown in Table 4 with different sample sizes or dimensions. The results strengthen the evidence that AQRM is more time-consuming due to the selection of tuning parameter.

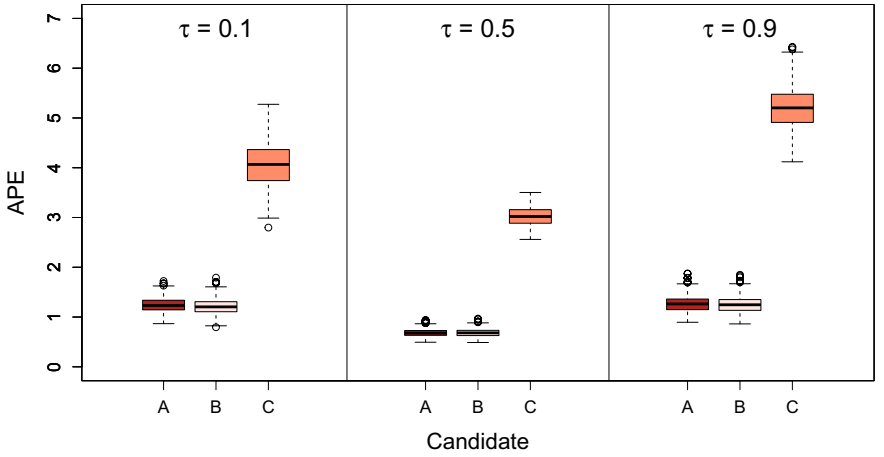
**Table 3** Simulation results for Case 3 with $n = 200$

| $\tau$ | Candidate A | Candidate B | Candidate C |
|---|---|---|---|
| Normalized QPE | | | |
| 0.1 | **1.005** | 1.063 | 2.528 |
| 0.2 | **1.008** | 1.024 | 2.816 |
| 0.3 | **1.007** | 1.009 | 2.624 |
| 0.4 | **1.009** | 1.013 | 2.950 |
| 0.5 | **1.009** | 1.029 | 2.792 |
| $\tau$ | | $\frac{\text{Candidate B}}{\text{Candidate A}}$ | $\frac{\text{Candidate C}}{\text{Candidate A}}$ |
| Paired t-test | | | |
| 0.1 | $t$ | 17.204 | 147.678 |
| | $p$-value | 0.000 | 0.000 |
| 0.2 | $t$ | 9.340 | 205.204 |
| | $p$-value | 0.000 | 0.000 |
| 0.3 | $t$ | 2.379 | 281.049 |
| | $p$-value | 0.000 | 0.018 |
| 0.4 | $t$ | 3.641 | 256.827 |
| | $p$-value | 0.000 | 0.000 |
| 0.5 | $t$ | 9.700 | 182.983 |
| | $p$-value | 0.000 | 0.000 |

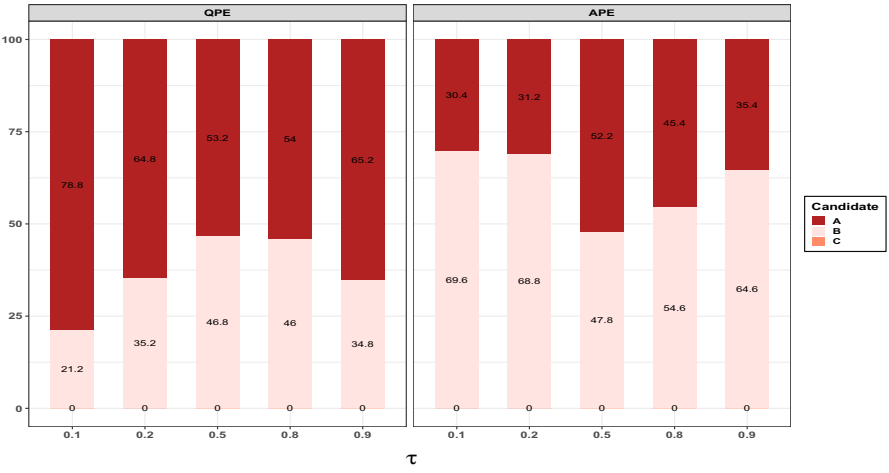
## 4 Analysis of SKCM data

In this example, we illustrate the proposed methods by analyzing the cutaneous melanoma data from The Cancer Genome Atlas (TCGA). This dataset is available at https://cancergenome.nih.gov/, provisional using R package *cgdsr*. TCGA data have a high quality and all-around omics measurements, enabling more accurate disease diagnosis and prognosis. The SKCM data contains information of 345 patients suffering from the skin cutaneous melanoma. For each patient, the value of Breslow's thickness (BRESLOW) is collected, which is a continuous variable and has been widely used as a prognostic indicator for melanoma. Besides, mRNA gene expressions that have been normalized (measurements on 18934 Z-scores), quantifying the relative expressions of tumor samples with respect to normal, as well as some environmental measurements such as Age, Gender, tumor pathological stage are also available for analysis. We take the continuous variable Age as the index variable $U$ to capture the gene-environment interactions. Besides, we carry out a logarithmic transformation to scale Breslow feature, a min–max normalization to Age feature and a standardized transformation to other features treated as predictors.

There have been several papers devoted to analyzing the SKCM data. Usually, the goal of the analysis is to identify the genetic factors affecting the Breslow's thickness. Considering the number of cancer-related genes is not expected to be large, we only include genes that may shed light on the molecular mechanisms. To deal with ultrahigh dimensionality, we could first adopt a marginal screening procedure




(a) Box-plot of APE




(b) Stacked bar plot of optimality rates

**Fig. 3** Simulation results for Case 3 with $n = 200$

**Table 4** The time cost (min) of CV-MA and AQRM

| | $n = 50$ | $n = 100$ | $n = 200$ |
|---|---|---|---|
| CV-MA | 0.371 | 1.013 | 3.728 |
| AQRM | 6.432 | 17.070 | 59.948 |
| | $p = 5$ | $p = 10$ | $p = 15$ |
| CV-MA | 0.432 | 1.014 | 1.565 |
| AQRM | 7.681 | 17.066 | 28.152 |

or model selection methods to screen out irrelevant genes. However, the important variables selected by different selection methods are quite different. Here, we just turn to sparsity oriented importance learning (SOIL) proposed in Ye et al. (2018) for help. In this study, we exclude the genes with missing values and use the genetic measurements selected by the methods proposed in Winnepenninckx et al. (2006), Li et al. (2015b), Wu et al. (2017), Chai et al. (2017), Li et al. (2018b), Xu et al. (2018) and as the candidate covariates $\mathbf{X}$ and adopt SOIL importance to narrow down the set of covariates for reaching a final model with sound considerations. Consequently, according to the variable importance measures, there are $p = 20$ variables with top 20 SOIL scores involved in our model averaging.

We compare the proposed methods with the fully varying coefficient model (FVC) that takes the coefficients of $\mathbf{X}$ as functions of variable $U$, as well as the method of the semiparametric varying coefficient model (SVC), where the coefficients are determined roughly as follows: we first estimate the coefficients using a fully varying coefficient model, then treat genes with obviously nonlinear coefficients as covariates of nonparametric components and others as parametric components. The estimates of varying coefficients for the fully varying coefficients model visually suggest that the coefficients of genes BUB1, CENPF, CEP295, D2HGDH, IFI44 and PTMA could be taken as linear functions. Finally, these genes are included in the semiparametric varying coefficient model in the form of constant

**Table 5** Results for SKCM data

| $\tau$ | SVC | CV-MA | AQRM | J-MA | FVC |
|---|---|---|---|---|---|
| Normalized QPE | | | | | |
| 0.1 | 2.303 | **1.008** | 1.013 | 1.067 | 2.972 |
| 0.2 | 1.956 | 1.007 | **1.004** | 1.138 | 2.302 |
| 0.3 | 1.757 | **1.002** | 1.008 | 1.125 | 2.010 |
| 0.4 | 1.578 | 1.006 | **1.005** | 1.067 | 1.867 |
| 0.5 | 1.528 | **1.004** | 1.007 | 1.047 | 1.832 |
| 0.6 | 1.546 | 1.004 | **1.002** | 1.043 | 1.862 |
| 0.7 | 1.564 | **1.003** | 1.006 | 1.043 | 1.873 |
| 0.8 | 1.580 | 1.007 | **1.005** | 1.044 | 1.834 |
| 0.9 | 1.841 | 1.028 | **1.023** | 1.076 | 2.085 |
| Violation rate | | | | | |
| 0.1 | 0.352 | 0.247 | 0.247 | 0.262 | 0.378 |
| 0.2 | 0.382 | 0.303 | 0.304 | 0.311 | 0.405 |
| 0.3 | 0.424 | 0.358 | 0.358 | 0.366 | 0.431 |
| 0.4 | 0.461 | 0.414 | 0.416 | 0.420 | 0.471 |
| 0.5 | 0.509 | 0.483 | 0.483 | 0.486 | 0.514 |
| 0.6 | 0.550 | 0.564 | 0.564 | 0.566 | 0.557 |
| 0.7 | 0.579 | 0.622 | 0.622 | 0.624 | 0.588 |
| 0.8 | 0.620 | 0.678 | 0.679 | 0.663 | 0.630 |
| 0.9 | 0.647 | 0.729 | 0.732 | 0.707 | 0.648 |
| MVR | 0.137 | 0.083 | 0.083 | 0.090 | 0.142 |

coefficients and the rest are in the form of functions. In addition, AQRM, J-MA are also considered to analyze the dataset for comparison.

We randomly split the data into two parts, a training dataset of size $n = 200$ and a testing sample of size $n_{\text{test}} = 145$. We fix the kernel function to be the Gaussian kernel for all local linear fitting and choose the bandwidth as $h = 1.06\widehat{\sigma}_u n^{-1/5}$ for simplicity. For quantiles $\tau = 0.1, 0.2, ..., 0.9$, we calculate QPE for all methods. With real data, since it is difficult to objectively evaluate the performance with APE, instead, we calculate the out-of-sample violation rate (Kuester et al. 2006): $\widehat{p}_\tau = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \mathbf{1}\left\{ Y_i < \widehat{Q}_\tau(\mathbf{X}_i, U_i) \right\}$, which should be close to $\tau$ if $\widehat{Q}_\tau(\mathbf{X}_i, U_i)$ correctly estimate the conditional quantile of $Y_i$. And we also calculate the mean violation rate $\text{MVR} = \frac{1}{9} \sum_{i=1}^{9} |\widehat{p}_{\tau_i} - \tau_i|$ to evaluate the overall performances. The results are presented in Table 5.

From the results, we find that among model averaging methods, the performance of CV-MA is analogous to that of AQRM and outperforms J-MA in terms of both normalized QPE and MVR. FVC produces the largest normalized QPE and MVR as expected. And SVC gives similar violation rate with FVC and is less close to $\tau$ than that of model averaging methods, which again supports that model averaging of semiparametric varying coefficients model can improve the prediction accuracy in the analysis of SKCM data. Overall, the proposed approach has a prominent superiority in prediction accuracy.

## 5 Concluding remarks

In this paper, we propose a new model averaging estimation for conditional quantile based on a set of semiparametric varying coefficient models. Each candidate model involves only one nonparametric component and thus can be easily fitted using univariate smoothing. Although all candidate models may be wrong, the averaging of all sub-models can improve the prediction accuracy. Moreover, the model averaging estimator shares a form like a fully varying coefficient model but is more robust against model misspecification due to the weighted coefficients that adjust the relative importance of the varying and constant coefficients for the same predictor. Based on leave-more-out cross-validation, we provide a weight choice criterion, which is shown to be asymptotically optimal in the sense of minimizing the quantile prediction error. Numerical studies show that the proposed method works very well compared to several alternatives.

Because there is no closed-form solution for quantile regression, and all the submodels under investigation may be incorrectly specified, it is more difficult to establish the asymptotic optimality of the averaging estimator. In this work, for the theoretical properties, we considered the fixed dimensional case. But the problems with diverging $p$ deserve further studies. In supplementary materials, we conduct simulation studies to investigate the asymptotic distribution of the CV-MA estimator and find out that the distribution of the final estimator is approaching a normal distribution. However, more effort is needed to formally examine the asymptotic distribution of the final estimator $\widehat{Q}_\tau(\mathbf{x}, u)$. In addition, the current paper considers only the

semiparametric varying coefficient model with simple random samples, and extensions to missing data or survival data (Lin et al. 2016) remain for future research.

## Appendix

### A.1. Proof of Theorem 1

We first introduce the following lemma, which is a direct result of Mack and Silverman (1982) and will be used in our proofs.

**Lemma 1** *Let* $(\mathbf{X}_1, Y_1), ..., (\mathbf{X}_n, Y_n)$ *be i.i.d. random vectors, where* $Y_1, ..., Y_n$ *are scalar random variables. Assume that* $E|Y|^r < \infty$ *and* $\sup_{\mathbf{x}} \int |y|^r f(\mathbf{x}, y)dy < \infty$, *where* $f$ *denotes the joint density of* $(\mathbf{X}, Y)$. *Let* $K$ *be a bounded positive function with bounded support, satisfying a Lipschitz condition. Then,*

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

*provided that* $n^{2\eta-1}h \to \infty$ *for some* $\eta < 1 - r^{-1}$.

Now, we prove the results of Theorem 1. First, we introduce Knight's identity (Knight 1988), which will be used in the following proof,

$$\rho_\tau(u+v) - \rho_\tau(u) = v\psi_\tau(u) + \int_0^{-v}[I(u \le z) - I(u \le 0)]dz. \tag{4}$$

For given $u$, $\tau$, and $j$, define $\varepsilon_{\tau,i} = Y_i - Q_\tau(\mathbf{X}_i, U_i)$, $r_{i(j)} = Q_\tau(\mathbf{X}_i, U_i) - \mathbf{X}_{i(j)}^\top \boldsymbol{\theta}_{(j)}^*$. Recall that $\boldsymbol{\theta}_{(j)}^* = (a_j^*, b_j^*, \boldsymbol{\beta}_{(j)}^{*\top})^\top$. To simplify the notation, we use a shorthand $\varepsilon_i$ and $\varepsilon$ for $\varepsilon_{\tau,i}$ and $\varepsilon_\tau$, respectively. For $\boldsymbol{\theta}_{(j)} \in \mathbb{R}^{p+1}$, we define

$$G_j(\tau, \boldsymbol{\theta}_{(j)}) = \sum_{i=1}^{n}\left[\rho_\tau(Y_i - \mathbf{X}_{i(j)}^\top \boldsymbol{\theta}_{(j)})K_{h_j}(U_i - u)\right].$$

We will show that for any $s > 0$, there is a constant $M > 0$ such that for all $n$ sufficiently large, we have

$$P\left\{\inf_{\|\mathbf{v}_{(j)}\|=M} G_j(\tau, \boldsymbol{\theta}_{(j)}^s) > G_j(\tau, \boldsymbol{\theta}_{(j)}^*)\right\} \ge 1 - s, \tag{5}$$

where $\boldsymbol{\theta}_{(j)}^s = \boldsymbol{\theta}_{(j)}^* + \delta_s \mathbf{v}_{(j)}$, $\mathbf{v}_{(j)} = (v_1, ..., v_{p+1})^\top$, and $\delta_s = o(1)$. By the Knight's identity, we obtain

$$
\begin{aligned}
&G_j(\tau,\boldsymbol{\theta}^s_{(j)}) - G_j(\tau,\boldsymbol{\theta}^*_{(j)}) \\
&= \sum_{i=1}^{n}\Big[\rho_\tau(Y_i - \mathbf{X}^\top_{i(j)}\boldsymbol{\theta}^s_{(j)})K_{h_j}(U_i-u)\Big] - \sum_{i=1}^{n}\Big[\rho_\tau(Y_i - \mathbf{X}^\top_{i(j)}\boldsymbol{\theta}^*_{(j)})K_{h_j}(U_i-u)\Big] \\
&= -\delta_s\sum_{i=1}^{n}K_{h_j}(U_i-u)\psi_\tau(\varepsilon_i + r_{i(j)})\mathbf{X}^\top_{i(j)}\mathbf{v}_{(j)} \\
&\quad + \sum_{i=1}^{n}K_{h_j}(U_i-u)\int_0^{\delta_s\mathbf{X}^\top_{i(j)}\mathbf{v}_{(j)}}[I(\varepsilon_i \le -r_{i(j)}+z) - I(\varepsilon_i \le -r_{i(j)})]dz \\
&\equiv G_{j,1}\big(\mathbf{v}_{(j)}\big) + G_{j,2}\big(\mathbf{v}_{(j)}\big) + G_{j,3}\big(\mathbf{v}_{(j)}\big),
\end{aligned}
$$

where

$$
\begin{aligned}
G_{j,1}\big(\mathbf{v}_{(j)}\big) &= -\delta_s\sum_{i=1}^{n}K_{h_j}(U_i-u)\psi_\tau\big(\varepsilon_i + r_{i(j)}\big)\mathbf{X}^\top_{i(j)}\mathbf{v}_{(j)}, \\
G_{j,2}\big(\mathbf{v}_{(j)}\big) &= \sum_{i=1}^{n}E\left[K_{h_j}(U_i-u)\int_0^{\delta_s\mathbf{X}^\top_{i(j)}\mathbf{v}_{(j)}}[I(\varepsilon_i \le -r_{i(j)}+z) - I(\varepsilon_i \le -r_{i(j)})]dz\Big|\mathbf{X}_i, U_i\right], \\
G_{j,3}\big(\mathbf{v}_{(j)}\big) &= \sum_{i=1}^{n}K_{h_j}(U_i-u)\int_0^{\delta_s\mathbf{X}^\top_{i(j)}\mathbf{v}_{(j)}}[I(\varepsilon_i \le -r_{i(j)}+z) - I(\varepsilon_i \le -r_{i(j)})]dz \\
&\quad - \sum_{i=1}^{n}E\left[K_{h_j}(U_i-u)\int_0^{\delta_s\mathbf{X}^\top_{i(j)}\mathbf{v}_{(j)}}[I(\varepsilon_i \le -r_{i(j)}+z) - I(\varepsilon_i \le -r_{i(j)})]dz\Big|\mathbf{X}_i, U_i\right].
\end{aligned}
$$

By equation (2), we have $E\big[G_{j,1}\big(\mathbf{v}_{(j)}\big)\big] = 0$. By Assumptions (A6) and (A9),

$$
E\big(G_{j,1}\big(\mathbf{v}_{(j)}\big)\big)^2 \le C_K^2\delta_s^2\sum_{i=1}^{n}\mathbf{v}^\top_{(j)}\mathbf{B}_{(j)}(u)\mathbf{v}_{(j)} \le C_K^2\overline{C}_{B_{(j)}}n\delta_s^2\|\mathbf{v}_{(j)}\|^2,
$$

where $C_K$ is a finite positive constant. Hence $G_{j,1}\big(\mathbf{v}_{(j)}\big) = O_p\Big(\overline{C}^{1/2}_{B_{(j)}}\delta_s\sqrt{n}\Big)\|\mathbf{v}_{(j)})\|$. By Taylor expansion, we have

$$
\begin{aligned}
G_{j,2}\big(\mathbf{v}_{(j)}\big) &= \frac{1}{2}\delta_s^2\mathbf{v}^\top_{(j)}\sum_{i=1}^{n}K_{h_j}(U_i-u)[f(-r_{i(j)}|\mathbf{X}_i,U_i) + o(1)]\mathbf{X}_{i(j)}\mathbf{X}^\top_{i(j)}\mathbf{v}_{(j)} \\
&= \frac{1}{2}\delta_s^2\mathbf{v}^\top_{(j)}n\big[f_U(u)\mathbf{A}_{(j)}(u) + o(1)\big]\mathbf{v}_{(j)} \\
&\ge \frac{n\delta_s^2 f_U(u)}{2}\underline{C}_{A_{(j)}}\|\mathbf{v}_{(j)}\|^2.
\end{aligned}
$$

Analogous to $G_{j,1}\big(\mathbf{v}_{(j)}\big)$, by Assumption (A8), we obtain that $G_{j,3}\big(\mathbf{v}_{(j)}\big) = O_p(\overline{C}^{1/2}_{A_{(j)}}\delta_s\sqrt{n})\|\mathbf{v}_{(j)}\|$. Thus we get (5), which implies that with probability approaching to 1, there exists a local minimum $\widehat{\boldsymbol{\theta}}_{(j)}$ in the ball $\mathcal{B}_{M,\delta_s} = \Big\{\boldsymbol{\theta}^*_{(j)} + \delta_s\mathbf{v}_{(j)} : \|\mathbf{v}_{(j)}\| \le M\Big\}$ such that $\|\widehat{\boldsymbol{\theta}}_{(j)} - \boldsymbol{\theta}^*_{(j)}\| = O_p(\delta_s) = o_p(1)$. By the

convexity of $G_j(\tau, \boldsymbol{\theta}_{(j)})$, $\widehat{\boldsymbol{\theta}}_{(j)}$ is also the global minimum. Thus, Theorem 1 is proved. □

### A.2. Proof of Theorem 2

For given $\tau$ and $j$, recall that $\varepsilon_i = Y_i - Q_\tau(\mathbf{X}_i, U_i)$, $r_{i(j)} = Q_\tau(\mathbf{X}_i, U_i) - \mathbf{X}_{i(j)}^\top \boldsymbol{\theta}_{(j)}^*$. Let $\widehat{\boldsymbol{\omega}}_j = \sqrt{nh_j}\Big(\widehat{a}_j - a_j^*, \widehat{\boldsymbol{\beta}}_{(j)}^\top - \boldsymbol{\beta}_{(j)}^{*\top}, h_j(\widehat{b}_j - b_j^*)\Big)^\top$. It follows from Theorem 1 in Cai and Xu (2008) that

$$\widehat{\boldsymbol{\omega}}_j = -f_U^{-1}(u)\mathbf{S}_j^{-1}(u)\mathbf{W}_{n,j}(u) + o_p(1),$$

where

$$\mathbf{W}_{n,j}(u) = \frac{1}{\sqrt{nh_j}} \sum_{i=1}^{n} K_{h_j}(U_i - u)\psi_\tau(\varepsilon_i + r_{i(j)})\mathbf{X}_{i(j)}^\circ,$$
$$\mathbf{S}_j(u) = E\Big[f(-r_{(j)}|\mathbf{X}, U)\mathbf{X}_{(j)}^\circ \mathbf{X}_{(j)}^{\circ\top}|U = u\Big],$$

where $\mathbf{X}_{i(j)}^\circ = \big(X_{ij}, X_{i1}, \ldots, X_{i(j-1)}, X_{i(j+1)}, \ldots, X_{ip}, (U_i - u)X_{ij}/h_j\big)^\top$. So we have

$$\sqrt{nh_j}\Big(\widehat{\boldsymbol{\theta}}_{(-j)} - \boldsymbol{\theta}_{(-j)}^*\Big) = -f_U^{-1}(u)\mathbf{C}_{(j)}^{-1}(u)\widetilde{\mathbf{W}}_{n,j}(u) + o_p(1), \tag{6}$$

where $\widetilde{\mathbf{W}}_{n,j}(u) = \frac{1}{\sqrt{nh_j}} \sum_{i=1}^n K_{h_j}(U_i - u)\psi_\tau(\varepsilon_i + r_{i(j)})\widetilde{\mathbf{X}}_{i(j)}$, and $\widetilde{\mathbf{X}}_{i(j)} = (X_{ij}, X_{i1}, ..., X_{i(j-1)}, X_{i(j+1)}, ..., X_{ip})^\top$. Noting that $E\Big(\widetilde{\mathbf{W}}_{n,j}(u)\Big) = 0$ by (2), and

$$\begin{aligned}
\mathrm{Var}\Big(\widetilde{\mathbf{W}}_{n,j}(u)\Big) &= \frac{1}{nh_j} \sum_{i=1}^{n} E\Big[K_{h_j}^2(U_i - u)\psi_\tau^2(\varepsilon_i + r_{i(j)})\widetilde{\mathbf{X}}_{i(j)}\widetilde{\mathbf{X}}_{i(j)}^\top\Big] \\
&= \frac{1}{n} \sum_{i=1}^{n} v_0 f_U(u) E[\psi_\tau^2(\varepsilon + r_{(j)})\widetilde{\mathbf{X}}_{(j)}\widetilde{\mathbf{X}}_{(j)}^\top|U = u] + o(1) \\
&= v_0 f_U(u)\mathbf{D}_{(j)}(u).
\end{aligned}$$

Then for any $\epsilon > 0$, define $\eta_{i(j)} = 1/\sqrt{nh_j}K_{h_j}(U_i - u)\psi_\tau(\varepsilon_i + r_{i(j)})\widetilde{\mathbf{X}}_{i(j)}$, we have

$$\begin{aligned}
&\sum_{i=1}^{n} E\big\{\|\eta_{i(j)}\|^2 I\big[\|\eta_{i(j)}\| \geq \epsilon\big]\big\} \\
&= nE\big\{\|\eta_{i(j)}\|^2 I[\|\eta_{i(j)}\| \geq \epsilon]\big\} \leq n\big\{E\|\eta_{i(j)}\|^4\big\}^{1/2}\{P(\|\eta_{i(j)}\| \geq \epsilon)\}^{1/2} \\
&\leq n\epsilon^{-2}E\|\eta_{i(j)}\|^4.
\end{aligned}$$

Furthermore, by Assumptions (A6) and (A10),

$$
\begin{aligned}
E\|\eta_{i(j)}\|^4 =& (nh_j)^{-2}E\left\{K_{h_j}^4(U_i-u)\left[\mathrm{tr}\left(\psi_\tau^2(\varepsilon_i+r_{i(j)})\widetilde{\mathbf{X}}_{i(j)}\widetilde{\mathbf{X}}_{i(j)}^\top\right)\right]^2\right\}\\
\le& (nh_j)^{-2}C_K^4E\left\{\left[\mathrm{tr}\left(\widetilde{\mathbf{X}}_{i(j)}\widetilde{\mathbf{X}}_{i(j)}^\top\right)\right]^2\right\}\\
\le& (nh_j)^{-2}C_K^4E\left\|\widetilde{\mathbf{X}}_{i(j)}\right\|^4 = O\big((nh_j)^{-2}\big).
\end{aligned}
$$

Thus, $\sum_{i=1}^n E\big\{\|\eta_{i(j)}\|^2 I\big[\|\eta_{i(j)}\|\ge\epsilon\big]\big\} = O\Big((nh_j^2)^{-1}\Big) = o(1)$. According to the Lindeberg–Feller central limit theorem, we obtain

$$
\widetilde{\mathbf{W}}_{n,j}(u)\xrightarrow{d}N\big(0, v_0 f_U(u)\mathbf{D}_{(j)}(u)\big).
$$

By the Slusky's theorem, we have

$$
\sqrt{nh_j}\Big(\widehat{\boldsymbol{\theta}}_{(-j)}-\boldsymbol{\theta}^*_{(-j)}\Big)\xrightarrow{d}N\left(\mathbf{0}, \frac{v_0}{f_U(u)}\mathbf{C}_{(j)}^{-1}(u)\mathbf{D}_{(j)}(u)\mathbf{C}_{(j)}^{-1}(u)\right).
$$

Therefore, the proof of Theorem 2 is completed. □

### A.3. Proof of Theorem 3

To show the results, it suffices to show that $\sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{CV_{n_0}(\mathbf{w})-QPE_n(\mathbf{w})}{QPE_n(\mathbf{w})}\right| = o_p(1)$. For notation simplicity, for a given $\tau$, let $\widehat{Q}_{j,n_0}(\cdot)=\widehat{Q}^{(j)}_{\tau,n_0}(\cdot)$, $\widehat{Q}_j(\cdot)=\widehat{Q}^{(j)}_\tau(\cdot)$. By the definition of $CV_{n_0}(\mathbf{w})$ and $QPE_n(\mathbf{w})$, we have

$$
\begin{aligned}
CV_{n_0}(\mathbf{w})-QPE_n(\mathbf{w}) =& \frac{1}{n-n_0}\sum_{i=n_0+1}^n\left[\rho_\tau\left(Y_i-\sum_{j=1}^p w_j\widehat{Q}_{j,n_0}(\mathbf{X}_i,U_i)\right)-\rho_\tau(\varepsilon_i)\right]\\
&+\big[E\rho_\tau(\varepsilon)-QPE_n(\mathbf{w})\big]+\frac{1}{n-n_0}\sum_{i=n_0+1}^n\big[\rho_\tau(\varepsilon_i)-E\rho_\tau(\varepsilon)\big].
\end{aligned}
$$

Noting that $E\left[\Big(Q_\tau(\mathbf{X},U)-\sum_{j=1}^p w_j\widehat{Q}_j(\mathbf{X},U)\Big)\psi_\tau(\varepsilon)\Big|\mathcal{D}_n\right]=0$ and

$$
\begin{aligned}
&E\left[\int_0^{\sum_{j=1}^p w_j\widehat{Q}_j(\mathbf{X},U)-Q_\tau(\mathbf{X},U)}[I(\varepsilon\le z)-I(\varepsilon\le 0)]dz\Big|\mathcal{D}_n\right]\\
&\quad=E_{\mathbf{X}_i,U_i}\left\{\int_0^{\sum_{j=1}^p w_j\widehat{Q}_j(\mathbf{X}_i,U_i)-Q_\tau(\mathbf{X}_i,U_i)}[F(z|\mathbf{X}_i,U_i)-F(0|\mathbf{X}_i,U_i)]dz\right\},
\end{aligned}
$$

where $E_{\mathbf{X}_i,U_i}$ denote the expectation with respect to $\{\mathbf{X}_i,U_i\}$. Together with the Knight's identity, we get the following decomposition expression

$$
CV_{n_0}(\mathbf{w})-QPE_n(\mathbf{w}) = CV_1(\mathbf{w})+CV_2(\mathbf{w})+CV_3(\mathbf{w})+CV_4(\mathbf{w})+CV_5,
$$

where

$$CV_1(\mathbf{w}) = \frac{1}{n-n_0}\sum_{i=n_0+1}^{n}\left[(Q_\tau(\mathbf{X}_i,U_i) - \sum_{j=1}^{p} w_j\widehat{Q}_{j,n_0}(\mathbf{X}_i,U_i))\psi_\tau(\varepsilon_i)\right],$$

$$CV_2(\mathbf{w}) = \frac{1}{n-n_0}\sum_{i=n_0+1}^{n}\int_0^{\sum_{j=1}^{p} w_j\widehat{Q}_{j,n_0}(\mathbf{X}_i,U_i)-Q_\tau(\mathbf{X}_i,U_i)} [I(\varepsilon_i \le z) - I(\varepsilon_i \le 0)] - [F(z|\mathbf{X}_i,U_i) - F(0|\mathbf{X}_i,U_i)]dz,$$

$$CV_3(\mathbf{w}) = \frac{1}{n-n_0}\sum_{i=n_0+1}^{n}\left[\int_0^{\sum_{j=1}^{p} w_j\widehat{Q}_{j,n_0}(\mathbf{X}_i,U_i)-Q_\tau(\mathbf{X}_i,U_i)} [F(z|\mathbf{X}_i,U_i) - F(0|\mathbf{X}_i,U_i)]dz - E_{\mathbf{X}_i,U_i}\left\{\int_0^{\sum_{j=1}^{p} w_j\widehat{Q}_{j,n_0}(\mathbf{X}_i,U_i)-Q_\tau(\mathbf{X}_i,U_i)} [F(z|\mathbf{X}_i,U_i) - F(0|\mathbf{X}_i,U_i)]dz\right\}\right],$$

$$CV_4(\mathbf{w}) = \frac{1}{n-n_0}\sum_{i=n_0+1}^{n}\left[E_{\mathbf{X}_i,U_i}\left\{\int_0^{\sum_{j=1}^{p} w_j\widehat{Q}_{j,n_0}(\mathbf{X}_i,U_i)-Q_\tau(\mathbf{X}_i,U_i)} [F(z|\mathbf{X}_i,U_i) - F(0|\mathbf{X}_i,U_i)]dz\right\} - E_{\mathbf{X}_i,U_i}\left\{\int_0^{\sum_{j=1}^{p} w_j\widehat{Q}_j(\mathbf{X}_i,U_i)-Q_\tau(\mathbf{X}_i,U_i)} [F(z|\mathbf{X}_i,U_i) - F(0|\mathbf{X}_i,U_i)]dz\right\}\right],$$

$$CV_5 = \frac{1}{n-n_0}\sum_{i=n_0+1}^{n}\left[\rho_\tau(\varepsilon_i) - E\rho_\tau(\varepsilon)\right].$$

Next, we will show that

(i) $\min_{\mathbf{w}\in\mathcal{H}} QPE_n(\mathbf{w}) \ge E[\rho_\tau(\varepsilon)] - o_p(1)$;

(ii) $\sup_{\mathbf{w}\in\mathcal{H}} |CV_1(\mathbf{w})| = o_p(1)$;

(iii) $\sup_{\mathbf{w}\in\mathcal{H}} |CV_2(\mathbf{w})| = o_p(1)$;

(iv) $\sup_{\mathbf{w}\in\mathcal{H}} |CV_3(\mathbf{w})| = o_p(1)$;

(v) $\sup_{\mathbf{w}\in\mathcal{H}} |CV_4(\mathbf{w})| = o_p(1)$;

(vi) $CV_5 = o_p(1)$.

(i). Using (4) again, define $Q_j^*(\mathbf{X},U) = \widetilde{\mathbf{X}}_{(j)}^\top \boldsymbol{\theta}_{(-j)}^*$, we get

$$
\begin{aligned}
&QPE_n(\mathbf{w}) - E[\rho_\tau(\varepsilon + r(\mathbf{w}))] \\
&= E\left\{ \rho_\tau\left( \varepsilon + r(\mathbf{w}) + \sum_{j=1}^{p} w_j[Q_j^*(\mathbf{X}, U) - \widehat{Q}_j(\mathbf{X}, U)] \right) - \rho_\tau(\varepsilon + r(\mathbf{w})) \middle| \mathcal{D}_n \right\} \\
&= E\left\{ \int_0^{\sum_{j=1}^{p} w_j[\widehat{Q}_j(\mathbf{X},U) - Q_j^*(\mathbf{X},U)]} [I(\varepsilon + r(\mathbf{w}) \le z) - I(\varepsilon + r(\mathbf{w}) \le 0)]dz \middle| \mathcal{D}_n \right\} \\
&= E_{\mathbf{X},U}\left\{ \int_0^{\sum_{j=1}^{p} w_j[\widehat{Q}_j(\mathbf{X},U) - Q_j^*(\mathbf{X},U)]} [F(z - r(\mathbf{w})|\mathbf{X}, U) - F(-r(\mathbf{w})|\mathbf{X}, U)]dz \right\},
\end{aligned}
$$

where $E_{\mathbf{X},U}$ denote the expectation with respect to $\{\mathbf{X}, U\}$. By Taylor's expansion and Jensen's inequality, we have that

$$
\begin{aligned}
&\left| QPE_n(\mathbf{w}) - E[\rho_\tau(\varepsilon + r(\mathbf{w}))] \right| \\
&= \left| E_{\mathbf{X},U}\left\{ \int_0^{\sum_{j=1}^{p} w_j[\widehat{Q}_j(\mathbf{X},U) - Q_j^*(\mathbf{X},U)]} zf(-r(\mathbf{w})|\mathbf{X}, U)dz \right\} + o_p(1) \right| \\
&= \left| E_{\mathbf{X},U}\left\{ \frac{1}{2} f(-r(\mathbf{w})|\mathbf{X}, U) \left[ \sum_{j=1}^{p} w_j[\widehat{Q}_j(\mathbf{X}, U) - Q_j^*(\mathbf{X}, U)] \right]^2 \right\} + o_p(1) \right| \\
&\le \left| E_{\mathbf{X},U}\left\{ \frac{1}{2} f(-r(\mathbf{w})|\mathbf{X}, U) \sum_{j=1}^{p} w_j \left[ \widehat{Q}_j(\mathbf{X}, U) - Q_j^*(\mathbf{X}, U) \right]^2 \right\} + o_p(1) \right| \\
&= \left| \frac{1}{2} E_{\mathbf{X},U}\left\{ \sum_{j=1}^{p} w_j (\widehat{\boldsymbol{\theta}}_{(-j)} - \boldsymbol{\theta}^*_{(-j)})^\top E\left[ f(-r(\mathbf{w})|\mathbf{X}, U) \widetilde{\mathbf{X}}_{(j)} \widetilde{\mathbf{X}}_{(j)}^\top \right] (\widehat{\boldsymbol{\theta}}_{(-j)} - \boldsymbol{\theta}^*_{(-j)}) \right\} + o_p(1) \right| \\
&\le \left| \frac{1}{2} \overline{C}_J \max_{1 \le j \le p} \| \widehat{\boldsymbol{\theta}}_{(-j)} - \boldsymbol{\theta}^*_{(-j)} \| + o_p(1) \right|,
\end{aligned}
$$

the last inequality follows from Assumption (A11). Now it follows from Theorem 1 that,

$$
\left| QPE_n(\mathbf{w}) - E[\rho_\tau(\varepsilon + r(\mathbf{w}))] \right| \le o_p(1). \tag{7}
$$

Using the fact that $D(t) = E[\rho_\tau(\varepsilon + t) - \rho_\tau(\varepsilon)]$ has a global minimum at $t = 0$, we have $\min_{\mathbf{w} \in \mathcal{H}} E\left[\rho_\tau(\varepsilon + r(\mathbf{w}))\right] \ge E[\rho_\tau(\varepsilon)]$. By combining (7), we get

$$
\min_{\mathbf{w} \in \mathcal{H}} QPE_n(\mathbf{w})) \ge \min_{\mathbf{w} \in \mathcal{H}} E[\rho_\tau(\varepsilon + r(\mathbf{w}))] - o_p(1) \ge E[\rho_\tau(\varepsilon)] - o_p(1).
$$

(ii). Define $Q_j^*(\mathbf{X}_i, U_i) = \widetilde{\mathbf{X}}_{i(j)}^\top \boldsymbol{\theta}^*_{(-j)}$. By simple calculation, we have the following decomposition,

$$
\begin{aligned}
CV_1(\mathbf{w}) = & \frac{1}{n-n_0}\sum_{i=n_0+1}^{n}\left[Q_\tau(\mathbf{X}_i,U_i)-\sum_{j=1}^{p}w_jQ_j^*(\mathbf{X}_i,U_i)\right]\psi_\tau(\varepsilon_i)\\
& -\frac{1}{n-n_0}\sum_{i=n_0+1}^{n}\sum_{j=1}^{p}w_j\left[\widehat{Q}_{j,n_0}(\mathbf{X}_i,U_i)-Q_j^*(\mathbf{X}_i,U_i)\right]\psi_\tau(\varepsilon_i)\\
= & CV_{11}(\mathbf{w})-CV_{12}(\mathbf{w}).
\end{aligned}
$$

It is easy to show that $E(CV_{11}(\mathbf{w}))=0$ and $\mathrm{Var}(CV_{11}(\mathbf{w}))=O(1/(n-n_0))$, which implies that $CV_{11}(\mathbf{w})=o_p(1)$ for each $\mathbf{w}\in\mathcal{H}$. To show the uniform convergence, we consider the function class $\mathcal{F}=\{g(\varepsilon_i,\mathbf{X}_i,U_i;\mathbf{w}):\mathbf{w}\in\mathcal{H}\}$, where $g(\varepsilon_i,\mathbf{X}_i,U_i;\mathbf{w})=\left[Q_\tau(\mathbf{X}_i,U_i)-\sum_{j=1}^{p}w_jQ_j^*(\mathbf{X}_i,U_i)\right]\psi_\tau(\varepsilon_i)$. On $\mathcal{H}$, we define the metric $|\cdot|_1$ as $|\mathbf{w}-\tilde{\mathbf{w}}|_1=\sum_{j=1}^{p}|w_j-\tilde{w}_j|$, for any $\mathbf{w}=(w_1,...,w_p)\in\mathcal{H}$ and $\tilde{\mathbf{w}}=(\tilde{w}_1,...,\tilde{w}_p)\in\mathcal{H}$. Then, the $\epsilon$-covering number of $\mathcal{H}$ with respect to $|\cdot|_1$ is $\mathcal{N}(\epsilon,\mathcal{H},|\cdot|_1)=O(1/\epsilon^{p-1})$. Further,

$$
\begin{aligned}
|g(\varepsilon_i,\mathbf{X}_i,U_i;\mathbf{w})-g(\varepsilon_i,\mathbf{X}_i,U_i;\tilde{\mathbf{w}})| = & \left|\sum_{j=1}^{p}(w_j-\tilde{w}_j)Q_j^*(\mathbf{X}_i,U_i)\psi_\tau(\varepsilon_i)\right|\\
\le & C_\theta|\mathbf{w}-\tilde{\mathbf{w}}|_1\max_{1\le j\le p}\|\widetilde{\mathbf{X}}_{i(j)}\|,
\end{aligned}
$$

where $C_\theta=p\max_{1\le j\le p}\|\boldsymbol{\theta}^*_{(-j)}\|=O(p^{3/2})$ and $E\max_{1\le j\le p}\|\widetilde{\mathbf{X}}_{i(j)}\|<\infty$ by Assumption (A3). For a fix $p$, this yields that the $\epsilon$-bracketing number of $\mathcal{F}$ with respect to the $L_1$-norm is $\mathcal{N}_{[]}(\epsilon,\mathcal{F},L_1(P))\le C/\epsilon^{p-1}$ for some constant $C$. By Theorem 2.4.1 of Van der Vaart and Wellner (1996), we conclude that $\mathcal{F}$ is Glivenko–Cantelli. And it follows from Glivenko–Cantelli theorem that $\sup_{\mathbf{w}\in\mathcal{H}}|CV_{11}(\mathbf{w})|=o_p(1)$. By the Cauchy–Schwarz inequality,

$$
\begin{aligned}
\sup_{\mathbf{w}\in\mathcal{H}}|CV_{12}(\mathbf{w})| \triangleq & \sup_{\mathbf{w}\in\mathcal{H}}\left|\frac{1}{n-n_0}\sum_{i=n_0+1}^{n}\sum_{j=1}^{p}w_j\left[\widehat{Q}_{j,n_0}(\mathbf{X}_i,U_i)-Q_j^*(\mathbf{X}_i,U_i)\right]\psi_\tau(\varepsilon_i)\right|\\
\le & \sup_{\mathbf{w}\in\mathcal{H}}\sum_{j=1}^{p}w_j\frac{1}{n-n_0}\sum_{i=n_0+1}^{n}\left|\left[\widetilde{\mathbf{X}}_{i(j)}^\top(\widehat{\boldsymbol{\theta}}_{(-j),n_0}-\boldsymbol{\theta}^*_{(-j)})\right]\psi_\tau(\varepsilon_i)\right|\\
\le & \sum_{j=1}^{p}\frac{1}{n-n_0}\sum_{i=n_0+1}^{n}\left|\widetilde{\mathbf{X}}_{i(j)}^\top(\widehat{\boldsymbol{\theta}}_{(-j),n_0}-\boldsymbol{\theta}^*_{(-j)})\right|\\
\le & \sum_{j=1}^{p}\max_{n_0+1\le i\le n}\|\widetilde{\mathbf{X}}_{i(j)}\|\|\widehat{\boldsymbol{\theta}}_{(-j),n_0}-\boldsymbol{\theta}^*_{(-j)}\|,
\end{aligned}
$$

where $\widehat{\boldsymbol{\theta}}_{(-j),n_0}=(\widehat{a}_{j,n_0},\widehat{\boldsymbol{\beta}}_{(j),n_0}^\top)^\top$, then by Theorem 1 and Assumption (A3), we get $\sup_{\mathbf{w}\in\mathcal{H}}|CV_{12}(\mathbf{w})|=o_p(1)$.

(iii) To prove (iii), we rewrite $CV_2(\mathbf{w})=CV_{21}(\mathbf{w})+CV_{22}(\mathbf{w})$, where

$$CV_{21}(\mathbf{w}) = \frac{1}{n-n_0}\sum_{i=n_0+1}^{n}\int_0^{\sum_{j=1}^p w_j Q_j^*(\mathbf{X}_i,U_i)-Q_\tau(\mathbf{X}_i,U_i)} [I(\varepsilon_i \le z) - I(\varepsilon_i \le 0) - F(z|\mathbf{X}_i,U_i) + F(0|\mathbf{X}_i,U_i)]dz,$$

$$CV_{22}(\mathbf{w}) = \frac{1}{n-n_0}\sum_{i=n_0+1}^{n}\int_{\sum_{j=1}^p w_j Q_j^*(\mathbf{X}_i,U_i)-Q_\tau(\mathbf{X}_i,U_i)}^{\sum_{j=1}^p w_j \widehat{Q}_{j,n_0}(\mathbf{X}_i,U_i)-Q_\tau(\mathbf{X}_i,U_i)} [I(\varepsilon_i \le z) - I(\varepsilon_i \le 0) - F(z|\mathbf{X}_i,U_i) + F(0|\mathbf{X}_i,U_i)]dz.$$

Noting that $E[CV_{21}(\mathbf{w})] = 0$, $\mathrm{Var}(CV_{21}(\mathbf{w})) = O(1/n-n_0)$. Analogous to the proof of $CV_{11}(\mathbf{w})$, we can show that $\sup_{w\in\mathcal{H}}|CV_{21}(\mathbf{w})| = o_p(1)$. On the other hand,

$$|CV_{22}(\mathbf{w})| \le \frac{2}{n-n_0}\sum_{i=n_0+1}^{n}\sum_{j=1}^{p}\left|w_j\left[\widehat{Q}_{j,n_0}(\mathbf{X}_i,U_i) - Q_j^*(\mathbf{X}_i,U_i)\right]\right|,$$

similar to $CV_{12}(\mathbf{w})$, we have $\sup_{\mathbf{w}\in\mathcal{H}}|CV_{12}(\mathbf{w})| = o_p(1)$.

(iv) We also decompose $CV_3(\mathbf{w}) = CV_{31}(\mathbf{w}) + CV_{32}(\mathbf{w})$ with

$$CV_{31}(\mathbf{w}) = \frac{1}{n-n_0}\sum_{i=n_0+1}^{n}\left\{\int_0^{\sum_{j=1}^p w_j Q_j^*(\mathbf{X}_i,U_i)-Q_\tau(\mathbf{X}_i,U_i)} [F(z|\mathbf{X}_i,U_i) - F(0|\mathbf{X}_i,U_i)]dz - E_{\mathbf{X}_i,U_i}\left[\int_0^{\sum_{j=1}^p w_j Q_j^*(\mathbf{X}_i,U_i)-Q_\tau(\mathbf{X}_i,U_i)} [F(z|\mathbf{X}_i,U_i) - F(0|\mathbf{X}_i,U_i)]dz\right]\right\},$$

$$CV_{32}(\mathbf{w}) = \frac{1}{n-n_0}\sum_{i=n_0+1}^{n}\left\{\int_{\sum_{j=1}^p w_j Q_j^*(\mathbf{X}_i,U_i)-Q_\tau(\mathbf{X}_i,U_i)}^{\sum_{j=1}^p w_j \widehat{Q}_{j,n_0}(\mathbf{X}_i,U_i)-Q_\tau(\mathbf{X}_i,U_i)} [F(z|\mathbf{X}_i,U_i) - F(0|\mathbf{X}_i,U_i)]dz - E_{\mathbf{X}_i,U_i}\left[\int_{\sum_{j=1}^p w_j Q_j^*(\mathbf{X}_i,U_i)-Q_\tau(\mathbf{X}_i,U_i)}^{\sum_{j=1}^p w_j \widehat{Q}_{j,n_0}(\mathbf{X}_i,U_i)-Q_\tau(\mathbf{X}_i,U_i)} [F(z|\mathbf{X}_i,U_i) - F(0|\mathbf{X}_i,U_i)]dz\right]\right\}.$$

Similar to the proof of $\sup_{\mathbf{w}\in\mathcal{H}}|CV_{11}(\mathbf{w})| = o_p(1)$, we can show that $\sup_{\mathbf{w}\in\mathcal{H}}|CV_{31}(\mathbf{w})| = o_p(1)$, the details are omitted here.

Noting that

$$\begin{aligned}|CV_{32}(\mathbf{w})| \le& \frac{1}{n-n_0}\sum_{i=n_0+1}^{n}\sum_{j=1}^{p} w_j\left|\widehat{Q}_{j,n_0}(\mathbf{X}_i,U_i) - Q_j^*(\mathbf{X}_i,U_i)\right| \\ &+ \frac{1}{n-n_0}\sum_{i=n_0+1}^{n}\sum_{j=1}^{p} w_j E_{\mathbf{X}_i,U_i}\left|\widehat{Q}_{j,n_0}(\mathbf{X}_i,U_i) - Q_j^*(\mathbf{X}_i,U_i)\right| \\ =& CV_{321}(\mathbf{w}) + CV_{322}(\mathbf{w}).\end{aligned}$$

We can prove that $\sup_{\mathbf{w}\in\mathcal{H}}|CV_{321}(\mathbf{w})| = o_p(1)$ as shown in $CV_{22}(\mathbf{w})$. Furthermore, by Cauchy–Schwarz inequality, we have

$$\begin{aligned}
&\sup_{\mathbf{w}\in\mathcal{H}} |CV_{322}(\mathbf{w})| \\
&\le \sup_{\mathbf{w}\in\mathcal{H}} \frac{1}{n-n_0} \sum_{i=n_0+1}^{n} \sum_{j=1}^{p} w_j \Big\{ (\widehat{\boldsymbol{\theta}}_{(-j),n_0} - \boldsymbol{\theta}^*_{(-j)})^\top E[\widetilde{\mathbf{X}}_{i(j)} \widetilde{\mathbf{X}}^\top_{i(j)}] (\widehat{\boldsymbol{\theta}}_{(-j),n_0} - \boldsymbol{\theta}^*_{(-j)}) \Big\}^{1/2} \\
&\le \max_{n_0+1\le i\le n} \max_{1\le j\le p} \lambda^{1/2}_{\max} E[\widetilde{\mathbf{X}}_{i(j)} \widetilde{\mathbf{X}}^\top_{i(j)}] \max_{1\le j\le p} \|\widehat{\boldsymbol{\theta}}_{(-j),n_0} - \boldsymbol{\theta}^*_{(-j)}\|.
\end{aligned}$$

By Assumption (A8) and Theorem 1, we have $\sup_{\mathbf{w}\in\mathcal{H}} |CV_{322}(\mathbf{w})| = o_p(1)$.

(v) To prove (v), we note that

$$\begin{aligned}
\sup_{\mathbf{w}\in\mathcal{H}} |CV_4(\mathbf{w})| \le & \frac{1}{n-n_0} \sum_{i=n_0+1}^{n} E_{\mathbf{X}_{i,U_i}} \left| \sum_{j=1}^{p} w_j \widehat{Q}_{j,n_0}(\mathbf{X}_i, U_i) - \sum_{j=1}^{p} w_j \widehat{Q}_j(\mathbf{X}_i, U_i) \right| \\
\le & \frac{1}{n-n_0} \sum_{i=n_0+1}^{n} \sum_{j=1}^{p} w_j E_{\mathbf{X}_{i,U_i}} \left| \widehat{Q}_{j,n_0}(\mathbf{X}_i, U_i) - \widehat{Q}_j(\mathbf{X}_i, U_i) \right|,
\end{aligned}$$

following the proof of $\sup_{\mathbf{w}\in\mathcal{H}} |CV_{322}(\mathbf{w})| = o_p(1)$, we obtain (v).

(vi) $CV_5 = o_p(1)$ follows from the weak law of large numbers.

Finally, we complete the proof of Theorem 3. □

### A.4. Proof of Theorem 4

According to the proof of Theorem 1, we can further obtain that $\|\widehat{\boldsymbol{\theta}}_{(j)} - \boldsymbol{\theta}^*_{(j)}\| = o_p(1)$ and $\|\widehat{\boldsymbol{\theta}}_{(j),n_0} - \boldsymbol{\theta}^*_{(j)}\| = o_p(1)$ uniformly for all $\tau \in \mathcal{T}$. That is to say, we have $\sup_{\tau\in\mathcal{T}} \|\widehat{\boldsymbol{\theta}}_{(j)} - \boldsymbol{\theta}^*_{(j)}\| = o_p(1)$ and $\sup_{\tau\in\mathcal{T}} \|\widehat{\boldsymbol{\theta}}_{(j),n_0} - \boldsymbol{\theta}^*_{(j)}\| = o_p(1)$.

In the following, we prove that $\widehat{\mathbf{w}}$ is asymptotically optimal uniformly for $\tau \in \mathcal{T}$. The proof is analogous to the proof of Theorem 3, but is more challenge due to the requirement of the asymptotic optimality of $\widehat{\mathbf{w}}$ to hold uniformly in the set of quantile indices. Specifically, we need to prove (i)-(vi) hold but with $\mathbf{w} \in \mathcal{H}$ replaced by $(\mathbf{w}, \tau) \in \mathcal{H} \times \mathcal{T}$ in Theorem 3.

(a) According to the proof of (i) in Theorem 3, it is easy to obtain that

$$\min_{\mathbf{w}\in\mathcal{H}} QPE_{\tau,n}(\mathbf{w}) \ge \min_{\mathbf{w}\in\mathcal{H}} E[\rho_\tau(\varepsilon + r(\mathbf{w}))] - o_p(1) \ge E[\rho_\tau(\varepsilon)] - o_p(1), \text{for all } \tau \in \mathcal{T},$$

which is to say that $\inf_{(\mathbf{w},\tau)\in\mathcal{H}\times\mathcal{T}} QPE_{\tau,n}(\mathbf{w}) \ge E[\rho_\tau(\varepsilon)] - o_p(1)$.

(b) We have $CV_1(\mathbf{w}) = CV_{11}(\mathbf{w}) - CV_{12}(\mathbf{w})$. To prove (b), it suffices to show that $\sup_{(\mathbf{w},\tau)\in\mathcal{H}\times\mathcal{T}} |CV_{11}(\mathbf{w})| = o_p(1)$ and $\sup_{(\mathbf{w},\tau)\in\mathcal{H}\times\mathcal{T}} |CV_{12}(\mathbf{w})| = o_p(1)$. Similar to the proof of (ii) in Theorem 2, to show the uniform convergence, we consider the class of functions $\mathcal{G} = \{g(\varepsilon_i, \mathbf{X}_i, U_i; \mathbf{w}, \tau) : (\mathbf{w}, \tau) \in \mathcal{H} \times \mathcal{T}\}$, where $g(\varepsilon_i, \mathbf{X}_i, U_i; \mathbf{w}, \tau) = \left[ Q_\tau(\mathbf{X}_i, U_i) - \sum_{j=1}^{p} w_j Q^*_j(\mathbf{X}_i, U_i) \right] \psi_\tau(\varepsilon_i)$. On $\mathcal{H} \times \mathcal{T}$, we define

the metric $|\cdot|_1^t$ as $|(\mathbf{w},t)-(\tilde{\mathbf{w}},1-t)|_1^t=\sum_{j=1}^p|w_j-\tilde{w}_j|+1-2t$. Then, the $\epsilon$-covering number $\mathcal{N}(\epsilon,\mathcal{H}\times\mathcal{T},|\cdot|_1^t)=O(1/\epsilon^{p-1})$. Further, the $\epsilon$-bracketing number $\mathcal{N}_{[]}(\epsilon,\mathcal{G},L_1(P))\le C/\epsilon^{p-1}$, and it follows from Glivenko–Cantelli theorem that $\sup_{(\mathbf{w},\tau)\in\mathcal{H}\times\mathcal{T}}|CV_{11}(\mathbf{w})|=o_p(1)$.

We also have

$$\sup_{(\mathbf{w},\tau)\in\mathcal{H}\times\mathcal{T}}|CV_{12}(\mathbf{w})|\le\sum_{j=1}^{p}\max_{n_0+1\le i\le n}\|\widetilde{\mathbf{X}}_{i(j)}\|\sup_{\tau\in\mathcal{T}}\|\widehat{\boldsymbol{\theta}}_{(-j),n_0}-\boldsymbol{\theta}^*_{(-j)}\|=o_p(1).$$

Hence

$$\sup_{(\mathbf{w},\tau)\in\mathcal{H}\times\mathcal{T}}|CV_1(\mathbf{w})|\le\sup_{(\mathbf{w},\tau)\in\mathcal{H}\times\mathcal{T}}|CV_{11}(\mathbf{w})|+\sup_{(\mathbf{w},\tau)\in\mathcal{H}\times\mathcal{T}}|CV_{12}(\mathbf{w})|=o_p(1).$$

Similarly, equations (iii), (iv), (v) and (vi) follow from the corresponding proof in Theorem 3 as well as the fact $\sup_{\tau\in\mathcal{T}}\|\widehat{\boldsymbol{\theta}}_{(j)}-\boldsymbol{\theta}^*_{(j)}\|=o_p(1)$ and $\sup_{\tau\in\mathcal{T}}\|\widehat{\boldsymbol{\theta}}_{(j),n_0}-\boldsymbol{\theta}^*_{(j)}\|=o_p(1)$. Therefore, we complete the proof of Theorem 4. □

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s10463-022-00857-z.

**Acknowledgements** Lin's work was supported by the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (No. 19XNB014).

## References

Angrist, J., Chernozhukov, V., Fernández-Val, I. (2006). Quantile regression under misspecification, with an application to the U.S. wage structure. *Econometrica, 74*, 539–563.

Cai, Z., Xiao, Z. (2012). Semiparametric quantile regression estimation in dynamic models with partially varying coefficients. *Journal of Econometrics, 167*, 413–425.

Cai, Z., Xu, X. (2008). Nonparametric quantile estimations for dynamic smooth coefficient models. *Journal of the American Statistical Association, 103*, 1595–1608.

Cai, Z., Chen, L., Fang, Y. (2018). A semiparametric quantile panel data model with an application to estimating the growth effect of FDI. *Journal of Econometrics, 206*, 531–553.

Chai, H., Shi, X., Zhang, Q., Zhao, Q., Huang, Y., Ma, S. (2017). Analysis of cancer gene expression data with an assisted robust marker identification approach. *Genetic Epidemiology, 41*, 779–789.

Fitzenberger, B., Koenker, R., Machado, J. (Eds.). (2002). *Economic application of quantile regression*. Heidelberg, Germany: Physica Verlag.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica, 75*, 1175–1189.

Hjort, N. L., Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association, 98*, 879–899.

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

Knight, K. (1998). Limiting distributions for $L_1$ regression estimators under general conditions. *Annals of Statistics, 26*, 755–770.

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

Kuester, K., Mittnik, S., Paolella, M. (2006). Value-at-risk prediction: A comparison of alternative strategies. *Journal of Financial Econometrics, 4*, 53–89.

Li, D., Linton, O., Lu, Z. (2015). A flexible semiparametric forecasting model for time series. *Journal of Econometrics, 187*, 345–357.
Li, G., Li, Y., Tsai, C. L. (2015). Quantile correlations and quantile autoregressive modeling. *Journal of the American Statistical Association, 110*, 246–261.
Li, J., Xia, X., Wong, W. K., Nott, D. (2018). Varying-coefficient semiparametric model averaging prediction. *Biometrics, 74*, 1417–1426.
Li, X., Ma, X., Zhang, J. (2018). Conditional quantile correlation screening procedure for ultrahigh-dimensional varying coefficient models. *Journal of Statistical Planning and Inference, 197*, 62–92.
Li, Y., Graubard, B. I., Korn, E. L. (2010). Application of nonparametric quantile regression to body mass index percentile curves from survey data. *Statistics in Medicine, 29*, 558–572.
Lian, H. (2015). Quantile regression for dynamic partially linear varying coefficient time series models. *Journal of Multivariate Analysis, 141*, 49–66.
Lin, H., Fei, Z., Li, Y. (2016). A semiparametrically efficient estimator of the time-varying effects for survival data with time-dependent treatment. *Scandinavian Journal of Statistics, 43*, 649–663.
Liu, J., Huang, J., Zhang, Y., Lan, Q., Rothman, N., Zheng, T., Ma, S. (2013). Identification of gene-environment interactions in cancer studies using penalization. *Genomics, 102*, 189–194.
Lu, X., Su, L. (2015). Jackknife model averaging for quantile regressions. *Journal of Econometrics, 188*, 40–58.
Ma, S., Yang, L., Romero, R., Cui, Y. (2011). Varying coefficient model for gene-environment interaction: A non-linear look. *Bioinformatics, 27*, 2119–2126.
Mack, Y., Silverman, B. (1982). Weak and strong uniform consistency of kernel regression estimates. *Probability Theory Related Fields, 61*, 405–415.
Nan, Y., Yang, Y. (2014). Variable selection diagnostics measures for high-dimensional regression. *Journal of Computational and Graphical Statistics, 23*, 636–656.
Shan, K., Yang, Y. (2009). Combining Regression Quantile Estimators. *Statistica Sinica, 19*, 1171–1191.
Sharafeldin, N., Slattery, M. L., Liu, Q., Franco-Villalobos, C., Caan, B. J., Potter, J. D., Yasui, Y. (2015). A candidate-pathway approach to identify gene-environment interactions: Analyses of colon cancer risk and survival. *Journal of the National Cancer Institute, 107*(9), djv160.
Shen, Y., Liang, H. (2017). Quantile regression for partially linear varying-coefficient model with censoring indicators missing at random. *Computational Statistics and Data Analysis, 117*, 1–18.
Silverman, B. W. (1986). Density Estimation for Statistics and Data Analysis. London: Chapman and Hall.
Stock, J., Watson, M. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting, 23*, 405–430.
Van der Vaart, A., Wellner, J. A. (1996). *Weak convergence and empirical Processes*: *with applications to statistics*. New York: Springer.
Wang, M., Zhang, X., Wan, A. T. K., you, K., Zou, G. (2021). Combination forecasts of output growth in a seven-country data set. *Biometrics, 2021*, 1–12.
Wheelock, D. C., Wilson, P. W. (2008). Non-parametric, unconditional quantile estimation for efficiency analysis with an application to Federal Reserve check processing operations. *Journal of Econometrics, 145*, 209–225.
Winnepenninckx, V., Lazar, V., Michiels, S., Dessen, P., Stas, M., Alonso, S. R., Avril, M., Romero, P. L., Robert, T., Balacescu, O., Eggermont, A. M., Lenoir, G., Sarasin, A., Tursz, T., Oord, J. J., Spatz, A. (2006). Gene expression profiling of primary cutaneous melanoma and clinical outcome. *Journal of the National Cancer Institute*, 98, 472–482.
Wu, M., Huang, J., Ma, S. (2017). Identifying gene-gene interactions using penalized tensor regression. *Statistics in Medicine, 37*, 598–610.
Xu, Y., Wu, M., Ma, S., Ahmed, S. (2018). Robust gene environment interaction analysis using penalized trimmed regression. *Journal of Statistical Computation and Simulation, 88*, 3502–3528.
Yang, Y. (2001). Adaptive Regression by Mixing. *Journal of the American Statistical Association, 96*, 574–588.
Yang, Y. (2007). Prediction/estimation with simple linear models: Is it really that simple? *Econometric Theory, 23*, 1–36.
Ye, C., Yang, Y., Yang, Y. (2018). Sparsity oriented importance learning for high-dimensional linear regression. *Journal of the American Statistical Association, 113*, 1797–1812.
Zhan, Z., Yang, Y. (2022). *Profile electoral college cross-validation Information Sciences, 586*, 24–40.

Zhu, R., Wan, A. T. K., Zhang, X., Zhou, G. (2019). A Mallows-type model averaging estimator for the varying coefficient partially linear model. *Journal of the American Statistical Association, 114*, 882–892.