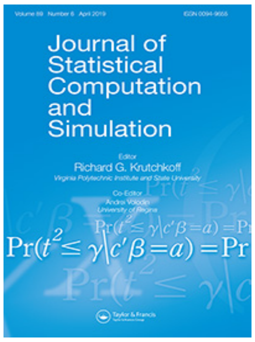


Volume 89 Number 6 April 2019 ISSN 0094-9655

Journal of Statistical Computation and Simulation

Editor
Richard G. Krutchkoff
Virginia Polytechnic Institute and State University

Co-Editor
Andrei Volodin
University of Regina

$\Pr(t^2 \leq \gamma | c'\beta = a) = \Pr$

Taylor & Francis

Taylor & Francis
Taylor & Francis Group

Journal of Statistical Computation and Simulation

ISSN: 0094-9655 (Print) 1563-5163 (Online) Journal homepage: https://www.tandfonline.com/loi/gscs20

Wavelet-based LASSO in functional linear quantile regression

Yafei Wang, Linglong Kong, Bei Jiang, Xingcai Zhou, Shimei Yu, Li Zhang & Giseon Heo

To cite this article: Yafei Wang, Linglong Kong, Bei Jiang, Xingcai Zhou, Shimei Yu, Li Zhang & Giseon Heo (2019) Wavelet-based LASSO in functional linear quantile regression, Journal of Statistical Computation and Simulation, 89:6, 1111-1130, DOI: 10.1080/00949655.2019.1583228

To link to this article: https://doi.org/10.1080/00949655.2019.1583228

Published online: 01 Mar 2019.

Submit your article to this journal

Article views: 36

View Crossmark data

JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION
2019, VOL. 89, NO. 6, 1111–1130
https://doi.org/10.1080/00949655.2019.1583228

Wavelet-based LASSO in functional linear quantile regression

Yafei Wang[a,b], Linglong Kong[b], Bei Jiang[b], Xingcai Zhou[c], Shimei Yu[b], Li Zhang[b] and Giseon Heo[b]

[a]College of Applied Sciences, Beijing University of Technology, Beijing, People's Republic of China; [b]Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Canada; [c]Institute of Statistics and Data Science, Nanjing Audit University, Nanjing, People's Republic of China

ABSTRACT
In this paper, we develop an efficient wavelet-based regularized linear quantile regression framework for coefficient estimations, where the responses are scalars and the predictors include both scalars and function. The framework consists of two important parts: wavelet transformation and regularized linear quantile regression. Wavelet transform can be used to approximate functional data through representing it by finite wavelet coefficients and effectively capturing its local features. Quantile regression is robust for response outliers and heavy-tailed errors. In addition, comparing with other methods it provides a more complete picture of how responses change conditional on covariates. Meanwhile, regularization can remove small wavelet coefficients to achieve sparsity and efficiency. A novel algorithm, Alternating Direction Method of Multipliers (ADMM) is derived to solve the optimization problems. We conduct numerical studies to investigate the finite sample performance of our method and applied it on real data from ADHD studies.

ARTICLE HISTORY
Received 5 June 2018
Accepted 12 February 2019

KEYWORDS
Quantile regression; wavelets; LASSO; functional data analysis; ADMM; ADHD

1. Introduction

With advances in technology, it is increasingly common to encounter data that are functional or curves in nature, for example neuroimaging data [1,2]. A common feature of many imaging techniques is that massive functional data are observed/calculated at the same design points, such as time for functional images (e.g. positron emission tomography (PET)) and arclength for structure imaging (e.g. diffusion tensor imaging (DTI)). In recent literature, extensive research has been focusing on the functional linear model where a scalar response is regressed on a functional predictor [3,4]. It has become a powerful statistical tool for functional data analysis. In our paper, we will consider the functional linear regression, where the responses such as the neurological or clinical outcomes (e.g. attention deficit hyperactivity disorder (ADHD) index) are modelled by a set of scalar covariates and functional covariates of interest (e.g. hemodynamic response functions (HRF)).

Denote $Y = (y_1,\ y_2,\ \ldots,\ y_n)^{\mathrm{T}}$ as a scalar vector response, $U = (\boldsymbol{u}_1,\ \boldsymbol{u}_2,\ \ldots\ \boldsymbol{u}_n)^{\mathrm{T}}$ as a $n \times s$ scalar matrix of predictors with i-th row $\boldsymbol{u}_i^{\mathrm{T}} = (u_{i1},\ u_{i2}\ \cdots, u_{is})$, $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_s)^{\mathrm{T}}$ as the corresponding scalar coefficient vector, and $Z(t) = (z_1(t), z_2(t),$

CONTACT Linglong Kong ✉ lkong@ualberta.ca

$\dots, z_n(t))^{\mathrm{T}}$ as the functional predictors, and $\eta(t)$ as the corresponding functional coefficient. In addition, α is the intercept and ε_i's are independent identically distributed (i.i.d.) random errors. Therefore, a functional linear regression model can be written as follows:

$$y_i = \alpha + \boldsymbol{u}_i{}^{\mathrm{T}}\boldsymbol{\delta} + \int z_i(t)\eta(t)\,\mathrm{d}t + \varepsilon_i. \tag{1}$$

In linear regression, the conditional mean of responses can be obtained by ordinary least squares (OLS) estimates, and it is optimal under the assumption that error follows Gaussian distribution. However, it is common to observe non-Gaussian distributed errors, including heavy-tailed ones, where OLS does not perform well [5]. Quantile regression has emerged as an important statistical methodology and has been used widely in various disciplines, such as biology, medicine, finance and economics. It is well known that the distribution of data is typically skewed or data contain some outliers, the median regression, a special case of quantile regression, provides more robust estimators than the mean regression. A comprehensive survey of the theory of quantile regression and its applications can be found in [6]. An alternative to model (1) is the functional linear quantile regression where the conditional quantiles of the functional responses are modelled by a set of scalar covariates and a functional covariate. We consider the quantile regression model as

$$Q_\tau(y_i|\boldsymbol{u}_i, z_i(t)) = \alpha_\tau + \boldsymbol{u}_i{}^{\mathrm{T}}\boldsymbol{\delta} + \int z_i(t)\eta(t)\,\mathrm{d}t, \tag{2}$$

for given $\tau \in (0,1)$, where $Q_\tau(y_i|\boldsymbol{u}_i, z_i(t))$ is the τ-th conditional quantile of y_i given covariates $\boldsymbol{u}_i$ and $z_i(t)$.

Many procedures have been proposed to approximate the functional coefficient $\eta(t)$, for example, functional principal component analysis (fPCA) based approaches [7–9], B-spline with penalties [10,11], methods combining fPCA and penalization [12], partial least squares (PLS) [4,13], and others. Among them, PLS bases can capture the information of both the response and covariates, while fPCA bases only use the information of covariates. As for smooth splines, the information of either response or explanatory variables is not considered. In order to provide a good approximation of the functional coefficients, a large number of bases should be chosen. However, this may cause overfitting [4] and calculation is always limited. Therefore, less finite bases are desired to capture the local information.

In this article, we prefer to use wavelets for efficiently approximating functions with a relatively small number of nonzero wavelet coefficients [5]. Because wavelets are used to transform the signals from time domain to frequency domain, where wavelet coefficients are independent so that we can apply regularization to encourage sparse representation. Its other advantage is the ability to capture the local information, including feature changes in space or time [14]. As well, it is computationally efficient [3]. For a large variety of functions, the wavelet decomposition allows good representation of the function by using only a relatively small number of wavelet coefficients. A comprehensive survey of wavelet applications in statistics can be found in [15,16]. We apply the Least Absolute Shrinkage and Selection Operator (LASSO) [17] to encourage sparse representation. Through discrete wavelet transform, functional variables can be represented by a set of wavelet coefficients.

Therefore, a problem of functional linear quantile regression can be converted into a problem of variable selection of the wavelet bases. The LASSO method is applied in the wavelet domain to select wavelet coefficients.

Our methodology consists of two important parts: wavelet transformation and regularized linear quantile regression. Wavelet transform can be used to approximate functional data through representing it by finite wavelet coefficients and effectively capturing its local features. Quantile regression is robust for response outliers and heavy-tailed errors. In addition, comparing with other methods it provides a more complete picture of how responses change conditional on covariates. Simultaneously, regularization can remove small wavelet coefficients to achieve sparsity and efficiency. Under mild conditions, the estimated functional coefficients converge to the true ones and the predicted response also convergences.

The rest of the article is organized as follows. In Section 2, we introduce the wavelet-based LASSO in functional linear quantile regression. In Section 3, we show the convergence of the estimated functional coefficients and predicted responses. We derive the ADMM algorithm to solve the optimization problem in Section 4. Section 5 provides finite sample performance through simulation studies and analyses a real data set from an ADHD study. Section 6 concludes the article with discusses and future research directions.

2. Wavelet-based LASSO in functional linear quantile regression

2.1. Functional linear regression with wavelet basis

Wavelets are basis functions that can be used to efficiently approximate functions with wavelet coefficients [5]. Given scaling function ϕ, and wavelet function ψ, we can construct a basis with dilation parameter j, and position parameter d as follows:

$$\phi_{jd}(t) = 2^{j/2}\phi(2^j t - d); \quad \psi_{jd}(t) = 2^{j/2}\psi(2^j t - d),$$

where the value of d varies according to different decomposition level in the functions. The power of $\phi(t)$ is more compact at low frequencies while the power of $\psi(t)$ concentrates at relatively high frequencies. Therefore, $\phi(t)$ is used to approximate the global properties and $\psi(t)$ is used to model the detailed local features. Given a specific decomposition level j_o, the orthonormal wavelet basis set is defined as $\{\phi_{j_0 d} : d = 0, \dots, 2^{j_0} - 1\} \cup \{\psi_{jd} : j \geq j_0, d = 0, \dots, 2^j - 1\}$.

We represent both the functional predictor $z(t)$ and the coefficient function $\eta(t)$ in model (2) in terms of wavelet bases. In practice, the functional predictors $z_i(t)$'s are sampled at equally spaced discrete m points. Consequently, we can calculate maximum of $J = \log_2^m - 1$ levels of decomposition via discrete wavelet transformation. Thus, given j_0, the functional predictors $z_i(t)$'s can be approximated by

$$z_i(t) = \sum_{d=0}^{2^{j_0}-1} x'_{i\phi}[j_0, d]\phi_{j_0 d}(t) + \sum_{j=j_0}^{J} \sum_{d=0}^{2^j-1} x_{i\psi}[j, d]\psi_{jd}(t),$$

where the wavelet coefficients are defined by

$$x'_{i\phi}[j_0, d] = \int z_i(t)\phi_{j_0 d}(t)\,\mathrm{d}t, \quad x_{i\psi}[j, d] = \int z_i(t)\psi_{jd}(t)\,\mathrm{d}t, \tag{3}$$

where $x'_{i\phi}[j_0, d]$ are called approximation coefficients and $x_{i\psi}[j, d]$ are detailed coefficients. Similarly, we can decompose the coefficient function $\eta(t)$ by discrete wavelet transformation using the same bases

$$\eta_i(t) = \sum_{d=0}^{2^{j_0}-1} \beta'_\phi[j_0, d]\phi_{j_0 d}(t) + \sum_{j=j_0}^{J} \sum_{d=0}^{2^j-1} \beta_\psi[j, d]\psi_{jd}(t),$$

where the wavelet coefficients are

$$\beta'_\phi[j_0, d] = \int \eta(t)\phi_{j_0 d}(t)\,\mathrm{d}t, \quad \beta_\psi[j, d] = \int \eta(t)\psi_{jd}(t)\,\mathrm{d}t. \tag{4}$$

At coarser level (small m), the coefficients β detect global features of the data while at finer levels (large m) they capture local features.

Due to the orthonormality of the wavelet bases, the quantile regression model (2) can be rewritten as

$$Q_\tau(y_i|\boldsymbol{u}_i, x_i) = \alpha_\tau + \boldsymbol{u}_i{}^{\mathrm{T}}\boldsymbol{\delta} + \boldsymbol{x}_i{}^{\mathrm{T}}\boldsymbol{\beta}, \tag{5}$$

where $\boldsymbol{x}_i^{\mathrm{T}} = (x_{i1}, x_{i2}, \ldots, x_{im})$ is a vector of predictor variables that are the derived wavelet coefficients, $i = 1, \ldots, n$, in (3), and $\boldsymbol{\beta}$ is a $m \times 1$ vector of coefficients in (4). The responses y_i, intercept α, the scalar predictor $\boldsymbol{u}_i$ and scalar coefficient $\boldsymbol{\delta}$ are defined the same as in (2).

2.2. Wavelet-based LASSO in functional linear quantile regression

Both functional predictor $z(t)$ and their corresponding functional coefficient $\eta(t)$ in quantile regression (2) are transformed to linear combinations of wavelet bases. The wavelet coefficients $\{x_{ik}\}_{k=1}^{m}$ become the predictors in transformed space and a few of m predictor variables are likely useful in predicting the response variable y. It is interesting to note that the functional regression problem may be viewed as a variable selection problem. The functional coefficient η is estimated by inverse discrete wavelet transform of a few important wavelet coefficients with the same wavelet bases and the same decomposition level. For more discussion, see [3,18,19].

Infinite dimensional nature of $z(t)$ and $\eta(t)$ in quantile regression model (2) has been taken care of via wavelet coefficients resulting finite and low dimensional representation (5), that is, m-dimension. We now aim to estimate the coefficients in (5) and at the same time select predictor variables that are effective in predicting response variable. LASSO method [17] is useful for this purpose because it imposes L_1 penalty on the coefficient vector $\boldsymbol{\beta}$ to encourage sparsity.

The coefficient vector $\boldsymbol{\delta}$ corresponding to scalar predictors is not affected by LASSO penalty. Putting all these together, the parameters $\alpha_\tau, \boldsymbol{\delta}$, and $\boldsymbol{\beta}$, can be estimated by minimizing the quantile loss function with shrinkage constraint. That is,

$$(\hat{\alpha}_\tau, \hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\beta}}) = \operatorname{argmin} \sum_{i=1}^{n} \rho_\tau(y_i - \alpha_\tau - \boldsymbol{u}_i{}^{\mathrm{T}}\boldsymbol{\delta} - \boldsymbol{x}_i{}^{\mathrm{T}}\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1, \tag{6}$$

where ρ_τ is the loss function in quantile regression, λ is a tuning parameter, $\|\cdot\|_1$ denotes L_1-norm, and $\lambda\|\boldsymbol{\beta}\|_1$ is the penalty function for regression shrinkage and variable selection.

There are other choices of penalty functions such as adaptive LASSO [20], SCAD [21], and MCP [22].

Instead of considering fixed quantile level, to improve efficiency we can model multiple quantile levels simultaneously, say, through composite quantile regression [23]. Under the condition that the effects of covariates are piecewise constant or continuous across different quantile levels, the composite quantile estimate is more efficient than the one from a single level; see [4,24–26]. In our setting we have

$$(\hat{\boldsymbol{\alpha}}_\tau,\ \hat{\boldsymbol{\delta}},\ \hat{\boldsymbol{\beta}}) = \operatorname{argmin} \sum_{i=1}^{n} \sum_{k=1}^{K} \rho_{\tau_k}(y_i - \alpha_{\tau_k} - \boldsymbol{u}_i{}^{\mathrm{T}}\boldsymbol{\delta} - \boldsymbol{x}_i{}^{\mathrm{T}}\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1, \tag{7}$$

where $\boldsymbol{\alpha}_\tau = (\alpha_{\tau_1} \ldots, \alpha_{\tau_K})^{\mathrm{T}}$ is a vector of intercepts. Typically, we can choose $K = 9$ and use equally spaced quantiles [23,24]. Note that quantile estimate (6) at a single level is just a special case of composite quantile estimate (7) with $K = 1$. In the following, we will focus on the composite quantile regression case of (7).

2.3. Tuning parameters selection

There are two tuning parameters, the penalty parameter λ and the decomposition level parameter J. The former controls the fitness of the model. When $\lambda \to 0$, the LASSO method becomes an ordinary quantile regression. When $\lambda \to \infty$, the LASSO penalty would set all the coefficients to zero. The later controls the optimal wavelet transformation. When J is small, the wavelet basis functions can only provide coarse approximation of functions and therefore lose detailed local features. When J is large, the approximation may pick up some redundancy noises while increases computational burdens. Therefore, it is important to choose appropriate unknown tuning parameters.

In general, three approaches can be used in tuning parameters selection: Akaike information criterion (AIC), Bayesian information criterion (BIC), and cross validation (CV) [3,17,20,21]. In this article, cross validation method is adopted. In particular, to speed up the computation which taking advantage of the cross-validation method, we prefer to use κ-fold cross validation. Commonly, $\kappa = 5$ or $\kappa = 10$ [17]. In this article, five-fold cross validation is used for tuning parameter selection. In the real data example, after tuning parameters selection, 10-fold cross validation is used to estimate the prediction errors.

3. Theoretical properties

In this section, we investigate the asymptotical behaviour of the wavelet-based LASSO estimators when both $n \to \infty$ and $m \to \infty$, meaning that the sample size n increases and the curves $z_i(t)$'s are also becoming more densely observed, respectively. Let m be the number of discrete points at which the functional predictors $z_i(t)$ are observed with the sample n. In order to derive the convergence rate of $\hat{\eta}(t)$ to $\eta(t)$ we need the following assumptions. Note that these assumptions are not necessary the weakest ones.

A1. The errors $\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed with distribution function F, its density function $f(\cdot)$ is bounded away from zero and infinity, and it has a continuous and uniformly bounded derivative at their τ-th quantiles.

A2. There exists constant M such that $\|z_i(t)\|_2 < M$ for all i.

A3. There are two constants c_1 and c_2 such that $(1/n)\sum_{i=1}^{n} C_i C_i^{\mathrm{T}}$ satisfies the eigenvalue condition

$$0 < c_1 < \lambda_{\min}\left\{\frac{1}{n}\sum_{i=1}^{n} C_i {C_i}^{\mathrm{T}}\right\} \leq \lambda_{\max}\left\{\frac{1}{n}\sum_{i=1}^{n} C_i {C_i}^{\mathrm{T}}\right\} < c_2 < \infty$$

where $C_i = [\boldsymbol{u}_i, \boldsymbol{x}_i]^{\mathrm{T}}$.

A4. $\eta(t)$ is a q times differentiable function in the Sobolev sense and the wavelet basis has p vanishing moments, where $p > q$.

A5. $\lambda = O(n^{-1/2})$, and $n = O_p(m^{4q})$

A6. $m/n \to 0$

The following theorem gives the convergence rate of the estimated functional coefficient $\hat{\eta}(t)$, which depends on both sample size n and the number of discrete points m.

Theorem 3.1: *Let $\hat{\eta}(t)$ be the estimator resulting from* (7). *If the assumptions A1–A6 hold, then*

$$\|\hat{\eta}(t) - \eta(t)\|_2^2 = O_p\left(\frac{m}{n}\right) + o_p\left(\frac{1}{m^{2q}}\right),$$

where the L_2 norm $\|\cdot\|_2$ is in function integration sense.

A detailed proof of this Theorem 3.1 is provided in the Appendix. The approximation error rate of $\hat{\eta}(t)$ towards $\eta(t)$ are controlled by two terms. The first term is of the same order of m/n which is a typical result of estimating, while the second term is of the lower order of $1/m^{2q}$ which is mainly due to approximation by wavelets. In particular, the approximation error rate is dominated by the second term if m^{2q+1} is of the lower order of n. Otherwise, it is dominated by the first term. Under further condition, we can have the following theorem for the prediction error bound:

Theorem 3.2: *Suppose $z(t)$ are square integrable functions on* $[0, 1]$ *and* $F^{-1}(\tau) = 0$. *If the assumptions A1–A6 hold, then*

$$|\hat{y} - y|^2 = O_p\left(\frac{m}{n}\right) + o_p\left(\frac{1}{m^{2q}}\right),$$

where $\hat{y} = \hat{\alpha}_\tau + \boldsymbol{u}_i^{\mathrm{T}}\hat{\boldsymbol{\delta}} + \int_0^1 z(t)\hat{\eta}(t)\,\mathrm{d}t$.

The proof follows that from Theorem 3.1 and the Cauchy–Schwarz inequality, the details of which are provided in the Appendix. Similarly as in Theorem 3.1, the prediction error rate depends on the same two terms from estimating and approximation by wavelets respectively, while the estimation errors caused by $\hat{\alpha}_\tau$ and $\hat{\boldsymbol{\delta}}$ are absorbed by the first term.

4. ADMM algorithm

The objective function in (6) is a sum of quantile loss function and penalty function. Both loss and penalty functions are convex, therefore, our minimization is a convex optimization problem. Convex optimization problems can be solved by some general techniques like interior point method [27] and the simplex method [6]. However, for large scale data, both methods usually lead to intense computation. We adopt a novel algorithm, alternating direction method of multipliers (ADMM) [28]. ADMM is a powerful algorithm for convex optimization problems especially that can be decomposed into sub-convex problems [29]. Although the ADMM originated in 1950s, it was mainly developed in 1970s [28,30]. Its convergence has been studied in many situations; see [29,31] for some examples. It has been popularized in recent years because of its efficiency on large scale problems and ability of solving multiple non-smooth terms in the objective function [29,31]. In this section, we reformulate our optimization problems of (6) and derive their ADMM algorithms based on the observation that they can be split into two sub-convex optimization problems and one of the sub-problems has a non-smooth function.

To apply the ADMM algorithm, we rewrite the objective function (6) in matrix from

$$(\hat{\boldsymbol{\alpha}_\tau}, \hat{\boldsymbol{\delta}}, \ \hat{\boldsymbol{\beta}},) = \text{argmin}\ \rho_\tau(\boldsymbol{y} - \boldsymbol{\alpha}_\tau - U\boldsymbol{\delta} - X\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1, \tag{8}$$

where $\rho_\tau(\boldsymbol{y}) = \sum_{i=1}^n \rho_\tau(y_i), \boldsymbol{y} = (y_1, \ldots, y_n)^{\mathrm{T}}, \boldsymbol{\alpha}_\tau = (\alpha_\tau, \ \ldots, \ \alpha_\tau)^{\mathrm{T}}$ is an n dimensional vector, and $\|\cdot\|_1$ stands for L_1 norm. For simplicity, we define $X^* = (\mathbf{1}_n, U, X)$, $\boldsymbol{\beta}^* = (\boldsymbol{\alpha}_\tau, \ \boldsymbol{\delta}^{\mathrm{T}}, \ \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$. The objective function in model (8) becomes

$$\begin{aligned} &\text{argmin}\ \rho_\tau(\boldsymbol{\beta}^*) + \lambda\|\boldsymbol{\beta}\|_1 \\ &\text{subject to } -\boldsymbol{r} - X^*\boldsymbol{\beta}^* = -\boldsymbol{y}, \end{aligned}$$

where $\rho_\tau(\boldsymbol{r})$ and $\lambda\|\boldsymbol{\beta}\|_1$ are two convex functions. The l-th iteration of ADMM is

$$\begin{aligned} \boldsymbol{r}^{l+1} &= \text{argmin}\ \rho_\tau(\boldsymbol{r}) + \frac{\varrho}{2}\|\boldsymbol{y} - \boldsymbol{r} - X\boldsymbol{\beta}^l + \boldsymbol{u}^l\|_2^2, \\ \boldsymbol{\beta}^{l+1} &= \text{argmin}\ \lambda\|\boldsymbol{\beta}\|_1 + \frac{\varrho}{2}\|\boldsymbol{y} - \boldsymbol{r}^l - X\boldsymbol{\beta}^l + 1 + \boldsymbol{u}^l\|_2^2, \\ \boldsymbol{u}^{l+1} &= \boldsymbol{u}^l + \varrho(\boldsymbol{y} - \boldsymbol{r}^{l+1} - X^*\boldsymbol{\beta}^{*\,l+1}), \end{aligned}$$

where ϱ is a tuning parameter and was chosen to be 1.2 in this article [29]. The first step can be simplified by the soft thresholding operator. That is,

$$\boldsymbol{r}^{l+1} = \boldsymbol{S}_{\frac{1}{\varrho}}\left(\boldsymbol{y} - X\boldsymbol{\beta}^l + \boldsymbol{u}^l - \frac{2\tau - 1}{\varrho}\right), \tag{9}$$

where $\boldsymbol{S}_\lambda(x) = (x - \lambda)_+ - (-x - \lambda)_+$ is a thresholding function. The second step is a standard L_1 penalized least square problem, which can be easily solved following LASSO algorithm or approximately solved by linearization at $\boldsymbol{\beta} = \boldsymbol{\beta}^l$ ending up with a closed form solution by soft thresholding [32]. The primal and dual residuals are

$$\boldsymbol{s}^{l+1} = \varrho X(\boldsymbol{\beta}^{l+1} - \boldsymbol{\beta}^{l}), \boldsymbol{t}^{l+1} = \boldsymbol{r}^{l+1} + X\boldsymbol{\beta}^{l+1} - y,$$

respectively. The termination criterion can be set as

$$\|\boldsymbol{s}^{l}\|_2 \leq \epsilon^{pri} \quad \text{and} \quad \|\boldsymbol{t}^{l}\|_2 \leq \epsilon^{dual}$$

where

$$\epsilon^{pri} = \sqrt{n}\epsilon^{abs} + \epsilon^{rel}\max\{\|\boldsymbol{r}^{l}\|_2, \|X^{*}\boldsymbol{\beta}^{*l}\|_2, \|\boldsymbol{y}\|_2\},$$
$$\epsilon^{dual} = \sqrt{n}\epsilon^{abs} + \epsilon^{rel}\|\boldsymbol{u}^{l}\|_2.$$

In our work, we choose ϵ^{abs} and ϵ^{rel} as 10^{-4} and 10^{-2} respectively.

5. Numerical studies

In this section, we conduct simulations to investigate the finite sample performance of the wavelet-based LASSO in functional linear quantile regression and composite quantile regression. For brevity in figures and tables, we use capital abbreviation QR standing for quantile regression, CQR standing for composite quantile regression, QR LASSO standing for quantile regression with a LASSO penalty, and CQR LASSO standing for composite quantile regression with a LASSO penalty.

5.1. Model set up

Considering the linear functional model (1),

$$y_i = \alpha + \boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{\delta} + \int z_i(t)\eta(t)\,\mathrm{d}t + \varepsilon_i.$$

We employ similar settings as those in [3]. In particular, $\boldsymbol{u}_i$ of dimension 2. That is, $\boldsymbol{u} = (u_1, u_2)^{\mathrm{T}}$, where $u_i \sim \mathrm{Uniform}(0, 1)$, and $u_2 \sim \mathrm{Bernolli}(0.5)$. Each functional predictor $z(t)$ comes from a stochastic Gaussian process with a mean of zero and the covariance function is $cov(z(t_1), z(t_2)) = t_1(1 - t_2)$, where $t_1 < t_2$. Two types of the coefficient functions are investigated, smooth function and bumpy function. The smooth function is $\omega(t) = 0.75\,\Phi\,(t, 20, 60) - 0.05\,\Phi\,(t, 50, 20)$, where $\Phi\,(t, \theta, \vartheta) = (\Gamma(\theta + \vartheta)(t^{\theta-1})((1 - t)^{\vartheta-1}))/(\Gamma(\theta)\Gamma(\vartheta))$. Bumps, one of Donoho and Johnstone test functions, is used as bumpy function.

Signal-to-noise ratio (SNR) is an important criterion to measure, it compares the level of a desired signal to the level of background noise. In this article, we choose, $SNR = \mu/\sigma$, where μ is the mean of the signal, and σ is the standard deviation of the noise. This definition is used because μ is known and σ can be controlled. When SNR is signal, the noise level is small and the signal on the signal is less than influenced the signal can be easily detected. SNR is big, on the other hand, when the signal is difficult to be discriminated from the noise. In our research, we control SNR to be with in $(1, 5)$.

The error term is decided in four types of distributions for each setting:

I. Standard normal distribution: $N(0,\ 1)$;
II. Mixed-mean normal distribution: $0.8N(0,\ 1.2) + 0.2N(10,\ 1.2)$;
III. Mixed-variance normal distribution: $0.8N(0,\ 0.5) + 0.2N(0,\ 5)$;
IV. Standard Cauchy distribution: $C(0, 1)$.

In the first three normal distribution settings, $SNR = 2$. We further investigate the performance of the proposed method with other SNR values. We adjust the coefficients through multiplying them by 0.5 and 1.5 to make $SNR = 1$ and $SNR = 3$ respectively. In the last setting, the variance is infinite and SNR is technically zero. we set its scale parameter to be 1 and the coefficients the same as other errors to represent the different levels of SNR. To simplify notations, we use SNR situation 1, SNR situation 2 and SNR situation 3 standing for three levels of SNR. Overall, there are 24 settings considering all the factors, and we repeat each setting 100 times. Furthermore, we set the sample size as 200 and the samples are captured at 128 ($s = 128$) equally spaced time points in the range of $(0,\ 1)$. In composite quantile regression, $K = 9$, i.e. $\alpha_\tau = (0.1,\ 0.2,\ \ldots,\ 0.9)$. Tuning parameter λ is selected from 20 grid points in an arithmetic sequence from $e^{-2.5}$ to $e^{2.5}$ by five-fold cross validation.

Various package in the software R was used to facilitate the computation. Wavelet transform of the functional data and the inverse procedure can be achieved through the package ‘rwt’ [33] and wavelet basis from Daubechies’ family is chosen with the filter number 4. The bumpy function can be performed through the ‘wavethresh’ package [34]. Quantile regression and composite quantile regression without a LASSO penalty can be analysed by the existing functions in the ‘quantreg’ package [35].

5.2. Simulation study results

We compare our methods with QR and CQR without any penalty. We use the mean square error (MSE) and the mean integrated squared error (ISE) from both prediction and estimation. respectively. In this article, we focus on functional coefficients only. We define MSE

Table 1. MSE with SNR situation 2.

		Bumpy function				Smooth function			
Dist	Method	Mean	Var	Med	MAD	Mean	Var	Med	MAD
I	QR	1.07	0.01	1.06	0.09	1.01	0.01	1.01	0.09
	QR LASSO	0.84	0.02	0.83	0.13	0.28	0.01	0.26	0.06
	CQR	0.88	0.01	0.87	0.08	0.84	0.01	0.84	0.07
	CQR LASSO	0.65	0.01	0.64	0.09	0.22	0.01	0.22	0.06
II	QR	0.87	0.01	0.87	0.08	0.94	0.01	0.93	0.07
	QR LASSO	0.69	0.01	0.67	0.11	0.24	0.01	0.23	0.05
	CQR	0.72	0.01	0.72	0.071	0.77	0.01	0.76	0.07
	CQR LASSO	0.52	0.01	0.52	0.07	0.21	0.01	0.19	0.05
III	QR	1.04	0.01	1.03	0.09	1.21	0.02	1.20	0.12
	QR LASSO	0.79	0.01	0.79	0.12	0.32	0.01	0.31	0.07
	CQR	0.87	0.01	0.86	0.08	1.01	0.01	1.01	0.09
	CQR LASSO	0.62	0.01	0.63	0.08	0.27	0.01	0.26	0.06
IV	QR	4.84	1.03	4.68	0.96	8.25	8.53	7.78	2.46
	QR LASSO	1.88	0.18	1.85	0.37	0.65	0.06	0.60	0.27
	CQR	4.25	0.81	4.13	0.89	7.56	5.97	7.38	2.32
	CQR LASSO	1.79	0.18	1.76	0.37	0.59	0.12	0.47	0.21

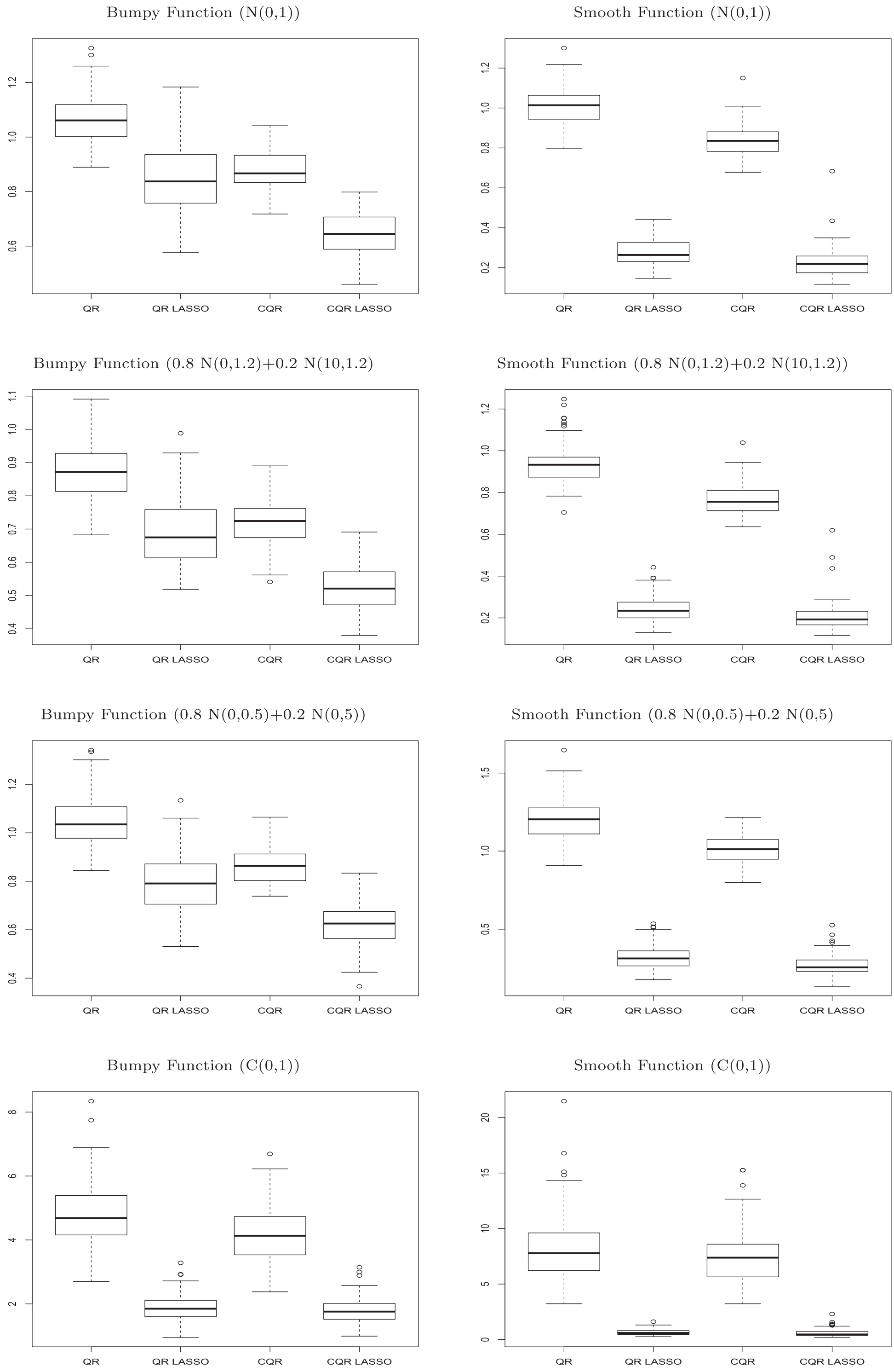


Figure 1. Boxplots of MSE for four distributions I–IV; bumpy (left) and smooth (right) functions.

Table 2. ISE with SNR situation 2.

		Bumpy Function				Smooth Function			
Dist	Method	Mean	Var	Med	MAD	Mean	Var	Med	MAD
I	QR	6.05	1.36	5.91	0.98	5.85	1.50	5.83	1.30
	QR LASSO	2.55	0.53	2.39	0.64	0.16	0.02	0.11	0.04
	CQR	4.98	1.06	4.93	1.06	4.44	0.89	4.28	1.01
	CQR LASSO	2.02	0.27	1.95	0.40	0.13	0.04	0.07	0.03
II	QR	4.81	0.73	4.71	0.88	5.39	1.35	5.44	1.28
	QR LASSO	2.20	0.50	2.06	0.45	0.11	0.01	0.09	0.03
	CQR	4.01	0.59	4.10	0.59	4.44	0.89	4.28	1.01
	CQR LASSO	1.64	0.24	1.50	0.36	0.13	0.07	0.08	0.03
III	QR	6.13	1.73	5.83	1.23	6.85	1.92	6.75	0.98
	QR LASSO	2.42	0.50	2.33	0.57	0.14	0.01	0.11	0.04
	CQR	5.06	0.98	4.96	1.10	5.74	1.60	5.51	0.80
	CQR LASSO	1.95	0.26	1.90	0.46	0.12	0.01	0.10	0.03
IV	QR	28.19	65.60	27.42	7.37	46.09	358.65	42.81	18.02
	QR LASSO	5.05	2.01	4.96	1.32	0.30	0.09	0.19	0.07
	CQR	24.98	58.76	24.60	7.23	42.56	267.42	41.29	15.31
	CQR LASSO	4.93	3.07	4.74	1.55	0.48	0.55	0.15	0.09

and ISE for the functional coefficients as

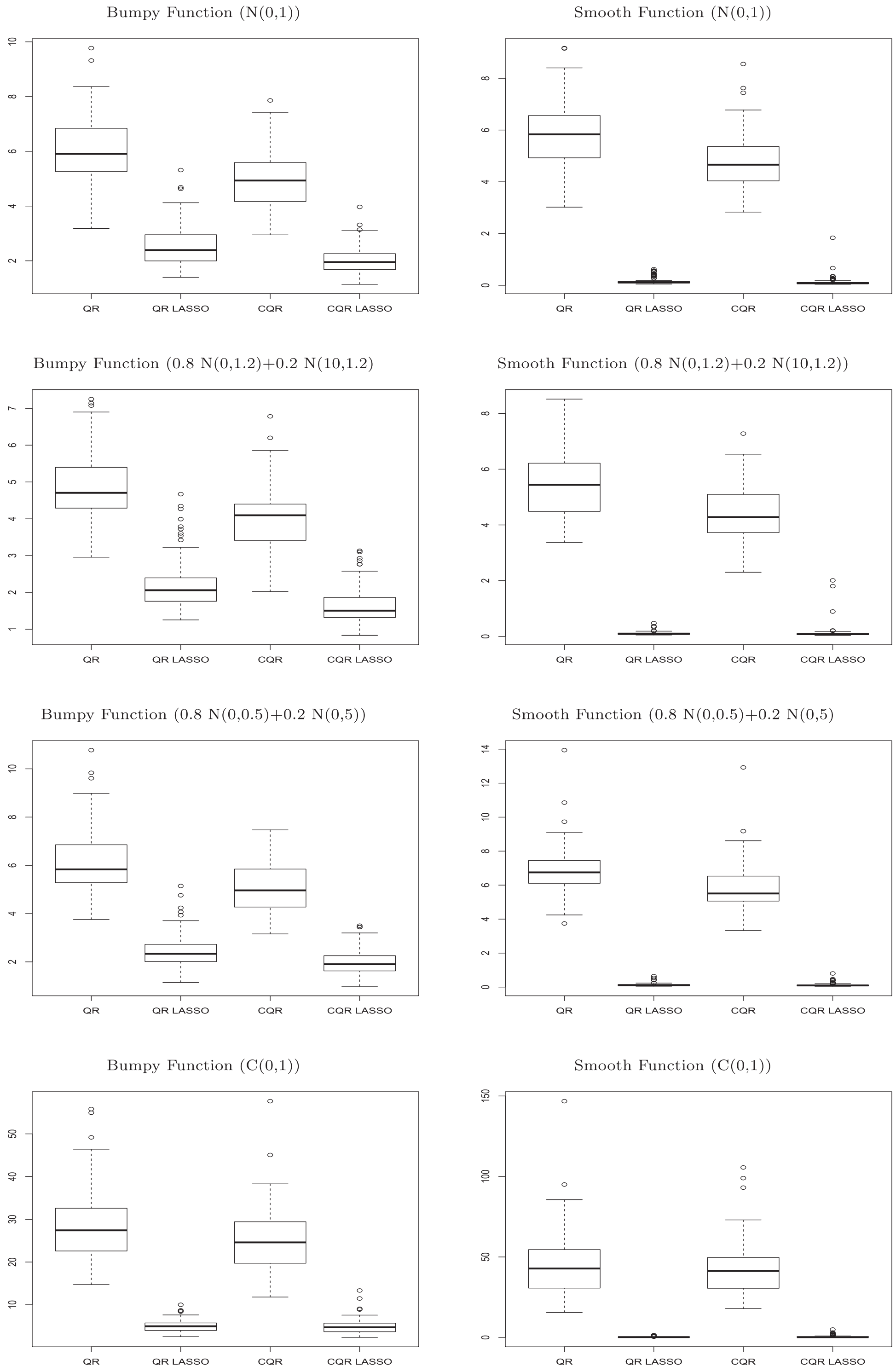
$$MSE = \frac{1}{n}\int_0^1 (\hat{\beta}(t) - \beta(t))^{\mathrm{T}} X^{\mathrm{T}}(t) X(t) (\hat{\beta}(t) - \beta(t))\,\mathrm{d}t,$$

$$ISE = \frac{1}{p}\int_0^1 (\hat{\beta}(t) - \beta(t))^{\mathrm{T}} (\hat{\beta}(t) - \beta(t))\,\mathrm{d}t.$$

To compare performance of the four methods, namely QR ($\tau = 0.5$), CQR, QR-LASSO ($\tau = 0.5$), and CQR-LASSO, we list the mean, variance, median and mean absolute deviation (MAD) of MSE and ISE for bumpy and smooth functions, as well as their boxplots. To save the space, we only report results from SNR situation 2. For the other two SNR situations, the results are both in favour more or less of our methods with LASSO penalty.

As shown in Table 1 and Figure 1, wavelet-based LASSO in functional linear quantile regression provides better estimation and prediction than wavelet-based functional linear quantile regression without a LASSO penalty. When errors follow normal distributions (including standard normal distribution and mixed normal distribution), medians of MSE with a LASSO penalty are about 0.2–0.4 less than those without a LASSO penalty in the bumpy function, and 0.7 less in the smooth function. Comparably, when errors are Cauchy distributed, the prediction performance has more obvious difference. To be specific, medians of MSE of LASSO methods are dropped by up to 90%. In addition, CQR has smaller MSE no matter with or without a LASSO penalty.

Table 2 and Figure 2 indicate the estimation performance. We observe that a LASSO penalty largely improves the estimation of the four methods in the bumpy function. The means and medians of ISE in LASSO methods are about 40% and 70% less than unpenalized methods with normal distributed errors. While in the smooth function, the means and medians of ISE in LASSO methods is much smaller, only about1% to 3% of those without a LASSO penalty. Moreover, when errors follow Cauchy distribution, means and medians of ISE without a LASSO penalty become very big, meaning that they almost lose their estimation ability and the results are not reliable. However, in LASSO methods, the values of ISE are still at the similar level of the values with normal errors, which confirms that OR



Bumpy Function (N(0,1))

Smooth Function (N(0,1))

Bumpy Function (0.8 N(0,1.2)+0.2 N(10,1.2)

Smooth Function (0.8 N(0,1.2)+0.2 N(10,1.2))

Bumpy Function (0.8 N(0,0.5)+0.2 N(0,5))

Smooth Function (0.8 N(0,0.5)+0.2 N(0,5)

Bumpy Function (C(0,1))

Smooth Function (C(0,1))

Figure 2. Boxplots of ISE for four distributions I–IV; bumpy (left) and smooth (right) functions.

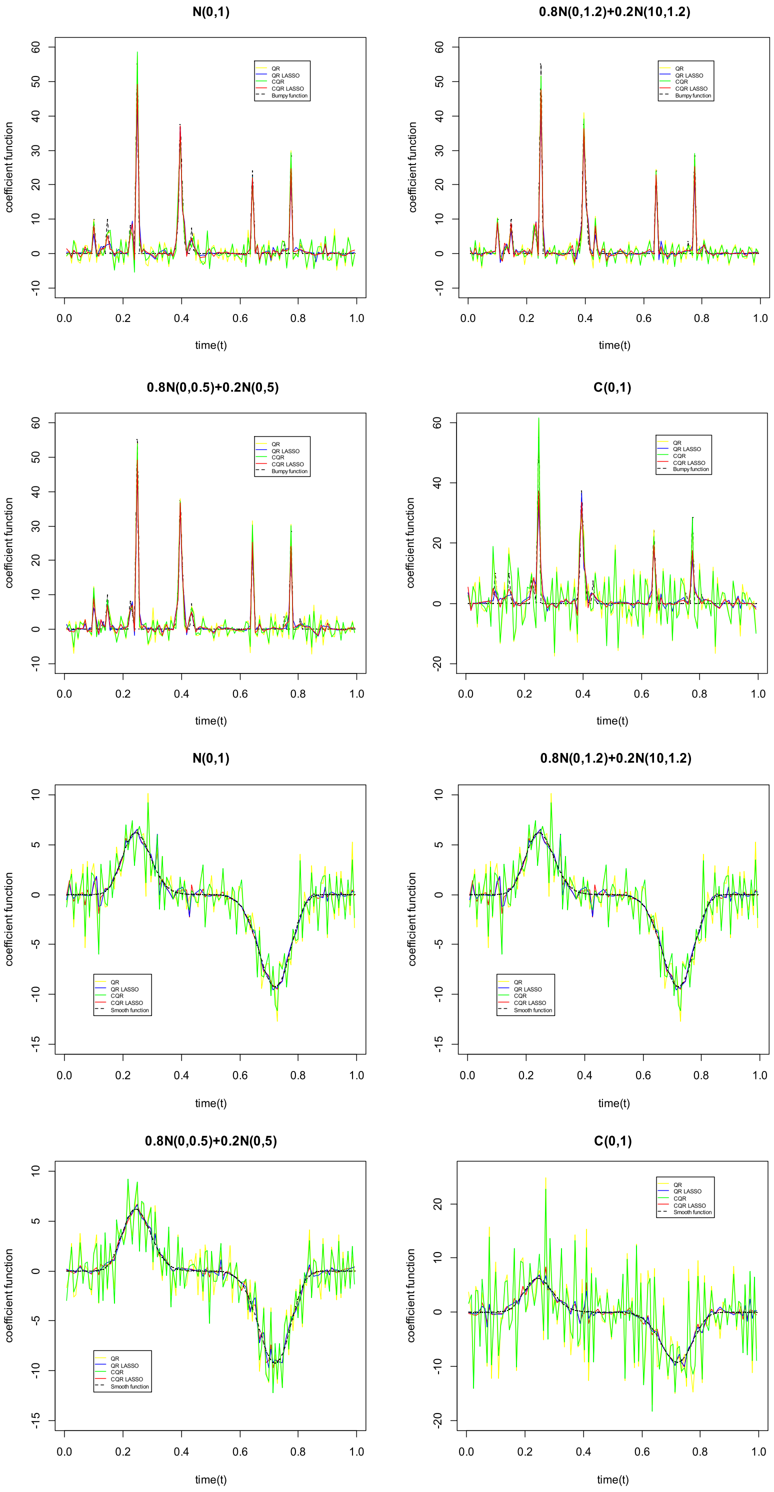


Figure 3. Estimate coefficient function in bumpy (first two rows) and smooth function (last two rows).

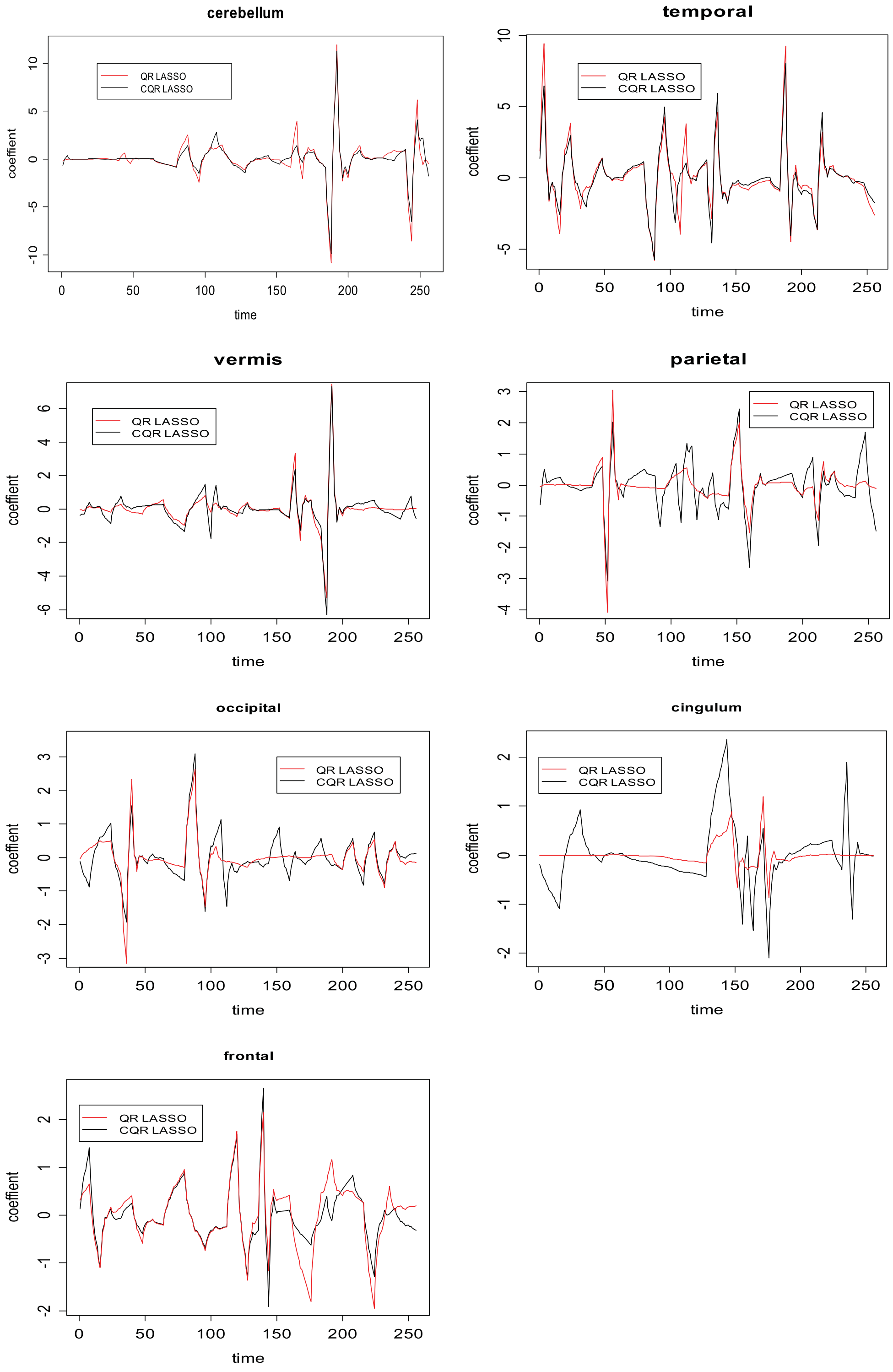


Figure 4. Estimate coefficient function for seven functional covariates.

and CQR with LASSO penalty performs more efficiently and stably with infinite-variance errors. In addition, CQR still performs better in most estimations than QR with or without a LASSO penalty. Figure 3 display the estimated coefficient functions of the four methods and overlaid with the true coefficients in SNR situation 2. LASSO methods give much closer estimates in both bumpy function and smooth function. In particular, for smooth function with Cauchy distribution errors, QR and CQR estimates with wavelet transform are very far from the true coefficients.

5.3. Real data example

As an illustration, we use our methods to analyse a data set on Attention deficit hyperactivity disorder (ADHD) from NYU site of ADHS-200 Global Competition, which are available at fcon-1000.projects.nitrc.org/ indi/adhd200. ADHD is the most common childhood psychiatric disorder and may be followed by a lifelong time. The disease symptom includes inattention, excessive motor hyperactivity or restlessness, and poor impulse control [36]. From New York University Child Study Center, atlas are the filtered preprocessed resting state data using the Anatomical Automatic Labelling (AAL) [37]. There are 172 time courses in the filtering and AAL has 116 Regions of Interest (ROIs) fractionated into functional space using nearest-neighbour interpolation [4].

In this article, we investigate the association between severity of ADHD and several selected regions while adjusting for some demographic and other covariates. The ADHD index (Conners Parent Rating Scale-Revised, long version (CPRS-LV)) serves as a continuous response, which refers to the behaviour score and reflects the severity of the ADHD disease. Seven regions of the brain are separately considered as functional covariates: the cerebellum, temporal, vermis, parietal, occipital, cingulum and frontal. Those regions have been previously found to be related to ADHD. All of them include as least four ROIs and we take average of the ROIs. The other scalar covariates, we consider, include gender, age, handedness, diagnosis status, medication status, Verbal IQ, Performance IQ and Full4 IQ. The raw data has 222 subjects. After data cleaning and quality control, we have 142 subjects left with. The 172 time courses are linearly interpolated to 256 equally spaced points for conveniently applying wavelet transformed.

We use wavelet-based LASSO in functional linear quantile regression on the ADHD data NYU site and estimate the functional covariates separately. Figure 4 shows the estimated functional coefficients with properly chosen tuning parameters. We conclude that the estimates of wavelet-based LASSO in composite quantile regression are consistent with that in quantile regression. In particular, the results are quite close in cerebellum, temporal, vermis and frontal. For parietal, occipital, and cingulum, the estimated coefficients in composite quantile regression have more small sharps. In addition, the estimated functional coefficients of the cerebellum have a similar pattern with that of vermis but in different magnitude. It means these two covariates may have similar effects in pattern on the disease but in different magnitude.

6. Conclusion

This article proposes an efficient approach to estimate functional coefficients in quantile regression. After applying the wavelet transform on functional data, we combine the LASSO method with quantile regression and composite quantile regression to select wavelet bases and estimate functional coefficients. Quantile regression and composite quantile regression have the advantage of resistant to the outliers and heavy tails. Meanwhile, LASSO methods effectively reduce the number of wavelet bases to enhance the strongest effects. Furthermore, it is more thorough to consider both functional covariates and scalar covariates in the model with only wavelet-based functional variables constrained in a LASSO penalty. In addition, the ADMM algorithm is derived to efficiently solve the optimization problem of quantile regression with a LASSO penalty.

Simulation studies show that wavelet-based quantile QR and CQR with a LASSO penalty are capable of providing estimates with less estimation errors as well as less prediction errors, even with infinite-variance errors. In addition, composite quantile regression performs better than quantile regression in functional data analysis in most situations, no matter with a LASSO penalty or without. We applied the proposed methods to the ADHD data at NYU site from ADHD-200 Global competition to study the association between ADHD index and seven regions of the brain. We observe that estimated coefficients from wavelet-based LASSO in quantile regression are similar to coefficients from wavelet-based LASSO in composite quantile regression, especially for cerebellum, temporal, vermis and frontal.

For the wavelet-based LASSO methods in functional linear quantile regression, there are several factors that may influence the performance. First, different wavelet bases may be chosen according to the nature of the application. In this article, we chose the Daubechies' family wavelet because of its good property in localization [3]. However, some other wavelet bases may have better ability to sparsely represent. Second, we derive the ADMM algorithm to solve the convex optimization problem, but the efficiency may be affected by the dimension of the finite sample. When the number of variables is bigger than the number of subjects, the results may be influenced. Third, the way to choose tunning parameters impacts the performance of the methods. In this article, five-fold cross validation is used to select the parameters, but AIC, BIC, and other methods can also be used for parameter selection. We tried AIC and BIC, the results are no better than cross-validation.

There are some directions that deserved further research. More complex structure can be imposed on the scalar effect part in (1), say imposing the single index structure [38] to allow more flexibility. For the LASSO penalty, we cannot make sure every coefficient equally penalized in the penalty, so adaptive LASSO [20] having the oracle properties may be a more appropriate way to shrink the coefficients. In our set up, there is only one functional covariate included in each model. However, to improve efficiency, multiple functional covariates need to be considered in one model. The group-LASSO method [39] may be an appropriate way to solve such problems. Moreover, if some of the functional covariates use common wavelet bases in the wavelet transform, they can be grouped. Sparse-group LASSO [40] can promote the desired sparsity pattern and regularize nicely within each group. However, the asymptotic properties and algorithm would need more investigation. This is our current research and we will discuss it in another manuscript.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

Drs Kong, Jiang and Heo acknowledge the support from the National Sciences and Engineering Research Council of Canada (NSERC) . Dr Kong's research is also partially supported by the Canadian Statistical Sciences Institute (CANSSI). Dr Zhou's research is partially supported by the China Postdoctoral Science Foundation (No. 2018T110422).

References

[1] Fass L. Imaging and cancer: a review. Mol Oncol. 2008;2(2):115–152.

[2] Friston KJ. Modalities, modes, and models in functional neuroimaging. Science. 2009;326 (5951):399–403.

[3] Zhao Y, Ogden RT, Reiss PT. Wavelet-based LASSO in functional linear regression. J Comput Graph Stat. 2012;21(3):600–617.

[4] Yu D, Kong L, Mizera I. Partial functional linear quantile regression for neuroimaging data analysis. Neurocomputing. 2016;195:74–87.

[5] Koenker R, Bassett G. Regression quantiles. Econometrica. 1978;46:33–50.

[6] Koenker R. Quantile regression. Cambridge: Cambridge University Press; 2005.

[7] Müller H, Stadtmüller U. Generalized functional linear models. Ann Stat. 2005;33:774–805.

[8] Müller H, Yao F. Functional additive models. J Am Stat Assoc. 2008;103:1534–1544.

[9] Delaigle A, Hall P, Apanasovich TV. Weighted least squares methods for prediction in the functional linear model. Electron J Stat. 2009;3:865–885.

[10] Marx BD, Eilers PHC. Generalized linear regression for sampled signals or curves: a p-spline approach. Technometrics. 1999;41:1–13.

[11] Cardot H, Ferraty F, Sarda P. Functional linear model. Stat Probab Lett. 1999;45:11–22.

[12] Reiss PT, Ogden RT. Functional principal component regression and functional partial least squares. J Am Stat Assoc. 2007;102:984–996.

[13] Delaigle A, Hall P. Methodology and theory for partial least squares applied to functional data. Ann Stat. 2012;40:322–352.

[14] Apenteng OO, Ismail NA. The impact of the wavelet propagation distribution on SEIRS modeling with delay. PLoS ONE. 2014;9(6):e98288.

[15] Härdle W, Kerkyacharian G, Picard D, et al. Wavelets: approximation and statistical applications. New York (NY): Springer-Verlag; 1998.

[16] Vidakovic B. Statistical modeling by wavelets. New York (NY): Wiley; 1999.

[17] Tibshirani R. Regression shrinkage and selection via the LASSO. J R Stat Soc B. 1996;58:267–288.

[18] Wang X, Nan B, Zhu J, et al. Regularized 3d functional regression for brain image data via haar wavelets. Ann Appl Stat. 2014;8(2):1045–1064.

[19] Zhao W, Zhang R, Liu J. Sparse group variable selection based on quantile hierarchical LASSO. J Appl Stat. 2014;41(8):1658–1677.

[20] Zou H. The adaptive LASSO and its oracle properties. J Am Stat Assoc. 2006;101(476): 1418–1429.

[21] Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc. 2001;96(456):1348–1360.

[22] Zhang CH, et al. Nearly unbiased variable selection under minimax concave penalty. Ann Stat. 2010;38(2):894–942.

[23] Zou H, Yuan M. Composite quantile regression and the oracle model selection theory. Ann Stat. 2008;36:1108–1126.

[24] Kai B, Li R, Zou H. Local composite quantile regression smoothing: an efficient and safe alternative to local polynomial regression. J R Stat Soc B. 2010;72(1):49–69.

[25] Fan J, Lv J. A selective overview of variable selection in high dimensional feature space. Stat Sin. 2010;20(1):101.

[26] Bradic J, Fan J, Wang W. Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. J R Stat Soc B. 2011;73(3):325–349.

[27] Koenker R, Park BJ. An interior point algorithm for nonlinear quantile regression. J Econom. 1996;71(1):265–283.

[28] Gabay D, Mercier B. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. Comput Math Appl. 1976;2(1):17–40.

[29] Boyd S, Parikh N, Chu E, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. Found Trends Mach Learn. 2011;3(1):1–122.

[30] Hestenes MR. Multiplier and gradient methods. J Optim Theory Appl. 1969;4(5):303–320.

[31] Lin Z, M C, Ma Y. The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices; 2013. ArXiv:1009.5055.

[32] Li X, Zhao T, Yuan X, et al. The flare package for high dimensional linear regression and precision matrix estimation in r. J Mach Learn Res. 2015;16(1):553–557.

[33] Roebuck P. group RUD. rwt: Rice wavelet toolbox wrapper; 2014. R package version 1.0.0; Available from: http://CRAN.R-project.org/package = rwt.

[34] Nason G. wavethresh: Wavelets statistics and transforms.; 2013. R package version 4.6.6; Available from: http://CRAN.R-project.org/package = wavethresh.

[35] Koenker R. quantreg: Quantile regression; 2015. R package version 5.11; Available from: http://CRAN.R-project.org/package = quantreg.

[36] Campbell SB. Handbook of developmental psychopathology. In: Sameroff A.J., Lewis M., Miller S.M., editors. Attention-deficit/hyperactivity disorder. In. Boston: Springer; 2000. p. 383–401.

[37] Tzourio-Mazoyer N, Landeau B, Papathanassiou D, et al. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage. 2002;15(1):273–289.

[38] Tang Q, Kong L. Quantile regression in functional linear semiparametric model. Statistics. 2017;51(6):1342–1358.

[39] Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. J R Stat Soc B. 2006;68(1):49–67.

[40] Simon N, Friedman J, Hastie T, et al. A sparse-group LASSO. J Comput Graph Stat. 2013;22(2):231–245.

[41] Mallat S. A wavelet tour of signal processing. Burlington: Elsevier; 2009.

Appendices

Appendix 1. Proof of Theorem 3.1

Proof: The proof based on articles [3,19]. First, we introduce some notations, the collection $\{\phi_{j_0k}, k = 1, \ldots, 2^{j_0}; \psi_{jk}, \quad j \geq j_0,\ k = 1, \ldots,\ 2^j\}$ is then an orthonormal basis of $L^2[0,1]$. Without loss of generality, the wavelet bases are ordered according to the scales from the coarsest level J_0 to the finest one. Let $V_m := Span\{\varphi_1,\ \ldots,\ \varphi_m\}$ be the space spanned by the first m basis function, for example, if $m = 2^{j_0+t}$ then the collection $\{\phi_{j_0k}, k = 1, \ldots, 2^{j_0};\ \psi_{jk},\ j_0 + t - 1 \geq j \geq j_0,\ k = 1, \ldots,\ 2^j\}$ is the basis of V_m. Let β_m^j be an $m \times 1$ parameter vector with elements $\beta_k = \langle \eta(t), \varphi_k \rangle$. In addition, let η_m be the functions reconstructed from the wavelet coefficients β_m. Here η_m is a linear approximation to η by the first m wavelet coefficients, while $\widehat{\eta_m}$ denotes the functions reconstructed from the wavelet coefficients $\widehat{\beta_m}$ from (7).

By the Parseval's theorem, we have $\|\hat{\eta} - \eta\|_{L_2}^2 = \|\widehat{\beta_m} - \beta_m\|_2^2 + \sum_{t=m+1}^{\infty} {\beta_t}^2$. To derive the convergence rate of $\hat{\eta}$ to η, we bound the error in estimating η_m by $\widehat{\eta_m}$ and the error in approximating η by η_m. By the Theorem 9.5 of [41], the linear approximation error goes to zero as

$$\sum_{t=m+1}^{\infty} {\beta_t}^2 = o(m^{-2q}) \tag{A1}$$

To obtain the result, we will show that for any given $\epsilon > 0$, there exists a constant C such that

$$\Pr\left\{\inf_{\|u\|=C} Q_n((\alpha_{\tau_1}^0, \ldots, \alpha_{\tau_K}^0, \delta^0, \beta^0) + r_n u) > Q_n((\alpha_{\tau_1}^0, \ldots, \alpha_{\tau_K}^0, \delta^0, \beta^0))\right\} \geq 1 - \epsilon \tag{A2}$$

where $r_n = \sqrt{m/n}$. This implies that there exists a local minimizer in the ball $\{(\alpha^0_{\tau_1}, \ldots, \alpha^0_{\tau_K}, \delta^0, \beta^0) + r_n u : \|u\| \leq C\}$ with probability at least $1 - \epsilon$. Hence, there exists a local minimizer such that $\|(\hat{\alpha}_{\tau_1}, \ldots, \hat{\alpha}_{\tau_K}, \hat{\delta}, \hat{\beta}) - (\alpha^0_{\tau_1}, \ldots, \alpha^0_{\tau_K}, \delta^0, \beta^0)\| = O_p(r_n)$. From this, we can also get $|\hat{\alpha}_{\tau_s} - \alpha_{\tau_s}| = O_p(r_n)$ for $k = 1, 2 \ldots K$. For any vector $v = [v_1, \ \ldots, v_K, v_\delta, v_\beta]$ with $\|v\| = C$, we have

$$\begin{aligned}
&Q_n((\alpha^0_{\tau_1}, \ldots, \alpha^0_{\tau_K}, \delta^0, \beta^0) + r_n v) - Q_n((\alpha^0_{\tau_1}, \ldots, \alpha^0_{\tau_K}, \delta^0, \beta^0)) \\
&\quad = L_n((\alpha^0_{\tau_1}, \ldots, \alpha^0_{\tau_K}, \delta^0, \beta^0) + r_n v) - L_n(\alpha^0_{\tau_1}, \ldots, \alpha^0_{\tau_K}, \delta^0, \beta^0) \\
&\quad +P_n(\beta^0 + r_n v_\beta) - P_n(\beta^0)
\end{aligned}$$

By using the Knight's identity,

$$\rho_\tau(u - v) - \rho_\tau(u) = -v\psi_\tau(u) + \int_0^v (I(u \leq t) - I(u \leq 0))\,\mathrm{d}t$$

with $\psi_\tau(u) = \tau - I(u < 0)$, we rewrite

$$\begin{aligned}
I &:= L_n((\alpha^0_{\tau_1}, \ldots, \alpha^0_{\tau_K}, \delta^0, \beta^0) + r_n v) - L_n(\alpha^0_{\tau_1}, \ldots, \alpha^0_{\tau_K}, \delta^0, \beta^0) \\
&= \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n}[\rho_{\tau_k}(e_{ki} - r_n(u_i^{\mathrm{T}} v_\delta + x_i^{\mathrm{T}} v_\beta + v_k)) - \rho_{\tau_k}(e_{ki})] \\
&= \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n}\left\{\int_0^{r_n(u_i^{\mathrm{T}} v_\delta + x_i^{\mathrm{T}} v_\beta + v_k)} (I(e_{ki} \leq t)\right. \\
&\quad \left. -I(e_{ki} \leq 0))\,\mathrm{d}t - [r_n(u_i^{\mathrm{T}} v_\delta + x_i^{\mathrm{T}} v_\beta + v_k)\psi_{\tau_k}(e_{ki})]\right\} \\
&:= I_1 + I_2,
\end{aligned}$$

where $e_{ki} = y_i - u_i^{\mathrm{T}}\delta^0 - x_i^{\mathrm{T}}\beta^0 - \alpha^0_{\tau_k}$, $I_1 = -(1/n)\sum_{k=1}^{K}\sum_{i=1}^{n}[r_n(u_i^{\mathrm{T}} v_\delta + x_i^{\mathrm{T}} v_\beta + v_k)\psi_{\tau_k}(e_{ki})]$, and $I_2 = (1/n)\sum_{k=1}^{K}\sum_{i=1}^{n}\int_0^{r_n(u_i^{\mathrm{T}} v_\delta + x_i^{\mathrm{T}} v_\beta + v_k)}(I(e_{ki} \leq t) - I(e_{ki} \leq 0))\,\mathrm{d}t$.

Note that $e_{ki} = y_i - u_i^{\mathrm{T}}\delta^0 - x_i^{\mathrm{T}}\beta^0 - \alpha^0_{\tau_k} = \epsilon_i - F^{-1}(\tau_k) + o(m^{-2q})$, hence we have $E(\psi_{\tau_k}(e_{ki}) = o(m^{-2q})$. It is easy to check that$|I_1| \leq r_n/n(\sum_{k=1}^{s} \| \sum_{i=1}^{n} \psi_{\tau_k}(e_{ki})[1, C_i]\|)$ and let $w_i = [1, C_i]$;

$$\begin{aligned}
E\left\|\sum_{i=1}^{n}\psi_{\tau_k}(e_{ki})w_i\right\|^2 &= E\left\|\sum_{j=1}^{m+1}\sum_{i=1}^{n}\sum_{l=1}^{n} w_{ij}w_{lj}\psi_{\tau_k}(e_{ki})\psi_{\tau_k}(e_{kl})\right\| \\
&= O_p(nm) + o_p(n^2 m^{1-4q}) \\
&= O_p(nm),
\end{aligned}$$

we have $I_1 \leq O_p((r_n/n)\sqrt{nm}) = O_p(r_n^2)$.

Using the same expression of e_{ki}, we obtain

$$E(I_2) = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n}\int_0^{r_n(u_i^{\mathrm{T}}v_\delta + x_i^{\mathrm{T}}v_\beta + v_k)} (\Pr(e_{ki} \le t) - \Pr(e_{ki} \le 0))\,\mathrm{d}t$$

$$= \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n}\int_0^{r_n(u_i^{\mathrm{T}}v_\delta + x_i^{\mathrm{T}}v_\beta + v_k)} (F(F^{-1}(\tau_k) + o(m^{-2q}) + t)$$

$$- F(F^{-1}(\tau_k) + o(m^{-2q})))\,\mathrm{d}t$$

$$= \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n}\int_0^{r_n(u_i^{\mathrm{T}}v_\delta + x_i^{\mathrm{T}}v_\beta + v_k)} \left(f(F^{-1}(\tau_k) + o(m^{-2q}))t + \frac{f'(\xi)}{2}t^2\right)\,\mathrm{d}t,$$

where ξ lies between $F^{-1}(\tau_k) + o(m^{-2q})$ and $F^{-1}(\tau_k) + o(m^{-2q}) + r_n(u_i^{\mathrm{T}}v_\delta + x_i^{\mathrm{T}}v_\beta + v_k)$.

Since there exists M such that $\|C_i\|_2^2 < M$, we have

$$\max_{1\le i\le n} |r_n({u_i}^{\mathrm{T}}v_\delta + {x_i}^{\mathrm{T}}v_\beta + v_k)| \to 0$$

$$E(I_2) = \frac{1}{2n}r_n^2\sum_{k=1}^{K}[f(F^{-1}(\tau_k) + o(m^{-2q}))(nv_k^2 + (v_\delta, v_\beta)^{\mathrm{T}}\sum_{i=1}^{n} C_i{C_i}^{\mathrm{T}}(v_\delta,\ v_\beta))$$

$$+ o_p(1)(nu_k^2 + (v_\delta, v_\beta)^{\mathrm{T}}\sum_{i=1}^{n} C_i{C_i}^{\mathrm{T}}(v_\delta, v_\beta)^{\mathrm{T}})].$$

Next, we will consider II:$= P_n(\beta^0 + r_nv_\beta) - P_n(\beta^0)$. since $r_n \to 0$, for $\|v\| \le C$ we have

$$|\beta^0 + r_nv_\beta| - |\beta^0| \le |r_nv_\beta|.$$

Therefore,

$$\begin{aligned} P_n(\beta^0 + r_nv_\beta) - P_n(\beta^0) &\le \lambda r_n|v_\beta|_1 \\ &\le \lambda r_n\sqrt{m}\|v_\beta\|_2 \\ &= O_p(r_n^2\|v_\beta\|_2). \end{aligned}$$

Since II is bounded by $r_n^2\|v_\beta\|_2$, we can choose a C such that the II is dominated by the term I_2 on $\|v\| = C$ uniformly. Therefore, we obtain

$$Q_n((\alpha^0_{\tau_1},\ldots,\alpha^0_{\tau_K},\delta^0,\beta^0) + r_nv) - Q_n((\alpha^0_{\tau_1},\ldots,\alpha^0_{\tau_K},\delta^0,\beta^0)) > 0$$

holds uniformly on $\|u\| = C$. This is the complete proof. ■

Appendix 2. Proof of Theorem 3.2

Proof: By the Cauchy–Schwarz inequality, we have

$$\begin{aligned} \left|\int_0^1 z(t)\eta(t)\,\mathrm{d}t - \int_0^1 z(t)\hat{\eta}(t)\,\mathrm{d}t\right| &\le \int_0^1 |z(t)||\eta(t) - \hat{\eta}(t)|\,\mathrm{d}t \\ &\le \left[\int_0^1 |z(t)|^2\,\mathrm{d}t\int_0^1 |\eta(t) - \hat{\eta}(t)|^2\,\mathrm{d}t\right]^{1/2} \\ &= O_p\left(\sqrt{\frac{m}{n}}\right) + o_p\left(\frac{1}{m^q}\right). \end{aligned}$$

Since $F^{-1}(\tau) = 0$, we obtain$|\alpha_\tau + \boldsymbol{u}^{\mathrm{T}}\boldsymbol{\delta} - \hat{\alpha}_t - \boldsymbol{u}^{\mathrm{T}}\hat{\boldsymbol{\delta}}| \le O_p(\sqrt{m/n}) + o_p(1/m^q)$ by the above proof. Therefore, we obtain the result. ■